
ISyE 2028 – Basic Statistical Methods - Fall 2015
Bonus Project: “Big” Data Analytics
Final Report

In a casual conversation the other day, I happened to know a friend of mine graduated with over 100,000 dollars loan from Georgia Tech. I was totally shocked by this number, although I knew high college tuition has always been an issue for many American families, and many students more or less carry some student loan. Imagine the pressure on a newly grad. That subject intrigues me. I want to find out the proportion of students graduated with loan, and the average amount of loan student graduated with. More importantly, I want to compare those data between public schools and private schools to see if there's a difference.

With these questions in mind, I began my research. A website I found particular useful for my project is ticas.org. It has a list of data showing the information about the colleges, including tuition, bachelor's degrees recipients, the proportion of students graduate with loan, and the average amount of student loan etc. The data is categorize by states and includes most official colleges in the state, public and private. My original plan was to randomly select three states to do the data analysis. However, I found that the data between each state could be quite different. For example, the average loan in Pennsylvania is \$33,264, while the average loan in California is \$21,382. With this level of uncertainty, randomly selecting three states is not a proper estimator for the entire country. To improve the accuracy, I decided to choose only one state---Pennsylvania, for my project.

My original assumption was that private schools has lower proportion of students graduated with loans, since students who go to private schools tend to have higher household income, and students from low income families tend to go to state schools. In addition, I expected to see private schools has higher average student loan, since private schools have higher tuition.

From the 33 samples of public schools in Pennsylvania, the average loan of graduates for 2014 has a mean of 33441.18, with a standard deviation of

5405.45. The proportion of students graduates with loan has a mean of 72.09%, with a standard deviation of 0.098. From the 33 samples of private schools in Pennsylvania, the average loan of graduates for 2014 has a mean of 31785.79, with a standard deviation of 6260.99. The proportion of students graduates with loan has a mean of 70.39%, with a standard deviation of 0.18.

From the data, we can see that the difference in the mean of the average loan of public schools and private schools is significant. Opposed of my assumption, the mean of average loan of public schools are 1656 dollars higher than that of private schools.

Use the formula for confidence interval :

$$\bar{x} - z_{\alpha/2} \frac{s}{\sqrt{n}} \leq \mu \leq \bar{x} + z_{\alpha/2} \frac{s}{\sqrt{n}}$$

Since $n > 30$, we can use z distribution to estimate t distribution. After computation, the 95% confidence interval for average student loan for public schools in Pennsylvania is [31596.88, 35285.48]. The 95% confidence interval for average student loan for private schools in Pennsylvania is [29649.59, 33921.99].

On the other hand, the mean of the proportion of students graduated with loan from public and private schools are very close. To find out the if the difference of the two means is significance enough, a hypothesis test is performed using $\alpha=0.05$.

1. Parameter of Interest: the mean of the proportion of students graduated with loan.
2. Null hypothesis: $H_0: \mu_1 - \mu_2 = 0$
3. Alternative hypothesis: $H_1: \mu_1 > \mu_2$
4. Test Statistics:

$$t_0^* = \frac{\bar{x}_1 - \bar{x}_2}{\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}}$$

After computation, $t_0^* = 0.56$

5.

$$v = \frac{\left(\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2} \right)^2}{\frac{(s_1^2 / n_1)^2}{n_1 - 1} + \frac{(s_2^2 / n_1)^2}{n_2 - 1}}$$

After computation, $v=49$

Reject H_0 if $t_0^* > t_{0.05,49} = 1.677$.

6. Conclusion: $t_0^* < t_{0.05,49}$. We do not reject H_0 .

Therefore, there is no strong evidence to conclude that the proportion of students graduated with loan in public school is higher than the proportion of students graduated with loan in private school in Pennsylvania.

In conclusion, public schools in Pennsylvania has higher average student loan than private schools, and the proportion of students graduated with loan in public and private schools are basically the same. I am a bit surprised when I see the results, since they are both different from my assumptions. I believe my calculations and conclusions are reliable and reasonable, since the debt data were reported by campuses to Peterson's Undergraduate Financial Aid and Undergraduate Databases. One thing that could have harmed my analysis is that some schools did not report the data, so that my samples could only be chosen from the ones that have reported the data, which made my samples biased.