Beyond subjective and objective in statistics[*]

Andrew Gelman[†] Christian Hennig[‡]

15 Feb 2015

Abstract

We propose to abandon the words "objectivity" and "subjectivity" in statistics discourse and 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. 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 acknowledgement 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

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, the process of scientific inference and discovery involves individual choices; indeed, scientists and the general public celebrate the brilliance and inspiration of greats such as Einstein, Darwin, and the like, recognizing the roles of their personalities and individual experiences in shaping their theories and discoveries, and philosophers of science have studied the interplay between personal attitudes and scientific theories (Kuhn, 1962). Thus it is clear that 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. 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. From the other direction, scientists who use methods labeled as objective often seem so intent on eliminating subjectivity from their analyses, that they end up censoring themselves. This happens, for example, when researchers rely on p-values but refuse to recognize that their analyses are contingent on data (as discussed by Simmons, Nelson, and Simonsohn, 2011, and Gelman and Loken, 2014): significance testing is used as a tool for 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.

[*] We thank Sebastian Weber for helpful comments.

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

[‡] Department of Statistical Science, University College London.

More generally, misguided concerns about subjectivity can lead researchers to avoid incorporating relevant and available information into their analyses.

A perhaps helpful analogy is to gender roles in social interactions. To get respect, women often need to choose between claiming stereotypically-male behaviors or affirming, or “taking back,” feminine roles. At the same time, men can find it difficult to step outside the restrictions implied by traditional masculinity. Rather than point and label, it can be better in such situations to identify the positive aspects of each sex role and then go from there. Similarly, good science contains both subjective and objective elements, and we think it would be best to understand how these perspectives can complement each other. Our point in this analogy is not to identify the female side with subjectivity and the male side with objectivity, nor are we seeking to engage with feminist philosophy; rather, we are merely alluding to another context, familiar to many of us within science and more generally, in which difficulties arise when behaviors are interpreted within a conventional dichotomy.

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, and that they are used with a variety of different meanings and are therefore prone to misunderstandings.

In Section 2 we lay out our proposed discourse. In Section 3 we review discussions about the meaning of objectivity and subjectivity in statistics, some other scientific areas and particularly philosophy. From this, we develop in Section 4 our perspective, including the heart of our paper, a detailed list of specific scientific virtues that are normally subsumed under the broader terms of objectivity and subjectivity. In Section 5 we apply our discourse to a discussion of the main streams of the foundations of statistics. This is followed by examples in Section 6 and a concluding discussion in Section 7.

2. Our proposal

We were motivated to write this paper because we felt that the statistical discourse on objectivity and subjectivity had become counterproductive. 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.

We propose when talking about statistics to abandon the words “objectivity” and “subjectivity” and 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 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 acknowledgement of multiple perspectives as complementary goals.

2.1. “Transparency,” “consensus,” “impartiality,” and “correspondence to observable reality,” instead of “objectivity”

Merriam-Webster defines “objective” as “based on facts rather than feelings or opinions: not influenced by feelings” and “existing outside of the mind: existing in the real world.” Science is practiced by human beings, who only have access to the real world through interpretation of their perceptions. Taking objectivity serious as an ideal, scientists need to make the sharing of their perceptions and interpretations possible. When applied to statistics, the implication is that the choices in the data analysis (including the prior distribution, if any, but also the model for the data, and the choice of what information to include in the first place) should be externally motivated based on factual, externally verifiable information. 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 (authors like van Fraassen, 1980, argue that analyses can be objective conditional on modeling decisions without being unique; see Section 3.3). 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, 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. But transparency is not enough. We hold that science aims at *consensus* in potentially free exchange (see Section 4.1 for elaboration), 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 motivations for the methodology and the judgments made and a clear explanation how this related to already existing knowledge. Following generally accepted rules and procedures supports the impression that results do not depend on the personalities of the 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. In any case, consensus can only be achieved if researchers appear *impartial* by taking into account competing perspectives, avoiding to favor prechosen hypotheses, and being open to criticism.

The "real world" outside the observer's mind plays a key role in usual concepts of objectivity. Finding out about the real world is seen by many as the major objective of science, and this suggests correspondence to reality as the ultimate source of scientific consensus. This idea is not without its problems and meets some philosophical opposition; see Section 3.3. 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. Still, as statisticians we are concerned with making statements about reality based on systematized observations, and this makes *correspondence to observed reality* a core concern regarding objectivity.

Formal statistical methods (Bayesian and non-Bayesian alike) 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 Bayesian priors are interpreted and used (see Sections 5.3–5.5), Bayesian priors fulfill or aid some or all of the virtues listed above; they 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. As classical likelihood functions, they depend upon some mixture of scientific judgment, feasibility, and convention, but in both cases, they can and are assigned based on hard data.

2.2. “Multiple perspectives” and “context-dependence,” instead of “subjectivity”

Merriam-Webster defines “subjective” as “relating to the way a person experiences things in his or her own mind” and “based on feelings or opinions rather than facts.” Science is normally seen as striving for objectivity, and therefore acknowledging subjectivity is not so popular in science. 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 recognize the information coming from differing perspectives.

In statistics the idea of subjectivity is often used for characterization of Bayesian inference and regularization—that is, estimates and statistical procedures that seek guidance from outside the data to improve accuracy and predictive power. For example, default priors on regression coefficients can be 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. 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. Sparsity assumptions can themselves 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.

Here we would like to consider subjectivity in a broader way. Indeed, there is no logical reason for subjectivity to go together with regularization in particular. 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 Bayes or regularized estimates as “subjective” misses the point.

Beyond this, even the choice of outcome measure in any given problem can be subject to debate. Consider, for example, the decision of whether to predict vote share in an election or score differential in a sporting event, or to directly predict the winner in either case (Gelman, 2014a). This modeling decision turns on statistical and substantive considerations, and the most important step here is to recognize that this choice of model exists, not to sweep it under the rug to avoid a possible accusation of subjectivity.

For another example, binomial-data confidence intervals based on $(y + 2)/(n + 4)$ give better confidence intervals than the classical intervals based on y/n (Agresti and Coull, 1998). Whereas the latter have a straightforward justification, the former are based on trading interval width against conservatism and involve 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 rather irrelevant which is considered more subjective.

Nonetheless, we do think there is *some* connection to subjectivity there, and the connection we see is that Bayes and, more generally, regularized approaches have tuning parameters which must be specified externally.[1] And this brings us to the question: Is the external specification of tuning parameters a bad thing (in that it introduces a subjective human element into the analysis) or a good thing (in allowing the user to include prior information)?

Here is where we would like to replace the concept of "subjectivity" with awareness of *multiple perspectives* and *context-dependence*. 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. To connect with the other half 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; actually, many tuning parameters are related to aims of the analysis and desirable characteristics of methods rather than to some assumed underlying "truth" about which there could be a "belief" (see Section 6.3). Rather, we should justify the choice based in some way on empirical data, embedding the choice in a statistical model that can ultimately be linked back to observable reality.

Alternatively, the choice of tuning parameter can be based on knowledge of the impact of the choice on results and a clear explanation why a certain impact is desired or not. In an application of 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).

2.3. 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 4.2. 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, 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.

[1] Yes, some tuning parameters can be estimated from data via hierarchical Bayes or cross-validation but that inference itself can at times be noisy, which just pushes the problem one step back.

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. 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," just as we would not limit classical statistical inference to "the analysis of simple random samples." Yes, Bayes 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, classical methods can be applied to all sorts of problems that do not involve random sampling. It's all about clarifying the foundation for using the mathematical models for a larger class of problems.

3. Objectivity and subjectivity in statistics and science

3.1. Discussions within statistics

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 for sure 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.

Other instances where the objectivity vs. subjectivity issue comes up in statistics are statistical methods that require tuning parameters (for example, the proportion of trimmed observations in trimmed means, or the threshold in wavelet smoothing); decision boundaries such as the significance level of tests; and decisions regarding inclusion, exclusion, and transformation of data in preparation for analysis.

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).

3.2. Discussions in other fields

Scholars in humanistic studies such as history and literary criticism have considered the ways in which differently-situated observers can give different interpretations to what Luc Sante calls the "factory of facts." In political arguments, controversies often arise over "cherry picking" or selective use of data, a concern we can map directly to the statistical principle of random or representative sampling, and the more general idea that information used in data collection be included in any statistical analysis (Rubin, 1978). In a different way, the concepts of transference and counter-transference, central to psychoanalysis, live at the boundary of personal impressions and measurable facts, all subject to the constraint that, as Philip K. Dick put it, "Reality is that which, when you stop believing in it, doesn't go away."

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. The interplay between objective and subjective effects also arises in political science, for example in the question of whether to attribute the political successes of a Ronald Reagan or a Bill Clinton to their charisma and appealing personalities, to their political negotiating skills, or simply to periods of economic prosperity that would have made a success out of just about any political leader. Again, these disputes link to controversies regarding research methods: a focus on objective, measurable factors can be narrow, but with a more subjective analysis it can be difficult to attain a scientific consensus. In fields such as social work it is clear that one must work with subjective realities in order to make objective progress (Saari, 2005) but this view is relevant to science more generally.

In the social and physical sciences alike (as well as in hybrid fields such as psychophysics), the twentieth century saw an intertwining of objectivity and subjectivity. From one direction, Heisenberg's uncertainty principle told us that, at the quantum level, measurement depends fundamentally on the observation process, an insight that is implicit in modern statistics and econometrics with likelihood functions, measurement-error models, and sampling and missing-data mechanisms being manifestations of observation models. So in that sense there is no pure objectivity. From the other direction, psychologists have continued their 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.

3.3. Concepts of objectivity

Discussions involving the antagonism between objectivity and subjectivity often suffer from the fact that objectivity means different things to different people, in statistics and elsewhere (much of the following discussion will focus on the term "objectivity"; subjectivity is often considered as the opposite of objectivity and as such implicitly defined). Ambiguity in these terms is often ignored. We believe that such discussions can become clearer by referring to the meanings that are relevant in any specific situation instead of using the ambiguous terms "objectivity" and "subjectivity" without further explanation.

The core of the current use of the term "objectivity" is the idea of impersonality of scientific statements and procedures. According to Daston and Galison (2007), the term has only been used in this way in science from the mid-nineteenth century; before then, "objective" and "subjective" were used with meanings almost opposite from the current ones and did not play a strong role in discussions about science.

The idea of independence of the individual subject can be applied in various ways. Megill (1994) listed 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." The latter somewhat surprisingly involves subjective contributions, because it refers 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. Statistics for example relies on the construction of well delimited populations and categories within which averages and probabilities can be defined; see Desrosieres (2002).

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 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. Instead, "structural objectivity" refers to mathematical and logical structures. Porter (1996) lists the ideal of impartiality of observers as another sense of objectivity, and highlights the important role of quantitative and formal reasoning for concepts of objectivity because of their potential for removing ambiguities.

To us, the most problematic aspect of the term "objectivity" is that it incorporates normative and descriptive aspects, and that these are often not clearly delimited. 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 depending on the specific situation the lack of flexibility caused by the impossibility of tuning may 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. Although some proponents of frequentism consider its objectivity (in the sense of impersonality, conditional on the definition of the reference set) as a virtue, 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.

The interpretation of objectivity as a scientific virtue is connected to what are seen to be the aims and values of science. Scientific realists hold that finding out the truth about the observer-independent reality is the major aim of science. This makes "absolute objectivity" as discussed above a core scientific ideal, as which it is still popular. But observer-independent reality is only accessible through human observations, and the realist ideal of objectivity has been branded as metaphysical, meaningless, and illusory by positivists, among which Porter (1996) counted Karl Pearson, and more contemporarily by empiricists such as van Fraassen (1980). In the latter groups, objectivity is seen as a virtue as well, although for them it does not refer to observer-independent reality but rather to a standardized, disciplined, and impartial application of scientific methodology enabling academic consensus about observations. Reference to observations is an element that the empiricist, positivist, and realist ideas of objectivity have in common; Mayo and Spanos (2010) interpreted objectivity in a realist manner, and they saw checking theories against experience by means of what they call "error statistics" as central tool to ensure objectivity. In contrast, van Fraassen (1980) took observability and the ability of theory to account for observed facts as objective from an *anti*-realist perspective. His construal of observability depends on the context, theory, and means of observation, and his concept of objectivity is conditional on these conditions of observation, assuming that at least acceptance of observations and observability given these conditions should not depend on the subject.

Daston and Galison (2007) portray the rise of "mechanical objectivity" as a scientific virtue in reaction to shortcomings of the earlier scientific ideal of "truth-to-nature," which refers to the idea that science should discover and present an underlying ideal and universal (Platonic) truth below the observed phenomena. The move towards mechanical objectivity, inspired by the development of photographic techniques, implied a shift of perspective; instead of producing pure and ideal "true" types the focus moved to capturing nature "as it is," with all irregularities and variations that had been suppressed by a science devoted to "truth-to-nature." Increasing insight in the shortcomings and the theory-dependence of supposedly objective observational techniques led to the virtue of "trained judgement" as a response to mechanical objectivity. According to Daston and Galison (2007), the later virtues did not simply replace the older ones, but rather supplemented them, so that nowadays all three still exist in science. Daston and Galison do not discuss statistics, but the statistical idea of modeling error causing variation around a true parameter can be seen as an attempt to integrate the mechanically objective observation of variation with the ideal of truth-to-nature. The idea from sampling theory of estimating a population quantity is more similar to the photographic idea of estimating nature as it is. One could also connect subjective and objective Bayesian perspectives with the concept of trained judgement. In Daston and Galison's framework, objectivity appears as one scientific virtue among others.

Objectivity has also been criticized on the grounds that, as attractive as it may seem as an ideal, it is illusory. This criticism has to refer to a specific interpretation of objectivity, and a weaker interpretation of objectivity may still seem to critics to be a good thing: van Fraassen agrees with Kuhn (1962) and others that "absolute objectivity" is an illusion and that access to reality is dependent of the observer, but he still holds that objectivity conditional on a system of reference is a virtue. But there is even criticism of the idea that objectivity, possible or not, is desirable. From a particular feminist point of view, MacKinnon (1987) wrote: "To look at the world objectively is to objectify it." Striving for objectivity itself is seen here as a specific and potentially harmful perspective, implying a denial of the specific conditions of an observer's point of view. A similar point was made by Feyerabend (1978). Maturana (1988) critically discussed the "explanatory path of objectivity-without-parenthesis" in which observers deny personal responsibility for their positions based on a supposedly privileged access to an objective reality; he accepted a more positive perspective-dependent use of the term called "objectivity-with-parenthesis."

4. Our perspective

4.1. Our attitude toward objectivity and subjectivity in science

The attitude taken in the present paper is based on Hennig (2010), which was in turn inspired by constructivist philosophy (Maturana, 1988, von Glasersfeld, 1995) and distinguishes personal reality, social reality, and observer-independent reality. 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.

According to Hennig (2010), 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.

This does not imply, however, that scientific disputes can never be resolved by scientific means. Yes, there is an element of "politics" involved in the adjudication of scholarly disagreements, but, as we shall discuss, the norm of *transparency* and other norms associated with both objectivity and subjectivity can advance such discussions. In general no particular observer has a privileged position but this does not mean that all positions are equal. We recognize subjectivity not to throw up our hands and give up on the possibility of scientific consensus but as a first step to exploring and, ideally, reconciling, the multiple perspectives that are inevitable in nearly any human inquiry.

Denying the existence of different legitimate subjective perspectives and of their potential to contribute to scientific enquiry cannot make sense in the name of objectivity. Heterogeneous points of view cannot be dealt with by imposing authority. Our attitude values the attempt to reach scientific agreement between different perspectives, but ideally such an agreement is reached by free exchange between the different points of view. In practice, however, agreement will not normally be universal, and in order to progress, science has to aim at a more restricted agreement between experts who have enough background knowledge to either make sure that the agreement about something new is in line with what was already established earlier, or to know that and how it requires a revision of existing knowledge. But the resulting agreement is still intended to be potentially open for everyone to join or to challenge. Therefore, in science there is always a tension between the ideal of general agreement and the reality of heterogeneous perspectives.

Furthermore our attitude to science is based on the idea that consensus is possible regarding stable and reliable statements about the observed reality (which may require elaborate measurement procedures), and that science aims at nontrivial knowledge in the sense that it makes statements about observable reality that can and should be checked and potentially falsified by observation.

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. This reality is a target of science, although it cannot be taken for granted that it is indeed independent of the observer. 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." This form of realism is not in contradiction to the criticism of realism by van Fraassen or the arguments against the desirability of certain forms of objectivity by constructivists or feminists as outlined above. Active scientific realism implies that finding out the truth about objective reality is not the ultimate aim of science, but that it rather supports human actions. This means that scientific methodology has to be assessed relative to the specific aims and actions connected to its use. Another irreducible subjective element in science, apart from multiple perspectives on reality, is therefore the aim of scientific inquiry, which cannot be standardized in an objective way. A typical statistical instance of this is how much prediction accuracy in a restricted setting is valued compared with parsimony and interpretability.

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. Objectivity as we see it is therefore a scientific ideal that can never fully be achieved. As much as science aims for objectivity, it has to acknowledge that it can only be built from a variety of subjective perspectives through communication.

4.2. A list of specific objective and subjective virtues

Calling an approach, a method, a statement, or a result "objective" is often a misleading marketing claim. Where this is not the case, it is still imprecise and ambiguous as a description. It is clearer and more useful to refer to its specific qualities that support the scientific aim of producing a view of reality that is stable and reliable and can be freely agreed upon by general observers. Virtues that are often referred to as "objective" include:

1. Transparency:

    (a) Clear and unambiguous definitions of concepts,
    (b) Open planning and following agreed protocols,
    (c) Full communication of reasoning, procedures, 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.

What about subjectivity? The term "subjective" is often used as opposite to "objective" and as such often meant to be opposed to scientific virtues, or to be something that cannot fully be avoided and that therefore has to be only grudgingly accepted.

But subjective perspectives are the building blocks for scientific consensus, and therefore there are also scientific virtues associated with subjectivity:

1. Awareness of multiple perspectives,

2. Awareness of context-dependence:
    (a) Recognition of dependence on specific contexts and aims,
    (b) Honest acknowledgement of the researcher's position, goals, experiences, and subjective point of view.

In the subsequent discussion we shall label the items in the above lists as O1a–O4b or O1–O4 for groups of items ("O" for "connected to objectivity"), and S1, S2 (S2a, S2b) for the items connected to subjectivity. Our intention is to sketch a system of virtues that allows a more precise and detailed discussion where issues of objectivity and subjectivity are at stake.

We are aware that in some situations some of these virtues may oppose each other, for example "consensus" can contradict "openness to multiple perspectives," but we think that this reflects an essential and unavoidable tension in 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 three 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 the following section we will apply the list to approaches to the foundation of statistics, but the items O1c and S2b rather apply to specific studies.

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. Thus, we take frequentism to be an interpretation of probability, which 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 4.2.

5.1. Frequentism

We label "frequentism" as 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. The term "independence" here should not be identified with formal stochastic independence, which is defined in terms of probabilities and therefore requires an interpretation of probability already in order to be meaningful, so it cannot be used to *define* an interpretation of probability, as von Mises (1957) recognized. However, stochastic independence is needed to model a sequence of experiments in such a way that the unobservable probability of an event can be connected to observed relative frequencies of occurrences under finite replication via the binomial distribution. In the wider sense, we call such probabilities "frequentist" when they formalize observer-independent tendencies or propensities of experiments to yield certain outcomes (see, for example, Gillies, 2000).

The frequentist mindset locates probabilities in the observer-independent world, so they are in this sense objective. This objectivity, however, is model-based, as an infinite amount of actual replicates cannot exist, and most researchers, in most settings, would be skeptical about truly identical replicates and true independence or, when it comes to observational studies, about whether observations can be seen as drawn in a purely random manner from an appropriate reference set.

The decision to adopt the frequentist interpretation of probability regarding a certain phenomenon therefore requires idealization. It cannot be justified in a fully objective way, which here means, referring to our list of virtues, that it can neither be enforced by observation, nor 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 (O2, O4). Once a frequentist model is adopted, however, it makes predictions about observations that can be checked, so the reference to the observable reality (O4) is clear.

There is some disagreement about whether the frequentist definition of probability is clear and unambiguous (O1a). 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, given that the circumstances of the experiment are well defined and regardless of whether frequencies indeed behave in the implied way. On the other hand, von Mises was not completely successful in his attempt to avoid involving stochastic independence and identity in the definition of frequentist probabilities through the concepts of the collective and the axiom of invariance under place selection rules (Fine, 1973), and the issue has never been completely resolved.

Frequentism implies that, in the observer-independent reality, true probabilities are unique, but there is considerable room for multiple perspectives (S1) regarding the definition of replicable experiments, collectives, or reference sets. The idea of replication is often constructed in a rather creative way. For example, frequentist time series models are used for time series data, implying 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 was 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 (S2a) 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.

From the perspective taken here and in Hennig (2010), the frequentist interpretation of probability can be adopted as an idealized model without having to believe that frequentist probabilities really exist in the observer-independent world. This can be justified, on a case-by-case basis, if it is seen as useful for the scientific aims in the given situation, for example because a specific frequentist model communicates (more or less) clearly the scientist's view of a certain phenomenon (O1a), and implies the means for testing this against observations (O4).

5.2. Error statistics

The term "error statistics" was coined by the philosopher Deborah Mayo (1996). We use it here to refer to an approach to statistical inference that is based on a frequentist interpretation of probability and methods that can be characterized and evaluated by error probabilities. Traditionally these would be the Type I and Type II errors of Neyman-Pearson hypothesis testing, but 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 (1996) introduced another key concept for error statistics, "severity," namely the probability of observing a result as much or less in line with a hypothesis than what was actually observed, given that the hypothesis is false. Severity is connected with, but not identical to, the power of tests. It serves to quantify the extent to which a test result can corroborate a hypothesis (keeping in mind that testing specific statistical hypothesis can only ever shed light on isolated aspects of a scientific theory of interest; and that a specific test can only corroborate a specific aspect of a hypothesized statistical model).

According to Mayo and Spanos (2010), objectivity is a core concern of error statistics, which is specifically driven by providing methodology for reproduction, testing, and falsification (O4b). Mayo (2014) defined objective scientific measurement as being “relevant,” “reliably capable,” and “able to learn from error,” which outlines the error-statistical rationale for consensus (O2c). Error statistical methodology is portrayed as “reliably capable” as far as its potential to produce inferential errors can be analyzed, and as far as the resulting error probabilities are low. The “ability to learn from error” refers to erroneous hypotheses, rejected by an error statistical procedure that optimally can pinpoint the reason for rejection and thus lead to an improvement of the hypothesis, rather than errors of the inferential method. 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, and it implies that for Mayo the reference to observations is central for objectivity.

Mayo’s “relevance” concerns the problem of inquiry of interest and is therefore related to virtue S2a, which we classified as related to subjectivity. As Mayo attempts to defend the objectivity of the error statistical approach against charges of subjectivity, she may not be happy about this classification, but we agree with her that this is an important virtue nonetheless, which, however, is not specifically connected to error statistics.

The error probability characteristics of error statistical methods rely, in general, on model assumptions. In principle, these model assumptions can be tested in an error statistical manner, too, and are therefore, according to Mayo, 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 (virtue O1b), 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 (2014) therefore suggests that it is misleading to hypothesize models or parameters to be “true,” and that one should instead take into account all models that are “adequate” for approximating the data in the sense that they are not rejected by tests based on features of the data the statistician is interested in, which does not require reference to unobservable true frequentist probabilities, but takes into account error probabilities as well. Such an approach is tied to the observations in a more direct way without making metaphysical assumptions about unobservable features of observer-independent reality (O1a, O4). However, it is possible that such a metaphysical assumption is implicitly still needed if the researcher wants to use “data approximating models” to learn about observer-independent reality, and that the class of all adequate models is too rich for meaningful inference (as in more standard frequentist treatments, Davies focuses on models with independent and identically distributed random variables or error terms). Earlier work on robust statistics (see Huber and Ronchetti, 2009) already introduced the idea of sets of models that neighbor a nominal model, from which the models in the neighborhood could not be reliably distinguished based on the data.

Even further flexibility in error statistical analyses comes from the fact that the assumption of a single true underlying distribution does not determine the parametric or nonparametric family of distributions, within which the true distribution is embedded. Although Neyman and Pearson derived optimal tests considering specific alternatives to the null hypothesis, many kinds of alternatives and test statistics could be of potential interest. Davies (2014) explicitly mentions the dependence of the choice of statistics for checking the adequacy of models on the context and the researcher's aims (S2a) instead of relying on Neyman-Pearson type optimality results.

Overall, there is no shortage of entry points for multiple perspectives (S1) in the error statistical approach. 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 and error statistical 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 (O2c). 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. Given the degree of idealization required for frequentism as discussed in Section 5.1, this is certainly a legitimate position.

In de Finetti's work the term "prior" refers to all probability assignments made before seeing the data, 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 4.2, de Finetti provided a clear definition of probability (O1a) based on principles that he sought to establish as generally acceptable (O2c). As opposed to 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. 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 (O2c)—but can be defended for honestly acknowledging that prior information often does not come in ways that allow a unique formalization (S2b). 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 (O1c). Such transparency is desirable in any statistical approach but is particularly relevant for subjective Bayesian models which cannot be rejected within the subjectivist paradigm in case of disagreement with observations.

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 (O3a).

However, priors in the subjectivist Bayesian conception are not open to falsification (O3b), 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 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 (1982) 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 subjective virtues S1 and S2b. 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. Furthermore, some problems of the frequentist approach criticized by de Finetti and his disciples stem from the unavoidable fact that useful mathematical models idealize and simplify personal and social perspectives on reality (see Hennig, 2010 and above), and the subjectivist Bayesian approach incurs such issues as well.

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 (O2c), 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 or working with informally specified prior information, even the more elementary key issue of specifying an objective prior distribution formalizing the absence of information is riddled with difficulties, and there are various principles for doing this which 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, Bernardo (1979) spoke of "reference priors," avoiding the term "objective," and emphasizing that it would be desirable to have a convention for such cases (O2b), 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.

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 the subjectivist Bayesian approach. 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 (but 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. This drive, together with some subjectivist honesty about the fact that despite trying hard in the vast majority of applications the resulting prior will not deserve the "objectivity" stamp and will still be subject to potential disagreement, can potentially combine the best of both of these traditional Bayesian worlds.

5.5. Falsificationist Bayesianism

For both subjectivist and objectivist Bayesians, following de Finetti (1974) and Jaynes (2003), 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. 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 improve Bayesian statistics regarding virtue O4b.

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 error statistical techniques (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. This reflects certain attitudes of Jaynes (2003), with the difference that Jaynes generally considered probability models as derivable from constraints of a physical system, whereas Gelman and Shalizi focus on examples in social or network science which are not governed by simple physical laws and thus where one cannot in general derive probability distributions from first principles, so that "priors" (in the sense that we are using the term in this paper, encompassing both the data model and the parameter model) are more clearly subjective.

A central issue for falsificationist Bayesianism is the meaning and use of the parameter prior, which can have various interpretations, which gives falsificationist Bayesianism a lot of flexibility for taking into account multiple perspectives, contexts, and aims (S1, S2a) but may be seen as a problem regarding clarity and unification (O1a, O2c). Frequentists may wonder whether a parameter prior is needed at all. Here are some potential benefits of incorporating a parameter prior:

- The parameter prior may formalize relevant prior information.
- The parameter prior may be a useful device for regularization.
- The parameter prior may formalize deliberately extreme points of view to explore sensitivity of the inference.
- The parameter prior may make transparent a point of view involved in an analysis.
- The parameter prior may facilitate a certain kind of behavior of the results that is connected to the aims of analysis (such as penalizing complexity or models on which it is difficult to act by giving them low prior weight).
- The Bayesian procedure involving a certain parameter prior may have better error statistical properties (such as the mean squared error of point estimates derived from the posterior) than a straightforward frequentist method, if such a method even exists.

- Often finding a Bayesian parameter prior which emulates a frequentist/error statistical method helps understanding the implications of the method.

Here are some ways to interpret the parameter prior:

- The parameter prior may be interpreted in a frequentist way, as formalizing a more or less idealized data generating process generating parameter values. The "generated" parameter values may not be directly observable, but in some applications the idea of having, at least indirectly, a sample of several parameter values from the parameter prior makes sense ("empirical Bayes"). In many other applications the idea is that only a single parameter from the parameter prior is actually realized, which then gives rise to all the observed data. Even in these applications one could in principle postulate a data generating process behind the parameter, of which only one realization is observable, and only indirectly. This is a rather bold idealization, but frequentists are no strangers to such idealizations either; see Section 5.1. A similarly 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.

  If parameter priors are interpreted in this sense, they can actually be tested and falsified using error statistical methods; see Gelman, Meng and Stern (1996). In situations with only one parameter realization, the power of such tests is low, though, and any kind of severe corroboration will be hard to achieve. Also, if there is only a single realization of an idealized parameter distribution, the information in the parameter posterior seems to rely strongly on idealization.

- If the quality of the inference is to be assessed by error statistical measures, the parameter prior may be seen as a purely technical device. In this case, however, the posterior distribution does not have a proper interpretation, and only well defined statistics with known error statistical properties such as the mean or mode of the parameter posterior should be interpreted.

- Assuming that frequentist probabilities from sampling models should be equal to the subjectivist or objectivist epistemic probabilities if it is known that the sampling model is true (which Lewis, 1980, called "the principal principle"), the parameter prior can still be interpreted as giving epistemic probabilities such as subjectivist betting rates, conditionally on the sampling model to hold, even if the sampling model is interpreted in a frequentist way. The possibility of rejecting the sampling model based on the data will invalidate both coherence and Cox's axioms, so that the foundation for the resulting epistemic probabilities becomes rather shaky. This does not necessarily have to stop an individual from interpreting and using them as betting rates, though.

Given such a variety of uses and meanings, it is crucial for applications of falsificationist Bayesianism that the choice of the parameter prior is clearly explained and motivated, so transparency is central here as well as for the other varieties of Bayesian statistics.

Overall, falsificationist Bayesianism combines the virtue of error statistical falsifiability with the virtues listed above as "subjective," doing so via a flexibility that may be seen by some as problematic regarding clarity and unification.

6. 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, with an intellectual "paper trail" 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 two 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.

6.1. 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).

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.

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 (O1) and some aspects of consensus (O2) and awareness of multiple perspectives (S1)—while recognizing it 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 (O4), and our awareness of context-dependence (S2) 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 (O2) and impartiality (O3). 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 (O1) and to compare to other possible prior distributions supported by other information, thus supporting impartiality (O3) and awareness of multiple perspectives (S1).

As with "institutional decision analysis" (Gelman et al., 2003, section 22.5), 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.

6.2. 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.

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).

We shall now express this discussion using the criteria from Section 4.2. Probability sampling has the clear advantage of transparency (O1) 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 (O2), at least in the United States, where such surveys have a long history and are accepted in marketing and opinion research. Impartiality (O3) and correspondence to observable reality (O4) 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).

The virtues associated with subjectivity are less apparent in our sample survey example, perhaps because there is clear agreement on the goal of estimating vote intention among potential voters. Stepping back a bit, though, one could consider the various adjustment schemes to represent awareness of context-dependence (S2) 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. Overall, though, multiple perspectives (S1) and context-dependence (S2) are less relevant here; they could play a more prominent role in related settings such as the use of probability samples or opt-in polls in marketing research.

6.3. 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 measure. 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 dissimilarity 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 kluge 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 "truly correct" 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 very 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 very strong temptation to ignore the issue and to just choose $c = 1$, which may look "natural" in the sense that it maps zero on to 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 a $c > 1$, 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 very often hidden in presentations of data analyses.

In Hennig and Liao (2013), we pointed out the issue and did some sensitivity analysis about the strength of the impact of the choice of c (O1, transparency). The way we picked the 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 (S2). We were also clear that alternative choices would amount to alternative perspectives rather than being just wrong (S1, O3).

In order to foster consensus and to make a connection to observable reality (O2, O4), it would be interesting to see how the value c could be connected to observations. Such a connection would necessarily be mediated by some model, as it is difficult to see how c itself could be given any kind of direct observable meaning. One approach that mixes objective and subjective reasoning would be to fit the resulting dissimilarity measure by choice of c and tuning constants required for the other variables to dissimilarity assessments between some sampled persons by experts. One could also think about real-life situations in which measurements can be taken that are about relating the amount of savings of a person to their socioeconomic status, although this is probably only conceivable in a very indirect manner.

In any case, we advocate the willingness to make decisions that cannot fully be justified on the grounds of consensus and impartiality, to be open about context-dependence and about the sensitivity of results to choices in the analysis. We also advocate to try hard to somehow find rationales for consensus and connections to observable data, to reduce the arbitrariness of such decisions. This relates to our general recommendation to have a "paper trail" justifying choices in data analysis and decision making.

It is however problematic to establish rationales for consensus that are based on ignoring the 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), for example trying to optimize "clusterability" of the dataset, or to prefer 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 skew and large savings amounts are rather sparsely distributed.

7. Discussion

The list in Section 4.2 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 violate these rules, which indicates to us that the underlying principles are subtle and motivates this paper. We hope that a change in names will clarify what can be done to improve statistical analyses in these two dimensions.

It is particularly common for scientific studies to be insufficiently transparent regarding choices made by the researchers and the reasons why what the researchers did was preferred to potential alternatives. This may come either by offering up models and analyses with no justification at all, or with a justification that is internal to a literature—essentially, doing something because it was done before. For example, in a study of menstrual cycle and voting patterns, Durante, Arsena, and Griskevicius (2013) define ovulation as days 7–14 of a 28-day menstrual cycle, with no clear justification of this choice beyond some references to earlier papers in the evolutionary psychology subfield. However, a quick check of more relevant sources reveals this choice to be questionable. For example, the website of the U.S. Department of Health and Human Services[2] places the dates of likely fertility as days 10–17, and other family planning sources give similar dates, for example, from

[2] `http://www.hhs.gov/opa/reproductive-health/contraception/natural-family-planning/`

Planned Parenthood[3], “After Day 11, hormones start working on the ripest egg to get it released from the ovary. Day 14, 15, or 16 is usually the day the egg is released (in a 28-day cycle).”

Durante et al. are transparent insofar as they cite the sources they used. But there is no visible attempt to explore the existing relevant knowledge. The authors appeal to consensus by just citing existing work, instead of acknowledging that scientific consensus has to be grounded in scientific observation and openness to a wide range of perspectives and possible objections. The reader is left wondering to what extent the authors' result is a consequence of choices that are not made transparent.

It is also routine for scientific research papers to fail to recognize other viewpoints. 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.

Scientific papers often present sensitivity analyses to get a sense of the robustness of their conclusions to small perturbations of their statistical models, but this is not the same as acknowledgment of multiple perspectives. For example, Chen et al. (2013) present a regression-discontinuity analysis to estimate the effects of air pollution on life expectancy in a set of Chinese cities. In their paper they present a sensitivity analysis in which the control variable (distance of the cities from a certain river) is included in linear, quadratic, cubic, quartic, and quintic forms. It can be valuable to present multiple versions of a model, but we consider this as elaborations on a single perspective. A multiple-perspective approach would allow for systematic variation unexplained by the model, which indeed seems reasonable in this case: if one government policy can have major differential effects on life expectancy in different cities, presumably other systematic differences between cities can have large effects as well. We believe that an acknowledgment of the dependence on the choice of models would have made the paper stronger and clarified the limitations of its empirical claims (Gelman and Zelizer, 2015).

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.

[3] `http://www.plannedparenthood.org/health-topics/birth-control/menstrual-cycle-22144.htm`

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.

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

Bernardo, J. M. (1979). Reference posterior distributions for Bayesian inference. *Journal of the Royal Statistical Society B* **41**, 113–147.

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.

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

Chen Y., Ebenstein, A., Greenstone, M., and Li, H. (2013). Evidence on the impact of sustained exposure to air pollution on life expectancy from China's Huai River policy. *Proceedings of the National Academy of Sciences* **110**, 12936—12941.

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. (1982) 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.

Durante, K., Arsena, A., and Griskevicius, V. (2013). The fluctuating female vote: Politics, religion, and the ovulatory cycle. *Psychological Science* **24**, 1007–1016.

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, Chichester.

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

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

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. (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., 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.*

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

Gelman, A., Meng, X. L., and Stern, H. S. (1996). Posterior predictive assessment of model fitness via realized discrepancies (with discussion). *Statistica Sinica* **6**, 733—807.

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.*

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

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, Series C (Applied Statistics)* **62**, 309–369.

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.

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

Kuhn, T. S. (1962). *The Structure of Scientific Revolutions.* University of Chicago Press.

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.

Mayo, D. G. (2014). Objective/subjective, dirty hands, and all that. Error Statistics Philosophy blog, 16 Jan. `http://errorstatistics.com/2014/01/16/objectivesubjective-dirty-hands-and-all-that-gelmanwasserman-blogolog/`

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

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. (1978). Bayesian inference for causal effects: The role of randomization. *Annals of Statistics* **6**, 34–58.

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

Saari, C. (2005). The contribution of relational theory to social work practice. *Smith College Studies in Social Work* **75**, 3–14.

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.

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.

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

von Glasersfeld, E. (1995). *Radical Constructivism: A Way of Knowing and Learning.* London: Falmer Press.

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

Wang, W., Rothschild, D., Goel, S., and Gelman, A. (2014). Forecasting elections with non-representative polls. *International Journal of Forecasting.*

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