ARTICLE IN PRESS

Journal of Mathematical Psychology ▮ (▮▮▮▮) ▮▮▮–▮▮▮

Contents lists available at ScienceDirect

Journal of Mathematical Psychology

journal homepage: www.elsevier.com/locate/jmp

The philosophy of Bayes factors and the quantification of statistical evidence

Richard D. Morey [a,b,*], Jan-Willem Romeijn [a], Jeffrey N. Rouder [c]

[a] University of Groningen, Netherlands
[b] Cardiff University, United Kingdom
[c] University of Missouri, United States

HIGHLIGHTS

- We discuss the general philosophical concept of evidence, connecting it to statistics.
- We outline how statistical evidence can be quantified using the Bayes factor.
- We discuss the philosophical details and difficulties in using the Bayes factor.

ARTICLE INFO

Article history:
Available online xxxx

Keywords:
Bayes factor
Hypothesis testing

ABSTRACT

A core aspect of science is using data to assess the degree to which data provide evidence for competing 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.

A core element of science is that data are used to argue for or against hypotheses or theories. Researchers assume that data – if properly analyzed – provide evidence, whether this evidence is used to understand global climate change (Lawrimore et al., 2011), examine whether the Higgs Boson exists (Low, Lykken, & Shaughnessy, 2012), explore the evolution of bacteria (Barrick et al., 2009), or to describe human reasoning (Kahneman & Tversky, 1972). Scientists using statistics often write as if evidence is quantifiable: one can have no evidence, weaker evidence, stronger evidence—but importantly, statistics in common use do not readily admit such interpretations. The use of significance tests and confidence intervals are cases in point (Berger & Sellke, 1987; Berger & Wolpert, 1988; Jeffreys, 1961; Wagenmakers, Lee, Lodewyckx, & Iverson, 2008). Instead, these statistics are designed to make decisions, such as rejecting a hypothesis, rather than providing for a measure of evidence. Consequently, statistical practice is beset by a difference between what statistics provide and what is desired from them.

In this paper, we explore a statistical notion that does allow for the desired interpretation as a measure of evidence: the Bayes factor (Good, 1979, 1985; Jeffreys, 1961; Kass & Raftery, 1995). Our central claim is that the computation of Bayes factors is an appropriate, appealing method for assessing the impact of data on the evaluation of hypotheses. Bayes factors present a useful and meaningful measure of evidence.

To arrive at the Bayes factor, we explore the concept of evidence more generally in Section 1. We make a number of reasoned choices for an account of evidence, identify certain properties that should be reflected in our account, and then show that an account using Bayes factors fits the bill. In Section 2.1 we give a detailed introduction into Bayesian statistics and the use of Bayes factors, giving particular attention to certain conceptual issues. In Section 3 we offer some examples of the use of Bayes factors as measure of evidence, and in Section 4 we consider critiques of this use of Bayes factors and difficulties inherent in their application.

1. Evidence

What is evidence? Our answer is that the evidence presented by data is given by the impact that the data have on our evaluation of

* Corresponding author at: Cardiff University, United Kingdom.
E-mail addresses: richarddmorey@gmail.com (R.D. Morey), j.w.romeijn@rug.nl (J.-W. Romeijn), rouderj@missouri.edu (J.N. Rouder).

http://dx.doi.org/10.1016/j.jmp.2015.11.001

a theory (e.g., Fox, 2011).[1] In what follows we develop an account that ties together three central notions in this answer (theory, evaluation, and the impact of data) and then motivate the use of Bayes factors in statistics. One important caveat: our exposition falls far short of a fully worked out theory of evidence, and we do not offer a defense of Bayes factors as the only statistical measure of it. We cannot treat evidence or Bayes factors in sufficient generality and detail to warrant such wide-scope conclusions; there may well be other suitable measures, e.g., model selection tools. We argue that Bayes factors reflect the key properties of a particular conception of evidence but we do not assess the competition.

1.1. Theory: empirical hypotheses

One possible goal of scientific inquiry is instrumental: it is enough to predict and control the world by means of some scientific system, e.g., a theory or a prediction device. The format of such a system is secondary to the goal. In particular, there is no reason to expect that system will employ general hypotheses on how the world works, or that it will involve evaluations of those hypotheses. But another important goal of science is epistemic: science offers us an adequate representation of the world, or at least one that lends itself for generating explanation as well as prediction and control. For such purposes, the evaluation of hypotheses seems indispensable. Of course, a system used for prediction and control might include evaluations of hypotheses as well. Our point is that in an instrumentalist view of science an evaluative mode (e.g., an interface with beliefs) is not mandatory while in an epistemic view it is.

The idea that scientific inquiry has epistemic implications 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 opinions; 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). It seems safe to say that we cannot altogether do away with epistemic goals in science.

An epistemic goal puts particular constraints on the format of scientific theory: it will have to allow for evaluations of how believable or plausible the theory is, and it must contain components that represent nature, or the world, in some manner. We call those components hypotheses.[2] There is a large 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 predictions of data. Scientists would regard a hypothesis that has no empirical consequences as problematic. Moreover, it is a deeply seated conviction among many scientists that the success of a theory should be determined on the basis of its ability to predict the data. In short, the hypotheses must have empirical content.

The foregoing claims may seem completely trivial to our current readers. However, they are all subject to controversy in the philosophy of science. There are long-standing debates on the nature, the use and the status of scientific theory. It is far from clear that scientific hypotheses are intended to represent something, and that they always have empirical content.[3] And a closer look at science also gives us a more nuanced view. 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, certainly when these tools are seen merely as pattern-seeking devices. The message here is that scientific theory need not always have components that do representational work. However, the account of evidence that motivates Bayes factors does rely on hypotheses as representational items, and does assume that these hypotheses have empirical content.[4]

1.2. Evaluations: belief and probability

As we have argued, the epistemic goals of science lead to a particular understanding of scientific theory: it consists of empirical hypotheses that somehow represent the world. Within statistical analysis, we indeed find that theory has this character: statistical hypotheses are distributions that represent a population, and they entail probability assignments over a sample space.[5] A further consequence of taking science as an epistemic enterprise was already briefly mentioned: scientific theory must allow for evaluations, and hence interface with our epistemic attitudes. These attitudes include expectations, convictions, opinions, commitments, assumptions, and more. But for ease of reference we will simply speak of beliefs in what follows. Now that we have identified the representational components of scientific theory as hypotheses, the requirement is that these hypotheses must feature in our beliefs. And 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 on the specifics of the items featuring in it. There are many ways of representing both the beliefs and the targets of beliefs. For example, when expressing the strength of our adherence to a belief we might take them as categorical, e.g., dichotomous between accepted and rejected, or graded in some way or other. Moreover, the beliefs need not concern the hypothesis in isolation. In an account of evidence, the beliefs might just as well pertain to relations between hypotheses and data. Consequently, the involvement of beliefs does not, by itself, impose that we assign probabilities to hypotheses. And it does not entail the use of Bayesian methods to the exclusion of others either. 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).

1 Although there is a large debate within the philosophy of science about the relation between data, facts, phenomena, and the like (e.g., Bogen & Woodward, 1988), we will align ourselves with scientific practice here and simply employ the term "data" without making further discriminations. It will lead us too far afield to add further considerations.

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

3 See, e.g., Bird (1998) and Psillos (1999) for introductions into the so-called realism debate.

4 Clearly this leaves open other motivations for using Bayes factors to evaluate neural networks and the like. Moreover, data-driven techniques are often used for informal hypothesis generation. While the formal evidence evaluation techniques discussed here may not be appropriate for such exploratory techniques, they may be appropriate for later products of such techniques.

5 Notice that the theoretical structure from which the statistical hypotheses arise may be far richer than the hypotheses themselves, 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. Romeijn (2013) offers a recent discussion of this point, placing hierarchical Bayesian models in the context of explanatory reasoning in science.