To appear in Entropy

Statistics as a social activity: Attitudes toward amalgamating evidence*

Andrew Gelman† Keith O'Rourke‡

22 Jul 2024

Abstract

Amalgamation of evidence in statistics is done in several ways. Within a study, multiple observations are combined by averaging or as factors in a likelihood or prediction algorithm. In multilevel modeling or Bayesian analysis, population or prior information are combined with data using the weighted averaging derived from probability modeling. In a scientific research project, inferences from data analysis are interpreted in light of mechanistic models and substantive theories. Within a scholarly or applied research community, data and conclusions from separate laboratories are amalgamated through a series of steps including peer review, meta-analysis, review articles, and replication studies.

These issues have been discussed for many years in the philosophy of science and statistics, gaining attention in recent decades first with the renewed popularity of Bayesian inference and then with concerns about the replication crisis in science. In this article, we review amalgamation of statistical evidence from different perspectives, connecting the foundations of statistics to the social processes of validation, criticism, and consensus building.

1. Aggregating information in a social context

Weighing and amalgamating evidence is a central problem in the process of science, giving rise to much debate on what methods are appropriate as well as exactly where, when and for what purposes they should be used. Within statistics, a central area of controversy is how to incorporate prior information in the context of objective and reproducible science. On the other hand, the weighing and amalgamating of evidence within a single isolated study (the multiple observations) in many default approaches in statistics is surprisingly often just automatic and implicit.

Vigorous debate on basic approaches in statistics likely comes as no surprise to statisticians and increasingly almost everyone else. Although there is much agreement on mathematical definitions of terms and procedures in statistics (what they are), as well as the discerning if particular instances meet these (is it this?), when it comes to the appropriate roles for these terms and procedures in facilitating scientific inquiry—their very purposes and what to make of them—it seems beyond agreement for the foreseeable future. The tools are largely agreed upon, but their appropriate use, where and for what purposes, not at all. For instance, there is a fair amount of agreement on what probabilities are, but not on what they can be used for. Many frequentists ban any use of probability in representing (uncertain) knowledge of unknown parameters. On the other hand, although almost all Bayesians would use probabilities to represent knowledge (or lack of it), some Bayesians would ban any testing or empirically based assessment of these. In the case of a single study, some statisticians would be concerned about properties of procedures that can be discerned if the procedure would be repeatedly applied infinitely often under similar kinds of studies or even exactly the same study conditions. Others argue this is not even sensible.

Gillies (1991, 2000) discusses a "pluralist" or "intersubjective" interpretation of probability, which is related to the concept of "institutional decision analysis," for which an important aspect

*To appear in *Entropy*. We thank three reviewers for helpful comments and the Office of Naval Research for grant N00014-15-1-2541.

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

‡O'Rourke Consulting, Ottawa, Ontario.

of inference is that decisions be justifiable, which motivates clear linkages between measurement, modeling assumptions, inferences, and decision recommendations. Once we situate science and data analysis in a social context, these challenges become clear. In the present paper we do not ourselves set up any agent-based models; rather, our goal is to provide a broader perspective on the problem of aggregation of evidence in statistics.

Going beyond a single isolated study, the system of scientific publication, criticism, and meta-analysis provides more general avenues for amalgamation of evidence between rather than just within a study, and here individual statistical analyses can be understood as (first) steps in this larger process. Perhaps unsurprisingly in this larger process there are more disagreements as opinions vary on what contextual (extra study) information can be plugged in, where and how. Should previous studies be amalgamated in a combined analysis, or just used to build a judgement informed prior or just used qualitatively to refine what analysis should done to make the study stand on its own as much as possible? In this article we lay out a general prospective on statistics as primarily about conjecturing, assessing, and adopting idealized representations of reality, predominantly using probability generating models for both parameters and data. That is, an explicit prior probability distribution to represent available but rough scientific conjectures of what values the unknown parameters might have been set to and a data generating probability distribution to represent how the recorded data likely came about if the unknown parameters' values were set to specific possible values. This contrasts with another prospective on statistics, as primarily being about discerning procedures with good properties that are uniform over a wide range of possible underlying realities and restricting use, especially in science, to just those procedures. Our perspective is perhaps more inviting of information aggregation, as reality likely has many commonalities that can be discerned and profitably gained from. We believe it is a perspective which can unify seemingly distinct statistical philosophies as well as provide some guidance to resolving the current replication crisis in science—as when claims fail to replicate—the methods used likely did not reflect reality well, if at all.

2. 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, targeting uniform type one error for all nulls and uniformly minimum type two errors for all alternatives. The Fisherian p-value, in contrast, evaluates the strength of evidence against a single null hypothesis without explicit reference to any alternative, targeting a uniform(0,1) distribution of p-values for all nulls. Another Fisherian approach, maximum likelihood, provides estimates within a parametric model (see Fraser, 1976, Cox, 2001, and Meng, 2009); meanwhile, Neyman-Pearson testing can be interpreted in Bayesian terms (Rubin, 1995, Li and Mealli, 2014). Bayesian inference can be viewed as a generalization of maximum likelihood but is anathema to many because of its assignment of probability distributions to parameters that are not the product of random processes. It targets probability distributions that represent current understanding of the realities and uncertainties involved. 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. They target lessening of assumptions (used to represent current understanding of the realities and uncertainties) involved and more identification of procedures with seemingly good properties. 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 (group variances assumed to be common to give a combined variance estimate with more degrees of freedom), χ^2 tests for discrete data (various choices of common parameters to assume in defining the expectations to test consistency with), linear regression models (assuming all observations have common slopes given the explanatory variables fit as well as common standard deviation), Cox models for survival data (common proportional hazard function assumed so that it cancels out), 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, and integration of the likelihood, while considered Bayesian, can also be interpreted more generally (Berger, Liseo, and Wolpert, 1999).

Modern data science has moved away from a mathematical framework of hypothesis testing and model building, toward a computationally-focused environment of prediction, external validation, and reproducibility (Donoho, 2024). From that perspective, we can think of evidence being combined not to evaluate theories but to form more effective predictions—but probability modeling can still be useful in constructing procedures that combine information efficiently while pointing to potential areas of sensitivity where predictions of interest are strongly influenced by particular decisions regarding pooling, partial pooling, or exclusion of evidence.

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. From that perspective, the power of 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 had decades of applied experience and would surely have agreed 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 have agrede that such models involve inevitable subjective choices. Rather, we believe Cox was 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. Perhaps default methods for combining evidence from diverse sources will be too hazardous? 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). 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.

Beyond this, when we move beyond simple textbook examples of experimentation and sampling, there is typically no default analysis available. There is no general way to deciding how to choose, combine, and transform regression predictors or features in a predictive model, and many scientific problems inherently involve integration of information from different sources, for example medical research interpreting clinical trial data in the light of biological models; or climate modeling combining data from tree rings, historical temperature data, and physical modeling; or election forecasting combining information from state polls, national polls, and forecasts based on economic and political conditions.

3. 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 (Wible, 1994), 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) 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. Fisher unfortunately may have undermined the appreciation of this with his claim,

> "It is usually convenient to tabulate its [the likelihood's] logarithm, since for independent bodies of data such as might be obtained by different investigators, the "combination of observations" requires only that the log-likelihoods be added." (Fisher, 1956)

This is technically correct if the data generating model (which defines the likelihood) is never questioned or assessed—but it should be. To do that adequately one needs all the individual raw data from all studies. Here, we quickly add that the prior's logarithm need only be added to the likelihood's logarithm to start a Bayesian analysis. Like the data generating model, the prior also needs to be questioned or assessed. 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 significance). 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. This quote from Peirce might

suffice: "I [Peirce] do not call the solitary studies of a single man a science. It is only when a group of men, more or less in intercommunication, are aiding and stimulating one another by their understanding of a particular group of studies as outsiders cannot understand them, that I call their life a science." but we also include two longer passages:

> "Science is to mean for us a mode of life whose single animating purpose is to find out the real truth, which pursues this purpose by a well-considered method, founded on thorough acquaintance with such scientific results already ascertained by others as may be available, and which seeks cooperation in the hope that the truth may be found, if not by any of the actual inquirers, yet ultimately by those who come after them and who shall make use of their results." (also in Pearce, 1902, 7.55)

> "But what I mean by a "science" (...) is the life devoted to the pursuit of truth according to the best known methods on the part of a group of men who understand one another's ideas and works as no outsider can. It is not what they have already found out which makes their business a science; it is that they are pursuing a branch of truth according, I will not say, to the best methods, but according to the best methods that are known at the time. I do not call the solitary studies of a single man a science. It is only when a group of men, more or less in intercommunication, are aiding and stimulating one another by their understanding of a particular group of studies as outsiders cannot understand them, that I call their life a science. It is not necessary that they should all be at work upon the same problem, or that all should be fully acquainted with all that it is needful for another of them to know; but their studies must be so closely allied that any one of them could take up the problem of any other after some months of special preparation and that each should understand pretty minutely what it is that each one of the other's work consists in; so that any two of them meeting together shall be thoroughly conversant with each other's ideas and the language he talks and should feel each other to be brethren." (MS 1334, pp. 11–14, 1905)

4. 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 such as Airy (1861). 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; 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, hopefully varying only in precision which can be readily assessed. On the other hand, qualitatively different data sources can vary in their bias, which may be very difficult to assess and properly correct for so that something is

actually common. Terms from psychometrics that make the same distinctions as bias and precision do here would be 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. A concrete but simple example that demonstrates practical controversies nicely would be the situation depicted in the wiki entry on Simpson's paradox (Schutz, 2017). The illustration of the quantitative version: a positive trend appears for two separate groups (blue and red), whereas a negative trend (black, dashed) appears when the groups are combined.

The illustration clearly depicts an underlying reality of exactly the same positive trend for two groups (both slopes equal to 1) but that happen to have different intercepts, one at about 5 and the other -7. A default application of regression modeling using the eight data points displayed in the illustration would likely specify a single intercept, slope, and standard deviation parameter. The incorrect single intercept here is a mistaken acceptance of commonness which destroys the evidence for common positive slopes by providing a single negative slope estimate of roughly -0.6, in addition providing a single incorrect intercept estimate of about 9. Specifying the correct commonness here—that of separate intercepts but a single common slope and single standard deviation parameter—captures (all the evidence for) the correct intercept and slope with no actual error, with resulting correct estimates of the intercepts of 5 and -7, slope of 1, and single standard deviation of 0. With realistic data, there would be errors of observation and the specifying of separate intercepts but along with incorrect separate slopes and again common standard deviation parameter would waste evidence providing two different slope estimates randomly varying about 1 and a biased-downward estimate of the standard deviation. One might further ask or question why the assumption of common standard deviation was being made. Simply convenience?

This simple contrived example from wiki reveals a lack of concern regarding the need to represent reality (correctly specifying common and non-common parameters) as well as one can or at least well enough for statistical procedures to provide reasonable answers. The usual training in statistics likely suggests the default use of a common intercept in multivariate regression analysis as well as the occasional need to consider interactions (with the statistical custom of always specifying separate intercepts for interaction terms—the lower order terms). But here, without the interaction, evidence is destroyed while with the interaction, it is wasted. The result is misleading descriptive or predictive inference with the former or inefficient descriptive or predictive inference with the latter.

Better descriptive or predictive inferences comes from better underlying representations of reality. For descriptive or predictive inference, reality as it is now (and likely to persist) is the only reality that needs to be represented well enough. On the other hand, for causal or transportable inference, both current realities and how those can be changed (causal) or local and remote realities (to transport between) need to be well represented (Pearl and Bareinboim, 2014). For instance, randomization creates two similar in expectation realities, one which presumably can be modified in a hopefully simple way (though maybe seldom as hopeful as additive). Causal and or transportable inference is of course much more challenging, but this should not lead to a disregard for representing a single current reality well for descriptive or predictive inference.

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, Keynes, 1911). 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 overviews and historical accounts, see Cox (1982), Stigler (1986), and Hald (1998). 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).

A repeated broadening of what was taken as common 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 being drawn from 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. To make this absolutely clear, each observation defines a likelihood (the probabilities of observing that very observation for the various parameter values the parameters can take), the study likelihood is a multiple of those single observation likelihoods (conditioned on other observations if observations are not independent) and the posterior is proportional to the prior multiplied by the combined study likelihood. That multiplication can be rearranged and re-expressed in any way that does not change that. Taking logarithms, the log posterior is proportional to the sum of the log prior and all the individual log likelihoods—a "weighted combination" with the "weights" determined by functional form of the prior density and individual likelihoods. Some authors object to the prior being in this combination, 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 Gelman and Hennig (2017).

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" (Gelman and Hennig, 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 inference 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. In just about any situation where a decision needs to be made, *some* choices need to be made regarding pooling of data.

Comparing the technique of nearest neighbors to linear regression will help clarify what we mean by unavoidable choices being made for pooling. For simplicity, consider a single x variable and its role in predicting a single y variable. A linear regression model conjectures a single common intercept and common slope for predicting the expected value of y from all values of x as well as a common standard deviation parameter. The probability model for all observations is taken to be normal$(a + bx, \sigma)$ and all observations provide evidence for just three parameters. In contrast, nearest neighbor regression tries to avoid specifying any commonality of expected values of y for differing values of x, instead allowing expected values to vary arbitrarily by neighborhoods. The technique identifies these neighborhoods from the observations and takes averages only within neighborhoods (never across). It usually specifies the size of these neighborhood. Taking the size of the neighborhoods as 2 requires that a single nearest neighbor is found for every observation and taken to have the same expectation—referred to as NN_1. Nearest neighbors does not actually avoid specifying some commonness, though, as NN_0 is not taken as an acceptable procedure (having no neighbors, all observations must be taken as islands on their own). So commonness of expectation between at least two observations is forced. Then to achieve better "good" properties commonness is then allowed over a larger number of observations depending on the data set—referred to as NN_k. Additionally, a common variance is usually assumed between neighborhoods (a secondary feature). That is to get combinations based on more than one observation, and variance estimates from more than a few isolated points in each neighborhood, you treat non-common points and non-common parameters "grudgingly" as common—simply to improve properties for estimating what you can.

This 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 constrained in some way such as being linear in regression or proportional hazards in survival analysis). 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 (e.g. applications having linear expectations or proportional hazards) with no direct justification for the goodness of the property nor guarantees of a particular application belonging to that class. That is, with no way to assess the goodness of the property or belonging within that appropriate 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 regard 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" (Gelman and Hennig, 2017). 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 (2017).

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 (Gelman and Hennig, 2017), "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 possible 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. 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 limiting probability generating models just for data while banning them for parameters 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.

5. 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. However, the mathematics and computing just defines and implements the tools (where there is much agreement), the purposes to which they should be put and what to make of what results from their use in particular applications is more than just mathematics and computing (and here there is little agreement.)

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 a longstanding skeptical tradition in the philosophy of science. 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 through recognizing, 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.

Chang, H. (2012). *Is Water H2O?: Evidence, Realism and Pluralism*. Springer.

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

Cox, D. R. (2001). Some remarks on likelihood factorization. *Lecture Notes—Monograph Series*, 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.

Donoho, D. (2024). Data science at the singularity (with discussion). *Harvard Data Science Review* **6** (1).

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*. Hafner.

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

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

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

Gelman, A., and Hennig, C. (2017). Beyond subjective and objective in statistics (with discussion and rejoinder). *Journal of the Royal Statistical Society A* **180**, 967–1033.

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. Springer.

Gillies, D. (1991). Intersubjective probability and confirmation theory. *British Journal of the Philosophy of Science* **42**, 513e533.

Gillies, D., (2000). *Philosophical Theories of Probability*. Routledge.

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.

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

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. (2017). Simple, scalable and accurate posterior interval estimation. *Biometrika* **104**, 665–680.

Li, F., and Mealli, F. (2014). A conversation with Donald B. Rubin. *Statistical Science* **29**, 439–457.

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.

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.

Peirce, C. S. (1902). The Charles S. Peirce Papers. 32 microfilm reels of the manuscripts kept in the Houghton Library. Cambridge: Harvard University Library, Photographic Service 1966.

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

Rubin, D. B. (1995). Bayes, Neyman, and calibration. Discussion of Berk, Western, and Weiss. *Sociological Methodology* **25**, 473–479.

Schutz (2017). Simpson's paradox continuous. `https://en.wikipedia.org/wiki/Simpson's_paradox`

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. *Royal Society Open Science* **3**, 160384.

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. CRC Press.

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