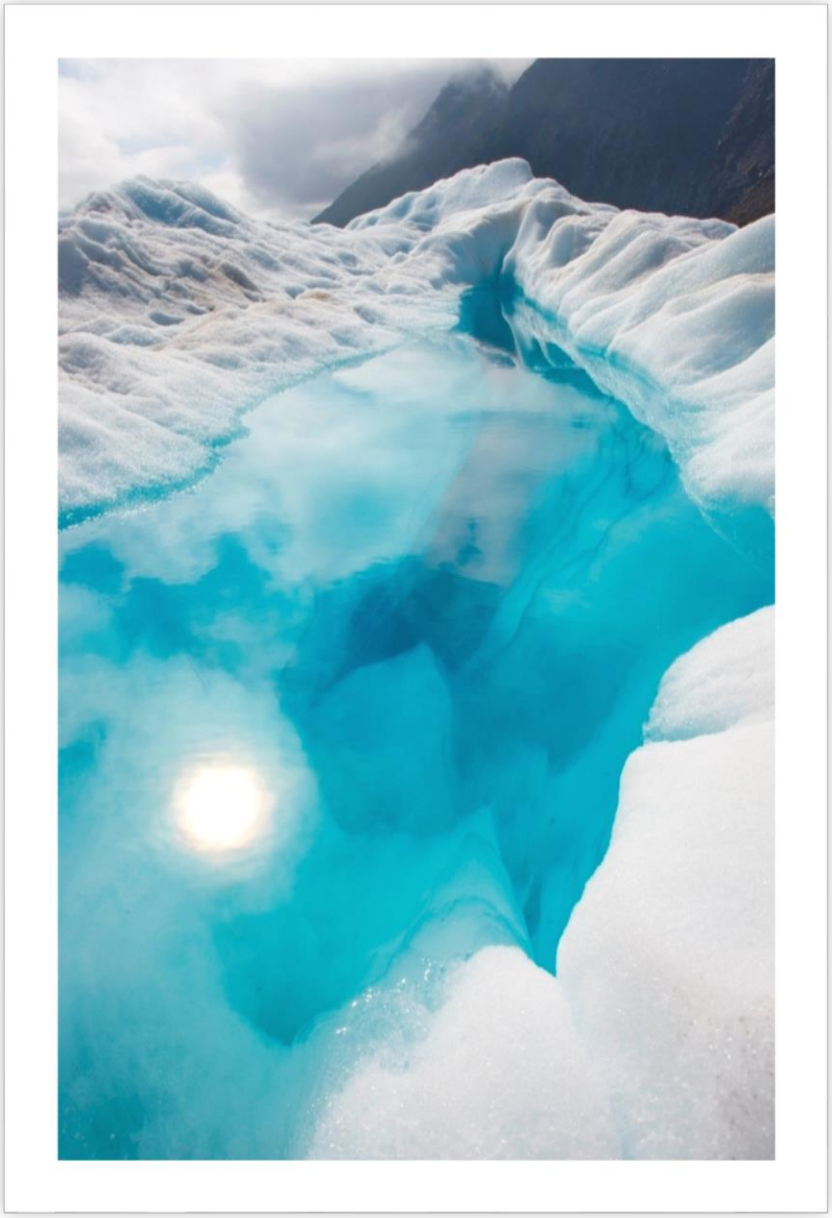


AP Statistics Final Project

ADMISSIONS RATE V. COMPLETION RATE

[redacted] | AP Statistics

Introduction

Being a Junior at WTW High School, many of my conversations with peers, parents, and family have been about College, and my readiness for it. Undoubtedly, my experience is similar to those of many of my classmates as they endure the pressure of post-secondary education, and their track to the rest of their lives.

Late last year, while looking into potential colleges to visit, and majors that I might be interested in. With President Obama's call for expansive improvements in education, the Department of Education released a fully featured, all-encompassing, interactive census of colleges, universities, and schools that operate in the U.S. The tool allowed me to easily define parameters that potential college-bound students would be interested in: including post-graduation median salary, acceptance rate, average annual cost; to name a few.

When assigned this project, I searched through many publicly available datasets, but found myself back at CollegeScorecard, on a subject that I was truly interested in. Admission rates, cost of attendance, retention rates, and standardized test scores have grown to be a part of students' daily thoughts (and anxieties).

The Department of Education (DOE) has made its datasets on all schools in the U.S. & Territories public from 1996 to 2013, and constantly works to make sure that students are able to attain the most updated and relevant information. With such an expansive dataset in-hand, it is critical to parse the data into a comprehensible subset. I used the DOE's 2013 dataset in order to perform my analysis.

In this study, I intend to focus on the relationship between the Admissions Rates and 4YR Completion Rates of post-secondary institutions as collected by the DOE. As both of my parameters are quantitative in nature, and I hope to find a correlation between the two parameters, I intend to perform a test of Regression on the sample dataset. I would assume that institutions that admit less students, and are considered to be 'more selective', would have higher 4YR completion rates than institutions that are less selective, and admit more students.

Procedure & Design

The Department of Education's (DOE) CollegeScorecard census contains extensive data about practically every statistic one could expect from a college; collected from all schools that operate within the U.S. & Territories.

With a population dataset of all colleges and universities in the U.S. in hand, choosing a variable to test was difficult. With hundreds of recorded parameters, deciphering the column notation's meaning using the DOE's data dictionary, and determining whether it was relevant or not, took countless hours. After trimming the dataset to the parameters: Institution Name, City, State, Admissions Rate, In-State Tuition, Out-of-State Tuition, and 4yr Completion rate. After much deliberation, I decided that an analysis of Admissions Rates and Completion Rates would be the most practical to study further.

Once I located my parameters of interest, I had to remove schools that did not report their Admissions Rates, Completion Rates, or both. After manual review of 7804 schools that the DOE published in their dataset, only 1771 schools had reported data for both Admissions Rates and Completion Rates.

In this case, nonresponse bias is unavoidable, as the reporting of school statistics is not mandatory for many institutions. Given the lack of mandated reporting in many institutions, it is also possible that some schools may avoid reporting unfavorable data. In my manual review, I could observe a pattern of niche schools, private for-profit institutions returning little to no information on their admissions or completion rates.

My sampling plan was one of a simple random sample, taken from a population of schools operating in the United States & Territories, where 150 numbers would be generated by Excel given a range from 1 to 1771 inclusive. The generated numbers would then be assigned to schools based on the alphabetic order in which the data was compiled by the DOE. Because there was no bias in the ordering of the schools when the dataset was compiled, and while parsing the dataset, it is reasonably safe to select schools based on random number generation.

Data Exploration

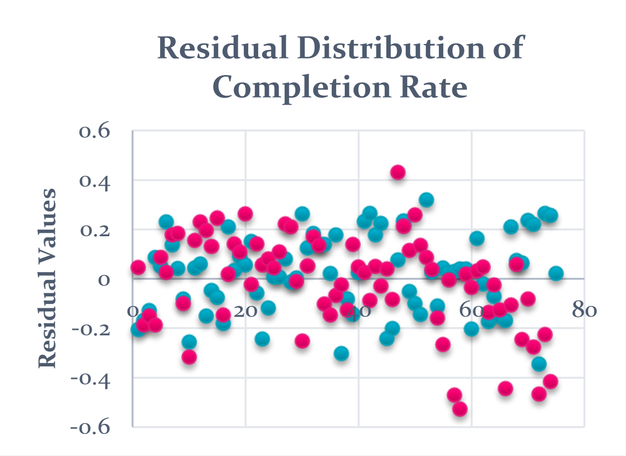
After compiling and sanitizing the dataset from schools that did not respond, I found the key descriptive statistics of each parameter.

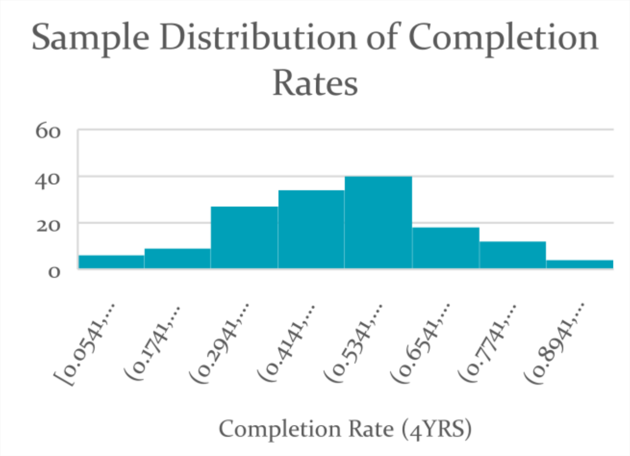
Before I could perform a linear regression analysis on my sample data, I made sure that:

1. The scatterplot of Admissions Rate and Completion Rate showed a linear relationship between the variables

2. That for each value of the Admissions Rate, the probability distribution of each Completion Rate had the same standard deviation as its complementary Admissions Rate

3. That each of the two parameters from each school would be independent of one another as proven by the quasi-random scatter of values on the scatterplot. In addition, the Completion Rates show a roughly normal distribution as shown by the histogram



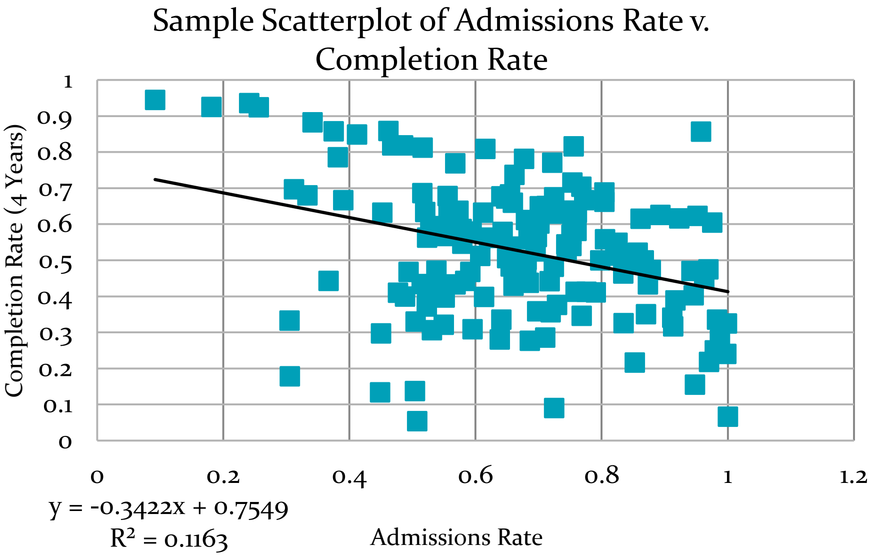


After checking conditions, I compiled some summary statistics on each of the parameters in the sample.

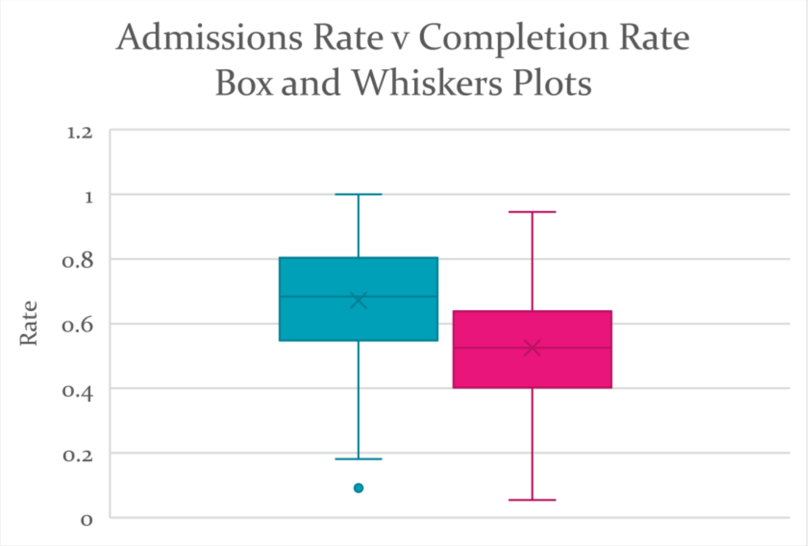
Admissions Rate		Completion Rate	
Mean	0.6722	Mean	0.5248
Standard Error	0.0153	Standard Error	0.0153
Median	0.6842	Median	0.5255
Mode	0.6411	Mode	0.5
Standard Deviation	0.187	Standard Deviation	0.1877
Sample Variance	0.035	Sample Variance	0.0352
Range	0.9082	Range	0.8911
Minimum	0.0918	Minimum	0.0541
Maximum	1	Maximum	0.9452
Sum	100.83	Sum	78.723
Count	150	Count	150

Data Exploration

After preliminary statistical analysis, I used Excel to build a scatterplot to compare the two parameters of interest; with Admissions Rate on the X-Axis and Completion Rate on the Y-Axis.



In addition to providing a preliminary look at the 150 schools and their distribution across the graph, I calculated the line of least squares regression for this scatter and found a slight negative trend to the data. In addition, I calculated an R^2 value of 0.1163, which gives us insight as to the strength of the correlation shown.



In drawing a box and whiskers plot for this data, we can see that there is slight overlap between the middle 50% of each distribution. There seems to be one outlier in the Admissions Rate sample data, but no outliers in the Completion Rate sample data. While Completion Rate looks to be roughly symmetrical, the plot for Admissions Rate is skewed to the left.

Analysis

I decided to perform a test of regression on my sample data, as the study was designed to find a relationship, if any, between Admissions Rates and Completion Rates in post-secondary education. After much data parsing and computer analysis, the data presented a few indicators worth noting. Given the general scope of this study, underlying factors that form the conclusion may not be clear.

When I built a scatterplot of all of the sample data for Admissions Rates and Completion Rates, I was at first surprised to see a negative trend to the data, but upon further inspection, I realized that the trend was as I had inferred; with more selective schools having higher completion rates than less selective schools. However, the scatter did return a correlation coefficient (R^2) of just 0.1163. This R^2 value tells us that this is a relatively weak correlation.

When comparing the two distributions' box and whisker plots, I found that the two distributions did have some overlap, and did have mean values that were close to one another. While building the plots for both sample datasets, I found that the plot for Admissions Rates was left-skewed, and had an outlier past its lowest (leftmost) value, and that the plot for Completion Rates was roughly normally distributed with no outliers. The condition of the Y parameter being normal allowed me to conduct the regression test.

The scatter is weakened more so by institutions that report having lower completion and admissions rates, than institutions with high admissions rates and high completion rates. This observation seems to be valid given that schools with high acceptance rates and high completion rates would be extremely favorable, and due to the lack of resources for the influx of students, the schools will be forced to reduce admissions rates once again.

Conclusion

As I had initially predicted, that sample data confirmed that there is in fact a negative trend between Admissions Rate and Completion Rate. The scatterplot of the sample of 150 institutions from a population of 1771 schools that reported the parameters of interest, showed the regression line ($y = -0.3422x + 0.7549$), showing a negative trend, however slight. In addition to the trend line that was generated, the scatterplot showed a correlation coefficient of $R^2 = 0.1163$, which means that there is a relatively weak correlation between Admissions Rate and Completion Rate for Institutions in the U.S.

With this analysis, we can affirm my prediction that higher completion rates are to be expected with schools that are more selective rather than schools with less selective admissions. The low r2 value can be explained by the multitude of factors that may influence a student, and by extension, their school's completion rate.

I feel as though this study was meaningful and opens the door to many other potential studies to be performed on the extensive dataset that the U.S. DOE has provided for the public with CollegeScorecard. The project was very time-intensive due to the immense amount of parsing required to shape the dataset into a valid, meaningful population to study. In the future, I would hope that more schools report their statistics to the DOE, to give students are more accurate insight into the institution.

Raw Sample Data (150 Schools)

Admissions Rates: (0.9703, 0.9796, 0.7293, 0.7011, 0.3123, 0.375, 0.7409, 0.8667, 0.7587, 0.6385, 0.7441, 0.857, 0.5231, 0.6558, 0.776, 0.4773, 0.5681, 0.4524, 0.6798, 0.5576, 0.6547, 0.6807, 0.6863, 0.494, 0.6565, 0.695, 0.5727, 0.5234, 0.5545, 0.7217, 0.7285, 0.9754, 0.6411, 0.7609, 0.5285, 0.9228, 0.45, 0.6844, 0.8349, 0.9625, 0.1813, 0.6154, 0.7726, 0.4681, 0.5953, 0.6411, 0.825, 0.5159, 0.917, 0.66, 0.5148, 0.7554, 0.5936, 0.5838, 0.643, 0.8546, 0.5778, 0.9423, 0.3333, 0.5229, 0.3816, 0.6811, 0.9975, 0.7908, 0.7189, 0.5509, 0.6617, 0.7523, 0.5572, 0.412, 0.7538, 1, 0.2411, 0.2561, 0.6217, 0.5496, 0.3904, 0.3668, 0.5429, 0.4884, 0.7604, 0.7442, 0.8932, 0.8044, 0.538, 0.3053, 0.8619, 0.4853, 0.9522, 0.6591, 0.3414, 0.6133, 0.8777, 0.713, 0.7536, 0.4616, 0.8731, 0.7539, 0.8532, 0.8051, 0.7523, 0.5156, 0.0918, 0.8041, 0.5791, 0.5051, 0.969, 0.7249, 0.7412, 0.9121, 0.7682, 0.7178, 0.6499, 0.5682, 0.6487, 0.8514, 0.6369, 0.9987, 0.6913, 0.9459, 0.8216, 0.5917, 0.9574, 0.7676, 0.5556, 0.6769, 0.7018, 0.6119, 0.6839, 0.6976, 0.5307, 0.8342, 0.3056, 0.5075, 0.7978, 0.6077, 0.8523, 0.6973, 0.9873, 0.7239, 0.9132, 0.5037, 0.8703, 0.5201, 0.8525, 0.9831, 0.9478, 0.4486, 0.7105, 0.7246)

Completion Rates: (0.2182, 0.25, 0.3768, 0.6022, 0.6972, 0.858, 0.6379, 0.5, 0.4129, 0.281, 0.5443, 0.5212, 0.4255, 0.4828, 0.4128, 0.4105, 0.7688, 0.6321, 0.6115, 0.6201, 0.683, 0.4638, 0.2769, 0.4679, 0.5381, 0.5247, 0.6376, 0.563, 0.5678, 0.7705, 0.6299, 0.6049, 0.6772, 0.6327, 0.5946, 0.6162, 0.2979, 0.4377, 0.3255, 0.45, 0.9254, 0.8094, 0.6667, 0.8187, 0.309, 0.3353, 0.5489, 0.8124, 0.3892, 0.4295, 0.4341, 0.816, 0.5797, 0.4435, 0.5789, 0.4891, 0.5868, 0.4708, 0.68, 0.372, 0.7861, 0.5012, 0.241, 0.4112, 0.3565, 0.3968, 0.7373, 0.5714, 0.6277, 0.8494, 0.7158, 0.0667, 0.9365, 0.9246, 0.5631, 0.3215, 0.6667, 0.4431, 0.419, 0.4, 0.5821, 0.5263, 0.6261, 0.6629, 0.4714, 0.3333, 0.6156, 0.819, 0.6231, 0.6598, 0.8832, 0.3984, 0.4733, 0.6511, 0.6047, 0.8584, 0.4333, 0.6378, 0.5189, 0.5587, 0.5398, 0.6864, 0.9452, 0.6885, 0.5464, 0.3303, 0.4747, 0.6754, 0.6364, 0.3409, 0.346, 0.4422, 0.5061, 0.4336, 0.6716, 0.5116, 0.5619, 0.3254, 0.5681, 0.4034, 0.5132, 0.4682, 0.8571, 0.7045, 0.6786, 0.7815, 0.6501, 0.6324, 0.5582, 0.3586, 0.3068, 0.4638, 0.1789, 0.0541, 0.5, 0.5123, 0.4943, 0.5646, 0.2794, 0.4824, 0.3168, 0.1379, 0.3509, 0.6338, 0.2167, 0.3359, 0.1552, 0.1343, 0.2857, 0.0909)

Sources

1. Data: https://collegescorecard.ed.gov/data/
2. Research: http://www.theatlantic.com/education/archive/2015/09/obamas-new-college-scorecard-flips-the-focus-of-rankings/405379/
3. Research: http://www.npr.org/sections/ed/2015/09/12/439742485/president-obamas-new-college-scorecard-is-a-torrent-of-data