A statistician intent on sharing research to promote better science

December 18 2013

For centuries, researchers in fields as disparate as astrophysics and political science have faced the same hurdle before they could win acceptance for their theories—their peers must replicate and verify their results.

"It's the only way we have to decide whether or not we are getting closer to the truth," says Victoria Stodden, an assistant professor of statistics. "Otherwise, how do we know if something is right?"

Yet, most papers—whether on a new solar system or the impact of adding more cops on the streets of St. Louis—are published without the [data](link) sets and computer codes used to generate the reported results.

Stodden, who arrived at Columbia in 2010 after earning a Ph.D. in statistics and a law degree at Stanford University, is at the forefront of a movement to convince journals, academics and policy makers alike to embrace a new era of [open access](link) data sharing. She has published widely on the subject, testified about it before Congress and is a primary collaborator behind ResearchCompendia.org, an online data repository that academics use to create companion websites for their papers to allow open access to [code](link) and data.

There's a lot at stake. Sharing code and data, Stodden argues, would not only dramatically speed up the pace at which academics could verify each other's work, but might produce new revelations and theories altogether.

"If we're not sharing code and data," Stodden says, "there's a lot of duplication. If I could get my hands on your data set and maybe combine it with some data I have, I can open up a whole new set of questions."

As an undergraduate at the University of Ottawa, Stodden was fascinated by policy issues and planned to pursue a doctorate in economics. But she soon realized that the questions that really fascinated her depended on the quality of the data she could get her hands on. If two states, for instance, have different welfare-to-work policies, reviewing the data would allow you to evaluate which one worked better.

So, instead of applying to grad school in economics, Stodden pursued statistics at Stanford, hoping to get the best possible "tool kit" to work with data. Fortuitously, her adviser required students to publish code and data with their academic papers, a policy that provided maximum

transparency and let readers delve more deeply into subjects.

Stodden went to law school as a path to policy research. But she quickly began to focus on the legal issues that stand in the way of academics interested in publishing data and code with their papers. She's been working to make such an approach standard practice ever since, and a number of academic institutions and government agencies are, too.

In 2011, the National Science Foundation began requiring researchers to include a "data management plan" with grant applications. More recently, the agency began requiring some applicants to describe how they plan to make software available.

The National Institute of Health has unveiled similar policies. And in February, the White House instructed federal funding agencies to develop plans to ensure public access to the results of federally sponsored research results, including data and publications.

Stodden has been working to facilitate and understand the impact of these policy changes and others like it. One of her first papers analyzed the legal barriers to sharing code and data, such as patent, copyright and intellectual property regulations. She came up with an approach that she called the "Reproducible Research Standard," which laid out practices that might help overcome the barriers. Among them: an automatic university approval process for reusing research data and software.

"It should be the case that if you make a really useful algorithm that other people can use in their research, that should accrue to your stature as a researcher," she says.

This past summer, Stodden published a paper in the journal *PLOS ONE* comparing data and code disclosure requirements at 170 academic journals and demonstrating that the norms are rapidly shifting.

Thirty journals made a data policy change between 2011 and 2012, 12 made changes in their software policies, and 36 made changes in their supplementary data policy.

A second part of Stodden's research involves analyzing rigorous statistical methods for reproducibility, and empirical modeling to understand which policies are most effective at promoting verification efforts. Among other questions Stodden plans to address are: How hard is it to get code from an author if the journal's policy is to make it available upon request? And are the data and code provided under these policies useful in reproducing the results? "If we're not sharing code and data, we're limiting avenues of inquiry," she says. "But this has to happen at the grassroots level. There has to be cooperation from researchers or it won't work."

Provided by Columbia University