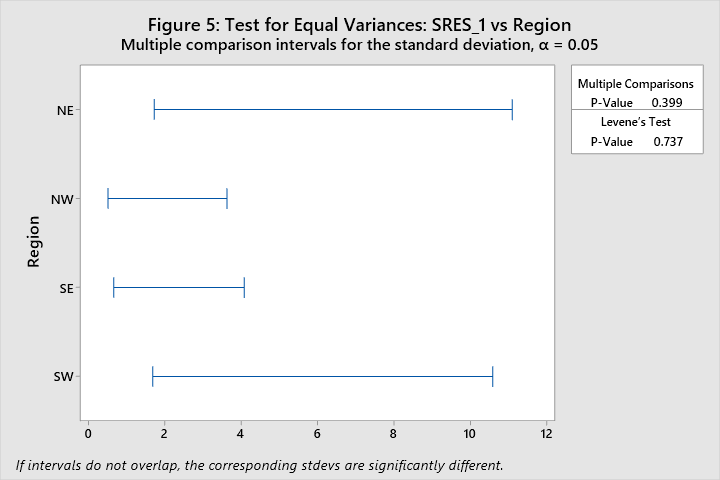
STAT 8120 – Applied Experimental Design



Lab 2 Report – Due 1/26/2020

Connor Armstrong

*The purpose of this report is to fulfill the requirement for Module 2 Lab, according to the supplied lab documentation, S8120Lab2d021318.pdf. SAS and Minitab are utilized in order to output some random and normally distributed data which will we analyzed in this report. These programs are also used as analytical tools to address the questions in the lab document.*

***Setting:*** *This lab asks students to analyze one-way Analysis of Variance (ANOVA) data. Different testing situations will be considered. Students will generate data having varying characteristics. Several groups or treatments are considered. These groups will have differing means, standard deviations or underlying distributions. Knowing the underlying group characteristics, students will attempt to use Minitab and/or ANOVA tools to detect real-world situations. Other statistical software packages may be used to perform the same tasks. Students are encouraged to ask additional “what if” questions and simulate their own data that can then be analyzed.*

***Questions:*** *Perform a one-way ANOVA on given groups each distributed as N(μ, σ2). [Note Minitab has σ NOT σ2 as input in simulations.] Generate each group in a column [label as “N(20,4)” etc.]. Use stacked data in GLM > Fit GLM > select std. residuals > select response, factor Group > Graph > 4 in 1 > ok. Perform in order: an ANOVA and test H0 for equal group means, Residual Analysis for assumptions (i)-(iii), and when F-test is significant Tukey’s pairwise comparison. Compare known population differences in means and variances with the analysis conclusions.*

1. *\*Consider n=20 with groups N(20, 4), N(21,4) and N(22, 4). [Label as N(20,4)20, etc.]*

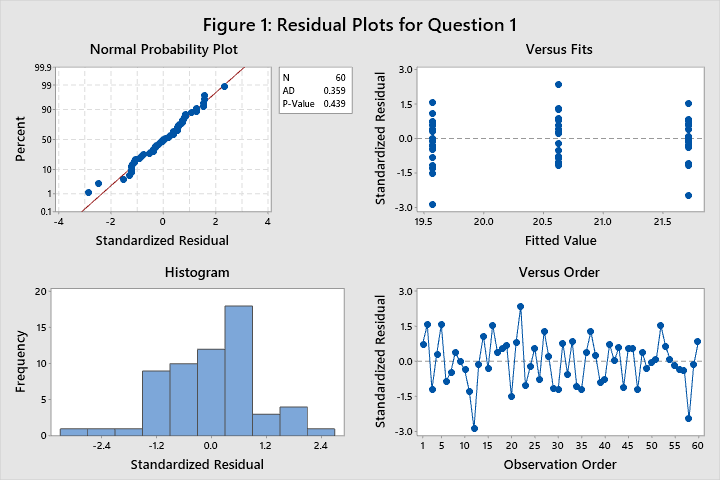
*\*If the ANOVA F-ratio is NOT significant or Residual Analysis fails, re-simulate the data. Do not transform.*

The groups were simulated using Minitab. See below for the ANOVA table:

**Table 1: Analysis of Variance where n=20**

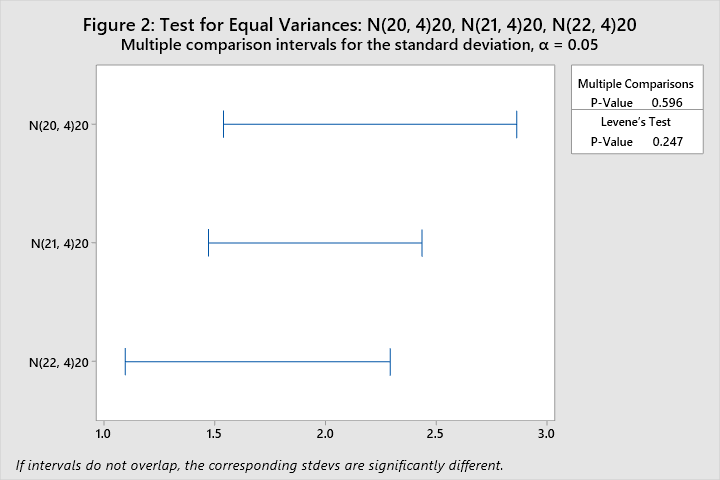
|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| **Source** | **DF** | **Adj SS** | **Adj MS** | **F-Value** | **P-Value** |
| Subscripts | 2 | 46.08 | 23.038 | 7.45 | 0.001 |
| Error | 57 | 176.21 | 3.091 |  |  |
| Total | 59 | 222.29 |  |  |  |

Given a p-value of 0.001, at a significance level of 0.05, there is sufficient evidence to reject the null hypothesis that the means of the 3 groups are equal. In order to accept this conclusion, it is necessary to validate the basic ANOVA Residual Analysis assumptions (i)-(iii). See Figure 1 below for the residual plots for this data.

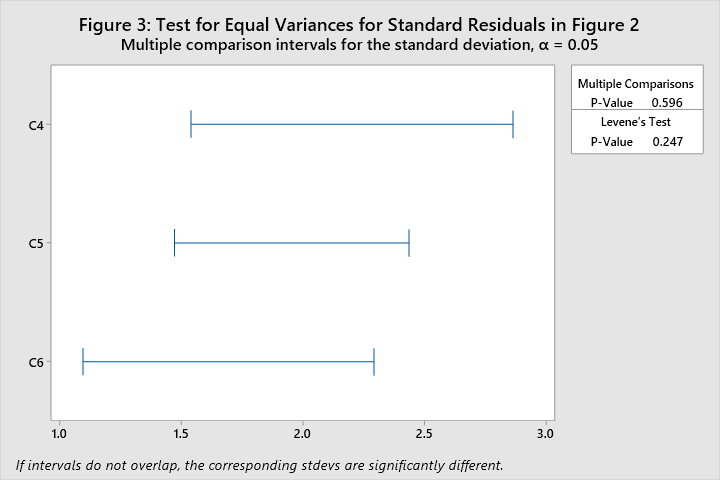


**Assumption 1, Normality** – The probability plot of the residuals is shown in figure 1. The data passed the Anderson-Darling test for normality with a p-value of 0.439. Therefore, the data is normally distributed with a confidence level of 95%.

**Assumption 2, Homogeneity of Variance –** Upon inspection of figure 2 below, the variances of the 3 groups pass Levene’s Test of homogeneity with a p-value of 0.247.



The plot above shows demonstrates the same amount of deviation as the plot below, which is a plot of the variances of the standard residuals of the data for question 1.



**Assumption 3, Outliers** – Figure 1, Versus fits does not show any outliers beyond +/- 3 sigma.

The following table is Tukey Pairwise Comparison Minitab output for the simulated data.

**Table 2: Tukey Pairwise Comparisons: Subscripts where n=20**

**Grouping Information Using the Tukey Method and 95% Confidence**

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **Subscripts** | **N** | **Mean** | **Grouping** | |
| N(22, 4)20 | 20 | 21.7174 | A |  |
| N(21, 4)20 | 20 | 20.6263 | A | B |
| N(20, 4)20 | 20 | 19.5710 |  | B |

*Means that do not share a letter are significantly different.*

The means for groups N(22, 4)20 and N(20, 4)20 are different, and the mean for the group N(21, 4)20 is not significantly different from the other two. This is a reasonable conclusion, given that we generated random data from 3 different normal distributions and the two distributions which were farther apart were given different “Grouping” designations in the Tukey model.

It is interesting to note that the group N(21, 4)20 was not determined to be different from the other two groups, even while it is known that the three means are in fact different. This is demonstrative of the limitations of ANOVA testing, and the sensitivity of Tukey method.

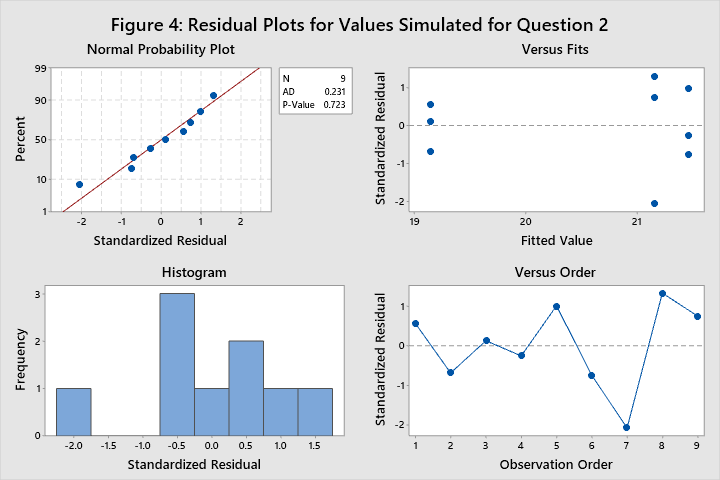
1. *+Consider same situation as problem (1), but let n=3 for each group. [Label as N(20,4)3, etc.]*

The data were simulated and processed using the General Linear Model function in Minitab. See output below:

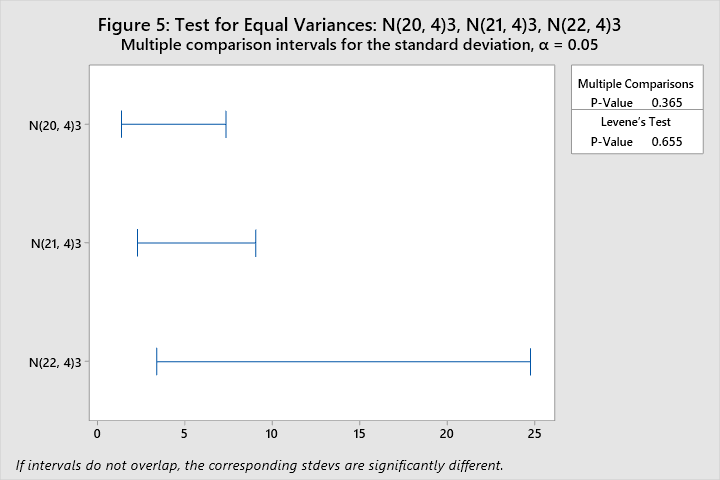
**Table 3: Analysis of Variance where n=3**

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| **Source** | **DF** | **Adj SS** | **Adj MS** | **F-Value** | **P-Value** |
| Distribution | 2 | 9.526 | 4.763 | 0.57 | 0.595 |
| Error | 6 | 50.488 | 8.415 |  |  |
| Total | 8 | 60.014 |  |  |  |

Given a p-value of 0.595, at a significance level of 0.05, there is not sufficient evidence to reject the null hypothesis that the means of the 3 groups are equal. In other words, the data does not support the alternate hypothesis that the means are different. It is likely that this is due to small sample sizes and relatively large variation in the groups given the means are close to one another. A larger sample size would provide more evidence that the means are in fact different for the 3 populations.



**Assumption verification**: The residuals are normal with p-value 0.723 of the A.D. test. There are no outliers beyond 3 standard deviations. Levene’s test demonstrated in Figure 5 below indicates that the variances are likely equal.



1. *\*Consider n=10 with 5 groups A-N(20, 1), B-N(20, 1), C-N( 21,1), D-N(22,4), E-N(23, 9)*

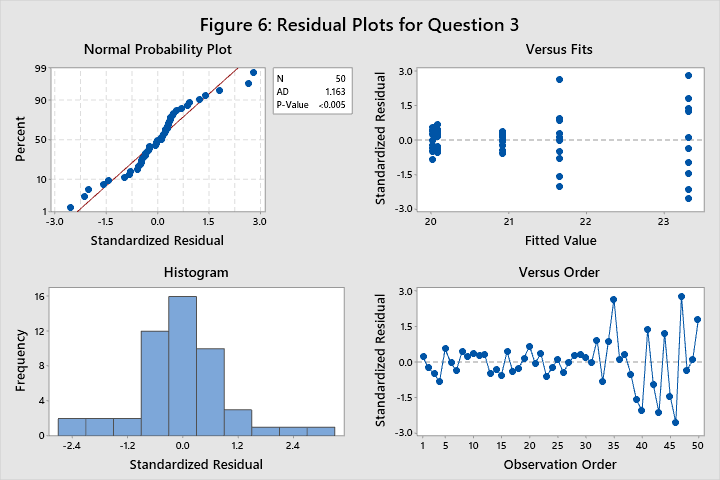
The data were simulated and processed using the General Linear Model function in Minitab. See output below:

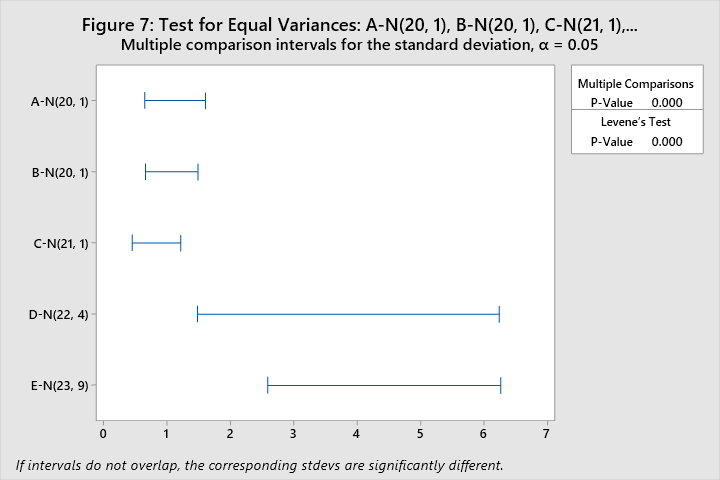
**Table 4: Analysis of Variance where n=10 and Variances are different**

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| **Source** | **DF** | **Adj SS** | **Adj MS** | **F-Value** | **P-Value** |
| Subscripts | 4 | 74.35 | 18.588 | 4.56 | 0.004 |
| Error | 45 | 183.56 | 4.079 |  |  |
| Total | 49 | 257.91 |  |  |  |

With a p-value of 0.004, at a significance level of 0.05, there is sufficient evidence to reject the null hypothesis that the means are not equivalent with a confidence level of 95%, subject to the validity of the assumptions.

**Assumption verification**: The data does not appear to be normal, given the p-value of the A.D. test is <0.005. There are no outliers beyond +/- 3 sigma as shown in the Versus fits plot in Figure 6 below. The data does not meet the homogeneity of variance criteria, as shown in Figure 7. The p-value for Levene’s test of homogeneity is 0.000, which indicates that the variances of the 5 groups are not likely all equivalent, which is a condition to accept the conclusion of the ANOVA.





1. *Summarize your conclusions from (1)-(3) in terms of separation of sample means (σ-distance) and sample size using a bullet point list of key observations.*

* The means of the groups in question 1 were relatively close, given the variation of the populations of the groups (Variance = 4). This did not prevent the ANOVA from detecting some differences, particularly between the two groups which were farthest apart. This was verified using Tukey’s Pairwise Comparison method.
* The means in question 2 were the same as question 1, but given the small sample sizes in question 2, there was not enough data to sufficiently detect the true population behaviors. This led to a large and not significant p-value.
* The differences in variances for the 5 groups in question 3 resulted in a failure to validate the assumption of homogeneity of variance. Given a failure to validate this assumption, among others, the next step would be to attempt to transform the data, which was prohibited in the problem statement.

1. *A manufacturer is evaluating Conductivity for four types of Coating. Use Minitab then repeat with SAS to evaluate coating types and evaluate assumptions. The data appear in Table 5. Based on the conclusions in question (4), what risks are there in your conclusions?*

**Table 5: Data for Question 5**

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Coating Type | Conductivity | | | |
| 1 | 143 | 141 | 150 | 146 |
| 2 | 152 | 149 | 137 | 143 |
| 3 | 134 | 136 | 132 | 127 |
| 4 | 129 | 127 | 132 | 129 |

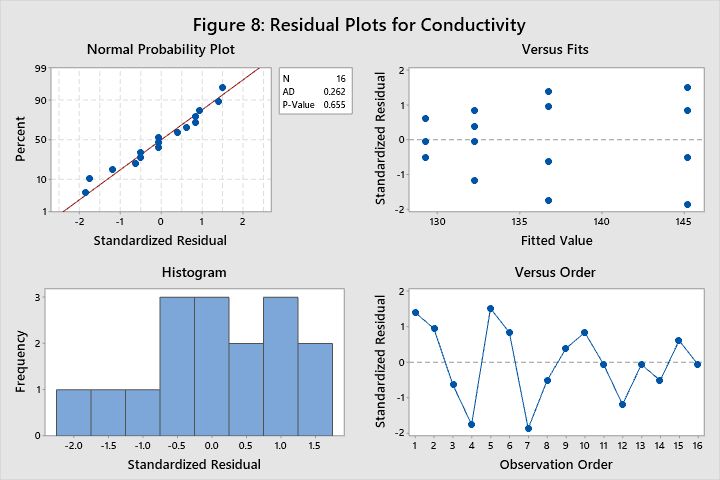
The data were simulated and processed using the General Linear Model function in Minitab. See output below:

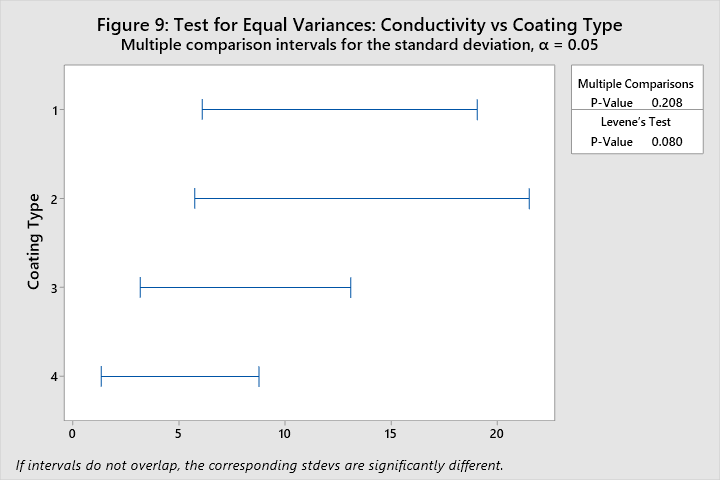
**Table 6: Analysis of Variance for Question 5**

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| **Source** | **DF** | **Adj SS** | **Adj MS** | **F-Value** | **P-Value** |
| Coating Type | 3 | 582.7 | 194.25 | 7.40 | 0.005 |
| Error | 12 | 315.0 | 26.25 |  |  |
| Total | 15 | 897.7 |  |  |  |

With a p-value of 0.005, at a significance level of 0.05, there is sufficient evidence to reject the null hypothesis that the means are not equivalent with a confidence level of 95%, subject to the validity of the assumptions. This conclusion should be questioned, considering the small sample sizes.

**Assumption verification**: The data passed the A.D. test for normality with a p-value of 0.655, as seen in Figure 8 below. There are no outliers beyond +/- 3 sigma as demonstrated in the versus fits plot in Figure 8. The variances passed Levene’s Test for homogeneity with a p-value of 0.080 (See Figure 9). This is a questionable p-value, and considering the implications of the application of the conclusion of the analysis, may warrant additional experimentation or verification.





The data was processed using SAS, evaluating an ANOVA in the same manner as above. The SAS generated ANOVA table (Table 7) is essentially identical to the Minitab generated ANOVA Table 6.

**Table 7: ANOVA for Question 5 Generated in SAS**

| **Source** | **DF** | **Sum of Squares** | **Mean Square** | **F Value** | **Pr > F** |
| --- | --- | --- | --- | --- | --- |
| **Model** | 3 | 582.7500000 | 194.2500000 | 7.40 | 0.0046 |
| **Error** | 12 | 315.0000000 | 26.2500000 |  |  |
| **Corrected Total** | 15 | 897.7500000 |  |  |  |

**Table 8: Levene’s Test for Homogeneity for Question 5 Generated in SAS**

| **Levene's Test for Homogeneity of Conductivity Variance ANOVA of Squared Deviations from Group Means** | | | | | |
| --- | --- | --- | --- | --- | --- |
| **Source** | **DF** | **Sum of Squares** | **Mean Square** | **F Value** | **Pr > F** |
| **Coating\_Type** | 3 | 2636.0 | 878.7 | 2.26 | 0.1335 |
| **Error** | 12 | 4660.8 | 388.4 |  |  |

**Figure 10: Boxplot of Question 5 Data with ANOVA Conclusion**



The assumptions for question 5 have been verified in SAS as well. Table 8 shows the results of Levene’s test, with a p-value of 0.13 > 0.05 demonstrates a failure to reject the notion that the variances are different. Figure 10 shows the boxplots for the different groups in Question 5, and there are no outliers shown in the figure. Table 9 below shows the results of normality tests performed in SAS. The p-value for the Anderson-Darling test is 0.1750, which indicates that the populations from which the data were sampled from is likely normal.

**Table 9: Normality Tests for Question 5 Evaluated in SAS**

| **Tests for Normality** | | | | |
| --- | --- | --- | --- | --- |
| **Test** | **Statistic** | | **p Value** | |
| **Shapiro-Wilk** | **W** | 0.911199 | **Pr < W** | 0.1217 |
| **Kolmogorov-Smirnov** | **D** | 0.158251 | **Pr > D** | >0.1500 |
| **Cramer-von Mises** | **W-Sq** | 0.079844 | **Pr > W-Sq** | 0.2009 |
| **Anderson-Darling** | **A-Sq** | 0.50989 | **Pr > A-Sq** | 0.1750 |

The complete SAS code can be seen in Appendix A.

**APPENDIX A: Complete SAS Code for Question 5**

/\*

STAT 8120 - Module 3 Lab

\*/

libname mod3 "C:\Users\conno\OneDrive\Desktop\STAT 8120 - Applied Experimental Design\Module 3";

**run**;

**proc** **import** datafile = "C:\Users\conno\OneDrive\Desktop\STAT 8120 - Applied Experimental Design\Module 3\q5.csv"

out = mod3.q5

DBMS = csv

Replace;

**run**;

ods rtf;

ods graphics on;

**proc** **anova** data =mod3.q5;

class coating\_type;

model conductivity = coating\_type;

means coating\_type / HOVTEST;

**run**;

**proc** **univariate** data = mod3.q5 normal plot;

var conductivity;

**run**;

ods graphics off;

ods rtf close;

**quit**;