Crossroads in Statistics Doctoral Education: Re-envisioning PhD Training

Rapporteurs: Deborah Nolan (chair), Jennifer Hill, Nicholas Horton, David Madigan

Prepared for: The National Science Foundation

Introduction

Statistics as a field has been under continuous evolution since its conception. In recent years especially, the field has changed dramatically, as it has leveraged new computational and data technologies and spread to nearly every field of human endeavor. As a result, our PhD programs no longer reflect the direction of the field and require urgent reform. In essence, we stand at an important crossroads where the decisions we make now will have far-reaching consequences for our students and for the field.

Statistics represents a vast and rapidly evolving body of foundational knowledge and experience integral to our data-centric society. Fundamental concepts continue to manifest themselves in novel ways, as the data we collect -- and how we collect it -- take new and diverse forms. For example, recent concerns regarding `algorithmic bias' fundamentally stem from issues of (un)representative sampling and have been shown to replicate existing biases; the widespread adoption of `A/B testing' within the high-tech industry reflects a resurgence of experimental design and randomized trials at scales previously unthinkable; omics-data have surfaced multiplicity challenges on previously inconceivable scale. At the same time, the need for new fundamental concepts in statistics is growing at a rapid pace. The rise of big-data and the substantial body of results surrounding inference and prediction in so-called `small n, large p' problems is particularly illustrative and relevant in the modern era.

To address these challenges, two NSF-funded workshops on "Statistics at a Crossroads" were organized in fall 2018. The report of the first workshop (https://www.amstat.org/ASA/News/Statistics-at-a-Crossroads-Recommendations-Are-Released.aspx) stressed the need for the profession to emphasize the central role of practice, focus on impact, reinforce the need for research for better practice, embrace grant challenges, develop broader metrics of methodology evaluation, and develop training in modern skills[1]. The second workshop[2] laid out the roadmap for a dramatic re-envisioning of PhD training. While the workshop attendees argued about some of the details, they were unanimous in the belief that an overhaul of statistics PhD training is long overdue.

[1] The report from the first workshop included some discussion of doctoral education, the need for changes in curriculum, ideal attributes for applicants to statistics doctoral programs, and the need for more effective training. These ideas were expanded upon and refined in the second workshop.

[2] The second workshop was funded by NSF Grant 1844975, Graduate Statistics Curriculum at a Crossroads (Deborah Nolan, PI). Members of the steering committee for the workshop were Jennifer Hill, Nick Horton, David Madigan, Deborah Nolan (chair), Donnalyn Roxey, and Daniela Witten. The list of workshop participants can be found in the Appendix.

In this report, we make six broad recommendations[3] for modernizing PhD programs and the university in order to better train the next-generation of statisticians:

1. Holistically integrate theory and practice
2. Build a village of advisors
3. Elevate computational and algorithmic thinking
4. Forge newer forms of industry relations
5. Deprecate the classroom
6. Broaden creative work in the professoriate

These recommendations advocate for core changes in our doctoral programs. They do not provide details on how to update curricula in courses. Instead, their aim is to point to essential qualities of PhD training that we think need to be addressed. While the recommendations are accompanied by specific examples, these examples are meant to spur ideas and not be overly prescriptive.

Underlying these recommendations is the consideration that fundamentally **the PhD is a research degree**. As research scientists, our graduates should be able to contribute to the solution of real-world data-centric problems through the creation of novel statistical objects (e.g., models, methods, visualizations) or the analyses of such objects.[4] Moreover, our graduates must have the ability to envision and conduct data-enabled research in a broad array of fields.

The recommendations outlined here primarily focus on post-coursework research training; in fact we believe that the precise coursework that students undertake can and should vary from program to program and should tailor itself to the intended research topic. Nonetheless, we include general recommendations for changing the nature of coursework at the graduate level.

Overall, these guidelines are ambitious in that they aim for a holistic integration in training, inspired by a cycle fundamental to the statistics discipline: *theory informs principles, principles inform practice, and practice in turn informs theory*. For example, formal asymptotic results, like central limit theorems and strong laws, inform our notions of the behavior of averages, which in turn inform our interpretation of and modeling of aggregate measurements in practice. Similarly, various notions of data-splitting popular in machine learning applications, necessitated by the practical need to calibrate sometimes large numbers of smoothing parameters, are informed by formal results characterizing specific versions, such as cross-validation.

---

[3] These recommendations are the outgrowth of the NSF-sponsored workshop on the topic. Additionally, feedback was sought from researchers in industry, computer science, and business and the statistics community via a pre-workshop webinar. Speakers at the webinar were Julie Novak Beckley, Netflix, Michael Rappa, Institute for Advanced Analytics, NCSU, Renata Rawlings-Goss, South Big Data Innovation Hub, and Padhraic Smyth, School of Information and Computer Science, UC Irvine.

[4] This is the stated premise in Section 5 Doctoral Education of the report from the first Cross Roads workshop report (p. 23).

Future PhD programs in statistics must prepare students to identify and solve novel and challenging non-standard data-rich problems. Emerging PhD statisticians should have strong job prospects not only in academia, but also in industry, science and policy, education, government, etc. Given that the preponderance of employment demand in data science generally over the next few decades is forecast to be largely in these latter sectors, our proposed evolution of statistics PhD training responds to a critical need in society at large. Furthermore, the breadth of opportunities available to PhD statisticians demands an increasing pool of applicants with diverse skills, interests, and backgrounds. We believe these recommendations will open greater access for participation in statistical science.

Recommendation #1. Holistically Integrate Theory and Practice

We acknowledge that a central task in modernizing PhD programs is to identify traditional and emerging statistical theory, principles, and methods essential to the training of doctoral students in statistics for conducting innovative and impactful research as they enter deeper into the digital age. Additionally, at its core, PhD training should provide reasonable exposure to the major frequentist, Bayesian, likelihood, information theoretic, and predictive perspectives; an understanding of their respective strengths and limitations; and appreciation of how they differ from each other. However, in this report, we wish to emphasize that the teaching and learning of topics should be closely interwoven with practice throughout the entire PhD training. Students need to build statistical insight both in their research *and* in early training in foundations, which takes place in the initial years of a PhD program or in Masters programs.

Many statisticians already espouse that theory and methods are best taught with real-data applications and projects, following the mantra that theory informs principles, which in turn guides practice. We argue that our PhD programs have not gone far enough in this direction. Our students need to develop expertise in the key steps in modeling and analysing data, the sequential and iterative nature of these steps, and appropriate computing platforms on which to carry them out. The detailed methods statisticians use for each of these steps largely depend on both the goals of the problem and the perspectives adopted; as such, the integration of theory and methods in training is essential. Real data-driven applications motivate and inspire theory and the development of methods that are more generally applicable.

A fundamental understanding of the key aspects of working with data-centric problems can be learned by drawing explicit connections between theory and practice while engaged with challenging authentic problems. For example, practical implementations of textbook sampling theory and experimental design often requires compromises and uncomfortable assumptions, and understanding the tradeoffs and implications is not found in books, but in practice. Similarly, at times statisticians have to proceed as if there are no unmeasured confounders and no selection bias, and they need to know when it is reasonable to proceed in this manner. Likewise, practice often surfaces subtle trade-offs between data quantity and data quality, but little theory exists to support the practitioner navigating this frontier. Gaining a deep understanding of these

and other fundamental statistical insights requires filling the gap between theory and practice in the curricula and beyond.

Example: Modern Masters Programs in Analytics[5]. Masters programs have been developed around the notion of a practicum that use the process behind practice as the organizing principle, integrating experiential learning through consulting and collaborative projects with pedagogy in asking questions and interpreting results, in data wrangling, workflow, and reproducibility, and in various forms of communication. The choice and organization of topics in courses in theory, methods, and computing are influenced by the practicum. This highly coordinated training is especially useful when a cohort takes all their courses together. A similar coordination in PhD training is both feasible and desirable, albeit with a different balance of theory, methods, and computing.

Recommendation #2. Build a Village of Advisors

Alternatives are needed to the traditional one-on-one advising model of PhD students. Research problems are increasingly complex and span multiple fields, and are better tackled by groups of students, postdoctoral scholars, and faculty. The demand for PhD-trained statisticians in industry continues to grow and industry tends to value researchers who have the additional skills needed to contribute as part of a team. It is the responsibility of faculty to advise graduate students and create a culture of enrichment that fosters the development of students' creativity and ability to conduct research.

Opportunities to receive advising and mentoring from multiple faculty, as well as peers and other collaborators is arguably a more robust approach to research training. The notion of co-authored dissertations should be within the realm of possibilities. We provide below three examples of how advising might take place outside the one-student-one-advisor model. Common across all three examples is the expectation for all group members to make presentations to the group and to give and get feedback from each other. Through these organizations, we can build a community where assistance comes not only from faculty advisors and where group members learn how to support the development of junior colleagues.

[5] This example is based on the Masters in Statistical Practice program at Boston University (http://www.bu.edu/mssp/) and the Masters in Analytics at North Carolina State University (https://analytics.ncsu.edu/)

Example: The Reading Group[6]*.* The structure of a reading group can be a useful research training mechanism. One way to organize a group is to choose a theme for the term, invite guest lectures, choose readings, and schedule presentations of the readings and on-going work related to the theme. Changing themes regularly draws different attendees and expose students to various viewpoints. Similarly, when several faculty, including junior faculty and faculty with different backgrounds, co-organize the reading group, then students benefit from observing how others more experienced than they master new material. Observing discussion between faculty and asking students to push themselves to think critically about what the organizers are saying has tremendous value. Additionally, the informality of the reading group provides a flexible environment that can respond to differences in member backgrounds with tangents that quickly fill in crucial missing pieces.

Example: Shared Advising. Here a group of faculty organize regular joint meetings with their students and postdocs. In addition to group meetings, faculty also meet regularly one-on-one with advisees; they also identify projects that teams of students might work on together without faculty leadership. This network of interactions creates a sense of community: students work together outside of group meetings, small collaborations form, and they help one another with scientific, statistical, and computational problems. Other training is achieved by requiring each member to present work in progress to the group, where they get feedback about technical aspects of their research and where they develop their communication skills in an informal and supportive environment. These meetings can lower the barrier for students to discuss their research with one another in other settings. Additionally, the students rotate responsibility for organizing the group meetings from semester to semester.

Example: The Science Lab. Yet another approach to advising is based on the notion of a science lab (e.g., in the biological sciences) with regular team meetings of all the faculty member's graduate students and postdocs. This group differs from the shared advising example described above in that it is led by a single faculty member. Here, graduate students advised by other faculty, students who have yet to begin research, other faculty and visitors are welcome to participate in the lab meetings. These additional students commit to attending the lab meeting for a term. Meetings include presentations of work in progress and more general discussion and training.

---

[6] The Causal Inference Reading Group at Berkeley has run semesterly since fall 2016. The group meets once a week for two hours and typically 3-5 junior faculty and postdocs convene the group. The size of the group varies, but typically includes 15-20 graduate students from across campus with the majority from statistics and biostatistics, and a few every term from economics, political science, math, and related disciplines. The group also includes undergrads, postdocs, and faculty who drop in on occasion. The expectation is that each graduate student (or pair of students) gives one paper presentation per term. These presentations vary from detailed formal talks to more casual 'chalk talks' and updates on research in progress.

We are encouraged by a number of institutions that have adopted collaborative advising models. We also acknowledge that many aspects of intellectual property and reward structures remain to be addressed to allow these changes to be adopted more widely.

Recommendation #3. Elevate Computational and Algorithmic Thinking

Computing is vitally important to data-centric science. “Computing” in this context refers not just to programming but to a design and analysis perspective that creates efficient and readable code, reproducible data analysis, and extensible software. Students have received math training for decades, and computational training needs to catch up. They must gain the capability to incorporate state-of-the-art computation in data-intensive projects, and develop new computational frameworks and paradigms for data science. A PhD statistician should have computing skills at the level needed to participate in collaborative science teams that deal with large data and intensive computations. In this regard, computational reasoning plays as fundamental a role as mathematical, statistical and design reasoning.

In order to move beyond the current minimum practices, we should expect students entering PhD programs to hold basic computational reasoning skills. This requirement seems achievable with the rise of data science at the undergraduate level, although the depth of preparation varies dramatically across institutions[7]. However, we also advocate that this training must continue at the graduate level. Within the PhD experience, we recommend extended (deeper) computational reasoning for novel data science problems so that our students can fully engage in research projects. As described in Recommendation #1, this training must be interwoven throughout PhD research training and structured spiraled curricula in ways that are only rarely seen as required components of statistics programs.

Finally, in order to realize these changes, academia must invest long-term in faculty or related research positions that support, train, and do research in statistical computing (see Recommendation #6).

Recommendation #4. Forge Newer forms of Industry Relations

The relationship between academia and industry is evolving rapidly. It is becoming increasingly common for faculty at some elite institutions to take extended industry leaves, begin startups,

[7] The NAS "Undergraduate Data Science: Opportunities and Options" report (2018, https://nas.edu/envisioningds) called for bachelor's level data scientists to have the following concepts/skills (part of a broader definition of "data acumen"):

- Computational foundations: Basic abstractions, Algorithmic thinking, Programming concepts, Data structures, and Simulations.
- Workflow and Reproducibility: Workflows and workflow systems, version control systems, reproducible analysis, documentation and code standards, source code (version) control systems, and collaboration.

and split their appointments with more than 50% of their time spent in industry[8]. The opportunities afforded by industry collaborations hold tremendous research potential for faculty, e.g., industry partnerships can provide access to data and computational resources that are not available within academia. As the relationship between academia and industry becomes increasingly interconnected at more institutions, we need to understand and manage these collaborations. This is largely uncharted territory and may have a huge impact on university culture, our education mission, and the form of graduate education.

One concern is that nimble entrepreneurial faculty who can make connections with industry researchers are most easily poached by industry and such poaching has the potential to gut the research core of the university. Another is that industry interests are typically more focused on projects with immediate payoff, which could be in conflict with the broader high-risk high-payoff long-term research in which faculty engage. Despite these concerns, we should be open to novel partnerships that may take new forms: dual advising with non-faculty serving on dissertation committees and adjunct appointments; industry-funded fellowships and scholarships with future employment in mind; admission of salaried company employees into PhD programs. Examples that may offer starting points to redefining industry-academia relations appear below.

Example: *Industry Research Exchange* – Industry sponsors of academic research could be organized so that faculty and graduate students carry out their research unfettered and with no intellectual property (IP) limitations on publication. The sponsors gain early access to this research through workshops where graduate students present their findings and sponsors provide feedback to the students and describe the problems that they face in practice. In this way, a sponsor gets early access to the latest industry developments and they get to influence the direction of the research by describing their practical considerations. This kind of exchange might be in a specified area of research and for a limited time period.

Example: Independent Research Commission. Conflicts between a company's desire to maintain data privacy and an academic's desire to publish freely are potentially accommodated

---

[8] The Computing Community Consortium report, Evolving Academia/Industry Relations in Computing Research (Morrisett, Patel, Rexford, and Zorn), https://www.cccblog.org/wp-content/uploads/2019/03/Industry-Interim-Report-w-footnotes.pdf reaches the following conclusions about the current trends in interaction between academia and industry in computing fields:

- In certain computing disciplines, such as artificial intelligence, we observe significant increases in the level of interaction between professors and companies, which take the form of extended joint appointments.
- Increasingly, companies are highly motivated to engage both professors and graduate students working in specific technical areas because companies view computing research and technical talent as a core aspect of their business success.
- This increasing connection between faculty, students, and companies has the potential to change (either positively or negatively) numerous things, including: the academic culture in computing research universities; the research topics that faculty and students pursue; the ability of universities to train undergraduate and graduate students; how companies and universities cooperate, share, and interact.

through a commission acting as an independent, trusted third party[9]. The commission would consist of respected scholars, and they would receive access to all relevant company information and systems. The commission would define specific research areas for the broader scholarly community, and invites academics to apply for grants to conduct research in these areas. The grant proposals would be vetted by the commission and funded by nonprofit foundations. The academic researchers who are awarded funding would be granted access to anonymized data to carry out their research, and they would not need to obtain approval from the company before publishing.

The innovations alluded to here raise complex intellectual property issues (see https://uidp.org for further important questions). Restrictions on publication may prove especially problematic but faculty and professional societies should push University General Counsel offices to embrace new models that can benefit all parties.

Recommendation #5: Deprecate the Classroom

Successful PhD training requires mastering a number of skills, some of which are taught explicitly, or formally, in courses and some of which have been taught more informally, on the side through individual mentors and advisors, if at all. Since PhD programs have different emphases and different student populations, we do not aim to prescribe the essential elements of formal statistics PhD training. Instead, we advocate for flexible curricula that provide the holistic integration of theory and practice as outlined in Recommendation #1 and that breaks free from administrative constraints imposed by the current university system of formal courses taught in term-length sequences.

As data science gains popularity, the field of statistics will experience a spillover effect, and students from nontraditional backgrounds will be interested in pursuing graduate studies in statistics[10]. This greater diversity in student backgrounds requires that statistics programs must be nimble to facilitate the integration of these students into research. Creating flexible PhD programs that allow more student choice is essential to the success and survival of the field.

PhD training often delays or ignores teaching the hardest part of statistics -- that which lies at the interface between theory and practice. The statistical reasoning behind practice often goes unmentioned or underemphasized in our formal courses. What lies in the gaps between theory

[9] This example is from the working paper “A new model for industry-academic partnerships” by King and Persily available at http://j.mp/2q1IQpH

[10] Historically, many statistics doctoral students have completed mathematics majors in college. This training serves many students well during their graduate coursework and has historically been helpful in passing qualifying/comprehensive exams. However, many mathematics programs do not provide depth in statistical practice, computation, and analytic workflow emphasized in other majors such as statistics and data science. As a result, traditional matriculants to doctoral programs in statistics have a solid foundation in mathematical rigor but less experience and depth in computation, statistical analysis, and communication.

and practice needs to be brought forward and articulated in our training, and students need multiple opportunities to work at these boundaries. The traditional classroom may not be the best place for this kind of essential learning to take place.

Addressing both the diversity in student background and the need for filling in the gaps between theory and practice both can be accomplished with more flexible, creative training. We need new models for teaching and learning that are not constrained by the traditional term-length courses. These alternatives might involve combinations of core semester-length courses tied to shorter courses, labs, and workshops, and they might utilize longer-term internships where students can dive deeper into real problems. The university system has the potential to remove the traditional course restrictions, and support such novel frameworks[11].

Example: Early Practical Work[12]. While taking preparatory coursework, students can begin to work on an applied project. Each student partners with a researcher at the university, outside the department, and a faculty member in statistics advises the student. This immersive experience can be embedded within the training program, and can be an extensive year-long project. The client provides data and regularly meets with the student to ensure the student's analysis is relevant to the problem. The student also meets regularly with the faculty advisor from statistics for guidance. Alternatively, students can work in vertically integrated teams that consist of advanced undergraduate, professional Masters, and beginning PhD students, and these teams can be horizontally integrated with backgrounds in statistics, computer science, and domain related to the project. These interactions with the client and faculty supervisor guide the student in learning how to answer domain questions with statistical analyses.

Example: Bootcamps[13]. Practical bootcamps and online courses in computational training abound and are obvious ways to fill in gaps in computing background. Many partnerships are now being formed between bootcamps and universities. We advocate that some of these bootcamps should downplay code recipes and templates and focus on the computational model and other foundations that underly technology. The idea being that if you understand the computational model, then you can reason about the code and the language and can more rapidly solve (and avoid) problems with your code. The fundamentals provide the foundation to learn new technologies.

---

[11] Such initiatives will likely require more faculty, not less and that many universities are already challenged to retain their existing faculty.

[12] This example is based on the PhD program at Carnegie Mellon University (see the Advanced Data Analysis Project description at http://www.stat.cmu.edu/sites/default/files/grad_handbook_18-19.pdf) and the program at Rice University (see the Data to Knowledge Lab at https://d2k.rice.edu/).

[13] This notion is based on the workshop led by Temple Lang at UC Davis. See https://github.com/dsidavis/RFundamentals/blob/master/Outline.md

Recommendation #6. Broaden Creative Work in the Professoriate

Superior intellectual attainment in research is an indispensable qualification for appointment and promotion in academia. However, the changes in PhD training recommended here, if successful, will lead to a new brand of researcher and there must be room in academia for these new researchers to succeed, if our field is to remain healthy and vibrant. As the field of statistics enters new areas of endeavor and the nature of creative work evolves, we expect faculty research to depart markedly from established academic patterns. In consequence, the evaluation of a faculty member's qualifications for advancement and promotion need to be broadly judged according to whether they are engaged in a program of work that is sound, productive, and creative. We refer the reader to Waller (2018) for an insightful discussion of faculty evaluation in a data science age[14] and to the Moore Sloan Data Environments report on career paths (http://msdse.org/themes/#careers).

Research products will particularly be expected to differ from established norms, and we need to give consideration to these new modes of research inquiry as our field evolves. Evidence of productive and creative activity should be sought in a faculty member's published research as well as other types of products. New forms of research output, e.g., software products, should continue to provide evidence that the candidate is effectively engaged in creative activity of high quality and significance. Additionally, as collaborative work becomes increasingly common, joint authorship and other products of joint effort will be evidence of intellectual activity, and we must make informed assessments of the contribution of faculty in these joint efforts. Exceptional care must be taken to apply the criteria for promotion and advancement with sufficient flexibility without relaxing high standards.

Example: Industry-Academia Joint Appointments. Industry collaborations have the potential to greatly enhance a faculty member's research agenda through access to important real-world problems, increased computing resources, and access to highly skilled research support staff. These advantages to industry affiliations also bring major challenges. Joint appointments, where faculty spend a significant fraction of their time in industry, may lead to the expectation that faculty align their research agenda (and their students' research) with the needs of the company or to problems with intellectual property and publication.

Example: Professional Practice. Contributions to the advancement of professional practice or professional education, including contributions to the advancement of equitable access and diversity in education, should be judged creative work when they present new ideas or original scholarly research. Recognition of these contributions will be key to the successful evolution of our field.

---

[14] Waller, L.A. (2018) Documenting and Evaluating Data Science Contributions in Academic Promotion in Departments of Statistics and Biostatistics, *The American Statistician*, 72:1, 11-19, DOI: 10.1080/00031305.2017.1375988

Summary

This report has put forth six recommendations that have the potential to significantly change research training in statistics in the university setting, the impetus being the rapidly changing landscape of research in statistics and data science. Now is the time to seriously revamp research training and to boldly innovate our education programs. We must find new ways to fill in the gaps between theory and practice with more flexible, creative training, and we must move beyond minimum practice to regard computational reasoning as fundamental to statistical research. Such ambitious innovation calls for developing new more collaborative modes of advising and new forms of research partnerships that reach beyond the university. We firmly believe that pushing the boundaries along these core dimensions will bring about a broadening of the diversity of researchers in our field and place the next generation of statisticians in a position to contribute meaningfully to evolving research programs.

The recommendations presented here advocate for significant changes to our doctoral programs. We urge faculty and departments to use these recommendations as a catalyst for innovation. We hope that faculty will consider them in the context of their local strengths and reimagine both the PhD training and research reward system at their universities.

Appendix: Workshop Participants

Genevera Allen, Rice University
Joe Blitzstein, Harvard University
Mine Cetinkaya-Rundel, University of Edinburgh
Tammy Greasby, The Trade Desk
Dorit Hammerling, Colorado School of Mines
Jennifer Hill, New York University
Jennifer Hoeting, Colorado State University
Nicholas Horton, Amherst College
Jackie Hughes-Oliver, North Carolina State University
Snehalata Huzurbazar, West Virginia University
Rafa Irizarry, Harvard University
Eric Kolaczyk, Boston University
Liza Levina, University of Michigan
Lauren McIntyre, University of Florida
David Madigan, Columbia University
Xiao-Li Meng, Harvard University
Deborah Nolan, UC Berkeley
Rebecca Nugent, Carnegie Mellon University
Roger Peng, Johns Hopkins University
Michael Rappa, North Carolina State University
Victoria Stodden, University of Illinois at Urbana-Champaign
Duncan Temple Lang, UC Davis
Marina Vannucci, Rice University
Tian Zheng, Columbia University