Beyond subjective and objective in statistics*

Andrew Gelman† Christian Hennig‡

21 June 2016

Abstract

Decisions in statistical data analysis are often justified, criticized, or avoided using concepts of objectivity and subjectivity. We argue that the words "objective" and "subjective" in statistics discourse are used in a mostly unhelpful way, and we propose to replace each of them with broader collections of attributes, with objectivity replaced by *transparency*, *consensus*, *impartiality*, and *correspondence to observable reality*, and subjectivity replaced by awareness of *multiple perspectives* and *context dependence*. The advantage of these reformulations is that the replacement terms do not oppose each other and that they give more specific guidance about what statistical science strives to achieve. Instead of debating over whether a given statistical method is subjective or objective (or normatively debating the relative merits of subjectivity and objectivity in statistical practice), we can recognize desirable attributes such as transparency and acknowledgment of multiple perspectives as complementary goals. We demonstrate the implications of our proposal with recent applied examples from pharmacology, election polling, and socioeconomic stratification.

1. Introduction

We can't do statistics without data, and as statisticians much of our efforts revolve around modeling the links between data and substantive constructs of interest. We might analyze national survey data on purchasing decisions as a way of estimating consumers' responses to economic conditions; or gather blood samples over time on a sample of patients with the goal of estimating the metabolism of a drug, with the ultimate goal of coming up with a more effective dosing schedule; or we might be performing a more exploratory analysis, seeking clusters in a multivariate dataset with the aim of discovering patterns not apparent in simple averages of raw data.

As applied researchers we are continually reminded of the value of integrating new data into an analysis, and the balance between data quality and quantity. In some settings it is possible to answer questions of interest using a single clean dataset, but more and more we are finding that this simple textbook approach does not work.

External information can come in many forms, including (a) recommendations on what variables to adjust for non-representativeness of a survey or imbalance in an experiment or observational study; (b) the extent to which outliers should be treated as regular, erroneous, or as indicating something that is meaningful but essentially different from the main body of observations; (c) issues of measurement, confounding, and substantively meaningful effect sizes; (d) population distributions that are used in poststratification, age adjustment, and other procedures that attempt to align inferences to a common population of interest; (e) restrictions such as smoothness or sparsity that serve to regularize estimates in high-dimensional settings; (f) the choice of the functional form in a regression model (which in economics might be chosen to work with a particular utility function, or in public health might be motivated based on success in similar studies in the literature); and

*We thank Sebastian Weber, Jay Kadane, Arthur Dempster, Michael Betancourt, Michael Zyphur, E. J. Wagenmakers, Deborah Mayo, James Berger, Prasanta Bandyopadhyay, Laurie Paul, Jan-Willem Romeijn, Gianluca Baio, Keith O'Rourke, Laurie Davies and the anonymous reviewers for helpful comments.

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

‡Department of Statistical Science, University College London.

(g) numerical information about particular parameters in a model. Of all these, only the final item is traditionally given the name "prior information" in a statistical analysis, but all can be useful in serious applied work. Other relevant information concerns not the data generating process but rather how the data and results of an analysis are to be used or interpreted.

We were motivated to write the present paper because we felt that our applied work, and that of others, was impeded because of the conventional framing of certain statistical analyses as subjective. It seemed to us that, rather than being in opposition, subjectivity and objectivity both had virtues that were relevant in making decisions about statistical analyses. We have earlier noted (Gelman and O'Rourke, 2015) that statisticians typically choose their procedures based on non-statistical criteria, and philosophical traditions and even the labels attached to particular concepts can affect real-world practice.

In Section 2 we show how the discussions of objectivity and subjectivity affect statistical practice and statistical thinking, followed by an outline of our own philosophical attitude to objectivity and subjectivity in science; an appendix provides an overview of what the philosophy of science has to say on the matter. In Section 3 we present our proposal, exploding objectivity and subjectivity into several specific virtues to guide statistical practice. In Section 4 we demonstrate the relevance of these ideas for three of our active applied research projects: a hierarchical population model in pharmacology, a procedure for adjustment of opt-in surveys, and a cluster analysis of data on socioeconomic stratification. In Section 5 we revisit fundamental approaches to statistics using the proposals of Section 3, demonstrating how they can elaborate advantages and disadvantages of the various approaches in a more helpful way than the traditional labeling of "objective" and "subjective." Section 6 contains a final discussion, including a list of issues in scientific publications that could be addressed using the virtues proposed here.

2. Objectivity and subjectivity

2.1. Objectivity, subjectivity, and decision making in statistics

Concepts of objectivity and subjectivity are often used to justify, criticize, avoid, or hide the decisions made in data analysis. Typically, the underlying idea is that science should be objective, understood as something like "independence of personal biases," without referring to any clear definition. Concepts of objectivity can be implicitly invoked when making choices, so that certain decisions are avoided or hidden in order to not open an analysis to charges of subjectivity.

For many statistical methods tuning constants need to be decided such as the proportion of trimmed observations when computing a trimmed mean or bandwidths for smoothing in density estimation or nonparametric regression; one could also interpret the conventional use of the 0.05 significance level as a kind of tuning parameter. In the statistical literature, methods are advertised by stating that they don't require any tuning decisions by the user. Often these choices are hidden so that users of statistics (particularly those without specific background knowledge in statistics) expect that for their data analysis task there is a unique correct statistical method. This expectation is exploited by the marketing strategies for some data analysis software packages such as Beyondcore that suggest that a default analysis is only one click away. While such an approach obviously tempts the user by its simplicity, it also appeals on the level of avoiding individual impact or subjectivity. Decisions that need to be made are taken out of the hand of the user and are made by the algorithm, removing an opportunity for manipulation but ignoring valuable information about the data and their background. This is in stark contrast to the typical trial-and-error way of building one or more statistical models with lots of subjective decisions starting from data pre-processing via data

exploration and choice of method on to the selection of how to present which results. Realistically, even one-click methods require user choices on data coding and data exclusion, and these inputs can have big impacts on end results such as p-values and confidence intervals (Steegen et al., 2006).

More mathematically oriented statisticians design and choose methods that optimize criteria such as unbiased minimum variance, often relying on restrictive assumptions. This can allow for elegant theoretical results with very restricted scope. When methods are to be compared in simulation studies, typically there is a huge variety of choices including data generating processes (distributions, parameters, dimensionality, etc.), performance measures, and tuning of competitors. This can easily discourage researchers from running such studies at all or at least beyond one or two illustrative toy setups, but despite their subjective flavor and the difficulty to find a path through a potentially confusing jungle of results, such studies can be informative and raise important issues. Already in 1962, Tukey criticized an obsession with mathematical optimization problems concentrating on one-dimensional criteria and simple parametric models in the name of objectivity, and stated that "in data analysis we must look to a very heavy emphasis on judgment."

Researchers often rely on the seeming objectivity of the $p < 0.05$ criterion without realizing that theory behind the p-value is invalidated when analysis is contingent on data (Simmons, Nelson, and Simonsohn, 2011, Gelman and Loken, 2014). Significance testing can be part of a misguided ideology that leads researchers to hide, even from themselves, the iterative searching process by which a scientific theory is mapped into a statistical model or choice of data analysis (Box, 1983). More generally, overreaction to concerns about subjectivity can lead researchers to avoid incorporating relevant and available information into their analyses and to not adapt analyses appropriately to their research questions and potential uses of their results.

Personal decision making cannot be avoided in statistical data analysis, and for want of approaches to justify such decisions, the pursuit of objectivity degenerates easily to a pursuit to merely *appear* objective. Scientists whose methods are branded as subjective have the awkward choice of either saying, No, we are really objective, or else embracing the subjective label and turning it into a principle, and the temptation is high to avoid this by hiding researcher degrees of freedom from the public unless they can be made to appear "objective." Such attitudes about objectivity and subjectivity can be an obstacle to good practice in data analysis and its communication, and we believe that researchers can be guided in a better way by a list of more specific scientific virtues when choosing and justifying their approaches.

The continuing interest in and discussion of objectivity and subjectivity in statistics is, we believe, a necessary product of a fundamental tension in science: On one hand, scientific claims should be impersonal in the sense that a scientific argument should be understandable by anyone with the necessary training, not just by the person promulgating it, and it should be possible for scientific claims to be evaluated and tested by outsiders. On the other hand, a reality assumed to be objective in the sense of being independent of its observers, is only accessible through observations made by observers and dependent on their perspectives; communication about the observations and the process of observation and measurement relies on language constructs. Thus objective and subjective elements arise in the practice of science, and similar considerations hold in statistics.

Within statistics, though, discourse on objectivity and subjectivity is at an impasse. Ideally these concepts would be part of a consideration of the role of different sorts of information and assumptions in statistical analysis, but instead they often seemed to be used in restrictive and misleading ways.

One problem is that the terms "objective" and "subjective" are loaded with so many associations and are often used in a mixed descriptive/normative way. For example, a statistical method that does not require the specification of any tuning parameters is objective in a descriptive sense (it

does not require decisions by the individual scientist). Often this is presented as an advantage of the method without further discussion, implying objectivity as a norm, but the lack of flexibility caused by the impossibility of tuning can actually be a disadvantage (and indeed can lead to subjectivity at a different point in the analysis, when the analyst must make the decision of whether to use an auto-tuned approach in a setting where its inferences do not appear to make sense). The frequentist interpretation of probability is objective in the sense that it locates probabilities in an objective world that exists independently of the observer, but the definition of these probabilities requires a subjective definition of a reference set. Some proponents of frequentism consider its objectivity (in the sense of impersonality, conditional on the definition of the reference set) as a virtue, but this property is ultimately only descriptive; it does not imply on its own that such probabilities indeed exist in the objective world, nor that they are a worthwhile target for scientific inquiry.

In discussions of the foundations of statistics, objectivity and subjectivity are seen as opposites. Objectivity is typically seen as a good thing; many see it as a major requirement for good science. Bayesian statistics is often presented as being subjective because of the choice of a prior distribution. Some Bayesians (notably Jaynes, 2003, and Berger, 2006) have advocated an objective approach, whereas others (notably de Finetti, 1974) have embraced subjectivity. It has been argued that the subjective/objective distinction is meaningless because all statistical methods, Bayesian or otherwise, require subjective choices, but the choice of prior distribution is sometimes held to be particularly subjective because, unlike the data model, it cannot be determined even in the asymptotic limit. In practice, subjective prior distributions often have well known empirical problems such as overconfidence (Alpert and Raiffa, 1984, Erev, Wallsten, and Budescu, 1994), which motivates efforts to check and calibrate Bayesian models (Rubin, 1984, Little, 2012) and to situate Bayesian inference within an error-statistical philosophy (Mayo, 1996, Gelman and Shalizi, 2013).

De Finetti can be credited with acknowledging honestly that subjective decisions cannot be avoided in statistics, but it is misleading to think that the required subjectivity always takes the form of prior belief. The confusion arises from two directions: first, prior distributions are not necessarily any more subjective than other aspects of a statistical model; indeed, in many applications priors can and are estimated from data frequencies (see Chapter 1 of Gelman, Carlin, et al., 2013, for several examples). Second, somewhat arbitrary choices come into many aspects of statistical models, Bayesian and otherwise, and therefore we think it is a mistake to consider the prior distribution as the exclusive gate at which subjectivity enters a statistical procedure.

On one hand, statistics is sometimes said to be the science of defaults: most applications of statistics are performed by non-statisticians who adapt existing general methods to their particular problems, and much of the research within the field of statistics involves devising, evaluating, and improving such generally applicable procedures (Gelman, 2014b). It is then seen as desirable that any required data-analytic decisions or tuning are performed in an objective manner, either determined somehow from the data or justified by some kind of optimality argument.

On the other hand, practitioners must apply their subjective judgment in the choice of what method to use, what assumptions to invoke, and what data to include in their analyses. Even using "no need for tuning" as a criterion for method selection or prioritizing bias, for example, or mean squared error, is a subjective decision. Settings that appear completely mechanical involve choice: for example, if a researcher has a checklist saying to apply linear regression for continuous data, logistic regression for binary data, and Poisson regression for count data, he or she still has the option to code a response as continuous or to use a threshold to define a binary classification. And such choices can be far from trivial; for example, when modeling elections or sports outcomes, one can simply predict the winner or instead predict the numerical point differential or vote margin. Modeling the binary outcome can be simpler to explain but in general will throw away information,

and subjective judgment arises in deciding what to do in this sort of problem (Gelman, 2013a). And in both classical and Bayesian statistics, subjective choices arise in defining the sample space and considering what information to condition on.

2.2. Objectivity, subjectivity, and quantification in scientific measurement

Another issue connected to objectivity and subjectivity relevant to statisticians has to do with where the data to be analyzed come from. There is an ideology widespread in many areas of science that sees quantification and numbers and their statistical analysis as key tools for objectivity. An important function of quantitative scientific measurement is the production of observations that are thought of as independent of individual points of view. But, even apart from the generally difficult issue of measurement validity, the focus on what can be quantified can narrow down what can be observed, and may not necessarily do the measured entities justice.

The social sciences have seen endless arguments over the relative importance of objective conditions and what Keynes (1936) called "animal spirits." In macroeconomics, for example, the debate has been between the monetarists who tend to characterize recessions as necessary consequences of underlying economic conditions (as measured, for example, by current account balances, business investment, and productivity), and the Keynesians who focus on more subjective factors such as stock market bubbles and firms' investment decisions. These disagreements also turn methodological, with much dispute, for example, over the virtues and defects of various attempts to objectively measure the supply and velocity of money, or consumer confidence, or various other inputs to economic models. In psychology, there is a big effort to scientifically measure personality traits and subjective states. For example, Kahneman (1999) defines "objective happiness" as "the average of utility over a period of time." Whether or not this definition makes much sense, it illustrates a movement in the social and behavioral sciences to measure, in supposedly objective manners, what might previously have been considered unmeasurable. Another example is the use of quantitative indicators for human rights in different countries; although it has been argued that it is of major importance that such indicators should be objective to have appropriate impact on political decision making (Candler et al., 2011), many aspects of their definition and methodology are subject to controversy and reflect specific political interests and views (Merry, 2011), and we think that it will help the debate to communicate such indicators transparently together with their limitations and the involved decisions rather than to sell them as objective and unquestionable.

Connected to quantification as a means of objectification is an attitude to statistics of many researchers in various areas who use standard routines in statistical software without much understanding of how the methods' assumptions and motivation relate to their specific research problem, in the expectation that the software can condense their research into a single summary (most often a p-value) that "objectifies" their results. This idea of objectivity is in stark contrast with the realisation by many of these researchers at some point that depending on individual inventiveness there are many ways to arrive at such a number.

See Porter (1996), Desrosieres (2002), Douglas (2009) for more discussion of the connection between quantification and objectivity. As with choices in statistical modeling and analysis, we believe that when considering measurement the objective/subjective antagonism is less helpful than a more detailed discussion of what quantification can achieve and what its limitations are.

2.3. Our attitude toward objectivity and subjectivity in science

Many users of the terms "objective" and "subjective" in discussions concerning statistics do not acknowledge that these terms are quite controversial in the philosophy of science (as is "realism")

and that they are used with a variety of different meanings and are therefore prone to misunderstandings. An overview is given in the Appendix.

The attitude taken in the present paper is based on Hennig (2010). According to this perspective, human inquiry starts from observations that are made by personal observers ("personal reality"). Through communication, people share observations and generate "social realities" that go beyond a personal point of view. These shared realities include for example measurement procedures that standardize observations, and mathematical models that connect observations to an abstract formal system that is meant to create a thought system cleaned from individually different point of views. Nevertheless, human beings only have access to "observer-independent reality" through personal observations and how these are brought together in social reality.

Science aims at arriving at a view of reality that is stable and reliable and can be agreed freely by general observers and is therefore as observer-independent as possible. In this sense we see objectivity as a scientific ideal. But at the same time we acknowledge what gave rise to the criticism of objectivity: the existence of different individual perspectives and also of perspectives that differ between social systems, and therefore the ultimate inaccessibility of a reality that is truly independent of observers, is a basic human condition. Objectivity can only be attributed by observers, and if observers disagree about what is objective, there is no privileged position from which this can be decided. Ideal objectivity can never be achieved.

How to resolve scientific disputes by scientific means without throwing up our hands and giving up on the possibility of scientific consensus is a key problem, and science should be guided by principles that at the same time aim at stable and reliable consensus as usually associated with "objectivity" while remaining open to a variety of perspectives, often associated with "subjectivity," exchange between which is needed in order to build a stable and reliable scientific world view.

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. This perspective allows an optimistic rendition of the pessimistic induction, which celebrates the fact that we can be successful in science without even knowing the truth. The standard realist argument from success to truth is shown to be ill-defined and flawed." Or, more informally "Reality is that which, when you stop believing in it, doesn't go away" (Dick, 1981). 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. This means that scientific methodology has to be assessed relative to the specific aims and actions connected to its use.

Because science aims at agreement, communication is central to science, as are transparency and techniques for supporting the clarity of communication. Among these techniques are formal and mathematical language, standardized measurement procedures, and scientific models. Such techniques provide a basis for scientific discussion and consensus, but at the same time the scientific consensus should not be based on authority and it always needs to be open to new points of view that challenge an established consensus. Therefore, in science there is always a tension between the ideal of general agreement and the reality of heterogeneous perspectives, and the virtues listed in Section 3 are meant to help statisticians navigating this tension.

3. Our proposal

In order to move the conversation toward principles of good science, we propose to replace, where ever possible, the words "objectivity" and "subjectivity" with broader collections of attributes, namely by *transparency*, *consensus*, *impartiality*, *correspondence to observable reality*, and *stability*, all related to objectivity; awareness of *multiple perspectives* and *context dependence*, related to subjectivity; and *investigation of stability*, related to both.

The advantage of this reformulation is that the replacement terms do not oppose each other. Instead of debating over whether a given statistical method is subjective or objective (or normatively debating the relative merits of subjectivity and objectivity in statistical practice), we can recognize attributes such as transparency and acknowledgment of multiple perspectives as complementary.

3.1. "Transparency," "consensus," "impartiality," and "correspondence to observable reality," instead of "objectivity"

Science is practiced by human beings, who only have access to the real world through interpretation of their perceptions. Taking objectivity seriously as an ideal, scientists need to make the sharing of their perceptions and interpretations possible. When applied to statistics, the implication is that choices in data analysis (including the prior distribution, if any, but also the model for the data, methodology, and the choice of what information to include in the first place) should be motivated based on factual, externally verifiable information and transparent criteria. This is similar to the idea of the concept of "institutional decision analysis" (Section 9.5 of Gelman, Carlin, et al., 2013), under which the mathematics of formal decision theory can be used to ensure that decisions can be justified based on clearly-stated criteria. Different stakeholders will disagree on decision criteria, and different scientists will differ on statistical modeling decisions, so, in general, there is no unique "objective" analysis, but we can aim at communicating and justifying analyses in ways that support scrutiny and eventually consensus. Similar thoughts have motivated the slogan "transparency is the new objectivity" in journalism (Weinberger, 2009).

In the context of statistical analysis, a key aspect of objectivity is therefore a process of *transparency*, in which the choices involved are justified based on external, potentially verifiable sources or at least transparent considerations (ideally accompanied by sensitivity analyses if such considerations leave alternative options open), a sort of "paper trail" leading from external information, through modeling assumptions and decisions about statistical analysis, all the way to inferences and decision recommendations. The current push of some journals to share data and computer code and the advent of tools to better organize code and projects such as Github and version control goes in this direction. Transparency also comprises spelling out explicit and implicit assumptions about the data production, some of which may be unverifiable.

But transparency is not enough. Science aims at stable *consensus* in potentially free exchange (see Section 2.3), which is one reason that the current crisis of non-replication is taken so seriously in psychology (Yong, 2012). Transparency contributes to this building of consensus by allowing scholars to trace the sources and information used in statistical reasoning (Gelman and Basbøll, 2013). Furthermore, scientific consensus, as far as it deserves to be called "objective," requires rationales, clear arguments, and motivation, along with elucidation of how this relates to already existing knowledge. Following generally accepted rules and procedures counters the dependence of results on the personalities of individual researchers, although there is always a danger that such generally accepted rules and procedures are inappropriate or suboptimal for the specific situation at hand. For such reasons, one might question the inclusion of consensus as a virtue. Its importance

stems from the impossibility to access observer-independent reality which means that exchange between observers is necessary to find out about what can be taken as real and stable. Consensus cannot be enforced; as a virtue it refers to behavior that facilitates consensus.

In any case, consensus can only be achieved if researchers attempt to be *impartial* by taking into account competing perspectives, avoiding to favor pre-chosen hypotheses, and being open to criticism. In the context of epidemiology, Greenland (2012) proposes transparency and neutrality as replacements for objectivity.

Going on, the world outside the observer's mind plays a key role in usual concepts of objectivity, and as explained in Section 2.3 we see it as a major target of science. We acknowledge that the "real world" is only accessible to human beings through observation, and that scientific observation and measurement cannot be independent of human preconceptions and theories. As statisticians we are concerned with making general statements based on systematized observations, and this makes *correspondence to observed reality* a core concern regarding objectivity. This is not meant to imply that empirical statements about observations are the only meaningful ones that can be made about reality; we think that scientific theories that cannot be verified (but potentially be falsified) by observations are meaningful thought constructs, particularly because observations are never "pure" and truly independent of thought constructs. Certainly in some cases the measurements, i.e., the observations the statistician deals with, require critical scrutiny before discussing any statistical analysis of them, see Section 2.2.

Formal statistical methods contribute to objectivity as far as they contribute to the fulfillment of these desiderata, particularly by making procedures and their implied rationales transparent and unambiguous.

For example, Bayesian statistics is commonly characterized as "subjective" by Bayesians and non-Bayesians alike. But depending on how exactly prior distributions are interpreted and used (see Sections 5.3–5.5), they fulfill or aid some or all of the virtues listed above. Priors make the researchers' prior point of view transparent; different approaches of interpreting them provide different rationales for consensus; "objective Bayesians" (see Section 5.4) try to make them impartial; and if suitably interpreted (see Section 5.5) they can be properly grounded in observations.

3.2. "Multiple perspectives" and "context dependence," instead of "subjectivity"

Science is normally seen as striving for objectivity, and therefore acknowledging subjectivity can be awkward. But as noted above already, reality and the facts are only accessible through individual personal experiences. Different people bring different information and different viewpoints to the table, and they will use scientific results in different ways. In order to enable clear communication and consensus, differing perspectives need to be acknowledged, which contributes to transparency and thus to objectivity. Therefore, subjectivity is important to the scientific process. Subjectivity is valuable in statistics in that it represents a way to incorporate the information coming from differing perspectives, which are the building blocks of scientific consensus.

We propose *awareness of multiple perspectives* and *context dependence* as key virtues making explicit the value of subjectivity. To the extent that subjectivity in statistics is a good thing, it is because information truly is dispersed, and, for any particular problem, different stakeholders have different goals. A counterproductive implication of the idea that science should be "objective" is that there is a tendency in the communication of statistical analyses to either avoid or hide decisions that cannot be made in an automatic, seemingly "objective" fashion by the available data. Given that all observations of reality depend on the perspective of an observer, interpreting science as striving for a unique ("objective") perspective is illusory. Multiple perspectives are a

reality to be reckoned with and should not be hidden. Furthermore, by avoiding personal decisions, researchers often waste opportunities to adapt their analyses appropriately to the context, the specific background and their specific research aims, and to communicate their perspective more clearly. Therefore we see the acknowledgment of multiple perspectives and context dependence as virtues, making clearer in which sense subjectivity can be productive and helpful.

The term "subjective" is often used to characterize aspects of certain statistical procedures that cannot be derived automatically from the data to be analyzed, such as Bayesian prior distributions, tuning parameters (for example, the proportion of trimmed observations in trimmed means, or the threshold in wavelet smoothing), or interpretations of data visualisation. Such decisions are entry points for multiple perspectives and context dependence. The first decisions of this kind are typically the choice of data to be analyzed and the family of statistical models to be fit.

To connect with the other part of our proposal, the recognition of different perspectives should be done in a transparent way. We should not say we set a tuning parameter to 2.5 (say) just because that is our belief. Rather, we should justify the choice explaining clearly how it supports the research aims. This could be by embedding the choice in a statistical model that can ultimately be linked back to observable reality and empirical data, or by reference to desirable characteristics (or avoidance of undesirable artifacts) of the methodology given the use of the chosen parameter; actually, many tuning parameters are related to such characteristics and aims of the analysis rather than to some assumed underlying "belief" (see Section 4.3). In many cases, such a justification may be imprecise, for example because background knowledge may be only qualitative and not quantitative or not precise enough to tell possible alternative choices apart, but often it can be argued that even then conscious tuning or specification of a prior distribution comes with benefits compared to using default methods of which the main attraction often is that seemingly "subjective" decisions can be avoided.

To consider an important example, regularization requires such decisions. Default priors on regression coefficients are used to express the belief that coefficients are typically close to zero, and from a non-Bayesian perspective, lasso shrinkage can be interpreted as encoding an external assumption of sparsity. Sparsity assumptions can be connected to an implicit or explicit model in which problems are in some sense being sampled from some distribution or probability measure of possible situations; see Section 5.5. This general perspective (which can be seen as Bayesian with an implicit prior on states of nature, or classical with an implicit reference set for the evaluation of statistical procedures) provides a potential basis to connect choices to experience; at least it makes transparent what kind of view of reality is encoded in the choices.

Tibshirani (2014) writes that enforcing sparsity is not primarily motivated by beliefs about the world, but rather by benefits such as computability and interpretability, hinting at the fact that considerations other than being "close to the real world" often play an important role in statistics and more generally in science. Even in areas such as social science where no underlying truly sparse structure exists, imposing sparsity can have advantages such as supporting stability (Gelman, 2013).

In a wider sense, if one is performing a linear or logistic regression, for example, and considering options of maximum likelihood, lasso, or hierarchical Bayes with a particular structure of priors, all of these choices are "subjective" in the sense of encoding aims regarding possible outputs and assumptions, and all are "objective" as far as these aims and assumptions are made transparent and the assumptions can be justified based on past data and ultimately be checked given enough future data. So the conventional labeling of Bayesian analyses or regularized estimates as "subjective" misses the point.

For another example, the binomial-data confidence interval based on $(y+2)/(n+4)$ gives better

coverage than the classical interval based on y/n (Agresti and Coull, 1998). Whereas the latter has a straightforward justification, the former is based on trading interval width against conservatism and involves some approximation and simplification, which the authors justify by the fact that the resulting formula can be presented in elementary courses. Debating whether this is more subjective than the classical approach, and whether this is a problem, is not helpful. Similarly, when comparing Bayesian estimates of public opinion using multilevel regression and poststratification to taking raw survey means (which indeed correspond to Bayesian analyses under unreasonable flat priors), it is irrelevant which is considered more subjective.

Tuning parameters can be set or estimated based on past data, and also based on understanding of the impact of the choice on results and a clear explanation why a certain impact is desired or not. In robust statistics, for example, the breakdown point of some methods can be tuned and may be chosen lower than the optimal 50%, because if there is a too large percentage of data deviating strongly from the majority, one may rather want the method to deliver a compromise between all observations, but if the percentage of outliers is quite low, one may rather want them to be disregarded, with borderline percentages depending on the application (particularly on to what extent outliers are interpreted as erroneous observations rather than as somewhat special but still relevant cases).

Here is an example in which awareness of multiple perspectives can help with a problem with impartiality. Simulation studies for comparing statistical methods are often run by the designers of one of the competing approaches, and even if this is not the case, the person running the study may have prior opinions about the competitors that may affect the study. There is simply no "objective" way how this can be avoided; taking into account multiple perspectives by for example asking designers of all competing methods to provide simulation setups might help here.

3.3. Stability

As outlined in Section 2.3, we believe that science aims at a stable and reliable view of reality. Human beings do not have direct access to observer-independent reality, but phenomena that remain stable when perceived through different channels, at different times, and that are confirmed as stable by different observers, are the best contenders to be attributed "objective existence."

The term *stability* is hard to find in philosophical accounts of objectivity, but it seems that Mayo's (1996) view of the growth of experimental knowledge through piecemeal testing of aspects of scientific theories and learning from error (which to her is a key feature of objectivity) implicitly aims at probing the stability of these theories. Stability is also connected to subjectivity in the sense that in the best case stability persists under inquiry from as many perspectives and in as many contexts as possible.

The accommodation and analysis of variability is something that statistical modeling brought to science, and in this sense statisticians investigate stability (of observations as well as of the statistics and estimators computed from them) all the time. An investigation of stability just based on variability assuming a parametric model is quite narrow, though, and there are many further sources of potential instabilities. Stability can refer to reproducibility of conclusions on new data, or to alternative analyses of the same data making different choices regarding for example tuning constants, Bayesian priors, transformations, resampling, removing outliers, or even completely different methodology as far as this aims at investigating the same issue (alternative analyses that can be interpreted as doing something essentially different cannot expected to deliver a similar result). On the most basic (but not always trivial) level, the same analysis on the same data should be replicable by different researchers. In statistical theory, basic variability assuming a

parametric model can be augmented by robustness against various violations of model assumptions and Bayesian sensitivity analysis.

There are many aspects of stability that can be investigated, and only so much can be expected from a single study or publication; the generation of reliable scientific knowledge generally requires investigation of phenomena from more points of view than that of a single researcher or team.

3.4. A list of specific virtues

To summarize the above discussion, here is a more detailed list of the virtues discussed above, which we think will improve on discussions in which approaches, analyses and arguments are branded "subjective" or "objective":

1. Transparency:
   (a) Clear and unambiguous definitions of concepts,
   (b) Open planning and following agreed protocols,
   (c) Full communication of reasoning, procedures, spelling out of (potentially unverifiable) assumptions and potential limitations;
2. Consensus:
   (a) Accounting for relevant knowledge and existing related work,
   (b) Following generally accepted rules where possible and reasonable,
   (c) Provision of rationales for consensus and unification;
3. Impartiality:
   (a) Thorough consideration of relevant and potentially competing theories and points of view,
   (b) Thorough consideration and if possible removal of potential biases: factors that may jeopardize consensus and the intended interpretation of results,
   (c) Openness to criticism and exchange;
4. Correspondence to observable reality:
   (a) Clear connection of concepts and models to observables,
   (b) Clear conditions for reproduction, testing, and falsification;
5. Awareness of multiple perspectives;
6. Awareness of context dependence:
   (a) Recognition of dependence on specific contexts and aims,
   (b) Honest acknowledgment of the researcher's position, goals, experiences, and subjective point of view;
7. Investigation of stability:
   (a) Consequences of alternative decisions and assumptions that could have been made in the analysis,
   (b) Variability and reproducibility of conclusions on new data.

In the subsequent discussion we refer to the item numbers in the above list starting by V for virtue, such as V4b for "clear conditions for reproduction, testing, and falsification."

We are aware that in some situations some of these virtues may oppose each other, for example "consensus" can contradict "awareness of multiple perspectives," and indeed dissent is essential to scientific progress. This tension between impersonal consensus and creative debate is an unavoidable aspect of science. Sometimes the consensus can only be that there are different legitimate points of view. Furthermore, the listed virtues are not all fully autonomous; clear reference to observations may be both a main rationale for consensus and a key contribution to transparency; and the subjective virtues contribute to both transparency and openness to criticism and exchange.

Not all items on the list apply to all situations. For example, in Section 5 we apply the list to the foundations of statistics, but some virtues (such as full communication of procedures) rather apply to specific studies.

4. Applied examples

In conventional statistics, assumptions are commonly minimized. Classical statistics and econometrics is often framed in terms of robustness, with the goal being methods that work with minimal assumptions. But the decisions about what information to include and how to frame the model—these are typically buried, not stated formally as assumptions but just baldly stated: "Here is the analysis we did ...," sometimes with the statement or implication that these have a theoretical basis but typically with little clear connection between subject-matter theory and details of measurements. From the other perspective, Bayesian analyses are often boldly assumption-based but with the implication that these assumptions, being subjective, need no justification and cannot be checked from data.

We would like statistical practice, Bayesian and otherwise, to move toward more transparency regarding the steps linking theory and data to models, and recognition of multiple perspectives in the information that is included in this paper trail and this model. In this section we show how we are trying to move in this direction in some of our recent research projects. We present these examples not as any sort of ideals but rather to demonstrate how we are grappling with these ideas and, in particular, the ways in which active awareness of the concepts of transparency, consensus, impartiality, correspondence to observable reality, multiple perspectives and context dependence is changing our applied work.

4.1. A hierarchical Bayesian model in pharmacology

Statistical inference in pharmacokinetics/pharmacodynamics involves many challenges: data are indirect and often noisy; the mathematical models are nonlinear and computationally expensive, requiring the solution of differential equations; and parameters vary by person but often with only a small amount of data on each experimental subject. Hierarchical models and Bayesian inference are often used to get a handle on the many levels of variation and uncertainty (see, for example, Sheiner, 1984, and Gelman, Bois, and Jiang, 1996).

One of us is currently working on a project in drug development involving a Bayesian model that was difficult to fit, even when using advanced statistical algorithms and software. Following the so-called folk theorem of statistical computing (Gelman, 2008), we suspected that the problems with computing could be attributed to a problem with our statistical model. In this case, the issue did not seem to be lack of fit, or a missing interaction, or unmodeled measurement error—problems we had seen in other settings of this sort. Rather, the fit appeared to be insufficiently constrained,

with the Bayesian fitting algorithm being stuck going through remote regions of parameter space that corresponded to implausible or unphysical parameter values.

In short, the model as written was only weakly identified, and the given data and priors were consistent with all sorts of parameter values that did not make scientific sense. Our iterative Bayesian computation had poor convergence—that is, the algorithm was having difficulty approximating the posterior distribution—and the simulations were going through zones of parameter space that were not consistent with the scientific understanding of our pharmacology colleagues.

To put it another way, our research team had access to prior information that had not been included in the model. So we took the time to specify a more informative prior. The initial model thus played the role of a placeholder or default which could be elaborated as needed, following the iterative prescription of falsificationist Bayesianism (Box, 1980, Gelman et al., 2013, Section 5.5).

In our experience, informative priors are not so common in applied Bayesian inference, and when they are used, they often seem to be presented without clear justification. In this instance, though, we decided to follow the principle of transparency and write a note explaining the genesis of each prior distribution. To give a sense of what we're talking about, we present a subset of these notes here:

- γ_1: mean of population distribution of $\log(\text{BVA}_j^{\text{latent}}/50)$, centered at 0 because the mean of the BVA values in the population should indeed be near 50. We set the prior sd to 0.2 which is close to $\log(60/50) = 0.18$ to indicate that we're pretty sure the mean is between 40 and 60.
- γ_2: mean of pop dist of $\log(k_j^{\text{in}}/k_j^{\text{out}})$, centered at 3.7 because we started with -2.1 for k^{in} and -5.9 for k^{out}, specified from the literature about the disease. We use a sd of 0.5 to represent a certain amount of ignorance: we're saying that our prior guess for the population mean of $k^{\text{in}}/k^{\text{out}}$ could easily be off by a factor of $\exp(0.5) = 1.6$.
- γ_3: mean of pop dist of $\log k_j^{\text{out}}$, centered at -5.8 with a sd of 0.8, which is the prior that we were given before, from the time scale of the natural disease progression.
- γ_4: $\log E_{\text{max}}^0$, centered at 0 with sd 2.0 because that's what we were given earlier.

The γ's here already represent a transformation of the original parameters, BVA (baseline visual acuity; this is a drug for treating vision problems), k_{in} and k_{out} (rate constants for differential equations that model the diffusion of the drug), and E_{max}^0, a saturation parameter in the model. One goal in this sort of work is to reparameterize to unbounded scales (so that normal distributions are more reasonable, and we can specify parameters based on location and scale) and to aim for approximate independence in the prior distribution because of the practical difficulties of eliciting prior correlations. The "literature about the disease" comes from previously published trials of other drugs for this disease; these trials also include control arms which give us information on the natural progression of visual acuity in the absence of any treatment.

We see this sort of painfully honest justification as a template for future Bayesian data analyses. The above snippet certainly does not represent an exemplar of best practices, but we see it as a "good enough" effort that presents our modeling decisions in the context in which they were made.

To label this prior specification as "objective" or "subjective" would miss the point. Rather, we see it as having some of the virtues of objectivity and subjectivity—notably, transparency (V1) and some aspects of consensus (V2) and awareness of multiple perspectives (V5)—while recognizing its clear imperfections and incompleteness. Other desirable features would derive from other aspects of the statistical analysis—for example, we use external validation to approach correspondence to observable reality (V4), and our awareness of context dependence (V6) comes from the placement of our analysis within the larger goal, which is to model dosing options for a particular drug.

One concern about our analysis which we have not yet thoroughly addressed is sensitivity to model assumptions. We have established that the prior distribution makes a difference but it is possible that different reasonable priors yield posteriors with greatly differing real-world implications, which would raise concern about consensus (V2) and impartiality (V3). Our response to such concerns, if this sensitivity is indeed a problem, would be to more carefully document our choice of prior, thus doubling down on the principle of transparency (V1) and to compare to other possible prior distributions supported by other information, thus supporting impartiality (V3) and awareness of multiple perspectives (V5).

The point is not that our particular choices of prior distributions are "correct" (whatever that means), or optimal, or even good, but rather that they are transparent, and in a transparent way connected to knowledge. Subsequent researchers—whether supportive, critical, or neutral regarding our methods and substantive findings—should be able to interpret our priors (and, by implication, our posterior inferences) as the result of some systematic process, a process open enough that it can be criticized and improved as appropriate.

4.2. Adjustments for pre-election polls

Wang et al. (2014) describe another of our recent applied Bayesian research projects, in this case a statistical analysis that allows highly stable estimates of public opinion by adjustment of data from non-random samples. The particular example used was an analysis of data from an opt-in survey conducted on the Microsoft Xbox video game platform, a technique that allowed the research team to, effectively, interview respondents in their living rooms, without ever needing to call or enter their houses.

The Xbox survey was performed during the two months before the 2012 U.S. presidential election. In addition to offering the potential practical benefits of performing a national survey using inexpensive data, this particular project made use of its large sample size and panel structure (repeated responses on many thousands of Americans) to learn something new about U.S. politics: we found that certain swings in the polls, which had been generally interpreted as representing large swings in public opinion, actually could be attributed to differential nonresponse, with Democrats and Republicans in turn being more or less likely to respond during periods where there was good or bad news about their candidate. This finding was consistent with some of the literature in political science (see Erikson, Panagopoulos, and Wlezien, 2004), but the Xbox study represented an important empirical confirmation (Gelman, Goel, et al., 2016).

Having established the potential importance of the work, we next consider its controversial aspects. For many decades, the gold standard in public opinion research has been probability sampling, in which the people being surveyed are selected at random from a list or lists (for example, selecting households at random from a list of addresses or telephone numbers and then selecting a person within each sampled household from a list of the adults who live there). From this standpoint, opt-in sampling of the sort employed in the Xbox survey lacks a theoretical foundation, and the estimates and standard errors thus obtained (and which we reported in our research papers) do not have a clear statistical interpretation.

This criticism—that inferences from opt-in surveys lack a theoretical foundation—is interesting to us here because it is *not* framed in terms of objectivity or subjectivity. We do use Bayesian methods for our survey adjustment but the criticism from certain survey practitioners is not about adjustment but rather about the data collection: they take the position that no good adjustment is possible for data collected from a non-probability sample.

As a practical matter, our response to this criticism is that nonresponse rates in national

random-digit-dialed telephone polls are currently in the range of 90%, which implies that real-world surveys of this sort are essentially opt-in samples in any case: If there is no theoretical justification for non-random samples then we are all dead, which leaves us all with the choice to either abandon statistical inference entirely when dealing with survey data, or to accept that our inferences are model-based and do our best (Gelman, 2014c).

Our Bayesian adjustment model (Wang et al., 2014) used prior information in two ways. First, population distributions of demographics, state of residence, and party identification were imputed using exit poll data from the previous election; from the survey sampling perspective this was a poststratification step, and from the political science perspective this represents an assumption of stability in the electorate from 2008 to 2012. The second aspect of prior information was encoded in our hierarchical logistic regression model, in which varying intercepts for states and for different demographic factors were modeled as exchangeable batches of parameters drawn from normal distributions. These assumptions are necessarily approximate and are thus ultimately justified on pragmatic grounds.

We shall now express this discussion using the criteria from Section 3.4. Probability sampling has the clear advantage of transparency (V1) in that the population and sampling mechanism can be clearly defined and accessible to outsiders, in a way that an opt-in survey such as the Xbox is not. In addition, the probability sampling has the benefits of consensus (V2), at least in the United States, where such surveys have a long history and are accepted in marketing and opinion research. Impartiality (V3) and correspondence to observable reality (V4) are less clearly present because of the concern with nonresponse, just noted. We would argue that the large sample size and repeated measurements of the Xbox data, coupled with our sophisticated hierarchical Bayesian adjustment scheme, put us well on the road to impartiality (through the use of multiple sources of information, including past election outcomes, used to correct for biases in the form of known differences between sample and observation) and correspondence to observable reality (in that the method can be used to estimate population quantities that could be validated from other sources).

Regarding the virtues associated with subjectivity, the various adjustment schemes represent awareness of context dependence (V6) in that the choice of variables to match in the population depend on the context of political polling, both in the sense of which aspects of the population are particularly relevant for this purpose, and in respecting the awareness of survey practitioners of what variables are predictive of nonresponse. The researcher's subjective point of view is involved in the choice of exactly what information to include in weighting adjustments and exactly what statistical model to fit in regression-based adjustment. Users of probability sampling on grounds of "objectivity" may shrink from using such judgments, and may therefore ignore valuable information from the context.

4.3. Transformation of variables in cluster analysis for socioeconomic stratification

Cluster analysis aims at grouping together similar objects and separating dissimilar ones, and as such is based, explicitly or implicitly, on some measure of dissimilarity. Defining such a measure, for example using some set of variables characterizing the objects to be clustered, can involve many decisions. Here we consider an example of Hennig and Liao (2013), where we clustered data from the 2007 U.S. Consumer Finances Survey, comprising variables on income, savings, housing, education, occupation, number of checking and savings accounts, and life insurance with the aim of data-based exploration of socioeconomic stratification. The choice of variables and the decisions of how they are selected, transformed, standardized, and weighted has a strong impact on the results of the cluster analysis. This impact depends to some extent on the clustering technique that

is afterwards applied to the resulting dissimilarities, but will typically be considerable, even for cluster analysis techniques that are not directly based on dissimilarities. One of the various issues discussed by Hennig and Liao (2013) was the transformation of the variables treated as continuous (namely income and savings amount), with the view of basing a cluster analysis on a Euclidean distance after transformation, standardization, and weighting of variables.

There is some literature on choosing transformations, but the usual aims of transformation, namely achieving approximate additivity, linearity, equal variances, or normality, are often not relevant for cluster analysis, where such assumptions only apply to model-based clustering, and only within the clusters, which are not known before transformation.

The rationale for transformation when setting up a dissimilarity measure for clustering is of a different kind. The measure needs to formalize appropriately which objects are to be treated as "similar" or "dissimilar" by the clustering methods, and should therefore be put into the same or different clusters, respectively. In other words, the formal dissimilarity between objects should match what could be called the "interpretative dissimilarity" between objects. This is an issue involving subject-matter knowledge that cannot be decided by the data alone.

Hennig and Liao (2013) argue that the interpretative dissimilarity between different savings amounts is governed rather by ratios than by differences, so that \$2 million of savings is seen as about as dissimilar from \$1 million, as \$2,000 is dissimilar from \$1,000. This implies a logarithmic transformation. We do not argue that there is a precise argument that privileges the log transformation over other transformations that achieve something similar, and one might argue from intuition that even taking logs may not be strong enough. We therefore recognize that any choice of transformation is a provisional device and only an approximation to an ideal "interpretative dissimilarity," even if such an ideal exists.

In the dataset, there are no negative savings values as there is no information on debts, but there are many people who report zero savings, and it is conventional to kludge the logarithmic transformation to become $x \mapsto \log(x + c)$ with some $c > 0$. Hennig and Liao then point out that, in this example, the choice of c has a considerable impact on clustering. The number of people with very small but nonzero savings in the dataset is rather small. Setting $c = 1$, for example, the transformation creates a substantial gap between the zero savings group and people with fairly low (but not very small) amounts of savings, and of course this choice is also sensitive to scaling (for example, savings might be coded in dollars, or in thousands of dollars). The subsequent cluster analysis (done by "partitioning around medoids"; Kaufman and Rousseeuw, 1990) would therefore separate the zero savings group strictly; no person with zero savings would appear together in a cluster with a person with nonzero savings. For larger values for c, the dissimilarity between the zero savings group and people with a low savings amount becomes effectively small enough that people with zero savings could appear in clusters together with other people, as long as values on other variables are similar enough.

We do not believe that there is a true value of c. Rather, clusterings arising from different choices of c are legitimate but imply different interpretations. The clustering for $c = 1$ is based on treating the zero savings group as special, whereas the clustering for $c = 200$, say, implies that a difference in savings between 0 and \$100 is taken as not such a big deal (although it is a bigger deal in any case than the difference between \$100 and \$200). Similar considerations hold for issues such as selecting and weighting variables and coding ordinal variables.

It can be frustrating to the novice in cluster analysis that such decisions for which there do not seem to be an objective basis can make such a difference, and there is apparently a strong temptation to ignore the issue and to just choose $c = 1$, which may look natural in the sense that it maps zero onto zero, or even to avoid transformation at all in order to avoid the discussion, so that no obvious

lack of objectivity strikes the reader. Having the aim of socioeconomic stratification in mind, though, it is easy to argue that clusterings that result from ignoring the issue are less desirable and useful than a clustering obtained from making a however imprecisely grounded decision choosing $c > 1$, therefore avoiding either separation of the zero savings group as a clustering artifact or an undue domination of the clustering by people with large savings in case of not applying any transformation at all.

We believe that this kind of tuning problem that cannot be interpreted as estimating an unknown true constant (and does therefore not lend itself naturally to an approach through a Bayesian prior) is not exclusive to cluster analysis, and is often hidden in presentations of data analyses.

Hennig and Liao (2013) pointed out the issue and did some sensitivity analysis about the strength of the impact of the choice of c (V7a). The way we picked c in that paper made clear reference to the context dependence, while being honest that the subject-matter knowledge in this case provided only weak guidelines for making this decision (V6). We were also clear that alternative choices would amount to alternative perspectives rather than being just wrong (V5, V3).

The issue how to foster consensus and to make a connection to observable reality (V2, V4) is of interest, but not treated here.

But it is problematic to establish rationales for consensus that are based on ignoring context and potentially multiple perspectives. There is a tendency in the cluster analysis literature to seek formal arguments for making such decisions automatically (see, for example, Everitt et al., 2011, Section 3.7, on variable weighting; it is hard to find anything systematic in the clustering literature on transformations), trying to optimize "clusterability" of the dataset, or preferring methods that are less sensitive to such decisions, because this amounts to making the decisions implicitly without giving the researchers access to them. In other words, the data are given the authority to determine not only which objects are similar (which is what we want them to do), but also what similarity should mean. The latter should be left to the researcher, although we acknowledge that the data can have a certain impact: for example the idea that dissimilarity of savings amounts is governed by ratios rather than differences is connected to (but not determined by) the fact that the distribution of savings amounts is skewed, with large savings amounts sparsely distributed.

4.4. Testing for homogeneity against clustering

An issue in Hennig and Liao (2013) was whether there is actually any meaningful clustering to be found in the data. Some sociologists suspect that, in many modern democratic societies, stratification may represent no more than a drawing of arbitrary borders through a continuum of socioeconomic conditions. We were interested in what the data have to say on this issue, and chose to address this by running a test of a homogeneity null hypothesis against a clustering alternative (knowing that there is some distance to go between the result of such an analysis and the "desired" sociological interpretation).

Had we been concerned primarily with appearing objective and the ease to achieve a significant result, probably we'd have chosen a likelihood ratio test of the null hypothesis of a standard homogeneity model (in the specific situation this could have been a Gaussian distribution for the continuous variables, an uncorrelated adjacent category ordinal logit model for ordinal variables and a locally independent multinomial model for categorical data) for a single mixture component in the framework of mixture models as provided, e.g., in the LatentGOLD software package (Vermunt and Magidson, 2016).

But even in the absence of meaningful clusters, real data don't follow such clean distributional shapes and therefore large enough datasets (including ours, with $n > 17,000$) will almost always

reject a simple homogeneity model. We therefore set out to build a null model that captured the features of the dataset such as the dependence between variables and marginal distributions of the categorical variables as well as possible, without involving anything that could be interpreted as clustering structure. As opposed to the categorical variables, the marginal distributions of the "continuous" variables such as the transformed savings amount were treated as potentially indicating clustering, and therefore the null model used nonparametric unimodal distributions for them. Data from this null model involving several characteristics estimated from the data could be simulated using parametric bootstrap.

As test statistic we used a cluster validity statistic of the clustering computed on the data, which was not model-based but dissimilarity-based. The idea behind this was that we wanted a test statistic which would measure the degree of clustering, so that we could find out how much "clustering" one could expect to see even if no meaningful clustering was present (under the null model). Actually we computed clusterings for various numbers of clusters. Rather than to somehow define a single p-value from aggregating all these clusterings (or selecting the "best" one), we decided to show a plot of the values of the validity statistic for the different numbers of clusters for the real dataset together with the corresponding results for many datasets simulated from the null model. The result of this showed clearly that a higher level of clustering was found in the real dataset.

In doing this, we deviated from classical significance test logic in several ways, by not using a test statistic that was optimal test against any specific alternative, by not arguing from a single p-value, and by using a null model that relied heavily on the data in order to try as hard as we can to model the data without clustering. Still, in case that the validity statistic values for the real data don't look clearly different from those of the bootstrapped dataset, this can be interpreted as no evidence in the data for real clustering, whereas the interpretation of a clear ("significant") difference depends on whether we can argue convincingly that the null model is as good as it gets at trying to model the data without clustering structure. Setting up a straw man null model for homogeneity and rejecting it would have been easy and not informative. The general principle is discussed in more detail in Hennig and Lin (2015), including real data examples where such a null model could not be rejected, as opposed to a straw man model.

The essence here is that we made quite a number of decisions that opened our analysis more clearly to the charge of "not being objective" than following a standard approach, for the sake of adapting the analysis better to the specific data in hand, and of giving the null hypothesis the best possible chance (the non-rejection of it would have been a non-discovery here; the role of it wasn't to be "accepted" as "true" anyway).

We tried to do good science, though, by checking as impartially and transparently as we could (V1, V3), whether the data support the idea of a real clustering (V4). This involved context dependent judgment (V6) and the transparent choice of a specific perspective (the chosen validity index) among a potential variety (V5), because we were after more qualitative statements than degrees of belief in certain models.

5. Decomposing subjectivity and objectivity in the foundations of statistics

In this section, we use the above list of virtues to revisit aspects of the discussion on fundamental approaches to statistics, for which the terms "subjective" and "objective" typically play a dominant role. We discuss what we perceive to be the major streams of the foundations of statistics, but within each of these streams there exist several different approaches, which we cannot cover completely in such a paper; rather we sketch the streams somewhat roughly and refer to only a single or a few leading authors for details where needed.

Here, we distinguish between interpretations of probability, and approaches for statistical inference. For example, "frequentism" as an interpretation of probability does not necessarily imply that Fisherian or Neyman-Pearson tests are preferred to Bayesian methods, despite the fact that frequentism is more often associated with the former than with the latter.

We shall go through several philosophies of statistical inference, for each laying out the connections we see to the virtues of objectivity and subjectivity outlined in Section 3.4.

Exercising awareness of multiple perspectives, we emphasize that we do not believe that one of these philosophies is the correct or best one, nor do we claim that reducing the different approaches to a single one would be desirable. What is lacking here is not unification, but rather, often, transparency about which interpretation of probabilistic outcomes is intended when applying statistical modeling to specific problems. Particularly, we think that, depending on the situation, both "aleatory" or "epistemic" approaches to modeling uncertainty are legitimate and worthwhile, referring to data generating processes in observer-independent reality on one hand and rational degrees of belief on the other.

We focus on approaches that are conventionally labeled as either Bayesian or frequentist, but we acknowledge that there are important perspectives on statistics that lie outside this traditional divide. Discussing them in detail would be worthwhile but is beyond our focus, and we hope that discussants of our paper will pick up these threads. Examples of other perspectives include *machine learning*, where the focus is on prediction rather than parameter estimation, thus there is more emphasis on correspondence to observable reality (V4) compared to other virtues; *alternative models of uncertainty* such as belief functions, imprecise probabilities, and fuzzy logic that aim to get around some of the limitations of probability theory (most notoriously, the difficulty of distinguishing between "known unknowns" and "unknown unknowns," or risk and uncertainty in the terminology of Knight, 1921); and *exploratory data analysis* (Tukey, 1977), which is sensitive to multiple perspectives (V5) and context dependence (V6), and tries to be more directly connected to the data than if it was mediated by probability models (V4a). Whether avoidance of probability modeling rather contributes to transparency (V1a) is rather problematic because implicit assumptions may not be spelled out (V1c) can be controversial.

5.1. Frequentist probabilities

"Frequentism" as an interpretation of probability refers, in a narrow sense, to the identification of the probability of an event in a certain experiment with a limiting relative frequency of occurrences if the experiment were to be carried out infinitely often in some kind of independent manner. Frequentist statistics is based on evaluating procedures based on a long-term average over a "reference set" of hypothetical replicated data sets. Different choices of reference sets are for example used by Fisher (1955) and Pearson (1955) when discussing permutation or χ^2 tests for 2×2 tables.

In the wider sense, we call probabilities "frequentist" when they formalize observer-independent tendencies or propensities of experiments to yield certain outcomes (see, for example, Gillies, 2000), which are thought of as replicable and yielding a behavior under infinite replication as suggested by what is assumed to be the "true" probability model.

The frequentist mindset locates probabilities in the observer-independent world, so they are in this sense objective (and often called "objective" in the literature, e.g., Kendall, 1949). This, however, doesn't guarantee that frequentist probabilities really exist; an infinite number of replicates cannot exist, and even a finite amount of real replicates will neither be perfectly identical nor perfectly independent. Ultimately the ideally infinite populations of replicates are constructed by the "statistician's imagination" (Fisher, 1955).

The decision to adopt the frequentist interpretation of probability regarding a certain phenomenon therefore requires idealization. It cannot be enforced by observation, and neither is there general enough consensus that this interpretation applies to any specific setup, although it is well discussed and supported in some physical settings such as radioactive decay (V2, V4). Once a frequentist model is adopted, however, it makes predictions about observations that can be checked, so the reference to the observable reality (V4) is clear.

There is some disagreement about whether the frequentist definition of probability is clear and unambiguous (V1a). On one hand, the idea of a tendency of an experiment to produce certain outcomes as manifested in observed and expected relative frequencies seems clear enough. On the other hand, it is difficult to avoid the circularity that would result from referring to independent and identical replicates when defining frequentist probabilities, because the standard definition of the terms "independent" and "identical" assumes a definition of probability already in place (see von Mises, 1957, for a prominent attempt to solve this, and Fine, 1973, for a criticism).

Frequentism implies that, in the observer-independent reality, true probabilities are unique, but there is considerable room for multiple perspectives (V5) regarding the definition of replicable experiments, collectives, or reference sets. The idea of replication is often constructed in a rather creative way. For example, in time series modeling the frequentist interpretation implies an underlying true distribution for every single time point, but there is no way to repeat observations independently at the same time point. This actually means that the effective sample size for time series data would be 1, if replication were not implicitly constructed in the statistical model, for example by assuming independent innovations in ARMA-type models. Such models, or, more precisely, certain aspects of such models, can be checked against the data, but even if such a check does not fail, it is still clear that there is no such thing in observable reality, even approximately, as a marginal "true" frequentist distribution of the value of the time series x_t at fixed t, as implied by the model, because x_t is strictly not replicable.

The issue that useful statistical models require a construction of replication (or exchangeability) on some level by the statistician, is, as we discuss below, not confined to frequentist models. In order to provide a rationale for the essential statistical task of pooling information from many observations to make inference relevant for future observations, all these observations need to be assumed to somehow represent the same process.

The appropriateness of such assumptions in a specific situation can often only be tested in a quite limited way by observations. All kinds of informal arguments can apply about why it is a good or bad idea to consider a certain set of observations (or unobservable implied entities such as error terms and latent variables) as independent and identically distributed frequentist replicates.

Unfortunately, although such an openness to multiple perspectives and potential context-dependence (V6a) can be seen as positive from our perspective, these issues involved in the choices of a frequentist reference set are often not clearly communicated and discussed. The existence of a true model with implied reference set is typically taken for granted by frequentists, motivated at least in part by the desire for objectivity.

5.2. Frequentist inference

This section is about inference from data about characteristics of an assumed true frequentist probability model. Traditionally, this comprises hypothesis tests, confidence intervals and parameter estimators, but is not limited to them, see below.

According to Mayo and Spanos (2010) and Cox and Mayo (2010), a fundamental feature of frequentist inference is the evaluation of error probabilities, i.e., probabilities of wrong decisions.

Traditionally these would be the Type I and Type II errors of Neyman-Pearson hypothesis testing, but the the error-statistical perspective could also apply to other constructs such as errors of sign and magnitude ("Type S" and "Type M" errors; Gelman and Carlin, 2014).

Mayo and co-authors see the ability to learn from error and to test models severely (in such a way that it would be hard for a model to pass a test if it was wrong regarding the specific aspect assessed by a test) against data as a major feature of objectivity, which is made possible by the frequentist interpretation of probability measures as "data generators." In our list of virtues, this feature is captured in V4b (reference to observations: reproduction, testing, falsification). The underlying idea, with which we agree, is that learning from error is a main driving force in science, a lifetime contract between the mode of statistical investigation and its object. This corresponds to Chang's active scientific realism mentioned above.

The error probability characteristics of methods for frequentist inference rely, in general, on model assumptions. In principle, these assumptions can be tested, too, and are therefore, according to Mayo and co-authors, no threat to the objectivity of the account. But this comes with two problems. Firstly, derivations of statistical inference based on error probabilities typically assume the model as fixed and do not account for prior model selection based on the data. This issue has recently attracted some research (for example, Berk et al., 2013), but this still requires a transparent listing of all the possible modeling decisions that could be made (V1b), which often is missing, and which may not even be desirable as long as the methods are used in an exploratory fashion (Gelman and Loken, 2014). Secondly, any dataset can be consistent with many models, which can lead to divergent inferences. Davies (2014) illustrates this with the analysis of a dataset on amounts of copper in drinking water, which can be fitted well by a Gaussian, a double exponential, and a comb distribution, but yields vastly different confidence intervals for the center of symmetry (which is assumed to be the target of inference) under these three models.

Davies suggests that it is misleading to hypothesize models or parameters to be "true." According to Davies, statistical modeling is about approximating the data in the sense that "adequate" models are not rejected by tests based on characteristics of the data the statistician is interested in (allowing for multiple perspectives and context dependence, V5, V6), i.e., they generate data that "look like" the observed data with respect to the chosen characteristics. Regarding these characteristics, according to Davies, there is no essential difference between parameter values and distributional shapes or structural assumptions, and therefore no conceptual separation as in traditional frequentist inference between checking model assumptions and inference about parameters assuming a parametric model to be true. Such an approach is tied to the observations in a more direct way without making metaphysical assumptions about unobservable features of observer-independent reality (V1a, V4). It is frequentist inference in the sense that the probability models are interpreted as "data generators."

Two further streams in frequentist inference are concerned about the restrictivity of parametric model assumptions. Robust statistics explores the stability (V7) of inferences in case that the "true" model is not equal to the nominal model but rather in some neighborhood, and strives to develop methods that are stable in this respect. There are various ways to define such neighborhoods and to measure robustness, so robustness considerations can bring in multiple perspectives (V5) but may cause problems with reaching consensus (V2).

Nonparametric statistics allows to remove bias (V3c) by minimizing assumptions regarding, e.g., distributional shapes (structural assumptions such as independence are still required). In some cases, particularly with small datasets, this has to be afforded by decreased stability (V7).

Overall, there is no shortage of entry points for multiple perspectives (V5) in frequentist inference. This could be seen as something positive, but it runs counter to some extent to the way the

approach is advertised as objective by some of its proponents. Many frequentist analyses could in our opinion benefit from acknowledging honestly their flexibility and the researcher's choices made, many of which cannot be determined by data alone.

5.3. Subjectivist Bayesianism

We call "subjectivist epistemic" the interpretation of probabilities as quantifications of strengths of belief of an individual, where probabilities can be interpreted as derived from, or implementable through, bets that are coherent in that no opponent can cause sure losses by setting up some combinations of bets. From this requirement of coherence, the usual probability axioms follow (V2c). Allowing conditional bets implies Bayes's theorem, and therefore, as far as inference concerns learning from observations about not (yet) observed hypotheses, Bayesian methodology is used for subjectivist epistemic probabilities, hence the term "subjectivist Bayesianism."

A major proponent of subjectivist Bayesianism was Bruno de Finetti (1974). De Finetti was not against objectivity in general. He viewed observed facts as objective, as well as mathematics and logic and certain formal conditions of random experiments such as the set of possible outcomes. But he viewed uncertainty as something subjective and he held that objective (frequentist) probabilities do not exist. He claimed that his subjectivist Bayesianism appropriately takes into account both the objective (see above) and subjective (opinions about unknown facts based on known evidence) components for probability evaluation.

In de Finetti's work the term "prior" refers to all probability assignments using information external to the data at hand, with no fundamental distinction between the "parameter prior" assigned to parameters in a model, and the form of the "sampling distribution" given a fixed parameter, in contrast to common Bayesian practice today, in which the term "prior" is used to refer only to the parameter prior. In the following discussion we shall use the term "priors" in de Finetti's general sense.

Regarding the list of virtues in Section 3.4, de Finetti provides a clear definition of probability (V1a) based on principles that he sought to establish as generally acceptable (V2c). Unlike objectivist Bayesians, subjectivist Bayesians do not attempt to enforce agreement regarding prior distributions, not even given the same evidence; still, de Finetti (1974) and other subjectivist Bayesians proposed rational principles for assigning prior probabilities. There is also some work on (partial) intersubjective agreement on prior specifications, e.g., Dawid (1982a), providing a rationale for consensus (V2c). The difference between the objectivist and subjectivist Bayesian point of view is rooted in the general tension in science explained above; the subjectivist approach can be criticized for not supporting agreement enough—conclusions based on one prior may be seen as irrelevant for somebody who holds another one (V2c)—but can be defended for honestly acknowledging that prior information often does not come in ways that allow a unique formalization (V6b). In any case it is vital that subjectivist Bayesians explain transparently how they arrive at their priors, so that other researchers can decide to what extent they can support the conclusions (V1c).

In de Finetti's conception, probability assessments, prior and posterior, can ultimately only concern observable events, because bets can only be evaluated if the experiment on which a bet is placed has an observable outcome, and so there is a clear connection to observables (V4a).

However, priors in the subjectivist Bayesian conception are not open to falsification (V4b), because by definition they have to be fixed before observation. Adjusting the prior after having observed the data to be analyzed violates coherence. The Bayesian system as derived from axioms such as coherence (as well as those used by objectivist Bayesians; see Section 5.4) is designed to cover all aspects of learning from data, including model selection and rejection, but this requires

that all potential later decisions are already incorporated in the prior, which itself is not interpreted as a testable statement about yet unknown observations. In particular this means that once a coherent subjectivist Bayesian has assessed a setup as exchangeable a priori, he or she cannot drop this assumption later, whatever the data are (think of observing twenty zeroes, then twenty ones, then ten further zeroes in a binary experiment). This is a major problem, because subjectivist Bayesians use de Finetti's theorem to justify working with parameter priors and sampling models under the assumption of exchangeability, which is commonplace in Bayesian statistics. Dawid (1982b) discussed calibration (quality of match between predictive probabilities and the frequency of predicted events to happen) of subjectivist Bayesians inferences, and he suggests that badly calibrated Bayesians could do well to adjust their future priors if this is needed to improve calibration, even at the cost of violating coherence.

Subjectivist Bayesianism scores well on the virtues V5 and V6b. But it is a limitation that the prior distribution exclusively formalizes belief; context and aims of the analysis do not enter unless they have implications about belief. In practice, an exhaustive elicitation of beliefs is rarely feasible, and mathematical and computational convenience often plays a role in setting up subjective priors, despite de Finetti's having famously accused frequentists of "adhockeries for mathematical convenience." Furthermore, the assumption of exchangeability will hardly ever precisely match an individual's beliefs in any situation—even if there is no specific reason against exchangeability in a specific setup, the implicit commitment to stick to it whatever will be observed seems too strong—but some kind of exchangeability assumption is required by Bayesians for the same reason for which frequentists need to rely on independence assumptions: some internal replication in the model is needed to allow generalization or extrapolation to future observations; see Section 5.1.

Summarizing, we view much of de Finetti's criticism of frequentism as legitimate, and subjectivist Bayesianism comes with a commendable honesty about the impact of subjective decisions and allows for flexibility accommodating multiple perspectives. But checking and falsification of the prior is not built into the approach, and this can get in the way of agreement between observers.

5.4. Objectivist Bayesianism

Given the way objectivity is often advertised as a key scientific virtue (often without specifying what exactly it means), it is not surprising that de Finetti's emphasis on subjectivity is not shared by all Bayesians, and that there have been many attempts to specify prior distributions in a more objective way. Currently the approach of E. T. Jaynes (2003) seems to be among the most popular. As with many of his predecessors such as Jeffreys and Carnap, Jaynes saw probability as a generalization of binary logic to uncertain propositions. Cox (1961) proved that given a certain list of supposedly common-sense desiderata for a "plausibility" measurement, all such measurements are equivalent, after suitable scaling, to probability measures. This theorem is the basis of Jaynes' objectivist Bayesianism, and the claim to objectivity comes from postulating that, given the same information, everybody should come to the same conclusions regarding plausibilities: prior and posterior probabilities (V2c), a statement with which subjectivist Bayesians disagree.

In practice, this objectivist ideal seems to be hard to achieve, and Jaynes (2003) admits that setting up objective priors including all information is an unsolved problem. One may wonder whether his ideal is achievable at all. For example, in chapter 21, he gives a full Bayesian "solution" to the problem of dealing with and identifying outliers, which assumes that prior models have to be specified for both "good" and "bad" data (between which therefore there has to be a proper distinction), including parameter priors for both models, as well as a prior probability for any number of observations to be "bad." It is hard to see, and no information about this is provided by

Jaynes himself, how it can be possible to translate the unspecific information of knowing of some outliers in many kinds of situations, some of which are more or less related, but none identical (say) to the problem at hand, into precise quantitative specifications as needed for Jaynes' approach in an objective way, all before seeing the data.

Setting aside the difficulties of working with informally specified prior information, a key issue of objectivist Bayesianism is the specification of an objective prior distribution formalizing the absence of information. Various principles for doing this have been proposed (maximum entropy, Jaynes, 2003; maximum missing Shannon information, Berger et al., 2009; and a set of desirable properties, Bayarri et al., 2012). Such principles have their difficulties and disagree in many cases (Kass and Wasserman, 1996). Objectivity seems to be an ambition rather than a description of what indeed can be achieved by setting up objectivist Bayesian priors. More modestly, therefore, Berger et al. (2009) use the term "reference priors," avoiding the term "objective," and emphasizing that it would be desirable to have a convention for such cases (V2b), but admitting that it may not be possible to prove any general approach for arriving at such a convention uniquely correct or optimal in any rational sense. However, the proposal and discussion of such principles certainly served transparency (V1a,c) and provided rationales for consensus (V2c).

Apart from the issue of the objectivity of the specification of the prior, by and large the objectivist Bayesian approach has similar advantages and disadvantages regarding our list of virtues as its subjectivist cousin. Particularly it comes with the same difficulties regarding the issue of falsifiability from observations. Prior probabilities are connected to logical analysis of the situation rather than to betting rates for future observations as in de Finetti's subjectivist approach, which makes the connection of objectivist Bayesian prior probabilities to observations even weaker than in the subjectivist Bayesian approach (probabilistic logic has applications other than statistical data analysis, for which this may not be a problem).

The merit of objectivist Bayesianism is that the approach comes with a much stronger drive to justify prior distributions in a transparent way using principles that are as clear and general as possible.

5.5. Falsificationist Bayesianism, and frequentist probabilities in Bayesian statistics

For both subjectivist and objectivist Bayesians, probability models including both parameter priors and sampling models do not model the data generating process, but rather represent plausibility or belief from a certain point of view. Plausibility and belief models can be modified by data in ways that are specified a priori, but they cannot be falsified by data.

In much applied Bayesian work, on the other hand, the sampling model is interpreted, explicitly or implicitly, as representing the data-generating process in a frequentist or similar way, and parameter priors and posteriors are interpreted as giving information about what is known about the "true" parameter values. It has been argued that such work does not directly run counter to the subjectivist or objectivist philosophy, because the "true parameter values" can often be interpreted as expected large sample functions given the prior model (Bernardo and Smith, 1994), but the way in which classical subjectivist or objectivist statistical data analysis is determined by the untestable prior assignments is seen as unsatisfactory by many statisticians.

In any case, the frequentist interpretation of a probability distribution as "data generator" is regularly used to investigate how Bayesian analyses perform under such assumptions, theoretically, often by analysis of asymptotic properties or by simulation. Wasserman (2006) called Bayesian methods with good frequentist properties "objective," referring to the "representing things in the observer-independent world"-sense of objectivity, but also providing a connection of Bayesian mod-

els to observables (V4a). Rubin (1984) discussed frequentist approaches for studying the characteristics of Bayesian methods under misspecified models, i.e., stability (V7).

The suggestion of testing aspects of the prior distribution by observations using error statistical techniques has been around for some time (Box, 1980). Gelman and Shalizi (2013) incorporate this in an outline of what we refer to here as "falsificationist Bayesianism," a philosophy that openly deviates from both objectivist and subjectivist Bayesianism, integrating Bayesian methodology with an interpretation of probability that can be seen as frequentist in a wide sense and with an error statistical approach to testing assumptions in a bid to satisfy virtue V4b.

Falsificationist Bayesianism follows the frequentist interpretation of the probabilities formalized by the sampling model given a true parameter, so that these models can be tested using frequentist inference (with the limitations that such techniques have, as discussed in Section 5.2). Gelman and Shalizi argue, as some frequentists do, that such models are idealizations and should not be believed to be literally true, but that the scientific process proceeds from simplified models through test and potential falsification by improving the models where they are found to be deficient.

To put it another way: it is desirable for Bayesian intervals to have close to nominal coverage both conditionally on any observables and unconditionally; the desire for this coverage leads naturally to calibration checks, which in turn motivates the modification or even rejection of models that are not well calibrated empirically. This process serves the correspondence to observable reality (V4) while at the same time putting more of a burden on transparency (V1) and stability (V7) in that the ultimate choice of model can depend on the decision of what aspects of the fitted model will be checked.

A key issue regarding transparency of falsificationist Bayes is how to interpret the parameter prior, which does not usually (if occasionally) refer to a real mechanism that produces frequencies. Major options are firstly to interpret the parameter prior in a frequentist way, as formalizing a more or less idealized data generating process generating parameter values. A bold idealization would be to view "all kinds of potential studies with the (statistically) same parameter" as the relevant population, even if the studies are about different topics with different variables, in which case more realizations exist, but it is hard to view a specific study of interest as a "random draw" from such a population.

Alternatively, the parameter prior may be seen as a purely technical device, serving aims such as regularization, without making any even idealized assumption it corresponds to anything that "exists" in the real world. In this case the posterior distribution does not have a proper direct interpretation, but statistics such as the posterior mean or uncertainty intervals could be interpreted based on their frequentist properties.

Overall, falsificationist Bayesianism combines the virtue of error statistical falsifiability with the virtues V5 and V6 connected to subjectivity. However, the flexibility of the falsificationist Bayesian approach—its openly iterative and tentative nature—creates problems regarding clarity and unification.

6. Discussion

6.1. Implications for statistical theory and practice

At the level of discourse, we would like to move beyond a subjective vs. objective shouting match. But our goals are larger than this. Gelman and Shalizi (2013) on the philosophy of Bayesian statistics sought not just to clear the air but also to provide philosophical and rhetorical space for Bayesians to feel free to check their models and for applied statisticians who were concerned

about model fit to feel comfortable with a Bayesian approach. In the present paper, our goals are for scientists and statisticians to achieve more of the specific positive qualities into which we decompose objectivity and subjectivity in Section 3.4. At the present time, we feel that concerns about objectivity are getting in the way of researchers trying out different ideas and considering different sources of inputs to their model, while an ideology of subjectivity is limiting the degree to which researchers are justifying and understanding their model.

There is a tendency for hardcore believers in objectivity to needlessly avoid the use of valuable external information in their analyses, and for subjectivists, but also for statisticians who want to make their results seem strong and uncontroversial, to leave their assumptions unexamined. We hope that our new framing of transparency, consensus, avoidance of bias, reference to observable reality, multiple perspectives, dependence on context and aims, investigation of stability, and honesty about the researcher's position and decisions will give researchers of all stripes the impetus and, indeed, permission, to integrate different sources of information into their analyses, to state their assumptions more clearly, and to trace these assumptions backward to past data that justify them and forward to future data that can be used to validate them.

Also, we believe that the pressure to appear objective has led to confusion and even dishonesty regarding data coding and analysis decisions which cannot be motivated in supposedly objective ways; see van Loo and Romeijn (2015) for a discussion of this point in the context of psychiatric diagnosis. We prefer to encourage a culture in which it is acceptable to be open about the reasons for which decisions are made, which may at times be mathematical convenience, or the aim of the study, rather than strong theory or hard data. It should be recognized openly that the aim of statistical modeling is not always to make the model as close as possible to observer-independent reality (which always requires idealization anyway), and that some decisions are made, for example, in order to make outcomes more easily interpretable for specific target audiences.

Our key points: (1) multiple perspectives correspond to multiple lines of reasoning, not merely to mindless and unjustified guesses; and (2) what is needed is not just a prior distribution or a tuning parameter, but a statistical approach in which these choices can be grounded, either empirically or by connecting them in a transparent way to the context and aim of the analysis.

For these reasons, *we do not think it at all accurate to limit Bayesian inference to "the analysis of subjective beliefs."* Yes, Bayesian analysis can be expressed in terms of subjective beliefs, but it can also be applied to other settings that have nothing to do with beliefs (except to the extent that all scientific inquiries are ultimately about what is believed about the world).

Similarly, *we would not limit classical statistical inference to "the analysis of simple random samples."* Classical methods of hypothesis testing, estimation, and data reduction can be applied to all sorts of problems that do not involve random sampling. There is no need to limit the applications of these methods to a narrow set of sampling or randomization problems; rather, it is important to clarify the foundation for using the mathematical models for a larger class of problems.

6.2. Beyond "objective" and "subjective"

The list in Section 3.4 is the core of the paper. The list may not be complete, and such a list may also be systematized in different ways. Particularly, we developed the list having particularly applied statistics in mind, and we may have missed aspects of objectivity and subjectivity that are not connected in some sense to statistics. In any case, we believe that the given list can be helpful in practice for researchers, for justifying and explaining their choices, and for recipients of research work, for checking to what extent the listed virtues are practiced in scientific work. A key issue here is transparency, which is required for checking all the other virtues. Another key issue is that

subjectivity in science is not something to be avoided at any cost, but that multiple perspectives and context dependence are actually basic conditions of scientific inquiry, which should be explicitly acknowledged and taken into account by researchers. We think that this is much more constructive than the simple objective/subjective duality.

We do not think this advice represents empty truisms of the "mom and apple pie" variety. In fact, we repeatedly encounter publications in top scientific journals that fall foul of these virtues, which indicates to us that the underlying principles are subtle

Instead of pointing at specific bad examples, here is a list of some common problems (discussed, for example, in Gelman, 2015, and Gelman and Zelizer, 2015), where we believe that exercising one or more of our listed virtues would improve matters:

- Presenting analyses that are contingent on data without explaining the exploration and selection process and without even acknowledging that it took place,
- Justifying decisions by reference to specific literature without acknowledging that what was cited may be controversial, not applicable in the given situation, or without proper justification in the cited literature as well (or not justifying the decisions at all),
- Failure to reflect on whether model assumptions are reasonable in the given situation, what impact it would have if they were violated, or whether alternative models and approaches could be reasonable as well,
- Choosing methods because they do not require tuning or are automatic and therefore seem "objective" without discussing whether the chosen methods can handle the data more appropriately in the given situation than alternative methods with tuning,
- Choosing methods for the main reason that they "do not require assumptions" without realizing that every method is based on implicit assumptions about how to treat the data appropriately, regardless of whether these are stated in terms of statistical models,
- Choosing Bayesian priors without justification or explanation of what they mean and imply,
- Using nonstandard methodology without justifying the deviation from standard approaches (where they exist),
- Using standard approaches without discussion of their appropriateness in a specific context.

Most of these have to do with the unwillingness to admit to having made decisions, to justify them, and to take into account alternative possible views that may be equally reasonable. In some sense perhaps this can be justified based on a sociological model of the scientific process in which each paper presents just one view, and then the different perspectives battle it out. But we think that this idea ignores the importance of communication and facilitating consensus for science. Scientists normally believe that each analysis aims at the truth, and if different analyses give different results, this is not because there are different conflicting truths but rather because different analysts have different aims, perspectives and access to different information. Letting the issue aside of whether it makes sense to talk of the existence of different truths or not, we see aiming at general agreement in free exchange as essential to science, and the more perspectives are taken into account, the more the scientific process is supported.

We see the listed virtues as ideals which in practice cannot generally be fully achieved in any real project. For example, tracing all assumptions to observations and making them checkable by observable data is impossible because one can always ask whether and why results from the specific observations used should generalize to other times and other situations. As mentioned in Section 5.1, ultimately a rationale for treating different situations as "identical and independent" or "exchangeable" needs to be constructed by human thought (people may appeal to historical successes

for justifying such idealizations, but this does not help much regarding specific applications). At some point—but, we hope, not too early—researchers have to resort to somewhat arbitrary choices that can be justified only by logic or convention, if that.

And it is likewise unrealistic to suppose that we can capture all the relevant perspectives on any scientific problem. Nonetheless, we believe it is useful to set these as goals which, in contrast to the inherently opposed concepts of "objectivity" and "subjectivity," can be approached together.

References

Agresti, A., and Coull, B. A. (1998). Approximate is better than exact for interval estimation of binomial proportions. *American Statistician* **52**, 119–126.

Alpert, M., and Raiffa, H. (1984). A progress report on the training of probability assessors. In *Judgment Under Uncertainty: Heuristics and Biases*, ed. Kahneman, D., Slovic, P., and Tversky, A., 294–305. Cambridge University Press.

Bayarri, M. J., Berger, J. O., Forte, A., and García-Donato, G. (2012). Criteria for Bayesian model choice with application to variable selection. *Annals of Statistics* **40**, 1550–1577.

Berger, J. O. (2006). The case for objective Bayesian analysis. *Bayesian Analysis* **1**, 385–402.

Berger, J. O., Bernardo, J. M., and Sun, D. (2009). The formal definition of reference priors. *Annals of Statistics* **37**, 905–938.

Bernardo, J. M., and Smith, A. F. M. (1994). *Bayesian Theory.* Chichester: Wiley.

Berk, R., Brown, L., Buja, A., Zhang, K., and Zhao, L. (2013). Valid post-selection inference. *Annals of Statistics* **41**, 802–837.

Box, G. E. P. (1980). Sampling and Bayes' inference in scientific modelling and robustness. *Journal of the Royal Statistical Society A* **143**, 383—430.

Box, G. E. P. (1983). An apology for ecumenism in statistics. In *Scientific Inference, Data Analysis, and Robustness*, ed. G.E.P. Box, T. Leonard, and C. F. Wu, 51–84. New York: Academic Press.

Candler, J., Holder, H., Hosali, S., Payne, A. M., Tsang T., and Vizard, P. (2011). *Human Rights Measurement Framework: Prototype Panels, Indicator Set and Evidence Base.* Research Report 81. Manchester: Equality and Human Rights Commission.

Chang, H. (2012). *Is Water H_2O? Evidence, Realism and Pluralism.* Dordrecht: Springer.

Cox, D. and Mayo, D. G. (2010). Objectivity and Conditionality in Frequentist Inference. In *Error and Inference*, ed. Mayo, D. G. and Spanos, A., 276–304. Cambridge University Press.

Cox, R. T. (1961). *The Algebra of Probable Inference.* Baltimore: Johns Hopkins University Press.

Daston, L., and Galison, P. (2007). *Objectivity.* New York: Zone Books.

Davies, P. L. (2014). *Data Analysis and Approximate Models.* Boca Raton, Fla.: CRC Press.

Dawid, A. P. (1982a). Intersubjective statistical models. In *Exchangeability in Probability and Statistics*, ed. Koch, G. and Spizichino, F., 217–232. Amsterdam: North Holland.

Dawid, A. P. (1982b). The well-calibrated Bayesian. *Journal of the American Statistical Association* **77**, 605–610.

de Finetti, B. (1974). *Theory of Probability.* New York: Wiley.

Desrosieres, A. (2002). *The Politics of Large Numbers.* Boston: Harvard University Press.

Dick, P. K. (1981). *VALIS.* New York: Bantam Books.

Douglas, H. (2004). The irreducible complexity of objectivity. *Synthese*, **138**, 453–473.

Douglas, H. (2009). *Science, Policy and the Value-Free Ideal.* University of Pittsburgh Press.

Erev, I., Wallsten, T. S., and Budescu, D. V. (1994). Simultaneous over- and underconfidence: The role of error in judgment processes. *Psychological Review* **101**, 519–527.

Erikson, R. S., Panagopoulos, C., and Wlezien, C. (2004). Likely (and unlikely) voters and the assessment of campaign dynamics. *Public Opinion Quarterly* **68**, 588–601.

Everitt, B. S., Landau, S., Leese, M. and Stahl, D. (2011), *Cluster Analysis*, fifth edition. Wiley.

Feyerabend. P. (1978). *Science in a Free Society.* London: New Left Books.

Fine, T. L. (1973). *Theories of Probability.* Waltham, Mass.: Academic Press.

Fisher, R. A. (1955). Statistical methods and scientific induction. *Journal of the Royal Statistical Society B* **17**, 69–78.

Fuchs, S. (1997). A sociological theory of objectivity. *Science Studies* **11**, 4–26.

Gelman, A. (2003). A Bayesian formulation of exploratory data analysis and goodness-of-fit testing. *International Statistical Review* **71**, 369–382.

Gelman, A. (2008). The folk theorem of statistical computing. Statistical Modeling, Causal Inference, and Social Science blog, 13 May. `http://andrewgelman.com/2008/05/13/the_folk_theore/`

Gelman, A. (2013). Whither the "bet on sparsity principle" in a nonsparse world? Statistical Modeling, Causal Inference, and Social Science blog, 25 Feb. `http://andrewgelman.com/2013/12/16/whither-the-bet-on-sparsity-principle-in-a-nonsparse-world/`

Gelman, A. (2014a). Basketball stats: Don't model the probability of win, model the expected score differential. Statistical Modeling, Causal Inference, and Social Science blog, 25 Feb. `http://andrewgelman.com/2014/02/25/basketball-stats-dont-model-probability-win-model-expected-score-differential/`

Gelman, A. (2014b). How do we choose our default methods? 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, 293–301. London: Chapman and Hall.

Gelman, A. (2014c). President of American Association of Buggy-Whip Manufacturers takes a strong stand against internal combustion engine, argues that the so-called "automobile" has "little grounding in theory" and that "results can vary widely based on the particular fuel that is used." Statistical Modeling, Causal Inference, and Social Science blog, `http://andrewgelman.com/2014/08/06/president-american-association-buggy-whip-manufacturers-takes-strong-stand-internal-combustion-engine-argues-called-automobile-little-grounding-theory/`

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

Gelman, A., and Basbøll, T. (2013). To throw away data: Plagiarism as a statistical crime. *American Scientist* **101**, 168–171.

Gelman, A., Bois, F. Y., and Jiang, J. (1996). Physiological pharmacokinetic analysis using population modeling and informative prior distributions. *Journal of the American Statistical Association* **91**, 1400–1412.

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., Carlin, J. B., Stern, H. S., Dunson, D., Vehtari, A., and Rubin, D. B. (2013). *Bayesian Data Analysis*, third edition. London: Chapman and Hall.

Gelman, A., Goel, S., Rivers, D., and Rothschild, D. (2016). The mythical swing voter. *Quarterly Journal of Political Science* **11**, 103–130.

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

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.

Gelman, A., and Shalizi, C. (2013). Philosophy and the practice of Bayesian statistics (with discussion). *British Journal of Mathematical and Statistical Psychology* **66**, 8–80.

Gelman, A., and Zelizer, A. (2015). Evidence on the deleterious impact of sustained use of polynomial regression on causal inference. *Research and Politics* **2**, 1–7.

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

Greenland, S. (2012). Transparency and disclosure, neutrality and balance: Shared values or just shared words? *Journal of Epidemiology and Community Health* **66**, 967–970.

Hacking, I. (2015). Let's not talk about objectivity. In *Objectivity in Science*, ed. F. Padovani et al. Boston Studies in the Philosophy and History of Science.

Hennig, C. (2010). Mathematical models and reality: A constructivist perspective. *Foundations of Science* **15**, 29–48.

Hennig, C., and Liao, T. F. (2013). How to find an appropriate clustering for mixed type variables with application to socioeconomic stratification (with discussion). *Journal of the Royal Statistical Science* **62**, 309–369.

Hennig, C. and Lin, C.-J. (2015). Flexible parametric bootstrap for testing homogeneity against clustering and assessing the number of clusters. *Statistics and Computing* **25**, 821–833.

Huber, P. J., and Ronchetti, E. M. (2009). *Robust Statistics*, second edition. New York: Wiley.

Jaynes, E. T. (2003). *Probability Theory: The Logic of Science.* Cambridge University Press.

Kahneman, D. (1999). Objective happiness. In *Well-being: Foundations of Hedonic Psychology*, 3–25. New York: Russell Sage Foundation Press.

Kass, R. E. and Wasserman, L. (1996). The selection of prior distributions by formal rules. *Journal of the American Statistical Association* **91**, 1343–1370.

Kendall, M. G. (1949). On the reconciliation of theories of probability. *Biometrika* **36**, 101–116.

Keynes, J. M. (1936). *The General Theory of Employment, Interest and Money.* Macmillan.

Knight, F. H. (1921). *Risk, Uncertainty, and Profit.* Boston: Hart, Schaffner and Marx.

Lewis, D. (1980). A subjectivist's guide to objective chance. In *Studies in Inductive Logic and Probability, Volume II*, ed. R. C. Jeffrey, 263-–293. Berkeley: University of California Press.

Linstone, H. A. (1989). Multiple perspectives: Concept, applications, and user guidelines. *Systems Practice* **2**, 307–3331.

Little, R. J. (2012). Calibrated Bayes, an alternative inferential paradigm for official statistics. *Journal of Official Statistics* **28**, 309–334.

MacKinnon, C. (1987). *Feminism Unmodified.* Boston: Harvard University Press.

Maturana, H. R. (1988). Reality: The search for objectivity or the quest for a compelling argument. *Irish Journal of Psychology* **9**, 25–82.

Mayo, D. G. (1996). *Error and the Growth of Experimental Knowledge.* University of Chicago Press.

Mayo, D. G. and Spanos, A. (2010). Introduction and background: The error-statistical philosophy. In *Error and Inference*, ed. Mayo, D. G. and Spanos, A., 15–27. Cambridge University Press.

Megill, A. (1994). Introduction: Four senses of objectivity. In *Rethinking Objectivity*, ed. A. Megill, 1–20. Durham, N.C.: Duke University Press.

Merry, S. E. (2011). Measuring the world: Indicators, human rights, and global governance. *Current Anthropology* **52** (S3), S83–S95.

Pearson, E. S. (1955). Statistical concepts in the relation to reality. *Journal of the Royal Statistical Society B* **17**, 204–207.

Pearson, K. (1911). *The Grammar of Science*. 2007 edition. New York: Cosimo.

Pollster.com (2004). Should pollsters weight by party identification? `http://www.pollster.com/faq/should_pollsters_weight_by_par.php`

Porter, T. M. (1996). *Trust in Numbers: The Pursuit of Objectivity in Science and Public Life*. Princeton University Press.

Rubin, D. B. (1984). Bayesianly justifiable and relevant frequency calculations for the applied statistician. *Annals of Statistics* **12**, 1151–1172.

Sheiner, L. B. (1984). The population approach to pharmacokinetic data analysis: Rationale and standard data analysis methods. *Drug Metabolism Reviews* **15**, 153–171.

Silberzahn, R., et al. (2015). Crowdsourcing data analysis: Do soccer referees give more red cards to dark skin toned players? Center for Open Science, `https://osf.io/j5v8f/`

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.

Steegen, S., Tuerlinckx, F., Gelman, A., and Vanpaemel, W. (2016). Increasing transparency through a multiverse analysis. *Perspectives on Psychological Science*.

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, 505–513. London: Chapman and Hall.

Tukey, J. W. (1962). The future of data analysis. *Annals of Mathematical Statistics* **33**, 1–67.

van Fraassen, B. (1980). *The Scientific Image*. Oxford University Press.

van Loo, H. M., and Romeijn, J. W. (2015). Psychiatric comorbidity: Fact or artifact? *Theoretical Medicine and Bioethics* **36**, 41–60.

Vermunt, J. K. and Magidson, J. (2016) *Technical Guide for Latent GOLD 5.1: Basic, Advanced, and Syntax.* Belmont, Mass.: Statistical Innovations Inc.

von Mises, R. (1957). *Probability, Statistics and Truth*, second English edition. New York: Dover.

Wang, W., Rothschild, D., Goel, S., and Gelman, A. (2015). Forecasting elections with non-representative polls. *International Journal of Forecasting* **31**, 980–991.

Wasserman, L. (2006). Frequentist Bayes is objective (comment on articles by Berger and by Goldstein). *Bayesian Analysis* **1**, 451–456.

Weinberger, D. (2009). Transparency is the new objectivity. Everything is Miscellaneous blog, 19 Jul. `http://www.everythingismiscellaneous.com/2009/07/19/transparency-is-the-new-objectivity/`

Yong, E. (2012). Nobel laureate challenges psychologists to clean up their act. *Nature News*, 3 Oct. `http://www.nature.com/news/nobel-laureate-challenges-psychologists-to-clean-up-their-act-1.11535`

Appendix: Objectivity in the philosophy of science

Megill (1994) lists four basic senses of objectivity: "absolute objectivity" in the sense of "representing the things as they really are" (independently of an observer), "disciplinary objectivity" referring to a consensus among experts within a discipline and highlighting the role of communication and negotiation, "procedural objectivity" in the sense of following rules that are independent of the individual researcher, and "dialectical objectivity," referring to active human "objectification" required to make phenomena communicable and measurable so that they can then be treated in an objective way so that different subjects can understand them in the same way. These idea appear under various names in many places in the literature. Porter (1996) lists the ideal of impartiality of observers as another sense of objectivity. Douglas (2004) distinguishes three modes of objectivity: human interaction with the world, individual thought processes, and processes to reach an agreement. Daston and Galison (2007) call the ideal of scientific images that attempt to capture reality in an unmanipulated way "mechanical objectivity" as opposed to "structural objectivity," which refers to mathematical and logical structures. The latter emerged from the insight of scientists and philosophers such as Helmholtz and Poincare that observation of reality cannot exclude the observer and will never be as reliable and pure as "mechanical objectivists" would hope.

More generally, pretty much all senses of objectivity have been criticized at some point in history for being unachievable, which often prompted the postulation of new scientific virtues and new senses of objectivity. For example, the realist ideal of "absolute objectivity" has been branded as metaphysical, meaningless, and illusory by positivists including Karl Pearson (1911), and more contemporarily by empiricists such as van Fraassen (1980). The latter takes observability and the ability of theory to account for observed facts as objective from an anti-realist perspective.

Some authors even criticize the idea that objectivity is a generally desirable virtue in science, e.g., for its implication of a denial of the specific conditions of an observer's point of view (Feyerabend, 1978, McKinnon, 1987, Maturana, 1988) and its use as a rhetorical device or tool of power (see Fuchs, 1997, for a critical overview of such ideas).

The core benefit of such controversies around objectivity and subjectivity for statisticians is the elaboration of aspects of good science, which should inform statistical data analysis and decision making. Hacking (2015) wrote a paper called "Let's not talk about objectivity," and with him we believe that for discussing the practice of statistics (or more generally science), the objectivity vs. subjectivity discourse should be replaced by looking at more specific virtues of scientific work, the awareness of which could have a more direct impact on the work of scientists. The virtues that we have listed in Section 3 are all connected either to senses of objectivity as summarized above, or to reasons for criticizing certain concepts of objectivity.