Paper 4683-2020

Solving Health Problems on the Academic Frontier: Training Students to Use SAS® Through Research

Charlotte Baker, Virginia Tech

ABSTRACT

SAS® is a valuable tool for data management and analysis in a variety of settings, and academia is no exception. As an educator of future public health professionals and a researcher with a data driven epidemiology laboratory, integrating various analytic tools into my courses and my lab is necessary. Students in my research laboratory complete data analysis to help solve large problems related to human and animal health issues, but many do not join the group with backgrounds in data analytics or statistics. It is absolutely imperative that these students get a well-versed education, and quickly, through-real world opportunities. This presentation discusses our methodology for training students, some of the benefits and struggles of training students in this manner, and demonstrate some of the overall competencies our students acquire through this experience.

INTRODUCTION

Outside of data nerds, why do people care about the nuances of using data to answer questions? Answers needed for many questions related to health, education, and other fields can be acquired from data analytics. We need to know why supply chains work, whether we should be concerned about disease, and whether performance in physical education affects classroom success for children. But who does analytics? We all analyze some kind of data in our day to day lives whether it is adding up the amount our groceries cost to calculating the time it takes to get ready in the morning once the initial alarm goes off. People may say "I'll never use Algebra again!" but in fact they do on a daily basis. Doing some type of data analysis for a job or career is a bit different. It takes training and practice to be a good data analyst and programmer no matter the field. With the rate of people retiring in fields like public health comes a great need to replace them without compromising quality and expertise. All workers essentially need to replace themselves and some of us are tasked with giving these replacement workers the foundation needed to assume their new roles. This particular paper will discuss how we go about designing the training for students and what thought goes into doing this; the benefits and struggles of training students through real life work; and how this particular method provides certain competencies to the future workforce.

THE IMPORTANCE OF TEACHING

Being a teacher of statistical programming and data analysis is a fascinating position. It requires staying abreast of current issues, software trainings, what software is best for a particular analysis, what type of education students come to you with, and needs of the field for where the students need to be. We cannot escape the conversation of whether we are all going to be replaced by machines and whether all of this information is completely useless to learn because the positions will be obsolete. In several fields, this threat is not as ominous as it seems, especially as methods have not yet been established and perfected to replace empathy and nuanced understanding of things human and how to interpret or act on gray areas. There is not an algorithm that replaces these things. Additionally, the broad idea of "data science" does not replace people that know the topics at hand and how to apply knowledge of a field to the answers gathered through data analysis. So, we must

keep doing as we have done – we must keep designing training opportunities that meet everyone's needs.

Our training related to statistical programming and data analytics starts with what we think people should learn. In our case, it is teaching students how to use a combination of common sense and technical skill to report to a larger audience about science and about injury and disease risk. They need to understand how to build and manipulate data sets; how to get descriptive statistics; how to look at relationships between variables; and how to spit that out in a way for all to understand and act upon.

WHAT PROGRAM(S) SHOULD BE TAUGHT?

After we determine what it is students need to know, we have to focus on the program(s) that will be used to teach them what they need to know. In terms of the statistical and visualization software tools, there is an extensive list to choose from – SAS, SPSS, Stata, R, Python, Epi Info, JMP, Tableau, Microsoft Excel, Apache Spark, MATLAB, Minitab, Hadoop, and more. How does one decide what to use? For one, it depends on the goals of the work to be accomplished; the tools in use where your students will go to work; and what access you have to software. If they will not be doing large data processing for an insurance company, is it useful to teach Hadoop as a first or second option? If they will be doing relational database work for a health agency, perhaps SQL is a good in between language. If students will be doing very little programming and analysis later and not anything beyond the most simplistic now, what is best to keep their interest and lower the level of everyone's frustration? Perhaps Epi Info or SPSS. Will you be doing machine learning or working with engineers? Perhaps Python needs to be in the toolbox.

Bottom line, you need to know –

- Who are you teaching?
- How in depth do you need to go?
- What work do you need done?
- Are there any curricular requirements you need to adhere to?
- What do you know how to use?
- What are you willing to learn?
- Who will be hiring these students?
- What type of software access do you have?

OUR LAB

Our epidemiology research lab consists of undergraduate**,** graduate, and professional students across multiple fields. Some have no statistical training, others have one to two courses, and some also have research methods training. About three quarters of students in the research lab are not interested in being data analysts, epidemiologists, or statisticians when they leave school and about half of the students come in with a basic understanding of R because of a curriculum requirement (which is taught by me, the leader of the lab, or elsewhere on campus). The students in the lab seek jobs in a number of public health related fields including epidemiology, medicine, dentistry, and veterinary medicine.

What Kinds of Needs Do We Have?

We use multiple data types and sources to achieve our goals including US Census data, data with complex sampling structures, clinical data, and survey data. Many of the nationally available public health data sets we use have pre-written SAS syntax at least for reading in the data, and others are available as SAS format data sets. For example, the National

Center for Health Statistics released SAS Input Statements for the ICD-9-CM Barell Matrix (Injury Diagnosis Matrix) but did not do the same for other languages. The syntax is approximately 230 lines long. Comparatively, the Centers for Disease Control and Prevention released multiple syntax and library files for the Behavioral Risk Factor Surveillance System (BRFSS) in SAS format but not for other languages. This does not mean that it is a requirement to analyze this information using SAS or that the syntax cannot be converted to another format such as R or SPSS. But given the complexity of the syntax, it can be simpler to leave as is especially for novice users. Other national data sources, such as the AHRQ's Health Care Utilization Project, release syntax in multiple formats – SAS, SPSS, and Stata.

As the lab director, I know multiple statistical languages. Our primary lab epidemiology and biostatistical collaborators know multiple languages. Rather than switch back and forth between multiple programs for the larger projects in the lab, it can be easier on my part to use one primary language and only use other languages for one-off projects or smaller projects that need to be accomplished very quickly. A robust software package understood by many people across fields is useful whether they are calculating simple incidences or predictive models for disease outcomes. We require analyses on most projects that cover descriptive statistics, basic comparisons, and adjusted comparisons using various regression methods. Given all of these reasons, we found SAS to be the top choice for teaching our research lab.

WE CHOSE SAS. NOW WHAT?

The plan matters for training people to use SAS. When teaching in a classroom setting, there can be specific guidelines and guiderails and you can use a specific training scenario that has a proven outcome. In our situation, however, we are training these students through real life experience. This means we start with the study design and create analyses around the study needs. It is expected to be dirty and messy but the expectations matter for lab productivity such as manuscripts and conference presentations. In public health, it also matters for improving the science. We have to build in time for the learning process but slowing the clock down too much decreases productivity. We have targets to hit and students need to learn to keep the pace.

THE OPTIONS ARE SO EXPANSIVE!

Now that we know the time constraints, we still have to select the SAS tools to use for the training and actual work. There are so many options – students have laptops, there are desktops in the lab. Some people have access to Apple computers and others have access to Windows computers. More importantly, many people are beginning to rely on tablets more than ever as their main computer or as their everyday carry. SAS 9.4 works natively on Windows. Most of the work our lab does has to happen in the laboratory due to data security so that is great. But what if students are at home doing practice exercises or writing syntax to test later in the lab? What about analyses using publicly available data like NHANES or BRFSS? SAS University Edition works on computers with MacOS as well as Windows and prevents students from needing to purchase their own license. At our institution students have a charge under $100 to purchase SAS 9.4 for a year, but that's a personal cost not a lab cost. The syntax is a little different in SAS University Edition than SAS 9.4 (because it uses a virtual machine) which also requires these novice users to keep up with those differences as they use both programs in parallel. This could be considered a fairly minor inconvenience to some.

In previous years, our lab used SAS Studio OnDemand for Academics, a web-based alternative for teaching purposes but it has too many limitations for our current situation. Whereas SAS University Edition has a workaround to use data sets larger than 10MB, using a data set such as NHANES in SAS Studio OnDemand for Academics requires breaking data

apart to be less than 10MB then combining it again once uploaded or using a very small data set which is not advantageous if trying to show students how to work with large data sets. The biggest benefit of this tool is to practice, not for full analyses. SAS University Edition is fully capable of completing many of the analyses in question, but is still different in layout and some syntax than SAS 9.4. What if students are using an Android or iOS tablet? SAS Studio OnDemand for Academics works but we've found that a server-based SAS 9.4 installation is better. Both require internet access so is not an option if internet is not available. If one just wants to type syntax for running later, the non-Wi-Fi options expand. Applications such as Kodex, CodeHub, or generic notes applications can all be used to type and save syntax.

WHAT DO WE USE TO TEACH SYNTAX?

SAS Institute, Inc, provides multiple training resources including webinars for learning SAS syntax. UCLA's Institute for Digital Research and Education has an expansive set of training explanations and result keys that walk even advanced users through how to use the software. For our specific research examples, we developed our own modules that we provide in person and via recordings made with a screen capture application and YouTube. All of our videos are hidden from general YouTube users and only are available to students and affiliated faculty with the research lab. These videos have needed to be re-recorded every 2-3 years as new examples are identified and student needs change but the accessibility means that they are available even when face to face lessons cannot be arranged.

THE PROJECT

The project being used this academic year in our laboratory to teach students SAS is one that involves the National Inpatient Sample, National Emergency Department Sample, and National Ambulatory Surgery Sample data from the Health Care Utilization Project from Agency for Healthcare Research and Quality. We need to identify how many people have diagnoses of herniated discs and sports or recreation related injuries in the same visit in the US. The students worked first to define the inclusion criteria and exclusion criteria, then learned how to clean the data sets for these complex survey samples to include just the relevant variables, and are in the midst of determining answers by data set (inpatient, emergency department, ambulatory surgery).

PROJECT GROUPS

Given the number of students in the lab working on the project, 12 students, they were assigned to small groups of 4 to analyze each set without replication. These groups allow the students to learn from each other without the overwhelming feeling of being lost in a larger group where 2-3 standout students would potentially be doing all of the work. Students were assigned to groups in a somewhat random fashion but the undergraduate students were intentionally distributed across the groups to also teach skills of working with the more academically advanced students. In addition to needing to know SAS skills, students needed to learn how to operate in small collaborative groups, how to understand and use ICD-10-CM codes, how to design studies, and how to critique the methods of others. Students each presented aspects of their analyses to the larger group and everyone was encouraged to provide feedback and provide suggestions for alternatives to gather answers. These are the skills mentioned earlier that are essential to gathering answers in public health.

The Struggles and Successes

The novice users encountered difficulties that we previously identified in past years. Forgetting the semicolon, overwriting data sets, and forgetting what part of the syntax did what are common mistakes for beginners. Given that students were working in groups,

these mistakes did not continue for long. Students became adept at identifying issues in the code of their teammates and then their own. Students regularly began commenting the syntax to remind their teammates to do so. The conversation increased about how things were happening in the project and what needed to happen next. We did not need to completely step out of the conversation and not critique the students but as time went on, it became easier to wait on students to critique each other and support each other without my input. Students that had learned other languages before were surprisingly not faster to pick up SAS than students that had never learned a statistical programming language. This could be because of the reality of teaching through a real project and the pressure that students felt to do well. Students took longer to admit confidence than we took to believe in their abilities and this is not uncommon to past experience.

CONCLUSION

We found that this particular real life scenario was advantageous to teaching students how to use SAS software. We found that with enough up front planning, it is easy to try and identify what roadblocks may be in the way but that answers to those roadblocks are not always obvious. Having a good group of students who are invested in learning and accomplishing something they can explain to others makes a huge difference in outcomes compared to teaching students who are simply learning for curricular reasons. We suggest that others who have the opportunity use more hands on, real consequence situations to teach students how to do statistical software programming.

ACKNOWLEDGMENTS

I would like to acknowledge the students of the I-SPY DATA Lab at Virginia Tech.

CONTACT INFORMATION

Your comments and questions are valued and encouraged. Contact the author at:

Charlotte Baker
Virginia Tech
sesug.ops.2018@gmail.com