Introduction

In this case study, we will be focusing on Cohort default rate, which is the percentage of Stafford Loan borrowers who enter repayment on certain loan programs and default on a loan. We will use a sample from 2015 cohort rates and compare the cohort rate across school type.

Data Summary

The data that we are analyzing for this study is from the Federal Student Aid in 2015 and we have records from 4873 higher education institutions, consisting of both foreign and domestic. The variables that we are using are:

OPEID: Office of Postsecondary Education Identifier

State: Institution's State abbreviation

Type: The code identifying the ownership control of the institution. This variable is separated into 6 different numerical values: 1 (public), 2 (private, nonprofit), 3 (proprietary), 5 (public, foreign), 6 (public, foreign), 7 (foreign, for-profit).

Num: Number of borrowers in default for 2015

Denom: Number of borrowers in repay for 2015

Drate: Official Cohort Default Rate for 2015; lower scores are better.

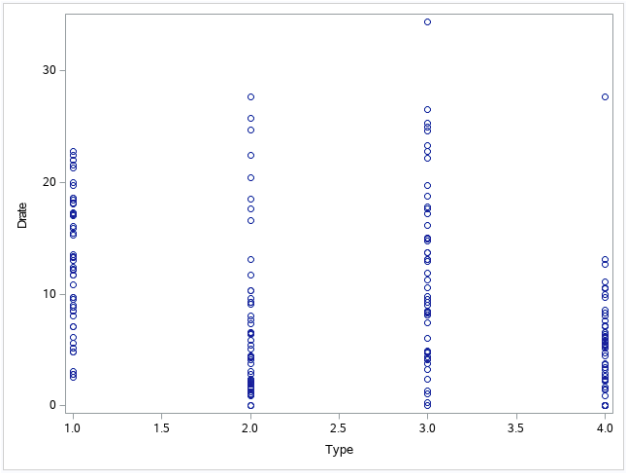
Region: The main institution's Department of Education regional code. This variable is separated into 11 different numerical values: 1 (Boston); 2 (NY); 3 (Philadelphia); 4 (Atlanta); 5 (Chicago); 6 (Dallas); 7 (Kansas City); 8 (Denver); 9 (San Francisco); 10 (Seattle); 11 (Foreign)

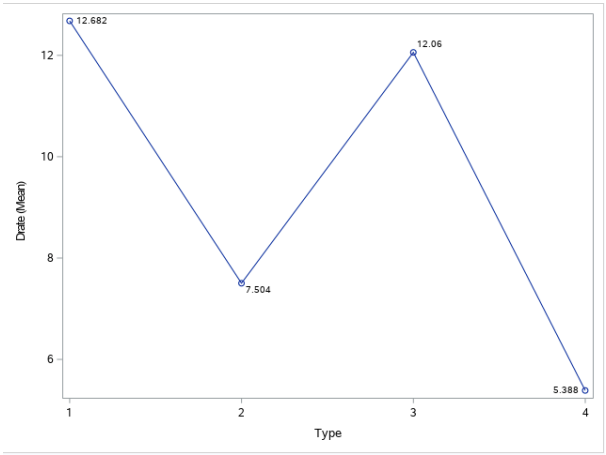
Data Cleaning and Sampling Process

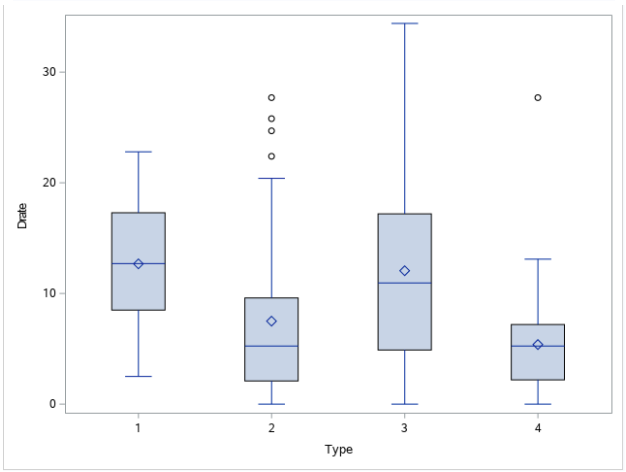
We did not have to remove any data for our bank subscription data as there are no missing values. We made dummy variables for education which has 4 levels. We set our base level as primary education. We will take a sample of 4521 observations as provided by our dataset.

Exploratory Data Analysis

We used a completely randomized design. We have three independent variables, balance, education, and age. We will take the average for each independent variable for both success and failure, and compare for each independent variable.







It appears there are difference in the mean Drate (Official Cohort Default Rate for 2015) for the four types of schools (ownership control of the institution), so we will perform a statistical analysis to compare the means, through the ANOVA procedure.

Analysis

Firstly we want to determine if there is a statistical difference between the four means (four types of schools):

F-Value = 14.57

Alpha = .05

P-Value = <.0001

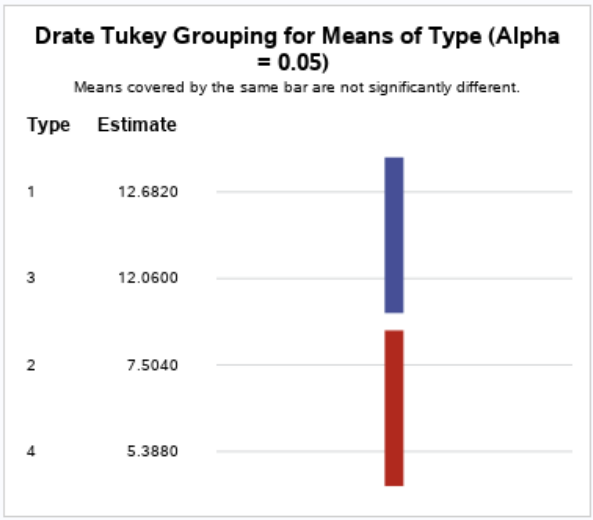
H_0: $\mu_1 = \mu_2 = \mu_3 = \mu_4$

H_a: $\mu i \neq \mu j$ *(for a pair of i and j) At least one of the pairs μ in the null are not equal.*

We reject the null hypothesis that μ_1 through μ_4 equal to each other. Our P-Value of <.0001 is less than our alpha .05. Thus, we concluded that there is a statistical difference between the four means.

We also want to determine the pairs of school types that are significantly different. We use an experiment error rate of 0.05.

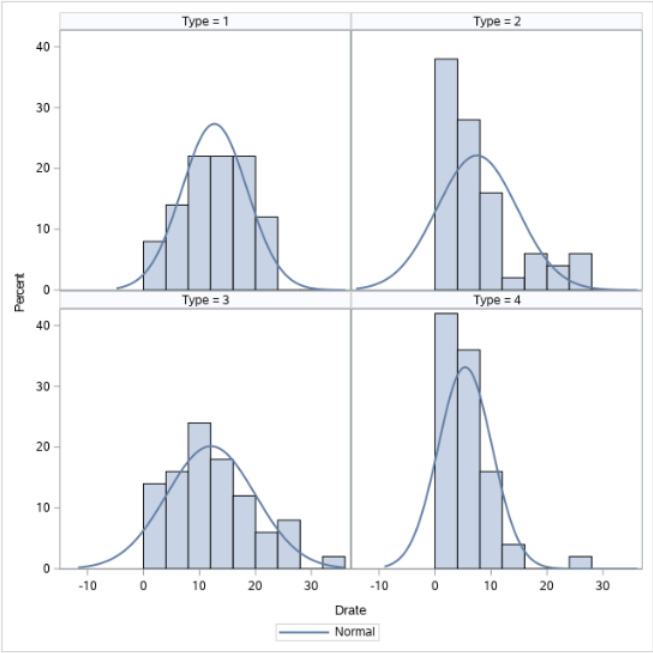
Critical Studentized Range: 3.66452



Public school and proprietary school is the only pair of school that do not have a significantly different rate.

Moreover, we are interested if these data violate the assumptions of ANOVA, so we check the assumptions for normality and equal variance.

Normality Assumption:



Types 2 (private) and 4 (foreign) violate the normality assumptions. However, our data is robust to violations of normality so we do not need to make any transformations to the model.

Constant Variance Assumption:
We are using Levene's test because our response was non-normal.
F-Value = 5.28
Alpha = .05
P-value = .0016
H_0: There is not a violation of the constant variance assumption
H_a: There is a violation of the constant variance assumption
We reject the null hypothesis that there is not a violation of the constant variance assumption. Our P-Value of .0016 is less than our alpha .05. Based on this, we can see that the experiment is not balanced and will need to do a transformation.

Lastly, we are interested in a 95% confidence interval for the difference in mean of the square root of Drate (Official Cohort Default Rate for 2015) between all types of schools.
F-Value = 16.87
Alpha = 0.05
P-value < 0.001

Comparisons significant at the 0.05 level are indicated by ***.				
TYPE Comparison	Difference Between Means	Simultaneous 95% Confidence Limits		
1 - 3	0.2102	-0.3928	0.8132	
1 - 2	1.0343	0.4313	1.6373	***
1 - 4	1.4301	0.8271	2.0331	***
3 - 1	-0.2102	-0.8132	0.3928	
3 - 2	0.8241	0.2211	1.4271	***
3 - 4	1.2199	0.6169	1.8229	***
2 - 1	-1.0343	-1.6373	-0.4313	***
2 - 3	-0.8241	-1.4271	-0.2211	***
2 - 4	0.3959	-0.2071	0.9989	
4 - 1	-1.4301	-2.0331	-0.8271	***
4 - 3	-1.2199	-1.8229	-0.6169	***
4 - 2	-0.3959	-0.9989	0.2071	

We conclude that comparisons that are significant at the 0.05 level are 1-2 , 1-4, 3-2, 3-4, 2-1, 2-3, 4-1, 4-3, because they don't include 0 as part of the confidence interval. Therefore we are 95% confident that the betas for these values are not 0. Where: 1 (public), 2 (private, nonprofit), 3 (proprietary), 4 (foreign).

Conclusion

Based on series of tests we conducted, we can conclude that school type does have an effect on the Official Cohort Default Rate for 2015. More specifically, after some adjustments to the model, we found that there is a difference between public and private schools, public and foreign schools, proprietary and private schools, proprietary and foreign schools, private and public schools, private and proprietary schools, foreign and public schools, and foreign and proprietary schools.

One of the improvements that we could make in our model is to increase the sample size which could possibly lead to more accurate results. We also could have possibly acquired the missing data points so that we would not have had to erase values. We only used completely randomized design, so we could have tested with other designs such as randomized block design to see if it would produce a different outcome. Another possible improvement we can conduct is to account for other factors than just the school types, such as the region.

SAS Codes

```
data cs4;
infile '/folders/myfolders/cohortscore1.csv' dlm=',' firstobs=2;
input OPEID STATE $ TYPE NUM DENOM DRATE REGION;
run;

data cs44;
 set cs4;
where DRATE ne . ;
run;

data cs444;
set cs44;
if TYPE ='5' then TYPE ='4';
if TYPE ='6' then TYPE ='4';
if TYPE ='7' then TYPE ='4';
run;

proc sort data=cs444;
by type;
run;

proc surveyselect data= cs444 method=srs n=50 seed=225 out=samples;
strata type;
run;

proc sgplot data=samples;
scatter x=type y=drate;
run;

proc sgplot data=samples;
        vline type /response=drate stat=mean markers datalabel;
run;

proc sgplot data=samples;
vbox drate/ category=type;
run;

data samples2;
set samples;
```

```
private=0;
proprietary=0;
foreign=0;
if type='2' then private=1;
if type='3' then proprietary=1;
if type='4' then foreign=1;
run;

proc reg data=samples2 plots=none;
model drate=private proprietary foreign;
test private,proprietary,foreign;
run;

proc anova data=samples2;
class type;
model drate=type;
means type/ tukey lines;
run;

proc sgpanel data=samples2;
panelby type;
histogram drate;
density drate /type=normal;
run;

proc anova data=samples2;
class type;
model drate=type;
means type/ hovtest=levene(type=abs);
run;



data samples3;
set samples2;
sqrtdrate=sqrt(drate);
lndrate=log(drate);
run;

proc anova data=samples3;
```

```
class type;
model sqrtdrate=type;
means type/ hovtest=levene(type=abs);
run;

proc anova data=samples3;
class type;
model lndrate=type;
means type/ hovtest=levene(type=abs);
run;

proc anova data=samples3;
class type;
model sqrtdrate=type;
means type/ tukey cldiff ;
run;
```