STAT 8325: Topics in Advanced Statistics:
Fall 2012
1025 SSW Bldng.
Mondays 1:10PM to 2:45PM

Instructor: Victoria Stodden
Office: 1101 SSW Bldng.
Phone: 212.851.2138
email: vcs2115@columbia.edu ; victoria@stodden.net
Instructor Office Hours: 12:25PM to 1:25PM, Wednesdays.

The aim of this course is to both verify and extend published computational statistical results, with the goal for students to produce publishable findings. The topics will include statistical issues from the papers covered, tools to facilitate reproducible computational science, such as version control, data structuring, and scripting, as well as topics in legal theory for research sharing. Students will be expected to present their work in class.

Readings and other supplemental material will be posted in courseworks.

Grading: 20% In-class participation; 30% Presentations; 50% Final paper

Policy: See http://www.columbia.edu/cu/gsas/rules/chapter-9/pages/honesty/.

List of Topics Covered: (subject to change)

Week 1: What is the "reproducibility movement" in statistical research? Who cares? Case studies.

Week 2: Reproducibility and the scientific method / Barriers to reproducible research

Week 3: Preliminary topic presentations

Week 4: Legal background

Week 5: Efforts in reproducibility and the policy landscape

Week 6: Technical aids 1: version control

Week 7: Technical aids 2: provenance and workflow tracking

Week 8: Mid-semester update presentations / first draft due

Week 9: Technical aids 3: scripting, web tools

Week 10: Statistical problems solving

Week 11: Statistical and technical problem solving / drafts due to other students for comments

Week 12, 13: Student presentations

Week 14: Student presentations / Replication of work by students / Crumple zone

Final Project:

The main focus of the course is the final project. Students will not be expected to incorporate every topic covered in class, rather use what is useful in producing a reproducible statistical analysis. The final project will replicate and extend previously published results, and deliver the work in such a way that others may use the code and data to regenerate the work done by the student for the class. The final project will result in a paper to be turned in for grading. If the student desires, the final project can be submitted to a journal.

At all steps help will be available to the student, on both statistical issues and technical issues of replication, and on final publication. Feedback will be given on presentations.

Tips on final project (adapted in part from Gary King's "Publication, Publication"):

1. Your paper should address a substantive problem in your field of interest and contain one or a few clear points; one point with several supporting points is better than a lot of unrelated points. Your point should unambiguously answer the question: Whose mind are you going to change about what? If that question isn't answered, then you're not making a contribution and there's little reason for he paper to be published.

2. Begin by locating an article in your field, acquiring the data used in the article, and replicating the specific numerical results in the tables and0or figures in that analysis. This article should have been published in a peer-reviewed scholarly journal, preferably within the last 4–5 years, the more recent and prominent the better. The better the article, journal, and author you choose, and the more often the article has been cited, the more likely your paper will be publishable.

Please beware: replicating an article, even if you secure access to the original data, is normally a highly uncertain and difficult process. Analyses that look neat and clean in published articles often prove to be far from that in reality. Most students find that prominent articles by leading scholars in the field contain errors, confusions, lack of essential information about how the analysis was conducted, and other problems. Some of these issues do not matter to substantive conclusions, and some do, but all make

replication more difficult. As such, completing the replication will likely be more troublesome and time consuming than you anticipate ~even after you adjust for the information in this sentence!! After you have done everything you can do on your own, you may need to contact the author of the article ~please do so respectfully and diplomatically! The remarkable difficulties students have in replicating published articles teaches more about the state of the literature, and conveys more about the sometimes shaky foundations of academic knowledge, than reading all the published literature one person could possibly consume on his or her own.

3. Please bring me a copy of the article you choose and ask for my views before proceeding. This will generate advice on what is unlikely to work, and might be useful for other reasons, but to be clear it is no guarantee that you will be able to replicate the work chosen and successfully complete the assignment. Your assignment is to pick an article according to the criteria above and to replicate it. The choice of the article is part of the assignment and so, just as happens to faculty researchers, you may need to change your choice of topic along the way depending on what you find or difficulties in replication and do it all again. (If you change articles, please bring the new article to me as well.)

4. If you decide that the conclusions of the original article are incorrect, then show why you think that but also what led the authors of the original article to think otherwise. You should never discuss it in the paper—directly or indirectly—but you should assume, unless you have overwhelming evidence to the contrary and maybe even then, that the authors were well-intentioned, smart, honest, and hard-working. Your article is about the author's findings, not about the author.

5. Clarify with precision the extent to which you were able to replicate the author's results. If you can't replicate the author's results even with the help of the author that is important information that needs to be on the public record, but it also means you can't build on this work to make further progress. And if you can't find out what the problem is, it might mean that you do not have a publishable paper and so might need to start with a different article. So try hard, and you may have to try very hard, to replicate.

6. Unlike almost all previous papers you may have written, do not allocate space in your paper in proportion to how much work you put in accomplishing each task. The point of this paper is to make your scholarly point, not to show how smart you are. This paper should not be about you or a report of what you did; it should be about what you contribute to our collective knowledge about the world. For example, a large fraction of your effort will probably go into replicating a prior result ~and thus getting up to the cutting edge of the field!, but only in rare cases will that take more than a page or two of your paper. Space in your paper should be allocated in proportion to how much of a contribution it makes to changing the minds of someone in the literature about something important.

7. After replicating the article, follow the logic of King, Tomz, and Wittenberg (2000) and try to improve the presentation of the original results. See whether you can find

useful, additional, or even contradictory information not discussed in the article without changing any assumptions in the original paper. If you are able to do this, then you need not defend anything other than your method of presentation, which would put you on very strong grounds in your claim for journal space.

8. Next, you should run some controlled methodological experiments designed to advance the state of knowledge about the substantive project. That is, make one improvement, or the smallest number of improvements possible to produce new results, and show the results so that we can attribute specific changes in substantive conclusions to particular methodological changes. (Improvements can include changing the way the author dealt with missing data, selection bias, omitted variable bias, the model specification, differential item functioning, the functional form, etc., adding control variables or better measures, extending the time series and conducting out-of-sample tests, applying a better statistical model, etc.) If you are able to produce an interesting substantive result that is different from the original article, with only one completely justifiable methodological change, then you only need to defend this change fully and carefully.

9. If you are able to improve or change the author's results in some important way with the minimal change necessary (and that is maximally justifiable), write that up separately. Then, in a separate section, go ahead and make all the changes you think are desirable and see what difference that makes to your results. But make sure the minimal changes necessary to produce the new conclusions are described and justified first with results fully presented. Once you've done that, then you're home free in your quest for journal space.

Ground Rules:

1. Papers should be no longer than about 20 pages (double-spaced, one-inch margins, 12pt, including figures, tables, and references). Think in terms of a short research note, not a full-length article. Journal space is scarce and so the longer the paper you write, the harder it will be to publish. If you can do it in 10 pages, so much the better.

2. We provide a formal way to provide you some advice along the way: In class, you will turn in a very early draft of your paper with the tables and figures in near final form but relatively little text. You'll also turn in a replication data set, just as faculty routinely do. We will then give this to another student, who will try to replicate your results (without talking with you). That student will then write a memo to you about your paper, with copy to me. In science, we compete to advance knowledge about the world, not to tear each other down. Thus, the purpose is to improve the student's work.

3. Do not ask the author of the published article whose work you are replicating for comments on your paper, and do not share it with him or her, or anyone outside of this class, until I have read it and you have revised it accordingly. This can be a sensitive topic.

4. After the paper is revised (for substance and style) to my satisfaction and yours, it will be much safer for you to go public, and going public then is essential. The procedure is, before you show it to anyone else outside this class, send a copy to the author of the work you're replicating or critiquing and respectively request comments. When you receive a response, you should revise, being as generous as possible, but only as you think is appropriate. Only at that point should you post the paper on your web site and make it fully public, which you certainly should do. If your contribution still stands, in your view, after receiving comments from a wider audience, you should then consider submitting the paper to a scholarly journal or presenting it at a conference. For information about where to submit your paper and how to do it, come by and we'll talk about it.

For instructions on the style of the paper, see the Style section of Gary King's "Publication, Publication" paper on courseworks. We will follow his description in this class.