# Final Project

Emma Resmini

2020-12-07

## Question of Interest

*Is there a difference in student loan debt between for-profit private insitutions and non-profit private institutions?*

Attending college/university is becoming increasingly expensive. (The school I attended before GMU had a hefty price tag of just under $50,000/year.) Many people have no choice but to take out student loans, which then plague them for years even after they've left school. Upon seeing this dataset, I found myself interested in the differences between non-profit and for-profit private institutions. For-profit institutions market themselves as schools that can prepare you to directly enter the workforce upon graduation. However, for-profit institutions are often accused of overcharging students, and encouraging them to on unrealistic amounts of debt while not adequately preparing them for the workforce.

*Columns used*: GRAD_DEBT_MDN, CONTROL, PREDDEG

This project will use statistical inference.

## Preprocessing

The entire college.rds dataset is not needed for this test. Therefore, we will first filter to only include rows whose value in the CONTROL column is 2 or 3. Then, using, the mutate function, make a new column of the same values, but now the values 2 and 3 are renamed "NonProfit" and "ForProfit", respectively. Next, mutate another column copying the values from the PREDDEG dataset, and recode the values 0-4 to their corresponding degree (as seen below). Finally, rename the column GRAD_DEBT_MDN to a more readable name, and select that and the two new mutated columns to form the new dataset.

```
For_Non <- college %>%

  filter(CONTROL == 2 | CONTROL == 3) %>%

  mutate(Ownership_Private = if_else(CONTROL == 2, "NonProfit", "ForProfit")) %>%


    mutate(Predominant_Degree = recode(PREDDEG,
        `0` = "Not_Classified",
```

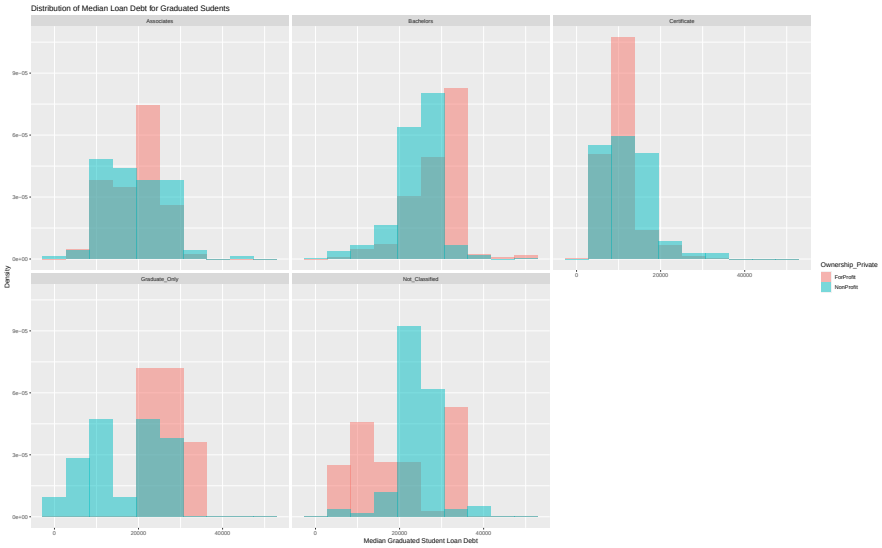
```
        '1' = "Certificate",
        '2' = "Associates",
        '3' = "Bachelors",
        '4' = "Graduate_Only"
    )
  ) %>%

  rename(Median_Debt_Grads = GRAD_DEBT_MDN) %>%

  select(Ownership_Private, Predominant_Degree, Median_Debt_Grads)
```

## Visualization

### Histogram

The histograms below show the distribution of median student debt for graduated students. With the histogram, we can determine the general shape of the distribution, and identify peak(s). We use the fill function to break the graph down into Non-Profit and For-Profit distributions so they can be compared, as well as facet-wrapping over the Predominant_Degree category so each individual degree program can be compared.

```
For_Non %>%
  ggplot() +
  geom_histogram(mapping = aes(
    x = Median_Debt_Grads, y = ..density.., fill = Ownership_Private),
                 position = "Identity", alpha = 0.5, bins = 10) +
  facet_wrap(~Predominant_Degree) +
  labs(title = "Distribution of Median Loan Debt for Graduated Sudents",
       x = "Median Graduated Student Loan Debt",
       y = "Density")
```



The first clear observation that can be made is the fact that each distribution shape for each type of degree is different. This likely contributes to the fact that different programs require different time commitments, and therefore students may accumulate more or less debt simply due to the amount of time in school.

For the rest of this analysis, we will be focusing on the median debt for *bachelor's programs*. We see the For-Profit distribution for bachelor's degrees is strongly left-skewed, while the Non-Profit distribution is roughly symmetrical. For-Profit median debt has a peak around $30,000 while Non-Profit median debt peaks earlier around $25,000, so already we can see that For-Profit schools may result in higher loan debt.

### Modify Dataset:

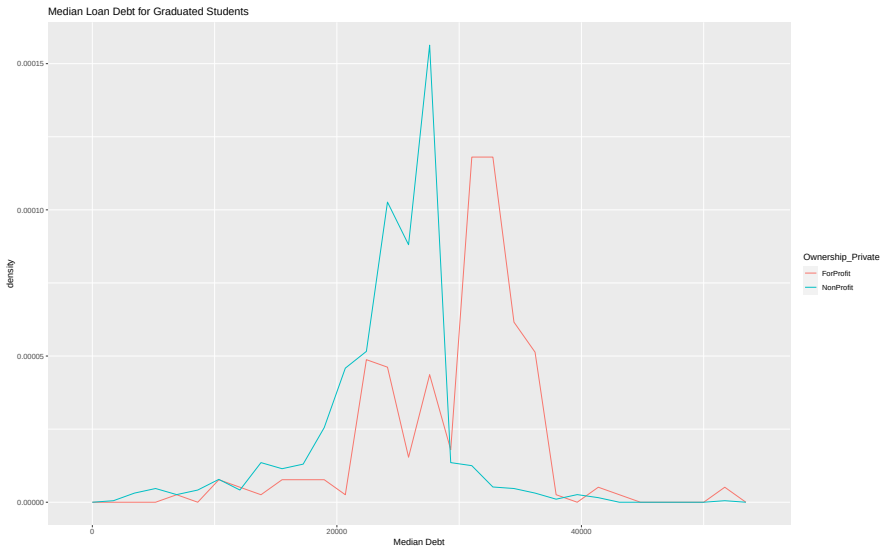
The dataset of interest is now being further filtered to only include rows whose predominantly-offered degree is bachelor's:

```
For_Non <- For_Non %>%
  filter(Predominant_Degree == "Bachelors")
```

### Frequency Polygon

Now that we are only looking at median student loan debt from bachelor programs, below we have a frequency polygon to compare the distributions between Non-Profit and For-Profit. Frequency polygons show similar information as histograms, but the use of lines eliminates some of the overlap that can make histograms difficult to read, as well as giving a more accurate presentation of the shape of the distribution.

```
For_Non %>%
  ggplot(mapping = aes(x = Median_Debt_Grads, y = ..density..)) +
  geom_freqpoly(mapping = aes(colour = Ownership_Private), bins = 30) +
  labs(title = "Median Loan Debt for Graduated Students",
       x = "Median Debt")
```



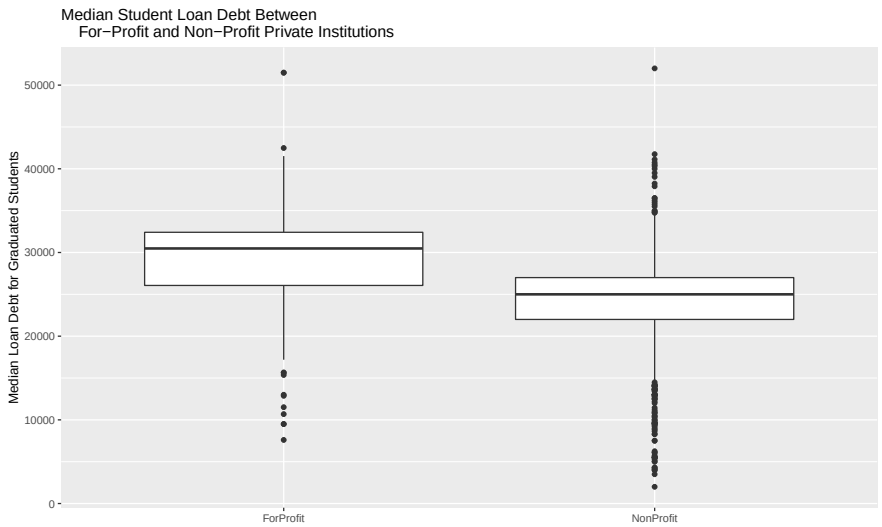
The For-Profit distribution has several peaks, while the Non-Profit distribution has few peaks as well as one main peak around $27,000. As we saw from the histograms, the distributions do not follow a similar pattern to each other, however both have thin tails on both sides of their peak(s).

### Boxplot

Next, we have boxplots, which can give a clearer picture of the spread of the distribution, as well as the presence of outliers, if any.

```
For_Non %>%
  ggplot() +
  geom_boxplot(mapping = aes(
    x = reorder(Ownership_Private, Median_Debt_Grads, FUN=median),
    y = Median_Debt_Grads)) +
  labs(
    title = "Median Student Loan Debt Between
    For-Profit and Non-Profit Private Institutions",
    y = "Median Loan Debt for Graduated Students",
    x = " ")
```

```
## Warning: Removed 165 rows containing non-finite values (stat_boxplot).
```



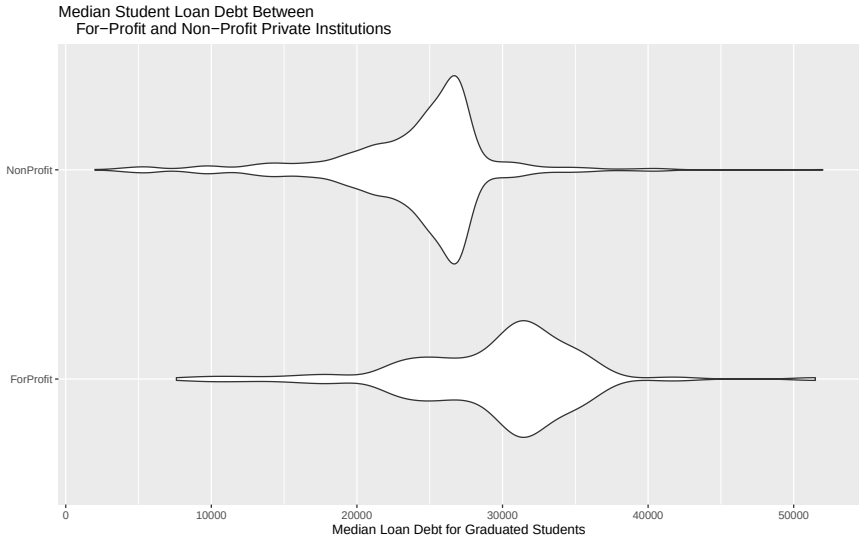

Starting with the For-Profit distribution, we immediately see that the median sits just above $30,000 and the interquartile range spans from approximately $26,000 - $35,000. Additionally, by evaluating the position of the median inside the box, we see the distribution is left-skewed. There are both lower and upper extreme values, however there are more low-value outliers.
The Non-Profit distribution has its median just around $25,000, and the interquartile range spans approximately from $22,000 to $27,000. The distribution skews slightly to the left, but could still be considered roughly symmetric. There are many outliers present, both lower extreme and upper extreme values.

### Violin Plot

The violin plot appears similar to the boxplot, but now we will be able to see density. In other words, seeing where lower counts and higher counts of median debt lie on the distribution (making this closer to a histogram, actually).

```
For_Non %>%
ggplot() +
  geom_violin(mapping = aes(x = Median_Debt_Grads, y = Ownership_Private))+
  labs(
    title = "Median Student Loan Debt Between
    For-Profit and Non-Profit Private Institutions",
       x = "Median Loan Debt for Graduated Students",
       y = " ")
```

```
## Warning: Removed 165 rows containing non-finite values (stat_ydensity).
```




The Non-Profit distribution has the highest density around $27,000. Also, we see very thin tails on both sides of the peak (another way to see the outliers from the boxplot). There is a quick increase in density going to the peak, and then a stark drop after the peak, which causes the thin tails.
The For-Profit distribution has its highest density around $32,000. This distribution has a more gradual rise in and fall in density compared to Non-Profit, but still also has thin tails on both sides.

### Summary Statistics

To generate summary statistics, we will need to produce the sum, standard deviation, mean, median, interquartile range (IQR), minimum, and maximum of each variable.

First, we want to see how many total rows of data we are handling. Since the original data was filtered by Predominant_Degree, we use group For_Non by that variable and use the summarise() function with n() to see our grand sum of 1,504.

```
For_Non %>%

  group_by(Predominant_Degree) %>%

  summarise(sum = n())
```

```
## `summarise()` ungrouping output (override with `.groups` argument)
```

| Predominant_Degree | sum |
|---|---|
| Bachelors | 1504 |

Next, we look at Median_Debt_Grads, grouped by Ownership_Private. Use n() for the sum of each category, mean(), median(), sd() for standard deviation, IQR() for interquartile range (middle 50%), min(), and max().

```
For_Non %>%

  group_by(Ownership_Private) %>%

  summarise(sum = n(),
  mean = mean(Median_Debt_Grads, na.rm = TRUE),
  median = median(Median_Debt_Grads, na.rm = TRUE),
  std = sd(Median_Debt_Grads, na.rm = TRUE),
  IQR = IQR(Median_Debt_Grads, na.rm = TRUE),
  min = min(Median_Debt_Grads, na.rm = TRUE),
  max = max(Median_Debt_Grads, na.rm = TRUE)
)
```

```
## `summarise()` ungrouping output (override with `.groups` argument)
```

| Ownership_Private | sum | mean | median | std | IQR | min | max |
|---|---|---|---|---|---|---|---|
| ForProfit | 244 | 29429.96 | 30481 | 6150.042 | 6353.125 | 7600 | 51488.5 |
| NonProfit | 1260 | 24012.78 | 25000 | 5399.549 | 5000.000 | 2000 | 52000.0 |

The number of For-Profit rows is 244, and the number of Non-Profit rows is 1,260.

When analyzing the median debt by ownership, we see that the For-Profit data has a higher mean, median, and standard deviation than the Non-Profit data. The larger standard deviation suggests more variation in the data compared to Non-Profit. Additionally, the For-Profit IQR (the middle 50%), 6353.125, has a wider range than the Non-Profit IQR, 5000. Interestingly, the overall range of the For-Profit data (max - min), 43888.5, is actually smaller than Non-Profit's, 50000. However, remember that the total sum For-Profit data rows is less than Non-Profit (1260, compared to 244).

This may explain why the Non-Profit data has a larger overall range, but less variation than the For-Profit data.

### Data Analysis

$H_0$: There is no difference in median student loan debt between bachelor's programs of non-profit private institutions and for-profit private institutions.

$H_a$: There is a difference in median student loan debt between bachelor's programs of non-profit private institutions and for-profit private institutions.

The hypothesis test will be two-sided, and will use a significance level ($\alpha$) of 0.05.

*Test Statistic*: -5,481

First, we are going to find the observed test statistic. Above, we saw the test statistic calculated by hand is -5,481, which was found by subtracting the medians of NonProfit and ForProfit. In the code chunk below, we use to specify() function to say which variable is the response, and which is the explanatory. Since we are looking at how the two different types of private institutions effect median loan debt for students who graduated, Median_Debt_Grads is the response variable and Ownership_Private is the explanatory.
Then, with the addition of the calculate() function, we will be given our test statistic. Since we are asking about median student loan debt, inside the function we write *stat = "diff in medians"*, and for this test the order will be NonProfit, ForProfit (meaning the difference is always NonProfit minus (-) ForProfit). Name this new data frame (ForNon_stat, for example, as shown below), and run the code.

```
ForNon_stat <- For_Non %>%
specify(Median_Debt_Grads ~ Ownership_Private) %>%
calculate(
  stat = "diff in medians", order = vctrs::vec_c("NonProfit", "ForProfit"))

ForNon_stat
```

| stat |
|---|
| -5481 |

The test statistic is *-5481*, identical to the number calculated by hand.

Now, we run the hypothesis test! First, we again use specify() with Median_Debt_Grads as the response variable and Ownership_Private as the explanatory variable.

Next, use the hypothesize() function and type *null = "independence"* to show that we are using two samples.

Third, within the generate() we specify how many times (reps) to repeat the test. We will use 10,000. Also, for "type", we choose *permute*, which will sample without replacement.

Finally, with again use the calculate() function, filling in with *stat = "diff in medians"*, and the order NonProfit, ForProfit. Name this new data frame (ForNon_null, for example, as shown below),

and run the code.

```
ForNon_null <- For_Non %>%
specify(Median_Debt_Grads ~ Ownership_Private) %>%
hypothesize(null = "independence") %>%
generate(reps = 10000, type = "permute") %>%
calculate(
  stat = "diff in medians", order = vctrs::vec_c("NonProfit", "ForProfit"))
```

Now, we obtain the p-value. In short, the p-value is the evidence against the null hypothesis. To see this value, we use get_p_value() with obs_stat set to ForNon_stat, and remember to mark the direction as "two_sided" since that is the type of hypothesis test that is being run.

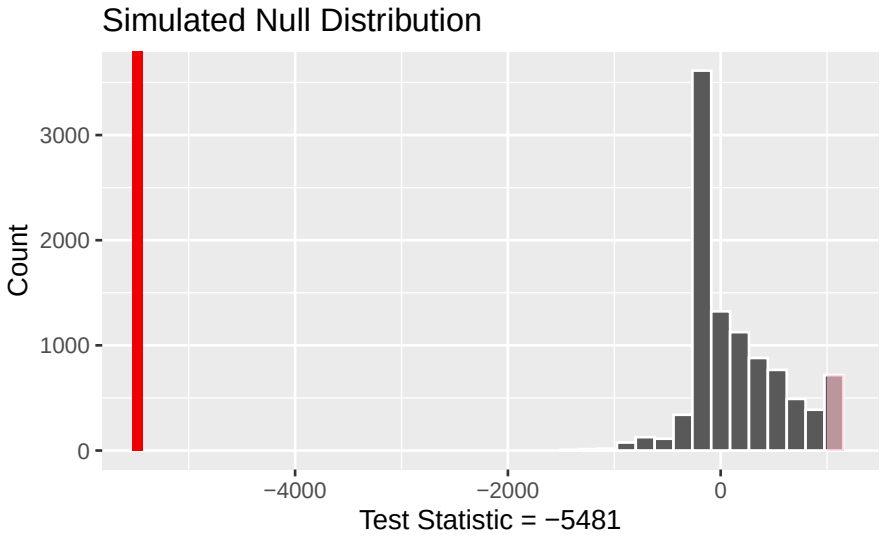
```
ForNon_null %>%
get_p_value(obs_stat = ForNon_stat, direction = "two_sided")
```

| p_value |
|---|
| 0 |

We obtain a p-value of 0. This means the probability of obtaining our test statistic (-5417.184), assuming the null hypothesis is true, is 0. The value is less than the established significance level, $\alpha = 0.05$, which means we can reject the null hypothesis.

To visualise the p-value, we use visualize() with the ForNon_null data frame, then shade_p_value() with ForNon_stat. Additionally, inside shade_p_value(), set the direction as "two_sided", and set the color and fill. Below, we see how far away the test statistic is from the null distribution.

```
visualize(ForNon_null) +
shade_p_value(ForNon_stat, "two_sided", color = "red2", fill = "pink") +
labs(
title = "Simulated Null Distribution",
x = "Test Statistic = -5481",
y = "Count"
)
```




From the hypothesis test, we reject the null hypothesis. There is sufficient evidence to conclude that there is a difference in median student loan debt between bachelor's programs of non-profit private institutions and for-profit private institutions.

*Note*: I debated whether to use "diff in medians" or "diff in means", and ultimately decided on medians since this test is answering the question of whether the two types of institutions have different *median* student loan debt. However, I did also run the test using "diff in means", and the output gave very similar results.

## Conclusion

As previously stated, sufficient evidence was gathered to reject the null hypothesis. There is reason to believe that there is a difference in median student loan debt between bachelor's programs of non-profit private institutions and for-profit private institutions. From the Visualization section, we saw the difference shapes of each category's distribution, and then with the summary statistics we saw that For-Profit had a higher mean and median than Non-Profit, as well as more variation. Finally, with the hypothesis test, we saw it would be highly unlikely for the median student loan debt between for-profit and non-profit institutions to be the same.

This test did not make a statement about non-profit and for-profit institutions in all regards. We only focused on bachelor's programs. As seen from the histograms in the Visualization section, the different degree programs have different distributions, and some program comparisons may show similar test results while others show the opposite, or one could choose to simply analyze institutions as a whole instead of breaking them down by degree. There are many approaches to determining if there really is a difference between non-profit and for-profit private institutions, and this hypothesis test only gave a possible answer to one part.