9.1: Handling Data

> *"Given enough eyeballs, all bugs are shallow."*
>
> *—Eric S. Raymond*

We've talked about the common mistakes made by scientists, and how the best way to spot them is a bit of outside scrutiny. Peer review provides some of this scrutiny, but a peer reviewer doesn't have the time to extensively re-analyze data and read code for typos – reviewers can only check that the methodology makes good sense. Sometimes they spot obvious errors, but subtle problems are usually missed.[52]

This is why many journals and professional societies require researchers to make their data available to other scientists on request. Full datasets are usually too large to print in the pages of a journal, so authors report their results and send the complete data to other scientists if they ask for a copy. Perhaps they will find an error or a pattern the original scientists missed.

Or so it goes in theory. In 2005, Jelte Wicherts and colleagues at the University of Amsterdam decided to analyze every recent article in several prominent journals of the American Psychological Association to learn about their statistical methods. They chose the APA partly because it requires authors to agree to share their data with other psychologists seeking to verify their claims.

Of the 249 studies they sought data for, they had only received data for 64 six months later. Almost three quarters of study authors never sent their data.[61]

Of course, scientists are busy people, and perhaps they simply didn't have the time to compile their datasets, produce documents describing what each variable means and how it was measured, and so on.

Wicherts and his colleagues decided they'd test this. They trawled through all the studies looking for common errors which could be spotted by reading the paper, such as inconsistent statistical results, misuse of various statistical tests, and ordinary typos. At least half of the papers had an error, usually minor, but 15% reported at least one statistically significant result which was only significant because of an error.

Next, they looked for a correlation between these errors and an unwillingness to share data. There was a clear relationship. Authors who refused to share their data were more likely to have committed an error in their paper, and their statistical evidence tended to be weaker.[60] Because most authors refused to share their data, Wicherts could not dig for deeper statistical errors, and many more may be lurking.

This is certainly not proof that authors hid their data out of fear their errors may be uncovered, or even that the authors knew about the errors at all. Correlation doesn't imply causation, but it does waggle its eyebrows suggestively and gesture furtively while mouthing "look over there."[1]

Footnotes

[1] Joke shamelessly stolen from the alternate text of http://xkcd.com/552/.