

# An analysis of the difference between STEM major and non STEM majors in the proportion of graduates who get a job that requires a college degree

Megan Grabill, Riyad Alam, SeBeom Lee

## **Introduction**

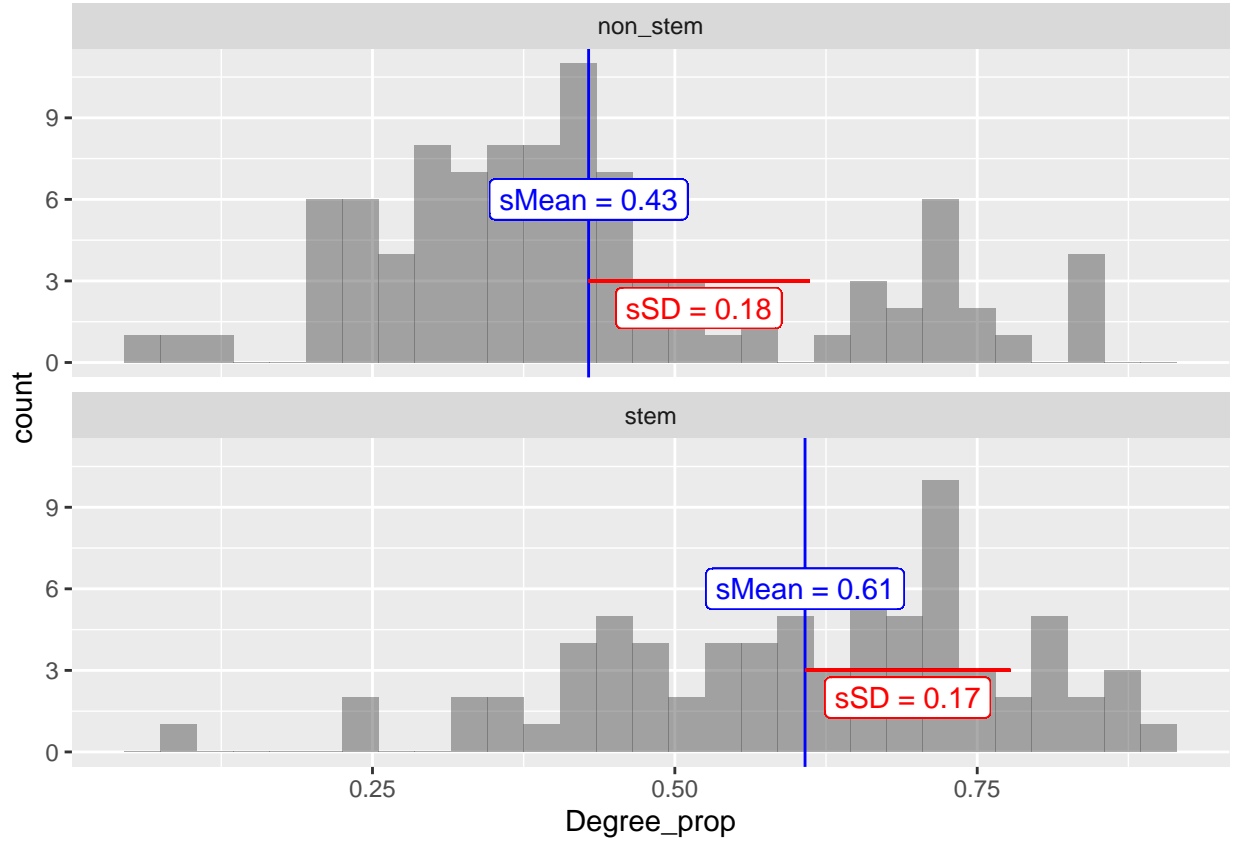
## **Data and Methods**

## **Results**

## **Parameters and Hypotheses**

## **Descriptive Analysis**

For the analysis, we divided the dataset into two groups. The group “STEM” has 76 cases, and the other group, “Non-stem” has 97 cases in it. We estimated the levels of the variable by calculating the proportion of the jobs that require a college degree out of the jobs that graduates with each major have. Figure shows the distribution of the variable in two groups.



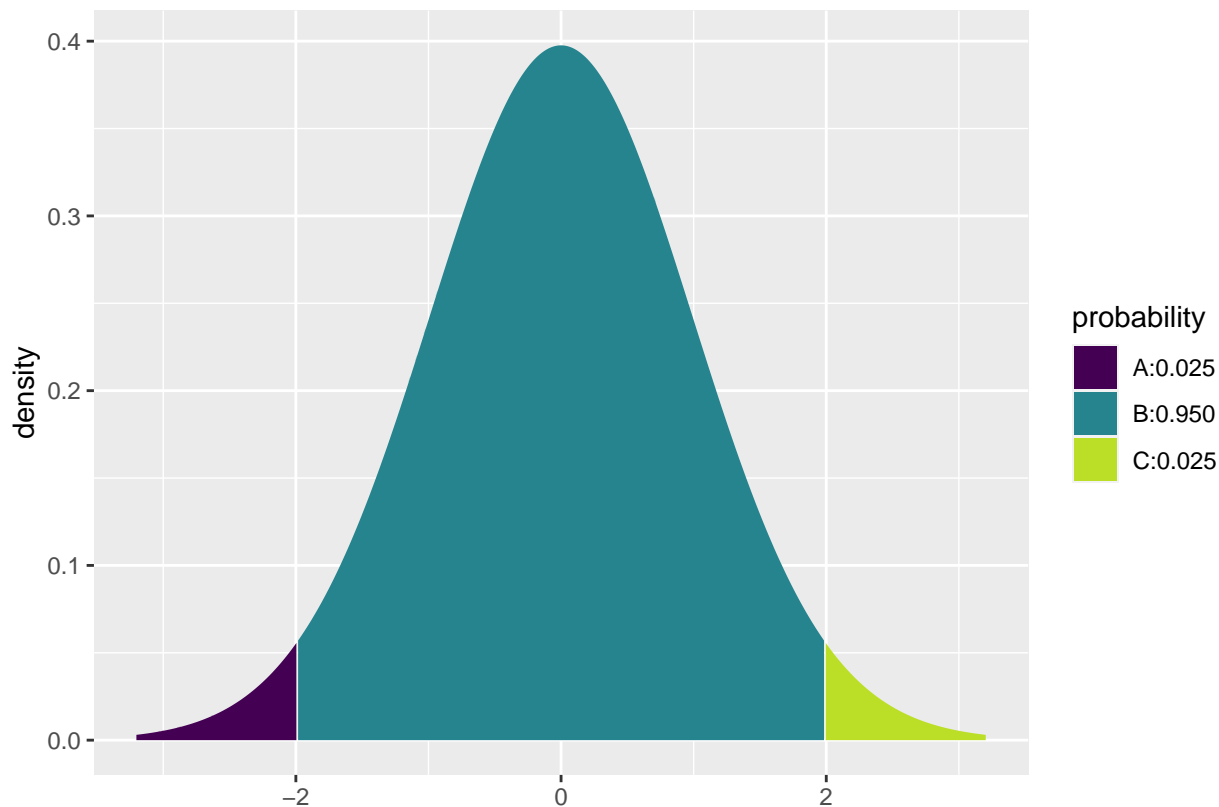
The mean of the variable in STEM group is 0.61 with a standard deviation of 0.17, and the mean of the variable in Non\_STEM group is 0.43 with a standard deviation of 0.18. There is one outlier in STEM group and 4 outliers in Non\_STEM group, but considering that the difference between the mean and the median is fairly small (stem: -0.029, non\_stem: 0.035), we decided that the effects of outliers is not extreme, and mean can be used as a representative statistic for each group.

Our statistic of interest  $x_{stem} - x_{non}$  is 0.18. The difference in mean is positive, which means that the mean of STEM group is bigger than that of Non-STEM group.

### Inferential Analysis

We want to find out the 95% confidence interval for the difference of means. We will use a t-distribution for the analysis, and it is appropriate since both samples we have are big enough. STEM group has 76 cases, which is bigger than 27, and Non\_STEM group has 96 cases, which is also bigger than 27.

To estimate the standardized endpoint for a 95% confidence interval, we use a t-distribution with degree of freedom of  $n_{stem} - 1 = 75$ .



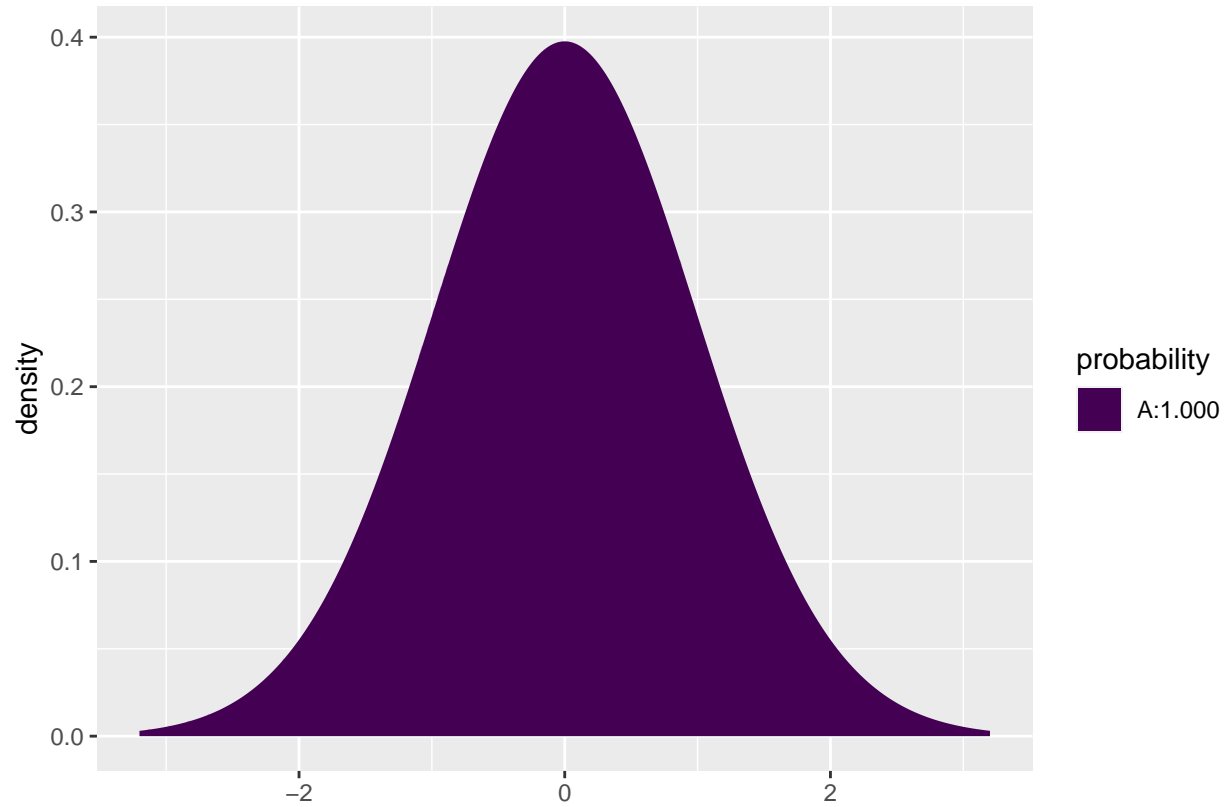
We could find out that the standardized endpoints are  $t_{end} = \pm 1.99$ .

To estimate the standard error, we used an equation  $\hat{SE} = \sqrt{\frac{s_{stem}^2}{n_{stem}} + \frac{s_{non}^2}{n_{non}}}$ . From the given sample, we could estimate the standard error of  $\hat{SE} = 0.03$ .

With the sample statistic  $x_{stem} - x_{non} = 0.18$  and standard error of  $\hat{SE} = 0.03$ , we calculated the 95% confidence interval using an equation of  $CI = SampleStat. \pm t_{end} \cdot \hat{SE}$ .

We could obtain a 95% confidence interval for the difference of means from 0.13 to 0.23, which means that we are 95% confident that the percentage of graduates who get a job that requires a college degree with STEM majors is from 12 to 23 percentage points higher than that of graduates with Non-STEM majors. One implication of the fact that the confidence interval does not include zero is that we are 95% confident that the difference between two groups is not zero.

Our null hypothesis is that  $H_0 : \mu_{stem} - \mu_{non} = 0$  and the alternative hypothesis is  $H_1 : \mu_{stem} - \mu_{non} > 0$ . To test the hypothesis we estimate how likely it is to observe the result as extreme as our observed sample. Using the sample statistic and the standard error estimated above, we set up a test statistic using an equation  $TestStat. = \frac{ObservedDiff. - NullDiff.}{SE}$ . With the observed difference of 0.18 null difference of 0 and standard error of 0.03, we calculated the test statistic of  $TestStat. = 6.64$ . Then, we estimated the P-value by a right-tailed test on a t-distribution with degree of freedom 76.



We obtained a P-value of  $2 \times 10^{-9}$ , which is extremely smaller than  $\alpha = 0.05$ , which means that if there were no difference between STEM group and Non-STEM group, the probability of observing result as extreme as our sample is almost 0. Therefore, we can conclude that there exists a statistically significant difference between two groups and reject the null hypothesis. Aligning with our alternative hypothesis, we conclude that the percentage of STEM major students who get a job that requires a college degree is significantly higher than that of non-STEM major students.

In addition, we could obtain a Cohen's d of 1.01 which can be considered relatively large.

## Discussion