Daphne Shaw, Nadia Jimenez, Brendon Ruiz

Stats 7 Lecture C

Prof. Lee Kucera

29 May 2022

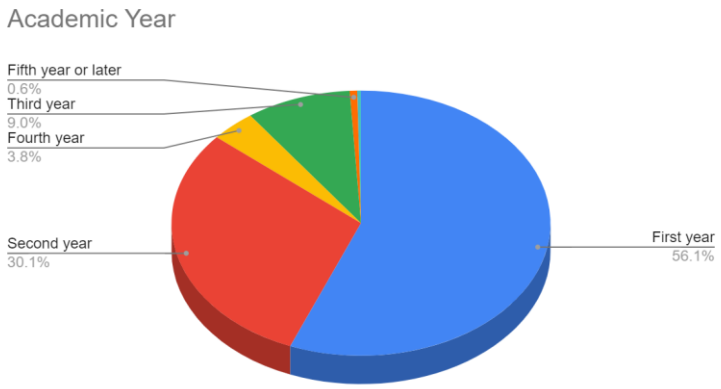
STEM and Social Sciences Major Distribution in Stats 7

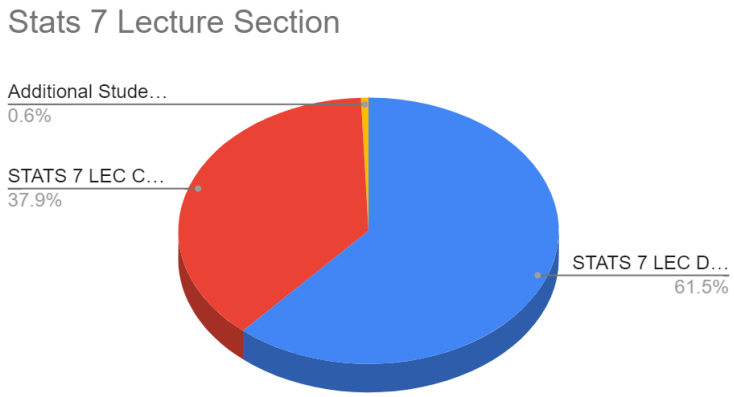
Our research question is, "Are there more STEM majors in a Stats 7 class than there are social science majors?" STEM stands for Science, Technology, Engineering, and Mathematics. In this study, STEM encompasses the schools of nursing, biological sciences, computer science, pharmacy, engineering, and physical sciences. The social sciences investigate human behaviors and interactions, both with other people and with their physical surroundings. Psychology, political science, and economics are examples of social sciences. We chose this topic since an individual's surroundings and the different circumstances to which they are exposed on a daily basis are constantly responsible for developing that individual. It's probable that our peers will or have experienced the same thing. When considering whether or not to devote so much time to a subject, it is useful to have a sense of what to expect by analyzing current demographic figures in the class. There has been previous data collected on major enrollment in UCI as a whole (https://datahub.oapir.uci.edu/Enrollment-Dashboard.php), so our goal for this study is to look at the major distribution between STEM and social sciences in a Stats 7 lecture in particular in order to reason what types of students are in/would select certain types of courses.

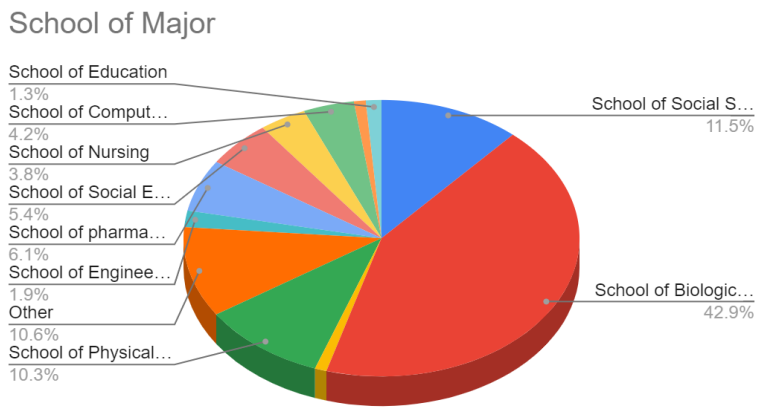
Since we wanted to explore whether the amount of majors were greater than social science majors, we formulated our null hypothesis as the amount of STEM majors in a Stats 7 lecture being not greater than the amount of social science majors. Our alternative hypothesis was that the amount of STEM majors in a Stats 7 lecture is greater than the amount of social

sciences majors. This allows us to explore the major demographic distribution in a Stats 7 lecture, and helps us to better understand what types of students take the class. To gather data, we initially used a Google Forms survey. However, we ended up using the Canvas class survey the TAs sent out, since we did not want overlap between our sample populations, and the Canvas survey included our questions. Afterwards, we organized and cleaned the data into a Google Sheets document, which can be found here. We used JMP to collect the data from the Canvas survey, Google Sheets to analyze our data since it allows us to create graphs easily, a TI-83 Graphing Calculator for statistical analysis, and online Confidence Interval calculators for data analysis. We ran into a few challenges, which included getting people to answer the survey initially with our Google Forms survey. However, we overcame this with the class survey. Another challenge was organizing data between academic years since the data from the JMP file showed both the names and the actual numbered academic year (freshman vs first year), which made Google Sheets categorize it incorrectly. We fixed this by manually combining the name of year and numbered academic year in our data so that it shows the correct values for students in each academic year (e.g. combining "freshman" and first year"). We also weren't sure which majors to include and exclude in the survey when formulating our research question, but ended up collecting all academic schools as part of our data, and defining STEM and social sciences clearly from those schools. There were a few outliers in our data, as there were majors that couldn't be classified as either STEM or social sciences (e.g. social ecology), as well as "additional students" and graduate students included in the sample. However, since these took up an extremely small portion of the data, they did not affect the overall results significantly. We decided to use a one proportion z test, for which the conditions we have checked: we did random sampling with the class survey, and both np_0 and $n(1-p_0)$ are at least 10.

Attached below are pie charts and bar graphs we created using Google Sheets to display results of the data we collected from our survey, which included three questions: lecture section, academic year, and school of major. The Google Sheet with the data is linked to the images.

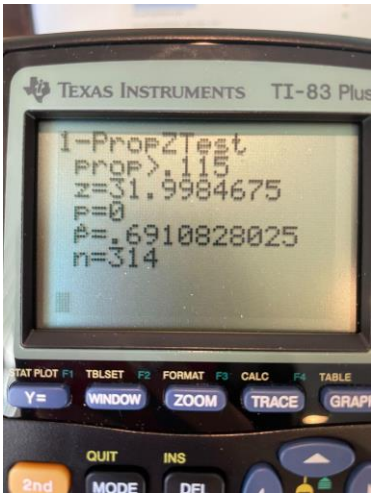




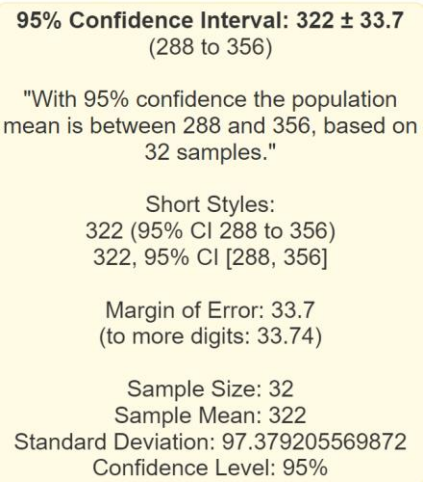


We used a **one proportion z test** to analyze the statistical significance of our data, since we only have one sample and we are trying to compare two proportions. In the below calculations, we set p_0 (hypothesized value) as 0.115, since that is the proportion of social science majors we calculated from our data according to the above school of major chart we created. We then set population proportion $p = 217$, since that is the amount of STEM majors (definition in introduction) we calculated from our data: 0.692 (proportion of STEM majors; $4.2\% + 3.8\% + 6.1\% + 1.9\% + 10.3\% + 42.9\% = 69.2\% = 0.692$ $0.692 * 314$ (total sample size)=217.288 STEM majors. However, since people cannot be decimals, we rounded down to 217. Below is our one proportion z test:

$H_0 : p = p_0$, H_a (alternative hypothesis, right tailed) : $p > p_0$. Putting the above values into a TI-83 and performing a 1 proportion z test, we obtained the following results:



Using a 0.05 level of significance, we can conclude from our above results that our data is statistically significant since our p value is 0. Since our P value is 0, this means we are able to reject the null hypothesis, and that our test is statistically significant. To determine if our data is practically significant, we will use confidence intervals. Since our data is qualitative, we will instead use enrollment data for Stats 7 lectures (from the UCI official Schedule of Classes page) from the past five years (2017 to 2022 academic school years) in comparison to our sample size of 314 to determine a 95% confidence interval (dataset can be found in Google Spreadsheet linked to graphs above, in F column), which means: with 95% confidence we can say that with 95% confidence the population mean is between 288 and 356, based on 32 samples (Figure 1 below). Since our sample size in our collected data falls within this interval, we can justify that our data has practical significance and can model average Stats 7 lectures at UCI (full confidence interval calculations below).



One potential limitation we failed to consider when sampling our data was that biological science majors are required to take a Stats course, so this could inherently skew our results since Bio majors would take up a large portion of STEM majors, causing inflation in the STEM majors data. There are also graduate students and additional students in the class, so our data does have outliers (grad students and additional students) that are not fully representative of the population of interest (undergrad students taking the Stats 7 class).

Overall, there was a good mixture from all the different schools that take a Stats course but the specific answer to our question was that there are a lot more STEM majors that take a Stats course compared to Social Science majors. Specific to our research question we were able to learn that Biological Science majors have a significantly higher number of students that take the Stats course due to major requirements. The overall research process had hoops we had to get through but not impossible to do. When collecting data we hadn't realized the lack of responses we would be receiving from our personal Google Forms. Having the survey get sent out through canvas as the class was helpful when collecting data although something that could've been done differently would've been simplifying the question to majors that only belong to STEM or Social Sciences that way we can get rid of unnecessary data and minimize our chances of any outliers.

We had the right idea of questions going out but we did need to be more specific with the majors and limit the class year options in order to get a better representation of our population. For anyone who would like to do their own research I would advise them to thoroughly consider the questions you need answered for your research. You don't want to be too specific to the point where you're missing information but you also don't want to be too broad where you have unnecessary information. Also, make sure you have an idea of the kind of testing you would like to do on your research so you know your populations and how to ask the questions. Further follow up questions for our research could be expanding gathering data for all the schools instead of just STEM and Social Science.

Works Cited

Confidence interval calculator. Math is Fun Advanced. (n.d.). Retrieved May 29, 2022, from https://www.mathsisfun.com/data/confidence-interval-calculator.html

Course Enrollment History/Statistics. Course enrollment history/statistics. (n.d.). Retrieved May 29, 2022, from https://www.reg.uci.edu/perl/EnrollHist.pl?dept_name=STATS&course_no=7&class_type=LEC&action=Submit

Enrollment. Office of Institutional Research. (n.d.). Retrieved May 29, 2022, from https://datahub.oapir.uci.edu/Enrollment-Dashboard.php

Jimenez, N., Shaw, D., & Ruiz, B. (2022, May 15). *Stats 7 Group Project Data*. Google Sheets: Sign-in. Retrieved May 29, 2022, from

https://docs.google.com/spreadsheets/d/1vPoOmY_UR0E68eDb3F9rsjRtUuo5xEgpPbzJR1QhjbQ/edit#gid=0