Reflections

Statistics for Applied Science

Chris Shepard

December 4, 2018

For our final project, the process from early beginnings to ultimate result was multi-layered, and sometimes complex.

To begin, each individual was to decide upon a research question that captured their interest, while still forming a sufficient foundation for a decent investigation. My own initial attempt at a quesiton involved potentially building a small piece of software to scrape relevant information from a website, in order to attempt to do an analysis.

This was ultimately insufficient for several reasons. Primarily, this was to be a group project, and there was no amount of buy-in for a project relying on something like like a custom web scraper to gather the necessary data. The two primary groups of thought for this project appeared to be "let's do a small survey of our fellow students", and "let's gather information from existing collections of data". Neither of these left room for the sort of data collection I had in mind.

Beyond this lack of buy-in, the focus of my attention was on a small website, with a niche audience. This meant that it would have been difficult to generalize to a larger population, which meant that any such investigation would be limited in its possible applicability.

The group that I became a part of came together primarily due to a common lack of commonality. None of us had research questions in mind that meshed with any other group, and so we ended up grouped together by default. After several days of discussion, we settled on an approximation of our question, hoping to compare the chosen majors of students against their resulting career.

In order to firm the question up, I looked into possible sources of information. Ultimately, I settled

upon the National Center for Education, which has a substantial amount of data collected from large surveys available through an easy-to-use web interface.

This choice led into a new set of problems. The chosen pieces of data consisted of relative percentages, with no clear relationship between the rows. My first attempt to rectify this was wrong-headed, in that I failed to notice this lack of a relationship and I simply treated percentages in the same column as summing to 100%. In so doing, I ended up with a set of numbers that added up to our expected sample size, but that were effectively arbitrary. We did not, at the time, have information about the relative proportion of respondants for one major versus another, and my attempt at turning the percentages into hard numbers assumed that the proportions were identical.

I overcame this issue by digging into the code books that came with the rest of the survey data. These spanned several hundred pages, and collected the details about each variable used in the larger survey. This happened to include the relative proportions of the majors, the key piece of information that we lacked up to this point. I proceeded to use this table of relative percentages, our original table of overall percentages, and the known estimated sample size to produce a table of hard values that preserved the sizes of the majors, relative to each other.

With our data pinned down, I was able to perform the analysis, and work out a conclusion for our question. At this point, we needed to simply assemble our poster and perform our presentation. One of my fellow group members and I worked to start the poster, stubbing out the framework for the material, and adding in the data table. Later on, I finished filling the poster out, which involved a few iterations of reformatting and resizing the information to fit the necessary pieces.

This was the bulk of the work accomplished. After multiple rounds of back-tracking and diving deep into the huge amount of information that was available to us, we managed to work through the process, beginning to end, and come to what I feel is a reasonable conclusion.