# The philosophy of Bayes factors and the quantification of statistical evidence


Richard D. Morey*

*University of Groningen*

Jan-Willem Romeijn*

*University of Groningen*

Jeffrey N. Rouder[1,*]

*University of Missouri*



**Abstract**

A core aspect of science is using data to assess the degree to which data provide evidence various claims, hypotheses, or theories. Evidence is by definition something that should change the credibility of a claim in a reasonable person's mind. However, common statistics, such as significance testing and confidence intervals have no interface with concepts of belief, and thus it is unclear how they relate to statistical evidence. We explore the concept of statistical evidence, and how it can be quantified using the Bayes factor. We also discuss the philosophical issues inherent in the use of the Bayes factor.



*Keywords:* Bayes factor, Hypothesis testing


A core element of science is that data are used to argue for or against hypotheses or theories. Researchers assume that data — if properly analysed — provide evidence, whether this evidence is used to understand global climate change (Lawrimore et al., 2011), examine whether the Higgs Boson exists


*Address correspondence concerning this article to Richard Morey.

*Email addresses:* richarddmorey@gmail.com (Richard D. Morey), j.w.romeijn@rug.nl (Jan-Willem Romeijn), rouderj@missouri.edu (Jeffrey N. Rouder)



[1]This research is supported by NSF Grants SES 1024080 and BCS 1240359.

Low et al. (2012), explore the evolution of bacteria (Barrick et al., 2009), or to describe human reasoning (Kahneman and Tversky, 1972). Scientists using statistics often write as if evidence is quantifiable: one can have no evidence, weaker evidence, stronger evidence — but importatly, statistics in common use, such as significance tests and confidence intervals, do not admit such interpretations (Berger and Sellke, 1987; Jeffreys, 1961; Wagenmakers et al., 2008; Berger and Wolpert, 1988). Instead, they are designed to make decisions, such as rejecting a hypothesis, rather than providing for a measure of evidence. Consequently, statistical practice is often beset by a difference between what statistics provide and what is desired from them.

In this paper, we explore a statistic that does have the desired interpretation as a measure of evidence for theories from data: the Bayes factor (Good, 1985, 1979; Jeffreys, 1961; Kass and Raftery, 1995). To arrive at the Bayes factor, however, we first explore the concept of evidence more generally. We show that formalizing evidence in a particular way — in a way that makes it useful, in fact — points to Bayesian statistics. We then describe how Bayes factors can be used in practice with an example, focusing the the philosophical issues that arise when using Bayes factors. Finally, in the discussion we consider critiques of Bayes factors as measures of evidence, and difficulties inherent in their application.

## 1. Evidence

What is evidence? One natural answer is that the evidence presented by data is the impact that the data have on our evaluation of a hypothesis (e.g., Fox, 2011). This is a straightforward general notion of evidence, popular among methodologists, epistemologists and philosophers of science alike. We will adopt and elaborate this view. Specifically, we review some philosophical ideas on the relation between scientific theory and empirical fact; or, in more scientific parlance, between hypotheses and data.[2]

Although our discussion is not specifically limited to statistics, its relevance for statistics easily becomes apparent. Our central claim is that the computation of Bayes factors is an appropriate, appealing method for as-

[2]Although there is a huge debate within the philosophy of science about the relation between data, facts, phenomena, and the like (e.g., Bogen and Woodward, 1988), we will align ourselves with scientific practice here and simply employ the term "data" without making further discriminations.

sessing the impact of data on the evaluation of hypotheses. In short, Bayes factors formalize a useful and meaningful notion of evidence. In order to show why Bayes factors are useful, we must develop a normative account of evidence that ties together notions central to evidence: hypotheses, evaluation, data. In particular, we develop a notion of evidence that relates to a particular goal of science and introduce Bayes factors in abstract terms, as a natural expression of this notion of evidence. Section 2.1 then provides a detailed introduction into the use of Bayes factors in statistics. In section 4, finally, we connect our notion of evidence to various possible merits and defects of Bayes factors in statistics.

### *1.1. Epistemic goals*

Scientific inquiry is concerned with many diverse goals. One possible goal of science, for instance, is that we look for reliable means to manipulate the world and bring about certain states of affairs. This goal fits with a pragmatic, instrumentalist attitude, according to which theory serves as an instrument: it is enough to have the means to predict the world on the basis of a distinct set of, preferably controllable, variables. The format of a predictive system is secondary to this goal. In particular, there is no conclusive reason to expect that the predictive system will employ general hypotheses on how the world works, or that it will involve beliefs about those hypotheses. For example, the predictive system could involve a neural network with nodes and links that do not bear any natural interpretation.

A second goal of science, which serves as the main focus of this article, is epistemic: science must offer us an adequate representation of the world, or at least one that lends itself for generating explanations as well as predictions. This goal puts some constraints on what the format of theory might be, and more generally on our account of evidence. For one, to serve representational goals scientific theory will have to interface with our beliefs. This is more than merely requiring that our theories interface with the principles that guide our actions. Of course, some principles guiding action are already cast in epistemic terms, e.g., standard decision theory, and this may be reason enough to engage with beliefs. Our point is that in an instrumentalist view of science the interface with belief is not mandatory, while in an epistemic view of science it is.

The idea that scientific inquiry has implications for belief is common among scientists. One important example of recent import is the debate over global climate change. The epistemic nature of this debate is hard to miss. Much attention has been given, for instance, to the *consensus* of climate scientists; that is, that nearly all climate scientists believe that global climate change is caused by humans. The available data is assumed to drive climate scientists beliefs; the fact of consensus then drives public opinion and policy on the topic. Those not believing with the consensus are called, pejoratively, "deniers" (Dunlap, 2013).

A major constraint that flows from the epistemic goals of the scientific enterprise concerns the format of scientific theory: namely, that it contains components that represent nature, or the world, in some manner. We call those components hypotheses, here denoted as $\mathbf{h}$.[3] There is a remarkable variety of structures that may all be classified as hypotheses in virtue of their role in representing the world. A hypothesis might be a distinct mechanism, the specification of a type of process, a particular class of solutions to some system of equations, and so on. For all hypotheses, however, an important requirement is that they entail, or at least make predictions regarding, data. Scientists would regard hypothesis that has no empirical consequences as problematic. According to a deeply seated conviction among many scientists, the success of a theory can only be determined on the basis of its ability to reproduce or match patterns in the data. Science is empirical, and the representational means of science must accordingly be empirical as well.

We should add that most of the above claims are subject to controversy. There is a long-standing debate in the philosophy of science that is concerned with the use and status of theory. It is far from clear that all theoretical structure is intended to represent, and that theoretical structure always has import for the empirical content of scientific theory. However, for our arguments it suffices that epistemic goals are not entirely absent in our scientific endeavors.[4]

### *1.2. Hypotheses and beliefs*

The foregoing considerations lead to a particular understanding of scientific theory: it consists of empirical hypotheses that somehow or other represent the world. Within statistical analysis, we indeed find that theory has

[3]In the philosophy of science literature, those structures are often referred to as models. But in a statistical context models have a specific meaning: sets of distributions over sample space that serve as input to a statistical analysis. To avoid confusion when we introduce statistical models later, we use the term "hypotheses".

[4]See, e.g., Psillos (1999); Bird (1998) for introductions into the so-called realism debate.

this character: statistical hypotheses are distributions that represent a population, and they entail probability assignments to events in a sample space. Notice that the theoretical structure from which the distribution arises may be far richer than the distribution itself, involving exemplars, stories, bits of metaphysics, and so on. In the philosophy of statistics, there is ongoing debate about the exact use of this theoretical superstructure, and the extent to which it can be detached from the empirical substructure.[5]

It may seem a trivial matter that scientific theory takes on the format of empirical hypotheses. But a closer look at science can give us a more nuanced view of what theory might be. Consider a statistical tool like principal component analysis, in which the variation among data points is used to identify salient linear combinations of manifest variables. Importantly, this is a data-driven technique that does not rely on any explicitly formulated hypothesis. The use of neural networks and other data-mining tools for identifying empirical patterns are also cases in point. The message here is that scientific theory need not always have components that do representational work. But the account of evidence that motivates Bayes factors does rely on hypotheses as representational items.

Another major consequence of the treating science as an epistemic enterprise, already touched on in the foregoing, is that scientific theory must interface with our epistemic attitudes. These attitudes include expectations, convictions, opinions, commitments, assumptions, and more, but for ease of reference we will speak of beliefs in what follows. Now that we have identified the representational components of scientific theory as hypotheses, the requirement is that hypotheses must feature in our beliefs. Our account of evidence must accommodate such a role.

The exact implications of the involvement of belief depend on what we take to be the nature of beliefs, and the specifics of the items featuring in it. There is not a uniquely best way of representing beliefs or the targets of beliefs. For example, when expressing the strength of our adherence to a belief, one extreme is to take them as categorical, e.g., dichotomous between accepted and rejected. But beliefs may be captured by more fine-grained formalizations, e.g., degrees of belief, imprecise probabilities, plausibility orderings and so on (see Halpern, 2003, for an overview). Moreover, the beliefs

[5]Romeijn (2013) offers a recent discussion of this point, placing hierarchical Bayesian models in the context of explanatory reasoning in science.

need not concern the hypothesis in isolation. We are seeking an account of evidence that accommodates the epistemic goals of science. But in such an account, the beliefs might just as well pertain to distinct pairs of hypotheses and data.

The upshot of this is that the involvement of hypotheses and beliefs does not, by itself, impose the use of Bayesian methods to the exclusion of others. Numerous interpretations of, and add-ons to, classical statistics have been developed to accommodate the need for an epistemic interpretation of results (for an overview see Romeijn, 2014). Nothing is said, as yet, about the kind of belief involved in the evaluation of hypotheses, and for good reasons: a normative account of evidence that is supposed to motivate a particular statistical method must not itself presuppose such a method.

### *1.3. Beliefs and probabilities*

The evaluation of empirical hypotheses consists in determining how well the hypotheses align with the data. But how can the data serve as evidence, i.e., how precisely do the data engage in our beliefs towards hypotheses? To answer this question, we first discuss a means of expressing beliefs. This sets the stage for a discussion of how beliefs and data interface.

Beliefs may be expressed in many ways. One important choice concerns the representation of the items about which we have beliefs. For example, we might frame our beliefs as pertaining to sentences, or some other kind of linguistic entity. A very general framework for beliefs presents them as as pertaining to elements from an algebra that represents events in, or facts about a target system. The beliefs themselves may then be formalized in terms of a function over the algebra, e.g., with truth value ascriptions or more fine-grained valuations. In what follows we will adopt this framework.

In philosophy, psychology, artificial intelligence, and in statistics, it is commonplace to formalize beliefs in terms of probability assignments over the algebra of events. In classical statistics the primary interpretation of these probabilities is, of course, different: they reflect frequencies in a population rather than beliefs. But even those frequencies are typically taken as a basis for expectations concerning random variables, and thus they relate to a particular kind of belief, albeit in a derivative way. For present purposes, the salient point is that if we decide to formalize beliefs—predictions, expectations, convictions, commitments—as part of an analysis of the evaluation of hypotheses, then there are convincing reasons for doing this in terms of probability assignments (Cox, 1946; de Finetti, 1995; Joyce, 1998; Ramsey, 1931).

The use of probabilities to express beliefs suggests a particular way of formalizing the evaluation of hypotheses by data. We express our beliefs in a probability assignment, i.e., by a measure function over an algebra. Items that obtain a probability, like data and possibly also hypotheses, are elements of this algebra. The relation between a hypothesis, denoted $\mathbf{h}$, and data, denoted $\mathbf{y}$, can thus be captured by certain valuations of this probability function. As will become apparent, a key role is reserved for the probability of the data on the assumption of a hypothesis, written $p_{\mathbf{h}}(\mathbf{y})$ or $p(\mathbf{y} \mid \mathbf{h})$ depending on the exact role given to hypotheses; in particular, classical statisticians might object to the appearance of $\mathbf{h}$ within the scope of the probability function $p$. If viewed as a function of the hypothesis, this expression is referred to as the (marginal) likelihood of the hypothesis $\mathbf{h}$ for the (known and fixed) data $\mathbf{y}$.

At this point it should be noted that the use of probability assignments puts further constraints on the nature of empirical hypotheses: the hypotheses must be such that a distinct probability assignment over possible data can be specified. In other words, the hypothesis must be *statistical*. Moreover, if the hypothesis under consideration is composite – meaning that it consists of a number of different distributions over sample space – then we must suppose a probability assignment over these distributions themselves in order to arrive at a single-valued probability over sample space. For instance, if we are interested in the probability $\theta$ that an unfair coin lands with heads showing, then the hypothesis $\theta > 0$, which specifies that the coin is biased toward heads, is such a composite hypothesis. Each possible value for $\theta$ implies a different sampling distribution over the number of heads. In addition to these sampling distributions we must have a weighting over all possible $\theta$ values. Without a probability assignment over these component distributions, the marginal likelihood of the hypothesis cannot be computed, thereby leaving the empirical content of the hypothesis unspecified.

So far we have argued that, insofar as scientific theory serves the goal of adequate representation, it involves beliefs concerning hypotheses. Following a deeply rooted assumption of empiricism, these beliefs are determined by the relations that obtain between hypotheses and data. And finally, we have argued that probability assignments offer a natural means for expressing these beliefs. Against this background, we will now investigate how data impacts on hypotheses and thereby turns into evidence. To motivate the use of likelihoods, we need a qualitative account of the relation between scientific theory and data.

### *1.4. Support: comparative and context-sensitive*

The data—in the context of statistics, elements from a sample space—do not present evidence all by themselves. The term evidence is suggestive of a context that turns dry database entries into something meaningful: that is, a context in which the data play a distinct role. To specify that context, we focus on the relative and comparative nature of support relations as a basis for our account of evidence. Subsequently we offer an account of evidence itself.

One way of adopting a belief about a hypothesis is by evaluating the hypothesis directly: e.g., by offering, in the light of the data, an absolute verdict regarding its truth or falsity. By contrast, we might also evaluate the relation between hypothesis and data, e.g., by forming a belief regarding the support that the data give to the hypothesis. The notion of support concerns a relation between hypothesis and data, and this is different from a belief that only pertains to the hypothesis in isolation. In statistics, for example, the notion of support hinges on the aforementioned probability that the hypothesis assigns to the data, written $p(\mathbf{y} \mid \mathbf{h})$.

Whether we opt for a verdict about a hypothesis itself, or for one that pertains to the relation between hypothesis and data, a crucial role is played by the alignment of hypotheses to those data. A natural way of spelling out this so-called empirical adequacy is by a measure of predictive accuracy. That is, hypotheses are scored and compared according to how well they predict the data. Notice that predictions based on a hypothesis have an epistemic nature—they are expectations—but that their standard formalization in terms of probability is usually motivated by the probabilistic nature of something non-epistemic: often hypotheses pertain to frequencies or chances, and the latter can be formalized using probability theory as well. The use of predictions for evaluating hypotheses thus involves two subtle conceptual steps. The probability $p(\mathbf{y} \mid \mathbf{h})$ refers to a chance ascription, which is then turned into an epistemic expectation, and subsequently into a score that expresses the support for the hypothesis by the data.

Apart from the relational nature of support, support can be considered in absolute or in relative terms. We might conceive of the support as something independent of the theoretical context in which our belief regarding the support is reached. For example, we may be tempted link the notion of support *solely* to how well the hypothesis predicts the data. It might appear that the predictive performance may be judged independently of how well other hypotheses – which may or may not be under consideration – predict those data. By contrast, we might also conceive of support as an essentially comparative affair. For example, we may consider one hypothesis to be better supported by the data than another because it predicts the data better, without saying anything about the absolute support that either receives from the data.

We think the comparative reading fits better with our intuitive understanding of support, namely as something context-sensitive. Indeed, we maintain that the data simply cannot offer support in absolute terms: they can only do so relative to rival hypotheses. Imagine that the hypothesis $\mathbf{h}$ predicts the empirical data $\mathbf{y}$ with very high probability. We will only say that the data $\mathbf{y}$ support the hypothesis $\mathbf{h}$ if other hypotheses $\mathbf{h}'$ do not predict the same data. If the other hypotheses also predict the data, perhaps because it is rather easy to predict them, then it seems that those data do not offer support either way. Moreover, even if the data are surprising in the sense that they have a low probability according to all the other hypotheses under consideration, then still, they are only surprising relative to those other hypotheses. In short, the notion of support seems to be dependent on what candidate hypotheses are being considered. We note, however, that relative support is a meaningful measure of the quality of a hypothesis, regardless of whether absolute support is considered attainable. We therefore advance a notion of support that is inherently relative, keeping open that relative support might lead to absolute support.

Summing up, we argued that support can be measured by predictive success, that it has a comparative and context-sensitive character, and that it may apply to hypotheses themselves or to the relations that obtains between hypotheses and data. In the remainder of this section, we will integrate these insights into an account of evidence and argue that Bayes factors offer a natural expression of this kind of evidence.

### *1.5. Bayes factors*

Let us return to the conception of evidence that was sketched at the start of this section. We stipulated that the evidence presented by the data is the impact that these data have on our evaluation of theory.[6] First, we associated theory with empirical hypotheses that have a role in representation. It was then argued that the evaluation of hypotheses involves beliefs, which were represented as probabilities and related to a notion of support. Finally, this impact will now be spelled out as the difference between our beliefs concerning hypotheses, before and after we received the data.

We can develop the idea of impact in several ways, depending on the contents of our beliefs about hypotheses. One option we have previously noted is to spell out the beliefs about hypotheses in terms of the relational notion of support. The evidence presented by a datum is then defined as the impact it has on the support relation. Using predictive accuracy, measured by the probability assignment $p(\mathbf{y} \mid \mathbf{h})$, as expression of support, we might formalize the evidence presented by a new datum $\mathbf{y}$ against the background knowledge $\mathbf{b}$, in terms of changes to the likelihoods upon receiving $\mathbf{y}$. This would lead to some expression involving $p(\mathbf{b})$ and $p(\mathbf{y} \cap \mathbf{b})$. A comparative version of that would also involve these terms for alternative hypotheses $\mathbf{h}'$.

In what follows we adopt a slightly different notion of evidence, in which hypotheses themselves are the subject of evaluation. Hence we look at the way in which data impact on the evaluation of hypotheses $\mathbf{h_i}$ as such. Ignoring background knowledge for notational ease, the evidence presented by the datum $\mathbf{y}$ can thus be formalized in terms of the change in the probability that we assign to the hypotheses, i.e., the change in probability prior and posterior to receiving the datum. To signal that these probabilities may be considered separate from the probability assignments over sample space, we denote priors and posteriors as $\pi(\mathbf{h_i})$ and $\pi(\mathbf{h_i} \mid \mathbf{y})$ respectively. A natural expression of the change between them is the ratio of prior and posterior.

The use of probability assignments over hypotheses means that we opt for a Bayesian notion of evidence. As is well known, Bayes' rule relates priors and posteriors as follows:

$$\frac{\pi_y(\mathbf{h_i})}{\pi(\mathbf{h_i})} = \frac{p(\mathbf{y} \mid \mathbf{h_i})}{p(\mathbf{y})},$$

[6] See Kelly (2014) for a quick presentation and some references to a discussion on the merits of this approach to evidence. Interestingly, others have argued that we can identify the meaning of a datum with the impact on our beliefs (cf. Veltman, 1996). This is suggestive of particular parallels between the concepts of evidence and meaning, but we will not delve into these here.

where $\pi$ indicates a prior belief function over hypotheses, $\pi_y$ indicates the belief function after observing data $\mathbf{y}$. In the above expression, the notion of evidence hinges entirely on the likelihoods $p(\mathbf{y}] \mid \mathbf{h_i})$ for the range of hypotheses $\mathbf{h_i}$ that are currently under consideration. In order to assess the relative evidence for two hypotheses $h_i$ and $h_j$, we may focusing on the ratio of priors and posteriors for two distinct hypotheses:

$$\frac{\pi_y(\mathbf{h_i})}{\pi_y(\mathbf{h_j})} = \frac{p(\mathbf{y} \mid \mathbf{h_i})}{p(\mathbf{y} \mid \mathbf{h_j})} \times \frac{\pi(\mathbf{h_i})}{\pi(\mathbf{h_j})}.$$

The crucial term – the one that measures the evidence – is the ratio of the probabilities of the data $\mathbf{y}$, conditional on the two hypotheses that are being compared. This ratio is known as the Bayes factor.

We can quickly see that the Bayes factor has the properties discussed in the foregoing, and that it is therefore a suitable expression of evidence. Obviously, the ratio

$$\frac{p(\mathbf{y} \mid \mathbf{h_i})}{p(\mathbf{y} \mid \mathbf{h_j})}$$

involves our beliefs concerning empirical hypotheses. More specifically, it directly involves an expression for the empirical support for the hypotheses. The support is expressed by predictive accuracy, in particular by the probability of the observed data under the hypotheses. Moreover, the evaluation is comparative, since we only look at the ratios: we express evidence as the factor between the ratio of priors and posteriors of two distinct hypotheses. The Bayes factor has all the properties we desired for an account of statistical evidence.

We now return briefly to the fact that we opted for a Bayesian account of evidence. We did so because we decided to spell out our beliefs regarding hypotheses directly, rather than focusing on our beliefs regarding the support relation. However, while our account of evidence involves probability assignments to hypotheses and is thereby typically Bayesian, the crucial expression involves probability assignments over data. It merely compares the support for the hypotheses that is offered by the datum under consideration. As a result, as long as hypotheses are not composite, our account of evidence can also be adopted by other statistical methodologies, certainly those that focus on our beliefs regarding support itself (e.g., Royall, 1997). Having said that, our own preference for a Bayesian notion of evidence should at this point be clear.

*1.6. The subjectivity of evidence*

Our notion of evidence hinges on the theoretical context: if we consider different hypotheses, our evidence changes as well. This points to a subjective element in evidence that affects statistical analyses in general.

An illustration from statistics may help to clarify this point, and put it in perspective. It is well-known that statistical procedures depend on modeling assumptions made at the outset. Hence, from one perspective, every statistical procedure is liable to model misspecification (Box, 1979). For instance, if we obtain observations that have a particular order structure but analyze those observations using a model of Bernoulli hypotheses, the order structure will simply go unnoticed. The data still present evidence for the hypotheses under consideration, but they do not provide evidence for an order structure, because there is no statistical context for identifying this order structure.

It may be argued that the context-sensitivity of evidence is more pronounced in Bayesian statistics, because a Bayesian inference is closed-minded about which hypotheses can be true: after the prior has been chosen, hypotheses with zero probability cannot enter the theory (cf. Dawid, 1982). As recently argued in Gelman and Shalizi (2013), classical statistical procedures are more open-minded in this respect: the theoretical context is not as fixed. For this reason, the context-sensitivity of evidence may seem a more pressing issue for Bayesians. However, as argued in Hacking (1965); Good (1988) among others, classical statistical procedures have a context-sensitivity of their own. It is well known that some classical procedures violate the likelihood principle. Roughly speaking, these procedures do not only depend on the actual data but also on data that, according to the hypotheses, could have been collected, but was not. The nature of this context sensitivity is different from the one that applies to Bayesian statistics, but it amounts to context sensitivity all the same.

The contextual and hence subjective character of evidence may raise some eyebrows. It might seem that the evidence that is presented by the data should not be in the eye of the beholder. We believe, however, that dependence on context is natural. To our mind, the context-sensitivity of evidence is an apt expression of the widely held view that empirical facts do not come wrapped in their appropriate interpretation. The same empirical facts will not have the same interpretation to all people in all situations, in all times. We ourselves play a crucial part in this interpretation, by framing the empirical facts in a theoretical context. This formative role for theory echoes ideas from the philosophy of science that trace back to Popper (1959) and Kuhn (1962).

## 2. Bayesian statistics: formalized statistical evidence

For previous section lays out a general way of approaching the relationship between evidence and rational belief change. The applications of such principles is broadly applicable to economic, legal, medical, and scientific reasoning. In some applications the principle concern is drawing inferences from quantitative data. *Bayesian statistics* is the application of the concepts of evidence and rational belief change to statistical scenarios.

Bayesian statistics is built atop two ideas: first, that the plausibility we assign to a hypothesis can be represented as a number between 0 and 1; and second, that Bayesian conditioning provides the rule by which we use the data to update beliefs. Let $\mathbf{y}$ be the data, $\boldsymbol{\theta}$ be a vector of parameters that characterizes the hypothesis, or the statistical model, $\mathbf{h}$ of the foregoing, and let $p(\mathbf{y} \mid \boldsymbol{\theta}))$ be the sampling distribution of the data given $\boldsymbol{\theta}$: that is, the statistical model for the data. Then Bayes conditioning implies that

$$\pi_{\mathbf{y}}(\boldsymbol{\theta}) = p(\boldsymbol{\theta} \mid \mathbf{y}) = \frac{p(\mathbf{y} \mid \boldsymbol{\theta})}{p(\mathbf{y})}\pi(\boldsymbol{\theta}).$$

This is Bayes' rule. A simple algebraic step yields the above variant, which we reproduce here:

$$\frac{\pi_{\mathbf{y}}(\boldsymbol{\theta})}{\pi(\boldsymbol{\theta})} = \frac{p(\mathbf{y} \mid \boldsymbol{\theta})}{p(\mathbf{y})}. \tag{1}$$

The left-hand side is a ratio indicating the change in belief for a specific $\theta$ due to seeing the data $\mathbf{y}$: that is, the weight of evidence. The right-hand side is the ratio of two predictions: the numerator is the predicted probability of the data $\mathbf{y}$ for $\boldsymbol{\theta}$, and the denominator is the average predicted probability of the data over all $\boldsymbol{\theta}$. Comparison of Eq. (1) with Eq. (1.5) shows that Eq. (1) reveals its link with the evidence. The evidence favors an explanation – in this case, a model with specific $\boldsymbol{\theta}$ – in proportion to how successfully it has predicted the observed data.

For convenience we denote evidence ratio

$$B(\boldsymbol{\theta}, \pi, \mathbf{y}) = \frac{p(\mathbf{y} \mid \boldsymbol{\theta})}{p(\mathbf{y})}.$$

as a function of $\boldsymbol{\theta}$, the prior beliefs $\pi$, and the data $\mathbf{y}$ that determines how beliefs should change across the values of $\boldsymbol{\theta}$, for any observed $\mathbf{y}$. As above, we use bold notation to indicate that the data, parameters, or both could be vectors. We should note that evidence ratio $B$ is not what is commonly referred to as a Bayes factor because it is a function of parameter values, $\boldsymbol{\theta}$. The connection between $B$ and Bayes factors is straightforward and will become apparent below.

To make our discussion more concrete, suppose we were interested in the probability of buttered toast falling butter-side down. Murphy's Law – which states that "anything that can go wrong will go wrong" – has been taken to imply that the buttered toast will tend to land buttered-side down (Matthews, 1995), rendering it inedible and soiling the floor[7]. We begin by assuming that toast flips have the same probability of landing butter-side down, and that the flips are independent, and thus the number of butter-down flips $y$ has a binomial distribution. There is some probability $\theta$ that represents the probability that the toast lands butter down. Figure 1 shows a possible distribution of beliefs, $\pi(\theta)$, about $\theta$; the distribution is unimodal and symmetric around $1/2$. Beliefs about $\theta$ are concentrated in the middle of the range, discounting the extreme probabilities. The choice of prior is a critical issue in Bayesian statistics; we use this prior for the sake of demonstration and defer discussion of choosing a prior.

In Bayesian statistics, most attention is centered on distributions of parameters, either before observing data (prior) or after observing data (posterior). We often speak loosely of these distributions as containing the knowledge we've gained from the data. However, it is important to remember that the parameter is inseparable from the underlying statistical model that links the parameter with the observable data, $p(\mathbf{y} \mid \boldsymbol{\theta})$. Jointly, the parameter and the data make predictions about future data. The parameters specify particular chances, or else they specify our expectations about future observations, and thereby they make precise a statistical hypothesis, i.e., a particular representation. As we argued above, an inference regarding a hypothesis should center on the degree to which a proposed constraint is successful in its predictions. With this in mind, we examine the ratio $B$ – a ratio of predictions

[7]There is ongoing debate over whether the toast could be eaten if left on the floor for less than five seconds (Dawson et al., 2007). We assume none of the readers of this article would consider such a thing.

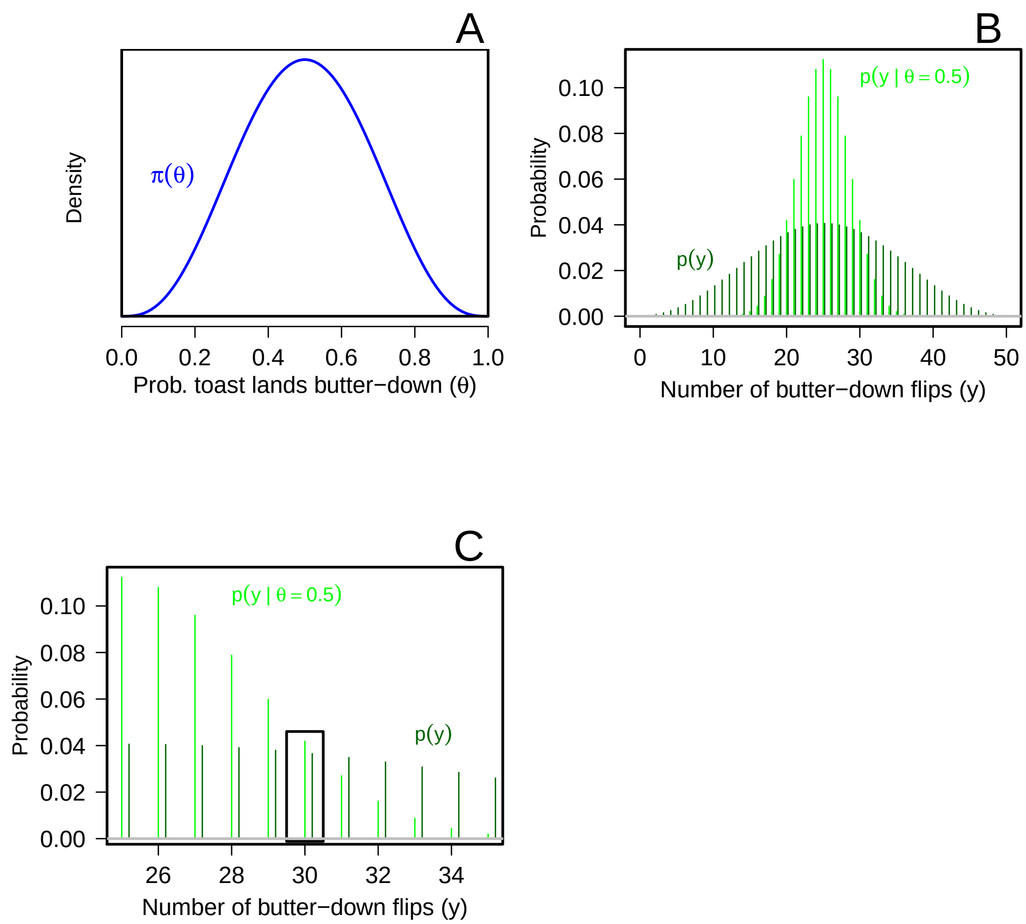



Figure 1: A: A prior distribution over the possible values $\theta$, the probability that toast lands butter-side down. B, C: Probability of outcomes under two models.

for data – in detail.

The function $B$ is a ratio of two probability functions. In the numerator is the probability of data $y$ given some specific value of $\theta$: that is, the numerator is a set of predictions for a specific model of the data. We can understand this as proposal: what predictions does this particular constraint make, and how successful are these predictions? For demonstration, we focus the specific $\theta = 0.5$. The light colored histogram in Figure 1B, labelled $p(y \mid \theta = 0.5)$, shows the predictions for the outcomes $y$ given $\theta = 0.5$, as derived from the binomial$(50, 0.5)$ probability mass function:

$$p(y \mid \theta = 0.5) = \binom{50}{y} 0.5^y (1 - 0.5)^{50-y}.$$

These predictions are centered around 25 butter-side down flips, as would be expected given that $\theta = 0.5$ and $N = 50$.

The denominator of the ratio $B$ is another set of predictions for the data: not for a specific $\theta$, but averaged over all $\theta$.

$$p(y) = \int_0^1 p(y \mid \theta)\pi(\theta)\, d\theta$$

The predictions $p(y)$ are called the *marginal* predictions, shown as the dark histogram in Figure 1B. These marginal predictions are necessarily more spread out than those of $\theta = 0.5$, because they do not commit to a specific $\theta$. Instead, they use the uncertainty in $\theta$ along with the binomial model to arrive at these marginal predictions. The spread of the predictions thus reflects all of the uncertainty about $\theta$ contained in the prior $\pi(\theta)$. The the marginal probability of the observed data – that is, when $y$ and $p(y)$ have a specific values – is called the marginal likelihood.

The ratio $B$ is thus the ratio of two competing models' predictions for the data. The numerator contains the predictions of the model where the parameter $\theta$ is constrained to a specific value, and the denominator contains the predictions of the full model, with all uncertainty from $\pi(\theta)$ included. For notational convenience, we call the restricted numerator model $\mathcal{M}_0$ and the full, denominator model $\mathcal{M}_1$. In statistics, models play the role of the hypotheses $\mathbf{h_i}$ discussed in the previous section.

Suppose we assign a research assistant to review hundreds of hours of security camera footage at a popular breakfast restaurant, she finds $N = 50$ instances where the toast fell onto the floor; in $y = 30$ of these instances, the toast landed butter down. We wish to assess the evidence in the data; or, put another way, we wish to assess how the data should transform $\pi(\theta)$ into a new belief based on $y$, $\pi_y(\theta)$. Eq. (1) tells us that the weight of evidence favoring the model $\mathcal{M}_0$ is precisely the degree to which it predicted $y = 30$ better than the full model, $\mathcal{M}_1$. Figure 1C (inside the rectangle) shows the probability of $y = 30$ under $\mathcal{M}_0$ and $\mathcal{M}_1$. Thus,

$$B = \frac{p(y = 30 \mid \theta = 0.5)}{p(y = 30)} = \frac{0.042}{0.037} = 1.145.$$

The plausibility of $\theta = 0.5$ has grown by about 15%, because the observation $y = 30$ was 15% more probable under $\mathcal{M}_0$ than $\mathcal{M}_1$.[8]

We can compute the factor $B$ for every value of $\theta$. The curve in Figure 2A the probability that $y = 30$ data under every point restriction of $\theta$; the horizontal line shows the marginal probability $p(y = 30)$. For each $\theta$, the height of the curve relative to the constant $p(y)$ gives the factor by which beliefs are updated in favor of that value of $\theta$. Where the curve is above the horizontal line (the shaded region), the value of the $\theta$ is more plausible, after observing the data; outside the shaded region, plausibility decreases. Figure 2B shows how all of these factors stretch the prior, making some regions higher and some regions lower. The effect is to transform the prior belief function $\pi(\theta)$ into a new belief function $\pi_y(\theta)$ which has been updated to reflect the observation $y$.

The prior and posterior are both shown in Figure 2C. Instead of being centered around $\theta = 0.5$, the new updated beliefs have been shifted consistent with the data proportion $y/N = 0.6$, and have smaller variance, showing the gain in knowledge from the sample size $N = 50$. Although simplistic, the example shows that the core feature of Bayesian statistics is that beliefs – modeled using probability – are driven by evidence weighed proportional to predictive success, as required by Bayes' theorem.

### *2.1. The Bayes factor*

Suppose that while your research assistant was collecting the data, you and several colleagues were brainstorming about possible outcomes. You

[8] We loosely speak of the plausibility of $\theta$ here but strictly speaking, because $\theta$ is continuous and $\pi(\theta)$ is a density function, we are referring to the collective plausibility of values in an arbitrarily small region around $\theta$.

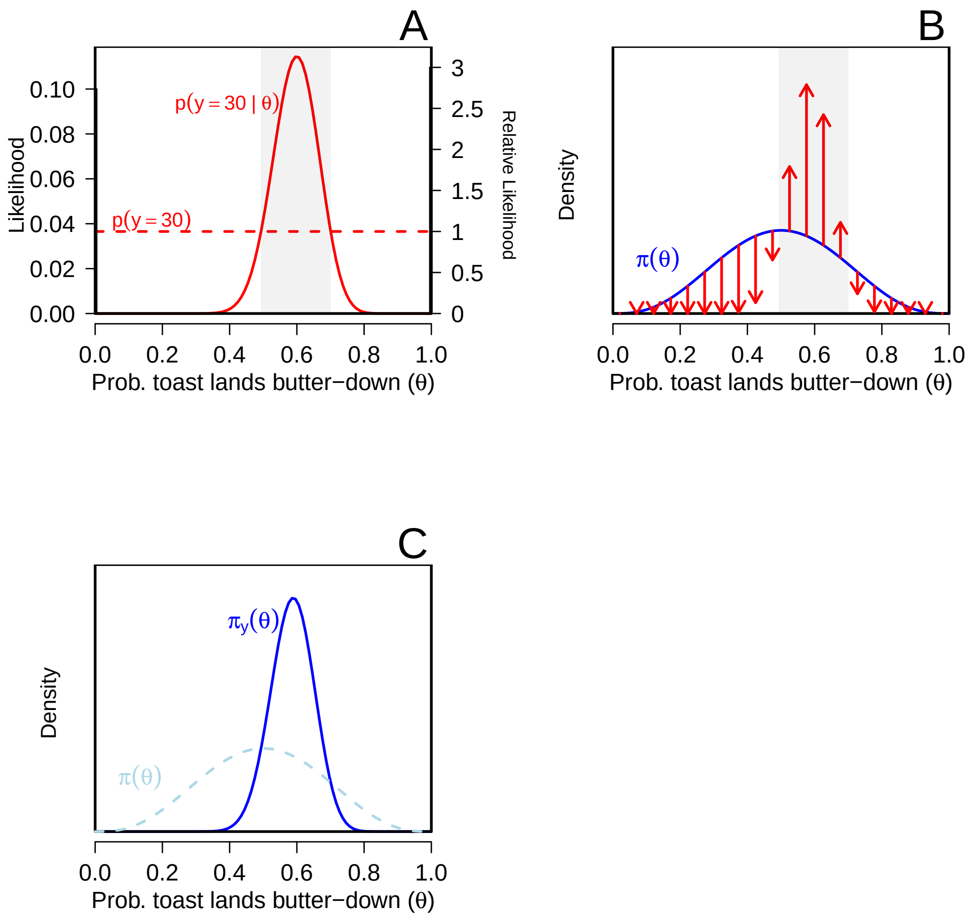



Figure 2: A: Likelihood function of $\theta$ given the observed data. Horizontal line shows the average, or marginal, likelihood. B: The transformation of the prior into the posterior through weighting by the likelihood. C: The prior and posterior. The shaded region in A and B shows the values of $\theta$ for which the evidence is positive.

assert that if Murphy's law is true, then $\theta > .5$; that is, anytime the toast falls, odds are that it will land butter-side down. A colleague points out, however, that the goal of the data collection is to assess Murphy's law. Murphy's law itself suggests that if Murphy's law is true, your attempt to test Murphy's law will fail. She claims that for the trials assessed by your research assistant, Murphy's law entails that $\theta < .5$. A second colleague thinks that the toast is probability biased, does not specify a direction of bias: that is, $\theta$ is could be any probability between 0 and 1. A third colleague thinks believes that $\theta = .5$: that is, the butter does not bias the toast at all.

You would like to assess the evidence for each of these hypotheses when your research assistant sends you the data. Because evidence is directly proportional to degree to which the observed outcomes were predicted, we need to posit predictions for each of the hypotheses. The predictions for $\theta = .5$ are the exactly those of $\mathcal{M}_0$, shown in Figure 1B, while the predictions of the unconstrained model are the same as those of $\mathcal{M}_1$. For $\theta < .5$ and $\theta > .5$, we must define plausible prior distributions over these ranges. For simplicity of demonstration, we assume that these prior distributions arise from restriction of the $\pi(\theta)$ in Figure 1A to the corresponding range (they each represent half of $\pi(\theta)$). We now have three models: $\mathcal{M}_0$, in which $\theta = .5$; $\mathcal{M}_+$, the "Murphy's law" hypothesis in which $\theta > .5$; and $\mathcal{M}_-$, the hypothesis in which our test of Murphy's law fails because $\theta < .5$.

Having defined each of the models in such a way that they have predictions for the outcomes, we can now outline how the evidence for each can be assessed. For any two models $\mathcal{M}_a$ and $\mathcal{M}_b$ we can define prior odds as the ratio of prior probabilities:

$$\frac{\pi(\mathcal{M}_a)}{\pi(\mathcal{M}_b)}$$

The prior odds are the degree to which one's beliefs favor the numerator model over the denominator model. If our beliefs are equivocal, the odds are 1; to the degree that the odds diverge from 1, the odds favor one model or the other. We can also define posterior odds; these are the degree to which beliefs will favor the numerator model over the denominator model after observing the data:

$$\frac{\pi_{\mathbf{y}}(\mathcal{M}_a)}{\pi_{\mathbf{y}}(\mathcal{M}_b)}$$

If we are interested in the evidence, then we want to know how the prior odds must be changed by the data to become the posterior odds. We again call this ratio $B$, and an application of Bayes' rule yields

$$B(\mathcal{M}_a, \mathcal{M}_b, \mathbf{y}) = \frac{\pi_{\mathbf{y}}(\mathcal{M}_a)}{\pi_{\mathbf{y}}(\mathcal{M}_b)} \bigg/ \frac{\pi(\mathcal{M}_a)}{\pi(\mathcal{M}_b)} = \frac{p(\mathbf{y} \mid \mathcal{M}_a)}{p(\mathbf{y} \mid \mathcal{M}_b)} \tag{2}$$

Here, $B$ – the relative evidence yielded by the data for $\mathcal{M}_a$ against $\mathcal{M}_b$ – is called the Bayes factor. Importantly, Eq. (2) has the same form as Eq. (1), which showed how a posterior distribution is formed from the combination of a prior distribution and the evidence. The ratio $B$ in Eq. (1) was formed from the rival predictions of a specific value of $\boldsymbol{\theta}$ against a general model in which all possible values of $\boldsymbol{\theta}$ were weighted by a prior. Eq. (2) generalizes this to any two models which predict data through a marginal likelihood.

We can now consider the evidence for each of our four models, $\mathcal{M}_0$, $\mathcal{M}_1$, $\mathcal{M}_-$, and $\mathcal{M}_+$. In fact, we have already computed the evidence for $\mathcal{M}_0$ against $\mathcal{M}_1$. The Bayes factor in this case is precisely factor by which the density of $\theta = .5$ increased against $\mathcal{M}_1$ in the previous section: 1.145. This is not an accident, of course; a posterior distribution is simply a prior distribution that has been transformed through comparison against the "background" model $\mathcal{M}_1$. If the Bayesian account of evidence is to be consistent, the evidence for $\mathcal{M}_0$ must be the same whether we are considering it as part of a posterior distribution or not.

Figure 3A shows the marginal predictions of three models, $\mathcal{M}_0$, $\mathcal{M}_-$, and $\mathcal{M}_+$. The predictions for $\mathcal{M}_0$ are the same as they were previously. For $\mathcal{M}_-$ and $\mathcal{M}_+$, we average the probability of the data over the

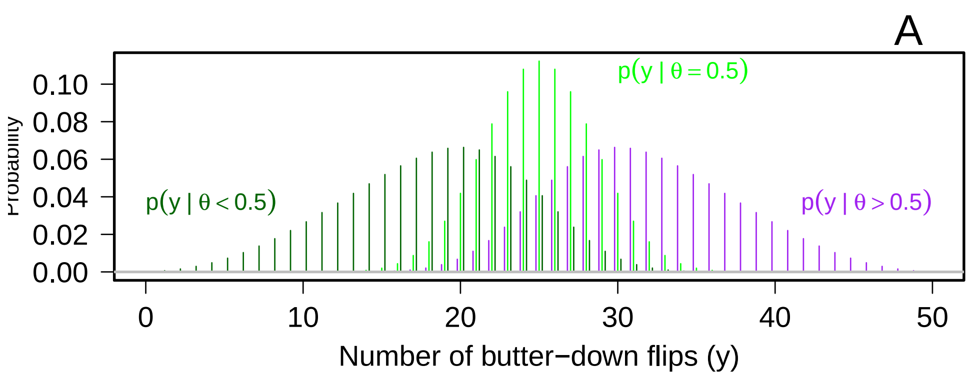
$$p(y \mid \mathcal{M}_+) = \int_{.5}^{1} p(y \mid \theta)\pi(\theta \mid \theta > .5)\, d\theta$$

and likewise for $\mathcal{M}_-$. As shown in Figure 3A, these marginal predictions are substantially more spread out than those $\mathcal{M}_0$ because they are formed from ranges of possible $\theta$ values. To assess the evidence provided by $y = 30$ we need only restrict our attention to the probability that each model assigned to the outcome. These probabilities are shown in Figure 3B.

The Bayes factor of $\mathcal{M}_+$ to $\mathcal{M}_0$ is

$$B(\mathcal{M}_+, \mathcal{M}_0, y) = \frac{p(y = 30 \mid \mathcal{M}_+)}{p(y = 30 \mid \mathcal{M}_0)} = \frac{0.066}{0.042} = 1.585,$$

The evidence favors $\mathcal{M}_+$ by a factor of 1.585 because $y = 30$ is 1.585 times as probable as $\mathcal{M}_+$ than under $\mathcal{M}_0$. Visually, this can be seen in Figure 1B



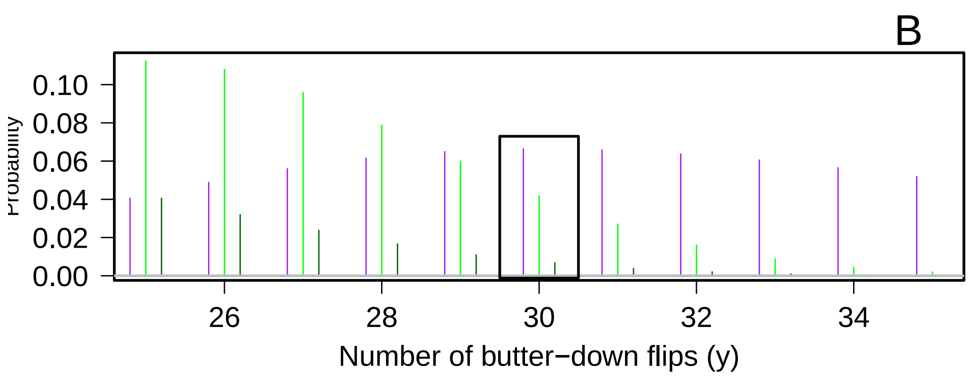




Figure 3: A: Probabilities of various outcomes under three hypotheses (see text). B: Same as A but showing only a subset of outcomes.

by the fact that the height of the bar for $\mathcal{M}_+$ is 58% higher than the one for $\mathcal{M}_0$. This Bayes factor means that to adjust for the evidence in $y = 30$, we would have to multiply our prior odds – whatever they are – by a factor of 1.585.

The Bayes factor favoring of $\mathcal{M}_+$ to $\mathcal{M}_-$ is much larger:

$$B(\mathcal{M}_+, \mathcal{M}_-, y) = \frac{p(y = 30 \mid \mathcal{M}_+)}{p(y = 30 \mid \mathcal{M}_-)} = \frac{0.066}{0.007} = 9.82,$$

indicating that the evidence favoring the "Murphy's law" hypothesis $\theta > .5$ over its complement $\theta < .5$ is much stronger than that favoring the "Murphy's law" hypothesis over the "unbiased toast" hypothesis $\theta = .5$.

Conceptually, the Bayes factor is simple: it is the ratio of the probabilities – or densities if the data are continuous – of the observed data under two models. It makes use of the same evidence that is used by Bayesian parameter estimation; in fact, Bayesian parameter estimation can be seen as a special case of Bayesian hypothesis testing, where many point alternatives are each compared to an assumed full model. Comparison of Eq. (1) and Eq (2) makes this clear.

Having defined the Bayes factor and its role in Bayesian statistics, we now move to an example that is closer to what one might encounter in research. We use this example to elucidate some of the finer philosophical points that arise from the use of the Bayes factor.

## 3. Examples

In this section, we illustrate how researchers may profitably use Bayes factors to assess the evidence for models from data using a realistic example. Consider the question of whether working memory abilities the same for men and women; that is that working memory is invariant to gender (e.g., Shibley Hyde, 2005). Although this research hypothesis can be stated in a straightforward manner, by itself this statement has no implications for the data. In order to test the hypothesis, we must instantiate the hypothesis as a statistical model. To show the statistical evidence for various theoretical positions, in the form of Bayes factors, may be compared, we first specify a general model framework. We then then instantiate competing theoretical positions as constraints within the framework.

To specify the general model framework, let $x_i$ and $y_i$ , $i = 1, \ldots, I$, be the scores for the $i$th woman and man, respectively. The modeling framework is:

$$x_i \sim \mathrm{N}(\mu + \sigma\delta/2, \sigma^2) \quad \text{and} \quad y_i \sim \mathrm{N}(\mu - \sigma\delta/2, \sigma^2), \tag{3}$$

where $\mu$ is a grand mean, $\delta$ is the standardized effect size $(\mu_x - \mu_y)/\sigma$, and $\sigma^2$ is the error variance.

The focus in this framework is $\delta$, the effect-size parameter. The theoretical position that working memory ability is invariant to gender can be instantiated within the framework by setting $\delta = 0$, shown in Figure 4A as the arrow. We denote the model as $\mathcal{M}_e$, where the $e$ is for equal abilities. With this setting, the Model $\mathcal{M}_e$ makes predictions about the data, which are best seen by considering $\hat{\delta}$, the observed effect size, $\hat{\delta} = (\bar{x} - \bar{y})/s$, where $\bar{x}$, $\bar{y}$, and $s$ are sample means and a pooled sample standard deviation, respectively. The prediction for $\hat{\delta}$ is

$$\hat{\delta}\sqrt{\frac{I}{2}} \sim T(\nu), \tag{4}$$

where $T$ is a $t$-distribution and $\nu = 2(I - 1)$ are the appropriate degrees-of-freedom for this example.[9] Predictions for sample effect size for Model $\mathcal{M}_e$ for $I = 40$ are shown in Figure 4B as the solid line. As can be seen, under the gender-invariant model of working memory performance, relatively small sample effect sizes are predicted.

Thus far, we have only specified a single model. In order to assess the evidence for $\mathcal{M}_e$, we must determine a model against which to compare. Because we have specified a general model framework, we can compare to alternative models in the same framework that do not encode the equality constraint. We consider the case of two teams of researchers, Team A and Team B who, after considerable thought, instantiate different alternatives.

Team A follows (Jeffreys, 1961) and (Rouder et al., 2009) who recommend using a Cauchy distribution to represent uncertainty about $\delta$:

$$\mathcal{M}_c: \quad \delta \sim \mathrm{Cauchy}(r),$$

[9] Prior distributions must be placed on $(\mu, \sigma^2)$. These two parameters are common across all models, and consequently the priors may be set quite broadly. We use the Jeffreys priors, $\pi(\mu, \sigma^2) \propto 1/\sigma^2$, and the predictions in (4) are derived under this choice. We note, however, that the distribution of the $t$ statistic depends only on the effect size, $\delta$, so by focusing on the $t$ statistic we make the prior assumptions for $\sigma^2$ and $\mu$ moot.

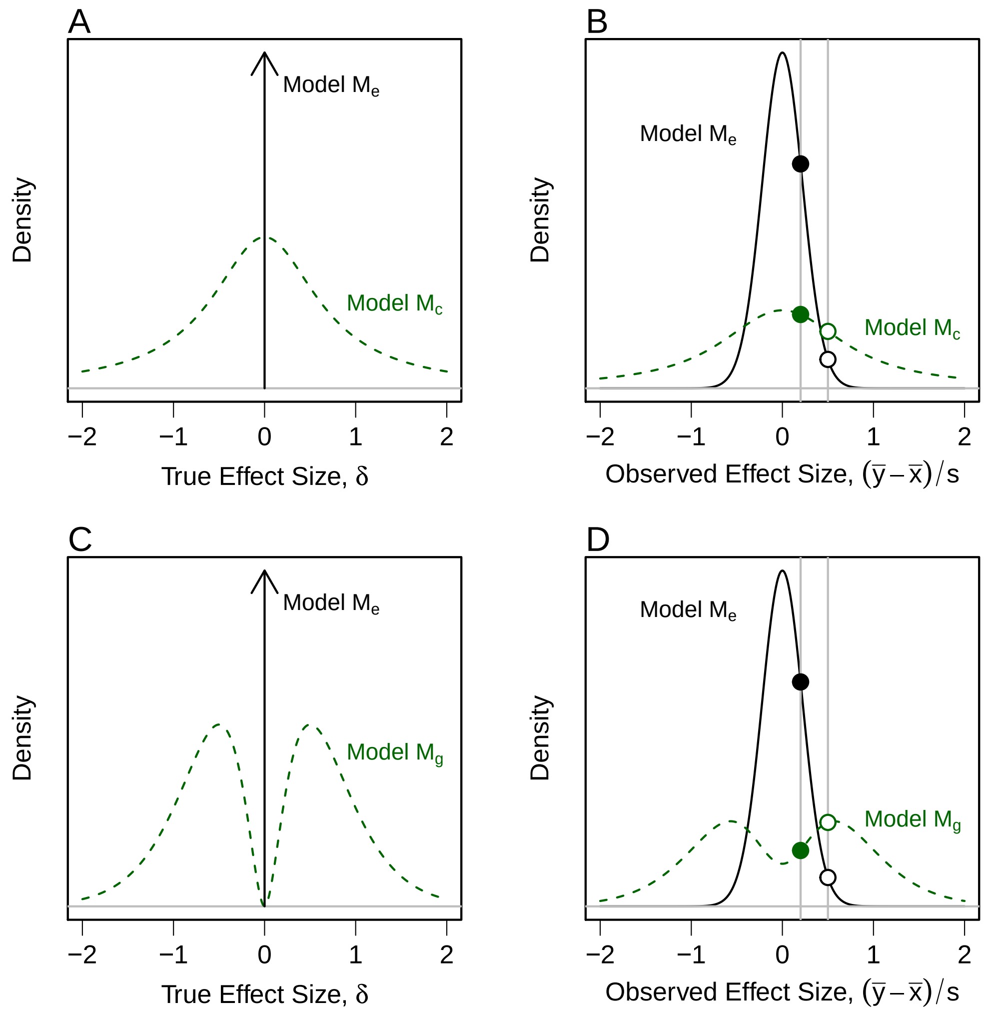



Figure 4: Models and predictions. **A.** Competing models on true effect size ($\delta$) used by Team A. **B.** Corresponding predictions for observed effect size. The filled and open points show the density values for observed effect sizes of $\hat{\delta} = .2$ and $\hat{\delta} = .5$, respectively. The ratio of these densities at an observed value is the Bayes factors, the evidence for one model relative another. **C.-D.** The models and corresponding predictions used by Team B, respectively.

where the Cauchy has a scale parameter, $r$, which describes the spread of effect sizes under the alternative.[10] The scale parameter $r$ must be set *a priori* and the team follows the recent advice of Morey and Rouder (Morey and Rouder, 2014) to set $r = \sqrt{2}/2$. With this setting the model on $\delta$, denoted $\mathcal{M}_c$ is shown in Figure 4A as the dashed line. As can be seen this model is a flexible alternative that has mass spread across small and large effects, but very large effect sizes are substantially less likely than smaller ones. The symmetry of the distribution encodes an *a priori* belief that it is as likely that women outperform men as that men outperform women. The corresponding prediction on sample effect size is shown in Figure 4B as the dashed line, and the model predicts a greater range of observed effect sizes than Model $\mathcal{M}_e$.

Team B considers a different alternative formed by representing their uncertainty about the effect size with a symmetric, but bimodal, distribution. This bimodal distribution is formed by joining gamma distributions in a back-to-back configuration as shown in Figure 4C as the dashed line. Similar bimodal priors were recommended by Johnson and Rossell (2010) and Morey and Rouder (2011). We denote this alternative as $\mathcal{M}_g$, and this alternative makes a commitment that if there are effects, they are moderate in value. [11] Compared to Team A's alternative, Team B's alternative has less mass for very large and very small magnitudes of effect size while retaining the symmetry constraint. A defense of such a prior could be that where gender effects are observed, say in mental rotation (see Matlin, 2003), they tend to be moderate in value. The corresponding prediction on sample effect size is shown in Figure 4B as the dashed line.

[10]The scaled Cauchy distribution has density

$$f(\delta) = \frac{1}{r\pi\left[1 + \left(\frac{\delta}{r}\right)^2\right]}$$

for $r > 0$.

[11]The density of the model on $\delta$ is

$$f(\delta) = \begin{cases} g(\delta, 3, 4)/2, & \delta \geq 0, \\ g(-\delta, 3, 4)/2, & \delta < 0, \end{cases}$$

where $g(\delta, \nu, \lambda)$ is the density function of a gamma distribution with shape $\nu$ and rate $\lambda$ evaluated at the value $\delta$.

It is critical to realize that neither Team A's nor Team B's choice need be considered more "correct" in their specification. Each team is interpreting the theoretical statement that men and women have different working memory capacities on average in good faith and their priors add value. In order to compute statistical evidence, choices such as these must be made. Hence, variation among priors is the reasonable and expected among analysts. It should be viewed as part of the everyday variation across researchers and research labs much as variations in experimental methods across laboratories are viewed as reasonable and expected. As with variations in experimental designs, so long as the choices made are transparent the answers will be interpretable.

Suppose the experiment resulted in an observed effect size of $\hat{\delta} = 0.2$, indicating that women somewhat outperformed men. For Team A, the predicted densities of observing $\hat{\delta}$ of 0.2 are shown as filled points in Figure 4B. The Bayes factor is the ratio of the predicted densities under $\mathcal{M}_e$ and $\mathcal{M}_c$. Because the density is 3.041 times higher under $\mathcal{M}_e$ than under $\mathcal{M}_c$, the evidence yielded by $\hat{\delta} = 0.2$ is a Bayes factor of 3.041. Team A can then state the evidence for the equality of working-memory performance by this same factor. Team B computes their Bayes factor analogously. Because the density is 4.018 times higher under $\mathcal{M}_e$ than under $\mathcal{M}_g$, the relative evidence yielded by $\hat{\delta} = 0.2$ is a Bayes factor of 4.018. Team B states evidence for the equality of working-memory performance by this factor. Although Team A and Team B reach the same conclusions, their evidence differs by a factor of 32%.

The open circles in Figure 4B show the same two analyses for a different hypothetical observed effect size, in this case $\hat{\delta} = 0.5$. The Bayes factors reached by Team A and Team B are about 2-to-1 and 3-to-1 in favor of a performance effect, and once again, these values differ.

Although it may appear problematic that two teams assessed the evidence in the same data differently, it is important to note that the two teams asked slightly different statistical questions; that is, the teams used different instantiations of the theoretically relevant statement into statistical models. Team A compared the null hypothesis $\delta = 0$ to their unimodal Cauchy prior, and Team B compared the null hypotheses to their bimodal prior. As we have argued, however, this dependence on context is a natural property of statistical evidence. Whereas the variation in modeling is expected and reasonable, so is the variation in evidence values. Data cannot impact different researchers in the same way across all contexts. We discuss this further in the next section.

## 4. Discussion

In this paper, we defined evidence in a straightforward way: the evidence presented by data is given by the change in belief that it affects. We formalized this definition and showed how it can be put to use in statistics. A Bayesian notion of evidence arises when it is assumed that "beliefs" are represented by probabilities, and that belief change is manifested by conditioning the probability of hypotheses on the data. These choices can be questioned, of course. If one wants to quantify statistical evidence in another manner, it would be necessary to flesh out other models that tie together hypothesis, data, and evaluation (e.g., fiducial statistics; Fisher, 1930).

Given the importance to scientists of quantifying statistical evidence, why have researchers not moved from frequentist techniques to other techniques more suited to their goals? There are several reasons for this. First, researchers believe, falsely, that currently popular methods serve their purposes (Gigerenzer et al., 2004; Oakes, 1986; Haller and Krauss, 2002; Hoekstra et al., ress). Second, there are several major critiques of Bayes factors that, thus far, have kept them from widespread usage. Here we outline some major critiques of Bayes factors that prevent them from being used as measures of evidence by working scientists: that Bayes factors are overly-sensitive to prior distributions, that prior distributions are too difficult to choose, and that Bayes factors depend on the true model being considered.

### *4.1. Sensitivity to prior distributions*

A number of authors have critiqued the use of Bayes factors for inference on the grounds that they are sensitive to the prior distribution chosen to represent the hypothesis (e.g., Aitkin, 1991; Liu and Aitkin, 2008; O'Hagan, 1995; Grünwald, 2000). In the example in Section 3, this was apparent: Team A and Team B chose different prior distributions over the effect size $\delta$. Each team had to decide what prior distribution best represented the alternative that women and men do have the same working memory ability on average. Although the two teams were nominally testing the same hypothesis, the Bayes factors computed by the two teams differed. This leads to the appearance that the Bayes factors are overly-dependent on the priors, which in turn causes the evidence to be arbitrary.

To some extent we defer this criticism to Bayesian statistics in general. As our development of the Bayes factor in Section 2 should make clear, the Bayes factor is neither less nor more dependent on the prior than any other Bayesian method. In fact, the transformation from prior to posterior is a special case of a Bayes factor analysis, where every point-restriction in a full model is compared to the full model itself. Any general critique of Bayes factors as a method is a critique of the foundations of Bayesian analysis itself. To avoid already well-trod ground, we refer the reader to other proponents of Bayesianism (Edwards et al., 1963; Jeffreys, 1961). In our account of evidence, we simply assume the Bayesian perspective.

It is important, however, to emphasize that the Bayes factor is not sensitive to prior distributions in all cases; the use of Bayes factors does not always require the specification of a prior distribution. Inspection of Eq. 2 reveals that the Bayes factor is solely a function of the probability of the data under the two hypotheses in question. Whenever the hypotheses are composite, these probabilities will be obtained through marginalizing over priors. But this is not the only way of obtaining predictions. It may so happen that the hypothesis, or model, under consideration does not involve any further parameters, and hence does not require any priors over the parameters (e.g., Jefferys and Berger, 1991)[12].

Even if the Bayes factors depend on the choice of a prior, a case can be made that this is as it should be. We obtain the marginal likelihoods of a model by taking an average of the likelihoods of the component hypotheses, weighted by the prior distribution. The prior distribution thus ensures that the model has a definite marginal likelihood, and thus establishes a bridge between the hypothesis and the data. Importantly, the Bayes factor is not dependent on the priors in any other way than through this marginal likelihood. Moreover, it is sensitive to the priors only insofar as the priors impact on the predictions of a model or a hypothesis. Arguably, this sensitivity of the Bayes factor to the priors is precisely what one would expect: the priors are included in the evaluation insofar as they have empirical content (see also Vanpaemel, 2010).

For users of classical significance testing, the above idea can at first be

[12]It may be thought that all modeling is accompanied by some degree of freedom but this need not be. A good example is given by statistical predictions about measurements of radioactive decay and subatomic particle spin. Predictions for these quantities can be derived from quantum mechanics, and they have unique distributions under the theory.

counter-intuitive. Consider a pair of standard classical hypotheses assuming known $\sigma$:

$$
\begin{aligned}
z &\sim \text{Normal}(\delta\sqrt{N}, 1) && (5)\\
\mathcal{H}_0 &: \delta = 0 && (6)\\
\mathcal{H}_a &: \delta \neq 0. && (7)
\end{aligned}
$$

The Bayes factor analysis cannot be run on this pair of hypotheses: one can never determine the support of this particular instantiation of $\mathcal{H}_a$, because it makes no predictions at all. In a classical significance test, by contrast, there are two possible outcomes: either we retain $\mathcal{H}_0$, or we reject it. One cannot make any positive claims about the evidence in favor of $\mathcal{H}_0$, and so the test is asymmetric, allowing only an argument for $\mathcal{H}_a$. A classical account of the evidence, in other words, is incomplete.

The use of Bayes factors requires that one instantiate hypotheses in such a way that they have constrained predictions for the data. One cannot test empty hypotheses such as "the population mean is not 100", because the marginal likelihood of such hypotheses is left indeterminate. But in order to arrive at a definite likelihood, we need a prior probability. And we believe that this is as it should be; any valid inference will hinge on the marginal data predictions, and hence on the choice of a prior. Even stronger, we believe that this prior dependence signals an important property of inference in general: evidence for or against a hypothesis should always be based on that hypothesis' empirical content – in our case: its predictions. However, because the choice of prior distributions is sometimes critical, we are required to put careful thought into this when we construct hypotheses.

### *4.2. Choosing prior distributions*

As we said, the use of Bayes factors forces the analyst to specify what the empirical content of a hypothesis is. But specifying the empirical content of a hypothesis may require substantial work. If used well, the Bayes factor rewards the analyst with an easily-interpretable measure of statistical evidence. If used badly, however, the Bayes factor is useless. Careless, automatic application of Bayes factors will inevitably lead to meaningless evidence measures that compare hypotheses not of interest to anyone. Solving the problem of careless, automatic application of Bayes factors is not trivial. For some relatively simple classes of models – e.g., linear models – it is possible to define flexible families of alternative models to compare (Liang et al., 2008; Rouder et al., 2012; Zellner and Siow, 1980).

However, for testing complex, non-nested models, the challenge of placing priors over unknown parameters is a serious impediment to the use of Bayes factors. There are several ways we might meet the challenge. One seemingly attractive way to instantiate the assumption that the values of the unknown parameters is irrelevant is to assume a so-called "non-informative" (possibly improper) prior over the parameter space. This sort of prior can be specially chosen to reflect indifference across possible values of the parameters (Bernardo, 1979; Berger and Bernardo, 1992; Jeffreys, 1961, 1946, e.g.,). However, given the development above, such a prior would be unwise. Bayes factors with improper priors have many issues stemming from the fact that the priors are not true probability distributions, and the marginal likelihood is not uniquely defined (Atkinson, 1978; Bartlett, 1957; Spiegelhalter and Smith, 1982).

Another approach to avoiding the arbitrariness of noninformative priors is to always specify "reasonable" priors. Lindley was a strong advocate of this approach. In his critique of O'Hagan's (1995), he wrote: "It is better to think about [the parameter] and what it means to the scientist. It is his prior that is needed, not the statistician's. No one who does this has an improper distribution." Although this approach is attractive in principle, in practice it can be daunting for a scientist to think of prior distributions. Some parameters can be difficult to interpret, and when there are hundreds or thousands of parameters in a statistical model, a scientist may not be able to realistically come up with priors (c.f. Goldstein, 2006; Berger, 2006, and discussion)

Another possible solution is to build a "default" prior for the parameters using the data itself. Because improper priors can yield proper posteriors given a minimal sample size, one could use a small part of the sample to compute the priors needed for the marginal likelihood to be defined for each model, then compute the Bayes factor as the ratio of the marginal likelihoods for the remaining data, given the priors built from the training data. Variations on this basic approach, called "partial Bayes factors," have been suggested by multiple authors, including Aitkin (1991); Atkinson (1978); Berger and Pericchi (1996, 1998); Spiegelhalter and Smith (1982). O'Hagan (1995) has suggested using a fraction of the likelihood itself as a prior. These approaches all attempt to circumvent, in some way, the problem of generating a reasonable prior for model comparison.

Discussion of the details of each of these statistics is outside the scope of this paper. However, we agree with the principle put forward by Berger and Pericchi (1996): "Methods that correspond to use of plausible default (proper) priors are preferable to those that do not correspond to any possible actual Bayesian analysis." Not all of the above default methods correspond to actual Bayesian analyses (see Berger and Pericchi, 1998, for discussion). The methods that correspond to a plausible default priors will have an interpretation in terms of statistical evidence for some pair of hypotheses; methods that do not correspond to any possible Bayesian analysis will not. Of course, even if a default method corresponds to a possible one must always ask whether the comparison offered by a default method is interesting.

### *4.3. Selection versus comparison, truth versus representation*

Bayes factors are often described as a model selection method; that is, one may compute the Bayes factors across a number of models, and select the model that has the highest Bayes factor as the "best" model. We have deliberately avoided discussion of model selection. In our minds, the most useful feature of the Bayes factor is its interpretation of the Bayes factor as a measure of evidence. Our view is that the concept of evidence is of paramount value. How one uses the evidence is a separate issue from the weighing of the evidence itself (see Fisher, 1955, for a similar point).

The distinction between model comparison and model selection is critically important. Selecting a model on the basis of a Bayes factor implies that one believes that the model is "good enough" in some way. However, as Gelman and Rubin (1995) point out, this cannot be argued on the basis of the Bayes factor alone. A model with the highest Bayes factor in a set of models may nonetheless fit badly. A model having the highest Bayes factor means nothing more than that the model had the highest amount of evidence in favor of it out of the models currently under consideration. However, a new model that could be considered may perform substantially better. We have stressed here and elsewhere that a model comparison perspective – as opposed to a model selection perspective – respects the fact that the evidence is always relative (Morey et al., 2013). This will not be so surprising to scientists, who are used to the tentative nature of scientific conclusions.

Finally, it has been argued the use of Bayes factors requires an implicit belief that one of the models under consideration is true (Gelman and Shalizi, 2013; Sanborn and Hills, 2014; Yu et al., 2014). Some statistical properties of Bayes factors — for instance, their convergence to the true model under regularity conditions — do depend on the "true" model model being in the set of considered models Schervish (1995). We believe, however, that in scientific practice the notion of true or false models is misguided. Statistical models are impoverished representations that attempt to capture an important aspect of a phenomenon. Although they may be used to generate propositions that can be true or false, by themselves they are not true or false. Or at least, put more carefully, their truth conditions are far from clear.

This may appear to threaten the entire enterprise of quantifying statistical evidence. After all, if models are not necessarily true or false, what does it mean to accumulate evidence for a model? We suggest that just as statistical models are proxies for real-world phenomena, statistical evidence is a proxy for real-world evidence. The applicability of the computed statistical evidence to the scientific question at hand will depend on a number of factors, including the degree to which the models compared correspond to the scientific question at hand (Morey et al., 2013). The rarefied property of statistics applies as much to statistical evidence as it does to other aspects of statistics. For instance, often statistical inferences are described as being about populations. However, the idea of a population is abstract, and a single, unique population – in the statistical sense – may not meaningfully exist. This, of course, does not not prevent the population from being a useful concept; likewise, that a model may not be true does not mean that statistical evidence for the model is not interesting. Careful consideration is required to know whether a statement of statistical evidence is useful in understanding the phenomenon of interest to the researcher.

## References

Aitkin, M. (1991). Posterior Bayes factors. *Journal of the Royal Statistical Society. Series B (Methodological)*, 53(1):111–142.

Atkinson, A. C. (1978). Posterior probabilities for choosing a regression model. *Biometrika*, 65(1):39–48.

Barrick, J. E., Yu, D. S., Yoon, S. H., Jeong, H., Oh, T. K., Schneider, D., Lenski, R. E., and Kim, J. F. (2009). Genome evolution and adaptation in a long-term experiment with Escherichia coli. *Nature*, 461(7268):1243–1247.

Bartlett, M. S. (1957). A comment on D.V. Lindley's statistical paradox. *Biometrika*, 44:533–534.

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

Berger, J. O. and Bernardo, J. M. (1992). On the development of reference priors. In *Bayesian Statistics 4. Proceedings of the Fourth Valencia International Meeting*, pages 35–49.

Berger, J. O. and Pericchi, L. R. (1996). The intrinsic Bayes factor for model selection and prediction. *Journal of the American Statistical Association*, 91(433):109–122.

Berger, J. O. and Pericchi, L. R. (1998). Accurate and stable Bayesian model selection: The median intrinsic Bayes factor. *Sankhyā: The Indian Journal of Statistics, Series B (1960-2002)*, 60(1):1–18.

Berger, J. O. and Sellke, T. (1987). Testing a point null hypothesis: The irreconcilability of p values and evidence. *Journal of the American Statistical Association*, 82(397):112–122.

Berger, J. O. and Wolpert, R. L. (1988). *The likelihood principle (2nd ed.)*. Institute of Mathematical Statistics, Hayward, CA.

Bernardo, J. M. (1979). Reference posterior distributions for Bayesian inference. *Journal of the Royal Statistical Society, Series B, Methodological*, 41:113–128.

Bird, A. (1998). *Philosophy of Science*. Routledge.

Bogen, J. and Woodward, J. (1988). Saving the phenomena. *Philosophical Review*, 97(3):303–352.

Box, G. E. P. (1979). Robustness in the strategy of scientific model building. In Launer, R. L. and Wilkinson, G. N., editors, *Robustness in Statistics: Proceedings of a Workshop*, pages 201–236. Academic Press.

Cox, R. T. (1946). Probability, frequency and reasonable expectation. *American Journal of Physics*, 14:1–13.

Dawid, P. (1982). The well-calibrated bayesian. *Journal of the American Statistical Association*, 77(379):605 – 610.

Dawson, P., Han, I., Cox, M., Black, C., and Simmons, L. (2007). Residence time and food contact time effects on transfer of Salmonella Typhimurium from tile, wood and carpet: testing the five-second rule. *Journal of Applied Microbiology*, 102(4):945–953.

de Finetti, B. (1995). The logic of probability. *Philosophical Studies*, 77:181–190.

Dunlap, R. E. (2013). Climate change skepticism and denial: An introduction. *American Behavioral Scientist*, 57(6):691–698.

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

Fisher, R. A. (1930). Inverse probability. *Proceedings of the Cambridge Philosophical Society*, 28:528–535.

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

Fox, J. (2011). Arguing about the evidence : A logical approach. In Dawid, P., Twining, W., and Vasilaski, M., editors, *Evidence, Inference and Enquiry*. The British Academy, London.

Gelman, A. and Rubin, D. B. (1995). Avoiding model selection in Bayesian social research. In Marsden, P. V., editor, *Sociological Methodology 1995*, number 165–173. Blackwell, Oxford, UK.

Gelman, A. and Shalizi, C. R. (2013). Philosophy and the practice of Bayesian statistics. *British Journal of Mathematical and Statistical Psychology*, 66:57–64.

Gigerenzer, G., Krauss, S., and Vitouch, O. (2004). The null ritual: What you always wanted to know about significance testing but were afraid to ask. In Kaplan, D., editor, *The Sage handbook of quantitative methodology for the social sciences*. Sage, Thousand Oaks, CA.

Goldstein, M. (2006). Subjective Bayesian analysis: Principles and practice. *Bayesian Analysis*, 1:403–420.

Good, I. (1988). The interface between statistics and philosophy of science. *Statistical Science*, 3(4):386–397.

Good, I. J. (1979). Studies in the History of Probability and Statistics. XXXVII A. M. Turing's Statistical Work in World War II. *Biometrika*, 66(2):393–396.

Good, I. J. (1985). Weight of evidence: A brief survey. In Bernardo, J. M., DeGroot, M. H., Lindley, D. V., and Smith, A. F. M., editors, *Bayesian Statistics 2*, pages 249–270, North-Holland. Elsevier Science Publishers B.V.

Grünwald, P. (2000). Model selection based on minimum description length. *Journal of Mathematical Psychology*, 44(1):133 – 152.

Hacking, I. (1965). *Logic of Statistical Inference*. Cambridge University Press, Cambridge, England.

Haller, H. and Krauss, S. (2002). Misinterpretations of significance: A problem students share with their teachers? *Methods of Psychological Research Online*, 7.

Halpern, J. (2003). *Reasoning about Uncertainty*. MIT press.

Hoekstra, R., Morey, R. D., Rouder, J. N., and Wagenmakers, E.-J. (in press). Robust misinterpretation of confidence intervals. *Psychonomic Bulletin and Review*.

Jefferys, W. H. and Berger, J. O. (1991). Sharpening Ockham's razor on a Bayesian strop. Technical Report.

Jeffreys, H. (1946). An invariant form for the prior probability in estimation problems. *Proceedings of the Royal Society of London. Series A, Mathematical and Physical Sciences*, 186(1007):453–461.

Jeffreys, H. (1961). *Theory of Probability (3rd Edition)*. Oxford University Press, New York.

Johnson, V. E. and Rossell, D. (2010). On the use of non-local prior desities in Bayesian hypothesis tests. *Journal of the Royal Statistical Society, Series B*, 72:143–170.

Joyce, J. M. (1998). A nonpragmatic vindication of probabilism. *Philosophy of Science*, 65:575–603.

Kahneman, D. and Tversky, A. (1972). Subjective probability: A judgment of representativeness. *Cognitive Psychology*, 3:430 – 454.

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

Kelly, T. (2014). Evidence. In Zalta, E. N., editor, *The Stanford Encyclopedia of Philosophy (Autumn 2014 Edition)*.

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

Lawrimore, J. H., Menne, M. J., Gleason, B. E., Williams, C. N., Wuertz, D. B., Vose, R. S., and Rennie, J. (2011). An overview of the Global Historical Climatology Network monthly mean temperature data set, version 3. *Journal of Geophysical Research: Atmospheres*, 116(D19):n/a–n/a.

Liang, F., Paulo, R., Molina, G., Clyde, M. A., and Berger, J. O. (2008). Mixtures of g-priors for Bayesian variable selection. *Journal of the American Statistical Association*, 103:410–423.

Liu, C. C. and Aitkin, M. (2008). Bayes factors: Prior sensitivity and model generalizability. *Journal of Mathematical Psychology*, 56:362–375.

Low, I., Lykken, J., and Shaughnessy, G. (2012). Have we observed the Higgs boson (imposter)? *Physical Review D - Particles, Fields, Gravitation and Cosmology*, 86.

Matlin, M. W. (2003). *The psychology of women*. Thompson/Wadsworth, Belmont, CA.

Matthews, R. A. J. (1995). Tumbling toast, Murphy's law, and fundamental constants. *European Journal of Physics*, 16:172–175.

Morey, R. D., Romeijn, J.-W., and Rouder, J. N. (2013). The humble Bayesian: model checking from a fully Bayesian perspective. *British Journal of Mathematical and Statistical Psychology*, 66:68–75.

Morey, R. D. and Rouder, J. N. (2011). Bayes factor approaches for testing interval null hypotheses. *Psychological Methods*, 16:406–419.

Morey, R. D. and Rouder, J. N. (2014). BayesFactor 0.9.6. Comprehensive R Archive Network.

Oakes, M. (1986). *Statistical inference: A commentary for the social and behavioral sciences*. Wiley, Chichester.

O'Hagan, A. (1995). Fractional Bayes factors for model comparison. *Journal of the Royal Statistical Society. Series B (Methodological)*, 57(1):99–138.

Popper, K. (1959). *Logic of Scientific Discovery*. Routledge Classics, London, 2002 elibrary edition.

Psillos, S. (1999). *Scientific Realism: How Science Tracks Truth*. Routledge.

Ramsey, F. P. (1931). Truth and probability. In Braithwaite, R., editor, *The Foundations of Mathematics and other Logical Essays*, pages 156–198. Harcourt, Brace and Company, New York. (1999 electronic edition).

Romeijn, J.-W. (2013). Abducted by bayesians. *Journal of Applied Logic*, 11(4):430–439.

Romeijn, J.-W. (2014). Philosophy of statistics. In Zalta, E. N., editor, *The Stanford Encyclopedia of Philosophy (Autumn 2014 Edition)*.

Rouder, J. N., Morey, R. D., Speckman, P. L., and Province, J. M. (2012). Default Bayes factors for ANOVA designs. *Journal of Mathematical Psychology*, 56:356–374.

Rouder, J. N., Speckman, P. L., Sun, D., Morey, R. D., and Iverson, G. (2009). Bayesian $t$-tests for accepting and rejecting the null hypothesis. *Psychonomic Bulletin and Review*, 16:225–237.

Royall, R. (1997). *Statistical Evidence: A Likelihood Paradigm*. CRC Press, New York.

Sanborn, A. N. and Hills, T. T. (2014). The frequentist implications of optional stopping on Bayesian hypothesis tests. *Psychonomic Bulletin & Review*, 21:283–300.

Schervish, M. J. (1995). *Theory of statistics*. Springer-Verlag, New York.

Shibley Hyde, J. (2005). The gender similarities hypothesis. *American Psychologist*, 60:581–592.

Spiegelhalter, D. J. and Smith, A. F. M. (1982). Bayes factors for linear and log-linear models with vague prior information. *Journal of the Royal Statistical Society. Series B (Methodological)*, 44(3):377–387.

Vanpaemel, W. (2010). Prior sensitivity in theory testing: An apologia for the Bayes factor. *Journal of Mathematical Psychology*, 54:491–498.

Veltman, F. (1996). Defaults in update semantics. *Journal of Philosophical Logic*, 25(3):221–261.

Wagenmakers, E.-J., Lee, M. D., Lodewyckx, T., and Iverson, G. (2008). Bayesian versus frequentist inference. In Hoijtink, H., Klugkist, I., and Boelen, P., editors, *Practical Bayesian Approaches to Testing Behavioral and Social Science Hypotheses*, pages 181–207, New York. Springer.

Yu, E. C., Sprenger, A. M., Thomas, R. P., and Dougherty, M. R. (2014). When decision heuristics and science collide. *Psychonomic Bulletin & Review*.

Zellner, A. and Siow, A. (1980). Posterior odds ratios for selected regression hypotheses. In Bernardo, J. M., DeGroot, M. H., Lindley, D. V., and Smith, A. F. M., editors, *Bayesian Statistics: Proceedings of the First International Meeting held in Valencia (Spain)*, pages 585–603. University of Valencia.