ERD

Examine.com

Research Digest

Issue 40, Vol 1 of 2 ◆ February 2018

INTERVIEW:

Andrew Gelman

Andrew Gelman is a professor of statistics and political science and director of the Applied Statistics Center at Columbia University. He has received the Outstanding Statistical Application award from the American Statistical Association, the award for best article published in the American Political Science Review, and the Council of Presidents of Statistical Societies award for outstanding contributions by a person under the age of 40. His books include Bayesian Data Analysis (with John Carlin, Hal Stern, David Dunson, Aki Vehtari, and Don Rubin), Teaching Statistics: A Bag of Tricks (with Deb Nolan), Data Analysis Using Regression and Multilevel/Hierarchical Models (with Jennifer Hill), Red State, Blue State, Rich State, Poor State: Why Americans Vote the Way They Do (with David Park, Boris Shor, and Jeronimo Cortina), and A Quantitative Tour of the Social Sciences (co-edited with Jeronimo Cortina).

Randomized controlled trials are often the gold standard for establishing causality in the health and social sciences because they reduce the impact of confounding by distributing confounders amongst groups, but they are not always practical. Some things simply cannot be studied through experimentation. In those cases, researchers often have to work with bench research and epidemiological research. How do researchers strengthen the case for causality in observational research, when there is potential for confounding, when effects are small, and influenced by random error (noise) and systematic error (bias)?
Economist Angus Deaton and philosopher Nancy Cartwright are not so sure we should think of randomized controlled trials as the gold standard. I'm inclined to share in their skepticism; see this discussion in the journal Social Science & Medicine.

You've been highly critical of some of the practices in the social sciences such as abusing statistical analyses to achieve statistical significance. Psychology, in particular, has received much criticism for its findings, which often can't be replicated. Do you think these problems are limited to psychology or are they also rampant in the medical and nutritional sciences?
Psychology is the easiest field to assess replicability because replication is typically easy and cheap: just find some more people and redo your experiment. In medical sciences, experimentation can be more expensive, and ethical issues arise when randomizing on a treatment that is already believed to be

effective. It's my guess that replication problems are no worse in psychology than in medicine and nutritional science; it's just that in psychology the problems are easy to find. It's to the credit of many researcher in all these fields that they are willing to face up to these difficulties.

What are your opinions on the practicality and usefulness of replication studies? Most replication studies are often not published (perhaps because journal editors don't think anyone will read them), but they don't seem to be very useful if there is a lack of quality control. For example, if a study lacks quality control in its measures and intended targets, it may produce positive results and replication studies may find the same effects because of things like measurement error and sampling error. Do you have any thoughts on this?

Replication studies are fine, and I think just about every study should be published, along with its raw data. That said, a lot of published studies are so hopeless that I don't see the need to replicate them: why bother? In short: if someone wants to go to the trouble of replicating a published study, go for it. Publish your data and we will all be the richer. And publish all criticisms too. Indeed, we should be more active about criticizing our own work. As it is, self-criticism can be difficult because it can get your paper rejected from the journal. Once we move to more regular post-publication review, we can start writing critical responses to our own published papers!

You consider yourself to be a Bayesian statistician. Could you explain how that differs from the conventional statistical philosophy taught and applied by most academic departments? Do you believe that differences in these philosophies actually produce notable differences even when followed properly? For example, if people practiced frequentist statistics without focusing excessively on controlling for long-term error rates and achieving statistical significance, would they still produce notable differences?

The two characteristics of Bayesian inference are: (a) representing all uncertainty and variation using probability distributions, and (b) using prior information when setting up a model. The benefits of Bayesian inference arise when (a) models are complex, so that much is lost by using simple point estimates, and (b) in settings where strong prior information is available. Lots of science has these characteristics. For more on Bayes, see my short article, "Bayes: What's it all about?", and our book, Bayesian Data Analysis.

In one of your blog posts you discussed using a within-person design, (where each person serves as their control) rather than using a between-groups design, when the data are likely to be influenced by random error (noise) and when there is likely to be a lot of variance. This could be a potentially important discussion, especially in fields like nutrition, where there tends to be a lot of variance. Could you explain your

> "It's my guess that replication problems are no worse in psychology than in medicine and nutritional science"

reasoning for believing that within-person designs may be more useful in these situations? Wouldn't you lose some of the benefits of a between-groups design such as accounting for things like regression towards the mean?

The short answer is that if you can do within-person comparisons, a lot of your variation will cancel out. Measurement error will always be a concern, but systematic variation between people is automatically accounted for in a within-person design. The risk is that if you apply multiple treatments to each person, there can be "carryover effects" so that the first treatment applied to a person alters the effect of the second treatment. There are also issues of cost and risk. But when within-person comparisons can be done, I think they're generally the way to go. For more, see here and here.

When researchers design a study, they perform power analyses where they try to figure out how many participants they need to see a particular effect when they are controlling for things like false negatives and false positives. You propose a different type of analysis known as a design analysis. Could you expand on this?

I think it's best if I point to my article on this, written with biostatistician John Carlin. Or, for a shorter treatment with examples, see section 2.1 of this recent paper.

Besides your fantastic blog, what resources would you recommend to laypersons who are trying to understand how to properly interpret statistics?

Thanks for the kind words. I'm not sure what is my advice to general laypersons, but here is my advice to journalists who are not experts in statistics or the scientific field in question, but have the chance to talk with the people who conducted a study in question. When you see a report of an interesting study, contact the authors and push them with hard questions: not just "Can you elaborate on the importance of this result?" but also "How might this result be criticized?", "What's the shakiest thing you're claiming?", "Who are the people who won't be convinced by this paper?", etc. Ask these questions in a polite way, not in any attempt to shoot the study down—but rather in the spirit of fuller understanding of the study. The best scientists will want to get things right and will be forthcoming in self-criticism. If a scientist won't or can't offer any serious objections to his or her work, then it's time to be suspicious. ◆

Andrew has done research on a wide range of topics, including: why it is rational to vote; why campaign polls are so variable when elections are so predictable; why redistricting is good for democracy; reversals of death sentences; police stops in New York City; the statistical challenges of estimating small effects; the probability that your vote will be decisive; seats and votes in Congress; social network structure; arsenic in Bangladesh; radon in your basement; toxicology; medical imaging; and methods in surveys, experimental design, statistical inference, computation, and graphics.