Deliverable 3

Nathaniel Jones

# **Table of Contents**

EXECUTIVE SUMMARY (Pg. 1-3)  
- **Hypothesis**  
- **Summary of Findings**  
- **Actionable Items**

# **Appendix**

1. DATA MANIPULATION (Pg. 4)
2. ARE THE VARIANCES EQUAL? (Pg. 4)
3. DO THE K-LEVELS FOLLOW A NORMAL DISTRIBUTION? (Pg. 5-7)
4. NONPARAMETRIC TEST: K-Sample Permutation (Pg. 8)
5. POST-HOC PROCEDURES (Pg. 8-9)

# **EXECUTIVE SUMMARY**

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

**Summary of Findings:**

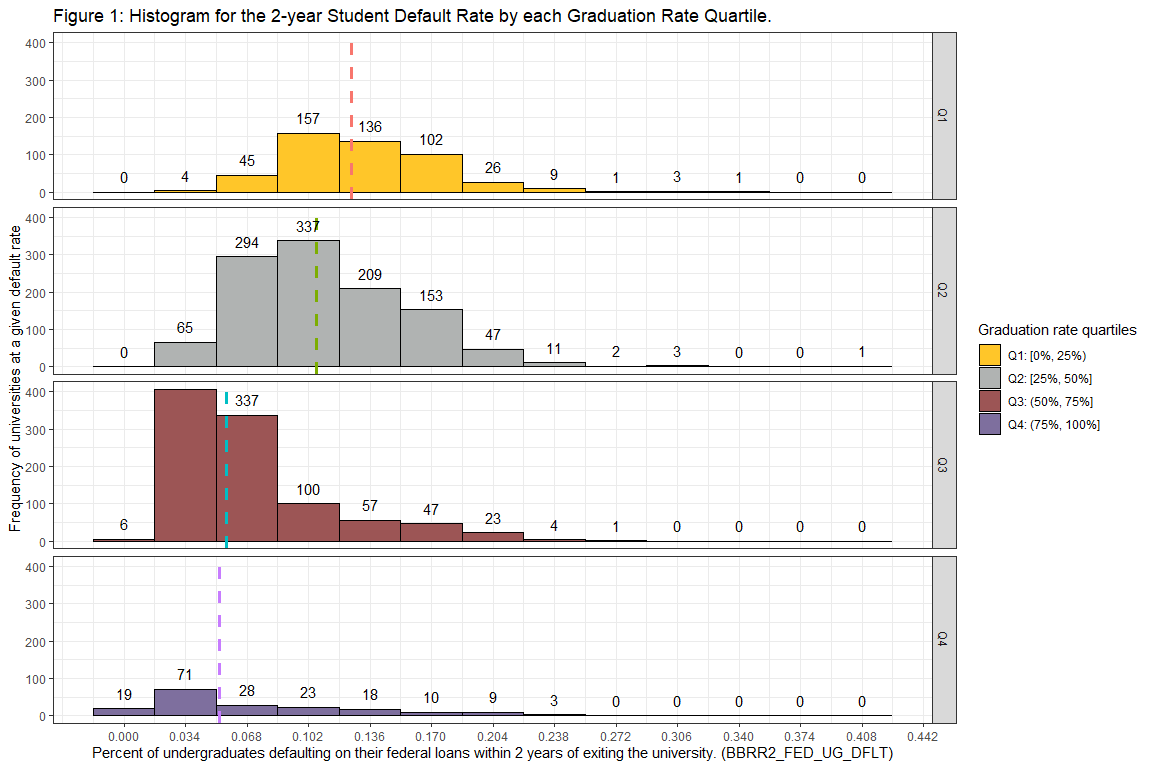
For post-secondary institutions, the “average” default rate for students who default within 2-years was calculated for the four-year graduation rate quartiles:

* less than or equal to 25% (Q1),
* greater than 25% to 50% (Q2),
* greater than 50% to 75% (Q3), and
* over 75% (Q4).

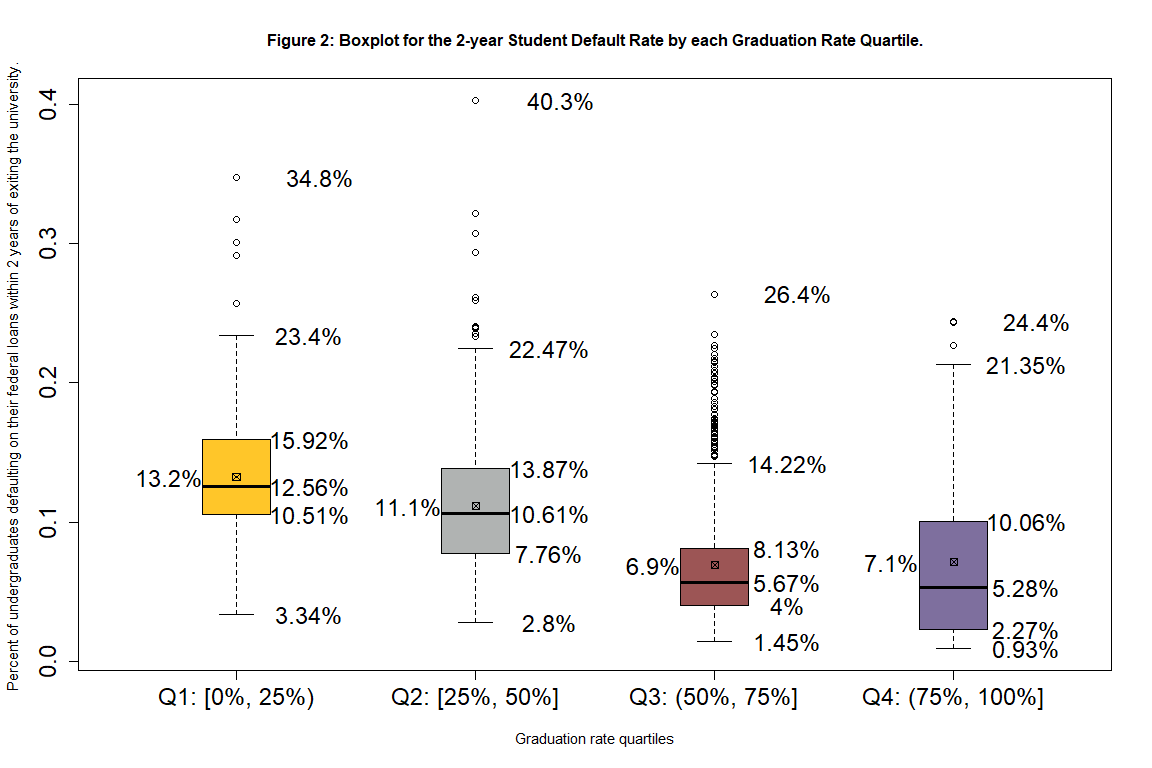
The nonparametric K-sample permutation test was used to detect difference between the graduation rate quartiles for the reason that the data displayed evidence of skewness and did not display evidence suggesting that it was normally distributed. There is a significant difference in the default rates for all four quartiles except the third and fourth quartiles.

**Actionable items:**

* The median 2-year default rate for undergraduates from institutions with higher graduation rates (Q3 and Q4) is about 5% for both quartiles.
* In comparison, the median default rates for institutions with low graduation rates (Q1 and Q2) were 13% and 11%, respectively.
* Thus, institutions who graduate more than 50% of their students tend to have significantly lower loan default rates.
* However, the default rate does not decrease for institutions that graduate more than 75% of their undergraduates.
* Notably, we see that the institutions in Q3 and Q4 are more skewed to the left side of the plot than institutions in Q1 and Q2.



* Although Q3 and Q4 have similar medians, the default loan rates are more variable for the institutions with graduation rates greater than 75% (Q4).
* In fact, Q4 institutions had a standard deviation of loan default rates of 5.87% while Q3 had a standard deviation of 4.34%.
* This suggests that institutions that achieve the highest possible graduation rate may not inherently yield the lowest possible 2-year loan default rate.



# **APPENDIX**

### **A. DATA MANIPULATION**

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. |

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 25%, greater than 25% to 50%, greater than 50% to 75%, and greater than 75%. They were then given the labels Quartile 1 (Q1), Quartile 2 (Q2), Quartile 3 (Q3), and Quartile 4 (Q4), respectively.

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
|  | Quartile 1 | Quartile 2 | Quartile 3 | Quartile 4 |
| Percent range | 0% <= x <= 25% | 25% < x <= 50% | 50% < x <= 75% | 75% < x <= 100% |

## **B. ARE THE VARIANCES EQUAL?**

The variance among the four quartiles is approximately equal. Table 1 displays the standard deviation for each of the 4 quartiles. The largest standard deviation (Q4, 0.5872) is only 1.42 times larger than the smallest standard deviation (Q1, 0.4136).

Since the cutoff point for considering the variance to be approximately equal is 2, our data has NOT violated the assumption of equal variance.

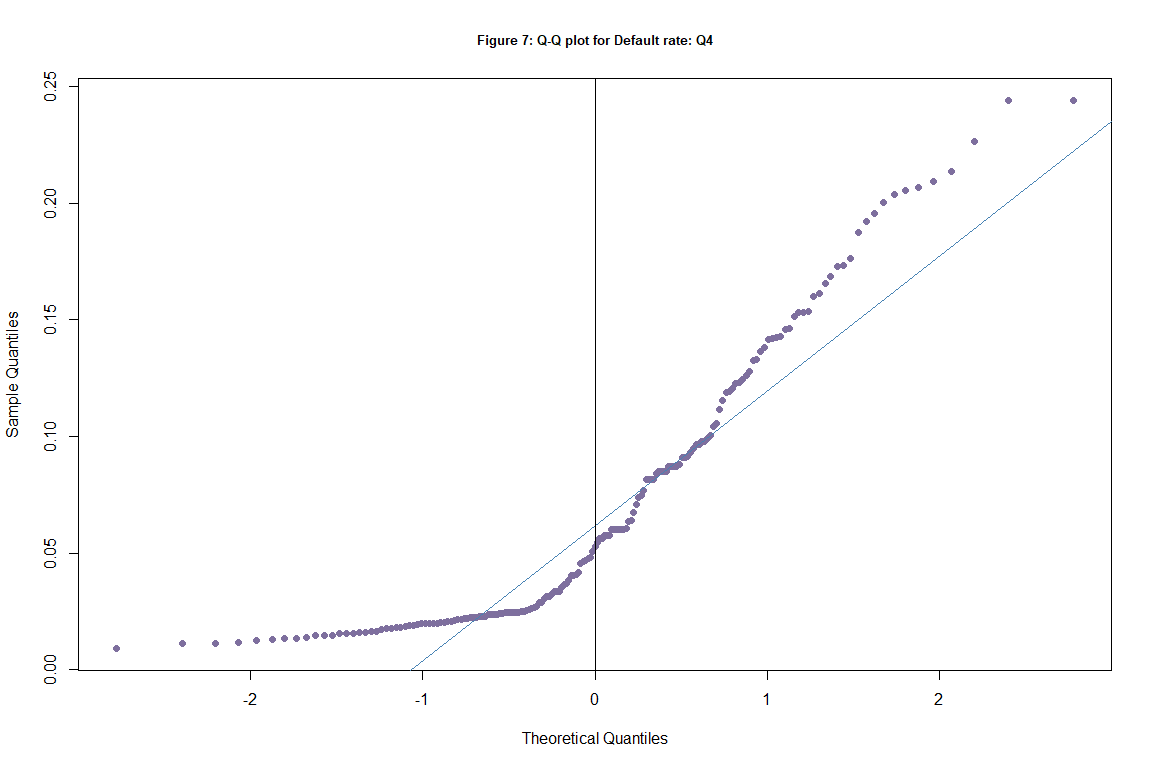
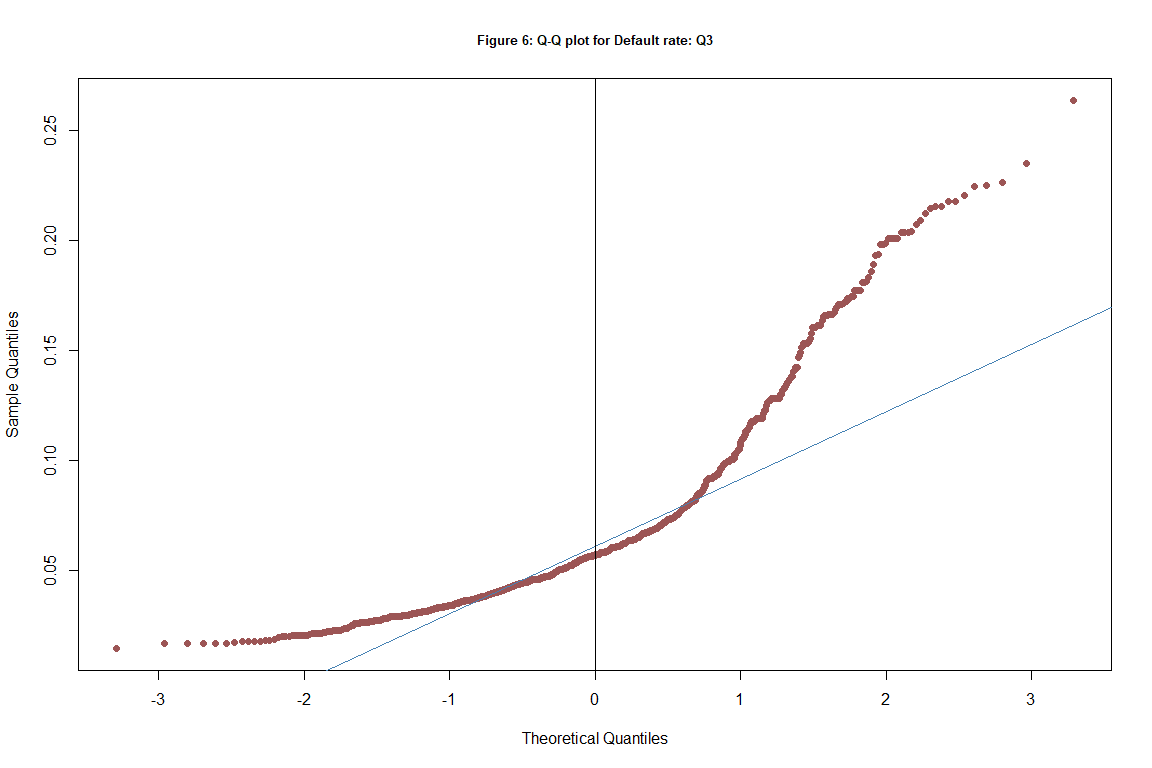
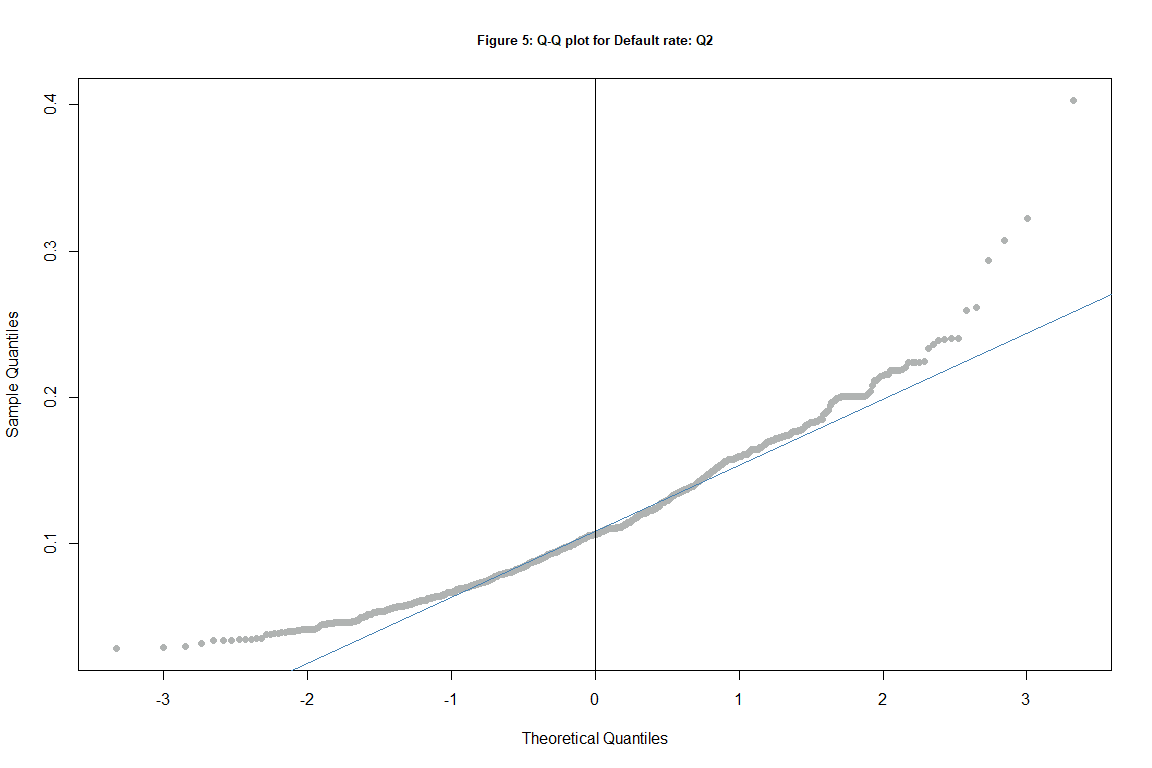
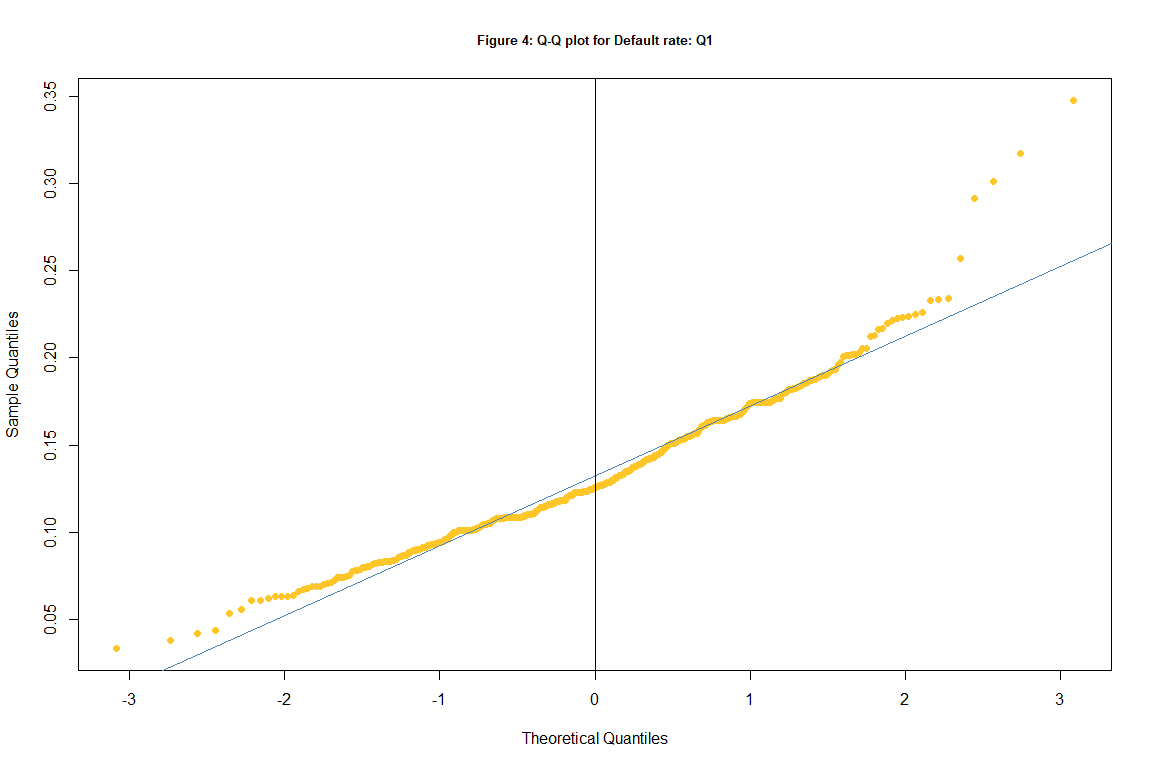
Table 1: Standard Deviations

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
|  | Quartile 1 | Quartile 2 | Quartile 3 | Quartile 4 |
| Percent range | 0% <= x <= 25% | 25% < x <= 50% | 50% < x <= 75% | 75% < x <= 100% |
| Default Rate | 0.04136 | 0.04557 | 0.04341 | 0.05872 |

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

The four quartiles display evidence of skewness. Figures 4, 5, 6, and 7 display Q-Q plots for the four graduation rate quartiles. Each of these plots display values on the left and right side to be above what we would expect (indicated by the blue line) if the data was normally distributed, indicating a significant deviation from normality. Since the interquartile variance is approximately equal and the data did not display evidence suggesting that it was normally distributed, the nonparametric K-sample permutation test will be used over the parametric ANOVA test and the Kruskal-Wallis test.

Notably, on the left side of each of the plots we see that the sample is above the blue line suggesting that there was more institutions with loan default rates closer to 0% than what we would expect if the data is normally distributed. For Q1 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 but with a slight deviation above the blue line. In comparison, Q2 in Figure 5, we can see on the left side of the plot that the deviation above the blue line is much greater than the deviation seen for Q1. Similarly, we can see in Figure 6 (for Q3) and 7 (for Q4) that the data further deviates from the blue line. This information suggests that institutions with higher graduation rates will have more institutions with loan default rates close to zero.



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

The results of this test showed that there is a difference in the default rate between each of the quartiles. 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 quartiles. Further analysis will be conducted via Tukey’s HSD, Fisher’s LSD, and Bonferroni’s correction in the next section.

##   
## K-Sample Asymptotic Permutation Test  
##   
## data: Quant\_var by Cat\_var  
## Chi Square = 645.02, df = 3, p-value < 2.2e-16

##   
## K-Sample Exact Permutation Test Estimated by Monte Carlo  
##   
## data: Quant\_var by Cat\_var  
## p-value = 9.999e-05  
##   
## p-value estimated from 10000 Monte Carlo replications  
## 99 percent confidence interval on p-value:  
## 0.0000000000 0.0005296914

## **E. 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 quartiles contained differences and which did not. From the output of the Tukey’s HSD test, we see that Quartiles 2 and 3 did not significantly differ from each other. Quartiles 0 and 1 did display a difference with each other as well as with Quartile 3. Thus, I am 95% confident that the true difference in the centers for Quartile 2 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 Quartile 3 and 3 but did find a difference between Quartiles 0, 1, and 2.

Table \*: Tukey HSD

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
|  | diff | lwr | upr | p adj |
| Q2-Q1 | -0.0209 | -0.0272 | -0.0146 | 0.0000 |
| Q3-Q1 | -0.0630 | -0.0694 | -0.0565 | 0.0000 |
| Q4-Q1 | -0.0611 | -0.0712 | -0.0510 | 0.0000 |
| Q3-Q2 | -0.0421 | -0.0472 | -0.0370 | 0.0000 |
| Q4-Q2 | -0.0402 | -0.0495 | -0.0309 | 0.0000 |
| Q4-Q3 | 0.0019 | -0.0074 | 0.0113 | 0.9518 |

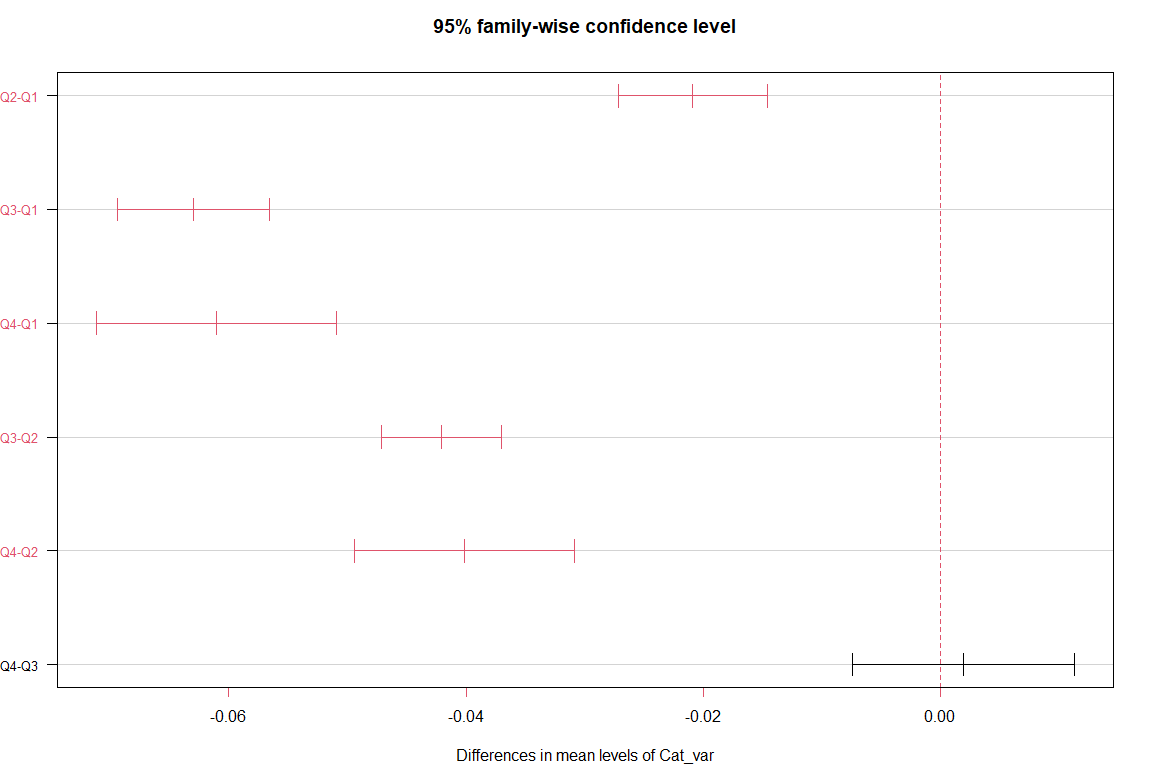


Table \*: Fishers LSD statistics

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
|  | MSerror | Df | Mean | CV |
|  | 0.002 | 2765 | 0.0975 | 46.2533 |

Table \*: Fishers LSD means by Quartile

|  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
|  | Quant\_var | std | r | LCL | UCL | Min | Max | Q25 | Q50 | Q75 |
| Q1 | 0.1323 | 0.0414 | 484 | 0.1283 | 0.1363 | 0.0334 | 0.3477 | 0.1052 | 0.1256 | 0.1592 |
| Q2 | 0.1114 | 0.0456 | 1122 | 0.1088 | 0.1140 | 0.0280 | 0.4030 | 0.0777 | 0.1061 | 0.1386 |
| Q3 | 0.0693 | 0.0434 | 982 | 0.0665 | 0.0721 | 0.0145 | 0.2636 | 0.0400 | 0.0567 | 0.0813 |
| Q4 | 0.0712 | 0.0587 | 181 | 0.0647 | 0.0778 | 0.0093 | 0.2439 | 0.0227 | 0.0528 | 0.1006 |

Table \*: Fishers LSD interquartile comparison

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
|  | difference | pvalue | signif. | LCL | UCL |
| Q1 - Q2 | 0.0209 | 0.0000 | \*\*\* | 0.0161 | 0.0257 |
| Q1 - Q3 | 0.0630 | 0.0000 | \*\*\* | 0.0581 | 0.0679 |
| Q1 - Q4 | 0.0611 | 0.0000 | \*\*\* | 0.0533 | 0.0688 |
| Q2 - Q3 | 0.0421 | 0.0000 | \*\*\* | 0.0382 | 0.0460 |
| Q2 - Q4 | 0.0402 | 0.0000 | \*\*\* | 0.0331 | 0.0473 |
| Q3 - Q4 | -0.0019 | 0.5961 |  | -0.0091 | 0.0052 |

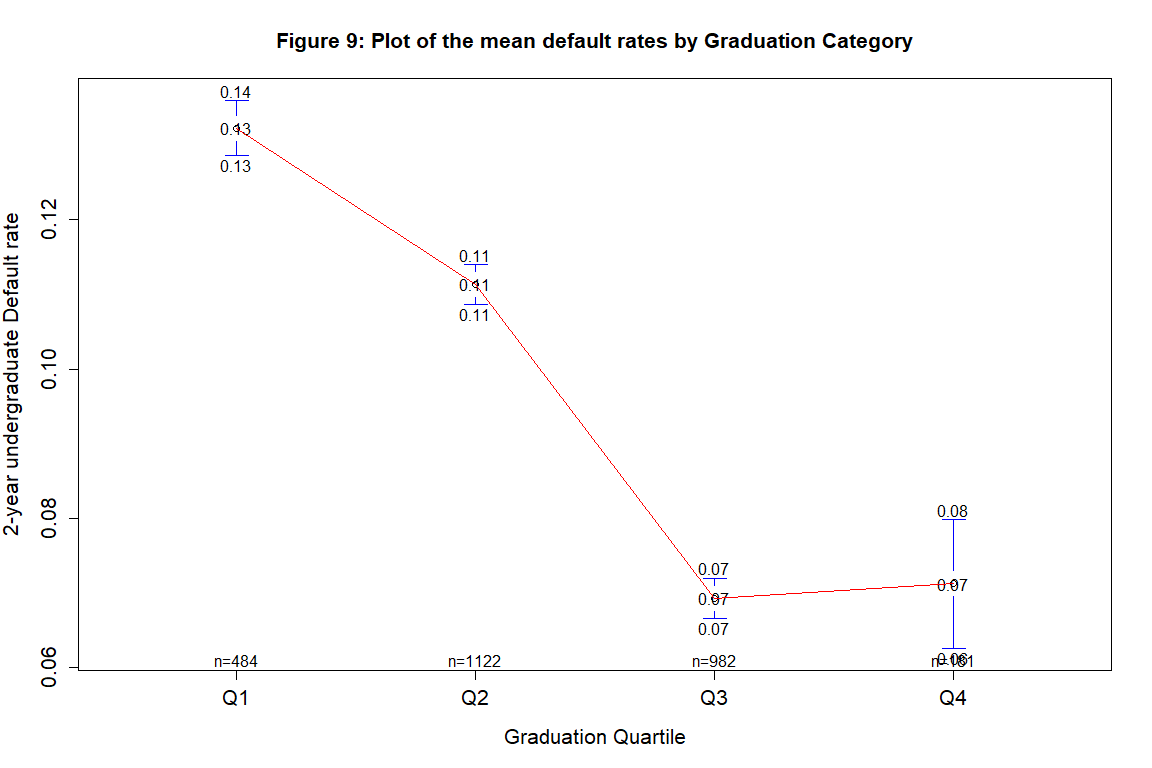


Table \*: Pairwise t-test

|  |  |  |  |
| --- | --- | --- | --- |
|  | Q1 | Q2 | Q3 |
| Q2 | 0 | NA | NA |
| Q3 | 0 | 0 | NA |
| Q4 | 0 | 0 | 0.5961272 |

Table \*: Pairwise t-test with Bonferroni correction

|  |  |  |  |
| --- | --- | --- | --- |
|  | Q1 | Q2 | Q3 |
| Q2 | 0 | NA | NA |
| Q3 | 0 | 0 | NA |
| Q4 | 0 | 0 | 1 |

## **Other unused 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 quartiles exhibited heavy skewness.