Ethics in statistical practice and communication: Five recommendations

Andrew Gelman[1]

23 May 2018

Abstract. Statistical analyses are often disputed, which is perhaps necessary given the fundamental difficulties of reasoning with uncertainty, and the demand for strong and certain conclusions from consumers of statistical analyses. Traditional advice on statistics and ethics focuses on professional integrity, accountability, and responsibility to collaborators and research subjects. All these are important but, along with these procedural steps, we argue that, when considering ethics, statisticians must wrestle with fundamental dilemmas regarding the analysis and communication of variation in data and uncertainty of inferences. The present article is an attempt to integrate procedural and methodological aspects of ethics in statistics. We organize our ideas around five recommendations for statistical practice and communication.

Keywords. Communication, ethics, uncertainty, variation

"I want to know if it's meant anything," Forlesen said. "If what I suffered—if it's been worth it."
"No," the little man said. "Yes. No. Yes. Yes. No. Yes. Yes. Maybe."
—Gene Wolfe (1974).

Ethics and statistics: Going beyond the basic principles of honesty and integrity

Statistics and ethics are intertwined, at least in the negative sense, given the famous saying about lies, damn lies, and statistics, and the well-known book, *How to Lie with*

[1] Department of Statistics and Department of Political Science, Columbia University, New York. We thank two reviewers for helpful comments and the U.S. Office of Naval Research grant N00014-15-1-2541 and Defense Advanced Research Projects Agency grant D17AC00001 for partial support of this work.

Statistics (which, ironically, was written by a journalist with little knowledge of statistics who later accepted thousands of dollars from cigarette companies and told a congressional hearing in 1965 that inferences in the Surgeon General's report on the dangers of smoking were fallacious).

The principle that one should present data as honestly as possible is a fine starting point but does not capture the dynamic nature of science communication: audiences interpret the statistics (and the paragraphs) they read in the context of their understanding of the world and their expectations of the author, who in turn has various goals of exposition and persuasion—and all of this is happening within a competitive publishing environment, in which authors of scientific papers and policy reports have incentives to make dramatic claims.

The result is that scientists aren't communicating their work to one another, let alone to general audiences, in terms appropriately geared to enlarging knowledge—they aren't doing science properly—and this is one of the recurring threats to the quality of our science communication environment. So the "science of science communication" has to be concerned with the quality of the norms and practices that regulate how scientists communicate to one another or simply do what they do. The field of statistics is a useful focus here as it is central to many disputes about science and policy, and ethics in statistics have received some recent attention (American Statistical Association, 2016).

Consider this fundamental paradox: *statistics is the science of uncertainty and variation, but data-based claims in the scientific literature tend to be stated deterministically* ("We have discovered . . . the effect of X on Y is . . . hypothesis H is rejected . . ."). Is statistical communication about exploration and discovery of the unexpected, or is it about making a persuasive, data-based case to back up an argument?

The answer to this question is, necessarily, each, at different times, and sometimes both at the same time. Just as you write in part in order to figure out what you trying to say—you are in the business of revealing your thoughts to yourself, even persuading yourself,

on the way to persuade others, so you do statistics not just to learn from data, but also to learn what you can learn from data, and to decide how to gather future data to help resolve key uncertainties.

Traditional advice on statistics and ethics focuses on professional integrity, accountability, and responsibility to collaborators and research subjects. All these are important but, along with these procedural steps, we argue that, when considering ethics, statisticians must wrestle with fundamental dilemmas regarding the analysis and communication of uncertainty and variation.

Five recommendations for statistical practice and communication

Gelman (2011) characterizes ethics in statistics as follows: "An ethics problem arises when you are considering an action that (a) benefits you or some cause you support, (b) hurts or reduces benefits to others, and (c) violates some rule. . . . Although any ethics violation can be framed to be ambiguous, this does not, and should not, negate the importance of ethics. Statisticians should be able to appreciate the necessity of decision making under uncertainty and ambiguity."

This is a general statement. The next question to ask is: "What are the ethical concerns specifically related to statistics, to data, to evidence-based communication?" In the present paper we consider five recommendations: (1) Open data and open methods; (2) Being clear about the information that goes into statistical procedures; (3) Creating a culture of respect for data; (4) Publishing criticisms; and (5) Respecting the limitations of statistics. These recommendations are not intended to be exhaustive, nor are do we presume to support them with rigorous quantitative analysis. Rather, they represent recommended directions for progress based on recent experiences.

1. Open data and open methods. Statistical conclusions are data-based and they can also be, notoriously, dependent on the methods used to analyze the data. An extreme example is the influential paper of Reinhart and Rogoff (2009) on the effects of deficit spending,

which was used to justify budget-cutting policies. In a notorious mistake, the authors had misaligned columns in an Excel spreadsheet so their results did not actually follow from their data. This highly consequential error was not detected until years after the article was published and later researchers went to the trouble of replicating the analysis (Herndon, Ash, and Pollin, 2014), illustrating how important it is to make data and data-analysis scripts available to others, providing more "eyes on the street," as it were.

There has been much (appropriate) concern about arbitrary decisions in data analysis, "researcher degrees of freedom" in the words of Simmons, Nelson, and Simonsohn (2011) that calls into question the many (most) published *p*-values in psychology, economics, medicine, etc.—but we should also be aware of researcher freedom in data coding, exclusion, and cleaning more generally. Open data and open methods implies a replicable "paper trail" leading from raw data, through processing and statistical analysis, to published conclusions.

Statistics professors promote quantitative measurement, controlled experimentation, careful adjustment in observational studies, and data-based decision making. But in teaching their own classes, they (we) tend to make decisions and inferences based on non-quantitative recall of uncontrolled interventions, just trying things out and seeing what we see, behavior that we would consider laughable and borderline unethical in social or health research.

Are we being unethical in not following our own advice, or in promulgating to others advice we do not ourselves follow? Not necessarily: it is a perfectly reasonable position to say that controlled experiments are appropriate in certain medical trials and public interventions but not in all aspects of our work. But in that case we should do a better job of understanding and explaining the conditions under which we do not believe controlled experimentation and statistical analysis to be appropriate.

2. Being clear about the information that goes into statistical procedures. Bayesian inference combines data with prior information, and some Bayesians would argue that it

is an ethical requirement to use such methods as otherwise information is being "left on the table" when making decisions. Others take the opposite position and argue that "there are situations where it is very clear that, whatever a scientist or statistician might do privately in looking at data, when they present their information to the public or government department or whatever, they should absolutely not use prior information, because the prior opinions on some of these prickly issues of public policy can often be highly contentious with different people with strong and very conflicting views" (Cox and Mayo, 2011).

Both these extreme views have problems, as discussed by Gelman (2012). The recommendation to always use prior information runs into difficulty when this prior information is disputed; in such settings it makes sense to present unvarnished results. But in many high-stakes settings it is impossible to make any use of data without a model that makes extensive use of prior information. Consider, for example, the reconstruction of historical climate from tree rings, which can only be done in the context of statistical models which themselves might be contentious. The models relating climate to tree rings are not quite physical models for tree growth and not quite curve fitting, but rather something in between: they are statistical models that are informed by physical considerations. As such, they rely of prior information, even if not in the conventional sense of prior distributions as discussed by Cox and Mayo in the quote above. Our point here is not that Bayes is better (or worse) but that, under any inferential philosophy, we should be able to identify what information is being used in methods.

In some settings, prior information is as strong as, or stronger, than the data from any given study. For example, Gertler et al. (2013) reported on an early-childhood intervention performed in an experiment in Jamaica that increased adult earnings (when the children grew up) by an estimated 42%, and the result was statistically significant, thus the data were consistent with effects between roughly 0% and an 80% increase in earnings. But prior knowledge of previous early-childhood interventions suggests that effects of 80%, or even 40%, are implausible. It is fine to present the results *from this particular study* without reference to any prior information, or to include such

information in a non-Bayesian way, as is done in power calculations, but we do not think it appropriate to offer policy recommendations from this one estimate in isolation. To return to our general recommendation, it is important to understand the implications of the method being used.

3. Creating a culture of respect for data. Opacity in data collection, analysis, and reporting is abetted and indeed encouraged by aspects of scholarly research culture: When it comes to data collection, institutional review boards can make it difficult to share one's own data or access others', and when it comes to reporting results, journals favor brevity over completeness. Even in this online age, top journals often aspire to the *Science/Nature* format of three-page articles. Details can appear in online appendixes but these usually focus not on the specifics of a study but rather on supplementary analyses to buttress the main paper's claims. Published articles typically focus on building a convincing case and giving a sense of certainty, not about making available all the information that would allow outsiders to check and replicate the research.

That said, honesty and transparency are not enough (Gelman, 2017). All the preregistration in the world won't save your study if the data are too remote from the questions of interest. Remoteness can come from mismatch of sample to population, lack of comparability between treatment and control groups, lack of realism of experimental conditions, or, most simply, biased and noisy measurements. For a study to be ethical it should be informative, which implies serious attention to measurement, design, and data collection. Without good and relevant measurements, protocols such as preregistration, random sampling, and random treatment assignment are empty shells.

As an institutional solution, top journals can publish papers that contain interesting or important data, without the requirement of innovative data analyses or conclusions. In addition to facilitating data availability, this step could also reduce the pressure on researchers to add unnecessary elaborations to their analyses or to hype their conclusions as a way of attaining publication. Public data are important in many fields of study (consider, for example, the U.S. Census, the Panel Study of Income Dynamics, the

National Election Study, and various weather and climate databases), so this proposal can be viewed as extending the culture of respect for data and applying it to individual studies.

4. Publication of criticisms. You don't need to be a philosopher to feel that it is unethical not to admit error, or to avoid facing evidence that you have erred. Statistical errors can be technical and hard to notice (and are sometimes even buried within conventional practices such as taking a statistically significant comparison as strong evidence in favor of a favored hypothesis). Institutions as well as individuals can be averse to admitting error, indeed scholarly publishing is often set up to suppress criticism. Journals are notoriously loath to retract articles or publish letters of correction.

For example, a couple years ago I was pointed to an article in the *American Sociological Review* that suffered from a serious case of selection bias. The article reported that students who paid for their own college educations performed better than those who were funded by their parents But the statistical analysis used to make this claim did not adjust for the fact that self-funded students who were not doing well would be more likely to drop out. Unfortunately it was not possible to correct this mistake in the journal where it appeared, as the editors judged the correction to be not worthy of publication.

A system of marginalizing criticism creates an incentive for authors to promote dramatic claims, with an upside when published in top journals and little downside if errors are later found. I'm sure that the author and editors in this particular case simply made an honest mistake in not catching the selection bias. Nonetheless, the system as a whole gives no clear incentives for the parties involved to be more careful.

Post-publication review outlets such as PubPeer and blogs may be changing this equation. This illustrates the dynamic relation between institutions and ethics that is a theme of the present article.

Researchers can do even better by criticizing their own work, as done by Nosek, Spies, and Motyl (2012), who performed an experiment to study "embodiment of political extremism." They continued: "Participants from the political left, right and center (N = 1,979) completed a perceptual judgment task in which words were presented in different shades of gray. . . . The results were stunning. Moderates perceived the shades of gray more accurately than extremists on the left and right ($p = .01$). Our conclusion: political extremists perceive the world in black-and-white, figuratively and literally."

Before publishing this result, though, the authors decided to collect new data and replicate their study: "We ran 1,300 participants, giving us .995 power to detect an effect of the original effect size at $\alpha = .05$."

And, then, the punch line: "The effect vanished ($p = .59$)."

How did this happen? The original statistically significant result was obtained via a data-dependent analysis procedure. The researchers compared accuracy of perception, but there are many other outcomes they could have looked at, for example there could have been a correlation with average perceived shade, or an interaction with age, sex, or various other logical moderators, or an effect just for Democrats or just for Republicans, and so forth. The replication, with its prechosen comparison, was not subject to this selection effect.

Motyl et al. discuss how to reform the scientific publication system to provide incentives for this self-critical behavior. But in the meantime you can do it yourself, just as they did!

More generally, you can make self-criticism part of your general practice by enabling others' criticisms of your work, via open data, clarity in assumptions, and the other steps listed above. Joining with others to criticize your own practices should strengthen your work. Our recommendations on facilitating criticisms are consistent with recent ethical

guidelines which call for prompt correction of errors and appropriate dissemination of the correction (American Statistical Association, 2016).

5. Respecting the limitations of statistics. Many fields of empirical research have become notorious for claims published in serious journals which make little sense (for example, the claim that people react differently to hurricanes with male and female names (see Frijters, 2014, and Malter, 2014) or the claim that women have dramatically different political preferences in different times of the month, or the claim that the subliminal image of a smiley face has large effects on attitudes on immigration policy (Gelman, 2015a)) but which are easily understood as the inevitable product of explicit or implicit searches for statistical significance with flexible hypotheses that are rich in researcher degrees of freedom (Simmons, Nelson, and Simonsohn, 2011). Unsurprisingly (given this statistical perspective), several high-profile research papers in social psychology have failed to replicate, for example the well-publicized claim in "embodied cognition" that college students walk more slowly after being subtly primed by being exposed to elderly-related words (Doyen et al., 2012).

Just to be clear: the above claims seem to many people (including the present author) to be *silly* but they are certainly not *impossible*, at least in a qualitative sense. For example, the literature on public opinion makes it highly implausible that women were experiencing during their monthly cycles a 20% swing in probability of supporting Barack Obama for president, as claimed by Durante, Arsena, and Griskevicius (2013). It is, however, possible that there is a tiny effect, essentially undetectable in the study in question given the precision of measurement of the relevant variables (see Gelman, 2015b).

So the error in that paper (and in the hurricanes paper and the others mentioned above) is that the data do not provide strong evidence for the authors' claims. These papers, and the system by which they are published and publicized, represent *a failure in science communication* in that they place an impossible burden on statistical data collection and analysis.

Moving away from systematic overconfidence

In statistics, we use mathematical analysis and stochastic simulation to evaluate the properties of proposed designs and data analyses. Recommendations for ethics are qualitative and cannot be evaluated in such formal ways. Nonetheless we believe there is value in the recommendations made in this paper. In particular we emphasize the links between ethical principles and the general statistical concepts of variation and uncertainty.

So far, this is just a story of statistical confusion perhaps abetted by incentives toward reporting dramatic claims on weak evidence. The ethics comes in if we think of this entire journal publication system as a sort of machine for laundering uncertainty: researchers start with junk data (for example, poorly-thought-out experiments on college students or surveys of online Mechanical Turk participants) and then work with the data, straining out the null results and reporting what is statistically significant, in a process analogous to the notorious mortgage lenders of the mid-2000s, who created high-value "tranches" out of subprime loans. The loan crisis precipitated an economic recession, and I doubt the replication crisis will trigger such a crash in science. But I see a crucial similarity in that technical methods (structured finance for the mortgages; statistical significance for the scientific research) were being used to create value out of thin air.

In the article, "The AAA Tranche of Subprime Science," Loken and Gelman (2014) concluded:

> When we as statisticians see researchers making strong conclusions based on analyses affected by selection bias, multiple comparisons, and other well-known threats to statistical validity, our first inclination might be to throw up our hands and feel we have not been good teachers, that we have not done a good enough job conveying our key principles to the scientific community.

> But maybe we should consider another, less comforting possibility, which is that our fundamental values have been conveyed all too well and the message we have been sending—all too successfully—is that statistics is a form of modern alchemy, transforming the uncertainty and variation of the laboratory and field measurements into clean scientific conclusions that can be taken as truth. . . .
>
> We have to make personal and political decisions about health care, the environment, and economics—to name only a few areas—in the face of uncertainty and variation. It's exactly because we have a tendency to think more categorically about things as being true or false, there or not there, that we need statistics. Quantitative research is our central tool for understanding variance and uncertainty and should not be used as a way to overstate confidence.

Ethics is, in this way, central to statistics and public policy. We use statistics to measure uncertainty and variation, but all too often we sell our methods as a sort of alchemy that will transform these into certainty. The first step to not fooling others is to not fool ourselves.

References

American Statistical Association (2016). Ethical guidelines for statistical practice. http://www.amstat.org/asa/files/pdfs/EthicalGuidelines.pdf

Cox, D. R., and Mayo, D. (2011). Statistical scientist meets a philosopher of science: A conversation. *Rationality, Markets and Morals* 2, 103-114.

Doyen, S., Klein, O., Pichon, C. L., and Cleeremans, A. (2012). Behavioral Priming: It's all in the mind, but whose mind? *PLoS ONE* 7, e29081.

Durante, K. M., Arsena, A. R., and Griskevicius, V. (2013). The fluctuating female vote: Politics, religion, and the ovulatory cycle. *Psychological Science* 24, 1007-1016.

Frijters, P. (2014). How to lie with statistics: The case of female hurricanes. http://clubtroppo.com.au/2014/06/11/how-to-lie-with-statistics-the-case-of-female-hurricanes/

Gelman, A. (2011). Ethics and statistics: Open data and open methods. *Chance* 24 (4), 51-53.

Gelman, A. (2012). Ethics and the statistical use of prior information. *Chance* 25 (4), 52-54.

Gelman, A. (2013). It's too hard to publish criticisms and obtain data for replication. *Chance* 26 (3), 49-52.

Gelman, A. (2015a). Disagreements about the strength of evidence. *Chance* 28, 55-59.

Gelman, A. (2015b). The connection between varying treatment effects and the crisis of unreplicable research: A Bayesian perspective. *Journal of Management* 41, 632-643.

Gelman, A. (2017). Honesty and transparency are not enough. *Chance* 30 (1), 37-39.

Gertler, P., Heckman, J., Pinto, R., Zanolini, A., Vermeerch, C., Walker, S., Chang, S. M., and Grantham-McGregor, S. (2013). Labor market returns to early childhood stimulation: A 20-year followup to an experimental intervention in Jamaica. IRLE Working Paper No. 142-13. http://irle.berkeley.edu/workingpapers/142-13.pdf

Herndon, T., Ash, M., and Pollin, R. (2014). Does high public debt consistently stifle economic growth? A critique of Reinhart and Rogoff. *Cambridge Journal of Economics* 38, 257-279.

Malter, D. (2014). Female hurricanes are not deadlier than male hurricanes 111, E3496.

Loken, E., and Gelman, A. (2014). The AAA tranche of subprime science. *Chance* 27 (1), 51-56.

Nosek, B. A., Spies, J. R., and Motyl, M. (2012). Scientific utopia II. Restructuring incentives and practices to promote truth over publishability. *Perspectives on Psychological Science* 7, 615-631.

Reinhart, C. M., and Rogoff, K. S. (2010). Growth in a time of debt. *American Economic Review* 100, 573-578.

Simmons, J., Nelson, L., and Simonsohn, U. (2011). False-positive psychology: Undisclosed flexibility in data collection and analysis allow presenting anything as significant. *Psychological Science* 22, 1359-1366.

Wolfe, G. (1974). Forlesen. In *Orbit 14*, ed. D. Knight. New York: Harper & Row.