STEM: Gender & Background

Alan Yue

Oct 3, 2022

Goal

It has long been the case that STEM (Science, Technology, Engineering, Mathematics) careers tend to outearn their non-STEM counterparts [1]. Thus, determining how information such as gender and background interrelate with feelings toward STEM is important because of the ramifications on income inequality. Many similar studies have focused on college-age students who've already chosen a major or looked at the gender differences in test scores in certain subjects [2]. The goal of our survey is to try and determine some of the root interactions making up the eventual pursuance of a STEM career by asking those in high school about their STEM preference and perceptions. This will allow us to gain a greater grasp over how the preference for STEM is moulded and possibly determine where to focus our efforts so as to positively shift any unjust imbalances in STEM over time.

Procedure

We are interested in determining the factors that drive those in high school to pursue STEM later on. However due to time, cost, and organizational constraints, it would be quite unwieldy to carry out our survey in multiple countries or even provinces. Thus, we pick the population of all high school students in Ontario as the target population.

To carry out the survey, we would obtain a list of all high schools in Ontario to use as clusters and take a simple random sample without replacement of some number of the schools that it would be feasible to contact (5 in our proposed design). Then we would contact each selected school and ask to obtain a list of all students (possibly in the form of student IDs so as to anonymize the list). We would take a simple random sample of some number of the students from each list (20 in our proposed design) and each chosen student would then be emailed a link to the survey. All students who respond become part of the overall sample.

Our first-stage sampling frame is the list of Ontario high schools and the second-stage sampling frames are the lists of students obtained for the chosen high schools. The frame population is all students who attend a high school in Ontario and the sampled population is all these students who would complete the survey if selected and whose school would be willing to provide some list of their students.

We believe that by using our outlined procedure, we are able to make appropriate cluster-based inferences on our population of interest, while also limiting the practical constraints of contacting too many schools to ask for student information. However, we would still be presented with an obstacle if a chosen school were to refuse to provide any list of students. Hopefully, the clusters would be homogeneous enough to each other and the general target population that this does not significantly bias our results after selecting a replacement school. Another source of bias comes from nonresponse if selected students choose not to answer the survey would have answered in some way that differed from the general target population.

Link to Survey

Select Questions

What is your household income in $CAD?

This question's purpose is to record income information and hopefully shed light on whether it's related to other survey questions such as interest in and pressure to pursue STEM. The respondent must enter a number between 0 and 1,000,000. This forces household incomes over a million to be lowered, but we believe this is uncommon enough to justify the cap to prevent erroneously large entries. Some respondents may not know their household income exactly and may have to estimate. Some respondents may feel uncomfortable sharing their household income if it is too low or too high, so we should approach these answers with caution.

Which do you enjoy more: STEM subjects or non-STEM subjects?

This question was included to simply record whether the respondents enjoyed STEM more than other subjects. They are given a binary choice to pick either STEM subjects or non-STEM subjects, which makes it easy for the respondent to pick one, but may not be nuanced enough of a question. Perhaps a respondent is passionate about the life sciences, but hates mathematics. However, we are primarily focusing on STEM as a whole and do not break down by further subjects in our analysis, so we keep it this way.

Rate your perceived academic ability in STEM subjects.

This prompt asks the respondent to rate their perceived academic ability in STEM on a scale of whole numbers from 1 to 5, with 1 being "Quite lacking" and 5 being "Quite proficient". Five numbers is somewhat arbitrary, but it represents a good range of options to choose from. We included this prompt as ability to excel in a field is a good reason to either choose to pursue or not pursue that field and we also wanted it to be the respondent's own personal perception of their ability, as this can presumably be a greater factor in one's personal choice than an objective measure of skill. However, this prompt suffers from the same drawback as the last question in that STEM may be too broad and a respondent may feel their ability ranges widely from subject to subject.

Data

In lieu of actually conducting the proposed procedure, we have simulated answers to the survey questions so that we can act as if 100 high school students in Ontario took the survey and show what subsequent analysis would look like. First, we randomly selected with equal probability a number of male students between 40 and 60 because it's safe to assume the genders are roughly equally split. This number is then subtracted from 100 to get the number of female students in our sample. We decided not to simulate students who identify as nonbinary because there is very sparse existing data on this in the context of what we're studying and we don't think we can make any guesses as to what survey answers would be like for such students. If the survey were to be actually carried out, one should prepare to handle results from nonbinary respondents.

We don't expect household income and gender of students to be correlated so we draw from a normal distribution with a mean of 70,000 and standard deviation of 20,000 to act as household income figures for both males and females. Income is likely not normally distributed, but we use the normal distribution and parameters we find to be reasonable for ease of simulation.

Whether a student intends to pursue STEM in the future is a binary response in our survey, so we simulate a Bernoulli trial for each student, with the probability of 'yes' being 0.1834 for males and 0.1287 for females. We use these probabilities to be more accurate to the real-world data analyzed by Chan, Handler, and Frenette [2], who found these to be the proportions for undergraduate students enrolled in a STEM program. Similarly, a student's preference for STEM is binary, so we simulate Bernoulli trials, with the probability of the preference for STEM being positive equal to 0.72 for males and 0.49 for females. These probabilities were taken from a study performed by Kans and Claesson [3, sec. 3], where Swedish secondary students of both genders rated their interest in STEM on a scale from 1 to 4. The proportions used correspond to the respondents who answered with a 3 or 4.

To simulate ratings for perceived academic ability in STEM, we sampled a whole number from 1 to 5 for each student, again using the Kans-Claesson study as an aid. In the study, students were asked to rate whether they were good at STEM from 1 to 4 [3, sec. 3]. We use the proportions of responses for each number as sampling weights for 1,2,4,5 and we assign 3 a weight of 0.2 to get weights of 0.07, 0.19, 0.2, 0.49, 0.25 for males for 1-5 respectively, and 0.16, 0.37, 0.2, 0.34, 0.12 for females. We simply picked random whole numbers from 1 to 5 for the pressure felt to excel in STEM rating because we couldn't find a real-world analogue for this data and we aren't using this variable in our analysis.

Because the data were entirely simulated, we didn't have to do any data cleaning, except ensuring that all income figures picked from the normal distribution were between 0 and 1,000,000 to match the imposed limitation on the income question responses.

We will be taking a closer look at the following variables:

Gender: gender of respondent

Household income ($CAD): continuous between 0 and 1,000,000

Preference for STEM: 1 if the respondent enjoys STEM more than non-STEM subjects, 0 otherwise

Intent to pursue STEM: 1 if the respondent intends to pursue STEM in the future, 0 otherwise

Table 1: Summary Statistics

Gender	female	male
Count	60	40
Mean household income (CAD)	71611.18	69351.56
Standard deviation household income (CAD)	19290.77	18279.28
Proportion preferring STEM	0.5166667	0.6000000
Standard deviation STEM preference	0.5039393	0.4961389
Proportion intending to pursue STEM	0.3166667	0.6000000
Standard deviation intending to pursue STEM	0.4691018	0.4961389

We can see from Table 1 that we have more females than males in our sample and there doesn't seem to be a difference in household income across the genders. It appears that males tend to prefer STEM more than females and are also more likely to want to pursue STEM in the future. We will try to explore some of the reasons for this in the next section.

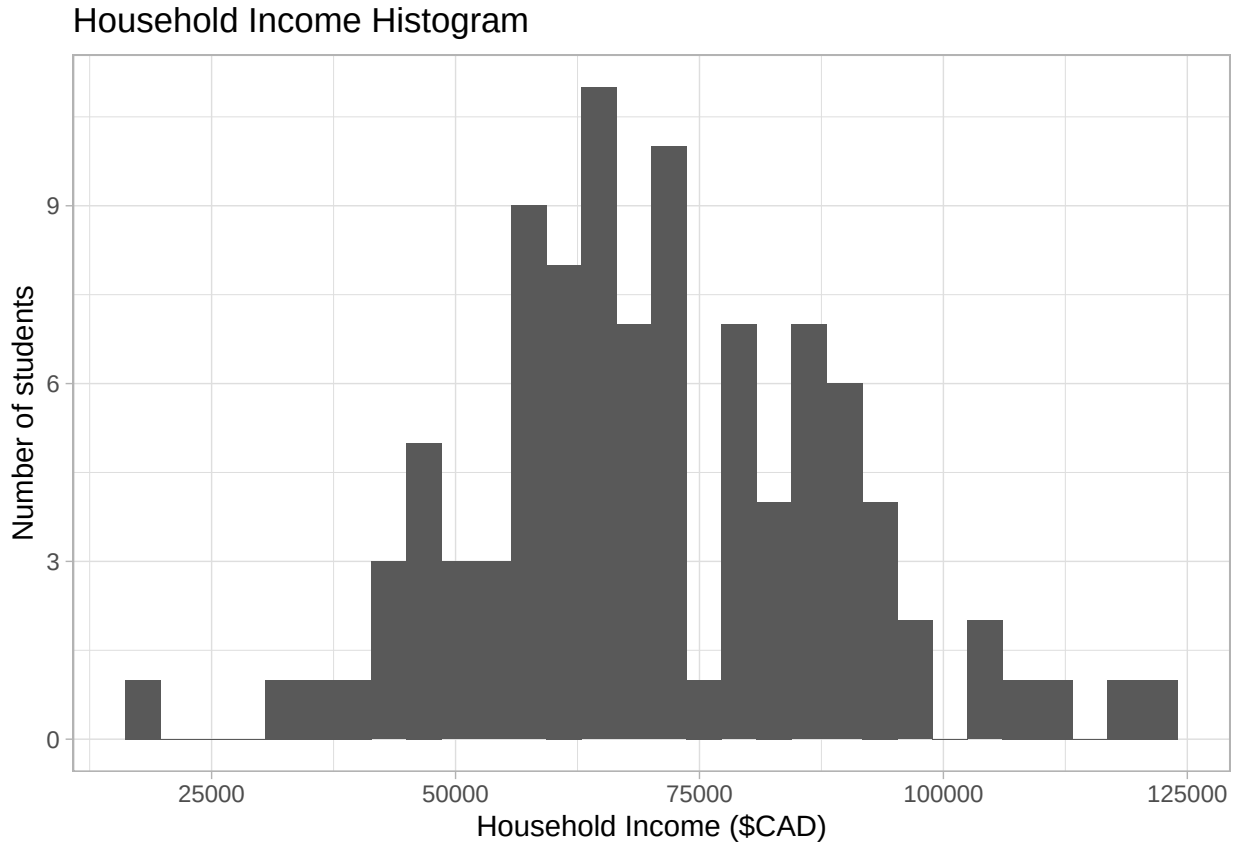

Figure 1: Household Income Distribution

We see from Figure 1 that household income appears to be roughly normally distributed with a mean of 70,000 and standard deviation of 20,000, which aligns with our simulation parameters. We suspect that this data would have more of a right skew had it been collected for real, as there could be some households who make far more than the mean or median. We will use this distribution in our analysis of income and intent to pursue STEM in the next section.

Methods

Confidence Interval

First, we will construct a confidence interval to get a plausible range of values for the difference in mean household income for those who intend to pursue STEM and those who don't. Our proposed procedure suggested using cluster sampling, but we will treat the students who intend to pursue STEM and students who don't as two separate and independently conducted simple random samples for simplicity of analysis. We assume that the frame population is large enough so that we can assume relative independence between students when sampling without replacement, letting us ignore the finite population correction that we would have had to use otherwise. Let's assume that household income is normally distributed for both groups of students with known standard deviation of 20,000, so that the difference of household incomes is also normally distributed.

We have considered all assumptions to construct the following confidence interval at a 95% confidence level [4, ch. 10]:

$$\bar{x} - \bar{y} \pm z_{0.025} * \sqrt{\frac{\sigma_x^2}{n_x} + \frac{\sigma_y^2}{n_y}} = 67581.94 - 73065.09 \pm 1.96 * \sqrt{\frac{20000^2}{43} + \frac{20000^2}{57}}$$

where $\bar{x}$ is the sample mean household income for students intending to pursue STEM,

$\bar{y}$ is the sample mean household income for students not intending to pursue STEM,

σ_x^2 is the population standard deviation of mean household income for students intending to pursue STEM,

σ_y^2 is the population standard deviation of mean household income for students not intending to pursue STEM,

n_x is the sample size of students intending to pursue STEM,

n_y is the sample size of students not intending to pursue STEM,

and $z_{0.025}$ is the 0.975th quantile of the normal distribution. We use this value because 95% of the distribution will be within the 0.025th and 0.975th quantiles.

Significance Test

We will now conduct a significance test on the difference in proportions of males who prefer STEM subjects and females who prefer STEM subjects. We will treat the male and female students as two separate and independently conducted simple random samples for simplicity of analysis. We again assume that the frame population is large enough so that we can assume relative independence between students when sampling without replacement, letting us ignore the finite population correction that we would have had to use otherwise. We check that we have at least 10 observations preferring STEM and at least 10 who don't in both samples so that the distribution of the difference in proportions is approximately normally distributed [4, ch. 10].

We will choose a standard significance level of $\alpha = 0.05$. Our hypotheses are:

$$H_0 : p_m = p_f$$
$$H_a : p_m > p_f$$

where p_m is the proportion of male Ontario high school students who prefer STEM and p_f is the proportion of female Ontario high school students who prefer STEM. We use H_a as our alternative hypothesis because we want to test the idea that the greater proportion of males in STEM careers [1] indicates male students

tend to enjoy STEM subjects more. Our null hypothesis or a priori assumption H_0 is that there is no difference in proportions.

Our test statistic is [4, ch. 10]:

$$z = \frac{\hat{p}_m - \hat{p}_f}{\sqrt{\hat{p}(1-\hat{p})(\frac{1}{n_m} + \frac{1}{n_f})}} = \frac{0.6 - 0.5166667}{\sqrt{0.55(0.45)(\frac{1}{40} + \frac{1}{60})}} = 0.8206099$$

where $\hat{p}_m$ is the sample proportion of male students preferring STEM,

$\hat{p}_f$ is the sample proportion of female students preferring STEM,

n_m is the sample size of male students,

n_f is the sample size of female students,

and the pooled estimate of the common proportion is $\hat{p} = \frac{n_m}{n_m+n_f}\hat{p}_m + \frac{n_f}{n_m+n_f}\hat{p}_f$. We use this because under H_0, we would have a common proportion between males and females and we want to see how unusual our test statistic is if this were true. If our test statistic is extreme enough to fall in the upper tail containing 5% of the standard normal distribution, we will choose to reject our null hypothesis in favor of the alternative.

Results

After computing the confidence interval in the previous section, we are 95% confident that the true difference in mean household incomes between Ontario high school students who intend to pursue STEM in the future and those who don't is in the interval (-13401, 2435). This means if we were to take repeated samples of students from the populations, 95% of the confidence intervals constructed in this manner should contain the true difference of means. Because 0 lies in the interval, we cannot conclude that household income is different for the two groups. This makes sense based on how we simulated the data because we drew income figures from the same distribution for both groups, but in real life it's possible that parents who are high earners either birth children who are predisposed to want to pursue STEM or foster an environment that encourages it. Previous studies have found a link between academic achievement in STEM and socioeconomic status [5]. Further exploring the link between household income and intent to pursue STEM gives us a better understanding of what influences the drive to pursue STEM and may have implications on which groups may be disadvantaged and should be the target of more STEM outreach.

The test statistic in the previous section gives us a p-value of 0.082308. This is greater than our significance level of $\alpha = 0.05$ so we don't have sufficient evidence to reject H_0, as our test statistic wasn't extreme enough to be considered unusual if H_0 was true. Thus, we can't make the claim that there exists a difference between the genders in the proportion of Ontario high school students who prefer STEM subjects. In our simulation and the study we based parameters off of, there was a difference, so this is somewhat unexpected. The randomness of simulation resulted in our sample proportions being closer to each other than the parameters we used to simulate. However, failing to reject the null hypothesis is not quite the same thing as accepting it as true. We would need future studies to explore the same hypotheses, as this could give an indication of how much of the gender disparity in STEM fields is due to personal preference and is less important to address. It should be noted however, that it's extremely difficult to disentangle causes in observational studies and reasons for this preference could be the result of implicit social pressures that should be addressed.

Bibliography

1. Richard Fry, Brian Kennedy, and Cary Funk. "STEM Jobs See Uneven Progress in Increasing Gender, Racial and Ethnic Diversity." (Apr. 1 2021). *Pew Research Center.* Retrieved Sep 28, 2022 from https://www.pewresearch.org/science/2021/04/01/stem-jobs-see-uneven-progress-in-increasing-gender-racial-and-ethnic-diversity/

2. Ping Ching Winnie Chan, Tomasz Handler, and Marc Frenette. "Gender differences in STEM enrolment and graduation: What are the roles of academic performance and preparation?" (Nov. 24, 2021). *Statistics Canada Economic and Social Reports.*

3. Mirka Kans and Lena Claesson. "Gender-Related Differences for Subject Interest and Academic Emotions for STEM Subjects among Swedish Upper Secondary School Students." *Education Sciences,* vol. 12, no. 8, 2022.

4. Jay L. Devore, Kenneth N. Berk, and Matthew A. Carlton. *Modern Mathematical Statistics with Applications.* 3rd ed. Springer, 2022.

5. Laura Betancur, Elizabeth Votruba-Drzal, and Christian Schunn. "Socioeconomic gaps in science achievement." *International Journal of STEM Education,* vol. 5, no. 38, 2018.