Statistical Analysis on the Department of Education's Graduation Results for Classes of 2005 to 2020 in New York City

Mohamad Ali Kalassina

Matthew Johnson

(Group 13)

Georgen Institute for Data Science

University of Rochester

Rochester, New York

DSCC462 – Computational Introduction to Statistics

Fall 2021

Submitted: December 06, 2021

Abstract

This study analyzes the data provided by the New York City Department of Education on graduating classes from 2005 to 2020 from a computational statistics perspective. Inferential results of six different statistical inquiries are presented and analyzed using R Studio as a statistical analysis tool. The inquiries target multiple attributes such as school size, racial composition, economic status of students across the five boroughs of New York City. Multiple statistical tests are utilized to reach conclusions that allow for a in depth understanding of the factors affecting graduation rates across Manhattan, Brooklyn, Queens, the Bronx, and Staten Island. The results of this analysis show that there is a discrepancy between the graduation rates of assigned genders (male and female), there exist similarities between graduation rates across different boroughs, economic disadvantage does not affect the graduation rate, in addition to some other insights into the components that play a role in graduation rates across the five boroughs of New York City.

Table of Contents

List of Tables

I. Introduction

New York City has pronounced diversity in its population when it comes to economic status, racial makeup, ethnic backgrounds, and other aspects and therefore its student body composition reflects this same diversity. According to the New York City Department of Education (2021), the NYC school system is constituted of 1,094,138 students across the five boroughs, making it the largest school district in the United States. With a graduation rate of 78.8% and a dropout rate of 5.8% in August 2020, there are a lot of questions to be answered about what the reasons behind this dropout rate are and what factors impact the graduation rate.

Each year the New York City Department of Education updates their statistical reports to show graduation rates at different levels (school, district, borough, and city) for different socio-economic factors including race, gender, economic situation, disability status, and English as a first language. The objective is to investigate the variables that have impact on graduation rates and dropout rates using the graduation results for the class of 2005 up until the most recent graduating class of 2020 (NYC Department of Education Graduation Results for Cohorts 2001 to 2016, 2021. There are six inquiries to be answered by performing different inferential tests on the data set and an in-depth analysis of these results will be performed in order to understand the factors that impact graduation rates and dropout rates, and other interesting insights.

The data was retrieved in different files, one for each of the city, borough, district, and school level. Below are some preliminary data descriptions:

- The boroughs attribute has all five boroughs of NYC: Brooklyn, Manhattan, The Bronx, Staten Island, and Queens
- The school attribute has a total of 543 non-charter schools
- The races attribute has 6 different values: Black, Asian, Hispanic, Native-American, Multi-Racial, and White
- The economic status attribute has 2 values: economically disadvantages and non-economically disadvantaged
- There are data points for 914,096 students

II. Methodology

The statistical tool used for the completion of this analysis is the R programming language using R Studio. The statistical inferential methods used to perform the analyses vary. For each of the six inquiries of interests, a hypothesis is defined and tested using the appropriate statistical technique(s), and the results are evaluated.

For the analysis of categorical variables (such as school size, graduation status, race, assigned gender), the below tests and techniques were utilized:

1. Chi-Square test for independence: This technique allows for the testing of independence for contingency tables created for the categorical variables of interest.
2. Inference of proportions: this technique allows for the comparison of proportions of two different groups.

For the analysis of numerical variables (size as cohort/class size, number of graduates, number of dropouts), the below tests and techniques were utilized:

1. Wilcoxon Rank Sum Test: this Non-parametric methods on means allows for testing on the median difference between two samples with unknown distributions
2. Welch T-test: t-test with two independent samples with unequal variance: used to determine the relationship between two numerical variables of two different samples of distinct variances
3. One-way ANOVA with multiple comparison: used to compare means of more than two groups of interest, the NYC boroughs in this specific study
4. Spearman's Correlation Coefficient: used to identify the correlation between two factors and whether they impact graduation rate in the same fashion

III. Data Analysis, Hypothesis Testing, and Results

A. Exploratory Data Analysis

Prior to deciding on which factors are of high significance and should be focused on in this statistical analysis, some exploratory data analysis was performed on the data set at hand. By viewing the summary of each category, we can see an average ten-point difference within each of the summary statistics between males and females. This is cause for inquiry and will be discussed using a test on proportions in the following section.

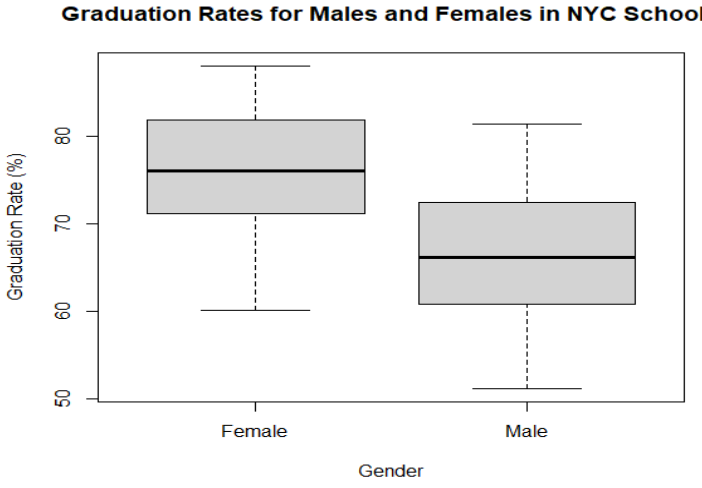
Female Graduation Rate Summary Statistics

```
    PerGrad
Min.   :60.10
1st Qu.:71.28
Median :76.05
Mean   :76.05
3rd Qu.:81.88
Max.   :88.10
```

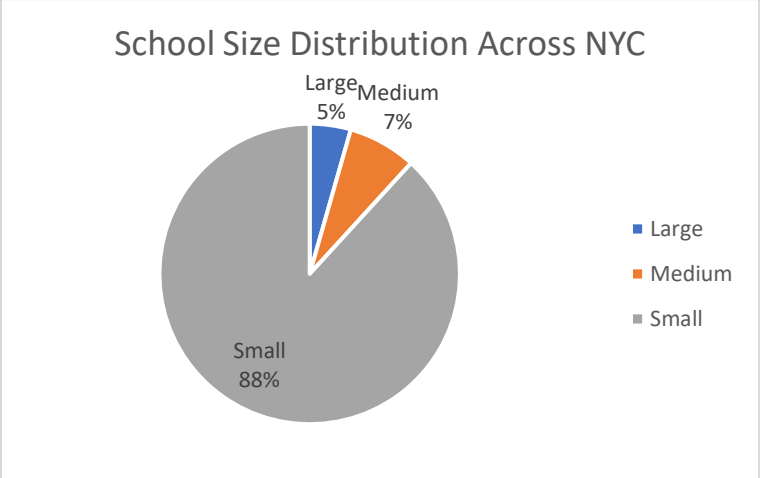
Male Graduation Rate Summary Statistics

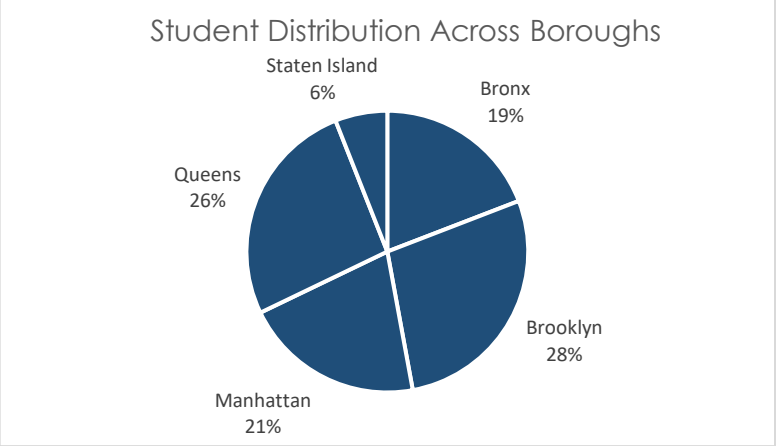
```
    PerGrad
Min.   :51.20
1st Qu.:60.88
Median :66.25
Mean   :66.21
3rd Qu.:72.50
Max.   :81.40
```

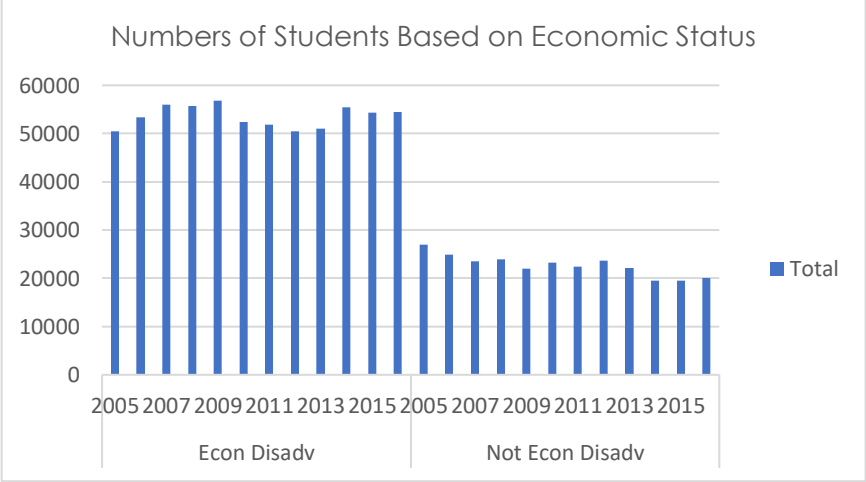
This is further explored by using the boxplots below, which lead us to believe that there is a difference in graduation rates between male and female students.



Furthermore, an interesting point of focus is the sizes of school across the city. Below is a summary of school sizes, where small schools have less than 250 students, medium schools have less than 850, and large schools have between 850 and 1500 students. This shows that the majority of schools in NYC are in fact small.







B. Hypothesis Testing & Results

This section outlines the six inquiries of interest, presents the hypothesis that allows for answering this inquiry, delineates the statistical technique used to test the hypothesis, and presents the results. The discussion around these results and what insights they provide are to be elaborated upon in the discussion section of this report.

a. Is the average graduation rate for non-white students different from that of white students?

The hypothesis being tested in this case is:

$$H_o: \mu_{white} - \mu_{non\text{-}white} = 0$$
$$H_1: \mu_{white} > \mu_{non\text{-}white}$$

To test this hypothesis, a ***non-parametric method on means***, the Wilcoxon Rank Sum Test, is performed on the data in table 6 appendix I. Below are the results of this test:

```
        Wilcoxon rank sum test

data:  sample1 and sample2
W = 141, p-value = 3.392e-05
alternative hypothesis: true location shift is greater than 0
```

Given that p-value <<< alpha, we reject the null hypothesis and conclude that the graduation rate for white students is in fact greater than that of non-white students (Asian, Black, Hispanic, Multi-Racial).

b. Based on the Class of 2020, are the school size and graduation status independent?

The hypothesis being tested in this case is:

H_o: school size and graduation status are independent
H_1: school size and graduation status are associated

To test this hypothesis, a ***Chi-Square test of independence*** is performed.
The class of 2020 cohort had a total of 71,155 students. The distribution of these students based on school size is summarized in the observed contingency table below which was retrieved from Table 1 and Table 2 in Appendix A:

Table 1:Observed Contingency Table for School Size vs Graduation Status

School Size	Graduated	Did Not Graduate
Large	16204	2035
Medium	10629	2229
Small	31748	8310
Grand Total	**58581**	**12574**

Given the above contingency table of observed values, a contingency table of expected values was calculated, and the Chi-Square test of independence was run on the data to give the below results:

```
        Pearson's Chi-squared test

data:  x
X-squared = 793.05, df = 2, p-value < 2.2e-16
```

Given the p-value << 0.05, the null hypothesis is rejected, and it is concluded that there in fact does exist an association between school size and graduation status.

c. *Do students who are not economically disadvantaged have a higher graduation rate from students who are economically disadvantaged?*

The hypothesis being tested in this case is:

$$H_0: \mu_{not\_disadvantage} - \mu_{disadvantaged} = 0$$
$$H_1: \mu_{not\_disadvantage} > \mu_{disadvantaged}$$

To test this hypothesis, a ***t-test with two independent samples of unequal variance (Welch t-test)*** is performed on the 12 different samples, one from each cohort from 2005 through 2016 (class of 2020). Table 5 in the appendix outlines the summary of this data. The results of the t-test are as follows:

```
        Welch Two Sample t-test

data:  nondisadvantaged and disavantaged
t = 0.17682, df = 21.163, p-value = 0.4307
alternative hypothesis: true difference in means is greater than 0
95 percent confidence interval:
 -4.520478       Inf
sample estimates:
mean of x mean of y
 70.21999  69.70207
```

Seeing that the p-value in this test result is much greater than alpha, we fail to reject the null hypothesis and conclude that students with economic disadvantage have the same graduation rate as students who do not have an economic disadvantage.

d. *Is the proportion of male students who graduate more than that of female students?*

The hypothesis being tested in this case is:

$$H_0: p_{females} = p_{males}$$
$$H_1: p_{females} > p_{males}$$

To test this hypothesis, a ***z-test for comparing two proportions*** is performed. This test was performed on all 914,096 students in the cohorts that have gender data attributed to them. Table 4 in the tables section provides the summary of this data. Tested against the standard significance level of 5%, below are the results of this test:

```
        2-sample test for equality of proportions without continuity correction

data:  x
X-squared = 2016.3, df = 1, p-value < 2.2e-16
alternative hypothesis: two.sided
95 percent confidence interval:
 -0.03699948 -0.03390676
sample estimates:
   prop 1    prop 2
0.5704356 0.6058888
```

Seeing that the p-value is significantly less than alpha, the null hypothesis is rejected, and we conclude that the proportion of female graduates is in fact higher than the proportion of male graduates across all five boroughs of New York City.

e. *Are graduation rates in the five boroughs of NYC equal?*

The hypothesis being tested in this case is:

H0: μ1 = μ2 = μ3

Ha: At least one of the boroughs has an average graduation rate different from at least one of the other boroughs

To test this hypothesis, a ***One-way ANOVA*** followed by ***multiple comparison*** is performed**,** and the results are presented below:

```
Analysis of Variance Table

Response: AnTest$PerGrad
                     Df Sum Sq Mean Sq F value    Pr(>F)
factor(AnTest$Borough)  4 1510.9  377.73  13.396 1.097e-07 ***
Residuals              55 1550.9   28.20
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

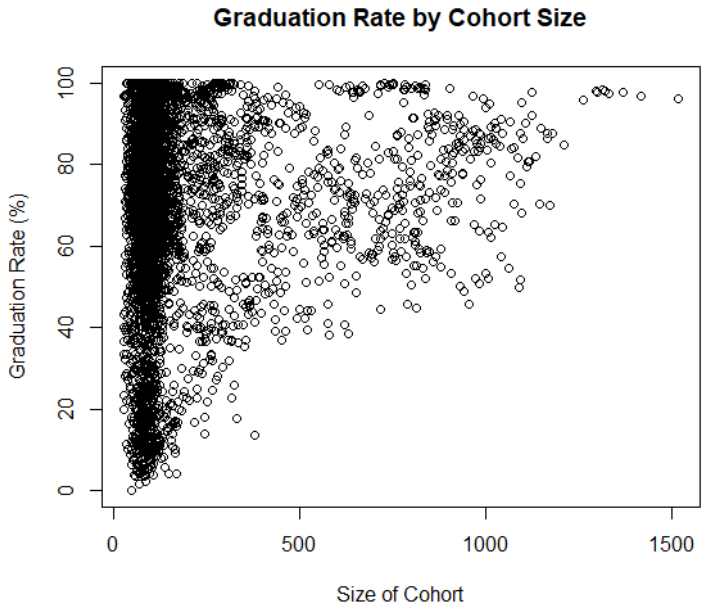
Observed is a p-value much less than the significance level for this test, therefore our null hypothesis is rejected, and we confirm that at least one of the NYC boroughs has a different graduation rate than the rest of the boroughs. In order to further analyze that, the Tukey multiple comparison of means is performed, and the results can be observed below.

```
  Tukey multiple comparisons of means
    95% family-wise confidence level

Fit: aov(formula = AnTest$PerGrad ~ factor(AnTest$Borough))

$`factor(AnTest$Borough)`
                             diff          lwr       upr     p adj
Brooklyn-Bronx           7.225000  1.110851377 13.339149 0.0128619
Manhattan-Bronx          9.183333  3.069184711 15.297482 0.0008035
Queens-Bronx            10.808333  4.694184711 16.922482 0.0000623
Staten Island-Bronx     15.300000  9.185851377 21.414149 0.0000000
Manhattan-Brooklyn       1.958333 -4.155815289  8.072482 0.8944067
Queens-Brooklyn          3.583333 -2.530815289  9.697482 0.4710160
Staten Island-Brooklyn   8.075000  1.960851377 14.189149 0.0040623
Queens-Manhattan         1.625000 -4.489148623  7.739149 0.9436196
Staten Island-Manhattan  6.116667  0.002518044 12.230815 0.0498572
Staten Island-Queens     4.491667 -1.622481956 10.605815 0.2468801
```

From the p-values generated above, we conclude that there are similar graduation rates in cases where the p-value is greater than the significance level. These cases are Manhattan and Brooklyn, Queens and Brooklyn, Queens and Manhattan, and Staten Island and Queens.

6. *Is the number of students in a school correlated to the number of dropouts?*

To see if there was a correlation between the size of the school and the overall graduation rate. The figure below plots these two variables against each other. Interestingly, this does show there is a relationship between graduation size and cohort size. However, it does not fit a traditional linear manner. It's possible that this relationship would best be described as a floor by which as cohort size increases, the minimum graduation rate increases as well. However, investigating this is outside the realm of this study.



The hypothesis being tested in this case is: **H0: ρ = 0 vs. H1: ρ != 0**

To test this hypothesis, a ***Spearman's Rank Correlation Coefficient*** test is performed. After running Spearman's correlation coefficient, the correlation coefficient turned out to be **R = 0.0933364**. Hence, we can conclude that there is not a strong linear relationship between these variables.

```
        Spearman's rank correlation rho

data:  cohortSize and gradRate
S = 2.0991e+10, p-value = 1.696e-11
alternative hypothesis: true rho is not equal to 0
sample estimates:
      rho
0.0933364
```

Furthermore, the p value above indicates that we reject the null hypothesis. Having looked at correlation in two different approaches, the conclusion is that there exists a correlation, however a weak one.

IV. Discussion

According to the exploratory data analysis and different hypotheses tested in the above section, many conclusions were made, some which were mere confirmations to studies that have been previously published, and others that presented new insights into the New York City school system and graduation rates. According to research published in early January 2021 by New York University's Steinhardt School of Culture, Education, and Human Development, the gap in graduation rates between Hispanic and Black students when comparted to their White and Asian counterparts has narrowed since 2004 (2021). The Wilcoxon Sum Rank Test performed on students of different races in the first inquiry where we concluded that the graduation rate for white students is in fact greater than that of non-white students shows that although the gap might be narrowing, the difference in the rates is still statistically significant.

Furthermore, our results are found to contradict a study conducted by Measure of America of the Social Science Research Council. Their brief highlights that a major factor behind the disparity in graduation rates between NYC boroughs is the economic disadvantage that some students suffer from (2016). On the contrary, our study proved through its third inquiry that students with economic disadvantage have the same graduation rate as students who do not have an economic disadvantage. This discrepancy between the two studies could be for one of two reasons.

The first is that our study is based on statistical analysis and theirs is based on qualitative measures and surveys. The second reason could be the fact that our study considers classes of 2016, 2017, 2018, 2019, and 2020 whereas their study stops at the class of 2015.

Another noticeable result is one shown in the exploratory data analysis, showing that the majority of NYC schools are small in size. This is actually in conformance with the fact that NYC Department of Education had strategized closing bigger schools and replacing them with smaller ones.

Some interesting insights that our analysis offers are the fact that although females and males are of equal proportions across the city's schools, females have a higher graduation rate than that of males. In addition to that, there are certain similarities in dropout and graduation rates between boroughs such as Manhattan and Brooklyn and Queens, while The Bronx proved to have a different graduation rate based on the ANOVA test performed. Overall, the results of this statistical analysis are important for understanding the current state of high schools in the five boroughs of New York City.

V. Conclusion

Being the largest school district in the United States, the NYC school district is home for more than 1 high school students. Although some of the five boroughs is the city share certain similarities in trends when it comes to graduation rates, there still exists a disparity between some of them. Furthermore, given the fact that students of certain demographics (whether it is white students or female students) have higher graduation rates than their counterparts, this highlights more disparity in the NYC schooling system. The outcomes of this study are insightful and should be looked upon by professionals and leaders in the education field, that of New York City specifically, in aim of narrowing down the gaps that exist between boroughs and students and achieving a better graduation rate overall.

References

[1] *DOE Data at a Glance*. NYC Department of Education. (2021). Retrieved from https://www.schools.nyc.gov/about-us/reports/doe-data-at-a-glance.

[2] *NYC Department of Education Graduation Results for Cohorts 2001 to 2016*. NYC Department of Education Info Hub. (2021, December 1). Retrieved from https://infohub.nyced.org/reports/academics/graduation-results.

[3] *How have NYC's High School graduation and college enrollment rates changed over time?* NYU Steinhardt. (2021, January 12). Retrieved December 5, 2021, from https://steinhardt.nyu.edu/research-alliance/research/spotlight-nyc-schools/how-have-nycs-high-school-graduation-and-college.

[4] (2016). (rep.). *HIGH SCHOOL GRADUATION IN NEW YORK CITY IS NEIGHBORHOOD STILL DESTINY?* Retrieved from https://ssrc-static.s3.amazonaws.com/wp-content/uploads/2016/04/27121634/MOA_HS_Brief.pdf.

Appendix A – Tables

Table 2: Cohort Size in Each of the School Types for the Class of 2020

School Size	Total Cohort
Large	18239
Medium	12858
Small	40058
Grand Total	**71155**

Table 3: Number of Students Graduated vs Not Graduated for the Class of 2020

Graduated	Did Not Graduate
16204	2035
10629	2229
31748	8310
58581	**12574**

Table 4: Graduation Information Based on Gender of Students Across all Cohorts

Gender	Total Cohort	Total Grads
Female	446136	335961
Male	467960	304393
Grand Total	**914096**	**640354**

Table 5: Economic Status for Students Across 12 Cohorts

Category	Cohort Year	% Grads	Not Econ Disadvantaged		
Econ Disadvantaged				2005	53.44197035
	2005	54.90985171		2006	63.38415985
	2006	66.66265488		2007	64.70087357
	2007	66.55098343		2008	66.42395859
	2008	65.06971283		2009	67.32970658
	2009	66.38227234		2010	70.27947464
	2010	68.94467316		2011	71.59451447
	2011	70.42955094		2012	73.06860733
	2012	73.17104645		2013	72.94371033
	2013	75.0076416		2014	78.35296326
	2014	74.79967346		2015	80.64116516
	2015	76.247789		2016	80.47876587
	2016	78.24893951			

Table 6: Cohort Graduation Rate based on Student Race

Cohort Year	Grad Rate	Cohort Year	Grad Rate
White		Non-White	
2005	76.3145706	2005	56.4996981
2006	77.8874878	2006	64.428061
2007	78.7951736	2007	64.5275897
2008	78.3294205	2008	62.1959386
2009	79.8188019	2009	61.1891197
2010	80.7161499	2010	69.3869096
2011	81.4755722	2011	70.9503677
2012	81.616481	2012	74.3149158
2013	82.5049439	2013	74.905753
2014	83.9705887	2014	75.214147
2015	84.949382	2015	77.4240346
2016	84.2536728	2016	77.9782596

Appendix B – Code

All code is provided within the .rmd file submitted along with this project folder.