# Expectations and Skills for Undergraduate Students Doing Research in Statistics and Data Science

**Abstract**

As more statistics undergraduates participate in summer and capstone research projects, it becomes paramount for teachers to prepare them adequately for the experiences. Indeed, the most successful research students are those with both a strong understanding of foundational statistical methods as well as a fluent ability to wrangle data. In this talk, I will give some examples of undergraduate research projects and the skills necessary for students to be most successful.

## Introduction

Undergraduate research has long been understood to be one of the strongest high-impact educational practices[[1]](#footnote-1). However, there are myriad paths to successful undergraduate research, and there is very little information on how to link successful research practices to the undergraduate curriculum. In recent years, the statistics and data science communities have come together to propose curriculum guidelines for undergraduate programs (GAISE[[2]](#footnote-2), Stats Guidelines[[3]](#footnote-3), Data Science Guidelines[[4]](#footnote-4)). The suggestions are well thought out and meant to balance competing demands of time/units, student background, and existing curricular structure. The guidelines have been written to encourage modernization of standard curricula to catch up with exciting new ideas being developed for the statistics and data science classroom. Of course, one important aspect of any curriculum is to set goals for the graduates of that program. That is, what skills are necessary for students coming out of a specific statistics or data science program who will enter either the workforce or a graduate program?

Less common in current curriculum guidelines are specific details about how the pieces of the curriculum work together. In particular, we are concerned here with how the course assignments and structure can support undergraduate research projects in statistics and data science.

Almost every summer since I began my faculty position, I have worked with students on projects related to my own research areas. Often those projects extend into year-long senior thesis projects or repeated summer projects. I have published with my students as first author or substantial contributors in *Statistical Applications in Genetics and Molecular Biology*, *BMC Bioinformatics*, *Briefings in Bioinformatics*, *Environmental and Ecological Statistics*, *Computational Geosciences*, and *Chance*. Certainly, some projects have been more successful than others; indeed, there are plenty of projects that did not result in a peer-reviewed publication. A project’s success depends on some things that are unrelated to the curriculum; e.g., how conducive the project is to undergraduate exploration, or whether the student is interested in the work, etc. But most of the success of the project **does** come directly from the experiences the student has prior to starting the research. Below I detail what I consider to be the curricular aspects most important to a successful research project in statistics or data science.

## Skills

As we all know, doing research takes a myriad of skills. The skills below are all taught to some degree or another within a standard curriculum. I submit, however, that it is the skills that typically get less attention that are most valuable for research.

Making an argument

The most important skill for successful research is the ability to make a convincing argument. Making an argument requires that there is a novel idea and a method by which to argue the point. Students can often approach the problem in different ways: theoretically (e.g., mathematical proof), simulation, or via a literature review. Knowing that any given idea **can** be argued in different formats is surprising/illustrative for students. By understanding that there are different creative ways to solve a problem, they are given the freedom to harness their own skills and comfort with the research project.

One of my recent projects involves creating prediction intervals for a random forest model. The novelty comes from the derivation of the appropriate standard error. There are a handful (not many) of papers on the topic, a few very theoretical papers and a few who approach the problem in a different applied way. My student and I have had to work through *how* our ideas add to the literature and *how* those ides can be synthesized into an argument. Our conversations circle back repeatedly to “what are we trying to argue and how can we argue that effectively?”

The classroom is an ideal place to demonstrate that there is almost never only one solution or argument to a problem. A simple comparison of mean versus median hypothesis testing shows that two tests can address the same underlying scientific hypothesis. Simulation studies (e.g., to determine coverage rates for different bootstrapped confidence intervals) give a way of understanding the theoretical outcome

The curricular aspect to making an argument relies on the professor continuing to ask a student “how do we know that?” or “how can we argue that one test is better than another?” Students are often so caught up in the weeds of learning the techniques that they fail to reflect on the process by which the method was developed or by which it has become popular.

Engaging with theory

The theoretical underpinnings of statistics have long been a cornerstone of most undergraduate programs. Certainly, my most productive research experiences have been those done with students who have strong knowledge of undergraduate topics like probability theory, distribution theory, maximum likelihood, and regression modeling. And although some of my projects have built on those theoretical constructs, it is the intuitive theoretical grounding of core principles in statistics (e.g., sampling distributions, interactions) that are imperative to a successful research project. Indeed, whether a student can derive a particular moment generating function is much less important than whether they understand that knowing every single moment of a distribution uniquely defines that distribution.

A solid theoretical background builds not only a strong intuition but also an ability to read the literature and place the research in a larger framework of knowledge. With undergraduates, I do not try to prove results with measure theoretic tools. Instead, I often work with the students to simulate scenarios and gain an understanding of what others have done within that more theoretical framework. In order for the students to have new insights, they must be able to understand the general structure in which their work resides.

A recent project of mine used canonical correlation analysis to identify correlated *pairs*  of linear combinations of variables. The setting is sufficiently complicated that it would be difficult to find the theoretical distribution underlying each correlated linear combination (keeping in mind that each pair is also correlated with other pairs), but the analysis is not useful unless there is a way for the practitioner to know whether a large correlation is actually statistically significant. We were able to derive a permutation algorithm to define significance (the method also doubled as a way to measure false positive and false negative rates). The permutation method, however, was not trivial to implement, and it required that the students understood how the distributional aspects are determined by both the linear combinations as well as the complex correlation structures.

I suggest that many theoretical statistics and data science courses are already providing the needed background to make a student researcher successful. However, when teaching, for example, the Neyman-Pearson lemma, the intuition behind how we know what we know (and why it matters!) is vastly more fundamental for the students’ future research capabilities than the detailed steps of the proof.

Working independently

There is likely no question that a student’s ability to work independently can be the most important aspect of a successful research project. As busy advisors, a needy student can be an inordinate drain on our time. And if we are providing guidance at every step, then we might as well be doing the research ourselves. (One way to cut down on contact hours is for a pair or trio of independent students who can be even more successful than an individual in working effectively on a successful research project.)

While absolutely important, one might argue that a student’s ability to work independently is not something that can be taught in a standard curriculum. I disagree. By providing the student with both **practice** working independently and time to **reflect**, all of our students can learn to generate their own ideas in a productive way.

I encourage all upper division (and dare I say lower division?) statistics and data science courses to contain project based assignments with some degree of autonomy[[5]](#footnote-5). That is, the project should have, as one of the assigned tasks, a directive for the student to do something independent[[6]](#footnote-6) (e.g., teach themselves a twist on a topic already covered in the course). Admittedly, projects can be time consuming to grade, especially in large classes. However, there are structures and tools to make projects more attractive for the instructor. For example, working in groups cuts down on the number of projects to grade, and peer assessment gives the students a sense of what other students have created. GitHub[[7]](#footnote-7) and GitHub Classroom[[8]](#footnote-8) are resources that streamline both group work and assessment and are nevertheless valuable skills for statistics and data science students. Additionally, Jenny Bryan has put together Git resources[[9]](#footnote-9) that are streamlined and easy to follow.

Along with doing the project, another aspect to developing independent research is providing structure for the student to reflect on their work. For any assignment (project or other), the student should be able to express **what** they did, **why** they did it, and what the **next** step should be. Reflecting on what they still do not understand can be incredibly valuable. Nick Horton (personal communication) requires his students to reflect via a Google form with a mechanism running in the background to inform him if they didn’t do it!

A quick reflection on a notecard or as part of the end of the assignment teaches the student to deliberate on what they have done and to think about the next steps in the process. A strong research student will come to your meetings with both work accomplished as well as ideas for moving forward. The process by which they generate ideas for moving forward is not innate and can be learned through repeated practice.

Wrangling data

I have included data wrangling as a core skill because it is more and more important to almost every research project I see (in my group and in others). Indeed, even the theoretical projects on which I am involved often use examples that require substantial data wrangling. Additionally, I continue to see data wrangling as a core tenet that is overlooked in many curricula.

Working with data is the only way to get good at working with data. Our students should be graduating with a fluency in programs like the dplyr[[10]](#footnote-10) R package. It is not only a vital aspect of having a successful research project, but it is the key to a successful career in any data related field.

Practice in data wrangling should come early and often, and the swirl[[11]](#footnote-11) R package and DataCamp[[12]](#footnote-12) platform use interactive tools to get students started. Requiring good data wrangling skills in your courses will benefit you and your students in the long run.

## Discussion

I’ve structured this article to focus on the learning goals associated with curricular choices in order to produce strong researchers (or new members of the workforce who can deal with the complex challenges that they will face in a world full of data). There are many other choices a student makes along the way that can contribute to a successful research project. For example, engagement in a field outside of statistics and data science can generate excitement for solving a particular applied research project. Or working with new software programs (e.g., the quo function in dplyr version 0.7.1 as of June 22, 2017) can give the student a sense of being part of a larger community of statisticians and data scientists.

Alas, the most important aspect of successful research is the degree to which YOU are excited about the project. If you love what you are doing, the student will sense that, and they will be just as engaged. So if there is something you want to work on, I implore you to assign an undergraduate student to the project regardless of their background or the curriculum from which they came.

This article grew out of an Invited Paper session “Modernizing the Undergraduate Statistics Curriculum” (session #480) at the Joint Statistical Meetings (JSM) in Baltimore, MD, 2017.

1. See high-impact practices suggested by the Association of American Colleges and Universities (AACU), <https://www.aacu.org/leap/hips> [↑](#footnote-ref-1)
2. http://www.amstat.org/asa/education/Guidelines-for-Assessment-and-Instruction-in-Statistics-Education-Reports.aspx [↑](#footnote-ref-2)
3. http://www.amstat.org/asa/education/Curriculum-Guidelines-for-Undergraduate-Programs-in-Statistical-Science.aspx [↑](#footnote-ref-3)
4. https://www.amstat.org/asa/files/pdfs/EDU-DataScienceGuidelines.pdf [↑](#footnote-ref-4)
5. Again, projects are a high-impact practice as defined by the AACU, https://www.aacu.org/leap/hips [↑](#footnote-ref-5)
6. Such an assignment has the added benefit of allowing you to use it on letters of recommendations. I sometimes provide details about the topic that a student has learned and communicated completely independently. [↑](#footnote-ref-6)
7. https://github.com/ [↑](#footnote-ref-7)
8. https://classroom.github.com/ [↑](#footnote-ref-8)
9. http://happygitwithr.com/ [↑](#footnote-ref-9)
10. https://cran.r-project.org/web/packages/dplyr/index.html [↑](#footnote-ref-10)
11. http://swirlstats.com/ [↑](#footnote-ref-11)
12. https://www.datacamp.com/ [↑](#footnote-ref-12)