Data Assignments in Substantive Courses: Getting Undergraduates Excited and Interested in Data Science

Lonna Rae Atkeson, *Florida State University, USA*

Teaching undergraduate methods is a difficult task in a major in which there are many math- and science-phobic students. Nevertheless, a foundation in scientific thinking, an understanding of basic statistics, and an introduction to data and principles of data science and its utility should be required curriculum and competency for a bachelor's degree in political science. To achieve this, I transformed my undergraduate public opinion and electoral behavior course into one that emphasizes and teaches students the power of data in understanding the political and social world.

It is important that social science methods training be integrated into substantive courses. Working with data is different from passively observing data presented during a lecture or in a reading, and data are more interesting when they are connected to substantive questions. Therefore, the course provides a social science methodological foundation in the principles of science (i.e., observation and comparison) in a concrete way connected to an interesting subject. Students complete the course not only with more knowledge of the subject than in traditional formats but also with self-confidence and competence in manipulating data in basic ways and asking and answering questions with simple statistics.

We begin each substantive topic by engaging in typical pedagogical approaches—lectures, readings, and discussions—but follow up with a lab assignment that substantively reinforces what we learned in class. Lab assignments must build on one another and must consider the students' ability and experience. In my experience, we can assume that majors have little knowledge or understanding of data, statistics, and the social scientific method, even if they have taken a course in one or all of those subjects. These students often are hesitant about numbers and often seem to have difficulty using their computer beyond basic email skills. Therefore, I assume that the lowest common denominator is no previous data or statistics course. It is important to remember that when students are learning about data science, there are many moving parts: computer program, computer code, syntax files, output files, data files, reading output, understanding output, writing about and discussing data, and applying it to the real world. These parts are all complex and multifaceted, and students need to learn through repetition and continuous interaction with the material—otherwise, they are overwhelmed. Simply stated, data analysis is difficult. Students do understand concepts such as averages and frequencies, which provide an initial toolkit to work with while they learn to manipulate data and become comfortable working within a program. Therefore, I start simple and become more complex.

During the first lab assignment, some students are completely lost and frustrated, but they endure. By the third lab assignment, they can begin without assistance from the instructor. By the sixth and final lab assignment, students tell me they are adding these skills to their resumés.

I use class time for working on labs and I provide extensive office hours to foster student success. Students do not feel comfortable asking for help, which is why in-class homework time is essential. As the teacher, I must actively engage with students, review their work, help them see small errors, and provide ongoing feedback and encouragement. Students are given all of the lab output related to assignments in a PDF document along with the assignment so that they can be sure that their results are correct. This reduces the number of errors that they make.

For each assignment, students must turn in their syntax file, lab answers, and appropriate data output. This is important: I do not want a data dump of every action they took and I do not want summary information about the variable in question if it is not relevant. Looking at the data and knowing which results to include to answer a question are important aspects of learning about how to use data to understand the political world.

Their syntax file and corresponding output also are important for finding mistakes and ensuring that the syntax would produce the outcomes that the students present. Because students have the expected output in a PDF file, they simply can cut and paste the answers. When I see mismatches between the syntax file and a student's output file, I know something is wrong and I speak with the student.

After each lab assignment is graded and returned to the students, I have a session in which I review the lab answers. An essential component of the course is teaching data literacy, and how to describe data and results is a critical part of that. Talking about and describing data in words is difficult. Students typically are unspecific and tell the reader to look at their output, which is not data literacy and is unacceptable. They often talk about point differences across categories as simply

doi:10.1017/S1049096521001165

"large" or "small"; however, over time, the lab assignments teach students to think about the scales and numbers relative to the variables in the data that they are working with and to use the data more directly to discuss output. Because students manipulate data and repeatedly examine the same variables in different contexts, the patterns within them become more concrete and more comfortable to them.

Because students manipulate data and examine the same variables repeatedly in different contexts, the patterns within them become more concrete and more comfortable to them.

This also is important because I allow students to resubmit previous assignments for more credit and a better grade. Students usually submit an assignment only twice, but some will submit a third time. Regrading homework that is lengthy and detailed is a high cost for the instructor, but I have found that making mistakes and fixing them is the best way to learn. Making mistakes without fixing them and simply moving on to the next assignment results in poor performance, loss of confidence, and a reduced motivation to engage the material. By offering opportunities for students to improve their grades, they will interact more with the materials resulting in improved data literacy and a better understanding of human decision making.

However, students must resubmit assignments in an orderly manner. Due dates must be given for the first revision shortly after the graded assignments are returned. Otherwise, students will wait until the last day of class to return assignments, thereby losing the opportunity to learn from their mistakes and apply that learning to subsequent assignments. Because the assignments are cumulative, they will fall behind in their understanding if they do not fix their mistakes shortly after an assignment is returned.

I use SPSS Statistics because it is available on all campus computing pods and costs only about $30 for students to purchase their own program for the semester. I have used other programs in the class (e.g., Excel and `Stata`) and I think any program could be adopted easily, including `R`. However, the program is simply a tool to learn about how data can provide powerful information about political behavior; after students have learned basic skills in any program, they can apply those skills elsewhere.

LAB ASSIGNMENTS

The lab assignments include only basic skills development in statistics, including frequencies, means (and t-tests), cross-tabulations, and cross-tabulations with controls, as well as the use of correlation coefficients and chi-square tests of independence. However, these basic skills, along with the ability to manipulate data, form the building blocks of statistics and prepare students for future courses and more complex data analysis (e.g., regression).

I introduce students to the latest American National Election Study (ANES) and the class Dropbox that contains the dataset along with each lab assignment and correct homework output for each lab. The introductory lecture covers how to open the dataset, the different data views, the concept of cases as rows and columns as variables, and a data matrix.

Next, we discuss sampling and statistical inference and why polls theoretically should be reliable and valid. Because polls can correct for sampling error, we discuss the importance of sample weights to correct the sample and make it more representative. We follow that discussion with an appropriate lab assignment designed to understand survey weighting, how it works, and its implications.

The lab assignments include only basic skills development in statistics, including frequencies, means (and t-tests), cross-tabulations, and cross-tabulations with controls, as well as the use of correlation coefficients and chi-square tests of independence.

The assignment is statistically simple using only frequencies, but this simplicity introduces students to the notions of cases and variables, sampling, sampling error, inferential statistics, population statistics, and representativeness.

The assignment requires students to make comparisons across weighted and unweighted variables and to determine how the sample was incorrect and how it had to be improved through weighting. For example, in the 2016 ANES, unweighted and weighted sex is hardly different but education levels change quite dramatically. This suggests that the ANES was impressive in attracting responses from adult US citizens in relation to their sex but not their education; thus, this simple assignment accomplishes much. Moreover, the repeated use of a simple frequency command provides students with easy first commands to focus on while they also are learning how to open data files, write and save data syntax files, and resolve other computer-related hardware and software issues.

The assignment also is important for understanding how point estimates from samples provide information about populations and subpopulations. Surprisingly, this is a difficult concept. Students become distracted by the fact that there are hardly any Blacks and Native Americans in the survey and —relatively speaking—so many whites. For some students, this is an equity issue and a reason to view surveys as suspect and

https://doi.org/10.1017/S1049096521001165 Published online by Cambridge University Press

inaccurate. Thus, the weighting exercise also introduces students to population statistics—the ethnic, age, and education breakdown of the American public—and the idea that samples must be representative to be both reliable and valid.

Students also must understand that weights result in the same frequencies for population variables even when those variables have different respondent sets. Therefore, in the assignment, they also compare the variable sex weighted by the pre-wave and then the post-wave variables. They observe that the N for their sample changes but that the weighted percentages are unchanged, and they have to explain why. This also introduces the concept of a panel survey.

The second lab assignment is about personality and political behavior. This assignment includes several repetitive data manipulations and is prefaced with two lectures and readings on Jonathan Haidt's (2012) *Moral Foundations Theory* and the Big Five personality traits. The assignment requires students to construct the Big Five from the Ten Item Personality Measure. To accomplish this task, they must reverse-code one variable in each set, eliminate missing data, and then create an index. I ask them to create a sum and mean index for each of the Big Five traits, after which they describe each set of variables by examining their means. Students also consider one simple substantive question, which is the difference in personality traits between men and women. They only look at the means of the index; they do not engage in any statistical tests across groups. The point is to build confidence in data manipulation and to practice looking at and discussing data. This last part is surprisingly difficult; scales can be big or small, and learning that a difference of one third of a point may be big or small depending on how many points are in the scale is not necessarily intuitive—which is the purpose of this assignment.

At this point, we turn our attention to the role of political knowledge in political behavior. Our goals for the third assignment are twofold. The political-knowledge assignment provides important new data-manipulation skills, including the calculation of a political-knowledge scale from a series of questions and a formal opportunity to review measures of central tendency: the mean, the median, and the mode. This analysis focuses on the variation in political knowledge and the differences in political knowledge among demographic groups, partisanship, ideologies, and political engagement defined as voting. This links the first assignment, in which we considered group proportions in our survey, with the second assignment, in which we used averages to examine differences across variables and across groups within variables.

The fourth assignment examines partisanship and ideology. Cross-tabulation and tests for statistical independence and correlation coefficients are introduced. I teach students the chi-square test and use Kendall's tau correlation coefficients to introduce correlation and strength of relationships. The advantage of a cross-tabulation is that students can see that the percentages and the statistics are reinforced by the cell data. A strong relationship shows significant differences in percentages across groups, whereas a weak relationship shows the opposite. Examining how the cell data relate to and affect the statistical tests increases students' understanding of concepts including strength of relationships, point estimates, prediction, and variation.

For example, one question focuses on the relationship between turnout and partisanship. We first cross-tabulate turnout by the traditional seven-point party-identification scale: we ask for column percentages and request the chi-square and Kendall tau output. The cross-tabulation output shows that strong Democrats and strong Republicans turn out at 90% and 94%, respectively, whereas weak and leaning partisans turn out between 82% and 87%, respectively, and eligible voters who do not identify with a party turn out at only 67%. The chi-square test shows that there is dependence—that is, knowing about a person's party tells us about their turnout—however, the tau output indicates that the strength of the relationship is 0.011 and $p>0.05$. Students can see that there is a relationship: it is clear in the cell percentages and the chi-square tests confirm it, but why does the tau output not capture it? Students then create a variable, which is the strength of partisanship that requires them to combine strong partisans, weak partisans, leaning partisans, and no partisans into an ordinal variable and then rerun the cross-tabulation. They now see that the chi-square and tau correlation confirm one another's findings. This brief assignment teaches students that statistics often rely on linearity and that nominal and ordinal variables are different. Clarifying the relationship and then rerunning the data and the statistical tests helps students to understand how the structure of data matters to statistical tests.

The fifth assignment focuses on issues and groups, introducing the concept of a control variable and the difference-of-means test. This builds on assignment 4, which focused on bivariate cross-tabulations and demonstrated that partisanship is a strong predictor of presidential vote choice. However, this assignment instead uses party as a control variable and gender and ethnicity as the focal variables. In the case of gender, the gender gap is shown in vote choice, which in 2016 exists but is somewhat weak. We have a significant chi-square and a significant but weak tau output (0.06), and the data in the cells show that women are about 7% more likely to vote Democratic than men. After we control for party, however, we find that there is no cell difference between men and women in their vote choice: 93% of Democratic men and 92% of Democratic women voted for Clinton and 92% of Republican men and 91% of Republican women voted for Trump. This is an insignificant chi-square value and an insignificant tau output. In the case of ethnicity, controlling for party reduces the strength of the relationship, but the relationship does not disappear as it does with gender. Thus, students see different patterns in the results, and these differences highlight different aspects of data literacy.

In this lab assignment, I also introduce the idea of context and how we can overlay context on our data to test additional hypotheses. In this case, we examine how being in a competitive presidential-election state versus a safe blue or red state affects political behavior. I ask students to identify which states were considered safe and which were toss-ups

https://doi.org/10.1017/S1049096521001165 Published online by Cambridge University Press

and then to code them as dummy variables. The assignment requires them to consider various political activities (e.g., turnout, political messaging on Facebook, giving money to a candidate, and discussing politics) and their relationship to context using t-tests. The students find significant and insignificant variables as well as relationships that seem intuitively inconsistent with the idea that being in a swing state with more mobilization and advertising leads to more activity. There is no significant difference in turnout or posting a political article on Facebook; however, students see that more money is raised and more political discussion occurs in non-swing states. Thus, they not only must look at the significance, they also must connect and interpret the means.

The sixth lab focuses on issue voting and polarization, bringing everything together. Students must manipulate a great deal of data and answer substantive questions using both t-tests and cross-tabulations. Instead of looking at questions one at a time, they crunch considerable data, create tables to summarize their results, and then consider the weight of the evidence as defined in their tables. I ask them to look at cleavages within and between the parties by using vote choice during the nomination campaign to obtain within-party differences and by using vote choice during the election to obtain cleavages across party. Because this is their final assignment and they will not be able to submit a revision, I offer an extra-credit question.

RESULTS

My measures of student success are anecdotal but reflect more than 25 years of teaching. By the end of the semester, students feel like they have accomplished something. They generally receive grades of As and Bs, which they justly earned through completing and fixing assignments. In addition, the material that they substantively covered is more concrete to them. The data assignments reinforced what we read and discussed, promoting a deeper understanding of the material. They also feel that they have a better understanding of data and how to manipulate it and know how to ask questions using data and how to read and interpret data results. Students have expressed this to me by indicating that they liked the course because it was "applied," "hands-on," and "challenging." Many students have requested letters of recommendation and have contacted me to discuss career options, including graduate school and jobs in public policy and data science. The first year that I changed political behavior to a data-intensive course, a student sent me this note during graduation ceremonies: "Dr. Atkeson, it was so meaningful to me to have a professor who trusts her students with challenging work. Thank you for trusting us and giving us the opportunity to actually work with data." ■

REFERENCE

Haidt, Jonathan. 2012. *The Righteous Mind: Why Good People Are Divided by Politics and Religion.* New York: Vintage Books.

https://doi.org/10.1017/S1049096521001165 Published online by Cambridge University Press