The AAA Tranche of Subprime Science

Andrew Gelman and Eric Loken

One of our ongoing themes when discussing scientific ethics is the central role of statistics in recognizing and communicating uncertainty. Unfortunately, statistics—and the scientific process more generally—often seems to be used more as a way of laundering uncertainty, processing data until researchers and consumers of research can feel safe acting as if various scientific hypotheses are unquestionably true.

A recent article in the *Economist* asked whether science was self-correcting, focusing on failures to replicate and the widespread concern that many areas of the scientific literature are not as reliable as we want to believe. This wasn't the first time we've read in the financial press about an overheated domain, awash in liquidity and expanding quickly, driven by its own internal logic, where risk was understated and underappreciated, but where professionals and consumers alike continued to buy in and operate with a faith that the good times would go on forever.

We have in mind an analogy with the notorious AAA-class bonds created during the mid-2000s that led to the subprime mortgage crisis. Lower-quality mortgages—that is, mortgages with high probability of default and, thus, high uncertainty—were packaged and transformed into financial instruments that were (in retrospect, falsely) characterized as low risk. There was a tremendous interest in these securities, not just among the most unscrupulous market manipulators, but in a world where a lot of money was looking for safe investments and investors were willing to believe the ratings agencies and brokers.

Similarly, the concerns about reliability and validity of published results come after years of rapid expansion in the world of scientific output. In published research studies, data of varying quality are thrown together, processed, and analyzed and formed into statistically significant aggregates that are combined into research papers. Any individual data point or small cluster of data might be suspect, but, when they are combined, the law of large numbers is supposed to ensure that the larger conclusions are stable.

How Is a Research Paper Like a Mortgage?

The analogy is anything but exact, but we see two equivalents in the modern scientific process to the aggregation and skimming that led to tranches of mortgages being declared AAA (high-quality) bonds. The first step is statistical significance. Out of the primordial soup of all possible data analyses, the statistically significant comparisons float to the top. They represent the high-certainty statements selected out of the many less-reliable claims. The second step is publication in a scientific journal, ideally a high-prestige outlet such as *Psychological Science, American Journal of Sociology,* or the *Proceedings of the National Academy of Sciences*—but, if not at a top journal, any outlet will do. The convention is to treat published claims as true unless demonstrated otherwise. The two-step process—first the achievement of statistical significance, then publication—corresponds with the movement of a scientific hypothesis from the hazy zone of uncertain speculation to presumed certainty.

By analogizing to the mortgage crisis, are we saying all research is over-valued? Not at all. Nor were all subprime mortgages destined to fail. The problem happened when reliable and less reliable components were bundled into a AAA tranche. The analogy might be to believe all papers published in *Nature* just because they are published in a top journal, or to believe the results of all published medical trials that have randomization in their designs.

Another way to look at this is to consider the roles played by statistical significance and peer review as a seal of approval for a scientific claim. On one hand, these hurdles to publication protect the general public from a proliferation of random claims. But once a claim has passed these tests, it is commonly considered conclusively proved. Or, at least, to be considered true until convincing evidence is presented to the contrary.

The trouble is that neither statistical significance nor peer review, as currently practiced, work quite the way they are supposed to:

- In theory, statistical significance (at the conventional 5% level) should occur at most one-twentieth of the time, if there is truly no underlying effect. In practice, though, researchers have many ways of processing their data ("researcher degrees of freedom," as discussed by Simmons, Nelson, and Simonsohn, 2011), so it is actually not difficult to obtain statistically significant comparisons in a data set, even in the presence of no true effect (or of a true effect that is close enough to be zero to be essentially undetectable given the available sample size in the study). Indeed, it is not difficult to come up with two or more statistically significant results, which, to a naïve interpreter of statistics, can seem to be overwhelming evidence against the null hypothesis. If misguided studies really had only a 1/20 chance of leading to statistical significance and only a 1/400 chance of resulting in two statistically significant findings, the threshold would have some value, as we do not think most researchers would like to undertake projects with 95% chances of failure. But, in the real world, where statistical significance can be obtained nearly all the time, it is not much of a barrier; rather, it serves as a motivation to distort studies and exaggerate effect sizes. This distortion can be done without any conscious design on the part of the researcher, merely by performing reasonable analyses that are contingent on data.

- Peer review plays a useful role in communicating differing views to the authors of scientific manuscripts and giving the opportunity to improve papers before publication, but it does not filter out the publication of mistakes (by which we mean lines of argument that are clearly in error given existing knowledge, not just claims that turn out not to be replicable in retrospect). Why is publication not an effective filter? First, there is a huge proliferation of journals. As those of us who write for scientific publication know, just about any paper is publishable if you try enough outlets. Second, journals select on novelty as much as correctness. Even (or perhaps especially) top journals can be receptive to newsworthy claims, even those that are not so strongly supported by data. And third, reviewers and publishers may be independent with respect to a specific article they review, but they are not independent in that they participate in related projects and in the broader scientific enterprise where publications are the currency for success and promotion.

Consider an analogy to the certification of food safety during a public health crisis. One might think of statistical significance and peer review as an institutional response, the equivalent of some sort of semi-public agency that would inspect and approve meat and vegetables as being safe for consumption. But suppose that food producers were free to submit sample after sample to the inspectors, and to manipulate their food samples before inspection? And suppose the inspection process was itself performed qualitatively (as with scientific peer review), with the option to resubmit food to a new inspection agency if it was turned down by the first group of inspectors? The result would be a deadly combination of unsafe food being sent out into the marketplace, but with inspection stickers that would lead naïve consumers to think the product was safe.

In recent years, applied quantitative researchers have written about many problems with the current system of *p*-values and publication (e.g., Vul et al., 2009; Wagenmakers et al., 2011; Button et al., 2013; Francis, 2013; and Humphreys, Sanchez, and Windt, 2013), and scholars in many fields have recommended replacing the current system of journals with open repositories of papers on the model of the Social Science Research Network or the ArXiv system used in physics (e.g., Kriegeskorte, 2009, and Wasserman, 2012). The point of the present article is not to present a blanket criticism of the current system, but rather to stress how it is used as a way of laundering uncertainty, creating AAA tranches of strong claims from masses of data and analyses of varying quality.

Put Your Money Where Your Mouth Is?

One proposed solution to the proliferation of claimed certainty in research papers is to see if people will back up their hypotheses with cold cash. Two famous examples so far of such challenges are the million-dollar prize offered by magician/skeptic James Randi for experimental evidence of the paranormal and the bet between economist Julian Simon and biologist Paul Ehrlich regarding the scarcity of natural resources (Sabin, 2013). More recently, it has been suggested that prediction markets for scientific hypotheses could be set up more generally (Wolfers and Zitzewitz, 2004, and Almenberg, Kittlitz, and Pfeiffer, 2009).

Would prediction markets (or something like them) help? It's hard to imagine them working out in practice. Indeed, the housing crisis was magnified by rampant speculation in derivatives that led to a multiplier effect. Allowing people to bet on the failure of other people's experiments just invites corruption, and the last thing social psychologists want to worry about is a point-shaving scandal. And there are already serious ways to bet on some areas of science. Hedge funds, for instance, can short the stock of biotech companies moving into phase II and phase III trials if they suspect earlier results were overstated and the next stages of research are thus underpowered.

More importantly, though, we believe that what many researchers in social science in particular are more likely to defend is a general *research hypothesis*, rather than the specific empirical findings.

On one hand, researchers are already betting—not just money (in the form of research funding) but also their scientific reputations—on the validity of their research. On the other hand, published claims are vague enough that all sorts of things can be considered as valid confirmations of a theory (just as it was said of Freudian psychology and Marxian economics that they can predict nothing but explain everything). And scientists who express great confidence in a given research area can get a bit more cautious when it comes to the specifics.

For example, our previous ethics column, "Is It Possible to Be an Ethicist Without Being Mean to People," considered the case of a controversial study, published in a top journal in psychology, claiming women at peak fertility were three times more likely to wear red or pink shirts, compared to women at other times during their menstrual cycles. After reading our published statistical criticism of this study in *Slate*, the researchers did not back down; instead, they gave reasons for why they believed their results (Tracy and Beall, 2013). But we do not think that they or others really believe the claimed effect of a factor of 3. For example, in an email exchange with a psychologist who criticized our criticisms, one of us repeatedly asked whether he believed women during their period of peak fertility are really three times more likely to wear red or pink shirts, and he repeatedly declined to answer this question.

What we think is happening here is that the authors of this study and their supporters separate the general scientific hypothesis (in this case, a certain sort of connection between fertility and behavior) from the specific claims made based on the data. We expect that, if forced to lay down their money, they would bet that, in a replication study, women in the specified days in their cycle would be less than three times more likely to wear red or pink, compared to women in other days of the cycle. Indeed, we would not be surprised if they would bet that the ratio would be less than two, or even less than 1.5. But we think they would still defend their hypothesis by saying, first, that all they care about is the existence of an effect and not its magnitude, and, second, that if this particular finding does not replicate, the non-replication could be explained by a sensitivity to experimental conditions.

In addition, betting cannot be applied easily to policy studies that cannot readily be replicated. For example, a recent longitudinal analysis of an early childhood intervention in Jamaica reported an effect of 42% in earnings (Gertler et al., 2013). The estimate was based on a randomized trial, but we suspect the effect size was being overestimated for the usual reason that selection on statistical significance induces a positive bias in the magnitude of any comparison, and the reported estimate represents just one possible comparison that could have been performed on these data (Gelman, 2013a). So, if the study could be redone under the same conditions, we would bet the observed difference would be less than 42%. And under new conditions (larger-scale, modern-day interventions in other countries), we would expect to see further attenuation and bet that effects would be even lower, if measured in a controlled study using pre-chosen criteria. Given the difficulty in setting up such a study, though, any such bet would be close to meaningless. Similarly, there might be no easy way of evaluating the sorts of estimates that appear from time to time in the newspapers based on large public-health studies.

That said, scientific prediction markets could be a step forward, just because it would facilitate clear predictive statements about replications. If a researcher believes in his or her theory, even while not willing to bet his or her published quantitative finding would reappear in a replication, that's fine, but it would be good to see such a statement openly made. We don't know that such bets would work well in practice—the biggest challenge would seem to be defining clearly enough the protocol for any replications—but we find it helpful to think in this framework, in that it forces us to consider, not just what is in a particular past data set, but also what might be happening in the general population.

Under the current system, scientific replication typically seems to be measured in terms of statistical significance: A result is considered confirmed if several later studies on the same topic (but with different details) are published in high-quality journals. We would like to see, in addition, a narrower standard of replication (statistically significant or otherwise) based on preregistered designs. An example of such an effort is Klein et al. (2013).

Embracing Variation and Accepting Uncertainty

We think researchers of all sorts (including statisticians, when we consider our own teaching methods; see Gelman and Loken, 2012) rely on two pre-scientific or pre-statistical ideas:

1. The idea that effects are "real" (and, implicitly, in the expected direction) or "not real." By believing this (or acting as if you believe it), you are denying the existence of *variation*. And, of course, if there really were no variation, it would be no big deal to discard data that don't fit your hypothesis.
2. The idea that a statistical analysis determines whether an effect is real. By believing this (or acting as if you believe it), you are denying the existence of *uncertainty*. And this can lead you to brush aside criticisms and think of issues such as selection bias as technicalities rather than serious concerns. In areas of science where *p*-values prevail, it is all too common to chase the asterisks that signal "statistical significance" and use that as a marker to justify more theoretical assertions. The statistical analysis is a step toward getting the AAA rating validating the general scientific idea that is for sale.

We think the problem is that often researchers do not admit uncertainty or variation; they think they've already made their discovery, and they think of various data-collection and data-analysis rules as technicalities that should not get in the way of science. After all, if you've published a paper with nine statistically significant results, it would seem like you've discovered a pattern that could only occur once in $(1/20)^9$ by chance, a probability that would seem too extreme to be seriously whittled away by minor methodological issues.

It is admirable, in some sense, for researchers to focus on the science and not get hung up on detailed technical criticisms, but the resulting attitude—the equation, "statistical significance" + "publication" = "truth"—can create problems given that social and environmental sciences are full of uncertainty and variation and that methodological issues are multiplicative (with concerns arising from choices of data to exclude, coding of variables, decisions about subdividing data, and so forth).

As with mortgage risks, all too often, the multiple tests in a statistical argument are correlated and much more sensitive to implicit assumptions than is realized. A few significant *p*-values give the illusion of a high rating to the whole set of results, and this illusion of certainty is then used to justify a discussion section devoted to the general scientific hypothesis.

Who Is Responsible for Bringing Change?

What are the ethical considerations for individuals to address? We need to remember the mortgage crisis was not just the responsibility of a few wild speculators. Instead, the good times rolled for many years, with millions of borrowers overextending their credit, many lenders willing to enable them, and many investors feeling it would be irresponsible not to take advantage of the low-risk, high-return investments on offer.

Many people felt the story was too good to be true, but it wasn't clear whose job it was to deflate the bubble. And the incentives for consumers, lenders, and politicians were heavily tilted to keeping the party going.

Everyone should think about their role in the current crisis about science. The advice below might seem like the usual list we've become accustomed to seeing over the last couple of years. These lessons, though, are not entirely easy to apply. Just as tightening credit standards had the effect of denying or delaying the pathway to home ownership for many, tightening the standards for statistical inference and publication would necessarily slow down many aspiring researchers. And just as thousands of mortgage brokers and lenders may have felt they had to go with the flow or fall behind, it's natural that many researchers might feel frustrated if held to what seems like higher standards than the prevailing norm. And it is no fun for statisticians, in their role as consultants and mentors, to dissuade researchers from being too enthusiastic about their results. After all, publications are the currency of the academic system, and, as any economic system, unilateral actions bear consequences. Here's our partial list of lessons from the housing crisis that might apply to statisticians and researchers in today's scientific climate:

1. *Get the details right.* Even when it wasn't outright deliberate fraud, the housing market ballooned in part because of shortcuts in vetting the credit worthiness of individual applicants for mortgages and in properly processing those mortgages. Documenting the details of research studies, including all the relevant decisions in the design and the data analysis, is important for being able to reproduce, evaluate, and replicate results.
2. *Make honest appraisals.* The frenzy of the housing bubble led to ever-inflated valuations that eventually became unsustainable. In the academic community, we already have grade inflation, and don't even try to suggest in a recommendation letter that a candidate doesn't walk on water three times a day. With the proliferation of data and publications, it's natural for

researchers to feel the pressure to distinguish their findings so they too can participate in the literature. There might be a distinction here between academic researchers and those tied more closely to industry, where the level of evidence required to change a policy is perhaps more closely tied to the importance of the decision. To put it another way, academic researchers may feel freer to inflate the importance of their findings because they are removed from decisionmaking. More honest appraisals of research findings would go a long way to tempering the science bubble.

3. *Don't pass along the risk.* One of the most egregious aspects of the housing bubble was that exotic financial instruments allowed people to keep passing the risk along to unsuspecting buyers. Or sometimes the new buyers were fully aware of the risk, but they deliberately intended to "flip" the investment. Those who created the financial instruments, or who bought them early, tended to pass them off before the value started to erode. In recent critiques of science, people have talked about the "winner's curse," which refers to the fact that the first demonstration of a new effect is often the largest and most eye-catching. From the perspective of science as a community practice, this is indeed a curse. But for the individual or team making the first "sale," it's not a near-term curse at all. Instead, it often can be handsomely rewarding, leaving the burden of diminishing returns to be absorbed by those who invest their time and money to studying further the effect. If we were to require replications of novel findings where they can be done reasonably, the discoverer would bear a more appropriate fraction of the risk, and consumers could be more confident in what they read.

4. *Keep the ratings agencies honest and transparent.* Much has been made of the failure of the ratings agencies in continuing to rate risky investments as AAA. Similarly, in improving the research review process, we should reflect on the way scientists are given incentives to increase production and use *p*-values and bold claims as leverage into the market. When publications are the currency of an expanding and competitive market, there is a risk of inflation.

5. *Be careful about categorical, yes/no reasoning.* You wouldn't simply report the yield on a financial instrument was positive or negative (although that is certainly a helpful distinction). Instead, you'd evaluate the expected return and the risk of loss. Similarly, research results should be reported and evaluated in a context of prior results, effect sizes, uncertainty, and variation. Indeed, we believe the standard statistical framework of false positives and false negatives (type 1 and type 2 errors) is misleading in its implicit assumption that effects are "there" or "not there."

What Message Are We Sending?

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 don't want to overstate this point—after all, the law of large numbers is a mathematical fact, and organizations as different as the Gallup poll, U.S. Census Bureau, and Las Vegas casinos have successfully used statistical principles to transmute the variability of individual cases into near-certainty in the aggregate for decades. And, to return to the analogy with which we began this article, financial instruments can reduce risk by aggregating large numbers of weakly correlated investments. In both cases, though, aggregation can run into problems if we depart too far from the assumptions of the models. For many scientific fields, especially those which—by necessity or choice—are studied using small samples and highly variable measurements, we think it necessary to accept large levels of uncertainty and to move beyond the paradigm in which effective certainty can be obtained via statistical significance and peer review. Just as with mortgage loans, subprime publication can be fine—we just need to be open about the associated uncertainty and variation.

Statisticians—as researchers, reviewers, consultants, and teachers—have a role to play in making sure our scientific literature doesn't become an overheated market of empirical findings where it becomes difficult to discern relative value. If we fail to respond to the current crisis, we might find ourselves facing a market correction in the perceived value of science and possibly a reduction in our collective capacity to generate further knowledge. From the other direction, we would be troubled to see a generalized skepticism of science take hold, in which even well-established findings are up for grabs. The reality is that 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. C

Further Reading

Almenberg, J., K. Kittlitz, and T. Pfeiffer. 2009. An experiment on prediction markets in science. *PLOS-One* 4(12), e8500.

Button, K. S., J. P. A. Ioannidis, C. Mokrysz, B. A. Nosek, J. Flint, E. S. J. Robinson, and M. R. Munafo. 2013. Power failure: Why small sample size undermines the reliability of neuroscience. *Nature Reviews Neuroscience* 14:365–376.

Economist. 2013. Unreliable research: Trouble at the lab. *Economist*, 19 Oct. *www.economist.com/news/briefing/21588057-scientists-think-science-self-correcting-alarming-degree-it-not-trouble*

Francis, G. 2013. Replication, statistical consistency, and publication bias (with discussion). *Journal of Mathematical Psychology* 57:153–169.

Gelman, A. 2013a. Childhood intervention and earnings. *Symposium Magazine*, November. *www.symposium-magazine.com/childhood-intervention-and-earnings*

Gelman, A. 2013b. Is it possible to be an ethicist without being mean to people? *CHANCE* 26(4):51–53.

Gelman, A. 2013c. Too good to be true. *Slate*, 24 Jul.

Gelman, A., and E. Loken. 2012. Statisticians: When we teach, we don't practice what we preach. *CHANCE* 25(1):47–48.

Gelman, A., and E. Loken. 2013. The garden of forking paths: Why multiple comparisons can be a problem, even when there is no 'fishing expedition' or 'p-hacking' and the research hypothesis was posited ahead of time. Technical report, Department of Statistics, Columbia University.

Gertler, P., J. Heckman, R. Pinto, A. Zanolini, C. Vermeerch, S. Walker, S. Chang, and S. Grantham-McGregor. 2013. Labor market returns to early childhood stimulation: A 20-year followup to an experimental intervention in Jamaica. *www.irle.berkeley.edu/workingpapers/142-13.pdf*

Humphreys, M., R. Sanchez, and P. Windt. 2013. Fishing, commitment, and communication: A proposal for comprehensive nonbinding research registration. *Political Analysis* 21:1–20.

Klein, R. A., et al. 2013. Investigating variation in replicability: The 'Many Labs' Replication Project, 14 Jun. *https://openscienceframework.org/project/WX7Ck*

Kriegeskorte, N. 2009. The future of scientific publishing: Open post-publication peer review. *http://futureofscipub.wordpress.com/2009/11/12/open-post-publication-peer-review*

Sabin, P. 2013. *The Bet: Paul Ehrlich, Julian Simon, and our gamble over Earth's future*. New Haven, Conn.: Yale University Press.

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

Tracy, J. L., and A. T. Beall. 2013. Too good does not always mean not true. 30 Jul. University of British Columbia Emotion and Self Lab. *http://ubc-emotionlab.ca/2013/07/too-good-does-not-always-mean-not-true*

Vul, E., C. Harris, P. Winkielman, and H. Pashler. 2009. Puzzlingly high correlations in fMRI studies of emotion, personality, and social cognition (with discussion). *Perspectives on Psychological Science* 4:274–324.

Wagenmakers, E. J., R. Wetzels, D. Borsboom, and H. L. J. van der Maas. 2011. Why psychologists must change the way they analyze their data: The case of psi: Comment on Bem (2011). *Journal of Personality and Social Psychology* 100:426–432.

Wasserman, L. 2012. A world without referees. *www.stat.cmu.edu/~larry/Peer-Review.pdf*

Wolfers, J., and E. Zitzewitz. 2004. Prediction markets. *Journal of Economic Perspectives* 18:107–126.

About the Authors

Andrew Gelman is a professor of statistics and political science and director of the Applied Statistics Center at Columbia University. He has received many awards, including the Outstanding Statistical Application Award from the American Statistical Association and the award for best article published in the *American Political Science Review*. He has coauthored many books; his most recent is *Red State, Blue State, Rich State, Poor State: Why Americans Vote the Way They Do.*

Eric Loken is a research associate professor of human development at Penn State, where he teaches graduate courses in regression and measurement. He studies latent variable models with applications in health and education and has co-founded two web-based assessment companies. The most recent is Criteria Corp, a pre-employment testing company.