Ethics in statistical practice and communication
Five recommendations

Ethics in statistics is about more than good practice. It extends to the communication of uncertainty and variation. **Andrew Gelman** presents five recommendations for dealing with fundamental dilemmas

"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."[1]

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 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 are not communicating their work to one another, let alone to general audiences, in terms appropriately geared to enlarging knowledge

Andrew Gelman is a professor of statistics and political science and director of the Applied Statistics Center at Columbia University.

– they are not doing science properly – and this is one of the recurring threats to the quality of our science communication environment.

Consider this paradox: statistics is the science of uncertainty and variation, but data-based claims in the scientific literature tend to be stated deterministically (e.g. "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 are trying to say, 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 when considering ethics, statisticians must also wrestle with fundamental dilemmas regarding the analysis and communication of uncertainty and variation.

In what follows, I make five recommendations for dealing with these dilemmas. These are not intended to be exhaustive, nor do I 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 analyse the data. An extreme example is the influential paper of Reinhart and Rogoff on the effects of deficit spending, which was used to justify budget-cutting policies.[2] In an infamous 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,[3] illustrating how important it is to make data and data-analysis scripts available to others – providing more "eyes on the street", as it were.

Statistics is the science of uncertainty, but data-based claims tend to be stated deterministically

There has been much (appropriate) concern about arbitrary decisions in data analysis – "researcher degrees of freedom"[4] – 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 imply 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 – behaviour 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 reasonable position to say that controlled experiments are appropriate in certain medical trials and public interventions, but not in all aspects of our work. However, 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. Be 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".[5]

Both these extreme views have problems.[6] 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 on prior information, even if not in the conventional sense of prior distributions as discussed by Cox and Mayo in the quote above. The 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.* 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.[7] 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 it is not appropriate to offer policy recommendations from this one estimate in isolation. Rather, it is important to understand the implications of the method being used.

3. Create 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 favour 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 appendices, 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 on making available all the information that would allow outsiders to check and replicate the research.

That said, honesty and transparency are not enough.[8] All the preregistration in the world will not 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 US 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 do not 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 favour of a favoured 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.

All too often we sell our methods as a sort of alchemy that will transform uncertainty into certainty

For example, a couple of 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 education 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 not to be worthy of publication.

A system of marginalising 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 am 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 criticising their own work, as done by Nosek, Spies, and Motyl, who performed an experiment to study "embodiment of political extremism".[9] Their initial finding: "Participants from the political left, right and center ($N = 1979$) 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 1300 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 pre-chosen comparison, was not subject to this selection effect.

Nosek *et al.* discuss how to reform the scientific publication system to provide incentives for this self-critical behaviour. 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 criticise your own practices should strengthen your work. These recommendations on facilitating criticisms are consistent with the American Statistical Association's recent ethical guidelines, which call for prompt correction of errors and appropriate dissemination of the correction (bit.ly/2ML136N).

5. Respect 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,[10] or the claim that women have dramatically different political preferences at different times of the

month, or the claim that the subliminal image of a smiley face has large effects on attitudes on immigration policy[11]) 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.[4] Unsurprisingly (given this statistical perspective), several high-profile research papers in social psychology have failed to replicate: for example, the well-publicised claim in "embodied cognition" that college students walk more slowly after being subtly primed by being exposed to elderly-related words.[12]

Just to be clear: the above claims seem to many people (including the present author) to be silly, but they are 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.[13] 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.[14]

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 publicised, 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 there is value in the recommendations made in this paper and in emphasising 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 towards 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 mortgages; statistical significance for scientific research) were being used to create value out of thin air.

In their article, "The AAA tranche of subprime science", Loken and Gelman 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.[15]

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. ■

Acknowledgements

The author wishes to thank two reviewers for helpful comments and the US Office of Naval Research grant N00014-15-1-2541 and Defense Advanced Research Projects Agency grant D17AC00001 for partial support of this work.

References

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

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

3. 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.

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

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

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

7. 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. Institute for Research on Labor and Employment, Berkeley, CA. http://irle.berkeley.edu/workingpapers/142-13.pdf

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

9. 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.

10. Malter, D. (2014) Female hurricanes are not deadlier than male hurricanes. *Proceedings of the National Academy of Sciences of the USA*, **111**, E3496.

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

12. 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.

13. 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.

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

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