entropy

MDPI

Article

Measuring and Controlling Bias for Some Bayesian Inferences and the Relation to Frequentist Criteria

Michael Evans * and Yang Guo

Department of Statistical Sciences, University of Toronto, Toronto, ON M5G 1Z5, Canada;
ygj.guo@mail.utoronto.ca
* Correspondence: mevans@utstat.utoronto.ca; Tel.: +1-416-287-7274

Abstract: A common concern with Bayesian methodology in scientific contexts is that inferences can be heavily influenced by subjective biases. As presented here, there are two types of bias for some quantity of interest: bias against and bias in favor. Based upon the principle of evidence, it is shown how to measure and control these biases for both hypothesis assessment and estimation problems. Optimality results are established for the principle of evidence as the basis of the approach to these problems. A close relationship is established between measuring bias in Bayesian inferences and frequentist properties that hold for any proper prior. This leads to a possible resolution to an apparent conflict between these approaches to statistical reasoning. Frequentism is seen as establishing figures of merit for a statistical study, while Bayes determines the inferences based upon statistical evidence.

Keywords: principle of evidence; bias against; bias in favor; plausible region; frequentism; confidence

check for updates

Citation: Evans, M.; Guo, Y. Measuring and Controlling Bias for Some Bayesian Inferences and the Relation to Frequentist Criteria. *Entropy* **2021**, *23*, 190. https://doi.org/10.3390/e23020190

Academic Editors: Jacinto Martín and María Isabel Parra Arévalo

Received: 9 January 2021
Accepted: 1 February 2021
Published: 4 February 2021

1. Introduction

A serious concern with Bayesian methodology is that the choice of the prior could result in conclusions that to some degree are predetermined before seeing the data. In certain circumstances, this is correct. This can be seen by considering the problem associated with what is known as the Jeffreys–Lindley paradox where posterior probabilities of hypotheses, as well as associated Bayes factors, will produce increasing support for the hypothesis as the prior becomes more diffuse. Thus, while one may feel that a very diffuse prior is putting in very little information, it is in fact biasing the results in favor of the hypothesis in the sense that that there is a significant prior probability that evidence will be found in favor of the hypothesized value when it is false. It has been argued, see [1,2], that the measurement and control of bias is a key element of a Bayesian analysis as, without it, and the assurance that bias is minimal, the validity of any inference is suspect.

While attempts have been made to avoid the Jeffreys–Lindley paradox through the choice of the prior, modifying the prior to avoid bias is contrary to the ideals of a Bayesian analysis which requires the elicitation of a prior based upon knowledge of the phenomenon under study. Why should one change such a prior because of bias? Indeed, there is bias in favor and bias against and typically choosing a prior to minimize one type of bias simply increases the other. Roughly speaking, in a hypothesis assessment problem, bias against means that there is a significant prior probability of finding evidence against a hypothesized value when it is true, and bias in favor means that there is a significant prior probability of finding evidence in favor of a hypothesized value when it is false. The real method for controlling bias of both types is through the amount of data collected. Bias can be measured post-hoc, and it then provides a way to assess the weight that should be given to the results of an analysis. For example, if a study concludes that there is evidence in favor of a hypothesis, but it can be shown that there was a high prior probability that such evidence would be obtained, then the results of such an analysis can't be considered to be reliable.

Previous discussion concerning bias for Bayesian methodology has focused on hypothesis assessment and, in many ways, this is a natural starting point. This paper is concerned with adding some aspects to those developments and to extending the approach to estimation and prediction problems as discussed in Section 3.3 where bias in favor and bias against are expressed in terms of a priori coverage probabilities. Furthermore, it is argued that measuring and controlling bias is essentially frequentist in nature. Although not the same, it is convenient to think of bias against in a hypothesis assessment problem as playing a role similar to the size in a frequentist hypothesis test or, in an estimation problem, playing a role similar to 1 minus the coverage probability of a confidence region. Bias in favor can be thought of as somewhat similar to power in a hypothesis assessment problem and simlar to the probability of a confidence region covering a false value in an estimation problem. Thus, consideration of bias leads to a degree of unification between different ways of thinking about statistical reasoning.

The measurement of bias, and thus its control, is dependent upon measuring evidence. The *principle of evidence* is adopted here: *evidence in favor of a specific value of an unknown occurs when the posterior probability of the value is greater than its prior probability, evidence against occurs when the posterior probability of the value is less than its prior probability and there is no evidence either way when these are equal.* The major part of what is discussed here depends only on this simple principle, but sometimes a numerical measure is needed and, for this, we use the *relative belief ratio* defined as the ratio of the posterior to prior probability. The relative belief ratio is related to the Bayes factor but has some nicer properties such as providing a measure of the evidence for each value of a parameter without the need to modify the prior.

The inferences discussed here are based on the relative belief ratio and these inferences are invariant to any 1–1, increasing function of this quantity. For example, the logarithm of the relative belief ratio can be used instead to derive inferences. The expected value of the logarithm of the relative belief ratio under the posterior is the relative entropy, also called the Kullback–Leibler divergence, between the posterior and prior. This is an object of considerable interest in and of itself and, from the perspective of measuring evidence, can be considered as a measure of how much evidence the observed data are providing about the unknown parameter value in question. This aspect does not play a role here, however, but indicates a close association between the measurement of statistical evidence and the concept of entropy. In addition, many divergence measures involve the relative belief ratio and play a role in [3], which is concerned with checking for prior-data conflict.

There is not much discussion in the Bayesian literature of the notion of bias in the sense that is meant here. There is considerable discussion, however, concerning the Jeffreys–Lindley paradox and our position is that bias plays a key role in the issues that arise. Relevant recent papers on this include [4–9], and these contain extensive background references. Ref. [10] is concerned with the validation of quantum theory using Bayesian methodology applied to well-known data sets, and the principle of evidence and an assessment of the bias play a key role in the argument.

As already noted, the approach to inference and the measurement of bias adopted here is dependent on the principle of evidence. This principle is not well-known in the statistical community and so Section 2 contains a discussion of this principle and why it is felt to be an appropriate basis for the development of a theory of inference. In Section 3, the concepts that underlie our approach to bias measurement are defined, and their properties are considered and illustrated via a simple example where the Jeffreys–Lindley paradox is relevant. In addition, it is seen that a well-known p-value does not satisfy the principle of evidence but can still be used to characterize evidence for or against provided the significance level goes to 0 with increasing sample size or increasing diffuseness of the prior. In Section 4, the relationship with frequentism is discussed and a number of optimality results are established for the approach taken here to measure and control bias. In Section 5, a variety of examples are considered and analyzed from the point-of-view of bias. All proofs of theorems are in the Appendix A.

2. Statistical Evidence

Attempts to develop a theory of inference based upon a definition, or at least provide a characterization, of statistical evidence exist in the statistical literature. For example, see [2,11–17]. The treatments in [12,14] have some aspects in common with the approach taken here, but there are also substantial differences. There is a significant amount of discussion of statistical evidence in the philosophy of science literature and this is much closer in spirit to the treatment here. For example, see [18] p. 6, where it is stated "for a fact e to be evidence that a hypothesis h is true, it is both necessary and sufficient for e to increase h's probability over its prior probability" which is what is called the principle of evidence here.

2.1. The Principle of Evidence

One characteristic of our, and the philosophical, treatment is that evidence is a probabilistic concept and thus a proper definition only requires a single probability model as opposed to a statistical model. This explains in part why our treatment requires a proper prior as then there is a joint probability model for the model parameter and data. The following two examples illustrate the relevance of characterizing evidence in such a context. Example 1 is a simple game of chance where the probabilities in question are unambiguous. The utility aspects of the game are ignored because these are irrelevant to the discussion of evidence but surely are relevant if some action like betting was involved. This is characteristic of the treatment here where loss functions play no role in the characterization of evidence but do play a role in determining actions when required as discussed in the well-known Example 2. The examples also illustrate that characterizing evidence in favor of or against is not enough, as it is necessary to also say something about the strength of the evidence.

Example 1. *Card game.*

Suppose that there are two players in a card game, labeled I and II, and each is dealt m cards, where $2 \leq m \leq 26$, from a randomly shuffled deck of 52 playing cards. Further suppose that player I, after seeing their hand, is concerned, for whatever reason dependent on the rules of the game, with the truth or falsity of the hypothesis H_0: player II has exactly two aces. It seems clear that the hand of player I will contain evidence concerning this. For example, if player I has three or four aces in their hand, then there is categorical evidence that H_0 is false. However, what about the evidence when the event observed is $C_k =$ "the number of aces in the hand of player I is k" with $k = 0, 1,$ or 2?

There are two questions to be answered: (i) is there evidence in favor of or against H_0 and (ii) how strong is this evidence? The prior probability $P(H_0)$ and posterior probability $P(H_0 \,|\, C_k)$ that H_0 is true are provided in Table 1 for various (k, m). What conclusions can be drawn from this table? In every case, other than $(m, k) = (25, 2), (26, 2)$, the conditional probability $P(H_0 \,|\, C_k)$ does not support H_0 being true. In fact, in many cases, some would argue that the value of this probability indicates evidence against H_0. This points to a significant problem with trying to use probabilities to determine evidence, as it is not at all clear what the cutoff should be to determine evidence for or against H_0. It seems clear, however, that if the data, here the observation that C_k is true, has increased belief in H_0 over initial beliefs, then there is evidence in the data pointing to the truth of H_0. Whether or not the posterior probability is greater than the prior probability is indicated by $RB(H_0 \,|\, C_k) = P(H_0 \,|\, C_k)/P(H_0)$, the relative belief ratio of H_0, being greater than 1. Certainly, it is intuitive that, when $k = 0$, then our belief in H_0 being true, a posteriori could increase, but, from the table and some reflection, it is clear that this cannot always be true as the amount of data, m in this case, grows. While $k = 0$ is evidence in favor of H_0, it is evidence against only for $m = 25, 26$. The relationship between the prior probabilities and posterior probabilities is somewhat subtle and not easy to predict, but a comparison of these quantities makes it clear when there is evidence in favor of H_0 and when there isn't. This answers question (i).

The measurement of the strength of evidence is not always obvious, but, in this case, effectively a binary event, the posterior probability of the event in question seems like a reasonable approach as it is measuring the belief that the event in question is true. Thus, if we get evidence in favor of H_0 and $P(H_0 \mid C_k)$ is small, then this suggests that the evidence can only be regarded as weak and similarly if there is evidence against H_0 and $P(H_0 \mid C_k)$ is large, then there is only weak evidence against H_0. Some might argue that a large value of $P(H_0 \mid C_k)$ should always be evidence in favor of H_0, but note that the data could contradict this by resulting in a decrease from a larger initial probability. Measuring strength in this way, the table indicates that there is strong evidence in favor of H_0 with $(m, k) = (25, 2), (26, 2)$ and weak to moderate evidence in favor otherwise when $RB(H_0 \mid C_k) > 1$. By contrast, there is typically quite strong evidence against H_0 in cases where $RB(H_0 \mid C_k) < 1$ with the exception of $(m, k) = (10, 1), (25, 1)$. Intuitively, it couldn't be expected that there would be strong evidence in favor of H_0 for small m, but there can still be evidence in favor. Note that a comparison, for $m = 2$ and 20, of the values of $RB(H_0 \mid C_0)$ illustrates that the relative belief ratio itself does not provide a measure of the strength of the evidence in favor. In general, the value of a relative belief ratio needs to be calibrated and the posterior probability of H_0 is a natural way to do this here.

Table 1. Probabilities and relative belief ratios for H_0 in Example 1.

	$P(H_0)$		$P(H_0 \mid C_k)$	$RB(H_0 \mid C_k)$
$m = 2$	0.0045	$k = 0$	0.0049	1.0824
		$k = 1$	0.0024	0.5412
		$k = 2$	0.0008	0.1804
$m = 5$	0.0399	$k = 0$	0.0483	1.2089
		$k = 1$	0.0259	0.6487
		$k = 2$	0.0093	0.2317
$m = 10$	0.1431	$k = 0$	0.1994	1.3934
		$k = 1$	0.1254	0.8765
		$k = 2$	0.0522	0.3652
$m = 20$	0.3481	$k = 0$	0.3487	1.0018
		$k = 1$	0.4597	1.3205
		$k = 2$	0.3831	1.1004
$m = 25$	0.3890	$k = 0$	0.0171	0.0439
		$k = 1$	0.2051	0.5274
		$k = 2$	0.8547	2.1974
$m = 26$	0.3902	$k = 0$	0.0000	0.0000
		$k = 1$	0.0000	0.0000
		$k = 2$	1.0000	2.5630

Example 2. *Prosecutor's fallacy.*

Assume a uniform probability distribution on a population of size N of which some member has committed a crime. DNA evidence has been left at the crime scene and suppose that this trait is shared by $m \ll N$ of the population. A prosecutor is criticized because they conclude that, because the trait is rare and a particular member possesses the trait, they are guilty. In fact, $P(\text{“has trait”} \mid \text{“guilty”}) = 1$ is misinterpreted as the probability of guilt rather than $P(\text{“guilty”} \mid \text{“has trait”}) = 1/m$, which is small if m is large. However, this probability does not reflect the evidence of guilt. If you have the trait, then clearly this is evidence in favor of guilt and indeed $RB(\text{“guilty”} \mid \text{“has trait”}) = N/m > 1$ and $P(\text{“guilty”} \mid \text{“has trait”}) = 1/m$. Thus, there is evidence of guilt, and the prosecutor is correct to conclude this. However, the evidence is weak whenever m is large and a conviction then does not seem appropriate. Since the posterior probability of "not guilty" is large whenever m is, it may seem obvious to conclude this. However, suppose that "guilty" corresponds to being a carrier of a highly infectious deadly disease and "has trait" corresponds to some positive, but not definitive, test for this. The same numbers should undoubtedly lead to a quarantine. Thus, the utilities determine the action taken and not just the evidence.

2.2. Confirmation Theory

As noted, discussion concerning statistical evidence has a long history, although mainly in the philosophy of science literature, where it is sometimes referred to as confirmation theory. An introduction to confirmation theory can be found in [19], but the history of this topic is much older. For example, see Appendix ix in [20] where, with x and y denoting events, the following is stated.

> If we are asked to give a criterion of the fact that the evidence y supports or corroborates a statement x, the most obvious reply is: that y *increases* the probability of x.

The book [20] references older papers and some sources cite [21] where the relative belief ratio $RB(A \,|\, B)$ is called the *coefficient of influence of* B *upon* A. In the Confirmation entry in [22], the definition of *probabilistic relevance confirmation* is what has been called here the principle of evidence. The following quote is from the third paragraph of this entry and it underlines the importance of this topic.

> Confirmation theory has proven a rather difficult endeavour. In principle, it would aim at providing understanding and guidance for tasks such as diagnosis, prediction, and learning in virtually any area of inquiry. However, popular accounts of confirmation have often been taken to run into trouble even when faced with toy philosophical examples. Be that as it may, there is at least one real-world kind of activity that has remained a prevalent target and benchmark, i.e., scientific reasoning, and especially key episodes from the history of modern and contemporary natural science. The motivation for this is easily figured out. Mature sciences seem to have been uniquely effective in relying on observed evidence to establish extremely general, powerful, and sophisticated theories. Indeed, being capable of receiving genuine support from empirical evidence is itself a very distinctive trait of scientific hypotheses as compared to other kinds of statements. A philosophical characterization of what science is would then seem to require an understanding of the logic of confirmation. In addition, thus, traditionally, confirmation theory has come to be a central concern of philosophers of science.

As far as we know, Ref. [2] summarizes one of the first attempts to use the principle of evidence as a basis for a theory of statistical inference. Some of the paradoxes/puzzles that arise in the philosophical literature, such as Hempel's the Raven paradox, are discussed there. Adding the measurement of the strength of evidence and the a priori measurement of bias to the principle of evidence leads to the resolution of many difficulties, see [2]. Whether one is convinced of the value of the principle of evidence or not, this is an idea that needs to be better known and investigated by statisticians.

2.3. Popper's Principle of Science as Falsification

Another aspect requiring comment is that the principle of evidence allows for finding either evidence against or evidence in favor of a hypothesis while, for example, a p-value cannot find evidence in favor. This one-sided aspect of a p-value is often justified by Popper's idea that the role of science lies in falsification of hypotheses and not their confirmation. In the context of Examples 1 and 2, this seems wrong as the hypothesis in question is either true or false, so it is desirable to be able to find evidence either way.

When applied to a statistical context, at least as formulated in Section 3, inferences about a quantity of interest are dependent on the choice of a statistical model and a prior. It is well understood that the model is typically false and it isn't meaningful to talk of the truth or falsity of the prior. Since there is only one chosen model, it can only be falsified via model checking rather than confirmed, namely, determining if the observed data are in the tails of every distribution in the model. Actually, all that is being asked in such a procedure is whether or not the model is at least reasonably compatible with the observed data. Similarly, the prior is checked through checking for prior-data conflict, namely, given that the model has passed its check, is there an indication that the true value lies in the tails of the prior. For example, see [3,23,24] for some discussion. Again, all that is being asked is whether or not the prior is at least reasonably compatible with the data.

For checking the model or checking the prior, there is one object that is being considered. Thus, it makes sense that only an indication that the entity in question is not appropriate is available, and a p-value can play a role in this aspect of a statistical argument. However, when making an inference, the model is accepted as being correct and, as such, one of the distributions in the model is true, and so it is natural to want to be able to find evidence in favor of or against a specific value of an object dependent on the true distribution. This situation is analogous to what arises in logic where a sound argument is distinguished from a valid argument. A logical argument is based upon premises and rules of inference like modus ponens. An argument is valid if the rules of logic are correctly applied to obtain the conclusions. However, an argument is sound only if the argument is valid and the premises are true. It is a basic rule of logical reasoning that one doesn't confound the correctness of the argument with the correctness of the premises. In the statistical context, there may indeed be problems with the model or prior, but the inference step, which assumes the correctness of the model and prior, needs to be able to find evidence in favor as well as evidence against a particular value of the object of interest. As part of the general approach as presented in [2], both model checking and checking for prior-data conflict are advocated before inference. If there are serious problems with either, then modifications of the ingredients are in order, but this is not the topic of this paper where it is assumed that the model and prior are acceptable. Thus, Popper's falsification idea plays a role but not in the inference step.

3. Evidence and Bias

For the discussion here, there is a model $\{f_\theta : \theta \in \Theta\}$, given by densities f_θ, for data x and a proper prior probability distribution given by density π. It is supposed that interest is in inferences about $\psi = \Psi(\theta)$, where $\Psi : \Theta \to \Psi$ is onto and for economy the same notation is used for the function and its range. For the most part, it is safe to assume all the probability distributions are discrete with results for the continuous case obtained by taking limits.

A measure of the evidence that $\psi \in \Psi$ is the true value is given by the relative belief ratio

$$RB_\Psi(\psi \,|\, x) = \lim_{\delta \to 0} \Pi_\Psi(N_\delta(\psi) \,|\, x)/\Pi_\Psi(N_\delta(\psi)) = \pi_\Psi(\psi \,|\, x)/\pi_\Psi(\psi) \tag{1}$$

where $\Pi_\Psi, \Pi_\Psi(\cdot \,|\, x)$ are the prior and posterior probability measures of Ψ with densities π_Ψ and $\pi_\Psi(\cdot \,|\, x)$, respectively, and $N_\delta(\psi)$ is a sequence of sets converging nicely to $\{\psi\}$. The last equality in (1) requires some conditions, but the prior density positive and continuous at ψ is enough. In addition, when $\Psi = I_A$ for $A \subset \Theta$, the indicator of A, then we write $RB(A \,|\, x)$ for $RB_\Psi(1 \,|\, x)$. Thus, $RB_\Psi(\psi \,|\, x) > 1$ implies evidence for the true value being ψ,, etc. It is also possible that a prior is dependent on previous data. In such a situation, it is natural to replace π_Ψ in (1) by the initial prior, as the posterior remains the same, but now the evidence measure is based on all of the observed data. There may be contexts, however, where the concern is only with the evidence provided by the additional data, for example, as when new data arise from random sampling from the relevant population(s), but the first dataset came from an observational study.

Any *valid* measure of evidence should satisfy the principle of evidence, namely, the existence of a cut-off value that determines evidence for or against as prescribed by the principle. Naturally, this cut-off is 1 for the relative belief ratio. The Bayes factor is also a valid measure of evidence and with the same cut-off. When $\Pi_\Psi(A) > 0$, then the Bayes factor of A equals $RB(A\,|\,x)/RB(A^c\,|\,x)$ and thus can be defined in terms of the relative belief ratio, but not conversely. In addition, $RB(A\,|\,x) > 1$ iff $RB(A^c\,|\,x) < 1$ and thus the Bayes factor is not really a comparison of the evidence for A being true with the evidence for its negation. In the continuous case, if we define the Bayes factor for ψ as a limit as in (1), then this limit equals $RB_\Psi(\psi\,|\,x)$. Further discussion on the choice of a measure of evidence can be found in [2] as there are other candidates beyond these two. One significant advantage for the relative belief ratio is that all inferences derived based on it are invariant under smooth reparameterizations. Furthermore, the relative belief ratio only serves to order the values of $\psi \in \Psi$ with respect to evidence, and the value $RB_\Psi(\psi\,|\,x)$ is not to be considered as measuring evidence on a universal scale. It is important to note that the discussion of bias here depends only on the principle of evidence and is the same no matter what valid measure of evidence is used.

Since the model and prior are subjectively chosen, the characterization and measurement of statistical evidence has a subjective component. This creates the possibility that these choices are biased, namely, they were chosen with some goal in mind other than letting the data determine the conclusions. Model checking and checking for prior-data conflict exposes these choices to criticism via the data, but these checks will not reveal inappropriate conduct like tailoring a model or prior based on the observed data. Perhaps a more important check on such behavior is to measure and control bias. As will now be shown, controlling the bias through the a priori determination of the amount of data collected can leave us with greater confidence that the data are the primary driver of whatever inferences are drawn, and this is surely the goal in scientific applications. Thus, while informed subjective choices are a good thing, there are also tools that can be used to mitigate concerns about subjectivity, as these allow an analysis to at least approach the scientific goal of an objective analysis. The lack of a precise definition of objectivity, and a clear methodology for attaining it, is not a failure since the issue can be addressed. This is a somewhat nuanced view of the objective/subjective concern and is perhaps more in line with the views on this topic as expressed in [25,26].

3.1. Bias in Hypothesis Assessment Problems

Suppose the problem of interest is to assess whether or not there is evidence in favor of or against $H_0 : \Psi(\theta) = \psi_*$, as is determined here by $RB_\Psi(\psi_*\,|\,x)$ being greater than or less than 1. It is to be noted that no restrictions, beyond propriety, are placed on priors here so Π could very well be a mixture of a prior on $H_0 \subset \Theta$ and a prior on H_0^c with H_0 assigned some positive mass as is commonly done in Bayesian testing problems. Certainly, such a prior is necessary when $\Psi = I_{H_0}$ and $\psi_* = 1$ so the relevant relative belief ratio is $RB(H_0\,|\,x)$. While this formulation is accommodated, there is no reason to insist that every hypothesis assessment be expressed this way. When $\Psi(\theta)$ is a quantity like a mean, variance, quantile, etc., it seems natural to compare the value $RB_\Psi(\psi_*\,|\,x)$ with each of the other possible values $RB_\Psi(\psi\,|\,x)$ for $\psi \in \Psi$ to calibrate, as is done subsequently via (2), how strong the evidence is concerning ψ_*.

The following example is carried along as it illustrates a number of things.

Example 3. *Location normal.*

Suppose $x = (x_1, \ldots, x_n)$ is i.i.d. $N(\mu, \sigma_0^2)$ with π a $N(\mu_0, \tau_0^2)$ prior. Then, $\mu\,|\,x \sim N(\left(n/\sigma_0^2 + 1/\tau_0^2\right)^{-1}(n\bar{x}/\sigma_0^2 + \mu_0/\tau_0^2), \left(n/\sigma_0^2 + 1/\tau_0^2\right)^{-1})$ and so $RB(\mu\,|\,x)$ equals

$$\left(1+\frac{n\tau_0^2}{\sigma_0^2}\right)^{1/2}\exp\left\{-\frac{1}{2}\left(1+\frac{\sigma_0^2}{n\tau_0^2}\right)^{-1}\left(\frac{\sqrt{n}(\bar{x}-\mu)}{\sigma_0}+\frac{\sigma_0(\mu_0-\mu)}{\sqrt{n}\tau_0^2}\right)^2+\frac{(\mu-\mu_0)^2}{2\tau_0^2}\right\}.$$

Observe that, as $\tau_0^2 \to \infty$, then $RB(\mu \,|\, x) \to \infty$ for every μ and in particular for a hypothesized value $H_0 = \{\mu_*\}$. Thus, it would appear that overwhelming evidence is obtained for the hypothesis when the prior is very diffuse, and this holds irrespective of what the data says. In addition, when the standardized value $\sqrt{n}|\bar{x} - \mu_*|/\sigma_0$ is fixed, then $RB(\mu_* \,|\, x) \to \infty$ as $n \to \infty$. These phenomena also occur if a Bayes factor (which equals $RB(\mu_* \,|\, x)$ in this case) or a posterior probability based upon a discrete prior mass at μ_*, are used to assess H_0. Accordingly, all these measures lead to a sharp disagreement with the frequentist p-value $2(1 - \Phi(\sqrt{n}|\bar{x} - \mu_*|/\sigma_0))$ when it is small. This is the Jeffreys–Lindley paradox, and it arises quite generally.

The Jeffreys–Lindley paradox shows that the strength of evidence cannot be measured strictly by the size of the measure of evidence. A logical way to assess strength is to compare the evidence for ψ_* with the evidence for the other values for ψ. The *strength* can then be measured by

$$\Pi_\Psi(RB_\Psi(\psi \,|\, x) \le RB_\Psi(\psi_* \,|\, x) \,|\, x), \tag{2}$$

the posterior probability that the true value has evidence no greater than the evidence for ψ_*. Thus, if $RB_\Psi(\psi_* \,|\, x) < 1$ and (2) is small, then there is strong evidence against ψ_*, while, if $RB_\Psi(\psi_* \,|\, x) > 1$ and (2) is large, then there is strong evidence in favor of ψ_*. The inequalities $\Pi_\Psi(\{\psi_*\} \,|\, x) \le \Pi_\Psi(RB_\Psi(\psi \,|\, x) \le RB_\Psi(\psi_* \,|\, x) \,|\, x) \le RB_\Psi(\psi_* \,|\, x)$ hold and thus, when $RB_\Psi(\psi_* \,|\, x)$ is small, there is strong evidence against ψ_* and, when $RB_\Psi(\psi_* \,|\, x) > 1$ and $\Pi_\Psi(\{\psi_*\} \,|\, x)$ is big, then there is strong evidence in favor of ψ_*. Note, however, that $\Pi_\Psi(\{\psi_*\} \,|\, x) \approx 1$ does not guarantee $RB_\Psi(\psi_* \,|\, x) > 1$ and, if $RB_\Psi(\psi_* \,|\, x) < 1$, this means that there is weak evidence against ψ_*. In addition, there is no reason why multiple measures of the strength of the evidence can't be used (see the discussion in Section 3.2). In fact, when Ψ is binary-valued, it is better to use $\Pi_\Psi(\{\psi_*\} \,|\, x)$ to measure the strength, as we did in Examples 1 and 2, and there are also some issues with (2) in the continuous case that can require a modification. These issues are ignored here, as the strength does not play a role when considering bias, and the reader can see [2] for further discussion. The important point is that it is necessary to calibrate the measure of evidence using probability to measure how strong belief in the evidence is and (2) is a reasonable way to do this in many contexts.

1. Example 3 *Location normal (continued).*

A simple calculation shows that, with $\sqrt{n}|\bar{x} - \mu_*|$ fixed, (2) then converges to $2(1 - \Phi(\sqrt{n}|\bar{x} - \mu_*|/\sigma_0))$ as $n\tau_0^2 \to \infty$. Thus, if the p-value is small, this indicates that a large value of $RB_\Psi(\mu_* \,|\, x)$ is only weak evidence in favor of μ_*. It is to be noted that the p-value $2(1 - \Phi(\sqrt{n}|\bar{x} - \mu_*|/\sigma_0))$ is not a valid measure of evidence as described here because there is no cut-off that corresponds to evidence for and evidence against. Thus, its appearance as a measure of the strength of the evidence is not circular.

Simple algebra shows (see the Appendix A), however, that

$$\begin{aligned}&2(1 - \Phi(\sqrt{n}|\bar{x} - \mu_*|/\sigma_0)) -\\ &2(1 - \Phi([\log(1 + n\tau_0^2/\sigma_0^2) + \left(1 + \sigma_0^2/n\tau_0^2\right)^{-1}(\bar{x} - \mu_0)^2/\tau_0^2]^{1/2}),\end{aligned}$$

a difference of two p-values, is a valid measure of evidence via the cut-off 0. From this, it is seen that the values of the first p-value $2(1 - \Phi(\sqrt{n}|\bar{x} - \mu_*|/\sigma_0)$ that lead to evidence against, generally become smaller as $n\tau_0^2 \to \infty$. For example, with $n = 10, \sigma_0^2 = 1, \mu_* = 0$ and $\sqrt{n}|\bar{x} - \mu_*|/\sigma_0 = 1.96$, the standard p-value equals 0.05. Setting $\mu_0 = 0$ and $\tau_0^2 = 1$, the second p-value equals 0.097 and thus there is evidence against $\mu_* = 0$, with $\tau_0^2 = 10$ being the second term equal to 0.031 and, with $\tau_0^2 = 100$, it equals 0.009, so there is evidence in favor of $\mu_* = 0$ in both cases. When n increases, these values become smaller, as, with $n = 50$, the first p-value equal to 0.05 is always evidence in favor. Similar results are obtained with a uniform prior on $(-m, m)$, reflecting perhaps a desire to treat many values equivalently, as $m \to \infty$ or $n \to \infty$. For example, with $m = 10$ and $n = 10, \sigma_0^2 = 1$, $\mu_* = 0, \sqrt{n}|\bar{x} - \mu_*|/\sigma_0 = 1.96$, then the second p-value equals 0.002, and there is evidence in favor of $\mu_* = 0$. These findings are similar to those in [27,28].

It is very simple to elicit (μ_0, τ_0^2) based on prescribing an interval that contains the true μ with some high probability such as 99.9%, taking μ_0 to be the mid-point and so τ_0^2 is determined. There is no reason to take τ_0^2 to be arbitrarily large. However, one still wonders if the choice made is inducing some kind of bias into the problem as taking τ_0^2 too large clearly does.

Certainly, default choices of priors should be avoided when possible, but even when eliciting, how can we know if the chosen prior is inducing bias? To assess this, a numerical measure is required. The principle of evidence suggests that *bias against* H_0 is measured by

$$M(RB_\Psi(\psi_* \,|\, X) \le 1 \,|\, \psi_*) \tag{3}$$

where $M(\cdot \,|\, \psi_*)$ is the prior predictive distribution of the data given that the hypothesis is true. Thus, (3) is the prior probability that evidence in favor of ψ_* will not be obtained when ψ_* is the true value. If (3) is large, then there is an a priori bias against H_0.

For the bias in favor of H_0, it is necessary to assess if evidence against H_0 will not be obtained with high prior probability even when H_0 is false. One possibility is to measure *bias in favor* by

$$\begin{aligned}&\int_{\Psi\backslash\{\psi_*\}} M(RB_\Psi(\psi_* \,|\, X) \ge 1 \,|\, \psi)\, \Pi_\Psi(d\psi) \\ &= M(RB_\Psi(\psi_* \,|\, X) \ge 1) - M(RB_\Psi(\psi_* \,|\, X) \ge 1 \,|\, \psi_*)\Pi_\Psi(\{\psi_*\}),\end{aligned} \tag{4}$$

the prior probability of not obtaining evidence against ψ_* when it is false. When $\Pi_\Psi(\{\psi_*\}) = 0$, (4) equals $M(RB_\Psi(\psi_* \,|\, X) \ge 1)$, where M is the prior predictive for the data. For continuous parameters, it can be argued that it doesn't make sense to consider values of ψ so close to ψ_* that they are practically indistinguishable. Suppose that there is a measure of distance d_Ψ on Ψ and a value $\delta > 0$ such that, if $d_\Psi(\psi_*, \psi) < \delta$, then ψ_* and ψ are indistinguishable in the application. The *bias in favor* of H_0 is then measured by replacing $\Psi\backslash\{\psi_*\}$ in (4) by $\{\psi : d_\Psi(\psi_*, \psi) \ge \delta\}$ leading to the upper bound

$$\sup_{\{\psi: d_\Psi(\psi_*, \psi) \ge \delta\}} M(RB_\Psi(\psi_* \,|\, X) \ge 1 \,|\, \psi). \tag{5}$$

Typically, $M(RB_\Psi(\psi_* \,|\, X) \ge 1 \,|\, \psi)$ decreases as ψ moves away from ψ_* so (5) can be computed by finding the supremum over the set $\{\psi : d_\Psi(\psi_*, \psi) = \delta\}$ and, when ψ is real-valued and d_Ψ is Euclidian distance, this set equals $\{\psi_* - \delta, \psi_* + \delta\}$.

It is to be noted that the measures of bias given by (3)–(5) do not depend on using the relative belief ratio to measure evidence. Any valid measure of evidence will determine the same values when the relevant cut-off is substituted for 1. It is only (2) that depends on the specific choice of the relative belief ratio as the measure of evidence.

Under general circumstances, see [2], both biases will converge to 0 as the amount of data increases and thus they can be controlled by the amount of data collected. There is no point in reporting the results of an analysis when there is a lot of bias unless the evidence contradicts the bias.

2. Example 3 *Location normal (continued).*

Under $M(\cdot \,|\, \mu)$, then $\bar{x} \sim N(\mu, \tau_0^2 + \sigma_0^2/n)$. Thus, putting

$$a(\mu, \mu_0, \tau_0^2, \sigma_0^2, n) = \sigma_0(\mu - \mu_0)/\sqrt{n\tau_0^2},$$
$$b(\mu, \mu_0, \tau_0^2, \sigma_0^2, n) = \{(1 + \sigma_0^2/n\tau_0^2)[\log(1 + n\tau_0^2/\sigma_0^2) + (\mu - \mu_0)^2/\tau_0^2]\}^{1/2},$$

then (3) is given by

$$M(RB(\mu\,|\,X) \le 1\,|\,\mu) = 1-\Phi\Big(a(\mu,\mu_0,\tau_0^2,\sigma_0^2,n) + b(\mu,\mu_0,\tau_0^2,\sigma_0^2,n)\Big)+ \Phi\Big(a(\mu,\mu_0,\tau_0^2,\sigma_0^2,n) - b(\mu,\mu_0,\tau_0^2,\sigma_0^2,n)\Big). \quad (6)$$

This goes to 0 as $n \to \infty$ or as $\tau_0^2 \to \infty$. Thus, bias against can be controlled by sample size n or by the diffuseness of the prior although, as subsequently shown, a diffuse prior induces bias in favor. It is also the case that (6) converges to 0 when $\mu_0 \to \pm\infty$ or when $\sigma_0/\sqrt{n}\tau_0$ is fixed and $\tau_0 \to 0$. Thus, it would appear that using a prior with a location quite different than the hypothesized value or a prior that was much more concentrated than the sampling distribution can be used to lower bias against. These are situations, however, where one can expect to have prior-data conflict after observing the data.

The entries in Table 2 record the bias against for a specific case and illustrate that increasing n does indeed reduce bias. The entries also show that bias against can be greater when the prior is centered on the hypothesis. Figure 1 contains a plot of the bias against $H_0 = \{\mu\}$, as a function of μ, when using a $N(0,1)$ prior. Note that the maximum bias against occurs at the mean of the prior (and equals 0.143), and this typically occurs when $\sigma_0^2/n\tau_0^2 < 1$, namely, when the data are more concentrated than the prior. Figure 1 also contains a plot of the bias against when using a prior more concentrated than the data distribution. That the bias against is maximized, as a function of the hypothesized mean μ, when μ equals the value associated with the strongest belief under the prior, seems odd. This phenomenon arises quite often, and the mathematical explanation for this is that the greater the amount of prior probability assigned to a value, the harder it is for the posterior probability to increase and so it is quite logical when considering evidence. It will be seen that this phenomenon is very convenient for the control of bias in estimation problems and could be used as an argument for using a prior centered on the hypothesis, although this is not necessary as beliefs may be different.

Table 2. Bias against (3) the hypothesis $H_0 = \{0\}$ with a $N(\mu_0,\tau_0^2)$ prior for different sample sizes n with $\sigma_0 = 1$.

n	$\mu_0 = 1, \tau_0 = 1$	$\mu_0 = 0, \tau_0 = 1$
5	0.095	0.143
10	0.065	0.104
20	0.044	0.074
50	0.026	0.045
100	0.018	0.031

Now, consider (5), namely, bias in favor of $H_0 = \{\mu_*\}$. Putting

$$c(\mu_*,\mu,\mu_0,\tau_0^2,\sigma_0^2,n) = \sqrt{n}(\mu_* - \mu)/\sigma_0 + a(\mu_*,\mu_0,\tau_0^2,\sigma_0^2,n),$$

then (5) equals $\max M(RB(\mu_*\,|\,X) \ge 1\,|\,\mu_* \pm \delta)$ where

$$M(RB(\mu_*\,|\,X) \ge 1\,|\,\mu) = \Phi\Big(c(\mu_*,\mu,\mu_0,\tau_0^2,\sigma_0^2,n) + b(\mu_*,\mu_0,\tau_0^2,\sigma_0^2,n)\Big) - \Phi\Big(c(\mu_*,\mu,\mu_0,\tau_0^2,\sigma_0^2,n) - b(\mu_*,\mu_0,\tau_0^2,\sigma_0^2,n)\Big) \quad (7)$$

which converges to 0 as $n \to \infty$ and also as $\mu \to \pm\infty$. However, (7) converges to 1 as $\tau_0^2 \to \infty$, so, if the prior is too diffuse, there will be bias in favor of μ_*. Thus, resolving the Jeffreys–Lindley paradox requires choosing the sample size n, after choosing the prior, so that (7) is suitably small. Note that choosing τ_0^2 to be larger reduces bias against but increases bias in favor and so generally bias cannot be avoided by choice of prior. Figure 2 is a plot of $M(RB(\mu_* \mid X) \geq 1 \mid \mu)$ for a particular case and this strictly decreases as μ moves away from μ_*.

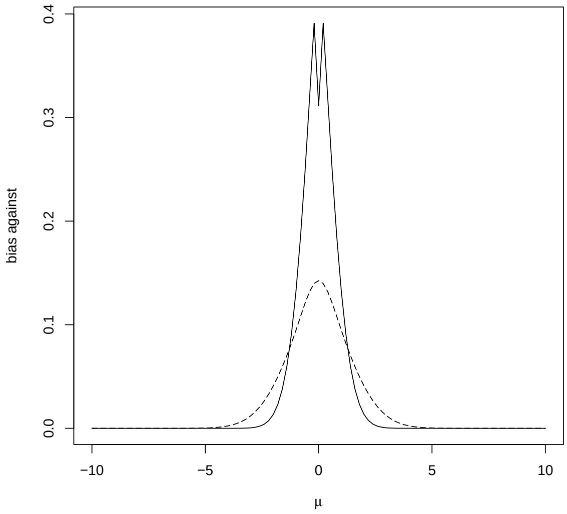


Figure 1. Plot of bias against $H_0 = \{\mu\}$ with a $N(0,1)$ prior (- - -) and a $N(0,0.01)$ prior (—) with $n = 5, \sigma_0 = 1$.

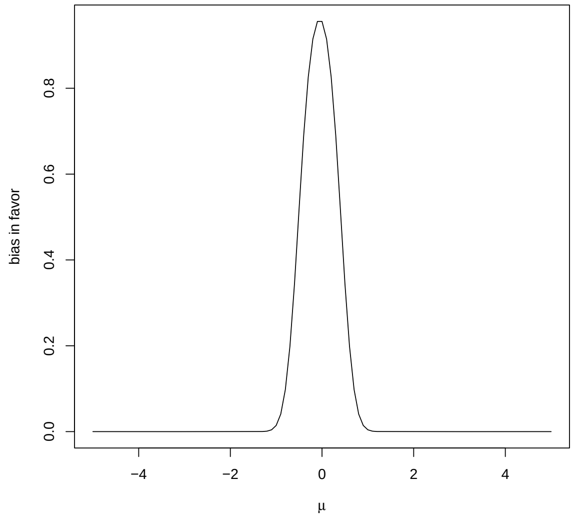


Figure 2. Plot of $M(RB(0 \mid X) \geq 1 \mid \mu)$ when $n = 20, \mu_0 = 1, \tau_0 = 1, \sigma_0 = 1$.

In Table 3, we have recorded some specific values of the bias in favor using (4) and using (5) where d_Ψ is Euclidean distance. It is seen that bias in favor can be quite serious for small samples. When using (5), this can be mitigated by making δ larger. For example, with $(\mu_0, \tau_0) = (0, 1), \delta = 1.0, n = 20$, the bias in favor equals 0.004. Note, however, that δ is not chosen to make the bias in favor small; rather, it is determined in an application as the difference from the null that is just practically important. The virtues of a suitable value of δ are readily apparent as (5) is much smaller than (4) for larger n.

A comparison of Tables 2 and 3 shows that a study whose purpose is to demonstrate evidence in favor of H_0 is much more demanding than one whose purpose is to determine whether or not there is evidence against H_0. As a cautionary note too, it is worth reminding the reader that bias is not to be used in the selection of a prior. The prior is to be selected by elicitation and the biases measured for that prior. If one or both biases are too large, then that is telling us that more data are needed to ensure that the conclusions drawn are primarily driven by the data and not the prior. It is tempting to look at Tables 2 and 3 and compare the priors, but this is not the way to proceed and it can be seen that choosing a prior to minimize one bias simply increases the other. It is also the case that bias can be measured when a default proper prior is chosen, see Example 3, as is often done when considering sparsity inducing priors, but the discussion here will focus on the ideal where elicitation can be carried out. One can argue that bias is also model dependent and that is certainly true so, while our focus is on the prior, in reality, the biases are a measure of the model-prior combination. The same comment applies to the model, however, that bias measurements are not to be used to select a model.

Table 3. Bias in favor of the hypothesis $H_0 = \{0\}$ with a $N(\mu_0, \tau_0^2)$ prior for different sample sizes n with $\sigma_0 = 1$ using (4) (and using (5) with $\delta = 0.5$).

n	$(\mu_0, \tau_0) = (1, 1)$	$(\mu_0, \tau_0) = (0, 1)$
5	0.323 (0.871)	0.451 (0.631)
10	0.259 (0.747)	0.371 (0.516)
20	0.215 (0.519)	0.299 (0.327)
50	0.153 (0.125)	0.219 (0.062)
100	0.116 (0.006)	0.168 (0.002)

3.2. *The Role of the Difference that Matters δ*

The role and value of δ require some further discussion as some may find the need to specify this quantity controversial. The value of δ depends on the application as well as the characteristic of interest $\psi = \Psi(\theta)$. For the developments here, specifying δ is a necessary part of the investigation. There may well be contexts where the precise value of δ is unclear. That seems to suggest, however, that the investigator does not fully understand what ψ is as a real-world object and formal inference in such a context seems questionable, although perhaps some kind of exploratory analysis is reasonable. In a well-designed study, a measurement process is selected which, together with sampling from the population, determines the data. In deciding on the measurement process, and sample size, an investigator has to decide on the accuracy required and that is where δ enters the picture.

Consider a problem where an investigator is measuring the length of some quantity associated with each member of a population and wants to make inferences about the mean length ψ. If the investigator chooses to measure each length to the nearest cm, then there is no way that the true value of the mean can be known to an accuracy beyond ± 0.5 cm, even if the entire population is measured. As another example, suppose that ψ represents the proportion of individuals in a population infected with a virus. Surely, it is imperative to settle on how accurately we wish to know ψ and that will play a key role in a number of statistical activities like determining sample size for the consideration of a hypothesis concerning the true value of ψ. For example, does the application require that ψ be known within an absolute error of δ or within a relative error of δ? See [29] for discussion on this point in the context of logistic regression. To simply proceed to collect data and do a statistical analysis without taking such considerations into account does not seem like good practice.

While discussion of δ may be limited, it has certainly not disappeared from the statistical literature. For example, consider power studies where a δ is required. In addition, one of the many criticisms of the p-value arises because, for a large enough sample size, a difference may be detected that is of no importance. The general recommendation is to then quote a confidence interval to see if that is the case, but it is difficult to see how that is helpful unless one knows what difference δ matters. This has long been an issue when discussing testing problems, see [30], and yet it still seems unresolved as it is not always clear how to obtain an appropriate p-value that incorporates δ. One of the benefits of the approach here is that it is straightforward to incorporate δ into the analysis and, in fact, it often makes an analysis easier. Thus, specifying δ is a part of every well-designed statistical investigation.

3.3. Bias in Estimation Problems

The relative belief estimate of $\psi = \Psi(\theta)$ is the value that maximizes the measure of evidence, namely, $\psi(x) = \arg\sup RB_{\Psi}(\psi \,|\, x)$. It is easy to show that $RB_{\Psi}(\psi(x) \,|\, x) \geq 1$ with the inequality strict except in trivial contexts. The accuracy of this estimate can be measured by the "size" of the *plausible region* $Pl_{\Psi}(x) = \{\psi : RB_{\Psi}(\psi \,|\, x) > 1\}$, the set of values of ψ that have evidence in their favor and note $\psi(x) \in Pl_{\Psi}(x)$. To say that $\psi(x)$ is an accurate estimate requires that $Pl_{\Psi}(x)$ be "small", perhaps as measured by $Vol(Pl_{\Psi}(x))$, where Vol is some measure of volume, and also has high posterior content $\Pi_{\Psi}(Pl_{\Psi}(x) \,|\, x)$, which measures the belief that the true value is in $Pl_{\Psi}(x)$. Note that $Pl_{\Psi}(x)$ does not depend on the specific measure of evidence chosen, in this case the relative belief ratio. Any valid estimator must satisfy the principle of evidence and thus be in $Pl_{\Psi}(x)$. It is now argued that, in an estimation problem, bias is measured by various coverage probabilities for the plausible region.

Note too that, if there is evidence in favor of $H_0 : \Psi(\theta) = \psi_*$, then $\psi_* \in Pl_{\Psi}(x)$ and so represents the natural estimate of ψ provided there was a clear reason, like the assessment of a scientific theory, for assessing the evidence for this value. This assumes too that there isn't substantial bias in favor of ψ_*. The strength of the evidence in favor of ψ_* could then also be measured by the size of $Pl_{\Psi}(x)$. Similarly, if evidence against H_0 is obtained, then $\psi_* \in Im_{\Psi}(x) = \{\psi : RB_{\Psi}(\psi \,|\, x) < 1\}$ the *implausible region*, and there is strong evidence against H_0 provided $Im_{\Psi}(x)$ has small volume and large posterior probability. A virtue of this approach to measuring the strength of the evidence is that it does not depend upon using the relative belief ratio in hypothesis assessment problems.

The prior probability that the plausible region does not cover the true value measures bias against when estimating ψ. If this probability is large, then the estimate and the plausible region are a priori likely to be misleading as to the true value. The prior probability that $Pl_{\Psi}(x)$ doesn't contain $\psi = \Psi(\theta)$ when $\theta \sim \Pi, X \sim P_{\theta}$ is

$$E_{\Pi_{\Psi}}(M(\psi \notin Pl_{\Psi}(X) \,|\, \psi)) = E_{\Pi_{\Psi}}(M(RB_{\Psi}(\psi \,|\, X) \leq 1 \,|\, \psi)) \tag{8}$$

which is also the average bias against over all hypothesis testing problems $H_0 : \Psi(\theta) = \psi$. Note $1 - E_{\Pi_{\Psi}}(M(\psi \notin Pl_{\Psi}(X) \,|\, \psi)) = E_{\Pi_{\Psi}}(M(\psi \in Pl_{\Psi}(X) \,|\, \psi)) = E_M(\Pi_{\Psi}(Pl_{\Psi}(X) \,|\, X))$ which is the prior coverage probability of Pl_{Ψ}. In addition,

$$\sup_{\psi} M(\psi \notin Pl_{\Psi}(X) \,|\, \psi) = \sup_{\psi} M(RB_{\Psi}(\psi \,|\, X) \leq 1 \,|\, \psi), \tag{9}$$

is an upper bound on (8). Therefore, controlling (9) controls the bias against in estimation and all hypothesis assessment problems involving ψ. In addition,

$$1 - \sup_{\psi} M(\psi \notin Pl_{\Psi}(X) \,|\, \psi) = \inf_{\psi} M(\psi \in Pl_{\Psi}(X) \,|\, \psi) \leq E_M(\Pi_{\Psi}(Pl_{\Psi}(X) \,|\, X)).$$

Thus, using (9) implies lower bounds for the coverage probability and for the expected posterior content of the plausible region. In general, both (8) and (9) converge to 0 with increasing amounts of data. Thus, it is possible to control for bias against in estimation problems by the amount of data collected.

3. Example 3 *Location normal (continued).*

The value of $M(RB(\mu \,|\, X) \leq 1 \,|\, \mu)$ is given in (6) and examples are plotted in Figure 1. When $\mu \sim N(\mu_0, \tau_0^2)$, then $z = (\mu - \mu_0)/\tau_0 \sim N(0,1)$, so

$$E_{\Pi}(M(RB(\mu \,|\, X) \leq 1 \,|\, \mu)) =$$

$$1 - E\left[\begin{array}{c} \Phi\left(\frac{\sigma_0}{\sqrt{n}\tau_0}Z + \left\{\left(1 + \frac{\sigma_0^2}{n\tau_0^2}\right)\left[\log\left(1 + \frac{n\tau_0^2}{\sigma_0^2}\right) + Z^2\right]\right\}^{1/2}\right) + \\ \Phi\left(\frac{\sigma_0}{\sqrt{n}\tau_0}Z - \left\{\left(1 + \frac{\sigma_0^2}{n\tau_0^2}\right)\left[\log\left(1 + \frac{n\tau_0^2}{\sigma_0^2}\right) + Z^2\right]\right\}^{1/2}\right) \end{array}\right]$$

which is notably independent of the prior mean μ_0. The dominated convergence theorem implies $E_{\Pi}(M(RB(\mu \,|\, X) \le 1 \,|\, \mu)) \to 0$ as $n \to \infty$ or as $\tau_0^2 \to \infty$. Thus, provided $n\tau_0^2/\sigma_0^2$ is large enough, there is hardly any estimation bias against. Table 4 illustrates some values of this bias measure. Subtracting the probabilities in Table 4 from 1 gives the prior probability that the plausible region covers the true value and the expected posterior content of the plausible region. Thus, when $n = 20, \tau_0 = 1$, the prior probability of $Pl(x)$ containing the true value is $1 - 0.051 = 0.949$ so $Pl(x)$ is a 0.949 Bayesian confidence interval for μ.

Table 4. Average bias against $H_0 = 0$ when using a $N(0, \tau_0^2)$ prior for different sample sizes n.

n	$\tau_0 = 1$	$\tau_0 = 0.5$
5	0.107	0.193
10	0.075	0.146
20	0.051	0.107
50	0.031	0.067
100	0.021	0.046

To use (9), it is necessary to maximize $M(RB(\mu \,|\, X) \le 1 \,|\, \mu)$ as a function of μ and it is seen that, at least when the prior is not overly concentrated, this maximum occurs at μ_0. Figure 1 shows that, when using the $N(0, 1)$ prior, the maximum occurs at $\mu = 0$ when $n = 5$ and, from the second column of Table 2, the maximum equals 0.143. The average bias against is given by 0.107, as recorded in Table 4. Note that the maximum also occurs at $\mu = 0$ for the other values of n recorded in Table 2.

Bias in favor when estimating ψ occurs when the prior probability that Im_Ψ does not cover a false value is large, namely, when

$$\begin{aligned}
&\int_\Psi \int_{\Psi \setminus \{\psi_*\}} M(\psi_* \notin Im_\Psi(X) \,|\, \psi)\, \Pi_\Psi(d\psi)\, \Pi_\Psi(d\psi_*) \\
&= \int_\Psi \int_{\Psi \setminus \{\psi_*\}} M(RB_\Psi(\psi_* \,|\, X) \ge 1 \,|\, \psi)\, \Pi_\Psi(d\psi)\, \Pi_\Psi(d\psi_*)
\end{aligned} \tag{10}$$

is large as this would seem to imply that the plausible region will cover a randomly selected false value from the prior with high prior probability. Note that (10) is the prior mean of (4) and, in the continuous case, equals $\int_\Psi M(\psi_* \notin Im_\Psi(X))\, \Pi_\Psi(d\psi_*)$. As previously discussed, however, it often doesn't make sense to distinguish values of ψ that are close to ψ_*. The bias in favor for estimation can then be measured by

$$\begin{aligned}
&E_{\Pi_\Psi}\left(\sup_{\{\psi : d_\Psi(\psi, \psi_*) \ge \delta\}} M(\psi_* \notin Im_\Psi(X) \,|\, \psi) \right) \\
=&E_{\Pi_\Psi}\left(\sup_{\{\psi : d_\Psi(\psi, \psi_*) \ge \delta\}} M(RB_\Psi(\psi_* \,|\, X) \ge 1 \,|\, \psi) \right).
\end{aligned} \tag{11}$$

An upper bound on (11) is commonly equal to 1, as illustrated in Figure 3, and thus is not useful.

It is the size and posterior content of $Pl_\Psi(x)$ that provides a measure of the accuracy of the estimate $\psi(x)$. As previously discussed, the a priori expected posterior content of $Pl_\Psi(x)$ can be controlled by bias against. The a priori expected volume of $Pl_\Psi(x)$ satisfies

$$E_M(Vol(Pl_\Psi(X))) = \int_\Psi \int_\Psi M(\psi_* \in Pl_\Psi(X) \,|\, \psi)\, \Pi_\Psi(d\psi)\, Vol(d\psi_*). \tag{12}$$

Notice that, when $\Pi_\Psi(\{\psi\}) = 0$ for every ψ, this can be interpreted as a kind of average of the prior probabilities of the plausible region covering a false value.

4. Example 3 *Location normal (continued).*

It follows from (7) that

$$\sup M(RB(\mu_* \,|\, X) \geq 1 \,|\, \mu_* \pm \delta) \;=\; \sup\left\{ \begin{array}{c} \Phi(c(\mu_*, \mu_* \pm \delta, \mu_0, \tau_0^2, \sigma_0^2, n) + b(\mu_*, \mu_0, \tau_0^2, \sigma_0^2, n)) - \\ \Phi(c(\mu_*, \mu_* \pm \delta, \mu_0, \tau_0^2, \sigma_0^2, n) - b(\mu_*, \mu_0, \tau_0^2, \sigma_0^2, n)) \end{array} \right\}$$

Note that, as $\mu_* \to \pm\infty$, then $M(RB(\mu_* \,|\, X) \geq 1 \,|\, \mu_* \pm \delta) \to 1$ when $n\tau_0^2/\sigma_0^2 > 1$, see Figure 3, and converges to 0 if $n\tau_0^2/\sigma_0^2 < 1$, so it would appear that the better circumstance for guarding against bias in favor is when the prior is putting in more information than the data. As previously noted, however, this is a situation where we might expect prior data-conflict to arise and, except in exceptional circumstances, should be avoided. Table 5 contains values of (11) for this situation with different values of δ. Again, these values are just for illustrative purposes and are not to be used to compare or choose priors.

Table 5. Average bias in favor for estimation based on (11) when using a $N(0, \tau_0^2)$ prior for different sample sizes n and difference δ.

n	$(\mu_0, \tau_0) = (0, 1), \delta = 1.0$	$(\mu_0, \tau_0) = (0, 1), \delta = 0.5$
5	0.451	0.798
10	0.185	0.690
20	0.025	0.486
50	0.000	0.131
100	0.000	0.009

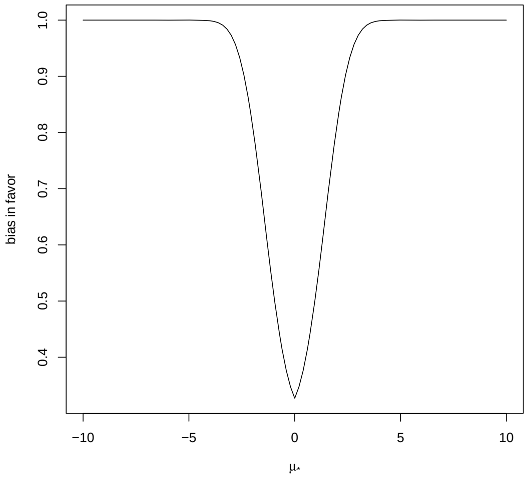


Figure 3. Bias in favor of μ maximized over $\mu \pm \delta$ based on a $N(0, 1)$ prior and $\sigma_0 = 1, n = 20, \delta = 0.5$.

Some elementary calculations give $Pl(x) = \bar{x} \pm w(\bar{x}, n, \sigma_0^2, \mu_0, \tau_0^2)$ with

$$w(\bar{x}, n, \sigma_0^2, \mu_0, \tau_0^2) = \frac{\sigma_0}{\sqrt{n}}\left(1 + \frac{n\tau_0^2}{\sigma_0^2}\right)^{-\frac{1}{2}} \left\{ \left(1 + \frac{n\tau_0^2}{\sigma_0^2}\right) \log\left(1 + \frac{n\tau_0^2}{\sigma_0^2}\right) + \left(\frac{\bar{x} - \mu_0}{\sigma_0/\sqrt{n}}\right)^2 \right\}^{\frac{1}{2}}$$

where $z = \sqrt{n}(\bar{x} - \mu_0)/\sigma_0 \sim N(0, 1)$ under M. It is notable that the prior distribution of the width is independent of the prior mean. Table 6 contains some expected half-widths together with the coverage probabilities of $Pl(x)$.

Table 6. Expected half-widths (coverages) of the plausible interval when using a $N(\mu_0, \tau_0^2)$ prior for different sample sizes n.

n	$\tau_0 = 1$	$\tau_0 = 0.5$
5	0.625 (0.893)	0.491 (0.807)
10	0.499 (0.925)	0.389 (0.854)
20	0.393 (0.949)	0.312 (0.893)
50	0.281 (0.969)	0.231 (0.933)
100	0.215 (0.979)	0.181 (0.954)

While the plausible region $Pl_\Psi(x)$ is advocated for assessing the accuracy of estimates, it is also possible to use a $\gamma-$relative belief credible region $C_\gamma(x) = \{\psi : RB_\Psi(\psi \,|\, x) \geq c_\gamma(x)\}$ where $c_\gamma(x) = \inf\{c : \Pi_\Psi(RB_\Psi(\psi \,|\, x) \geq c \,|\, x) \leq \gamma\}$. There is one proviso with this, however, as the principle of evidence requires that $\gamma \leq \Pi_\Psi(Pl_\Psi(x) \,|\, x)$; otherwise, $C_\gamma(x)$ will contain values of ψ for which there is evidence against. Notice that, while controlling the bias against allows control of the coverage probability of $Pl_\Psi(x)$, this does not control the coverage probability of a credible region since $\Pi_\Psi(Pl_\Psi(x) \,|\, x)$ is not known until the data are observed. For this reason, reporting the plausible region always seems necessary. All these regions are invariant under smooth reparameterizations and in [31] various optimality results are established for these credible regions.

4. Frequentist and Optimal Properties

Consider now the bias against $H_0 = \{\psi_*\}$, namely, $M(RB_\Psi(\psi_* \,|\, X) \leq 1 \,|\, \psi_*)$. If we repeatedly generate $\theta \sim \pi(\cdot \,|\, \psi_*), X \sim f_\theta$, then this probability is the long-run proportion of times that $RB_\Psi(\psi_* \,|\, X) \leq 1$. This frequentist interpretation depends on the conditional prior $\pi(\cdot \,|\, \psi_*)$ and, when $\Psi(\theta) = \theta$, there are no nuisance parameters, this is a "pure" frequentist probability. Even in the latter case, there is some dependence on the prior, however, as $RB(\theta_* \,|\, x) = f_{\theta_*}(x)/m(x)$ so x satisfies $RB_\Psi(\theta_* \,|\, x) \leq 1$ iff $f_{\theta_*}(x) \leq m(x)$, where $m(x) = \int_\Theta f_\theta(x)\, \Pi(d\theta)$. Thus, in general, the region $\{x : RB_\Psi(\psi_* \,|\, x) \leq 1\}$ depends on π, but the probability $M(RB_\Psi(\psi_* \,|\, X) \leq 1 \,|\, \psi_*)$ depends only on the conditional prior predictive given $\Psi(\theta) = \psi_*$, namely, $m(x \,|\, \psi_*) = \int_\Theta f_\theta(x)\, \Pi(d\theta \,|\, \psi_*)$, and not on the marginal prior π_Ψ on ψ. We refer to probabilities that depend only on $M(\cdot \mid \psi_*)$ as frequentist, for example, coverage probabilities are called confidences, and those that depend on the full prior π as Bayesian confidences. The frequentist label is similar to use of the confidence terminology when dealing with random effects' models as nuisance parameters have been integrated out.

Suppose now that some other general rule, not necessarily the principle of evidence, is used to determine whether there is evidence in favor of or against ψ_* and this leads to the set $D(\psi_*) \subset \mathcal{X}$ as those data sets that do not give evidence in favor of $H_0 = \{\psi_*\}$. The rules of potential interest will satisfy $M(D(\psi_*) \,|\, \psi_*) \leq M(RB_\Psi(\psi_* \,|\, X) \leq 1 \,|\, \psi_*)$ since this implies better performance a priori in terms of identifying when data has evidence in favor of H_0 via the set $D^c(\psi_*)$ than the principle of evidence. For example, $D(\psi_*) = \{x : RB_\Psi(\psi_* \,|\, x) \leq q\}$ for some $q < 1$ satisfies this, but note that a value satisfying $q < RB_\Psi(\psi_* \,|\, x) \leq 1$ violates the principle of evidence if it is claimed that there is evidence in favor of ψ_*. Putting $R(\psi_*) = \{x : RB_\Psi(\psi_* \,|\, x) \leq 1\}$ leads to the following result.

Theorem 1. *Consider $D(\psi_*) \subset \mathcal{X}$ satisfying $M(D(\psi_*) \,|\, \psi_*) \leq M(R(\psi_*) \,|\, \psi_*)$. (i) The prior probability $M(D(\psi_*))$ is maximized among such rules by $D(\psi_*) = R(\psi_*)$. (ii) If $\Pi_\Psi(\{\psi_*\}) = 0$, then $R(\psi_*)$ maximizes the prior probability of not obtaining evidence in favor of ψ_* when it is false and otherwise maximizes this probability among all rules satisfying $M(D(\psi_*) \,|\, \psi_*) = M(R(\psi_*) \,|\, \psi_*)$.*

When $\Pi_\Psi(\{\psi_*\}) \neq 0$, rules may exist having greater prior probability of not getting evidence in favor of ψ_* when it is false, but the price paid for this is the violation of the principle of evidence. In addition, when comparing rules based on their ability to distinguish falsity, it only seems fair that the rules perform the same under the truth. Thus, Theorem 1 is a general optimality result for the principle of evidence applied to hypothesis assessment when considering bias against.

Now, consider $C(x) = \{\psi : x \notin D(\psi)\}$, the set of ψ values for which there is evidence in their favor after observing x according to some alternative evidence rule. Since $M(\psi_* \notin C(X) \,|\, \psi) = M(D(\psi_*)) \,|\, \psi)$, then

$$\begin{aligned} E_{\Pi_\Psi}(M(\psi \in C(X)\,)\,|\,\psi)) &= 1 - E_{\Pi_\Psi}(M(\psi \notin C(X)\,|\,\psi)) = 1 - E_{\Pi_\Psi}(M(D(\psi)\,|\,\psi)) \\ &\geq 1 - E_{\Pi_\Psi}(M(R(\psi)\,|\,\psi)) = E_{\Pi_\Psi}(M(\psi \in Pl_\Psi(X)\,)\,|\,\psi)) \end{aligned}$$

and so the Bayesian coverage of C is at least as large as that of Pl_Ψ and thus represents a viable alternative to using Pl_Ψ. The following establishes an optimality result for Pl_Ψ.

Theorem 2. *(i) The prior probability that the region C doesn't cover a value ψ_* generated from the prior, namely, $E_{\Pi_\Psi}(M(\psi_* \notin C(X)))$, is maximized among all regions satisfying*

$$M(\psi_* \notin C(X) \,|\, \psi_*) \leq M(\psi_* \notin Pl_\Psi(X) \,|\, \psi_*)$$

for every ψ_*, by $C = Pl_\Psi$. (ii) If $\Pi_\Psi(\{\psi_*\}) = 0$ for all ψ_*, then Pl_Ψ maximizes the prior probability of not covering a false value and otherwise maximizes this probability among all C satisfying $M(\psi_* \notin C(X) \,|\, \psi_*) = M(\psi_* \notin Pl_\Psi(X) \,|\, \psi_*)$ for all ψ_*.

Again, when $\Pi_\Psi(\{\psi_*\}) \neq 0$, the existence of a region with better properties with respect to not covering false values than Pl_Ψ can't be ruled out, but, when considering such a property, it seems only fair to compare regions with the same coverage probability, and, in that case, Pl_Ψ is optimal. Thus, Theorem 2 is also a general optimality result for the principle of evidence applied to estimation when considering bias against. In addition, if there is a value $\psi_0 = \arg\inf_\psi M(\psi \in Pl_\Psi(X)\,)\,|\,\psi)$, then $\gamma_0 = M(\psi_0 \in Pl_\Psi(X)\,)\,|\,\psi_0)$ serves as a lower bound on the coverage probabilities, and thus Pl_Ψ is a γ_0-confidence region for ψ and this is a pure frequentist γ_0-confidence region when $\Psi(\theta) = \theta$. Since $M(\psi \in Pl_\Psi(X)\,)\,|\,\psi) = 1 - M(\psi \notin Pl_\Psi(X)\,)\,|\,\psi) = 1 - M(R(\psi)\,|\,\psi)$, then Example 3 shows that it is reasonable to expect that such a ψ_0 exists.

The principle of evidence leads to the following satisfying properties which connect the concept of bias as discussed here with the frequentist concept.

Theorem 3. *(i) Using the principle of evidence, the prior probability of getting evidence in favor of ψ_* when it is true is greater than or equal to the prior probability of getting evidence in favor of ψ_* given that ψ_* is false. (ii) The prior probability of Pl_Ψ covering the true value is always greater than or equal to the prior probability of Pl_Ψ covering a false value.*

The properties stated in Theorem 3 are similar to a property called unbiasedness for frequentist procedures. For example, a test is unbiased if the probability of rejecting a null is always larger when it is false than when it is true and a confidence region is unbiased if the probability of covering the true value is always greater than the probability of covering a false value. While the inferences discussed here are "unbiased" in this generalized sense, they could still be biased against or in favor in the sense of this paper, as it is the amount of data that controls this.

Now, consider bias in favor and suppose there is an alternative characterization of evidence that leads to the region $E(\psi_*)$ consisting of all data sets that do not lead to evidence against ψ_*. Putting $A(\psi_*) = \{x : RB_\Psi(\psi_* \,|\, x) \geq 1\}$, we restrict attention to regions satisfying $M(E(\psi_*) \,|\, \psi_*) \geq M(A(\psi_*) \,|\, \psi_*)$. Using (4) to measure bias in favor leads to the following results.

Theorem 4. *(i) The prior probability $M(E(\psi_*))$ is minimized among all $E(\psi_*) \subset \mathcal{X}$ satisfying $M(E(\psi_*) \,|\, \psi_*) \geq M(A(\psi_*) \,|\, \psi_*)$ by $E(\psi_*) = A(\psi_*)$. (ii) If $\Pi_\Psi(\{\psi_*\}) = 0$, then the set $A(\psi_*)$*

minimizes the prior probability of not obtaining evidence against ψ_* when it is false and otherwise minimizes this probability among all rules satisfying $M(E(\psi_*)\,|\,\psi_*) = M(A(\psi_*)\,|\,\psi_*)$.

Theorem 5. *(i) The prior probability region C covers a value ψ_* generated from the prior, namely, $E_{\Pi_\Psi}(M(\psi_* \in C(X)))$, is minimized among all regions satisfying $M(\psi_* \in C(X)\,|\,\psi_*) \geq M(\psi_* \in Pl_\Psi(X)\,|\,\psi_*)$ for every ψ_*, by $C = Pl_\Psi$. (ii) If $\Pi_\Psi(\{\psi_*\}) = 0$ for all ψ_*, then Pl_Ψ minimizes the prior probability of covering a false value and otherwise minimizes this probability among all rules satisfying $M(\psi_* \in C(X)\,|\,\psi_*) = M(\psi_* \in Pl_\Psi(X)\,|\,\psi_*)$ for all ψ_*.*

Thus, Theorems 4 and 5 are optimality results for the principle of evidence when considering bias in favor.

Clearly, the bias against H_0 is playing a role similar to size in frequentist statistics and the bias in favor is playing a role similar to power. A study that found evidence against H_0, but had a high bias against, or a study that found evidence in favor of H_0 but had a high bias in favor, could not be considered to be of high quality. Similarly, a study concerned with estimating a quantity of interest could not be considered of high quality if there is high bias against or in favor. There are some circumstances, however, where some bias is perhaps not an issue. For example, in a situation where sparsity is to be expected, then, allowing for high bias in favor of certain hypotheses accompanied by low bias against, may be tolerable, although this does reduce the reliability of any hypotheses where evidence is found in favor.

The concept of a *severe test* is introduced in [32], and this has a similar motivation to measuring bias. This is described now with some small modifications that allow for a more general discussion than the special situations used in the reference. Suppose $d(x)$ is the test statistic for an test of size α so that $H_0 : \Psi(\theta) = \psi_0$ is rejected when $d(x) > c_\alpha$ and accepted otherwise. A deviation γ^* that is *substantively important* is specified. When the test leads to the acceptance of H_0, the severity of the test is assessed by the *attained power* $P_\theta(d(X) > d(x)\,|\,x)$ for θ values satisfying $d_\Psi(\psi_0, \Psi(\theta)) \geq \gamma^*$, where d_Ψ is a distance measure on Ψ. To get a single number for the severity measure, it makes sense to use $\inf_{\{\theta:d_\Psi(\psi_0,\Psi(\theta))=\gamma^*\}} P_\theta(d(X) > d(x)\,|\,x)$ as generally $P_\theta(d(X) > d(x)\,|\,x)$ will increase as $d_\Psi(\psi_0, \Psi(\theta))$ increases. The hypothesis H_0 is accepted with *high severity* when the attained power is high. The motivation for adding this measure of the test is that it claimed that it is incorrect to simply accept H_0 when $d(x) \leq c_\alpha$ unless the probability of obtaining a value of the test statistic as least as large as that observed is high when the hypothesis is meaningfully false. When H_0 is rejected, then the severity of the test is measured by $P_\theta(d(X) \leq d(x)\,|\,x)$ for θ values satisfying $d_\Psi(\psi_0, \Psi(\theta)) < \gamma^*$ and, to obtain a single number one could use $\sup_{\{\theta:d_\Psi(\psi_0,\Psi(\theta))\leq\gamma^*\}} P_\theta(d(X) \leq d(x)\,|\,x)$. It is then required that this probability be small to claim a rejection with high severity.

The use of the γ^* quantity seems identical to the difference that matters δ and we agree that this is an essential aspect of a statistical analysis. In hypothesis assessment, this guards against "the large n problem" where large sample sizes will detect deviations from H_0 that are not practically meaningful. There are, however, numerous differences with the discussion of bias here. The severity approach is expressed within the context where either H_0 or H_0^c is accepted and the relative belief approach is more general than this binary classification. The testing approach suffers from the lack of a clear choice of α to determine the cut-off, and this is not the case for the principle of evidence. The bias measures are frequentist performance characteristics, albeit somewhat dependent on the prior, but the measures of severity are conditional on the observed x leaving one wondering about their frequentist performance characteristics, see [33] for more discussion on this point. The assessment of H_0 via relative belief is based on the observed data and datasets not observed are irrelevant, at least for the expression of the evidence. The relevance of unobserved data are for us better addressed a priori where such considerations lead to an assessment of the merits of the study, but these play no role in the actual inferences. The major difference is that a proper prior is required here as this leads to a characterization of evidence via the principle of evidence.

5. Examples

A number of examples are now considered.

Example 4. *Binomial proportion.*

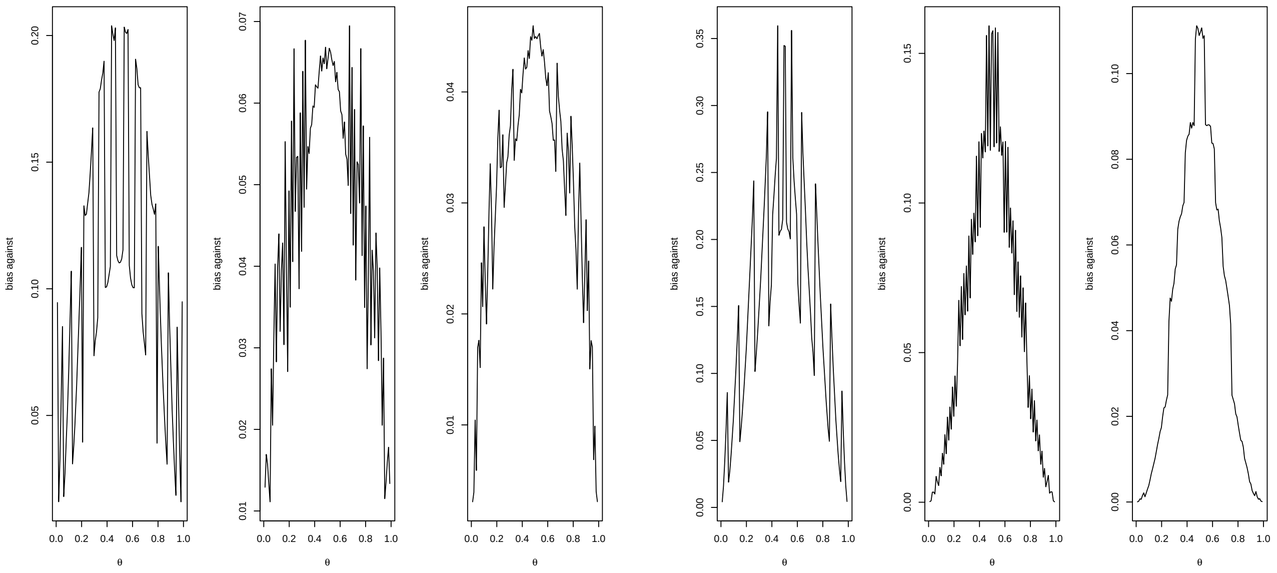
Suppose $x = (x_1, \dots, x_n)$ is a sample from the Bernoulli(θ) with $\theta \in [0,1]$ unknown so $n\bar{x} \sim$ binomial(n, θ) and interest is in θ. For the prior, let $\theta \sim$ beta(α_0, β_0) where the hyperparameters are elicited as in, for example [34], so $\theta \,|\, n\bar{x} \sim$ beta$(\alpha_0 + n\bar{x}, \beta_0 + n(1-\bar{x}))$. Then,

$$RB(\theta \,|\, n\bar{x}) = \frac{\Gamma(\alpha_0 + \beta_0 + n)}{\Gamma(\alpha_0 + n\bar{x})\Gamma(\beta_0 + n(1-\bar{x}))} \frac{\Gamma(\alpha_0)\Gamma(\beta_0)}{\Gamma(\alpha_0 + \beta_0)} \theta^{n\bar{x}}(1-\theta)^{n(1-\bar{x})}$$

is unimodal with mode at $\bar{x}$, so $Pl(x)$ is an interval containing $\bar{x}$. Note that $M(\cdot \,|\, \theta)$ is the binomial(n, θ) probability measure and the bias against θ is given by $M(RB(\theta \,|\, n\bar{x}) \leq 1 \,|\, \theta)$ while the bias in favor of θ, using (5), is given by $\max M(RB(\theta \,|\, n\bar{x}) \geq 1 \,|\, \theta \pm \delta)$for $\theta \in [\delta, 1-\delta]$.

Consider first the prior given by $(\alpha_0, \beta_0) = (1,1)$. Figure 4a gives the plots of the bias against for $n = 10$ (max. = 0.21, average = 0.11), $n = 50$ (max.= 0.07, average = 0.05) and $n = 100$ (max. = 0.05, average = 0.03). Therefore, when $n = 10$, then $Pl(x)$ is a 0.79-confidence interval for θ; when $n = 50$, it is a 0.93-confidence interval for θ and, when $n = 100$, it is a 0.95-confidence interval for θ. For the informative prior given by $(\alpha_0, \beta_0) = (5,5)$, Figure 4b gives the plots of the bias against for $n = 10$ (max. = 0.36, average = 0.21), $n = 50$ (max. = 0.16, average = 0.10) and $n = 100$ (max. = 0.11, average = 0.07). Thus, when $n = 10$, then $Pl(x)$ is a 0.64-confidence interval for θ, when $n = 50$, it is a 0.84-confidence interval for θ and, when $n = 100$, it is a 0.93-confidence interval for θ. One feature immediately stands out, namely, when using a more informative prior the bias against increases. As previously explained, this phenomenon occurs because when the prior probability of θ is small, it is much easier to obtain evidence in favor than when the prior probability of θ is large.

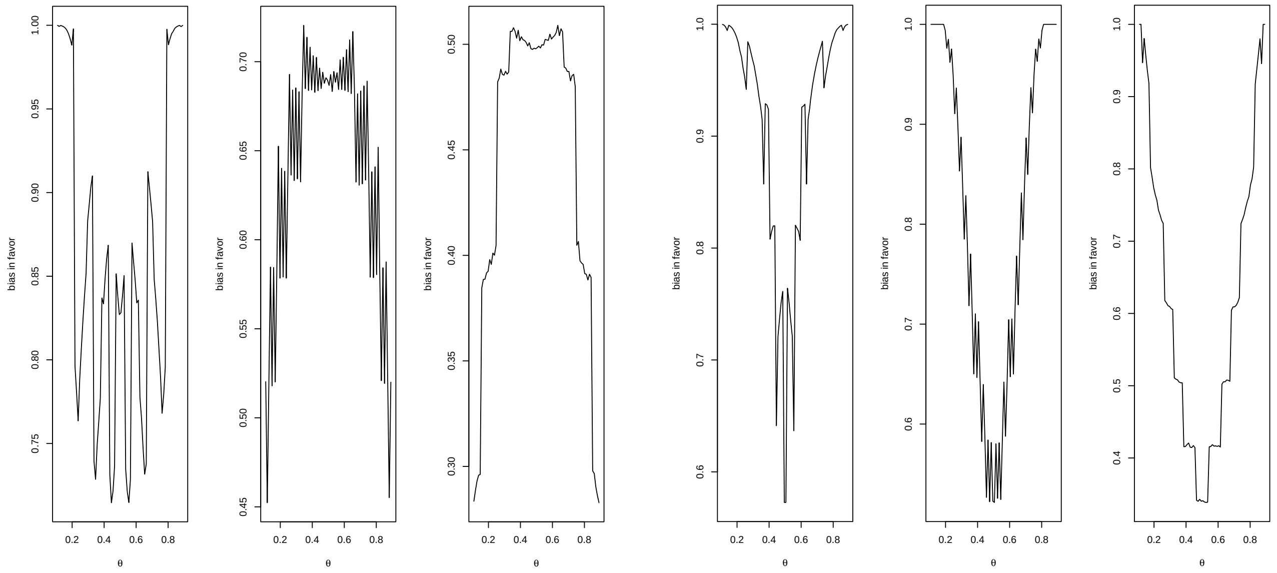
Now, consider bias in favor using (11). When $(\alpha_0, \beta_0) = (1,1)$ and $\delta = 0.1$, Figure 5a gives the plots of the bias in favor for $n = 10$ (max. = 1.00, average = 0.84), $n = 50$ (max. = 0.72, average = 0.51) and $n = 100$ (max. = 0.50, average = 0.35). Therefore, when $n = 10$, the maximum probability that $Pl(x)$ contains a false value at least δ away from the true value is 1, when $n = 50$ this probability is 0.72 and, when $n = 100$, it is a 0.50. When $(\alpha_0, \beta_0) = (5,5)$, Figure 5b gives the plots of the bias in favor for $n = 10$ (max. = 1.00, average = 0.68), for $n = 50$ (max. = 1.00, average = 0.71) and for $n = 100$ (max. = 1.00, average = 0.49). Thus, in this case, the maximum probability that $Pl(x)$ contains a false value at least δ away from the true value is always 1, but, when averaged with respect to the prior, the values are considerably less. It is necessary to either increase n or δ to decrease bias in favor. For example, with $(\alpha_0, \beta_0) = (5,5)$, $\delta = 0.1$ and $n = 400$, the maximum bias in favor is 0.02 and the average bias in favor is 0.02 and, when $n = 600$, these quantities equal 0 to two decimals. When $\delta = 0.2$ and $n = 50$, the maximum bias in favor is 0.29 and the average bias in favor is 0.11 and, when $n = 100$, the maximum bias in favor is 0.01 and the average bias in favor is 0.01.



(**a**) Using a beta(1,1) prior.

(**b**) Using a beta(5,5) prior.

Figure 4. Plots of bias against at θ for $n = 10, 50, 100$ in Example 4.



(**a**) Using a beta(1,1) prior.

(**b**) Using a beta(5,5) prior.

Figure 5. The bias in favor at θ for $n = 10, 50, 100$ with $\delta = 0.1$ in Example 4.

Another interesting case is when the prior is taken to be Jeffreys prior which in this case is the beta$(1/2, 1/2)$ distribution. This reference prior, see [35], is proper and thus can be used with the principle of evidence. The prior does represent somewhat extreme beliefs, however, as 28.7% of the beliefs are that $\theta \in (0, 0.05) \cup (0.95, 1)$. The corresponding biases against are for $n = 10$ (max. = 0.24, average = 0.07), $n = 50$ (max. = 0.09, average = 0.03) and $n = 100$ (max. = 0.07, average = 0.02). The biases in favor are, using (11) with $\delta = 0.1$, for $n = 10$ (max. = 1.00, average = 0.73), $n = 50$ (max. = 0.72, average = 0.59) and $n = 100$ (max. = 0.54, average = 0.41). Although the plots of the bias functions can be seen to be quite different than those for the beta(1,1) prior, the summary values presented are very similar. The beta(1/2,1/2) prior does a bit better with respect to bias against but a bit worse with respect to bias in favor. This reinforces the point that the biases do not serve as a basis for the choice of the prior.

The strange oscillatory nature of the plots for the binomial is difficult to understand but is a common feature with such calculations. For example, Ref. [36] studies the coverage probabilities for various confidence intervals for the binomial, and the following comment is made "The oscillation in the coverage probability is caused by the discreteness of the binomial distribution, more precisely, the lattice structure of the binomial distribution", which still doesn't fully explain the phenomenon.

Example 5. *Location-scale normal quantiles.*

Suppose $x = (x_1, \ldots, x_n)$ is a sample from $N(\mu, \sigma^2)$ with $(\mu, \sigma^2) \in R^1 \times (0, \infty)$ unknown with prior $\mu \,|\, \sigma^2 \sim N(\mu_0, \tau_0^2\sigma^2), \sigma^{-2} \sim \text{gamma}_{\text{rate}}(\alpha_0, \beta_0)$. The hyperparameters $(\mu_0, \tau_0^2, \alpha_0, \beta_0)$ can be obtained via an elicitation as, for example, discussed in Evans and Tomal (2018) for the more general regression model. This example is easily generalized to the regression context. A MSS is $T(x) = (\bar{x}, ||x - \bar{x}\mathbf{1}||^2)$, where $\mathbf{1} = (1, \ldots, 1)'$, with the posterior distribution given by $\mu \,|\, \sigma^2, T(x) \sim N(\mu_{0x}, (n + 1/\tau_0^2)^{-1}\sigma^2), \sigma^{-2} \,|\, T(x) \sim \text{gamma}_{rate}(\alpha_0 + n/2, \beta_{0x})$, where $\mu_{0x} = (n + 1/\tau_0^2)^{-1}(n\bar{x} + \mu_0/\tau_0^2)$ and

$$\beta_{0x} = \beta_0 + ||x - \bar{x}\mathbf{1}||^2/2 + n(\bar{x} - \mu_0)^2/2(n\tau_0^2 + 1).$$

Suppose interest is in the γ-th quantile $\psi = \Psi(\mu, \sigma^2) = \mu + \sigma z_\gamma$, where $z_\gamma = \Phi^{-1}(\gamma)$. To determine the bias for or against ψ, we need the prior and posterior densities of ψ for which there is not a closed form. It is easy, however, to work with the discretized ψ by simply generating from the prior and posterior of (μ, σ^2), estimate the contents of the relevant intervals and then approximate the relative belief ratio using these. Thus, we are essentially approximating the densities by density histograms here, although alternative density estimates could be used. A natural approach to the discretization is to base it on the prior mean $E(\psi) = \mu_0 + \beta_0^{1/2}(\Gamma(\alpha_0 - 1/2)/\Gamma(\alpha_0))z_\gamma$ and variance $Var(\psi) = E(\psi^2) - (E(\psi))^2$ where $E(\psi^2) = (z_\gamma^2 + \tau_0^2)\beta_0/(\alpha_0 - 1)$. Thus, for a given δ, we discretize using $2k + 1$ intervals $(E(\psi) + i\delta, E(\psi) + (i + 1)\delta]$ where $k = cSD(\psi)/\delta$ and c is chosen so that the collection of intervals covers the effective support of ψ which is easily assessed as part of the simulation. For example, with the prior given by hyperparameters $\mu_0 = 0, \tau_0^2 = 1, \alpha_0 = 2, \beta_0 = 1$ and $\gamma = 0.5, \delta = 0.1, c = 5$, then $k = 50$ and, on generating 10^5 values from the prior, these intervals contained 99,699 of the values and with $c = 6$, then $k = 60$, and these intervals contained 99,901 of the generated values. Similar results are obtained for more extreme quantiles because the intervals shift.

For the bias against for estimation, the value of $M(RB_\Psi(\psi \,|\, X) \le 1 \,|\, \psi)$ is needed for a range of ψ values. For this, we need to generate from the conditional prior distribution of T given $\Psi(\mu, \sigma^2) = \psi$, and an algorithm for generating from the conditional prior of (μ, σ^2) given ψ is needed. Putting $\nu = 1/\sigma^2$, the transformation $(\mu, \nu) \to (\psi, \nu) = (\mu + \nu^{-1/2}z_\gamma, \nu)$ has Jacobian equal to 1, so the conditional prior distribution of $\nu \,|\, \psi$ has density proportional to $\nu^{\alpha_0 - 1/2}\exp\{-\beta_0\nu\}\exp\{-\nu(\psi - \mu_0 - \nu^{-1/2}z_\gamma)^2/2\tau_0^2\}$. The following gives a rejection algorithm for generating from this distribution:

1. generate $\nu \sim \text{gamma}(\alpha_0 + 1/2, \beta_0)$,
2. generate $u \sim \text{unif}(0, 1)$ independent of ν,
3. if $u \le \exp\{-\nu(\psi - \mu_0 - \nu^{-1/2}z_\gamma)^2/2\tau_0^2\}$ return ν, else go to 1.

As ψ moves away from the prior expected value $E(\psi)$, this algorithm becomes less efficient, but, even when the expected number of iterations is 86 (when $\gamma = 0.95, \psi = 12$), generating a sample of 10^4 is almost instantaneous. Figure 6 is a plot of the conditional prior of ν given that $\psi = 2$. After generating ν, then generate $||x - \bar{x}1||^2 \sim \nu^{-1}\text{chi-squared}(n - 1)$ and $\bar{x} \sim N(\psi - \nu^{-1/2}z_\gamma, \nu^{-1}/n)$ to complete the generation of a value from $M_T(\cdot \,|\, \psi)$.

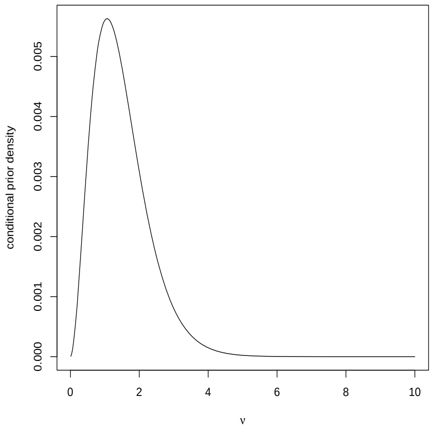


Figure 6. Conditional prior density of $\nu = 1/\sigma^2$ given $\psi = 2$ when $\gamma = 0.95$ and $\mu_0 = 0, \tau_0^2 = 1,$ $\alpha_0 = 2, \beta_0 = 1$ in Example 5.

The bias against as a function of $\psi = \mu + \sigma z_{0.95}$, has maximum value 0.151 when $n = 10$ and so $Pl_\Psi(x)$ is a 0.849-confidence region for ψ while the average bias against is 0.104 implying that the Bayesian coverage is 0.896. Table 7 gives the coverages for other values of n as well. Figure 7 is a plot of the bias in favor as a function of ψ with $\delta = \pm 0.5$ and $n = 10$. The jitter in the right tail is a result of Monte Carlo sampling error, but this error is not of significance as bias measurements are not required to be known to high accuracy. The average bias in favor is 0.629. When $n = 50$, the average bias in favor is 0.335.

Table 7. Coverage probabilities for $Pl_\psi(x)$ for the 0.95 quantile in Example 5.

n	Frequentist Coverage	Bayesian Coverage
10	0.849	0.896
20	0.895	0.927
50	0.934	0.958
100	0.955	0.973

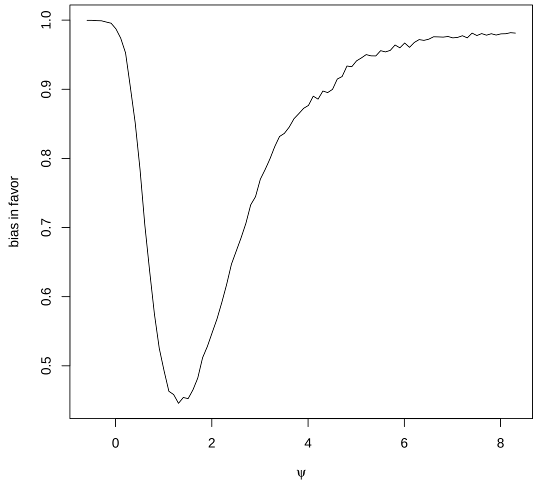


Figure 7. The bias in favor as a function of ψ when $\gamma = 0.95, n = 10, \delta = 0.5$ and using a prior with hyperparameters $\mu_0 = 0, \tau_0^2 = 1, \alpha_0 = 2, \beta_0 = 1$ in Example 5.

The case $\gamma = 0.50$, so $\psi = \Psi(\mu, \sigma^2) = \mu$ is also of interest. For $n = 10$, then $Pl_\Psi(x)$ has 0.878 frequentist coverage and 0.926 Bayesian coverage; when $n = 20$, the coverages are 0.916 and 0.952 while, when $n = 50$, the coverages are 0.950 and 0.973. When $n = 10, \delta = 0.5$, the average bias in favor is 0.619; when $n = 20$, this is 0.4206 and, for $n = 100$, the average bias in favor is 0.091.

Example 6. *Normal Regression—Prediction.*

Prediction problems have some unique aspects when compared to inferences about parameters. To see this, consider first the location normal model of Example 3, and the problem is to make an inference about a future value $y \sim N(\mu, \sigma_0^2)$. The prior predictive distribution is $y \sim N(\mu_0, \tau_0^2 + \sigma_0^2)$ and the posterior predictive is $y \sim N(\mu_x, \sigma_n^2 + \sigma_0^2)$ where $\mu_x = \sigma_n^2(n\bar{x}/\sigma_0^2 + \mu_0/\tau_0^2), \sigma_n^2 = \left(n/\sigma_0^2 + 1/\tau_0^2\right)^{-1}$ so

$$RB(y \,|\, \bar{x}) = \left(\frac{\tau_0^2 + \sigma_0^2}{\sigma_n^2 + \sigma_0^2}\right)^{1/2} \exp\left\{-\frac{1}{2}\left[\frac{(y-\mu_x)^2}{\sigma_n^2 + \sigma_0^2} - \frac{(y-\mu_0)^2}{\tau_0^2 + \sigma_0^2}\right]\right\}.$$

For a given y, the bias against is $M(RB(y \,|\, \bar{x}) \le 1 \,|\, y)$ and, for this, we need the conditional prior predictive of $\bar{x} \,|\, y$. The joint prior predictive is $(\bar{x}, y) \sim N_2(\mu_0 1_2, \Sigma_0)$, where

$$\Sigma_0 = \begin{pmatrix} \tau_0^2 + \sigma_0^2/n & \tau_0^2 \\ \tau_0^2 & \tau_0^2 + \sigma_0^2 \end{pmatrix}$$

and so $\bar{x} \,|\, y \sim N(\mu_0 + \tau_0^2(y-\mu_0)/(\tau_0^2+\sigma_0^2), \sigma_0^2(\tau_0^2/(\tau_0^2+\sigma_0^2) + 1/n))$. From this, we see that, as $n \to \infty$, the conditional prior distribution of $\mu_x \,|\, y$ converges to the

$$N\Big(\mu_0 + \tau_0^2(y-\mu_0)/(\tau_0^2+\sigma_0^2), \sigma_0^2\tau_0^2/(\tau_0^2+\sigma_0^2)\Big)$$

distribution. Thus, with $Z \sim N(0,1)$, $r = \tau_0^2/\sigma_0^2$, and

$$d((y-\mu_0)/\sigma_0, r) = (1+1/r)\log(1+r) + r^{-1}(y-\mu_0)^2/\sigma_0^2),$$

then

$$M(RB(y \,|\, \bar{x}) \le 1 \,|\, y) \to 1 - P(Z \in [r^{-1/2}(1+r)^{-1/2}(y-\mu_0)/\sigma_0 \pm d^{1/2}((y-\mu_0)/\sigma_0, r)])$$

as $n \to \infty$. Thus, the bias against does not go to 0 as $n \to \infty$, and there is a limiting lower bound to the prior probability that evidence in favor of a specific y will not be obtained. This baseline is dependent on both $(y-\mu_0)/\sigma_0$ and r. As $r = \tau_0^2/\sigma_0^2 \to \infty$, this baseline bias against goes to 0 and so it is necessary to ensure that the prior variance is not too small. Table 8 gives some values for the bias against, and it is seen that, if τ_0^2/σ_0^2 is too small, then there is substantial bias against even when y is a reasonable value from the distribution. When $\tau_0^2/\sigma_0^2 = 1, (y-\mu_0)/\sigma_0 = 0$ and $n = 10$, the bias against is computed to be 0.248, which is quite close to the baseline, so increasing sample size will not reduce bias against by much and similar results are obtained for the other cases.

Table 8. Baseline bias against values for prediction for location normal in Example 6.

τ_0^2/σ_0^2	Bias against $(y-\mu_0)/\sigma_0 = 0$	BIAS against $(y-\mu_0)/\sigma_0 = 1$
1	0.239	0.213
10	0.104	0.100
100	0.031	0.031
1/2	0.270	0.263
1/100	0.316	0.460

Now consider bias in favor of y, namely, $M(RB(y \,|\, \bar{x}) \ge 1 \,|\, y \pm \delta)$ for some choice of δ. False values for y correspond to values in the tails so we consider, for example, $y + \delta$ as a value in the central region of the prior and then a large value of δ puts y in the tails. Again, the bias in favor has a baseline value as $n \to \infty$. A similar argument leads to the bias in favor of y satisfying

$$\begin{aligned}&M(RB(y\,|\,\bar{x}) \geq 1\,|\,y \pm \delta) \to \\ &P\left(Z \in \left[r^{-1/2}(1+r)^{-1/2}\left(\frac{y-\mu_0}{\sigma_0} \pm r\frac{\delta}{\sigma_0}\right) \pm d^{1/2}\left(\frac{y-\mu_0}{\sigma_0}, r\right)\right]\right).\end{aligned}$$

Figure 8 is a plot of $\sup M(RB(y\,|\,\bar{x}) \geq 1\,|\,y \pm \delta)$. Thus, the bias in favor is low for central values of y, but, once again, there is a trade-off as when r increases the bias in favor goes to 1.

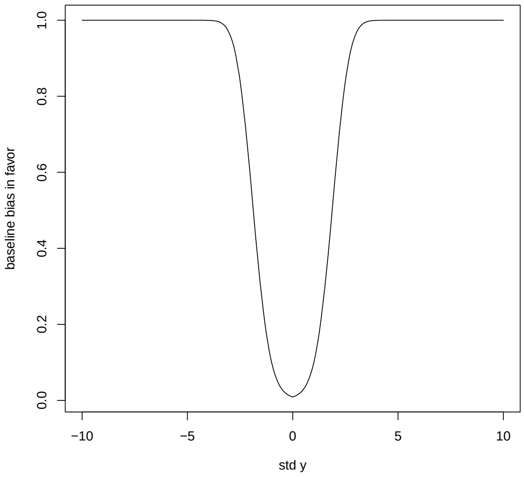


Figure 8. Plot of the baseline bias in favor for values of $(y-\mu_0)/\sigma_0$ when $\tau_0^2/\sigma_0^2 = 1$ when $\delta = 5$ in Example 6.

Prediction plays a bigger role in regression problems, but we can expect the same issues to apply as in the location problem. Suppose $y \sim N_n(X\beta, \sigma^2 I)$, where $X \in R^{n\times k}$ is of rank k, $(\beta, \sigma^2) \in R^k \times (0, \infty)$ is unknown, our interest is in predicting a future value $y_{new} \sim N(w^t\beta, \sigma^2)$ for some fixed known w and, putting $\nu = 1/\sigma^2$, the conjugate prior $\beta\,|\,\nu \sim N_k(\beta_0, \nu^{-1}\Sigma_0)\,\nu \sim \text{gamma}_{\text{rate}}(\alpha_0, \eta_0)$ is used. Specifying the hyperparameters $(\beta_0, \Sigma_0, \alpha_0, \eta_0)$ can be carried out using elicitation as discussed in [37].

For the bias calculations, it is necessary to generate values of the MSS $(b, s^2) = ((X^tX)^{-1}X^ty, ||y - Xb||^2)$ from the conditional prior predictive $M(\cdot\,|\,y_{new})$. This is accomplished by generating from the conditional prior of $(\beta, \nu)\,|\,y_{new}$ and then generating $b \sim N_k(\beta, \nu^{-1}(X^tX)^{-1})$ independent of $s^2 \sim \nu^{-1}$ chi-squared$(n-k)$. The conditional prior of $(\beta, \nu)\,|\,y_{new}$ is proportional to

$$\begin{aligned}&\nu^{\alpha_0 - 1/2}\exp\{-\eta_0(y_{new})\nu\}\times \\ &\nu^{k/2}\exp\left\{-\frac{\nu}{2}\left(\beta - \left(\Sigma_0^{-1} + ww^t\right)^{-1}(\Sigma_0^{-1}\beta_0 + y_{new}w)\right)^t\left(\Sigma_0^{-1} + ww^t\right)(\cdot)\right\}\end{aligned}$$

where

$$\begin{aligned}(\Sigma_0^{-1} + ww^t)^{-1} &= \Sigma_0 - (1 + w^t\Sigma_0 w)^{-1}\Sigma_0 ww^t\Sigma_0, \eta_0(y_{new}) \\ &= \eta_0 + (1 + w^t\Sigma_0 w)^{-1}(w^t\beta - y_{new})^2/2.\end{aligned}$$

Thus, generating $(\beta, \nu)\,|\,y_{new}$ is accomplished via $\nu \sim \text{gamma}_{\text{rate}}(\alpha_0 + 1/2, \eta_0(y_{new}))$,

$$\beta\,|\,\nu \sim N_k\left(\left(I - \frac{\Sigma_0 ww^t}{1 + w^t\Sigma_0 w}\right)(\beta_0 + y_{new}\Sigma_0 w), \nu^{-1}\left(\Sigma_0 - \frac{\Sigma_0 ww^t\Sigma_0}{1 + w^t\Sigma_0 w}\right)\right).$$

For each generated (b, s^2), it is necessary to compute the relative belief ratio $RB(y_{new} \mid b, s^2)$ and determine if it is less than or equal to 1. There are closed forms for the prior and conditional densities of y_{new} since $y_{new} \sim w^t\beta_0 + \{\eta_0(1 + w^t\Sigma_0 w)/\alpha_0\}^{1/2} t_{2\alpha_0}, y_{new} \mid (b, s^2) \sim w^t\beta_0(b, s^2) + \{\eta_0(b, s^2)(1 + w^t(\Sigma_0^{-1} + X^tX)^{-1}w)/(\alpha_0 + n/2)\}^{1/2} t_{2\alpha_0+n}$ where t_λ denotes a Student(λ) random variable and $\beta_0(b, s^2) = (\Sigma_0^{-1} + X^tX)^{-1}(\Sigma_0^{-1}\beta_0 + X^tXb), \eta_0(b, s^2) = \eta_0 + [s^2 + ||Xb||^2 + ||\Sigma_0^{-1}\beta_0||^2 - \beta_0(b, s^2)^t(\Sigma_0^{-1} + X^tX)\beta_0(b, s^2)]/2$. These results permit the calculation of the biases as in the location problem.

6. Conclusions

There are several conclusions that can be drawn from the discussion here. First, it is necessary to take bias into account when considering Bayesian procedures and currently this is generally not being done. Depending on the purpose of the study, some values concerning both bias against and bias in favor need to be quoted as these are figures of merit for the study. The approach to Bayesian inferences via a characterization of evidence makes this relatively straight-forward conceptually. Second, frequentism can play a role in the approach to Bayesian statistical reasoning via relative belief, not through the inferences, but rather through determining the biases and then controlling these through the amount of data collected. Overall, this makes sense because, before the data are seen, it is natural to be concerned about what inferences can be reliably drawn. Once the data are observed, however, it is the evidence in this data set that matters and not the evidence in the data sets not seen. Still, if we ignore the latter, it may be that the existence of bias makes the inferences drawn of very low quality. Third, the results concerning the standard p-value in Example 3 can be seen to apply quite generally, and this makes any discussion about how to characterize and measure evidence of considerable importance. The principle of evidence makes a substantial contribution in this regard as was shown in a variety of results. The major purpose of this paper, however, is to deal with a key criticism of Bayesian methodology, namely that inferences can be biased because of their dependence on the subjective beliefs of the analyst. This criticism is accepted, but we also assert that this can be dealt with in a logical and scientific fashion as has been demonstrated in this paper.

Author Contributions: Conceptualization, M.E. and Y.G.; methodology, M.E. and Y.G.; software, M.E. and Y.G.; validation, M.E. and Y.G.; formal analysis, M.E. and Y.G.; investigation, M.E. and Y.G.; writing—original draft preparation, M.E.; writing—review and editing, M.E. and Y.G.; supervision, M.E.; funding acquisition, M.E. Both authors have read and agreed to the published version of the manuscript.

Funding: The work of Evans was supported by a grant 10671 from the Natural Sciences and Engineering Research Council of Canada.

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: Not applicable.

Acknowledgments: The authors thank three reviewers for their constructive comments.

Conflicts of Interest: The authors declare no conflict of interest.

Appendix A

Proof that the difference of p-values in Example 3 is a valid measure of evidence. The Savage–Dickey ratio result implies that $RB_\Psi(\psi \mid x) = m_\psi(x)/m(x)$, where m denotes the prior predictive density of x, and m_ψ denotes the conditional prior predictive density of x given that $\Psi(\theta) = \psi$. Furthermore, the data can be reduced to the minimal sufficient statistic. In Example 1, the prior predictive of $\bar{x}$ is $N(\mu_0, \tau_0^2 + \sigma_0^2/n)$, and the prior predictive given μ is $N(\mu, \sigma_0^2/n)$. Therefore,

$$RB(\mu \mid x) = (1 + n\tau_0^2/\sigma_0^2)^{1/2} \exp\{-n(\bar{x} - \mu)^2/2\sigma_0^2 + (\bar{x} - \mu_0)^2/2(\tau_0^2 + \sigma_0^2/n)\}$$

so $RB(\mu_* \,|\, x) \le 1$ iff

$$\begin{aligned}
&\frac{n(\bar{x}-\mu_*)^2}{\sigma_0^2} \ge \log\left(1+\frac{n\tau_0^2}{\sigma_0^2}\right) + \frac{(\bar{x}-\mu_0)^2}{2(\tau_0^2+\sigma_0^2/n)} \text{ iff } \Phi\left(\frac{\sqrt{n}|\bar{x}-\mu_*|}{\sigma_0}\right) \ge \\
&\Phi\left(\left\{\log\left(1+\frac{n\tau_0^2}{\sigma_0^2}\right)+\left(1+\frac{\sigma_0^2}{n\tau_0^2}\right)^{-1}\frac{(\bar{x}-\mu_0)^2}{\tau_0^2}\right\}^{1/2}\right) \text{ iff} \\
&2\left(1-\Phi\left(\frac{\sqrt{n}|\bar{x}-\mu_*|}{\sigma_0}\right)\right) - \\
&2\left(1-\Phi\left(\left\{\log\left(1+\frac{n\tau_0^2}{\sigma_0^2}\right)+\left(1+\frac{\sigma_0^2}{n\tau_0^2}\right)^{-1}\frac{(\bar{x}-\mu_0)^2}{\tau_0^2}\right\}^{1/2}\right)\right) \le 0.
\end{aligned}$$

Proof of Theorem 1. The Savage–Dickey ratio result implies $RB_\Psi(\psi_* \,|\, x) = m_{\psi_*}(x)/m(x)$ and note $R(\psi_*) = \{x : m_{\psi_*}(x) \le m(x)\}$. Now, put

$$\begin{aligned}
\mathcal{X}_1 &= \{x : I_{R(\psi_*)}(x) - I_{D(\psi_*)}(x) < 0\} \\
&= \{x : I_{R(\psi_*)}(x) - I_{D(\psi_*)}(x) < 0, m_{\psi_*}(x) > m(x)\} \\
\mathcal{X}_2 &= \{x : I_{R(\psi_*)}(x) - I_{D(\psi_*)}(x) > 0\} \\
&= \{x : I_{R(\psi_*)}(x) - I_{D(\psi_*)}(x) \ge 0, m_{\psi_*}(x) \le m(x)\}.
\end{aligned}$$

Then,

$$\begin{aligned}
M(R(\psi_*)) - M(D(\psi_*)) = &\int_{\mathcal{X}_1} (I_{R(\psi_*)}(x) - I_{D(\psi_*)}(x))\, M(dx) + \\
&\int_{\mathcal{X}_2} (I_{R(\psi_*)}(x) - I_{D(\psi_*)}(x))\, M(dx) \\
&\ge M(R(\psi_*) \,|\, \psi_*) - M(D(\psi_*) \,|\, \psi_*) \ge 0
\end{aligned}$$

establishing (i). In addition,

$$M(D(\psi_*)) = M(D(\psi_*) \,|\, \psi_*)\Pi_\Psi(\{\psi_*\}) + \int_{\Psi\setminus\{\psi_*\}} M(D(\psi_*) \,|\, \psi)\, \Pi_\Psi(d\psi)$$

and the integral is the prior probability of not getting evidence in favor of ψ_* when it is false, and this establishes (ii). □

Proof of Theorem 2. Now,

$$\begin{aligned}
&E_{\Pi_\Psi}(M(\psi_* \notin C(X))) = E_{\Pi_\Psi^2}(M(\psi_* \notin C(X) \,|\, \psi)) \\
&= E_{\Pi_\Psi^2}(M(D(\psi_*)) \,|\, \psi)) = \int_\Psi M(D(\psi_*))\, \Pi_\Psi(d\psi_*)
\end{aligned}$$

and (i) follows from Theorem 1. In addition,

$$\begin{aligned}
&\int_\Psi M(D(\psi_*))\, \Pi_\Psi(d\psi_*) = E_{\Pi_\Psi}(\int_\Psi M(D(\psi_*) \,|\, \psi)\, \Pi_\Psi(d\psi)) \\
&= E_{\Pi_\Psi}(M(D(\psi_*) \,|\, \psi_*)\Pi_\Psi(\{\psi_*\})) + E_{\Pi_\Psi}(\int_{\Psi\setminus\{\psi_*\}} M(D(\psi_*) \,|\, \psi)\, \Pi_\Psi(d\psi)) \\
&= E_{\Pi_\Psi}(M(\psi_* \notin C(X) \,|\, \psi_*)\Pi_\Psi(\{\psi_*\})) + \\
&E_{\Pi_\Psi}(\int_{\Psi\setminus\{\psi_*\}} M(\psi_* \notin C(X) \,|\, \psi)\, \Pi_\Psi(d\psi))
\end{aligned}$$

establishing (ii). □

Proof of Theorem 3. Now,

$$\begin{aligned} M(R(\psi_*)\,|\,\psi_*) &= \int I_{R(\psi_*)}(x)\,M_{\psi_*}(dx) \le \int I_{R(\psi_*)}(x)\,M(dx) = M(R(\psi_*)) \\ &= \int_\Psi M(R(\psi_*)\,|\,\psi)\,\Pi(d\psi) = M(R(\psi_*)\,|\,\psi_*)\Pi_\Psi(\{\psi_*\}) \\ &+ \int_{\Psi\setminus\{\psi_*\}} M(R(\psi_*)\,|\,\psi)\,\Pi_\Psi(d\psi) \end{aligned}$$

so $\Pi_\Psi(\{\psi_*\}^c)M(R(\psi_*)\,|\,\psi_*) \le \int_{\Psi\setminus\{\psi_*\}} M(R(\psi_*)\,|\,\psi)\,\Pi_\Psi(d\psi)$ which implies (i). Furthermore, (ii) is implied by

$$\begin{aligned} &E_{\Pi_\Psi}(M(\psi_* \notin Pl_\Psi(X)\,|\,\psi_*)) = E_{\Pi_\Psi}(M(R(\psi_*)\,|\,\psi_*)) \\ &\le E_{\Pi_\Psi}(\int_{\Psi\setminus\{\psi_*\}} M(R(\psi_*)\,|\,\psi)\,\Pi_\Psi(d\psi)/\Pi_\Psi(\{\psi_*\}^c)) \\ &= E_{\Pi_\Psi}(\int_{\Psi\setminus\{\psi_*\}} M(\psi_* \notin Pl_\Psi(X)\,|\,\psi)\,\Pi_\Psi(d\psi)/\Pi_\Psi(\{\psi_*\}^c). \end{aligned}$$

□

Proof of Theorem 4. It is easy to see that the proof of Theorem 1 can be modified to show that, among all regions, $D^{int}(\psi_*) \subset \mathcal{X}$ satisfying $M(D^{int}(\psi_*)\,|\,\psi_*) \le M(RB_\Psi(\psi_*\,|\,X) < 1\,|\,\psi_*)$ the prior probability $M(D^{int}(\psi_*))$ is maximized by $D^{int}(\psi_*) = \{x : RB_\Psi(\psi_*\,|\,x) < 1\}$. This implies that (i) and (ii) are similar. □

Proof of Theorem 5. Now,

$$\begin{aligned} &E_{\Pi_\Psi}(M(\psi_* \in C(X))) = E_{\Pi_\Psi^2}(M(\psi_* \in C(X)\,|\,\psi)) \\ &= E_{\Pi_\Psi^2}(M(D^c(\psi_*))\,|\,\psi)) = E_{\Pi_\Psi}(M(D^c(\psi_*)) \end{aligned}$$

and (i) follows from Theorem 1 (i). In addition, (ii) is implied by

$$\begin{aligned} E_{\Pi_\Psi}(M(D^c(\psi_*)) &= \int_\Psi M(D^c(\psi_*)\,|\,\psi_*)\Pi_\Psi(\{\psi_*\})\,\Pi_\Psi(d\psi_*) + \\ &\int_\Psi \int_{\Psi\setminus\{\psi_*\}} M(D^c(\psi_*)\,|\,\psi)\,\Pi_\Psi(d\psi)\,\Pi_\Psi(d\psi_*) \\ &= \int_\Psi M(\psi_* \in C(X)\,|\,\psi_*)\Pi_\Psi(\{\psi_*\})\,\Pi_\Psi(d\psi_*) + \\ &\int_\Psi \int_{\Psi\setminus\{\psi_*\}} M(\psi_* \in C(X)\ |\,\psi)\,\Pi_\Psi(d\psi)\,\Pi_\Psi(d\psi_*). \end{aligned}$$

□

References

1. Baskurt, Z.; Evans, M. Hypothesis assessment and inequalities for Bayes factors and relative belief ratios. *Bayesian Anal.* **2013**, *8*, 569–590. [CrossRef]
2. Evans, M. Measuring Statistical Evidence Using Relative Belief. In *Monographs on Statistics and Applied Probability 144*; CRC Press: Boca Raton, FL, USA, 2015.
3. Nott, D.; Wang, X.; Evans, M.; Englert, B.-G. Checking for prior-data conflict using prior to posterior divergences. *Stat. Sci.* **2020**, *35*, 234–253. [CrossRef]
4. Robert, C.P. On the Jeffreys–Lindley paradox. *Philos. Sci.* **2014**, *81*, 216–232. [CrossRef]
5. Shafer, G. Lindley's paradox (with discussion). *J. Am. Stat. Assoc.* **1982**, *77*, 325–351. [CrossRef]
6. Spanos, A. Who should be afraid of the Jeffreys–Lindley paradox? *Philos. Sci.* **2013**, *80*, 73–93. [CrossRef]
7. Sprenger, J. Testing a precise null hypothesis: The case of Lindley's paradox. *Philos. Sci.* **2013**, *80*, 733–744. [CrossRef]
8. Cousins, R.D. The Jeffreys–Lindley paradox and discovery criteria in high energy physics. *Synthese* **2017**, *194*, 395–432. [CrossRef]
9. Villa, C.; Walker, S. On the mathematics of the Jeffreys–Lindley paradox. *Commun. Stat. Theory Methods* **2017**, *46*, 12290–12298. [CrossRef]

10. Gu, Y.; Li, W.; Evans, M.; Englert, B.-G. Very strong evidence in favor of quantum mechanics and against local hidden variables from a Bayesian analysis. *Phys. Rev. A* **2019**, *99*, 022112. [CrossRef]
11. Birnbaum, A. The anomalous concept of statistical evidence:axioms, interpretations and elementary exposition. In *IMM NYU-332*; Courant Institute of Mathematical Sciences: New York, NY, USA, 1964.
12. Aitkin, M. *Statistical Inference: An Integrated Bayesian/Likelihood Approach*; Chapman and Hall/CRC: Boca Raton, FL, USA, 2010.
13. Morey, R.; Romeijn, J.-W.; Rouder, J. The philosophy of Bayes factors and the quantification of statistical evidence. *J. Math. Psychol.* **2016**, *72*, 6–18. [CrossRef]
14. Royall, R. *Statistical Evidence: A Likelihood Paradigm*; Chapman and Hall/CRC: Boca Raton, FL, USA, 1997.
15. Shafer, G. *A Mathematical Theory of Evidence*; Princeton University Press: Princeton, NJ, USA, 1976.
16. Thompson, B. The Nature of Statistical Evidence. In *Lecture Notes in Statistics 189*; Springer: Berlin/Heidelberg, Germany, 2007.
17. Vieland, V.J.; Seok, S.-J. Statistical evidence measured on a properly calibrated scale for multinomial hypothesis comparisons. *Entropy* **2016**, *18*, 114. [CrossRef]
18. Achinstein, P. *The Book of Evidence*; Oxford University Press: Oxford, UK, 2001.
19. Salmon, W. Confirmation. *Sci. Am.* **1973**, *228*, 75–81. [CrossRef]
20. Popper, K. *The Logic of Scientific Discovery*; Harper Torchbooks: New York, NY, USA, 1968.
21. Keynes, J.M. *A Treatise on Probability*; Wildside Press LLC: Rockville, MD, USA, 1921.
22. Stanford Encyclopedia of Philosophy. Confirmation. 2020. Available online: https://plato.stanford.edu/ (accessed on 3 February 2021)
23. Evans, M.; Jang, G.-H. A limit result for the prior predictive applied to checking for prior-data conflict. *Stat. Probab. Lett.* **2011**, *81*, 1034–1038. [CrossRef]
24. Evans, M.; Moshonov, H. Checking for prior-data conflict. *Bayesian Anal.* **2006**, *1*, 893–914. [CrossRef]
25. Gelman, A.; Hennig, C. Beyond subjective and objective in statistics. *J. R. Stat. Soc. A* **2017**, *180*, 967–1033. [CrossRef]
26. Gelman, A.; Shalizi, C.R. Philosophy and the practice of Bayesian statistics. *Br. J. Math. Stat. Psychol.* **2013**, *66*, 8–38. [CrossRef] [PubMed]
27. Berger, J.O.; Delampady, M. Testing precise hypotheses. *Stat. Sci.* **1987**, *2*, 317–335. [CrossRef]
28. Berger, J.O.; Selke, T. Testing a point null hypothesis: The irreconcilability of p values and evidence. *J. Am. Assoc.* **1987**, *82*, 112–122. [CrossRef]
29. Al-Labadi, L.; Baskurt, Z.; Evans, M. Goodness of fit for the logistic regression model using relative belief. *J. Stat. Appl.* **2017**, *4*, 17. [CrossRef]
30. Boring, E. Mathematical vs. statistical significance. *Psychol. Bull.* **1919**, *16*, 335–338. [CrossRef]
31. Evans, M.; Guttman, I.; Swartz, T. Optimality and computations for relative surprise inferences. *Can. J. Stat.* **2006**, *34*, 113–129. [CrossRef]
32. Spanos, A.; Mayo, D. Severe testing as a basic concept in a Neyman–Pearson philosophy of induction. *Br. J. Philos. Sci.* **2006**, *57*, 323–357.
33. Rochefort-Maranda, G. Inflated effect sizes and underpowered tests: How the severity measure of evidence is affected by the winner's curse. *Phil. Stud.* **2020**, *178*, 133–145. [CrossRef]
34. Evans, M.; Guttman, I.; Li, P. Prior elicitation, assessment and inference with a Dirichlet prior. *Entropy* **2017**, *19*, 564. [CrossRef]
35. Berger, J.O.; Bernardo, J.M.; Sun, D. The formal definition of reference priors. *Ann. Stat.* **2009**, *37*, 905–938. [CrossRef]
36. Brown, L.D.; Cai, T.; DasGupta, A. Interval estimation for a binomial proportion. *Stat. Sci.* **2001**, *16*, 101–133.
37. Evans, M.; Tomal, J. Multiple testing via relative belief ratios. *Facets* **2018**, *3*, 563–583. [CrossRef]