The Objectivity of Subjective Bayesian Inference

Jan Sprenger*

December 7, 2015

Abstract

Subjective Bayesianism is a major school of uncertain reasoning and statistical inference. Yet, it is often criticized for an apparent lack of objectivity. By and large, these criticisms come in three different forms. First, the lack of constraints on prior probabilities, second, the entanglement of statistical evidence and degree of belief, third, the apparent blindness to bias in experimental design. This paper responds to the above criticisms and argues in addition that frequentist statistics is no more objective than Bayesian statistics. In particular, the various arguments are related to different senses of scientific objectivity that philosophers have worked out in recent years.

1 Introduction

Subjective Bayesianism is a major school of uncertain reasoning and statistical inference that is steadfastly gaining popularity. It is based on the subjective interpretation of probability and describes how prior degree of belief in a scientific hypothesis is updated to posterior degree of belief. Since degrees of belief obey the axioms of probability, there is a straightforward connection between the mathematical theory of probability and the epistemological question of which hypothesis is confirmed by the evidence.

Yet, subjective Bayesian inference is often criticized for foundational reasons. More often than not, these criticisms take issue with the apparent lack of *objectivity*: "a notion of probability as personalistic degree of belief [...], by its very nature, is not focused

*Contact information: Tilburg Center for Logic, Ethics and Philosophy of Science (TiLPS), Tilburg University, P.O. Box 90153, 5000 LE Tilburg, The Netherlands. Email: j.sprenger@uvt.nl. Webpage: www.laeuferpaar.de

on the extraction and presentation of evidence of a public and objective kind" (Cox and Mayo, 2010, 298). This view is echoed in writings of well-known statisticians and philosophers of science such as Fisher (1956), Mayo (1996), Popper (2002) and Senn (2001, 2011).

Objectivity, however, is essential for a method of statistical inference, for it conveys an image of epistemic authority and strenghtens our trust in science. The 2009 "Climategate" affair and the recent "replication crisis" in psychology (i.e., the widespread failure to replicate experimental results due to various forms of bias, see Makel et al., 2012), illustrate how an apparent lack of objectivity weakens trust in scientific findings.

The objectivity-related criticisms of Bayesian inference come, by and large, in three different forms. First, the lack of constraints on prior probabilities, second, the entanglement of statistical evidence and degree of belief, third, the apparent blindness to bias in experimental design. In the light of these objections, one is tempted to conclude that Bayesian inference cannot produce objective knowledge, is not suitable for scientific communication and is therefore inferior to frequentist inference.

This paper responds to the above criticisms and debunks the view that subjective Bayesian inference cannot provide objective evidence (Section 2–4). I also argue that frequentist statistics is no more objective than Bayesian statistics. In particular, it augments the problem of publication bias in science (Section 5). The final section concludes and embeds our discussion into a broader debate about different senses of scientific objectivity (Section 6). Not all of the anti-Bayesian arguments do not refer to one and the same conception of scientific objectivity (see Douglas, 2004; Reiss and Sprenger, 2014, for a survey). Similarly, the defense of Bayesian inference concedes that Bayesian inference is—like any method of inference—not fully objective in every possible sense (e.g., intersubjective agreement), but it stresses that Bayesian inference promotes various important senses of objectivity.

The paper does not claim originality for all rebuttals and counterarguments. Partly, they are anticipated in the extant scientific literature. However, this paper takes a more foundational perspective than most publications in the scientific or statistical literature, and it relates technical and methodological arguments to the different senses of scientific objectivity worked out by philosophers. It systematizes the debate about the objectivity of Bayesian reasoning and shows how philosophical insights can be used in scientific practice.

2 Objection 1: The Choice of the Prior Distribution

Bayesian inference is based on prior probability distributions. Assume that you are interested in assessing a hypothesis $H_0 \in \mathcal{H}$. You represent your prior belief in H_0 by means of a probability distribution over the entire space of hypotheses $\mathcal{H}$. Assume further that your data D follow a definite probability distribution $p(D|H)$ under all possible hypotheses in $\mathcal{H}$. Then, your posterior degree of belief in the null hypothesis H_0 can be calculated by the formula

$$p(H_0|D) = \frac{p(H_0)p(D|H_0)}{p(D)} \tag{1}$$

where $p(D)$ is the marginal probability of data D. On the basis of the posterior probability $p(H_0|D)$, a Bayesian can then form a theoretical judgment about H_0 or make a practical decision. For example, if H_0 is the hypothesis that a new medical drug is not more efficacious than a placebo, and if H_0 is sufficiently probable given the data, then we will not pursue further development of the drug.

Subjective Bayesians such as Ramsey (1926) and de Finetti (1972) have stressed that in principle, *any* coherent prior probability distribution can be defended as rational. However, this attitude seems to jeopardize any claims to objectivity that subjective Bayesians could possibly make. Often, there is not sufficient background knowledge to establish consensus on prior probabilities. But if the choice of the prior is unconstrained, it is not clear what kind of epistemic warrant a Bayesian inference provides. After all, the choice of the prior can hide all kind of pernicious values, e.g., financial interests of the experiment sponsor. This is particularly worrying in sensitive subjects such as medicine, where the need for impartial inference methods is particularly high, due to the manifest financial interests in clinical trials and the ethical consequences of wrong decisions. As the medical methodologist Lemuel Moyé writes:

> Without specific safeguards, use of Bayesian procedures will set the stage for the entry of non-fact-based information that, unable to make it through the "evidence-based" front door, will sneak in through the back door of "prior distributions". There, it will wield its influence, perhaps wreaking havoc on the research's interpretation. (Moyé, 2008, 476)

The objection claims that Bayesians can bias the final result in their preferred direction by choosing an appropriate prior. This objection is thus based on the value-free ideal

that the core business of scientific reasoning, namely evaluating evidence, assessing and accepting theories, should be free of non-cognitive values and individual biases—a requirement that Bayesian inference seems to violate blatantly. Adherence to the value-free ideal has, however, in one form or another, been upheld as a trademark of scientific objectivity (e.g., Lacey, 1999; Reiss and Sprenger, 2014), and for practitioners, it plays an even greater role due to regulatory constraints and conflicts of interests. Even if one recognizes the philosophical problems with the value-free ideal, values should still not be allowed to *replace* scientific evidence (Douglas, 2008, 2009). How can Bayesian inference be safeguarded against this danger?

It is tempting to argue that various "objective Bayesian" techniques provide relief for the Bayesian. After all, these approaches build on a highly developed mathematical theory in order to uniquely determine prior probabilities where theoretical background knowledge and empirical track record provide none (Bernardo, 1979; Williamson, 2010). However, this escape route presupposes a waterproof philosophical justifcation of the objective Bayesian approaches, which is firstly difficult to achieve and secondly beside the scope of this paper (=a defense of subjective Bayesian reasoning). The point of this paper is rather to explain why *subjective* Bayesian inference can make claims to objectivity.

So a different defense is required. In fact, I will provide three of them.

The first defense notes that subjective opinion need not be the same as individual bias. Two medical doctors may, on the basis of their experience, give a different judgment about what might be a good therapy for a patient with a given set of symptoms. The fact that they disagree does not mean that one of them or both are biased in a certain way: they may just have enjoyed a different training, come from different disciplinary perspectives or have different experience in dealing with those symptoms. Prior probability distributions provide a way to make explicit a judgment that is fed by individual expertise and track record.

The second defense notes that also prior probabilities are open to rational criticism. Whenever a prior distribution is used, be its shape conventional or peculiar, the researcher should justify her particular choice and explain which considerations (theoretical and empirical ones) led her to this choice. This is also explicit in regulations for medical trials, such as the guidelines for the use of Bayesian statistics, issued by the Food and Drug Administration of the United States:

> We recommend you be prepared to clinically and statistically justify your

> choices of prior information. In addition, we recommend that you perform sensitivity analysis to check the robustness of your models to different choices of prior distributions. (US Food and Drug Administration, 2010)

The above quote hints to a second requirement in Bayesian reasoning: to perform a sensitivity analysis on the choice of the prior and to check whether the main result of the research remains intact under different prior assumptions. Such an analysis also contributes to scientific objectivity in terms of "convergent objectivity" (Douglas, 2004, 2009, 2011), according to which a scientific result can claim to be objective when it is validated from different assumptions and perspectives.

The bottom line of this defense is that the choice of the prior is just like any other modeling assumption in science open to criticism. Frequentist statisticians are no better off in this respect: they have to decide on the sample size, whether to perform a one-sided or a two-sided test, whether to use a uniform or a mixed effects model, whether to run a parametric or a non-parametric test, and so on. All these assumptions reveal to a certain extent prior expectations about the likely values of the unknown parameter—e.g., a small sample size makes sense if we are interested in detecting large effects, but not if we are interested in small effects. In fact, being explicit about the prior assumptions in the Bayesian framework makes it easier to criticize a particular choice, contributing to scientific objectivity in the sense of a process that is transparently conducted and open to rational criticism (Longino, 1990).

The third defense observes that the role of priors in Bayesian inference differs from the cliché picture that "any posterior can be justified by a suitable choice of the prior". This would indeed be true if posterior probabilities were the Bayesians' preferred measure of evidence. However, while posterior probabilities are indeed apt for individual decision-making, Bayesians typically use Bayes factors (Kass and Raftery, 1995) for the quantification and communication of scientific evidence. For examples from cognitive science, see Lee and Wagenmakers (2013) and Wetzels and Wagenmakers (2012). They are defined as the ratio of posterior odds and prior odds, are the standard choice for expressing the support for H_0 over the alternative H_1. Equivalently, the Bayes factor can be expressed as the integrated likelihood of H_0 over H_1 with data x.

$$B_{01}(x) \quad := \quad \frac{p(H_0|x)}{p(H_1|x)} \cdot \frac{p(H_1)}{p(H_0)} = \frac{\int_{\theta \in \Theta_0} p(x|\theta) p(\theta) d\theta}{\int_{\theta \in \Theta_1} p(x|\theta) p(\theta) d\theta} \tag{2}$$

In the second formulation, we see that the Bayes factor expresses the ratio of the probabilities of the data under H_0 and H_1, that is, the likelihoods. When these are composite hypotheses, the Bayes factor is still affected by the prior probability distriution, but in a more indirect way than the criticism claims: only the *relative weight of the constituents of H_0 and H_1* affects the weight of evidence. That is, we cannot manipulate the Bayes factor in favor of the null hypothesis by assigning it sufficiently high prior probability vis-à-vis the alternatives. Thus, the Bayesian cannot easily replace evidence with values.

In a nutshell, screwing up a Bayesian analysis with a biased prior is as easy or difficult as screwing up a non-Bayesian analysis with biased modeling assumptions. We now move to the next objection: that Bayesians mix up belief and evidence.

3 Objection 2: Belief vs. Evidence

The second objection contends that scientific reasoning, and statistical analysis in particular, is not about assessing the subjective probability of hypotheses, but about finding out whether a certain effect is real or due to chance. The task of science is to state the objective *evidence* for the truth of the hypothesis. In this view, the Bayesian statistician commits a category mistake: she tries to answer a question that scientists are not (and should not be) interested in. Statistical reasoning is about the truth of hypotheses and the evidence for them, not about subjective plausibility judgments. Ronald A. Fisher, one of the fathers of modern statistics, forcefully articulated this view:

> Advocates of inverse probabilities [ascribing probabilities to scientific hypotheses given some data, J.S.] are forced to regard mathematical probability, not as an objective quantity measured by observable frequencies, but as measuring merely psychological tendencies, theorems respecting which are useless for scientific purposes (Fisher, 1935, 6–7)

Royall (1997, 4) makes a similar distinction between three major questions in statistical analysis: "What should we believe?", "What should we do?" and "What is the evidence?". A good answer to one of them need not be a good answer to another question. In subjective Bayesian inference, belief and evidence seem to be entangled, however.

Underlying this objection is the idea of "concordant objectivity" Douglas (2009, 126–127) that assessments of evidence have to be intersubjectively agreed. As Quine

(1992, 5) stated it: "The requirement of intersubjectivity is what makes science objective." However, the "psychological tendencies" that correspond to personal degrees of belief do not fulfil this requirement. How can scientists uphold an image of objectivity and intersubjective agreement if their (Bayesian) data analysis yield different strengths of statistical evidence?

Many philosophers and scientists share the view that subjective Bayesian inference falls short of this goal. Williamson (2007) notes that "full objectivity—i.e. a single probability function that fits available evidence" cannot be achieved in the subjective Bayesian framework. Bem et al. (2011) quote an anonymous referee of their paper as saying

> I have great sympathy for the Bayesian position [...] The problem in implementing Bayesian statistics for scientific publications, however, is that such analyses are inherently subjective, by definition [...] with no objectively right answer as to what priors are appropriate. I do not see that as useful scientifically.

In other words, even if the priors are not contaminated by extra-scientific values (see Section 2), they still mirror individual perspectives. The divergence between these perspectives prevents us to reach intersubjective agreement on the observed evidence and makes the interpretation and communication of evidence a very delicate matter.

Our strategy in dealing with this objection is twofold: First, we will show how Bayesian measures of evidence justify objective claims. Second, it will be argued that frequentist inference is no guarantee for intersubjective agreement on the evidence. By contrast, the same data can look very differently if two different frequentists analyze them.

Let us first look at the standard Bayesian measure of evidence, the Bayes factor. As already mentioned, the Bayes factor in favor of H_0 over H_1 with data x is defined as follows:

$$B_{01}(x) \quad := \quad \frac{\int_{\theta \in H_0} p(x|\theta) p(\theta) d\theta}{\int_{\theta \in H_1} p(x|\theta) p(\theta) d\theta} \tag{3}$$

It has already been said that the Bayes factor only depends on the relative prior weight of the *components* of H_0 and H_1, but not on the prior probability of H_0 and H_1 as a whole. It is important to realize that this dependency is benign and not pernicious. Imagine the frequent case that we are testing the null hypothesis that a certain inter-

vention, e.g., taking vitamin C tablets as a cure for the common flu, has no effect at all: $H_0 : \theta = 0$ and $H_1 : \theta \neq 0$, where θ is the variable denoting the effect size. Of course, it is implausible that the effect of the vitamin C intervention is *exactly* zero: the tables will cause a biochemical reaction in the human body even if it is negligibly small. The test aims at finding out whether we can use the null hypothesis as a simple, precise, but strictly speaking wrong idealization of a complex reality. However, in order to assess whether a finding is evidence for or against H_0, we need to know which effect sizes are plausible at all. Only if this is clarified, we can state meaningfully that the observed results speak in favor of or against the null hypothesis.

This argument echoes the old insight that the term "evidence in favor of a hypothesis" does not make sense unless relativized to an alternative (Hacking, 1965; Spielman, 1974; Royall, 1997). The alternative $H_1 : \theta \neq 0$ is underspecified unless we make judgments of relative plausibility over the individual components of the hypothesis. In particular, the likelihood of H_1, $p(E|H_1)$, cannot be calculated otherwise. In the absence of clear theoretical guidance or a track record of past data, the Bayesian will plug in the subjective judgment of a knowledgeable scientist. It is hard to imagine what else he or she should do. Asking for fully intersubjective evidence in such a case demands a stronger conclusion than our epistemic situation warrants. Plausibility judgments are thus no danger to scientific objectivity, but required for a meaningful statistical analysis.

We now turn to the negative part of our response—arguing that frequentists suffer from problems which are greater than those of Bayesians.

The most widespread method of frequentist statistics is to test null hypotheses and assess the evidence with the help of *p-values* or observed *significance levels*. Given a statistic $z(X)$ that measures the disagreement of data X with the hypothesized parameter value, the p-value is equal to the probability (under the null) that z takes an even more extreme value than the one it has actually taken:

$$p := p_{H_0}(z|X| > z(x)) \tag{4}$$

where X denotes a random variable standing for the observed data and x the actually observed data. For example, if $p = 0.02$, this means that if the null hypothesis is true, a result that diverges even more from the null hypothesis than the actual result would only be expected in 2% of all cases.

What does such a p-value or "observed level of significance" mean from an inferential perspective? According to the classical frequentist school, the smaller the p-value, the stronger the evidence against the null hypothesis, and the less are we justified to believe that it is true:

> [...the feeling induced by a test of significance has an objective basis in that the probability statement on which it is based is a fact communicable to and verifiable by other rational minds. The level of significance in such cases fulfills the conditions of a measure of the rational grounds for the disbelief [in the null hypothesis] it engenders. (Fisher, 1956, 43)

One question to ask is whether p-values have indeed any connection to rational (dis)belief in the null hypothesis (see also Berger and Sellke, 1987). It is well-known that p-values alone do not suffice to infer to the improbability of the null hypothesis; some assumption on the prior plausibility of the null hypothesis has to be made. This "base rate fallacy" continues to haunt statistical practictioners, as observed in the surveys by Oakes (1986) and Fidler (2007). But my point is a different one. It claims that p-values are not suitable as measures of evidence even independent of their abuse in practice.

First, in a very large study, almost any effect size, even a negligibly small one, will be sufficient to trigger a "significant finding" and a very low p-value. Whereas p-values are indicative of larger—and *scientifically* more significant—effect sizes in smaller trials (e.g. Lindley, 1957; Sprenger, 2013; Robert, 2014). That is, p-values are a highly contextual measure of evidence whereas Bayes factors always denote the ratio by which a result is more expected under H_0 than under H_1. That is, Bayes factors measure the discriminatory power of the evidence. They are therefore much easier to interpret across different contexts and in this sense also more objective.

Second, even when the observed results are very unexpected under the null, they do not tell strongly against the null unless those results are more likely under some alternative hypothesis. For each experimental result, however unlikely it is, we can construct an alternative hypothesis that explains it well. And when, like in many cases of real hypothesis testing, there is *always* an alternative hypothesis which explains the data better than the null hypothesis (e.g., because the alternative comprises a large spectrum of parameter values), judgments of plausibility are required to determine whether a given low p-value really justifies rejection of the null. For example, rejecting the null hypothesis that a soothsayer has no extrasensory capacities—and inferring to the existence of such capacities—requires an extremely low p-value since we do not

have any scientifically plausible theory of extrasensory powers. On the other hand, the fairly implausible null hypothesis that exercise and body weight are uncorrelated may be appropriately rejected at the standard 5% level ($p < 0.05$). What I am arguing here is the following: requiring different observed significance levels in different contexts is nothing but a hidden judgment of plausibility which the frequentist makes as well as the Bayesian. Frequentists possess a higher degree of "procedural objectivity" (Douglas, 2009, 125–126) than Bayesians because their inference is more standardized, but the appearance of impartiality actually works backwards: valuable factors for statistical inference cannot be accommodated into their standardized procedures. As Fine (1998, 14) put it: "Bias and the impersonal are quite happy companions."

Third, frequentists may draw very different consequences from the same data. For example, when we conduct a simple experiment to learn about a parameter in a Bernoulli ("success of failure") trial such as the toss of a coin, we can either choose a Binomial design (fixed sample size) or a Negative Binomial design (fixed number of failures). For the same data set (e.g., nine successes and three failures), the calculated p-values for the null hypothesis $H_0 : \theta = 0.5$ against the one-sided alternative $H_1 : \theta > 0.5$ will be different. The p-value in the Binomial design (N=12) will be above the conventional significance cutoff $p = 0.05$, whereas the Negative Binomial design will interpret the same data as significant evidence against the null hypothesis. When both trials terminate at the same moment, with the same data set and still reach different conclusions, the *intentions in the head of the experimenter*—what they would have done if different data had been observed—influence the strength of the evidence (Edwards et al., 1963).

The dependency on these counterfactual intentions undermines the advantage in concordant objectivity and freedom of idiosyncratic bias that the frequentist claims to have over the Bayesian. Imagine two scientists arguing about the proper evidential interpretation of an experiment and then discovering that they only disagree because they had two different sampling protocols in mind! Frequentist inference does not provide an objective "view from nowhere in particular" (Nagel, 1986) any more than Bayesian inference does. The problem of retrieving the "correct" experimental design for calculating p-values is especially salient in more complex examples, such as scanning for correlations in a large set of variables, where exploratory data analysis is hard to separate from quantifying evidence and proper statistical inference.

Taken together, we see that the claim that p-values provide a more objective, intersubjectively compelling measure of evidence than Bayes factors is not tenable. It

is an illusion to neatly separate statistical evidence from judgments of plausibility. In fact, the Bayesian offers a more coherent, transparent and stringent picture than her frequentist counterpart. We now move to the next objection which maintains that Bayesian inference is blind to bias in experimental design.

4 Objection 3: Experimental Design and Error Control

The third objection can best be motivated with an example from medicine. Randomized Controlled Trials (RCTs) are currently the gold standard within evidence-based medicine (see Worrall, 2008, for a critical discussion). They are usually conducted as *sequential trials* allowing for monitoring for early signs of effectiveness or harm. In sequential trials, data are typically *monitored* as they accumulate. That is, we have interim looks at the data and we may decide to stop the trial before the planned sample size is reached. By terminating a trial when overwhelming evidence for the effectiveness or harmfulness of a new drug is available, the prohibitive costs of a medical trial can be limited and in-trial patients are protected against receiving inferior treatments.

However, such truncated trials are often seen as problematic. In a review of 134 trials stopped early for benefit, Montori et al. (2005) point to an inverse correlation between sample size and treatment effect: the smaller the sample size achieved by the trial at the moment of stopping, the larger the estimate it provided for the effect. These findings are supported by a more recent study by Bassler et al. (2010) where truncated trials report significantly higher effects than trials that were not stopped early. While the authors of these studies do not object to monitoring and truncating trials in general, they advocate that results (e.g., effect size estimates) from such trials be treated with caution. Truncating a trial seems to introduce a bias toward overestimating effect sizes. A good measure of statistical evidence should take this into account.

Bayesian measures of evidence such as the Bayes factor do not depend on sampling protocol or experimental design. By decoupling statistical inference method from the sampling protocol and the experimental design, the Bayesian is unable to discover that an experiment was conducted in a biased way. Indeed, critics of Bayesian inference such as Deborah Mayo (1996) complain that ignoring the sampling protocol (e.g., treating a truncated trial like a fixed sample trial) "can lead to a high probability of error, and [...] this high error probability is not reflected in the interpretation of data" (Mayo and Kruse, 2001) on the Bayesian and related accounts. In the context of medical research, the Bayesian seems to provide *carte blanche* for implementing any design

that favors the interests of the sponsor (e.g., a pharmaceutical company) rather than finding out the true efficacy of the drug.

The first thing to note is that higher effect sizes in truncated trials are not surprising, but *predictable* (Goodman et al., 2010): highly efficacious treatments will naturally be more prone to early termination for benefit. That is, when the actual effect is large, we will more probably observe a large effect in the population and decide to terminate the trial. Hence, the observed difference in estimated effect size is precisely what we should expect. Comparing truncated to completed trials amounts, as highlighted by Berry et al. (2010), to selecting the trials to be compared on the basis of their outcome.

In this context, prior knowledge or empirically-based prior expectations are highly relevant for sound decision-making. Imagine that we are interested in the relative risk reduction which a medical drug provides. A Bayesian represents her uncertainty by means of a prior probability distribution over that quantity. By means of Bayes' Theorem, this distribution is updated to a posterior probability distribution that synthesizes the observed evidence with the background knowledge. Then, the Bayesian framework naturally accounts for the intuition that truncated trials should be treated with caution: for the same observed effect size, small sample sizes change the prior distribution less than large sample sizes. The posterior distribution visualizes these differences in an intuitive way that can be directly used for decision-making (Goodman, 2007; Nardini and Sprenger, 2013).

In other words, the worry about Bayesian inference as unable to detect bias presupposes a frequentist understanding of the evidence. By amalgamating prior expectations with observed evidence for the purpose of decision-making, Bayesians automatically correct for the smaller sample size that truncated trials possess.

Second, that Bayes factors do not depend on the sampling protocol does not mean that Bayesians should just ignore all matters of experimental design. Procedural objectivity in the form of following certain regulatory constraints and standard procedures can be helpful to eliminate certain forms of institutional bias. In fact, guidelines for the use of Bayesian statistics (such as the ones issued by the Food and Drug Administration) stress that Bayesians should be as conscious and diligent in matters of experimental design as frequentists. For instance, also from a Bayesian perspective, a test with high type I and type II errors is evidently a bad test. The point of disagreement is different: while the frequentist bases her post-experimental evaluation of the evidence on the pre-experimental design and the properties of the entire experiment, the Bayesian considers these properties as essential for obtaining valid data, but as

orthogonal to the question of how to interpret them once they are in.

This concludes our discussion of the standard arguments against the lack of objectivity in Bayesian inference. We will now turn the tables and show that the use of frequentist inference is one of the probable causes of publication bias that many fields of science, especially psychology, suffer from.

5 Counterargument: Frequentism and Publication Bias

In recent years, the topic of *publication bias* in science has been in the limelight (e.g. Ioannidis, 2005; Ioannidis and Trikalinos, 2007; Francis, 2014; Francis et al., 2014). Authors mainly explore the effects of the so-called *file drawer* effect (Rosenthal, 1979) on science: the fact that results which fail to support an interesting hypotheses are rarely, if ever published. In the context of frequentist significance testing, this means that experimental data which fail to be significant at the conventional 5% level often don't make it past the peer review process—either because the referees fail to see an interesting result in non-significant data or because of self-censoring: authors don't submit such studies in the first place.

In his influential 2005 paper, "Why most published research findings are false", John Ioannidis sets up a simple mathematical model of how the failure to take into account non-significant findings biases our scientific knowledge. Let R be the ratio of the number of true causal relationships (corresponding to a false null hypothesis) to false causal relationships (=the null effect) in a scientific discipline. Assume that a causal relationship is tested with type I error rate α (that is, the probability of falsely rejecting the null and inferring that there is a genuine effect) and power $1-\beta$ (that is, the probability of inferring the alternative when it is true is $1-\beta$). In that case, the probability that a significant finding is true is

$$p = R\,\frac{1-\beta}{R\,(1-\beta)+\alpha} \tag{5}$$

which implies that a significant finding is more probably true than false if and only if $(1-\beta)R > \alpha$. Taking the conventional $\alpha = 0.05$ significance threshold and the already quite optimistic power of 80%, we see in Figure 1 that for a sufficiently small ratio of true to false causal hypotheses (that is, those with a negligible effect size), there may be more false than true published research findings.

Hence, what gets published under the heading of scientific findings or scientific

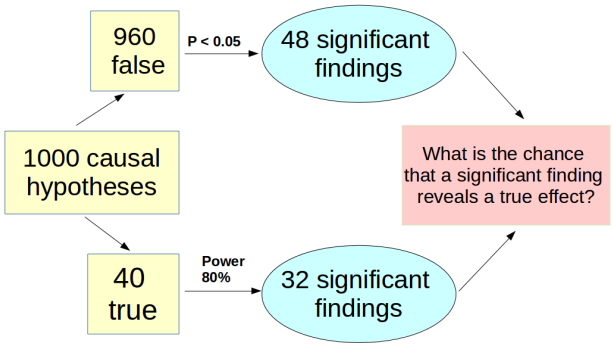

Figure 1: An illustration of publication bias through the file drawer effect in a concrete example.

knowledge is very different from the set of true hypotheses. By contrast, published findings may sketch a very biased picture, where a substantial part of published findings is not indicative of a real effect. As a consequence, many published findings are not replicable. This effect is the more pronounced the more significant or impressive an effect has been in the original publication (e.g., Makel et al., 2012). When several research teams independently test one and the same null hypothesis, publication bias can be especially painful since only the significant results will be published while the other ones end up in the file drawer. Even if the null hypothesis is literally true and most teams draw this conclusion, the *published* studies will indicate the presence of an effect and severely bias the available evidence.

Of course, non-significant results sometimes get published in practice. This may happen when they report an unsuccessful replication of a previously observed spectacular effect (e.g., the failure of Galak et al. (2012) to replicate the *psi* effect postulated by Bem (2011)), or when the journal employs a pre-registration practice where the papers are peer-reviewed before data collection. Nowadays, some high-ranking journals encourage this practice and publish results of pre-registered studies—in psychology, for instance *Social Psychology*, *Psychonomic Bulletin & Review* or *Perspectives on Psychological Science*). But overall, this is still rather an exception than the rule.

Admittedly, the choice to publish results (almost) only if they are significant is nothing that the frequentist methodology can be blamed for. No way of analyzing

data and quantifying evidence will eliminate publication bias: it is a very human tendency to publish exciting results (=strong evidence for a hypothesis) rather than boring results (=weak evidence). This problem needs to be tackled at the level of the reviewing process, for instance, by deciding on the acceptance of papers before data collection, rather than in the foundations of statistics. My point is, however, that the frequentist framework, with its dichotomous distinction between "significant" and "non-significant" results encourages and reinforces these bad practices. Only results that speak *against* the null hypothesis allow a quantification of the evidence in terms of p-values, whereas non-significant results (with p-values greater than .05) do not have a valid statistical interpretation, e.g., supporting the null hypothesis. This makes it especially difficult to publish results in favor of the null: when $p > .05$, there is no gradation of the strength of the evidence (and frequentists agree that large p-values are essentially meaningless), unlike in the "significant" case, where levels of evidence are measured according to whether $p < .05$, $p < .01$ or $p < .001$.

This is different in Bayesian inference. While the concept of "strong evidence" means "strong evidence against the null" in frequentist statistics, Bayesians can publish strong evidence against and for the null hypothesis. Even if we have a peer review system that demands a certain strength of the evidence in published research reports, Bayesian statistics is less entangled with publication bias than frequentist statistics. While a lot of inconclusive findings will still be suppressed, at least the publishable findings will be balanced in the sense that they are representative of the ratio of true to false null hypotheses.

The point I make here is almost Kuhnian: in the frequentist paradigm, it is impossible to say that strong evidence in favor of the null hypothesis has been found. Indeed, null hypotheses play an important role in science: they are simple, precise, easily testable and often express theoretically important claims such as equality of means in two populations, additivity of factors, causal independence, etc. (e.g., Gallistel, 2009; Morey et al., 2015). Finding evidence for the null hypotheses may sometimes be less spectacular than rejecting it, but especially if we are interested in scientific objectivity, we need an instrument to evaluate the evidence in their favor. The frequentist methodology, with its one-sided way of quantifying statistical evidence, therefore promotes publication bias and the file drawer effect, whereas Bayesian inference stands orthogonal to it.

6 Conclusion: A Digression on Scientific Objectivity

The concept of scientific objectivity is a notoriously difficult one, having various aspects and interpretations. It is a commonly shared view, though, that objective conclusions support the epistemic authority of science, distinguishing it from religion or political ideology. No wonder that statistical approaches are also valued according to their ability to provide an image of objectivity. Objective reasoning can manifest itself in different ways, e.g., leading to intersubjective agreement on evidence, freedom of extra-scientific values and idiosyncratic bias, standardization of inference procedures, a priority for evidence over values, responsiveness to criticism, and so on. The standard criticisms of Bayesian inference relate to selected aspects of the complex notion of scientific objectivity.

First, there is the idea that subjective Bayesian inference is particularly vulnerable to the intrusion of extra-scientific values since there is apparently no restriction on choosing prior probabilities. However, this objection overlooks that value-ladenness is to some extent inevitable in all forms of statistical inference (e.g., Rudner, 1953). Also methodologists in the frequentist camp such as Fisher (1956) and Mayo (1996) have emphasized the necessity to apply informed judgment in statistical inference, rather than just following automatic procedures. But even without the comparison to other frameworks, Bayesians have the resources to counter the objection. Prior degrees of belief can incorporate valuable expertise and background information, they can be fruitfully criticized like any statistical model assumption, and they can be separated from statistical evidence (Section 2).

Second, there is the inability to state scientific evidence in a way that is not entangled with (possibly idiosyncratic) judgments of plausibility. On the face of it, frequentist inference seems to have an edge here. But a closer look reveals that frequentist inference, if it wants to be menaningful, requires the same—albeit implicit—subjective assumptions about the plausibility of hypotheses as Bayesian inference. In the highly standardized picture of frequentist inference, important factors for assessing scientific evidence drop out of the picture. Therefore, Bayes factors are a convincing alternative which are also transparent about the subjective compoment in inference (Section 3 and 4). It is also argued that frequentist methods of evidence augment the problem of publication bias whereas Bayesian methods may help to make meaningful evidential statements in favor of the null hypothesis and thereby alleviate the problem (Section 5).

Third, also from the point of view of establishing intersubjective agreement, Bayes factors do not perform worse than their frequentist counterpart since they describe the degree to which the evidence changes one's prior probabilities, rather than the posterior probabilities themselves (Section 3).

Finally, I would like to gloss aspects of objectivity that relate to interaction and mutual criticism in a research community. Here, Helen Longino (1990) has forcefully argued that scientific objectivity is also about the structure of scientific discourse: the possibility of openly criticizing each other's assumptions, providing a floor for the exchange of rational arguments, etc. In this respect, Bayesian inference fares much better than frequentist inference, which only provides an image of schein-objectivity: Bayesian inference is perfectly honest and transparent about the assumptions it makes and distinguishes clearly between prior belief, evidence, and conclusions (=posterior belief). This allows for a straightforward detection of inappropriate bias, such as prior assumptions that heavily favor a particular hypothesis. Moreover, it provides a coherent framework for assessing what happens when the prior assumptions on a parameter value are varied. The dependency on individual degrees of beliefs, hidden and implicit in other schools of statistical inference, can be seen as an asset of subjective Bayesianism from the vantage point of scientific objectivity.

In the light of these arguments, claims that subjective Bayesians cannot quantify evidence in an objective way must be rejected as unjustified. They rely on a too simplified picture of scientific objectivity, on a caricature of Bayesian inference and on a blind eye regarding the shortcomings of classical, frequentist inference.

I would like to conclude with an apology. The treatment of various deep issues in statistical methodology in this paper was by necessity superficial. However, the paper was not intended as a principled comparison of subjective Bayesian and frequentist inference. Rather, I wanted to motivate that subjective Bayesian reasoning is an appropriate tool for quantifying evidence and for making objective scientific inferences. Common language may suggest the contrary, but this impression is based on a naïve and outdated conception of objective inference as "free of any subjective component". On a more appropriate reading of scientific objectivity that takes into account the diversity and complexity of that concept, subjective Bayesian inference is no less objective than its frequentist counterpart, perhaps even more.

References

Bassler, D., Briel, M., Montori, V. M., Lane, M., Glasziou, P., Zhou, Q., Heels-Ansdell, D., Walter, S. D., Guyatt, G. H., N Flynn, D., and Others (2010). Stopping randomized trials early for benefit and estimation of treatment effects. *JAMA*, 303(12):1180–1187.

Bem, D. J. (2011). Feeling the Future: Experimental Evidence for Anomalous Retroactive Influences on Cognition and Affect. *Journal of Personality and Social Psychology*, 100(3):407–425.

Bem, D. J., Utts, J., and Johnson, W. O. (2011). Must psychologists change the way they analyze their data? *Journal of personality and social psychology*, 101(4):716–719.

Berger, J. O. and Sellke, T. (1987). Testing a point null hypothesis: the irreconcilability of P values and evidence. *Journal of the American Statistical Association*, 82:112–122.

Bernardo, J. M. (1979). Reference Posterior Distributions for Bayesian Inference. *Journal of the Royal Statistical Society. Series B (Methodological)*, 41:113—-147.

Berry, S. M., Carlin, B. P., and Connor, J. (2010). Bias and Trials Stopped Early for Benefit. *JAMA*, 304(2):156.

Cox, D. and Mayo, D. G. (2010). Objectivity and Conditionality in Frequentist Inference. In Mayo, D. G. and Spanos, A., editors, *Error and Inference: Recent Exchanges on Experimental Reasoning, Reliability and the Objectivity and Rationality of Science*, chapter 2, pages 276–304. Cambridge University Press, Cambridge.

de Finetti, B. (1972). *Probability, Induction and Statistics: the Art of Guessing*. John Wiley & Sons, New York.

Douglas, H. (2004). The irreducible complexity of objectivity. *Synthese*, 138(3):453–473.

Douglas, H. (2008). The Role of Values in Expert Reasoning. *Public Affairs Quarterly*, 22:1–18.

Douglas, H. (2009). *Science, policy, and the value-free ideal*. Pittsburgh University Press, Pittsburgh.

Douglas, H. (2011). The SAGE Handbook of the Philosophy of Social Sciences. pages 283–306. SAGE Publications, London.

Edwards, W., Lindman, H., and Savage, L. J. (1963). Bayesian statistical inference for psychological research. *Psychological Review*, 70:193–242.

Fidler, F. (2007). *From Statistical Significance to Effect Estimation: Statistical reform in psychology, medicine and ecology*.

Fine, A. (1998). The Viewpoint of No-One in Particular. *Proceedings and Addresses of the American Philosophical Association*, 72(2):7–20.

Fisher, R. A. (1935). *The design of experiments.* Oliver & Boyd, Edinburgh.

Fisher, R. A. (1956). *Statistical methods and scientific inference.* Hafner, New York.

Francis, G. (2014). The frequency of excess success for articles in Psychological Science. *Psychonomic Bulletin & Review*, 21(5):1180–1187.

Francis, G., Tanzman, J., and Matthews, W. J. (2014). Excess success for psychology articles in the journal science. *PloS one*, 9(12):e114255.

Galak, J., LeBoeuf, R. A., Nelson, L. D., and Simmons, J. P. (2012). Correcting the Past: Failures to Replicate Psi. *Journal of Personality and Social Psychology*, 103:933—-948.

Gallistel, C. R. (2009). The importance of proving the null. *Psychological review*, 116(2):439—-453.

Goodman, S. J., Berry, D., and Wittes, J. (2010). Bias and Trials Stopped Early for Benefit. *JAMA*, 304(2):157.

Goodman, S. N. (2007). Stopping at nothing? Some dilemmas of data monitoring in clinical trials. *Annals of Internal Medicine*, 146(12):882.

Hacking, I. (1965). *Logic of statistical inference*. Cambridge University Press, Cambridge.

Ioannidis, J. and Trikalinos, T. (2007). An exploratory test for an excess of significant findings. *Clinical trials*, 4(3):245—–253.

Ioannidis, J. P. A. (2005). Why most published research findings are false. *PLoS Medicine*, 2:e124.

Kass, R. E. and Raftery, A. E. (1995). Bayes Factors. *Journal of the American Statistical Association*, 90:773–795.

Lacey, H. (1999). *Is Science Value Free? Values and Scientific Understanding*. Routledge, London.

Lee, M. D. and Wagenmakers, E.-J. (2013). *Bayesian Cognitive Modeling: A Practical Course*. Cambridge University Press, Cambridge.

Lindley, D. V. (1957). A statistical paradox. *Biometrika*, 44:187–192.

Longino, H. (1990). *Science as Social Knowledge: Values and Objectivity in Scientific Inquiry*. Princeton University Press, Princeton, NJ.

Makel, M. C., Plucker, J. A., and Hegarty, B. (2012). Replications in Psychology Research: How Often Do They Really Occur? *Perspectives on Psychological Science*, 7(6):537–542.

Mayo, D. G. (1996). *Error and the growth of experimental knowledge*. University of Chicago Press, Chicago.

Mayo, D. G. and Kruse, M. (2001). Principles of inference and their consequences. In Corfield, D. and Williamson, J., editors, *Foundations of Bayesianism*. Springer, New York.

Montori, V. M., Devereaux, P. J., Adhikari, N. K. J., Burns, K. E. A., Eggert, C. H., Briel, M., Lacchetti, C., Leung, T. W., Darling, E., Bryant, D. M., Bucher, H. C., Schünemann, H. J., Meade, M. O., Cook, D. J., Erwin, P. J., Sood, A., Sood, R., Lo, B., Thompson, C. A., Zhou, Q., Mills, E., and Guyatt, G. H. (2005). Randomized trials stopped early for benefit: a systematic review. *JAMA : the journal of the American Medical Association*, 294:2203–2209.

Morey, R. D., Hoekstra, R., Rouder, J. N., Lee, M. D., and Wagenmakers, E.-J. (2015). The fallacy of placing confidence in confidence intervals. *Psychonomic Bulletin & Review*.

Moyé, L. A. (2008). Bayesians in clinical trials: Asleep at the switch. *Statistics in Medicine*, 27:469–482.

Nagel, T. (1986). *The View from Nowhere*. Oxford University Press, Oxford.

Nardini, C. and Sprenger, J. (2013). Bias and Conditioning in Sequential Medical Trials. *Philosophy of Science*, 80:1053–1064.

Oakes, M. W. (1986). *Statistical inference*. Wiley, New York.

Popper, K. R. (2002). *Logic of Scientific Discovery*. Routledge, London.

Quine, W. (1992). *Pursuit of Truth*. Harvard University Press, Cambridge MA.

Ramsey, F. P. (1926). Truth and Probability. In Mellor, D. H., editor, *Philosophical Papers*, pages 52–94. Cambridge University Press, Cambridge.

Reiss, J. and Sprenger, J. (2014). Scientific Objectivity. In Zalta, E. N., editor, *Stanford Encyclopedia of Philosophy*.

Robert, C. (2014). On the Jeffreys-Lindley Paradox. *Philosophy of Science*, 81:216—-232.

Rosenthal, R. (1979). The file drawer problem and tolerance for null results. *Psychological Bulletin*, 86(3):638–641.

Royall, R. (1997). *Statistical Evidence: A Likelihood Paradigm*. Chapman & Hall, London.

Rudner, R. (1953). The scientist qua scientist makes value judgments. *Philosophy of Science*, 10:1—-6.

Senn, S. (2001). Two cheers for P-values? *Journal of epidemiology and biostatistics*, 6(2):193–204; discussion 205–210.

Senn, S. (2011). You may believe you are a Bayesian but you are probably wrong. *Rationality, Markets and Morals*, 2:48–66.

Spielman, S. (1974). The logic of tests of significance. *Philosophy of Science*, 41(3):211–226.

Sprenger, J. (2013). Testing a Precise Null Hypothesis: The Case of Lindley's Paradox. *Philosophy of Science*, 80:733–744.

US Food and Drug Administration (2010). Guidance for the Use of Bayesian Statistics in Medical Device Clinical Trials.

Wetzels, R. and Wagenmakers, E.-J. (2012). A default Bayesian hypothesis test for correlations and partial correlations. *Psychonomic Bulletin & Review*, 19:1057–1064.

Williamson, J. (2007). Motivating Objective Bayesianism: From Empirical Constraints to Objective Probabilities. In Harper, W. and Wheeler, G., editors, *Probability and*

Inference: Essays in Honour of Henry E. Kyburg Jr., pages 151–179. College Publications, London.

Williamson, J. (2010). *In Defence of Objective Bayesianism*. Oxford University Press, Oxford.

Worrall, J. (2008). Evidence and Ethics in Medicine. *Perspectives in Biology and Medicine*, 51(3):418–431.