Syllabus/Orientation to the Class, 36-757

Cosma Rohilla Shalizi

23 August 2010

36-757 is the first half of the preparation class for the advanced data analysis exam. It is intended for students pursuing the doctorate in statistics who have finished with their lower-level coursework and exams.

Class Mechanics

Regular class meetings will be approximately one week per month, namely the weeks of:
23 August
13 September
11 October
8 November
and 29 November.
(These dates are subject to change, with warning.)

After the first week, during regular meetings students will give ≈ 15 minute presentations on their work to date (and especially on progress since the last meeting), followed by discussion of the presentation and work by the rest of the class.

Other weeks I will be available in my office (Baker Hall 229C) during class times. I expect to hear from you, *briefly*, every week about your progress. If I do not hear from you, I will harass you.

Office Hours If you need to see me at other times, or privately, please send me e-mail to make an appointment.

Grading Grading will be based on (1) your presentations, (2) your participation when others present, and (3) a written end-of-semester progress report (≥ 5 pp.).

Orientation to the Class

To understand ADA, it helps to realize some

The Ph.D. is an apprenticeship in scholarship. The skills of a successful scholar are not, paradoxically, pieces of knowledge which you can acquire just by reading, but habits which can only be learned by practice. In an apprenticeship, the learner works under the supervision of experts, watching how the experts perform, and having their own acts guided and corrected by the experts. This sort of learning is largely a trial-and-error process, but the experts have already made, and learned from, many mistakes, so they can guide the apprentice around and through those errors, making the learning process faster.[1] As the apprenticeship progresses, the teachers provide less and less control, and the apprentice is more and more autonomous, until the process culminates in a project which shows the apprentice has *independent* mastery of all the skills: this is, or should be, your dissertation.

The goal of research, broadly speaking, is to improve humanity's common understanding of the most important aspects of the world. This is a collaborative process: you build off the work others have done, you work with other specialists, you position your own contributions so they help others. This means that communication is vital: work that sits in a desk (or on an electronic desktop), however polished, simply does not count, because it is not available to the scientific community. Research is also an entrepreneurial process: you have to develop your own line of research, which is distinct from, but complementary to, everyone else's. Pursuing research as a profession means entering a "reputation economy", where your access to jobs and resources depends on your reputation among your scientific peers, which in turn depends on their judgment of your contributions to the field.

What all this leads up to is that one of the habits you acquire in your apprenticeship is *problem formulation*, which is distinct from *problem solving.* ADA may be your first serious exposure to problem formulation, which has as an important component *re*-formulating the problem in light of experience.

This is how ADA fits into the apprenticeship process. You had a taste of research already in Statistical Practice, but you were pretty much handed a well-defined problem, albeit one with real data, and just had to solve it. To switch metaphors, this was like riding a tricycle, which is intrinsically stable and almost impossible to hurt yourself on. ADA is like a bicycle with training wheels: it's a little more stable than the real thing, but just a little.[2] In particular, one of the crucial skills of research is *problem formulation*, as opposed to *problem solving*; ADA is, most likely, your first real experience of this.

In fact, the department is hoping to see you acquire a number of abilities over the course of this year:

- Finding a research project and formulating a clear statistical problem
- Figuring out how to approach the problem

[1] Ideally, in fact, the apprentice makes their own, unique, novel and interesting, mistakes.

[2] To continue the metaphor, during the dissertation the training wheels come off, but your parents are still watching; a post-doc is a completely unsupervised bicycle; and on the tenure track you get a motorcycle with an engine, and can do real harm to yourself *and* others.

- Understanding the data in depth
- Finding and using appropriate methods
- Find and absorbing relevant knowledge from the literature
- Persisting with the problem enough to solve it
- *Re*-formulating the problem as you gain experience with it
- Working with other scientists
- Communicating your findings with other scientists

We also hope that ADA will help you get a sense of what you would *like* to do your dissertation on.

Experience has shown us that the more time you put into the ADA project, the better, and that there is basically nothing which matters more than this. Therefore, you need to find a project and begin working with the data *as soon as possible*. I need to approve the project, and can help if you are stuck for one, but the primary responsibility for getting a project and getting started with it is yours.

Your main resource for understanding the data and the scientific problem that statistics is supposed to help solve is the outside investigator. Likewise, the main resource for understanding and using appropriate methods is your statistics supervisor. (In both cases, you are supposed to be capable of reading and running down references on your own, however.)

Part of your job is to take your investigator's scientific problem, and translate it into a *statistical* problem. They may already have some idea about how to do this; you should respect this, but not feel bound by it, since you know more about statistics than they do. You will almost certainly discover that the initial formulation of the statistical problem is not quite right; you need to work with them to refine it as you go along.

It can hardly be emphasized too much that it is not enough to come up with a problem and solve it. Your findings must be communicated to the rest of the relevant scientific community in such a way that they (1) understand and (2) care. The two traditional modes of communication are the oral presentation (a seminar or conference talk) and the written paper. Both of these are skilled performances, where the necessary habits are acquired through practice. We will drill repeatedly on the oral presentation, and next semester will add in more drilling on the written paper.

My role in all of this I have four major functions in ADA.

1. Make sure you get a project started fast, and that the project stays on track.
2. Give an outside view on how the project is going.

3. Offer resources on statistical issues, scientific communication, the research process, and general kvetching, over and above what your statistics supervisor is doing.

4. Finally, and in extreme cases, I do rescue work when the project turns into a disaster. This is hardly ever necessary, and I hope it won't be this semester, but if everything looks like it's collapsing around you, let me know and I'll pull you out of it.

Some reading, for your copious free time

Robert P. Abelson, *Statistics as Principled Argument* Many of the examples are from psychology, but the ideas are more general. (Lawrence Erlbaum Associates, 1995)

Philip Agre, "Networking on the Network" (`http://vlsicad.ucsd.edu/Research/Advice/network.html`) This is partly about making contacts through electronic media, but much more generally it's about becoming a member of a learned profession. Agre has a number of other valuable how-to essays on his homepage, including "How to Be a Leader in Your Field" and "Hosting a Speaker".

Wayne C. Booth, Gregory Colomb and Joseph M. Williams, *The Craft of Research* A nice guide to defining problems, gathering information, and writing up your findings in a comprehensible and convincing way. (3rd edition, University of Chicago Press, 2008)

Peter B. Medawar, *Pluto's Republic.* See especially the essays "The Art of the Soluble" and "Is the Scientific Paper a Fraud?" Medawar's *Advice to a Young Scientist* is also very good. (Oxford University Press, 1982)

Max Weber, "Science as a Vocation" Famous lecture from 1918, reprinted in many different collections, and available online in many places.

Joseph M. Williams, *Style: Towards Clarity and Grace* This is the best book I have seen on practical writing advice, backed up by actual research. (University of Chicago Press, 1990)

John Ziman, *Real Science: What It Is, and What It Means* One of the best books I have read about how science actually works, and why it works that way. Similarly recommendable are Philip Kitcher's *The Advancement of Science* and Stephen Toulmin's *Human Understanding.* (Cambridge University Press, 2000)