Deliverable 3

Nathaniel Jones

# **Executive Summary**

**Hypothesis:** Is there a relationship between loan default rate and graduation rate quartiles?"

1. VISUALS
2. ARE THE VARIANCES EQUAL?
3. DO THE K-LEVELS FOLLOW A NORMAL DISTRIBUTION?
4. NONPARAMETRIC TEST: K-Sample Permutation
5. POST-HOC PROCEDURES

            A difference in the percent of students that default within 2-years has been found for post-secondary institutions with graduation rates that are less than 25%, 25% to less than 50%, and 50% to 75%. No difference was found between institutions with a graduation rate greater than 75% and institutions with a graduation rate between 50% and 75%. The nonparametric K-sample permutation test was used to detect difference between the graduation rate quartiles for the reason that the data displayed evidence of skewness and did not display evidence suggesting that it was normally distributed. After further analysis, it was found that:

* Institutions with a graduation rate greater than 50% had a median default rate close to 5% while institutions with a graduation rate of 50% or less displayed a median default rate higher than 10%.
* No difference was found between institutions with a graduation rate greater than 75% and institutions with graduation rates greater than 50% but less than or equal to 75%.
* The largest difference in center was found between the second and third quartile. I am 95% confident that the true shift in location is between 0.03824 and 0.04597 for institutions with a graduation rate greater than 25% but less than or equal to 50% and institutions with a graduation rate greater than 50% but less than or equal to 75%.

There were 2,789 observations analyzed in this report. The variables used are defined as:

|  |  |
| --- | --- |
| Variable label | Definition |
| BBRR2\_FED\_UG\_DFLT | The percent of undergraduates that default on their federal loan within 2-years. |
| OMAWDP8\_ALL | The percent of undergraduates that graduated within 8-years. |

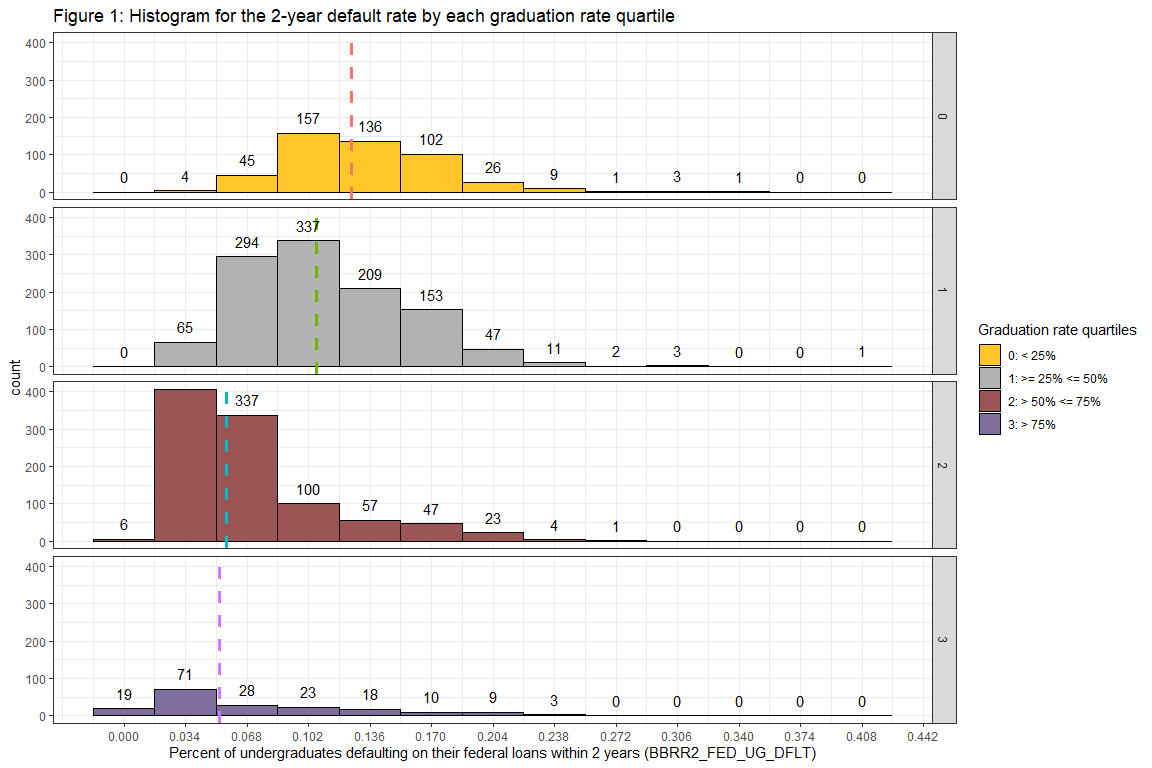
## **Data Manipulation**

A categorical variable was created by splitting the percent of undergraduate students that graduate from a post-secondary institution within 8 years into four equal sized bins. These bins ranged from 0% to less than 25%, 25% to 50%, greater than 50% to 75%, and greater than 75%. They were then given the labels Group 0, Group 1, Group 2, and Group 3, respectively.

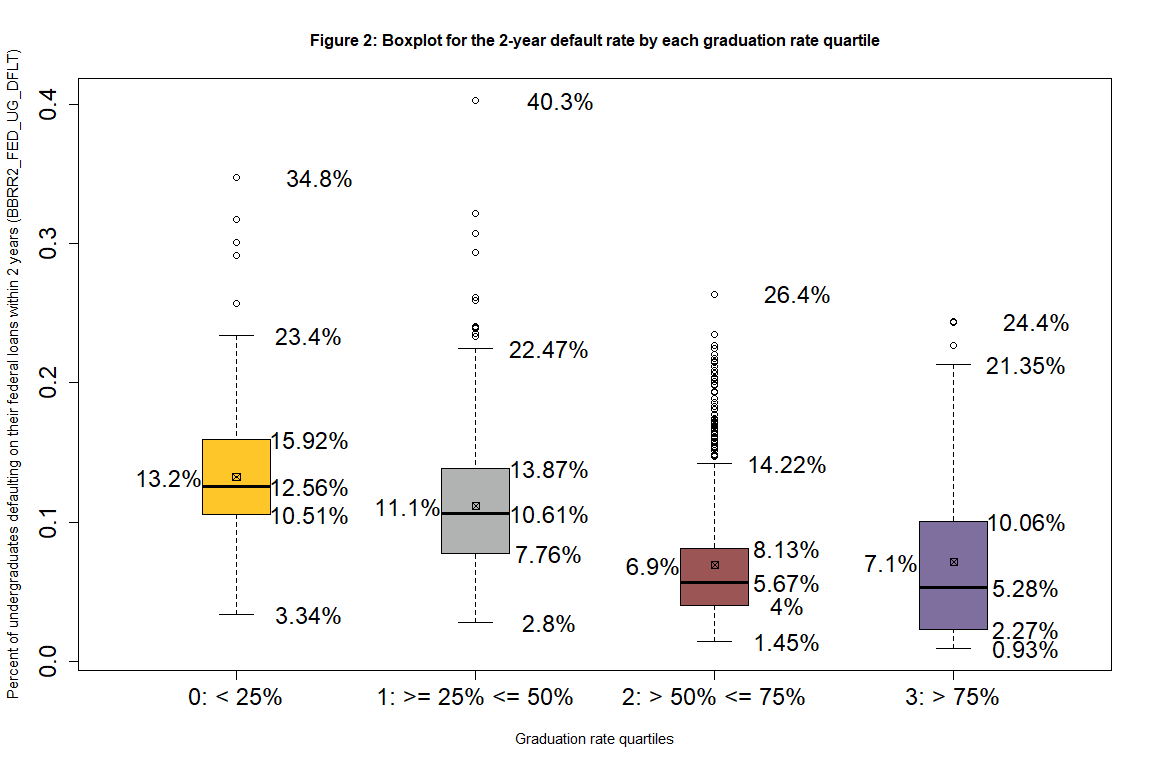
|  |  |  |  |  |
| --- | --- | --- | --- | --- |
|  | Group 0 | Group 1 | Group 2 | Group 3 |
| Percent range | less than 25% | 25% to 50% | greater than 50% to 75% | greater than 75% |

## **VISUALS**

**HISTOGRAM:** Figure 1 displays a histogram for the 2-year default rate at a post secondary institution with a graduation rate that is either < 25% (in gold), >= 25% to <= 50% (in grey), > 50% to <= 75% (in maroon), or > 75% (in purple). Embedded onto each facet is a line that represents the median for that group. From this plot, we can see that the median lines for Group 2, in maroon, and Group 3, in purple, are close to a default rate of 0.05 while Group 0, in gold, and Group 1, in grey, have a median default rate larger than 0.10. There is very little difference in the location of the median line can be seen between Group 2 (grey) and Group 3 (purple) indicating that institutions that graduate more than 50% have similar median default rates, at around 5%. Visually, we can see a significant location difference between Group 1 and Group 2. Specifically, we a see a lower default rate of 5% at institutions with a graduation rate greater 50%, and a higher default rate of 10% at institutions with a graduation rate of 50% or less.



**BOXPLOT:** Figure 2 displays the boxplot for the default rates within each group. One can see from this plot that the variance between each group varies. Group 2, in maroon, contains institutions with a graduation rate greater than 50% and less than or equal to 75% while Group 3, in purple, represents institutions with a graduation rate higher than 75%. The median default rate for institutions in Group 2 is 5.67% while the median default rate for Group 3 is 5.28%. The interquartile range for Group 2’s default rate runs between 4% and 8.13%. Compare this to the interquartile range for Group 3 which runs from 2.27% to 10.06%. These two IQRs are approximately centered in the same location but range over different intervals. Notably, the full range for Group 2 begins at a default rate 1.45% and ends at 26.4% while the full range of default rates for Group 3 was between 0.93% and 24.4%. So, although the interquartile range was smaller for Group 2, the full range of Group 2 was larger than Group 3 which had a larger interquartile range. Group 1, in grey, represents the institutions with a graduation rate greater than or equal to 25% and less than or equal to 50%. The median default rate for this group was 10.61% with a full range going from 2.8% to 40.3%. With this information, an assumption of equal variance seems to be violated by the varying ranges and interquartile ranges for the 4 groups default rates. This will be numerically checked in the next section.



## **ARE THE VARIANCES EQUAL?**

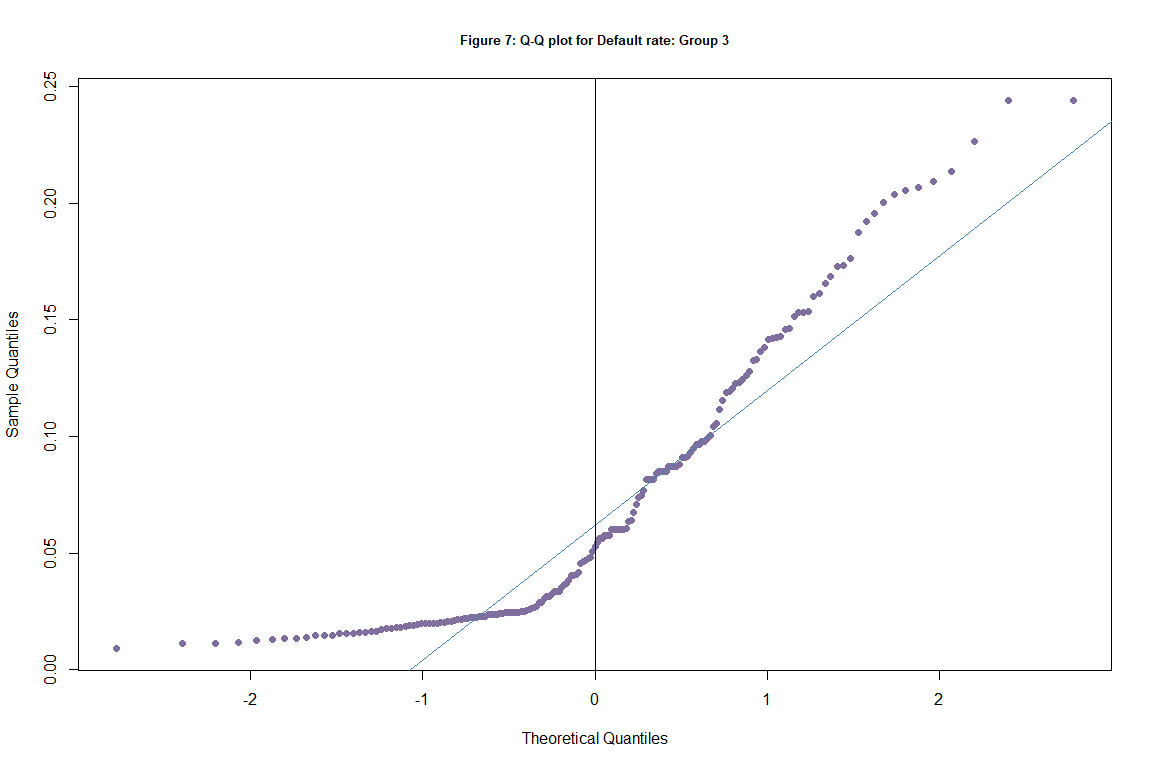
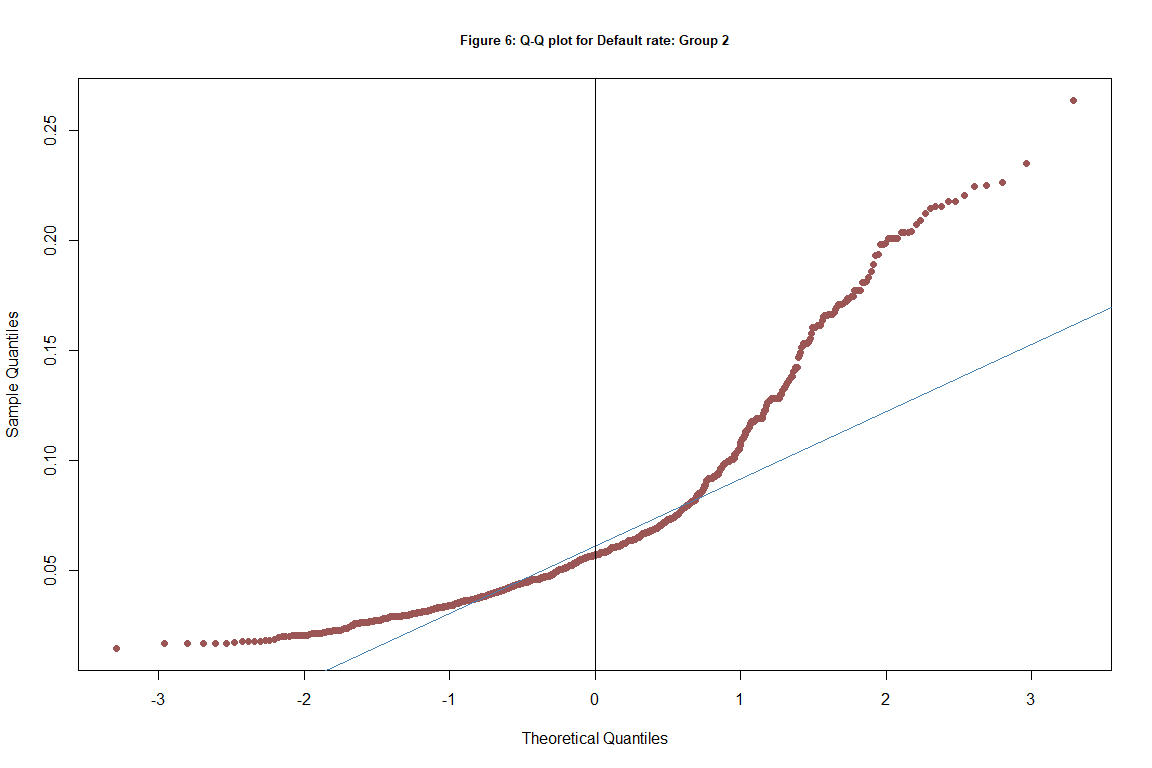
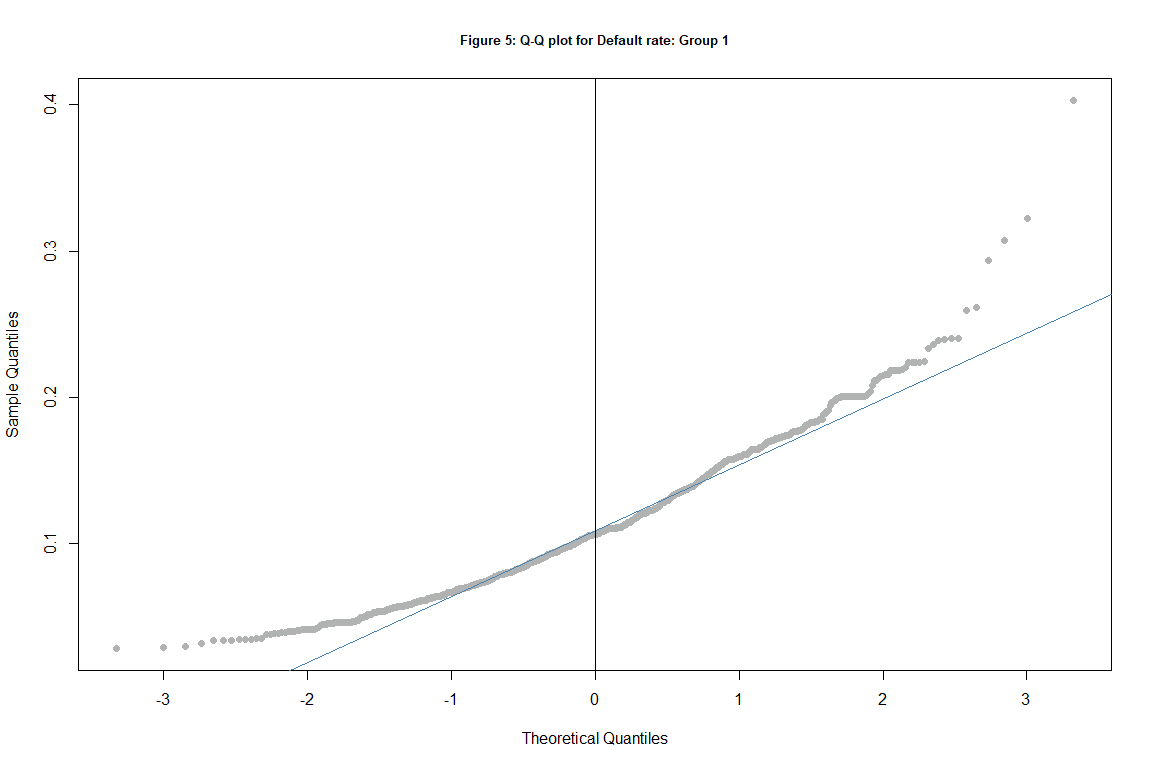
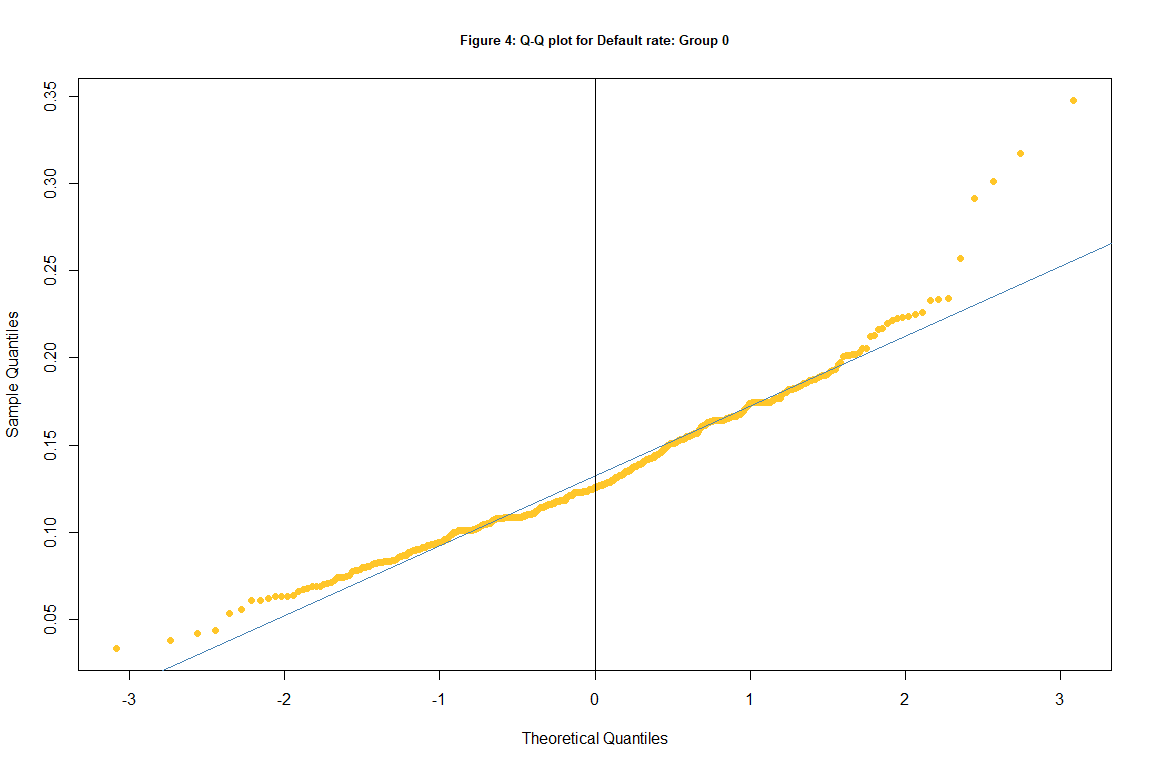
            Yes, the variance among the four groups is approximately equal. Table 1 displays the standard deviation for each of the 4 groups. The largest standard deviation was 0.5872 in Group 3 which corresponds to institutions that have graduation rate greater than 75%. The smallest standard deviation was 0.4134 in Group 0, which corresponds to institutions that have a graduation rate less than 25%. This indicates that the largest standard deviation is only 1.42 times larger than the smallest standard deviation. Since the cutoff point for considering the variance to be approximately equal is 2, our data has NOT violated the assumption of equal variance. Therefore, since we have found that the variances are approximately equal, this report will not use the Kruskal-Wallis. Instead, an investigation into the distribution will be conducted.

Table 1: Standard Deviations

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
|  | Group 0 | Group 1 | Group 2 | Group 3 |
| Percent range | less than 25% | 25% to 50% | greater than 50% to 75% | greater than 75% |
| Default Rate | 0.04136 | 0.04557 | 0.04341 | 0.05872 |

## **DO THE K-LEVELS FOLLOW A NORMAL DISTRIBUTION?**

            No, the four groups display evidence of skewness. Figures 4, 5, 6, and 7 display Q-Q plots for the four groups. For Group 0 in Figure 4, we can see that the left side of the plot mostly follows what would be expected if the data was normally distributed. On the right side, we see the data stray from the normal line, depicted in blue. This suggests that the data is heavily skewed. Furthermore, in Figure 5 we see that Group 1 displays a similar skewness on the right as seen in Group 0 but it is also skewed on the left side, suggesting that the extreme values on both sides of Group 1’s distribution was higher than what was expected if the data was normally distributed. Figures 6 and 7 similarly, but to a much greater degree, show what was found in Figure 5. These figures correspond to Groups 2 and 3 respectively and indicate that the distribution of these groups are heavily skewed on the right side with some skewness on the left side. Since the four groups displayed distributions that may not be normally distributed, a K-sample permutation test will be used over the ANOVA. Although the central limit theorem may be invoked by the sheer size of the sample, the parametric ANOVA test may yield less power than the nonparametric permutation test due to the skewness of the distribution.

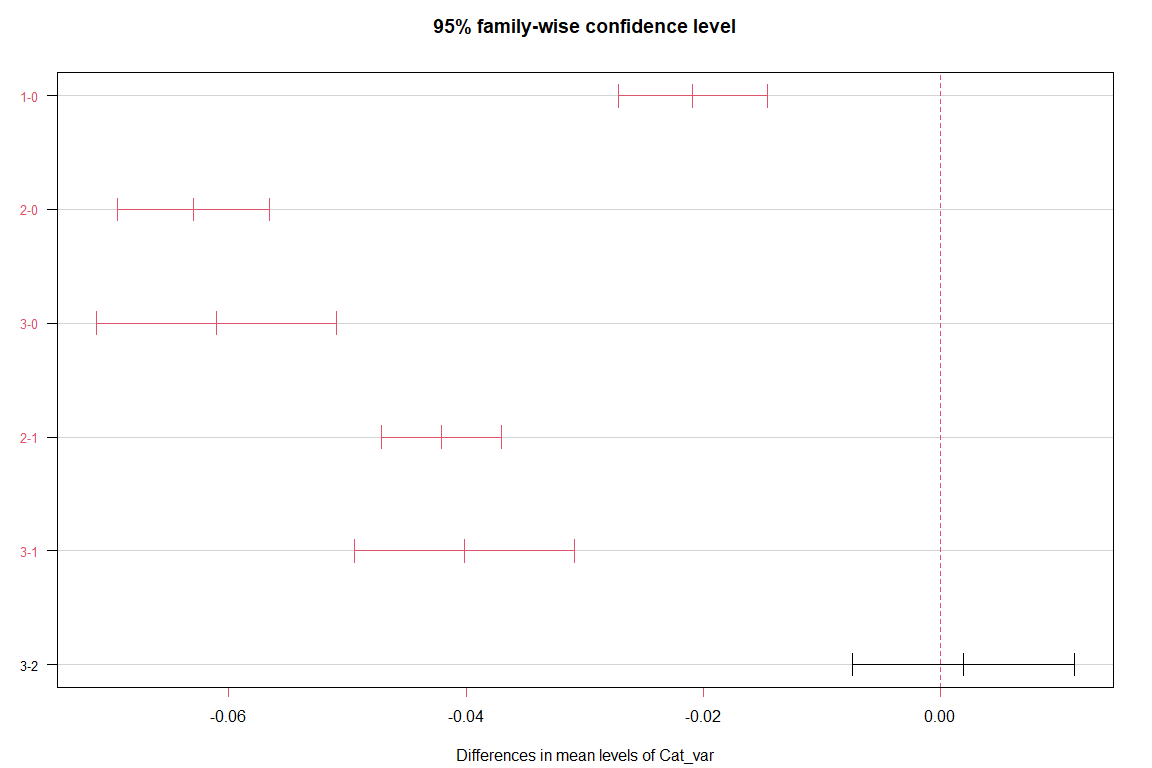


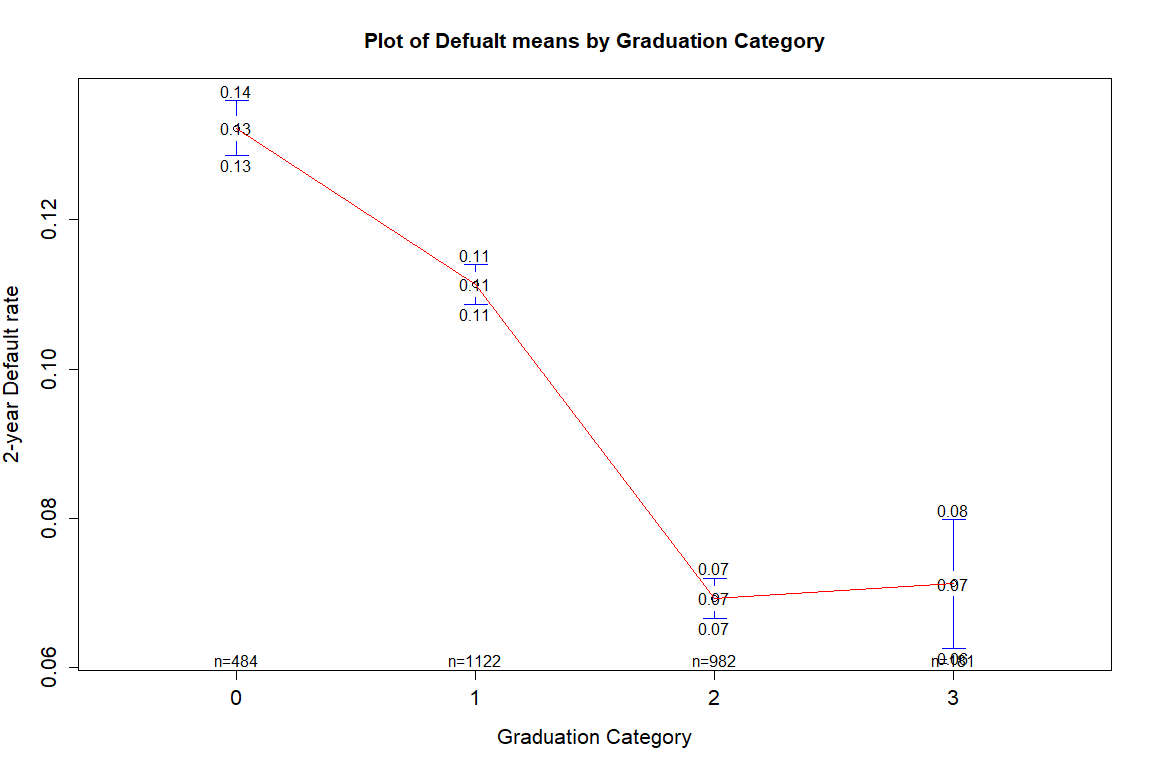
## **NONPARAMETRIC TEST: K-Sample Permutation**

            The results of this test showed that there is a difference in the default rate between each of the groups. In fact, I am 99% confident that the true p-value is between 0 and 0.0005. This indicates that at least one of the distributions has a shift in location that is significantly different from the other groups. Further analysis will be conducted via Tukey’s HSD, Fisher’s LSD, and Bonferroni’s correction in the next section.

## **POST-HOC PROCEDURES**

            Three post-hoc procedures were performed: Tukey’s HSD, Fisher’s LSD, and Bonferroni’s correction. These procedures are used to investigate which groups contained differences and which did not. From the output of the Tukey’s HSD test, we see that Groups 2 and 3 did not significantly differ from each other. Groups 0 and 1 did display a difference with each other as well as with Group 2. Thus, I am 95% confident that the true difference in the centers for Group 1 and 2 is between -0.047 and -0.037. The results of the Tukey’s HSD test can be visualized by the figure below. Fisher’s LSD and Bonferroni’s correction did not disagree with the Tukey’s HSD test. These tests found no difference between Group 2 and 3 but did find a difference between Groups 0, 1, and 2.





## **Other tests**

            The Kruskal-Wallis test was not used since the variance of this sample was found to be approximately equal. The K-sample permutation test was chosen to be used over the ANOVA because each of the groups exhibited heavy skewness.