Project Part 3

Research Question

How does the average proportion of first generation students who completed all four years of their undergraduate degree at the university they are enrolled in compare across public and private universities in the United States?

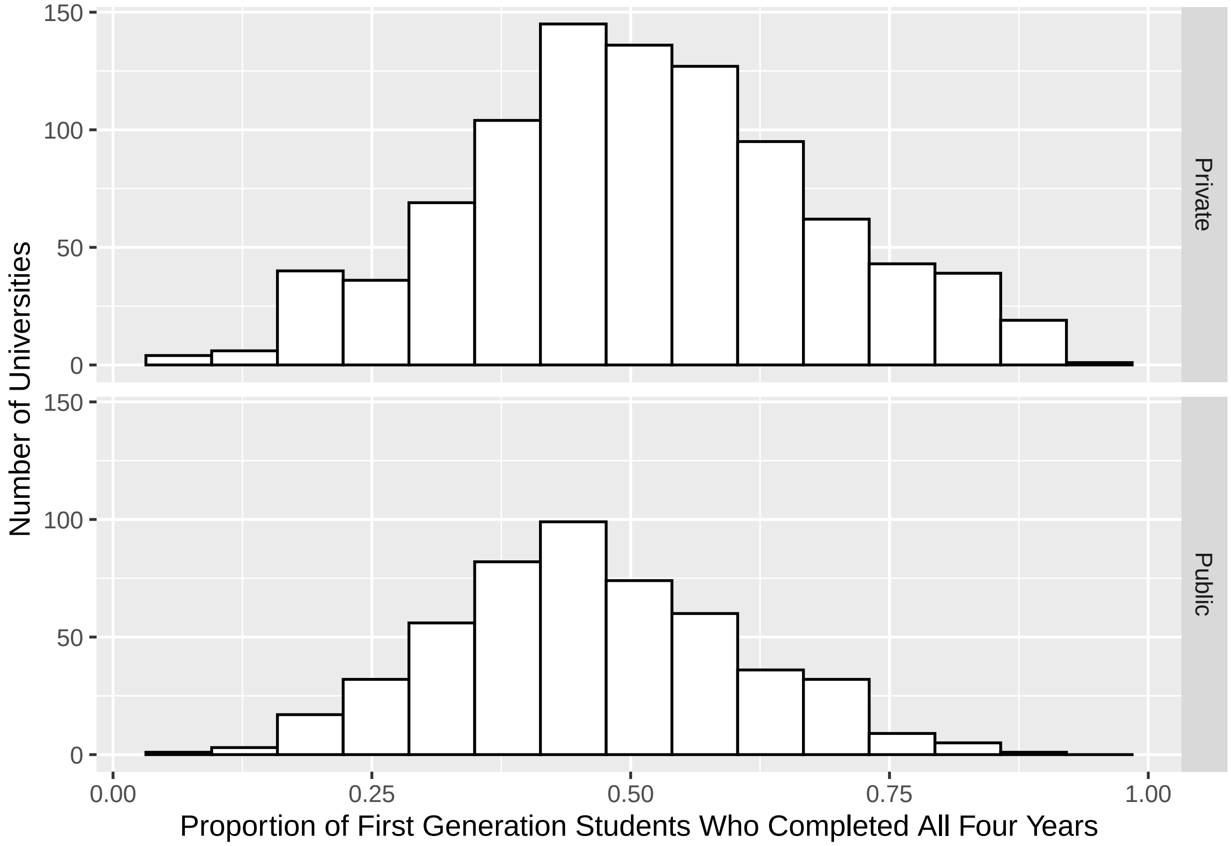
Data Description & Appropriateness

The data were collected from two primary sources, the 2019-2020 U.S. Department of Education College Scorecard and the 2020 The State Higher Education Executive Officers Association (SHEEO) State Higher Education Finance (SHEF) Report (Reference 1 & Reference 2). When the two data sets were merged into one data set of U.S. university attributes on the basis of state and rows with missing values were removed, the number of universities decreased from 6,694 to a sample of 1,433. The data are appropriate to address the research question because they include the variables of interest: the proportion of first-generation students who completed within four years at their university and whether the university is public or private. The data set was also subset to represent the two sample groups: public universities, with 507 observations, and private universities, with 926 observations.

About the Test

I used a two-sample t-test to address my research question, which is appropriate because I am comparing the means of a quantitative variable (the proportion of first generation students who completed all four years at their enrolled university) across two groups (public and private universities) with unknown population standard deviations. The parameter of interest for this test is the difference of the two population means and the sample statistic is the difference of the two sample means. The test statistic is calculated by dividing the difference between the first group's sample mean and the second group's sample mean by the pooled standard deviation multiplied by the square root of the sum of one divided by the first group's sample size and one divided by the second group's sample size (Reference 3). The pooled standard deviation is the square root of the sum of the first group's sample size minus 1, multiplied by its squared standard deviation, and the second group's sample size minus 1, multiplied by its squared standard deviation, divided by the two sample sizes plus 2 (Reference 3). The degrees of freedom is the first group's sample size plus the second group's sample size minus 2 (Reference 3). These characteristics are appropriate, as the parameter and statistic can be clearly stated, and the test statistic and degrees of freedom are able to be calculated since I know the sample means, standard deviations, and sizes for each group.

The assumptions of the two-sample t-test are the two populations are normal, independent, and have equal variances (Reference 3). As displayed in the two histograms below, the sample distributions of the proportion of first generation students who completed all four years for both public and private universities look normally distributed. The two groups' variances are also about equal since the spread of the two distributions are roughly similar. The two groups are also independent because the proportion of first generation students who completed all four years at public universities intuitively does not depend on that of private universities.



The null hypothesis is there is no difference between the average proportions of first generation students who completed all four years at private universities and at public universities in the United States. The alternative hypothesis is the average proportion of first generation students who completed all four years at their university is higher at private universities than at public universities in the United States.

Test Results and Conclusion

```
t.test(my_samp_private, my_samp_public, mu=0, alternative="greater", var.equal = TRUE)
```

```
##
##  Two Sample t-test
##
## data:  my_samp_private and my_samp_public
## t = 5.549, df = 1431, p-value = 1.709e-08
## alternative hypothesis: true difference in means is greater than 0
## 95 percent confidence interval:
##  0.0345989       Inf
## sample estimates:
## mean of x mean of y
## 0.5094866 0.4602974
```

The test statistic is 5.549 and the p-value is 1.709e-08, and the mean proportion of first generation students who completed all four years is about 0.51 for private universities and 0.46 for public universities. Because the p-value is less than the significance level of 0.05, I reject the null hypothesis. There is sufficient evidence to conclude that the average proportion of first generation students who completed all four years at their university is higher at private universities than at public universities in the United States.

Generally speaking, this outcome could be due to private universities often having more university specific resources and prestige than public universities, as I suspected when formulating my hypothesis. These factors could affect the decisions of first-generation students, an often vulnerable group of college students. These results and reasons can be generalized beyond the data, as the two groups are representative samples of public and private universities in the U.S., as private universities have many of the same attributes, as described earlier, and so do public universities, as they generally charge lower tuition for in-state students and depend more on government funding.

.

References

1. https://collegescorecard.ed.gov/data/
2. https://shef.sheeo.org/about/
3. https://www.statology.org/two-sample-t-test/