Attitudes toward amalgamating evidence in statistics*

Andrew Gelman† Keith O'Rourke‡

6 Sep 2016

Abstract

Weighing of evidence is a central problem in statistics, and there is much debate on what methods are appropriate for the task. At the theoretical level, one can distinguish between hypothesis testing and Bayesian philosophies; in practice, either of these approaches has various points in which contextual information can be plugged in. Beyond this, the system of scientific publication, criticism, and meta-analysis provides another avenue for amalgamation of evidence, and individual statistical analyses can be understood as steps in this larger process. In this article we lay out an interpretation of statistics as information aggregation, a perspective which can unify seemingly distinct statistical philosophies and also provides some guidance to resolving the current replication crisis in science.

1. Statistics as amalgamation of evidence

One of the frustrating—and fascinating—aspects of statistics, compared to many other modern sciences, is its profusion of seemingly incompatible philosophies. The Neyman-Pearson approach is centered around defining procedures for discriminating between hypotheses. The Fisherian p-value, in contrast, evaluates the strength of evidence against a single null hypothesis without explicit reference to any alternative. Another Fisherian approach, maximum likelihood, provides estimates within a parametric model. Bayesian inference can be viewed as a generalization of maximum likelihood but is anathema to many because of iis assignment of probability distributions to parameters that are not the product of ramdon processes. Nonparametric approaches such as bootstrap and lasso have traditionally been shoehorned into the frameworks of hypothesis testing and interval estimation, but in recent years the machine learning approach has focused not on those classical problems but rather on pure prediction. The decision of what information is to be combined is often dictated by probability models or inferential algorithms that themselves are chosen largely by convention. This occurs for basic users who are taught to use t-tests for continuous data, χ^2 tests for continuous data, Cox models for survival data, etc., but even experienced statisticians often do not seem to be clear as to where the choices are made of which information to combine in their data analysis.

Even amid the diversity of statistical methods and philosophies, though, all these approaches involve the amalgamation of evidence. This goes for the simplest models of random sampling and independent identically distributed data; to slightly more elaborate models with hierarchical, time-series, and spatial structure; to elaborate multistage deep learning algorithms combining thousands of predictors or features. Even something as basic as Fisherian p-values or likelihood-ratio testing can be seen as a way to use the accumulation of data—that is, the piling-up of evidence—to draw increasingly certain conclusions.

It has been said that the most important aspect of a statistical method is not what it does with the data but rather what data it uses (Gelman, 2015). From that perspective, the power of

*For a special issue of *Synthese* on Amalgamating Evidence in the Sciences. We thank Samuel Fletcher for inviting this article and the Office of Naval Research for partial support of this work.

†Department of Statistics and Department of Political Science, Columbia University, New York.

‡O'Rourke Consulting, Ottawa, Ontario.

Bayesian, regularization, and machine-learning methods is that they can incorporate large amounts of data into analysis and decision making

At the same time, as datasets become larger and more diverse, there is an increasing need to model and adjust for differences between sample (that is, available data) and population, and between treatment and control groups in causal analysis. Amalgamation of evidence is important but it is not trivial; it is not just a matter of throwing data into a blender. One must evaluate data quality to decide what to include. Or, more generally, one must weight and adjust data in light of what is known about the quality and representativeness of measurements and in light of the consistency of different data sources with available research hypotheses. Implicitly these procedures can be seen as deriving from different probabilistic data-generating models and prior distributions, but in our discussion here we focus on the information included in data analysis, not the algorithms used to construct inferences or the models underlying these algorithms.

Some of the fiercest debates in statistical theory and practice involve the use of prior information. For example, the well-respected statistician David Cox wrote,

> "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)

We expressed disagreement, pointing for example to a problem on "the politically controversial problem of reconstructing historical climate from tree rings":

> "We have a lot of prior information on the processes under which tree rings grow and how they are measured. I don't think anyone would want to just take raw numbers from core samples as a climate estimate! All the tools from Statistical Methods for Research Workers won't take you from tree rings to temperature estimates. You need some scientific knowledge and prior information on where these measurements came from." (Gelman, 2012)

Cox has decades of applied experience and would surely agree that prior information, in the form of physical/biological models, are essential to making climate-related decisions based on tree rings, and we are sure he would also agree that such models involve inevitable subjective choices. Rather, we believe Cox is concerned about the way that Bayesian methods can be abused, what one might call the "moral hazard" involved in a statistical method in which all modeling decisions are up for grabs. In addition is the concern that, in most settings, including the tree-ring example, expressing prior information as probability distributions can lead, paradoxically, to a false sense of certainty. Hence the preference of Cox and others for inclusion of prior information in a more piecemeal, case-by-case manner. From this perspective, the smoothness and apparent all-encompassing nature of Bayesian inference is itself a hazard.

The paradox is that flexibility is required to combine evidence from diverse sources, but if that flexibility is abused, the ultimate conclusions of the analysis can be dictated by the analyst rather than by the data. This is a concern with Bayesian inference with overconfident priors and with classical inference when "p-hacking," "researcher degrees of freedom," and "the garden of forking paths" give users the opportunity to find statistical significance from virtually any dataset (Simmons, Nelson, and Simonsohn, 2011, Gelman and Loken, 2014a). And there is also the choice of what statistical method to use, a decision that is typically not based itself on statistical evidence

(Gelman and O'Rourke, 2013). We offer no general solution here but we think it useful to formulate all statistical methods as data aggregators of one sort of another and to be open about the evidence used to form any particular statistical conclusion—and also the available evidence that, for one reason or another, has been "left on the table" and is not yet incorporated into our inferences.

2. Amalgamation of evidence in the scientific process

Statistical modeling typically focuses on a particular set or stream of data which leads to some inference or decision. But it can also be helpful to think more "sociologically" of an evolution-like mechanism involving thousands of research hypotheses, millions of scientists, and processes of publication, publicity, career rewards, and replication which lead not just to specific conclusions but also to strands of research, subfields, and allocations of research effort: as C. S. Peirce might have put it, communal science that is and remains profitable. Particularly in the field of psychology there has been much recent discussion of the replicability (or lack thereof) of published research claims, and similar concerns have been raised among medical research. As Peirce (1879) wrote,

> "The theory here given rests on the supposition that the object of the investigation is the ascertainment of truth. When the investigation is made for the purpose of attaining personal distinction, the economics of the problem are entirely different. But that seems to be well enough understood by those engaged in that sort of investigation."

But the current de facto procedure, in which studies are summarized by statistically-significant estimates, has technical problems of bias and inefficiency even if we assume all researchers are acting altruistically.

Considering the entire academic research enterprise—the processes of peer review, publication, replication, and meta-analysis—as a grand collective effort of information aggregation, we join a long string of concern from Peirce through Ioannidis (2016) in seeing major problems with incentives and structure, and where simple technical fixes such as weighting studies by appraised quality can be disastrous (Greenland and O'Rourke, 2001). Smaldino and McElreath (2016) offer a simplified but suggestive model of problems with the current system of incentives and publication. On one hand, the diversity of research labs must represent a strength, a potential escape from the groupthink that is associated with central planning. But, from the statistical standpoint, much information is lost by dividing our data into small pieces and summarizing each by a p-value. This would be an inefficient procedure even if p-values were computed as described in the textbooks based on pre-specified tests, but problems of drastic overestimation of effect sizes (Type M or "magnitude" errors; Gelman and Carlin, 2014) become even worse given the documented ability of researchers at all levels to attain statistical significance virtually at will. Systematic overestimation of effect sizes creates a vicious cycle in which new studies are incorrectly anticipated to have a high probability of being successful (Button et al., 2013), leading to further data whose significance is overstated.

A cleaner approach would be to analyze larger data sets directly, not by postprocessing published estimates and p-values but by modeling larger and more diverse sets of raw data. This gives direct access to more efficient statistical analyses and also more ability to check model assumptions. Again, we see a statistical and societal advantage to explicit recognition that inference arises from amalgamation of evidence, and more openness to the sources of this evidence and possible biases,

To step back from data analysis to the scientific enterprise more generally, various specific reforms of science have been proposed, including post-publication review, preregistered replication, and publication/career credit for data quality (rather than just for novelty and statistical signif-

icance). We find it helpful to follow Peirce and think of these as steps in a larger process rather than merely attempts to minimize false positives in isolated studies.

3. Connections to philosophy of science and the history of statistics as a quest for principled amalgamation

Statistical science has evolved from the growing awareness, extraction, and assessment of commonness in the midst of diversity. Not only can physical laws (or, as social scientists say, "law-like relationships") be uncovered from noisy data, in the manner of Gauss, Laplace, and their followers. Also, variation itself can be categorized and thought of as a form of commonality: that was the key insight of Galton, Pearson, and other statisticians who in the late 19th century applied the concept of the probability distribution to biological variation. We have argued that in recent years that this insight has been oversold, now that researchers have the demonstrated ability to extract large, statistically significant, yet spurious and unreplicable findings from just about any set of data (Gelman and Loken, 2014b); that said, from a historical point of view, the idea that variation can itself be quantified is central to any statistical understanding of modern social and biological sciences.

Here we focus on methods of quantifying commonness among different empirical studies and their reported observations. Commonness refers to studies aiming at the same target (aspect of reality) as well as being qualitatively similar evidence of that target. Quantitatively different data sources can vary in their bias and precision (or, to use terms from psychometrics, validity and reliability).

Awareness of commonness can lead to an increase in evidence regarding the target; disregarding commonness wastes evidence; and mistaken acceptance of commonness destroys otherwise available evidence. It is the tension between these last two processes that drives many of the theoretical and practical controversies within statistics.

Statistical science historically emerged out of the conjecture, assessment and reasoned acceptance of the commonness of observations made by different members of the community of astronomers. Among a set of apparently related observations, some combination was conjectured to be better than just enumerating the set, but a justification for how to weight observations, whether repeatedly made by the same astronomer or by different astronomers, was completely lacking and desperately sought. Astronomers and others would often reflect on how to determine which was dataset was the best (thus implicitly assigning weights of 0 to all the remaining data), anticipating that was the obvious solution, but they had yet to learn that, as Stigler (2016) put it, "the details of individual observations had to be, in effect, erased to reveal a better indication than any single observation could on its own." In the modern world of social media we similarly speak of the wisdom of crowds, an idea which is often illustrated using an example of Galton (1907).

The problem of information aggregation attracted the attention of brightest minds at the time, mathematicians and philosophers including Laplace and Gauss, and its resolution finally came from a recognition of a common object being measured by all and the reasonableness of a common error probability model for all—regardless of whether the same or different astronomers were making the observations. That involved a model both for common target of reality ("the" aspect of reality the observation was attempting to get right) and a common observational error that is the same for all. According to Stigler (1986), it was the idea of "dealing with observations made by various other observers under different conditions—that actually 'spurred' on the development." The probabilistic error model, along with the willingness to use it on data from multiple sources, was the key technological insight needed. In gambling, probability models had provided a means to

determine the best bet regarding outcomes from games and or devices that had common chance outcome mechanisms; in contrast, in astronomy the error probability model representing common errors provided a means to determine the best combination for some target taken as common and hence the best weights for the combination of observations (O'Rourke, 2002). In much of statistical practice, probability models provide a formal mathematical basis for amalgamating and assessing commonness which then sets out the best combinations for various purposes. For a thorough historical account see Hald (1998) or Stigler (1986). More recently, machine learning methods have moved to more algorithmic, less model-based approaches—not from any perceived defect with the probability models but rather for computational reasons when dealing with "big data"—but, again, the principle remains that data from different sources can and will be pooled in a single procedure (unless trivially based on single observations).

This repeated broadening can be briefly outlined here as starting with the above-mentioned initial recognition of a common object being measured and the reasonableness of a common error model that implied the weights for the best combination. The next step is extending or revising to still a common object being measured but now a differing error model, one that allows for a source of error that affects all observations taken on that day, but that itself was represented as being drawn from a common distribution of error distributions (a recognition of a commonness at a higher level). This extension/revision implies different weights for the best combination. Earlier, in a different context than astronomy (ratios of male to female births in different cities), the reasonableness of a common error model was kept but the object being measured itself was not taken as common, but instead being conjectured/represented as being a draw from common distribution of objects. That is, the objects themselves that were being measured were allowed to vary but in line with a shared probability distribution. At this point the purposeful designing or bringing-about of commonness in the observations' underlying distributions emerges. An early instance was Peirce's recognition that random sampling and random assignment of treatments induce a common distribution for sample and population, or treatment and control group. Nowadays we might frame all these problems using multilevel models with variance at the observation level and, in the astronomy context, variance components for individual measurement methods, astronomers, and other factors that could induce systematic error.

In Bayesian inference, the prior density is just multiplied by the factors of the likelihood which quantify the information coming from the data (conditional on the assumed class of models). The prior can then be seen, mathematically, as just one more data point. Some authors object to this, using what could be seen as an apple and oranges argument, arguing that now what is being amalgamated is of a different nature. Reid and Cox (2015) express concerns with "merg[ing] seamlessly what may be highly personal assessments with evidence from data possibly collected with great care," instead preferring to use prior information "largely or entirely qualitatively." We disagree and rather see this seemingly outright refusal to consider possible representations of commonness between prior and observations as simply "blocking inquiry" by disallowing a possibly profitable to science representation of the unknown that may well be a "powerful aid to the formation of true and fruitful conceptions," to paraphrase Peirce. For the purpose of the present paper, it is not necessary to resolve this disagreement but just to point out that it can viewed as a question of amalgamation of evidence (rather than as a dispute of objectivity vs. subjectivity, which is how Bayesian/non-Bayesian debates are often framed, for further discussion of this topic see Hennig and Gelman, 2016)

From our perspective one can "interpret the parameter prior in a frequentist way, as formalizing a more or less idealized data generating process generating parameter values" (Hennig and Gelman, 2016). One of earliest to concretely express this view was Francis Galton who constructed a physical

machine to clearly demonstrate both parameter and observation generation. It involved a two stage quincunx. The top level represented the generation or setting of the unknown parameter (the prior) and the second level the generation of a single noisy observation of each observed object's value (the data generating model or likelihood). By tracing back from a chosen value of noisy observations (the slot the pellet ended up in) and identifying all the various values of unknowns parameters that had generated them, a crude sample from the posterior is identified and obtained. Though clunky and limited (just single unknown parameter with a singe observation from each) it fully demonstrated how Bayesian inferences uses probability generating models, both for parameter values and observations, to amalgamate commonness between observations and then those observations and the prior).

There are real risks of taking things as common in a sense that in reality they are not, whether between the parameter generating process and the data generating process or among the data generating process for different observations themselves. We used the phrase "conjecture, assessment, and reasoned acceptance of the commonness" to emphasize that. But similar scientific judgment is required in deciding how to combine measurements—the "likelihood" part of the model—and we do not see the risks of model error as being qualitatively different when considering data-combination rules as when considering how to express prior information; see also Evans (2016) on this point.

Bayesian models "domesticate" uncertainty by turning it into (probabilistically represented) variation; in the jargon of economics, transforming Knightian uncertainty into quantifiable risk. Such procedures gain statistical efficiency at the cost of making mathematical assumptions about the distributions and more importantly, independence of error terms (strong replication) and thus induce skepticism among many potential users; however, alternative approaches that appear to avoid such assumptions can generally be seen to be performing information aggregation in some other way, for example avoiding pooling across data sources but then averaging over time (Gelman, 2013). In just about any situation where a decision needs to be made, *some* choices need to be made regarding pooling of data.

An alternative approach to statistics avoids relying on probability models, instead aiming for procedures that work well under weak assumptions—for example, instead of assuming a distribution is Gaussian, you would just want the procedure to work well under some conditions on the smoothness of the second derivative of the log density function. These approaches also evolved in astronomy, with Legendre developing least squares regression without requiring the probability generating models that Gauss had assumed and used to get the exact same technique.

Instead of requiring probability model assumptions, this approach requires a choice of good properties (why minimize squared error?) over a class of problems to be dealt with (where values of unknowns are constant or linear?). Probability models make representations that try to get at some aspect of reality that cannot be directly assessed but do provide indirect checks on their adequacy. On the other hand, alternative approaches choose properties to be optimal under for a given class of applications with no direct justification for the goodness of the property or guarantees of an application belonging to that class. That is, with no way to assess the goodness of the property or belonging within a class, without making some representation of reality to average or maximize over.

Given sufficient flexibility, data aggregation can always be seen as appropriate, but if the data to be combined are too different—and if there is no good model to bridge these differences—there will be little or any practical gain from pooling, and indeed there can be a risk if analysts might use inappropriately strong models that do not sufficiently account for variation among data sources.

With regards to exactly when observations have something in common amongst them so that aggregation can be applied to useful effect, there is always some judgment involving "replication (or

exchangeability) on some level by the statistician" (Hennig and Gelman, 2016). For a replication to be a true replication and not a mere duplication, there must not be complete dependence, and for a replication to be strong there must be as much independence as is possible. Often data-analytic procedures are set up in terms of observations that can be taken as independent under reasonable assumptions. It is these unit-of-analysis contributions that we wish to understand how to conjecture, extract, and assess of commonness from. In astronomy, the units of analysis were simply individual observations and they were understood as being independent.

An extreme case often arising in social science is when differing scales (for example, aggressiveness, anger, etc.) are used for assessing treatment effects in different randomized experiments. It can be challenging, especially given what is reported in such studies, to specify probability generating models for these different outcomes that had common parameters. This points to the interplay between design of experiments, data collection, and analysis, as expressed for example by Cox (2016). Cleaner data collection puts less of a burden on analysis; conversely, the sorts of "big data" which arise from social media, etc., are messy and require more assumptions in order to make causal inferences and generalize from sample to population. This in turn increases computational requirements, both from sample size and model complexity, and helps explain why much of the work of modern applied and theoretical statistics centers on algorithms and computing. Again, this is all happening within the context of information aggregation; see, for example, Li, Srivastava, and Dunson (2016).

By identifying a target of getting reality right, and an aspect of that reality being common as part of what makes commonness applicable, we are placing ourselves in the larger philosophical community defined by Peirce, Ramsay and others. As we put it elsewhere (Hennig and Gelman, 2016), "Although there is no objective access to observer-independent reality, we acknowledge that there is an almost universal human experience of a reality perceived as located outside the observer and as not controllable by the observer. We see this reality as a target of science, which makes observed reality a main guiding light for science. We are therefore 'active scientific realists' in the sense of Chang (2012), who writes: 'I take reality as whatever is not subject to one's will, and knowledge as an ability to act without being frustrated by resistance from reality' and 'Active scientific realism implies that finding out the truth about objective reality is not the ultimate aim of science, but that science rather aims at supporting human actions.'" We add here that we strive for more than just not being frustrated by resistance from reality; rather, we want our findings and claims that aim at truth to be "beliefs which succeed for reasons connected to the way things are" (Misak, 2016).

The classical view of statistics, briefly mentioned before, of being primarily about procedures to get estimates, tests, confidence intervals, etc. with certain good properties (often common properties for all unknowns) has limitations when moving beyond simple settings. We believe scientific research would be more effective if statistics was viewed instead as primarily about conjecturing, assessing, and adopting idealized representations of reality, predominantly using probability generating models for both parameters and data that can make the most out of commonness, for example using hierarchical models with group-level predictors so that unexplained group-level variance is low and more information can be pooled from different sources (Gelman, 2006). It seems to be already widely supported for probability generating models for data "[providing an] explicit description in idealized form of the physical, biological, ... data generating process," that is essentially "to hypothesize a data generating mechanism that produces observations as if from some physical probabilistic mechanisms" (Reid and Cox, 2015.) We have argued that this approach is too restrictive for much of science.

Our belief in the efficacy of information aggregation, using continuous parameters to determine

the level of partial pooling, is supported by a belief that reality though never directly accessible is continuous, that different experiments, treatments, and outcomes are connected somehow rather than distinct severed islands on their own. Differing considerations and purposes can then be brought to bear on what best combinations (estimates, summaries) follow. From a slightly different direction Tibshirani (2014) argues that enforcing sparsity is not primarily motivated by beliefs about the world, but rather by benefits such as computability and interpretability, indicating how considerations other than correspondence to reality often play an important role in statistics and more generally in science. Tibshirani's view fits squarely within the alternative "classical," or non-Bayesian, approach in which techniques are chosen based on various robust operational properties rather than being viewed as approximations of reality. With this in mind, when we indicated that we considered generating models as an idealization, we need to point out that they could be in fact just be fictions—useful fictions if they lead to an ability to act without being frustrated by resistance from reality. Sometimes fictions do seem turn out to be connected with how things actually are. But if they are just accidental, with anything more than just in the short term, we suspect these will not be as profitable for scientific practice as by definition science (unendingly) tries to get reality right, or at least less wrong.

4. Conclusion

The foundations of statistics remain controversial, even among its leading practitioners, in a way that biology, say, or chemistry or physics are no longer. In many ways, statistics looks more like social sciences such as sociology, economics, and political science which are riven by deep ideological divisions—but with the difference that statistics is a field of mathematics and computing in which ideology does not seem to play any obvious role.

Whatever the historical sources and ultimate resolutions of the debates within the field of statistics, we see the combination of evidence as central to *any* statistical method, and we view methods as stronger to the extend that they can incorporate diverse sources of information, weighting or adjusting appropriately to account for inevitable problems of data quality and representativeness.

Furthermore, we see statistical concepts of data integration, and the quantification of uncertainty and variation, as central to serious understanding and reforms of the currently-broken system of scientific publication and promotion.

Finally, all these concerns relate to longstanding questions in the philosophy of science, following a skeptical tradition of Peirce, Meehl, and others. Ironically, various modern abuses of statistics such as the chase for statistical significance or, more generally, the deterministic thinking that leads researchers to establish certitude beyond the capabilities of their data, arise from skeptical ideas in statistics such as Fisher's warnings about overinterpreting chance variation or the Neyman-Pearson-Wald rigorizing of certain stylized statistical decision problems.

When amalgamating evidence we typically are at least one step beyond available theory—it only *feels* like amalgamation if it cannot be done automatically—but we should not let this stop us from trying. It is though formalizing and modeling our attempts at combining information—and by recording and learning from our failures—that we will do better.

References

Airy, G. B. (1861). *On the Algebraical and Numerical Theory of Errors of Observations and the Combination of Observations*. Macmillan.

Berger, J. O., Liseo, B., and Wolpert, R. L. (1999). Integrated likelihood methods for eliminating nuisance parameters. *Statistical Science* **14**, 1–28.

Button, K. S., Ioannidis, J. P. A., Mokrysz, C., Nosek, B. A., Flint, J., Robinson, E. S. J., and Munafo, M. R. (2013). *Nature Reviews Neuroscience* **14**, 365–376.

Cox, D. R. (1982). Combination of data. In *Encyclopedia of Statistical Science*, ed. S. Kotz and N. L. Johnson. New York: Wiley.

Cox, D. R. (2001). Some remarks on likelihood factorization. *Institute of Mathematical Statistics Lecture Notes—Monograph Series* **36**, 165–172.

Cox, D. R. (2016). The design of empirical studies: Toward a unified view. *European Journal of Epidemiology* **31**, 217–228.

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

Evans, M. (2016). Measuring statistical evidence using relative belief. *Computational and Structural Biotechnology Journal* **14**, 91–96.

Fisher, R. A. (1956). *Statistical Methods and Scientific Inference.*

Fraser, D. F. (1976). *Probability and Statistics: Theory and Applications.* Duxbury Press.

Galton, F. (1907). Vox populi. *Nature* **75**, 450–451.

Gelman, A. (2006). Multilevel (hierarchical) modeling: What it can and cannot do. *Technometrics* **48**, 241–251.

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

Gelman, A. (2013). Everyone's trading bias for variance at some point, it's just done at different places in the analyses. Statistical Modeling, Causal Inference, and Social Science blog, 14 Mar. `http://andrewgelman.com/2013/03/14/everyones-trading-bias-for-variance-at-some-point-its-just-done-at-different-places-in-the-analyses/`

Gelman, A., (2015). Regression: What's it all about? Review of *Bayesian and Frequentist Regression Methods*, by Jon Wakefield. *Statistics in Medicine.*

Gelman, A., and Carlin, J. B. (2014). Beyond power calculations: Assessing Type S (sign) and Type M (magnitude) errors. *Perspectives on Psychological Science* **9**, 641–651.

Gelman, A., and Loken, E. (2014a). The statistical crisis in science. *American Scientist* **102**, 460–465.

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

Gelman, A., and O'Rourke, K. (2015). Convincing evidence. In *Roles, Trust, and Reputation in Social Media Knowledge Markets*, ed. Sorin Matei and Elisa Bertino. New York: Springer.

Greenland, S., and O'Rourke, K. (2001). On the bias produced by quality scores in meta-analysis, and a hierarchical view of proposed solutions. *Biostatistics* **2**, 463–471.

Greenland, S., and O'Rourke, K. (2008). Meta-analysis. In *Modern Epidemiology*, ed. Rothman, K. J., Greenland, S., and Lash, T. L. Lippincott Williams and Wilkins.

Gelman, A., and Shalizi, C. R. (2013). Philosophy and the practice of Bayesian statistics (with discussion). *British Journal of Mathematical and Statistical Psychology* **66**, 8–38.

Gelman, A., and Hennig, C. Beyond subjective and objective in statistics. `http://www.stat.columbia.edu/~gelman/research/unpublished/objectivityr3.pdf`

Hald, A. (1998). *A History of Mathematical Statistics from 1750 to 1930.* Wiley.

Havenel, J. (2008). Peirce's clarifications of continuity. *Transactions of the Charles S. Peirce Society: A Quarterly Journal in American Philosophy* **44**, 86–133.

Ioannidis, J. P. A. (2016). Evidence-based medicine has been hijacked: A report to David Sackett. *Journal of Clinical Epidemiology* **73**, 82–86.

Keynes, J. M. (1911). The principal averages and the laws of error which lead to them. *Journal of the Royal Statistical Society* **74**, 322–331.

Li, C., Srivastava, S. and Dunson, D. B. (2016). Simple, scalable and accurate posterior interval estimation. `https://arxiv.org/abs/1605.04029`

Meng, X. L. (2009). Decoding the h-likelihood. *Statistical Science*, **24**, 280–293.

Misak, C. (2016). *Cambridge Pragmatism: From Peirce and James to Ramsey and Wittgenstein.* Oxford University Press.

O'Rourke, K. (2002). Meta-analytical themes in the history of statistics: 1700 to 1938. *Pakistan Journal of Statistics* **18**, 285–300.

O'Rourke, K. (2007). An historical perspective on meta-analysis: dealing quantitatively with varying study results. *Journal of the Royal Society of Medicine* **100**, 579–582.

O'Rourke, K., and Altman, D. G. (2005). Bayesian random effects meta-analysis of trials with binary outcomes: Methods for the absolute risk difference and relative risk scales. *Statistics in Medicine* **24**, 2733–2742.

Owen, A. B. (2009). Karl Pearson's meta-analysis revisited. *Annals of Statistics* **37**, 3867–3892.

Pearl, J., and Bareinboim, E. (2014). External validity: From do-calculus to transportability across populations. *Statistical Science* **29**, 579–595.

Peirce, C. S. (1879). Note on the theory of the economy of research. *Report of the Superintendent of the United States Coast Survey Report*, 197–201.

Reid, N., and Cox, D. R. (2015). On some principles of statistical inference. *International Statistical Review* **83** 293–308.

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.

Smaldino, P., and McElreath, R. (2016). The natural selection of bad science. `http://arxiv.org/pdf/1605.09511v1.pdf`

Stigler, S. M. (1986). *The History of Statistics: The Measurement of Uncertainty Before 1900.* Harvard University Press.

Stigler, S. M. (2016). *The Seven Pillars of Statistical Wisdom.* Harvard University Press.

Tibshirani, R. J. (2014). In praise of sparsity and convexity. In *Past, Present, and Future of Statistical Science*, ed. X. Lin, C. Genest, D. L. Banks, G. Molenberghs, D. W. Scott, and J. L. Wang, 497–505. London: Chapman and Hall.

Warn, D. E., Thompson, S. G., and Spiegelhalter, D. J. (2002). Bayesian random effects meta-analysis of trials with binary outcomes: methods for the absolute risk difference and relative risk scales. *Statistics in Medicine* **21**, 1601–1623.

Wible, J. R. (1994). Charles Sanders Peirce's economy of research. *Journal of Economic Methodology* **1**, 135–160.