Case Study 4, STAT 3220-001, Prof. Varanyak
Graham Gilliam, Timothy Kim

Analysis of Cohort Default Rates Across School Types

Introduction

Student loans often compose a significant proportion of a borrower's financial debt and can take years to pay off. For some students, loans are one of the largest barriers to higher education. According to the US Department of Education, the national federal student loan cohort default rate decreased by 6.1% from the 2014 to 2015 Fiscal Years, reaching the lowest percentages in history since the metric's inception in 2012[1] based on statistical analysis of the 2015 cohort rates from a sample from a larger dataset of higher education institutions (domestic and international). Accounting for potentially relevant variables such as state, region, type of institution, and number of borrowers in default or repay can provide insight on national student loan cohort default rates, and help generally decrease this value and guide both students and lawmakers. The objectives of this study are as follows:

1. Determine if the cohort rate differs across school types, and check to make sure the test's assumptions are met. Does the sample data provide evidence that the cohort rates are different?
2. Determine which specific school types differ. What school types in particular differ from one another?

Data Summary

The initial data contained information from 4,873 higher education institutions. The variables in this dataset relevant to our study are as follows:

1. Default Rate [DRATE]: the official cohort default rate for 2015, found as the ratio of number of borrowers in default over the number of borrowers in repay
2. Institution Type [TYPE]: the category of educational institution. The initial data contained 6 potential categories: public, private non-profit, proprietary, foreign public, foreign non-profit, and foreign for-profit.

For simplicity the three foreign institution types were combined into one umbrella "foreign" category. Thus, there were essentially four institution types being examined: Public (1), Private Non-Profit (2), Proprietary (3), and Foreign (4). The next step was to clean up our data by removing any rows missing data from our initial data set. We then took a 50 institution sample from each of the 4 categories that were used to study our research questions. Our sample was explored, and the following summary statistics were calculated for the default rate and summarized in Table 1.

Table 1

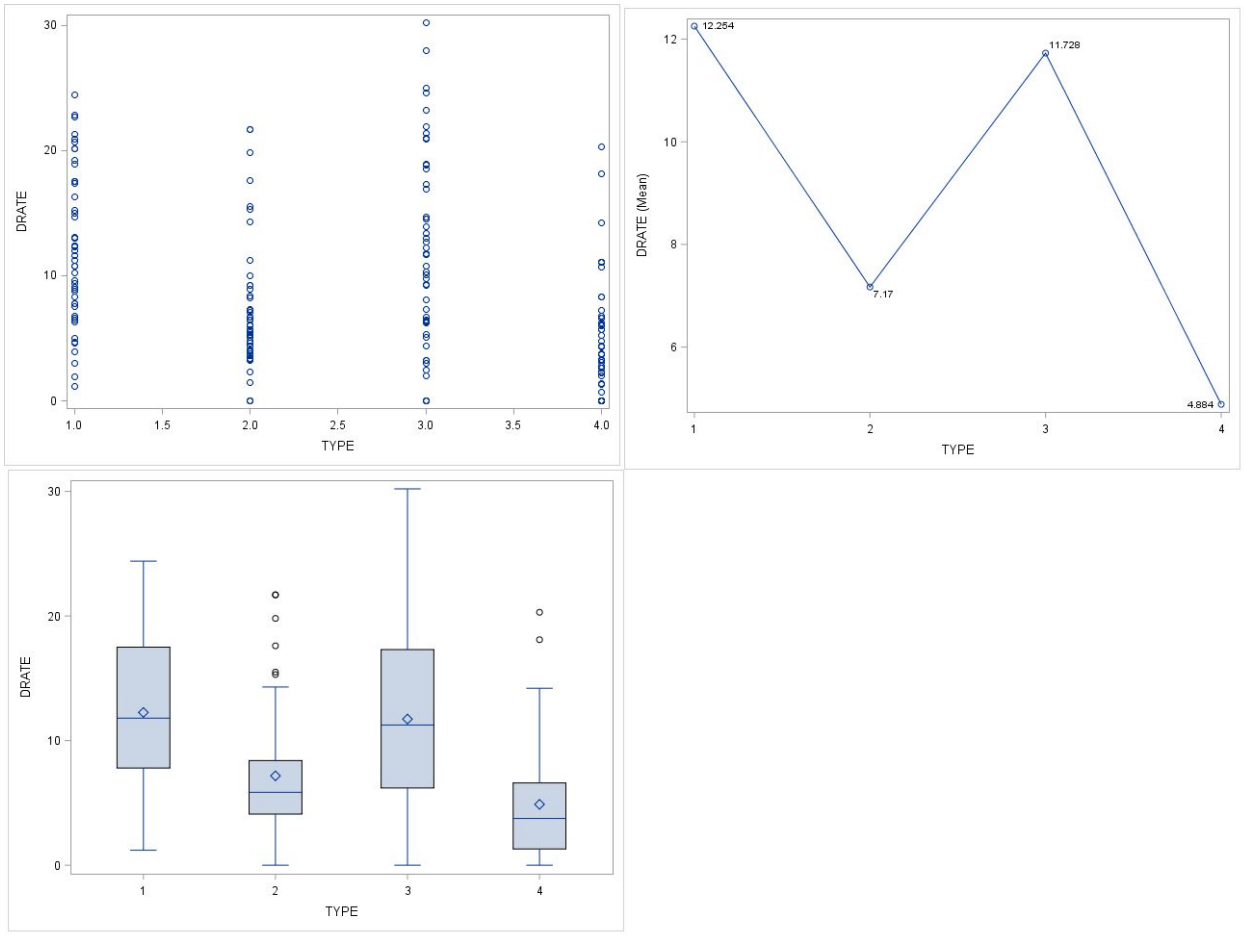
Variable	N	Mean	Standard Deviation	Median	Q1	Q3
Sale Price	200	9.009	6.741	6.950	3.950	13.000

Just by looking at this table, we can already begin to get some information about our sample. As the median is less than the mean, it is likely that we might not have a completely

[1] https://www.ed.gov/news/press-releases/national-student-loan-cohort-default-rate-falls

normal sample, which can be ignored in our analysis if the effect is small enough. What this table doesn't help answer is our main research question, and for that we created three graphs: a scatterplot, a boxplot, and a line plot of default rate means, and summarized the findings from these in Figure 1.

Figure 1



From any of the three graphs, there seems to be a strong case for there being a difference between the mean default rates for each of the four institution categories. Interestingly, from the scatterplot, we see that proprietary institutions have a very large variance, while the other three school types seem to be much more compact in distribution. The boxplot reveals that the mean default rate is higher than the median for every institution type, furthering the possibility of right skew. The sample means for each group also certainly differ. Looking at the line plot representing sample means, it looks as if public and proprietary schools have a larger default rate than foreign or non-profit institutions. With this information in mind, a model was constructed hypothesizing that each institution affects the mean default rate, with public institutions as the base effect. For this model values were assigned by the following:

$x_1 = 1$ if non-profit, 0 otherwise

$x_2 = 1$ if proprietary, 0 otherwise

$x_3 = 1$ if foreign, 0 otherwise.

This yields the following prediction equation: $E(y) = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3$.

Analysis

The initial assumption that all institution types have the same effect on mean default rate gives the null hypothesis $H_o : \beta_1 = \beta_2 = \beta_3 = 0$. Therefore, the reduced model to test is: $E(y) = \beta_0$. SAS yields an F statistic of 17.57 with a p-value of <0.001. Therefore, the null hypothesis can be strongly rejected in favor of the alternative hypothesis, which states that at least one β is not equal to 0. In the context of ANOVA, this means that at least one pair of institution type differs in mean default rate. With this knowledge, a Post-Hoc Analysis was then conducted to determine which of the 12 possible permutations of the 4 institution types differed from each other. The following output was given from SAS when the Tukey's Studentized Range Test was ran at an alpha of 0.05 and summarized in Table 2.

Table 2

Alpha	0.05
Error Degrees of Freedom	196
Error Mean Square	36.35519
Critical Value of Studentized Range	3.66452
Minimum Significant Difference	3.1248

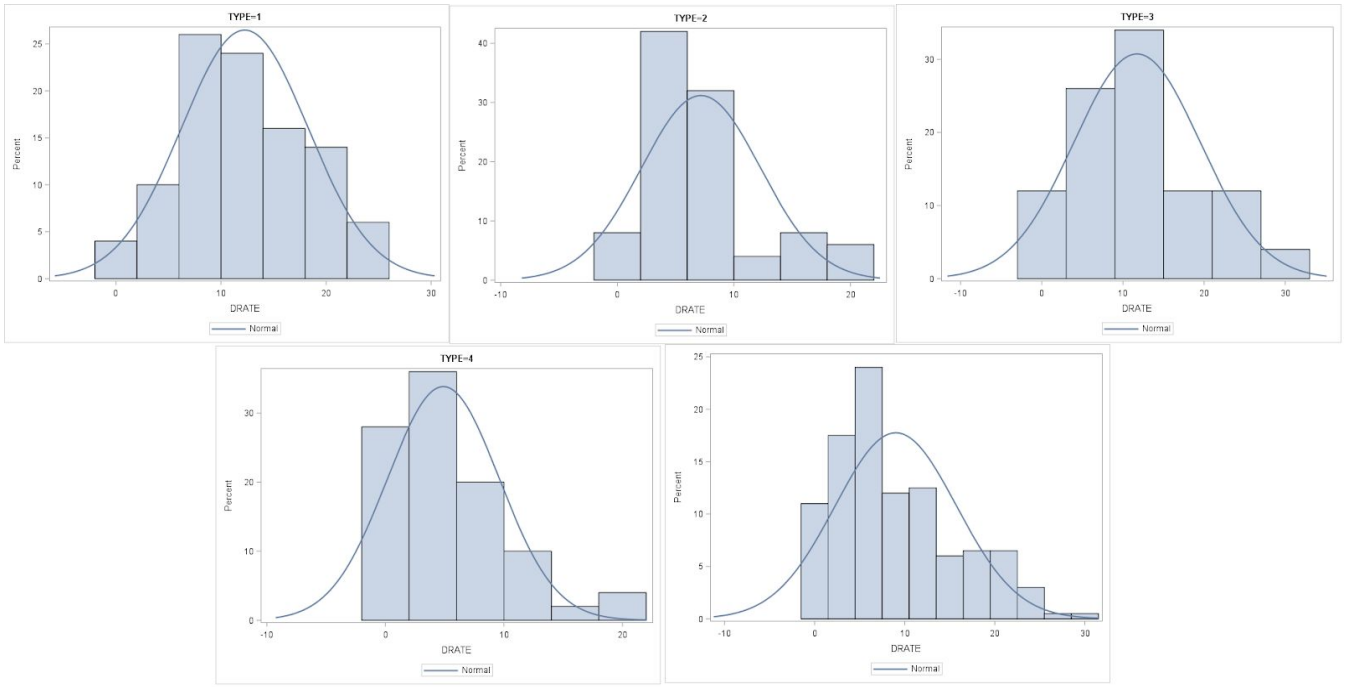
Comparisons significant at the 0.05 level are indicated by ***.

TYPE Comparison	Difference Between Means	Simultaneous 95% Confidence Limits		
1 - 3	0.526	-2.599	3.651	
1 - 2	5.084	1.959	8.209	***
1 - 4	7.370	4.245	10.495	***
3 - 1	-0.526	-3.651	2.599	
3 - 2	4.558	1.433	7.683	***
3 - 4	6.844	3.719	9.969	***
2 - 1	-5.084	-8.209	-1.959	***
2 - 3	-4.558	-7.683	-1.433	***
2 - 4	2.286	-0.839	5.411	
4 - 1	-7.370	-10.495	-4.245	***
4 - 3	-6.844	-9.969	-3.719	***
4 - 2	-2.286	-5.411	0.839	

This analysis illustrates that the pair of public and proprietary institutions and the pair of non-profit and foreign institutions are not significantly different, but that the two pairs differ significantly from each other, which matches with what was observed during exploratory data analysis. From here, the assumptions of the ANOVA test- normality of each treatment and constant variance- were tested.

For ANOVA, each sample has to behave fairly normally, so histograms of default rate for each institution type were constructed in Figure 2. For reference, a histogram of all 200 samples was also made.

Figure 2



From the output illustrated in Figure 2, we see that some of the institution types, public and proprietary, appear to be fairly normal. However, institution types 2 and 4 show substantial skew, as does the overall population. Because tests of ANOVA are normally fairly robust against deviations from normality, and the sample size is decently large, we chose to ignore this mild violation and move on to a more serious one.

Since the data from each treatment is non-normal, Levene's Test was used to test the significance of the unequal variance observed in Figure 1. The Levene's Test gave a statistic of 6.26 with a p-value of 0.004, making this unequal variance very significant. The default rate was then transformed with a logarithm and square root and re-tested. When using a square-root transformed default rate, the Levene's Test still gives a significant statistic, 3.28 with a p-value of 0.022. When compared to the logarithm transformation, the Levene's Test yields a statistic of 1.21, with a p-value of 0.31. While both transformations greatly improve the error, the logarithmic transformation gives equal variance, and is therefore used for the next step of the analysis.

Now that the response variable better meets the assumptions of ANOVA, we re-examined the initial test results to make sure that the unequal variance didn't skew the findings. With the same steps and null hypothesis as before, we reached the same conclusions. We then definitively concluded that public and proprietary schools have equal means, and that private non-profit and foreign institutions have equal means, and adjusted our model to match these findings. We removed β_2 and β_3 from our model and modified the definition of the sole dummy variable, x = 1 if foreign or non-profit, 0 otherwise. This gave a new model equation: $E(y) = \beta_0 + \beta_1 x$. When the model is run in SAS, the prediction equation is found to be: $E(y) = 11.991 - 5.964x$.

Conclusion/Results

Interpreting the Prediction Equation

The analysis yielded a a model equation representing the two significant institution means: $E(y) = 11.991 - 5.964x$. The result of the analysis showed that groups 1 and 3, which represented public and proprietary schools, respectively, had roughly equal mean default rates, and that groups 2 and 4, which represented private non-profit and foreign schools, respectively, were also roughly equal. These pairs, however, had mean cohort default rates that are substantially different. To interpret this equation, each institution will either have a mean cohort default rate of 11.991%, or (11.991-5.964)%. The interpretation for each institution type is represented in Table 3.

Table 3

Institution	Public	Non-Profit	Proprietary	Foreign
Mean Cohort Default Rate (%)	11.991	6.027	11.991	6.027

As explained previously, the public and proprietary institutions have the same mean cohort default rate, 11.991%, while non-profit and foreign institutions have the same mean default rate, 6.027%. This does not, however, tell us why the means are the same. Public institutions likely accept more low-income students, who are more likely to default on loans, than a foreign institution. Non-profit schools, which we found to be slightly skewed, might have a low average due to smaller student bodies.

Conclusion

The test and model created from the sample data suggest that each of the four institution types has a different cohort default rate. The difference between institution types seems to be large enough to justify making distinctions between the overall mean CDR and those of its subsets. Thus, it might not be wise to necessarily average all cohort rates and claim that they are dropping- it could just be one of the represented institution types that is making strides in helping its students pay off loans, while the rest of them piggyback off hasty generalizations.

SAS Code