Introduction to Bayesian Inference for Psychology

Alexander Etz[a] **and Joachim Vandekerckhove**[a,1]

This is a preprint of a manuscript to appear in *Psychonomic Bulletin and Review*.

We introduce the fundamental tenets of Bayesian inference, which derive from two basic laws of probability theory. We cover the interpretation of probabilities, discrete and continuous versions of Bayes' rule, parameter estimation, and model comparison. Using seven worked examples, we illustrate these principles and set up some of the technical background for the rest of this special issue of *Psychonomic Bulletin & Review*. Supplemental material is available via https://osf.io/wskex/.

Introduction | Bayesian estimation | Bayesian inference

> Dark and difficult times lie ahead. Soon we must all face the choice between what is right and what is easy.
>
> A. P. W. B. Dumbledore

1. Introduction

Bayesian methods by themselves are neither dark nor, we believe, particularly difficult. In some ways, however, they are radically different from classical statistical methods and as such, rely on a slightly different way of thinking that may appear unusual at first. Bayesian estimation of parameters will usually not result in a single estimate, but will yield a range of estimates with varying plausibilities associated with them; and Bayesian hypothesis testing will rarely result in the falsification of a theory but rather in a redistribution of probability between competing accounts.

Bayesian methods are also not new, with their first use dating back to the 18th century. Nor are they new to psychology: They were introduced to the field over 50 years ago, in what today remains a remarkably insightful exposition by Ward Edwards, Harold Lindman, and L. J. Savage (1963).

Nonetheless, until recently Bayesian methods have not been particularly mainstream in the social sciences, so the recent increase in their adoption means they are new to most practitioners – and for many psychologists, learning about new statistical techniques can evoke understandable feelings of anxiety or trepidation. At the same time, recent revelations regarding the reproducibility of psychological science (e.g., [10, 43]) have spurred interest in the statistical methods that find use in the field.

In the present article, we provide a gentle technical introduction to Bayesian inference (and set up the rest of this special issue of *Psychonomic Bulletin & Review*), starting from first principles. We will first provide a short overview involving the definition of probability, the basic laws of probability theory (the *product* and *sum* rules of probability), and how Bayes' rule and its applications emerge from these two simple laws. We will then illustrate how the laws of probability can and should be used for *inference*: to draw conclusions from observed data. We do not shy away from showing formulas and mathematical exposition, but where possible we connect them to a visual aid, either in a figure or a table, to make the concepts they represent more tangible. We also provide examples after each main section to illustrate how these ideas can be put into practice. Most of the ideas outlined in this paper only require mathematical competence at the level of college algebra; as will be seen, many of the formulas are obtained by rearranging equations in creative ways such that the quantity of interest is on the left hand side of an equality.

At any point, readers more interested in the bigger picture than the technical details can safely skip the equations and focus on the examples and discussion. However, the use of verbal explanations only suffices to gain a superficial understanding of the underlying ideas and implications, so we provide mathematical formulas for those readers who are interested in a deeper appreciation. Throughout the text, we occasionally use footnotes to provide extra notational clarification for readers who may not be as well-versed with mathematical exposition.

While we maintain that the mathematical underpinnings serve understanding of these methods in important ways, we should also point out that recent developments regarding Bayesian statistical software packages (e.g., [39, 58, 64, 65]) have made it possible to perform many kinds of Bayesian analyses without the need to carry out any of the technical mathematical derivations. The mathematical basis we present here remains, of course, more general.

First, however, we will take some time to discuss a subtle semantic confusion between two interpretations of the key concept "probability." The hurried reader may safely skip the section that follows (and advance to "The Product and Sum Rules of Probability"), knowing only that we use the word "probability" to mean "a degree of belief": a quantity that indicates how strongly we believe something to be true.

What is probability?. Throughout this text, we will be dealing with the concept of *probability*. This presents an immediate philosophical problem, because the word "probability" is in some sense ambiguous: it will occasionally switch from one meaning to another and this difference in meaning is sometimes consequential.

In one meaning—sometimes called the *epistemic*[*]—probability is a *degree of belief*: it is a number between zero and one that

---

[a] University of California, Irvine
All authors contributed to the final draft.
[1] To whom correspondence should be addressed. E-mail: joachim@uci.edu.

The authors would like to thank J. P. de Ruiter, Stephan Franke, Zita Oravecz, for helpful comments, Brian Clayton for the Illustration, and J. K. Rowling for the Harry Potter universe. The authors were supported by NSF grants #1230118 and #1534472 from the Methods, Measurements, and Statistics panel and by John Templeton Foundation grant #48192. AE was further supported by the National Science Foundation Graduate Research Fellowship Program (#DGE-1321846).

[*] From Greek *epistēmē*, meaning knowledge.

quantifies how strongly we should think something to be true based on the relevant information we have. In other words, probability is a mathematical language for expressing our uncertainty. This kind of probability is inherently subjective—because it depends on the information that *you* have available—and reasonable people may reasonably differ in the probabilities that they assign to events (or propositions). Under the epistemic interpretation, there is hence no such thing as *the* probability—there is only *your* probability (34). Your probability can be thought of as characterizing your state of incomplete knowledge, and in that sense probability does not exist beyond your mind.

We may for example say "There is a 60% probability that the United Kingdom will be outside the European Union on December 31, 2018." Someone who believes there is a 60% probability this event will occur should be willing to wager *up to* \$6 against \$4 on the event, because their expected gain would be *at least* $60\% \times (+4\$) + 40\% \times (-6\$)$, which is zero. In other words, betting more than \$6 would be unsound because they would expect to lose money, and to take such an action would not *cohere* with what they believe. Of course, in scientific practice one is rarely forced to actually make such bets, but it would be unfortunate if our probabilities (and hence our inferences) could not be acted on with confidence if such an occasion were to arise (18).

The fact that epistemic probabilities of events are subjective does not mean that they are *arbitrary*. Probabilities are not acts of will; they are subjective merely in the sense that they may differ from one individual to the next. That is just to say that different people bring different information to a given problem. Moreover, if different people update their beliefs in a rational way, then as data accumulate they will gradually approach agreement (unless they have a priori ruled out the point of agreement entirely; see, e.g., 26). In fact, it can be shown that the only way that our pre-data beliefs (whatever those may be) will cohere with our post-data beliefs is to use probability to represent our uncertainty and update our beliefs according to the laws of probability (34).

In another meaning—the *physical* or *aleatory*[†] interpretation—probability is a statement of an *expected frequency over many repetitions of a procedure*. A statement of aleatory probability might be "If I flip a fair coin very many times, the ratio of flips on which the coin will come up heads is 50%. Thus, the probability that a fair coin will come up heads is 50%." These statements express properties of the *long-run behavior* of well-defined processes, but they can not speak to singular events; they require assumptions about physical repeatability and independence among repetitions. It is important to grasp that these frequencies are seen as being a real part of the physical world, in that "the relative frequencies of a die falling this way or that way are 'persistent' and constitute this die's measurable properties, comparable to its size and weight" (42, p. 99). Neyman's quote provides an interesting contrast to the epistemic interpretation. Italian probabilist and influential Bayesian statistician Bruno de Finetti famously began his treatise *Theory of Probability* by stating "Probability does not exist" and that "the abandonment of superstitious beliefs about the existence of the Phlogiston, the Cosmic Ether, Absolute Space and Time, . . . or Fairies and Witches was an essential step along the road to scientific thinking. Probability, too, if regarded as something endowed with some kind of objective existence, is no less a misleading misconception, an illusory attempt to exteriorize or materialize our true probabilistic beliefs" (5, p. x). This is not to say that we cannot build models that assign probabilities to the outcomes of physical processes, only that they are necessarily abstractions.

It is clear that these two interpretations of probability are not the same. There are many situations to which the aleatory definition does not apply and thus probabilities could not be determined: we will not see repeated instances of December 31, 2018, in which the UK could be inside or outside the EU, we will only see one such event. Similarly, "what is the probability that *this* coin, *on the very next flip*, will come up heads?" is not something to which an aleatory probability applies: there are no long-run frequencies to consider *if there is only one flip that matters*.

Aleatory probability may—in some cases—be a valid *conceptual* interpretation of probability, but it is rarely ever an *operational* interpretation (see 21, 68, 69): it cannot apply to singular events such as the truth or falsity of a scientific theory, so we simply cannot speak of aleatory probabilities when wrestling with the uncertainty we face in scientific practice. That is to say, we may validly use aleatory probability to *think about* probability in an abstract way, but not to make statements about real-world observed events such as experimental outcomes.

In contrast, epistemic probability applies to any event that we care to consider—be it singular or repetitive—and if we have relevant information about real-world frequencies then we can choose to use that information to inform our beliefs. If repetition is possible and we find it reasonable to assume that the chance a coin comes up heads on a given toss does not change based on the outcome of previous tosses, then a Bayesian could reasonably believe both (a) that on the next toss there is a 50% chance it comes up heads; *and* (b) 50% of tosses will result in heads in a very long series of flips. Hence, epistemic probability is both a *conceptual* interpretation of probability and an *operational* interpretation. Epistemic probability can be seen as an extension of aleatory probability that applies to all the cases where the latter would apply and to countless cases where it could not.

Why this matters. We argue that the distinction above is directly relevant for empirical psychology. In the overwhelming majority of cases, psychologists are interested in making probabilistic statements about singular events: *this* theory is either true or not; *this* effect is either positive or negative; *this* effect size is probably between x and y; and either *this* model or the other is more likely given the data. Seldom are we merely interested in the frequency with which a well-defined process will achieve a certain outcome. Even arbitrarily long sequences of faithful replications of empirical studies serve to address a *singular* question: "is *this* theory correct?" We might reasonably define a certain behavioral model and assign parameters (even parameters that are probabilities) to it, and then examine its long-run behavior. This is a valid aleatory question. However, it is not an inferential procedure: it describes the behavior of an idealized model but does not provide us with inferences with regard to that model. We might also wonder how frequently a researcher will make errors of inference (however defined) under certain conditions, but this is a purely academic exercise; unless the proportion of errors is 0 or 1, such a long-run frequency alone does not allow us to determine the probability the researcher actually made an error regarding any *singular* finding – regarding *this* coin, *this* effect, or *this* hypothesis. By contrast, epistemic probability expresses degrees of belief regarding specific, individual, *singular* events, and for that reason should be the default for scientific inference.

In the next section, we will introduce the basic rules of probability theory. These rules are agnostic to our conception of probability—they hold equally for epistemic and aleatory

---

[†] From Latin *alea*, meaning dice.

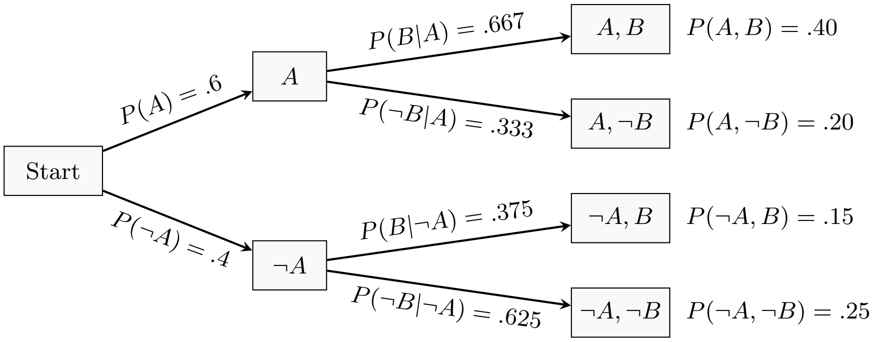

Fig. 1. An illustration of the Product Rule of probability: The probability of the joint events on the right end of the diagram is obtained by multiplying the probabilities along the path that leads to it. The paths indicate where and how we are progressively splitting the initial probability into smaller subsets. A suggested exercise to test understanding and gain familiarity with the rules is to construct the equivalent path diagram (i.e., that in which the joint probabilities are identical) starting on the left with a fork that depends on the event B instead of A.

probability—but throughout the rest of this paper and particularly in the examples, we will, unless otherwise noted, use an epistemic interpretation of the word "probability."

The Product and Sum Rules of Probability. Here we will introduce the two cardinal rules of probability theory from which essentially all of Bayesian inference derives. However, before we venture into the laws of probability, there are notational conventions to draw. First, we will use $P(A)$ to denote the probability of some event A, where A is a statement that can be true or false (e.g., A could be "it will rain today", "the UK will be outside the EU on December 31, 2018", or "the 20th digit of π is 3"). Next, we will use $(B|A)$ to denote the *conditional* event: the probability that B is true *given that A is true* (e.g., B could be "it will rain tomorrow") is $P(B|A)$: the probability that it will rain tomorrow given that it rained today. Third, we will use (A, B) to denote a *joint* event: the probability that A and B are both true is $P(A, B)$. The joint probability $P(A, B)$ is of course equal to that of the joint probability $P(B, A)$: the event "it rains tomorrow and today" is logically the same as "it rains today and tomorrow." Finally, we will use $(\neg A)$ to refer to the negation of A: the probability A is false is $P(\neg A)$. These notations can be combined: if C and D represent the events "it is hurricane season" and "it rained yesterday," respectively, then $P(A, B|\neg C, \neg D)$ is the probability that it rains today and tomorrow, given that $(\neg C)$ it is not hurricane season and that $(\neg D)$ it did not rain yesterday (i.e., both C and D are not true).

With this notation in mind, we introduce the **Product Rule of Probability**:

$$\begin{aligned} P(A, B) &= P(B)P(A|B) \\ &= P(A)P(B|A). \end{aligned} \qquad [1]$$

In words: the probability that A and B are both true is equal to the probability of B multiplied by the conditional probability of A *assuming B is true*. Due to symmetry, this is also equal to the probability of A multiplied by the conditional probability of B *assuming A is true*. The probability it rains today and tomorrow is the probability it first rains today multiplied by the probability it rains tomorrow *given that we know it rained today*.

If we assume A and B are statistically independent then $P(B)$ equals $P(B|A)$, since knowing A happens tells us nothing about the chance B happens. In such cases, the product rule simplifies as follows:

$$P(A, B) = P(A)P(B|A) = P(A)P(B). \qquad [2]$$

Keeping with our example, this would mean calculating the probability it rains both today and tomorrow in such a way that knowledge of whether or not it rained today has no bearing on how strongly we should believe it will rain tomorrow.

Understanding the **Sum Rule of Probability** requires one further concept: the *disjoint set*. A disjoint set is nothing more than a collection of mutually exclusive events. To simplify the exposition, we will also assume that exactly one of these events must be true although that is not part of the common definition of such a set. The simplest example of a disjoint set is some event and its denial:[‡] $\{B, \neg B\}$. If B represents the event "It will rain tomorrow," then $\neg B$ represents the event "It will not rain tomorrow." One and only one of these events must occur, so together they form a disjoint set. If A represents the event "It will rain today," and $\neg A$ represents "It will not rain today" (another disjoint set), then there are four possible pairs of these events, one of which must be true: (A, B), $(A, \neg B)$, $(\neg A, B)$, and $(\neg A, \neg B)$. The probability of a single one of the singular events, say B, can be found by adding up the probabilities of all of the joint events that contain B as follows:

$$P(B) = P(A, B) + P(\neg A, B).$$

In words, the probability that it rains tomorrow is the sum of two joint probabilities: (1) the probability it rains today and tomorrow, and (2) the probability it does not rain today but does rain tomorrow.

In general, if $\{A_1, A_2, \ldots, A_K\}$ is a disjoint set, the Sum Rule of Probability states:

$$\begin{aligned} P(B) &= P(A_1, B) + P(A_2, B) + \ldots + P(A_K, B) \\ &= \sum_{k=1}^{K} P(A_K, B). \end{aligned} \qquad [3]$$

That is, to find the probability of event B alone you add up all the joint probabilities that involve both B and one element of a disjoint set. Intuitively, it is clear that if one of $\{A_1, A_2, \ldots, A_K\}$ *must* be true, then the probability that *one of these and* B is true is equal to the base probability that B is true.

In the context of empirical data collection, the disjoint set of possible outcomes is often called the *sample space*.

An illustration of the Product Rule of Probability is shown by the path diagram in Figure 1. Every fork indicates the start of a

---

[‡] We use curly braces $\{\ldots\}$ to indicate a set of events. Other common examples of disjoint sets are the possible outcomes of a coin flip: {heads, tails}, or the possible outcomes of a roll of a six-sided die: $\{1, 2, 3, 4, 5, 6\}$. A particularly useful example is the truth of some model $\mathcal{M}$, which must be either true or false: $\{\mathcal{M}, \neg\mathcal{M}\}$.

Table 1. The event A is that it rains today. The event B is that it rains tomorrow. Sum across rows to find $P(A)$, sum down columns to find $P(B)$. One can also divide $P(A,B)$ by $P(A)$ to find $P(B|A)$, as shown in the next section.

	B	$\neg B$		B or $\neg B$
A	$P(A,B) = .40$	$P(A,\neg B) = .20$	$\Rightarrow$	$P(A) = .60$
$\neg A$	$P(\neg A,B) = .15$	$P(\neg A,\neg B) = .25$	$\Rightarrow$	$P(\neg A) = .40$
A or $\neg A$	$P(B) = .55$	$P(\neg B) = .45$		1.00

disjoint set, with each of the elements of that set represented by the branches extending out. The lines indicate the probability of selecting each element from within the set. Starting from the left, one can trace this diagram to find the joint probability of, say, A and B. At the *Start* fork there is a probability of .6 of going along the top arrow to event A (a similar diagram could of course be drawn that starts with B): The probability it rains today is .6. Then there is a probability of .667 of going along the next top fork to event (A,B): The probability it rains tomorrow given it rained today is .667. Hence, of the initial .6 probability assigned to A, two-thirds of it forks into (A,B), so the probability of (A,B) is $.6 \times .667 = .40$: Given that it rained today, the probability it rains tomorrow is .667, so the probability it rains both today *and* tomorrow is .4. The probability of any joint event at the end of a path can be found by multiplying the probabilities of all the forks it takes to get there.

An illustration of the Sum Rule of Probability is shown in Table 1, which tabulates the probabilities of all the joint events found through Figure 1 in the main cells. For example, adding up all of the joint probabilities across the row denoted A gives $P(A)$. Adding up all of the joint probabilities down the column denoted B gives $P(B)$. This can also be seen by noting that in Figure 1, the probabilities of the two child forks leaving from A, namely (A,B) and $(A,\neg B)$, add up to the probability indicated in the initial fork leading to A. This is true for any value of $P(B|A)$ (and $P(\neg B|A) = 1 - P(B|A)$).

2. What is Bayesian inference?

> Together [the Sum and Product Rules] solve the problem of inference, or, better, they provide a framework for its solution.
>
> D. V. Lindley (34)

Bayesian inference is the application of the product and sum rules to real problems of inference. Applications of Bayesian inference are creative ways of looking at a problem through the lens of these two rules. The rules form the basis of a mature philosophy of scientific learning proposed by Dorothy Wrinch and Sir Harold Jeffreys ((24, 25, 70); see also (36)). Together, the two rules allow us to calculate probabilities and perform scientific inference in an incredible variety of circumstances. We begin by illustrating one combination of the two rules that is especially useful for scientific inference: Bayesian hypothesis testing.

Bayes' Rule. Call event $\mathcal{M}$ (the truth of) an hypothesis that a researcher holds and call $\neg\mathcal{M}$ a competing hypothesis. Together these can form a disjoint set: $\{\mathcal{M},\neg\mathcal{M}\}$. The set $\{\mathcal{M},\neg\mathcal{M}\}$ is necessarily disjoint if $\neg\mathcal{M}$ is simply the denial of $\mathcal{M}$, but in practice the set of hypotheses can contain any number of models spanning a wide range of theoretical accounts. In such a scenario, it is important to keep in mind that we cannot make inferential statements about any model not included in the set.

Before any data are collected, the researcher has some level of prior belief in these competing hypotheses, which manifest as *prior probabilities* and are denoted $P(\mathcal{M})$ and $P(\neg\mathcal{M})$. The hypotheses are well-defined if they make a specific prediction about the probability of each experimental outcome X through the *likelihood functions* $P(X|\mathcal{M})$ and $P(X|\neg\mathcal{M})$. Likelihoods can be thought of as how strongly the data X are implied by an hypothesis. *Conditional* on the truth of an hypothesis, likelihood functions specify the probability of a given outcome and are usually easiest to interpret in relation to other hypotheses' likelihoods. Of interest, of course, is the probability that $\mathcal{M}$ is true, given the data X, or $P(\mathcal{M}|X)$.

By simple rearrangement of the factors of the Product Rule shown in the first line of Equation 1, $P(\mathcal{M},X) = P(X)P(\mathcal{M}|X)$, we can derive that

$$P(\mathcal{M}|X) = \frac{P(\mathcal{M},X)}{P(X)}.$$

Due to the symmetric nature of the Product Rule, we can reformulate the joint event in the numerator above by applying the product rule again as in the second line in Equation 1, $P(\mathcal{M},X) = P(\mathcal{M})P(X|\mathcal{M})$, and we see that this is equivalent to

$$P(\mathcal{M}|X) = \frac{P(\mathcal{M})P(X|\mathcal{M})}{P(X)}. \qquad [4]$$

Equation 4 is one common formulation of **Bayes' Rule**, and analogous versions can be written for each of the other competing hypotheses; for example, Bayes' Rule for $\neg\mathcal{M}$ is

$$P(\neg\mathcal{M}|X) = \frac{P(\neg\mathcal{M})P(X|\neg\mathcal{M})}{P(X)}.$$

The probability of an hypothesis given the data is equal to the probability of the hypothesis before seeing the data, multiplied by the probability that the data occur if that hypothesis is true, divided by the prior predictive probability of the observed data (see below). In the way that $P(\mathcal{M})$ and $P(\neg\mathcal{M})$ are called prior probabilities because they capture our knowledge prior to seeing the data X, so $P(\mathcal{M}|X)$ and $P(\neg\mathcal{M}|X)$ are called the *posterior probabilities*.

The prior predictive probability $P(X)$**.** Many of the quantities in Equation 4 we know: we must have some prior probability (belief or prior information) that the hypothesis is true if we are even considering the hypothesis at all, and if the hypothesis is well-described it will attach a particular probability to the observed data. What remains is the denominator: the prior predictive probability $P(X)$—the probability of observing a given outcome in the experiment, which can be thought of as the average probability of the outcome implied by the hypotheses, weighted by the prior probability of each hypothesis. $P(X)$ can be obtained through the sum rule by adding the probabilities of the joint events $P(X,\mathcal{M})$ and $P(X,\neg\mathcal{M})$, as in Equation 3, each of which is obtained through an application of the product rule, so we obtain the following expression:

$$\begin{aligned} P(X) &= P(X,\mathcal{M}) + P(X,\neg\mathcal{M}) \\ &= P(\mathcal{M})P(X|\mathcal{M}) + P(\neg\mathcal{M})P(X|\neg\mathcal{M}), \end{aligned} \qquad [5]$$

which amounts to adding up the right-hand side numerator of Bayes' Rule for all competing hypotheses, giving a weighted-average probability of observing the outcome X.

Now that we have a way to compute $P(X)$ in Equation 5, we can plug the result into the denominator of Equation 4 as follows:

$$P(\mathcal{M}|X) = \frac{P(\mathcal{M})P(X|\mathcal{M})}{P(\mathcal{M})P(X|\mathcal{M}) + P(\neg\mathcal{M})P(X|\neg\mathcal{M})}. \qquad [6]$$

Equation 6 is for the case where we are only considering one hypothesis and its complement. More generally,

$$P(\mathcal{M}_i|X) = \frac{P(\mathcal{M}_i)P(X|\mathcal{M}_i)}{\sum_{k=1}^{K} P(\mathcal{M}_k)P(X|\mathcal{M}_k)}, \qquad [7]$$

for the case where we are considering K competing and mutually-exclusive hypotheses (i.e., hypotheses that form a disjoint set), one of which is $\mathcal{M}_i$.

Quantifying evidence. Now that we have, in one equation, factors that correspond to our knowledge before—$P(\mathcal{M})$—and after—$P(\mathcal{M}|X)$—seeing the data, we can address a slightly alternative question: *How much did we learn due to the data X?* Consider that every quantity in Equation 7 is either a prior belief in an hypothesis, or the probability that the data would occur under a certain hypothesis—all known quantities. If we divide both sides of Equation 7 by $P(\mathcal{M}_i)$,

$$\frac{P(\mathcal{M}_i|X)}{P(\mathcal{M}_i)} = \frac{P(X|\mathcal{M}_i)}{\sum_{k=1}^{K} P(\mathcal{M}_k)P(X|\mathcal{M}_k)}, \qquad [8]$$

we see that after observing outcome X, the ratio of an hypothesis's posterior probability to its prior probability is larger than 1 (i.e., its probability goes up) if the probability it attaches to the observed outcome is greater than a weighted-average of all such probabilities – averaged across all candidate hypotheses, using the respective prior probabilities as weights.

If we are concerned with only two hypotheses, a particularly interesting application of Bayes' Rule becomes possible. After collecting data we are left with the posterior probability of two hypotheses, $P(\mathcal{M}|X)$ and $P(\neg\mathcal{M}|X)$. If we form a ratio of these probabilities we can quantify our *relative belief* in one hypothesis vis-à-vis the other, or what is known as the posterior odds: $P(\mathcal{M}|X)/P(\neg\mathcal{M}|X)$. If $P(\mathcal{M}|X) = .75$ and $P(\neg\mathcal{M}|X) = .25$, the posterior odds are $.75/.25 = 3$, or $3{:}1$ ("three to one") in favor of $\mathcal{M}$ over $\neg\mathcal{M}$. Since the posterior probability of an hypothesis is equal to the fraction in the right-hand side of Equation 6, we can calculate the posterior odds as a ratio of two right-hand sides of Bayes' Rule as follows:

$$\frac{P(\mathcal{M}|X)}{P(\neg\mathcal{M}|X)} = \frac{\dfrac{P(\mathcal{M})P(X|\mathcal{M})}{P(\mathcal{M})P(X|\mathcal{M}) + P(\neg\mathcal{M})P(X|\neg\mathcal{M})}}{\dfrac{P(\neg\mathcal{M})P(X|\neg\mathcal{M})}{P(\mathcal{M})P(X|\mathcal{M}) + P(\neg\mathcal{M})P(X|\neg\mathcal{M})}},$$

which can be reduced to a simple expression (since the denominators cancel out),

$$\underbrace{\frac{P(\mathcal{M}|X)}{P(\neg\mathcal{M}|X)}}_{\text{Posterior odds}} = \underbrace{\frac{P(\mathcal{M})}{P(\neg\mathcal{M})}}_{\text{Prior odds}} \times \underbrace{\frac{P(X|\mathcal{M})}{P(X|\neg\mathcal{M})}}_{\text{Bayes factor}}. \qquad [9]$$

The final factor—the Bayes factor—can be interpreted as *the extent to which the data sway our relative belief from one hypothesis to the other*, which is determined by comparing the hypotheses' abilities to predict the observed data. If the data are more probable under $\mathcal{M}$ than under $\neg\mathcal{M}$ (i.e., if $P(X|\mathcal{M})$ is larger than $P(X|\neg\mathcal{M})$) then $\mathcal{M}$ does the better job predicting the data, and the posterior odds will favor $\mathcal{M}$ more strongly than the prior odds.

It is important to distinguish Bayes factors from posterior probabilities. Both are useful in their own role – posterior probabilities to determine our total belief after taking into account the data and to draw conclusions, and Bayes factors as a learning factor that tells us how much evidence the data have delivered. It is often the case that a Bayes factor favors $\mathcal{M}$ over $\neg\mathcal{M}$ while at the same time the posterior probability of $\neg\mathcal{M}$ remains greater than $\mathcal{M}$. As Jeffreys, in his seminal paper introducing the Bayes factor as a method of inference, explains: "If . . . the [effect] examined is one that previous considerations make unlikely to exist, then we are entitled to ask for a greater increase of the probability before we accept it," and moreover, "To raise the probability of a proposition from 0.01 to 0.1 does not make it the most likely alternative" (22, p. 221). This distinction is especially relevant to today's publishing environment, where there exists an incentive to publish counterintuitive results – whose very description as counterintuitive implies most researchers would not have expected them to be true. Consider as an extreme example Bem (1) who presented data consistent with the hypothesis that some humans can predict future random events. While Bem's data may indeed provide positive evidence for that hypothesis (50), it is staggeringly improbable a priori and the evidence in the data does not stack up to the strong priors many of us will have regarding extrasensory perception – extraordinary claims require extraordinary evidence.

Since Bayes factors quantify statistical evidence, they can serve two (closely related) purposes. First, evidence can be applied to defeat prior odds: supposing that prior to the data we believe that $\neg\mathcal{M}$ is three times more likely than $\mathcal{M}$ (i.e., the prior ratio favoring $\neg\mathcal{M}$ is 3, or its prior probability is 75%), we need a Bayes factor favoring $\mathcal{M}$ that is greater than 3 so that $\mathcal{M}$ will end up the more likely hypothesis. Second, evidence can be applied to achieve a desired level of certainty: supposing that we desire a high degree of certainty before making any practical decision (say, at least 95% certainty or a posterior ratio of at least 19) and supposing the same prior ratio as before, then we would require a Bayes factor of $19 \times 3 = 57$ to defeat the prior odds and obtain this high degree of certainty. These practical considerations (often left implicit) are formalized by utility (loss) functions in *Bayesian decision theory*. We will not go into Bayesian decision theory in depth here; introductions can be found in Lindley (32) or Winkler (68), and an advanced introduction is available in Robert (46).

In this section, we have derived Bayes' Rule as a necessary consequence of the laws of probability. The rule allows us to update our belief regarding an hypothesis in response to data. Our beliefs after taking account the data are captured in the *posterior probability*, and the amount of updating is given by the *Bayes factor*. We now move to some applied examples that illustrate how this simple rule pertains to cases of inference.

Example 1: "The happy herbologist". At Hogwarts School of Witchcraft and Wizardry,[§] professor Pomona Sprout leads the Herbology Department (see Illustration). In the Department's greenhouses, she cultivates crops of a magical plant called *green codacle* – a flowering plant that when consumed causes a witch or wizard to feel euphoric and relaxed. Professor Sybill Trelawney, the professor of Divination, is an avid user of green codacle and frequently visits Professor Sprout's laboratory to sample the latest harvest.

However, it has turned out that one in a thousand codacle plants is afflicted with a mutation that changes its effects: Consuming those rare plants causes unpleasant side effects such as paranoia, anxiety, and spontaneous levitation. In order to evaluate the quality of her crops, Professor Sprout has developed a mutation-detecting spell. The new spell has a 99% chance to accurately detect an

---

[§] With our apologies to J. K. Rowling.

Illustration. Professor Pomona Sprout is Chair of the Herbology Department at Hogwarts School of Witchcraft and Wizardry. ©Brian Clayton, used with permission.

existing mutation, but also has a 2% chance to falsely indicate that a healthy plant is a mutant. When Professor Sprout presents her results at a School colloquium, Trelawney asks two questions: What is the probability that a codacle plant is a mutant, when your spell says that it is? And what is the probability the plant is a mutant, when your spell says that it is healthy? Trelawney's interest is in knowing how much trust to put into Professor Sprout's spell.

Call the event that a specific plant is a mutant $\mathcal{M}$, and that it is healthy $\neg\mathcal{M}$. Call the event that Professor Sprout's spell diagnoses a plant as a mutant D, and that it diagnoses it healthy $\neg D$. Professor Trelawney's interest is in the probability that the plant is indeed a mutant given that it has been diagnosed as a mutant, or $P(\mathcal{M}|D)$, and the probability the plant is a mutant given it has been diagnosed healthy, or $P(\mathcal{M}|\neg D)$. Professor Trelawney, who is an accomplished statistician, has all the relevant information to apply Bayes' Rule (Equation 7 above) to find these probabilities. She knows the prior probability that a plant is a mutant is $P(\mathcal{M}) = .001$, and thus the prior probability that a plant is not a mutant is $P(\neg\mathcal{M}) = 1 - P(\mathcal{M}) = .999$. The probability of a correct mutant diagnosis given the plant is a mutant is $P(D|\mathcal{M}) = .99$, and the probability of an erroneous healthy diagnosis given the plant is a mutant is thus $P(\neg D|\mathcal{M}) = 1 - P(D|\mathcal{M}) = .01$. When the plant is healthy, the spell incorrectly diagnoses it as a mutant with probability $P(D|\neg\mathcal{M}) = .02$, and correctly diagnoses the plant as healthy with probability $P(\neg D|\neg\mathcal{M}) = 1 - P(D|\neg\mathcal{M}) = .98$.

When Professor Sprout's spell gives a mutant diagnosis, the posterior probability that the plant is really a mutant is given by Bayes' Rule:

$$P(\mathcal{M}|D) = \frac{P(\mathcal{M})P(D|\mathcal{M})}{P(\mathcal{M})P(D|\mathcal{M}) + P(\neg\mathcal{M})P(D|\neg\mathcal{M})}.$$

Professor Trelawney can now consult Figure 2 to find that the posterior probability the plant is a mutant given a mutant diagnosis is:

$$P(\mathcal{M}|D) = \frac{.001 \times .99}{.001 \times .99 + .999 \times .02} \approx .047.$$

A mutant diagnosis from Professor Sprout's spell raises the probability the plant is a mutant from $.001$ to roughly $.047$. This means that when a plant is diagnosed as a mutant, the posterior probability the plant is *not* a mutant is $P(\neg\mathcal{M}|D) \approx 1 - .047 = .953$. The low prior probability that a plant is a mutant means that, even with the spell having 99% accuracy to correctly diagnose a mutant plant as such, a plant diagnosed as a mutant is still probably safe to eat – nevertheless, Professor Trelawney will think twice.

Analogous calculations show that the posterior probability that a plant is a dangerous mutant, given it is diagnosed as healthy, is:

$$P(\mathcal{M}|\neg D) = \frac{.001 \times .01}{.001 \times .01 + .999 \times .98} \approx .000010.$$

The posterior probability that a plant is a dangerous mutant despite being diagnosed as healthy is quite small, so Trelawney can be relatively confident she is eating a healthy plant after professor Sprout's spell returns a healthy diagnosis.

A major advantage of using Bayes' Rule in this way is that it gracefully extends to more complex scenarios. Consider the perhaps disappointing value of $P(\mathcal{M}|D)$: a mutant diagnosis only raises the posterior probability to just under 5%. Suppose, however, that Trelawney knows that Professor Sprout's diagnosis (D_S) is statistically independent from the diagnosis of her talented research associate Neville Longbottom (D_L) – meaning that for any given state of nature $\mathcal{M}$ or $\neg\mathcal{M}$, Longbottom's diagnosis does not depend on Sprout's. Further suppose that both Sprout and Longbottom return the mutant diagnosis (and for simplicity we also assume Longbottom's spells are equally as accurate as Sprout's). To find the posterior probability the plant is a mutant after two independent mutant diagnoses, $P(\mathcal{M}|D_S, D_L)$, Trelawney can apply a fundamental principle in Bayesian inference: **Yesterday's posterior is today's prior** (34).

Since we take diagnosis D_S and diagnosis D_L as conditionally independent, we know that $P(D_L|\mathcal{M}, D_S) = P(D_L|\mathcal{M})$ and $P(D_L|\neg\mathcal{M}, D_S) = P(D_L|\neg\mathcal{M})$, giving

$$\begin{aligned}&P(\mathcal{M}|D_S, D_L)\\ &\quad= \frac{P(\mathcal{M}|D_S)P(D_L|\mathcal{M})}{P(\mathcal{M}|D_S)P(D_L|\mathcal{M}) + P(\neg\mathcal{M}|D_S)P(D_L|\neg\mathcal{M})}\\ &\quad= \frac{.047 \times .99}{.047 \times .99 + .953 \times .02} \approx .71,\end{aligned}$$

where the probability the plant is a mutant *prior to Longbottom's diagnosis* D_L, $P(\mathcal{M}|D_S)$, is the probability it is a mutant *posterior to Sprout's diagnosis* D_S. This illustrates the value of multiple independent sources of evidence: a plant that has twice been independently diagnosed as a mutant is quite likely to be one. A third independent diagnosis would put the posterior probability over 99%. Note that, crucially, we would have obtained precisely the same final probability of $.71$ had we updated $P(\mathcal{M})$ to $P(\mathcal{M}|D_S, D_L)$ all at once. This is easily confirmed when we consider the two diagnoses as a joint event (D_S, D_L) and use the conditional probability $P(D_S, D_L|\mathcal{M}) = P(D_S|\mathcal{M}) \times P(D_L|\mathcal{M})$ (as in Equation 2) to update $P(\mathcal{M})$ to $P(\mathcal{M}|D_S, D_L)$ in a single step.

Discussion It is instructive to consider some parallels of this (admittedly fictional) example to current practices in social science. The scenario is similar in setup to a null-hypothesis significance testing scenario in which one defines the null hypothesis $\mathcal{H}_0$ (e.g., that there is no effect of some manipulation) and its negation $\mathcal{H}_1$ (that there is an effect), and the end goal is to make a choice between two possible decisions $\{D, \neg D\}$; D means deciding to reject $\mathcal{H}_0$, and $\neg D$ means deciding not to reject $\mathcal{H}_0$. In the example above the rate at which we falsely reject the null hypothesis (i.e.,

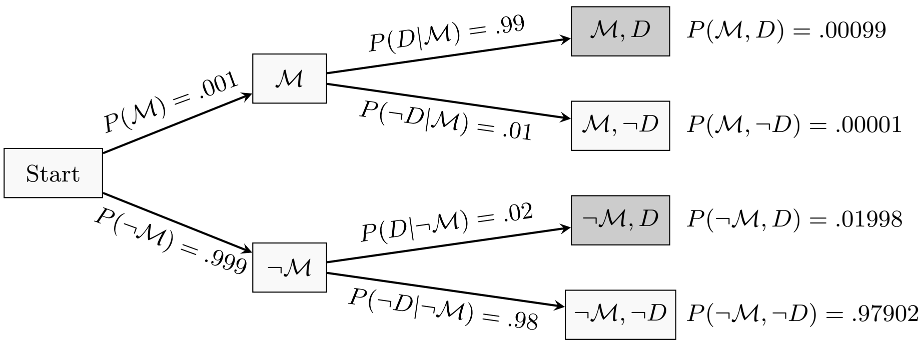
Start
$P(\mathcal{M}) = .001$
$P(\neg\mathcal{M}) = .999$
$\mathcal{M}$
$\neg\mathcal{M}$
$P(D|\mathcal{M}) = .99$
$P(\neg D|\mathcal{M}) = .01$
$P(D|\neg\mathcal{M}) = .02$
$P(\neg D|\neg\mathcal{M}) = .98$
$\mathcal{M}, D$
$\mathcal{M}, \neg D$
$\neg\mathcal{M}, D$
$\neg\mathcal{M}, \neg D$
$P(\mathcal{M}, D) = .00099$
$P(\mathcal{M}, \neg D) = .00001$
$P(\neg\mathcal{M}, D) = .01998$
$P(\neg\mathcal{M}, \neg D) = .97902$

Fig. 2. The event $\mathcal{M}$ is that a given codacle plant is a mutant. The event D is that Professor Sprout's spell returns a mutant diagnosis. A mutant diagnosis D is in fact observed, so the only paths that remain relevant are those that lead to a mutant diagnosis (events $(\mathcal{M}, D)$ and $(\neg\mathcal{M}, D)$, shaded). Professor Trelawney takes the following steps to find the posterior probability the plant is a mutant given the mutant diagnosis: Multiply $P(\mathcal{M})$ by $P(D|\mathcal{M})$ to find $P(\mathcal{M}, D)$; multiply $P(\neg\mathcal{M})$ by $P(D|\neg\mathcal{M})$ to find $P(\neg\mathcal{M}, D)$; add $P(\mathcal{M}, D)$ and $P(\neg\mathcal{M}, D)$ to find $P(D)$; divide $P(\mathcal{M}, D)$ by $P(D)$ to find $P(\mathcal{M}|D)$. Professor Trelawney's question can be rephrased as: of the total probability remaining in the diagram after D is observed – which is equal to $P(D)$ – what proportion of it originated at the $\mathcal{M}$ node? The results of Professor Trelawney's calculations are given in the text.

deciding to reject it when in fact it is true) is given by $P(D|\neg\mathcal{M}) = .02$ – this is what is commonly called the false alarm rate. The rate at which we correctly reject the null hypothesis (i.e., rejecting it if it is false) is $P(D|\mathcal{M}) = .99$. However, even with a low false alarm rate and a very high correct rejection rate, a null hypothesis rejection may not necessarily provide enough evidence to overcome the low prior probability an alternative hypothesis might have.

Example 2: "A curse on your hat". At the start of every school year, new Hogwarts students participate in the centuries-old Sorting ceremony, during which they are assigned to one of the four Houses of the School: Gryffindor, Hufflepuff, Ravenclaw, or Slytherin. The assignment is performed by the Sorting Hat, a pointy hat which, when placed on a student's head, analyzes their abilities and personality before loudly calling out the House that it determines as the best fit for the student. For hundreds of years the Sorting Hat has assigned students to houses with perfect accuracy and in perfect balance (one-quarter to each House).

Unfortunately, the Hat was damaged by a stray curse during a violent episode at the School. As a result of the dark spell, the Hat will now occasionally blurt out "Slytherin!" even when the student's proper alliance is elsewhere. Now, the Hat places exactly 40% of first-years in Slytherin instead of the usual 25%, and each of the other Houses get only 20% of the cohort.

To attempt to correct the House assignment, Professor Cuthbert Binns has developed a written test—the *Placement Accuracy Remedy for Students Erroneously Labeled* or P.A.R.S.E.L. test—on which true Slytherins will tend to score *Excellent* (S_E), while Ravenclaws will tend to score *Outstanding* (S_O), Gryffindors *Acceptable* (S_A), and Hufflepuffs *Poor* (S_P). Benchmark tests on students who were Sorted before the Hat was damaged have revealed the approximate distribution of P.A.R.S.E.L. scores within each House (see Table 2). The test is administered to all students who are sorted into Slytherin House by the damaged Sorting Hat, and their score determines the House to which they are assigned. Headmistress Minerva McGonagall, who is a Gryffindor, asks Professor Binns to determine the probability that a student who was sorted into Slytherin and scored *Excellent* on the P.A.R.S.E.L. test actually belongs in Gryffindor.

The solution relies on the repeated and judicious application of the Sum and Product Rules, until an expression appears with the desired quantity on the left hand side and only known quantities on the right hand side. To begin, Professor Binns writes down Bayes' Rule (remembering that a joint event like (D_S, S_E) can be treated like any other event):

$$P(\mathcal{M}_G|D_S, S_E) = \frac{P(\mathcal{M}_G)P(D_S, S_E|\mathcal{M}_G)}{P(D_S, S_E)}$$

Here, $\mathcal{M}_G$ means that the true House assignment is Gryffindor, D_S means that the Sorting Hat placed them in Slytherin, and S_E means the student scored *Excellent* on the P.A.R.S.E.L. test.

In most simple cases, we often have knowledge of simple probabilities, of the form $P(A)$ and $P(B|A)$, while the probabilities of joint events (A, B) are harder to obtain. For Professor Binns' problem, we can overcome this difficulty by using the Product Rule to unpack the joint event in the numerator:[¶]

$$P(\mathcal{M}_G|D_S, S_E) = \frac{P(\mathcal{M}_G)P(S_E|\mathcal{M}_G)P(D_S|S_E, \mathcal{M}_G)}{P(D_S, S_E)}.$$

Now we discover the probability $P(D_S|S_E, \mathcal{M}_G)$ in the numerator. Since the cursed hat's recommendation does not add any information about the P.A.R.S.E.L. score above and beyond the student's true House affiliation (i.e., it is *conditionally independent*; the test score is not entirely independent of the hat's recommendation since the hat is often right about the student's correct affiliation and the affiliation influences the test score), we can simplify this conditional probability: $P(D_S|S_E, \mathcal{M}_G) = P(D_S|\mathcal{M}_G)$. Note that the numerator now only contains known quantities: $P(S_E|\mathcal{M}_G)$ can be read off as 0.05 from Table 2; $P(D_S|\mathcal{M}_G)$ is the probability that a true Gryffindor is erroneously sorted into Slytherin, and since that happens to one in five true Gryffindors (because the proportion sorted into Gryffindor went down from 25% to 20%), $P(D_S|\mathcal{M}_G)$ must be 0.20; and $P(\mathcal{M}_G)$ is the base probability that a student is a Gryffindor, which we know to be one in four. Thus,

$$\begin{aligned} P(\mathcal{M}_G|D_S, S_E) &= \frac{P(\mathcal{M}_G)P(S_E|\mathcal{M}_G)P(D_S|\mathcal{M}_G)}{P(D_S, S_E)} \\ &= \frac{0.25 \times 0.05 \times 0.20}{P(D_S, S_E)}. \end{aligned}$$

This leaves us having to find $P(D_S, S_E)$, the prior predictive probability that a student would be Sorted into Slytherin and score

---

[¶] Note that this is an application of the Product Rule to the scenario where both events are conditional on $\mathcal{M}_G$: $P(D_S, S_E|\mathcal{M}_G) = P(S_E|\mathcal{M}_G)P(D_S|S_E, \mathcal{M}_G)$.

Table 2. Probability of each P.A.R.S.E.L. score by true House affiliation. Each value indicates the conditional probability $P(S|\mathcal{M})$, that is, the probability that a student from house $\mathcal{M}$ obtains score S.

	Excellent (S_E)	Outstanding (S_O)	Acceptable (S_A)	Poor (S_P)
Slytherin ($\mathcal{M}_S$)	0.80	0.10	0.05	0.05
Gryffindor ($\mathcal{M}_G$)	0.05	0.20	0.70	0.05
Ravenclaw ($\mathcal{M}_R$)	0.05	0.80	0.15	0.00
Hufflepuff ($\mathcal{M}_H$)	0.00	0.10	0.25	0.65

Excellent on the P.A.R.S.E.L. test. Here, the Sum Rule will help us out, because we can find the right hand side numerator for each type of student in the same way we did for true Gryffindors above – we can find $P(D_S, S_E|\mathcal{M}_i)$ for any House $i = S, G, R, H$. Hence (from Equation 3),

$$\begin{aligned}
P(D_S, S_E) &= \sum_i P(\mathcal{M}_i)P(S_E|\mathcal{M}_i)P(D_S|\mathcal{M}_i)\\
&= P(\mathcal{M}_S)P(S_E|\mathcal{M}_S)P(D_S|\mathcal{M}_S)\\
&\quad + P(\mathcal{M}_G)P(S_E|\mathcal{M}_G)P(D_S|\mathcal{M}_G)\\
&\quad + P(\mathcal{M}_R)P(S_E|\mathcal{M}_R)P(D_S|\mathcal{M}_R)\\
&\quad + P(\mathcal{M}_H)P(S_E|\mathcal{M}_H)P(D_S|\mathcal{M}_H)\\
&= 0.25 \times 0.80 \times 1.00\\
&\quad + 0.25 \times 0.05 \times 0.20\\
&\quad + 0.25 \times 0.05 \times 0.20\\
&\quad + 0.25 \times 0.00 \times 0.20\\
&= 0.2050.
\end{aligned}$$

So finally, we arrive at:

$$P(\mathcal{M}_G|D_S, S_E) = \frac{0.0025}{0.2050} = 0.0122,$$

which allows Professor Binns to return to the Headmistress with good news: There is only around a 1% probability that a student who is Sorted into Slytherin and scores *Excellent* on the P.A.R.S.E.L. test is actually a Gryffindor. Binns claims that the probability that such a student is a true Slytherin is over 95%, and that the combined procedure—that consists of first letting the Sorting Hat judge and then giving Slytherin-placed students a P.A.R.S.E.L. test and rehousing them by their score—will correctly place students of any House with at least 90% probability. For example, he explains, a true Ravenclaw would be sorted into their correct House by the Hat with 80% ($P(D_R|\mathcal{M}_R)$) probability, and would be placed into Slytherin with 20% probability. In the second case, the student would be given the P.A.R.S.E.L. test, in which they would obtain an *Outstanding* with 80% ($P(S_O|\mathcal{M}_R)$) probability. Hence, they would be placed in their correct House with probability $P(D_R|\mathcal{M}_R) + P(D_S|\mathcal{M}_R) \times P(S_O|\mathcal{M}_R) = 0.80 + 0.20 \times 0.80 = 0.96$.

Discussion The Sorting Hat example introduces two extensions from the first. Here, there are not two but four possible "models" – whereas statistical inference is often seen as a choice problem between two alternatives, probabilistic inference naturally extends to any number of alternative hypotheses. The extension that allows for the evaluation of multiple hypotheses did not require the ad hoc formulation of any new rules, but relied entirely on the same basic rules of probability.

The example additionally underscores an inferential facility that we believe is vastly underused in social science: we selected between models making use of two *qualitatively different* sources of information. The two sources of information were individually insufficient but jointly powerful: the Hat placement is only 80% accurate in most cases, and the written test was only 50% accurate for the Ravenclaw case, but together they are 90% accurate. Again, this extension is novel only in that we had not yet considered it – the fact that information from multiple sources can be so combined requires no new facts and is merely a consequence of the two fundamental rules of probability.

3. Probability theory in the continuous case

> In Bayesian parameter estimation, both the prior and posterior distributions represent, not any measurable property of the parameter, but only our own state of knowledge about it. The width of the [posterior] distribution... indicates the range of values that are consistent with our prior information and data, and which honesty therefore compels us to admit as possible values.
>
> E. T. Jaynes (20)

The full power of probabilistic inference will come to light when we generalize from discrete events A with probabilities $P(A)$, to continuous parameters a with probability densities $p(a)$.[‖] Probability densities are different from probabilities in many ways. Densities express how much probability exists "near" a particular value of a, while the probability of any particular value of a in a continuous range is zero. Probability densities cannot be negative but they can be larger than 1, and they translate to probabilities through the mathematical operation of integration (i.e., calculating the area under a function over a certain interval). Possibly the most well-known distribution in psychology is the theoretical distribution of IQ in the population, which is shown in Figure 3.

By definition, the total area under a probability density function is 1:

$$1 = \int_A p(a)da,$$

where capitalized A indicates that the integration is over the entire range of possible values for the parameter that appears at the end – in this case a. The range A is hence a disjoint set of possible values for a. For instance, if a is the mean of a normal distribution, A indicates the range of real numbers from $-\infty$ to ∞; if a is the rate parameter for a binomial distribution, A indicates the range of real numbers between 0 and 1. The symbol da is called the *differential* and the function that appears between the integration sign and the differential is called the *integrand* – in this case $p(a)$.

We can consider how much probability is contained within smaller sets of values within the range A; for example, when dealing with IQ in the population, we could consider the integral over only the values of a that are less than 81, which would equal

---

[‖] When we say a parameter is "continuous" we mean it could take any one of the infinite number of values comprising some continuum. For example, this would apply to values that follow a normal distribution.

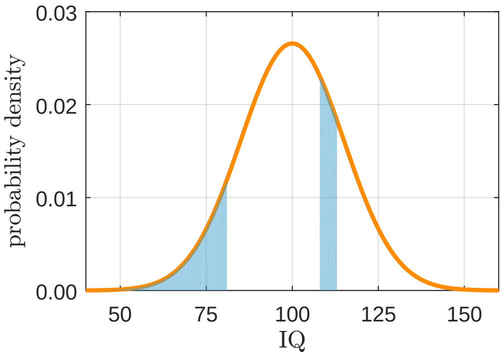

Fig. 3. An example of a probability density function (PDF). PDFs express the relative plausibility of different values and can be used to determine the probability that a value lies in any interval. The PDF shown here is the theoretical distribution of IQ in the population: a normal distribution (a.k.a. Gaussian distribution) with mean 100 and standard deviation 15. In this distribution, the filled region to the left of 81 has an area of approximately 0.10, indicating that for a random member of the population, there is a 10% chance their IQ is below 81. Similarly, the narrow shaded region on the right extends from 108 to 113 and also has an area of 0.10, meaning that a random member has a 10% probability of falling in that region.

the probability that a is less than 81:[**]

$$P(a < 81) = \int_{-\infty}^{81} p(a)da.$$

In Figure 3, the shaded area on the left indicates the probability density over the region $(-\infty, 81)$.

The fundamental rules of probability theory in the discrete case—the sum and product rules—have continuous analogues. The continuous form of the product rule is essentially the same as in the discrete case: $p(a, b) = p(a)p(b|a)$, where $p(a)$ is the density of the continuous parameter a and $p(b|a)$ denotes the *conditional density* of b (i.e., the density of b assuming a particular value of a). As in the discrete case of Equation 1, it is true that $p(a, b) = p(a)p(b|a) = p(b)p(a|b)$, and that $p(a, b) = p(a)p(b)$ if we consider a and b to be statistically independent. For the continuous sum rule, the summation in Equation 3 is replaced by an integration over the entire parameter space B:

$$p(a) = \int_B p(a, b)db.$$

Because this operation can be visualized as a function over two dimensions ($p(a, b)$ is a function that varies over a and b simultaneously) that is being collapsed into the one-dimensional margin ($p(a)$ varies only over a), this operation is alternatively called *marginalization*, *integrating over* b, or *integrating out* b.

Using these continuous forms of the sum and product rules, we can derive a continuous form of Bayes' Rule by successively applying the continuous sum and product rules to the numerator and denominator (analogously to Equation 7):

$$\begin{aligned} p(a|b) &= \frac{p(a,b)}{p(b)} = \frac{p(a)p(b|a)}{p(b)} \\ &= \frac{p(a)p(b|a)}{\int_A p(a)p(b|a)da}. \end{aligned} \qquad [10]$$

Since the product in the numerator is divided by its own integral, the total area under the posterior distribution always equals 1; this guarantees that the posterior is always a proper distribution if the prior and likelihood are proper distributions. It should be noted that by "continuous form of Bayes' Rule" we mean that the prior and posterior distributions for the model parameter(s) are continuous – the sample data can still be discrete, as in Example 3 below.

One application of Bayesian methods to continuous parameters is *estimation*. If θ (theta) is a parameter of interest (say, the success probability of a participant in a task), then information about the relative plausibility of different values of θ is given by the probability density $p(\theta)$. If new information becomes available, for example in the form of new data x, the density can be updated and made conditional on x:

$$p(\theta|x) = \frac{p(\theta)p(x|\theta)}{p(x)} = \frac{p(\theta)p(x|\theta)}{\int_\Theta p(\theta)p(x|\theta)d\theta}. \qquad [11]$$

Since in the context of scientific learning these two densities typically represent our knowledge of a parameter θ before and after taking into account the new data x, $p(\theta)$ is often called the *prior density* and $p(\theta|x)$ the *posterior density*. Obtaining the posterior density involves the evaluation of Equation 11 and requires one to define a likelihood function $p(x|\theta)$, which indicates how strongly the data x are implied by every possible value of the parameter θ.

The numerator on the right hand side of Equation 11, $p(\theta)p(x|\theta)$, is a product of the prior distribution and the likelihood function, and it completely determines the shape of the posterior distribution (note that the denominator in that equation is not a function of the parameter θ; even though the parameter seems to feature in the integrand, it is in fact "integrated out" so that the denominator depends only on the data x). For this reason, many authors prefer to ignore the denominator of Equation 11 and simply write the posterior density as proportional to the numerator, as in $p(\theta|x) \propto p(\theta)p(x|\theta)$. We do not, because this conceals the critical role the denominator plays in a predictive interpretation of Bayesian inference.

The denominator $p(x)$ is the weighted-average probability density of the data x, where the form of the prior distribution determines the weights. This normalizing constant is the continuous analogue of the prior predictive distribution, often alternatively referred to as the *marginal likelihood* or the Bayesian *evidence*.[††] Consider that, in a similar fashion to the discrete case, we can rearrange Equation 11 as follows—dividing each side by $p(\theta)$—to illuminate in an alternative way how Bayes' rule operates in updating the prior distribution $p(\theta)$ to a posterior distribution $p(\theta|x)$:

$$\frac{p(\theta|x)}{p(\theta)} = \frac{p(x|\theta)}{p(x)} = \frac{p(x|\theta)}{\int_\Theta p(\theta)p(x|\theta)d\theta}. \qquad [12]$$

On the left hand side, we see the ratio of the posterior to the prior density. Effectively, this tells us for each value of θ how much more or less plausible that value became due to seeing the data x. The equation shows that this ratio is determined by how well that specific value of θ predicted the data, in comparison to the weighted-average predictive accuracy across all values in the range Θ. In other words, **parameter values that exceed the average predictive accuracy across all values in Θ have their densities increased, while parameter values that predict**

[**] Strictly speaking, this integral is the probability that a is less than *or equal to* 81, but the probability of any single point in a continuous distribution is 0. By the sum rule, $P(a \leq 81) = P(a < 81) + P(a = 81)$, which simplifies to $P(a \leq 81) = P(a < 81) + 0$.

[††] We particularly like Evans's take on the term Bayesian *evidence*: "For evidence, as expressed by observed data in statistical problems, is what causes beliefs to change and so we can measure evidence by measuring change in belief" (12, p. 243).

worse than the average have their densities decreased (see 41, 66).

While the discrete form of Bayes' rule has natural applications in hypothesis testing, the continuous form more naturally lends itself to parameter estimation. Examples of such questions are: "What is the probability that the regression weight β is positive?" and "What is the probability that the difference between these means is between $\delta = -.3$ and $\delta = .3$?" These questions can be addressed in a straightforward way, using only the product and sum rules of probability.

Example 3: "Perfection of the puking pastille". In the secretive research and development laboratory of *Weasley's Wizarding Wheezes*, George Weasley works to develop gag toys and prank foods for the entertainment of young witches and wizards. In a recent project, Weasley is studying the effects of his store's famous *puking pastilles*, which cause immediate vomiting when consumed. The target audience is Hogwarts students who need an excuse to leave class and enjoy making terrible messes.

Shortly after the pastilles hit Weasley's store shelves, customers began to report that puking pastilles cause not one, but multiple "expulsion events." To learn more about this unknown behavior, George turns to his sister Ginny and together they decide to set up an exploratory study. From scattered customer reports, George believes the expulsion rate to be between three to five events per hour, but he intends to collect data to determine the rate more precisely. At the start of this project, George has no distinct hypotheses to compare – he is interested only in estimating the expulsion rate.

Since the data x are counts of the number of expulsion events within an interval of time, Ginny decides that the appropriate model for the data (i.e., likelihood function) is a Poisson distribution (see top panel of Figure 4):

$$p(x|\lambda) = \frac{1}{x!}\exp\left(-\lambda\right)\lambda^{x}, \qquad [13]$$

with the λ (lambda) parameter representing the expected number of events within the time interval (note $\exp(-\lambda)$ is simply a clearer way to write $e^{-\lambda}$).

A useful prior distribution for Poisson rates is the Gamma distribution (15, Appendix A):[‡‡]

$$p(\lambda|a,b) = \frac{b^{a}}{\Gamma(a)}\exp\left(-\lambda b\right)\lambda^{a-1}, \qquad [14]$$

A visual representation of the Gamma distribution is given in the second panel of Figure 4. A Gamma distribution has two parameters that determine its form, namely shape (a) and scale (b).[§§] The Gamma distribution is useful here for two reasons: first, it has the right *support*, meaning that it provides nonzero density for all possible values for the rate (in this case all positive real numbers); and second, it is *conjugate* with the Poisson distribution, a technical property to be explained below.

Before collecting further data, the Weasleys make sure to specify what they believe to be reasonable values based on the reports George has heard. In the second panel of Figure 4, Ginny set the prior parameters to $a = 2$ and $b = 0.2$ by drawing the shape of the distribution for many parameter combinations and selecting a curve that closely resembles George's prior information: Values between three and five are most likely, but the true value of the expulsion rate could conceivably be much higher.

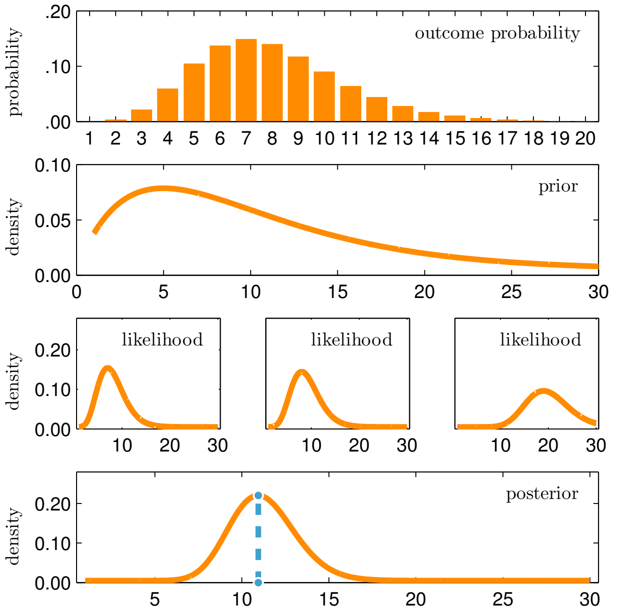

Fig. 4. Top row: An example Poisson distribution. The function is $p(x|\lambda = 7)$ as defined in Equation 13. The height of each bar indicates the probability of that particular outcome (e.g., number of expulsion events). **Second row:** The prior distribution of λ; a Gamma distribution with parameters $a = 2$ and $b = 0.2$. This is the initial state of the Weasley's knowledge of the expulsion rate λ (the expected number of expulsion events per hour). **Third row:** The likelihood functions associated with $x_1 = 7$ (left), $x_2 = 8$ (center), and $x_3 = 19$ (right). **Bottom row:** The posterior distribution of λ; a Gamma distribution with parameters $a = 36$ and $b = 3.2$. This is the final state of knowledge regarding λ.

Three volunteers are easily found, administered one puking pastille each, and monitored for one hour. The observed event frequencies are $x_1 = 7$, $x_2 = 8$, and $x_3 = 19$.

With the prior density (Equation 14) and the likelihood (Equation 13) known, Ginny can use Bayes' rule as in Equation 10 to derive the posterior distribution of λ, conditional on the new data points $X_n = (x_1, x_2, x_3)$. She will assume the $n = 3$ data points are independent given λ, so that their likelihoods may be multiplied.[¶¶] This leaves her with the following expression for the posterior density of $(\lambda|X_n, a, b)$:

$$p(\lambda|X_n,a,b) = \frac{\frac{b^{a}}{\Gamma(a)}\exp\left(-\lambda b\right)\lambda^{a-1}\prod_{i=1}^{n=3}\frac{1}{x_i!}\exp\left(-\lambda\right)\lambda^{x_i}}{\int_{\Lambda}\frac{b^{a}}{\Gamma(a)}\exp\left(-\lambda b\right)\lambda^{a-1}\prod_{i=1}^{n=3}\frac{1}{x_i!}\exp\left(-\lambda\right)\lambda^{x_i}d\lambda}.$$

This expression may look daunting, but Ginny Weasley is not easily intimidated. She goes through the following algebraic steps to simplify the expression: (1) collect all factors that do not depend on λ (which, notably, includes the entire denominator) and call

---

[‡‡] Recall that $x! = x \times (x-1) \times \cdots \times 1$ (where $x!$ is read as "the factorial of x," or simply "x factorial"). Similarly, the Gamma function $\Gamma(a)$ is equal to $(a-1)! = (a-1)\times(a-2)\times\cdots\times 1$ when a is an integer. Unlike a factorial, however, the Gamma function is more flexible in that it can be applied to non-integers.

[§§] To ease readability we use Greek letters for the parameters of a likelihood function and Roman letters for the parameters of prior (posterior) distributions. The parameters that characterize a distribution can be found on the right side of the conditional bar; for instance, the likelihood function $p(x|\lambda)$ has parameter λ, whereas the prior distribution $p(\lambda|a,b)$ has parameters (a,b).

[¶¶] The likelihood function of the combined data is $p(X_n|\lambda) = p(x_1|\lambda) \times p(x_2|\lambda) \times p(x_3|\lambda)$, which we write using the more compact product notation, $\prod_{i=1}^{n=3} p(x_i|\lambda)$, in the following equations to save space. Similarly, $\prod_{i=1}^{n=3}\exp\left(-\lambda\right)\lambda^{x_i} = \exp(-3\lambda)\lambda^{(x_1+x_2+x_3)}$.

them $Q(X_n)$, and (2) combine exponents with like bases:

$$\begin{aligned} p(\lambda|X_n,a,b) &= Q(X_n)\exp\left(-\lambda b\right)\lambda^{a-1}\times\prod_{i=1}^{n=3}\exp\left(-\lambda\right)\lambda^{x_i} \\ &= Q(X_n)\exp\left[-\lambda(b+n)\right]\lambda^{\left(a+\sum_{i=1}^{n=3}x_i\right)-1}. \end{aligned}$$

Note the most magical result that is obtained here! Comparing the last equation to Equation 14, it turns out that these have *exactly the same form*. Renaming $(b+n)$ to $\hat{b}$ and $\left(a+\sum_i^n x_i\right)$ to $\hat{a}$ makes this especially clear:

$$p(\lambda|X_n,a,b)=\frac{\hat{b}^{\hat{a}}}{\Gamma\left(\hat{a}\right)}\exp\left(-\lambda\hat{b}\right)\lambda^{\hat{a}-1}=p(\lambda|\hat{a},\hat{b}).$$

Here, Ginny has completed the distribution by replacing the scaling constant $Q(X_n)$ with the scaling constant of the Gamma distribution – after all, we know that the outcome must be a probability density, and each density has a unique scaling constant that ensures the total area under it is 1.

The posterior distribution $p(\lambda|X_n,a,b)$ thus turns out to be equal to the prior distribution with updated parameters $\hat{b}=b+n$ and $\hat{a}=a+\sum_{i=1}^n x_i$. Differently put,

$$p(\lambda|X_n,a,b)=p\left(\lambda\mid a+\sum_{i=1}^n x_i, b+n\right). \qquad [15]$$

This amazing property, where the prior and posterior distributions have the same form, results from the special relationship between the Gamma distribution and the Poisson distribution: *conjugacy*. The bottom panel of Figure 4 shows the much more concentrated posterior density for λ: a Gamma distribution with parameters $\hat{a}=36$ and $\hat{b}=3.2$.

When priors and likelihoods are conjugate, three main advantages follow. First, it is easy to express the posterior density because it has the same form as the prior density (as seen in Equation 15). Second, it is straightforward to calculate means and other summary statistics of the posterior density. For example, the mean of a Gamma distribution has a simple formula: a/b. Thus, George and Ginny's prior density for λ has a mean of $a/b=2/.2=10$, and their posterior density for λ has a mean of $\hat{a}/\hat{b}=36/3.2=11.25$. The prior and posterior densities' respective modes are $(a-1)/b=5$ and $(\hat{a}-1)/\hat{b}=35/3.2\approx 11$, as can be seen from Figure 4. Third, it is straightforward to update the posterior distribution sequentially as more data become available.

Discussion Social scientists estimate model parameters in a wide variety of settings. Indeed, a focus on estimation is the core of the *New Statistics* ((4); see also (28)). The *puking pastilles* example illustrates how Bayesian parameter estimation is a direct consequence of the rules of probability theory, and this relationship licenses a number of interpretations that the *New Statistics* does not allow. Specifically, the basis in probability theory allows George and Ginny to (1) point at the most plausible values for the rate of expulsion events and (2) provide an interval that contains the expulsion rate with a certain probability (e.g., a Gamma distribution calculator shows that λ is between 8.3 and 14.5 with 90% probability).

The applications of parameter estimation often involve exploratory settings: no theories are being tested and a distributional model of the data is assumed for descriptive convenience. Nevertheless, parameter estimation can be used to adjudicate between theories under certain special circumstances: if a theory or hypothesis makes a particular prediction about a parameter's value or range, then estimation can take a dual role of hypothesis testing. In the social sciences most measurements have a natural reference point of zero, so this type of hypothesis will usually be in the form of a directional prediction for an effect. In our example, suppose that George was specifically interested in whether λ was less than 10. Under his prior distribution for λ, the probability of that being the case was 59.4%. After seeing the data, the probability λ is less than 10 decreased to 26.2%.

Estimating the mean of a normal distribution. By far the most common distribution used in statistical testing in social science, the normal distribution deserves discussion of its own. The normal distribution has a number of interesting properties—some of them rather unique—but we discuss it here because it is a particularly appropriate choice for modeling unconstrained, continuous data. The mathematical form of the normal distribution is

$$\begin{aligned} p(x|\mu,\sigma) &= N(x|\mu,\sigma^2) \\ &= \frac{1}{\sqrt{2\pi\sigma^2}}\exp\left[-\frac{1}{2}\left(\frac{x-\mu}{\sigma}\right)^2\right], \end{aligned}$$

with the μ (mu) parameter representing the average (*mean*) of the population from which we are sampling and σ (sigma) the amount of dispersion (*standard deviation*) in the population. We will follow the convention that the normal distribution is parameterized with the *variance* σ^2. An example normal distribution is drawn in Figure 3.

One property that makes the normal distribution useful is that it is *self-conjugate*: The combination of a normal prior density and normal likelihood function is itself a normal distribution, which greatly simplifies the derivation of posterior densities. Using Equation 10, and given some data set $X_n=(x_1,x_2,...,x_n)$, we can derive the following expression for the posterior density $(\mu|X_n,a,b)$:

$$p(\mu|X_n,a,b)=\frac{N(\mu|a,b^2)\times\prod_i^n N(x_i|\mu,\sigma^2)}{\int_M N(\mu|a,b^2)\times\prod_i^n N(x_i|\mu,\sigma^2)d\mu}$$

Knowing that the product of normal distributions is also a normal distribution (up to a scaling factor), it is only a matter of tedious algebra to derive the posterior distribution of μ. We do not reproduce the algebraic steps here – the detailed derivation can be found in Gelman et al. (15) and Raiffa and Schlaifer (45), among many other places. The posterior is

$$p(\mu|X_n,a,b)=N\left(\mu|\hat{a},\hat{b}^2\right),$$

where

$$\hat{b}^2=\frac{1}{\frac{n}{\sigma^2}+\frac{1}{b^2}}$$

and

$$\begin{aligned} \hat{a} &= \left(\frac{\hat{b}^2}{b^2}\right)a+\left(\frac{\hat{b}^2}{\sigma^2/n}\right)\bar{x} \\ &= W^2a+\left(1-W^2\right)\bar{x}, \end{aligned}$$

where $\bar{x}$ refers to the mean of the sample.

Carefully inspecting these equations can be instructive. To find $\hat{b}$, the standard deviation (i.e., spread) of the posterior distribution of μ, we must compare the spread of the prior distribution, b, to the *standard error of the sample*, $\sigma/\sqrt{n}$. The formula for $\hat{b}$ represents how our uncertainty about the value of μ is reduced due to the information gained in the sample. If the sample is noisy, such that the standard error of the sample is large compared to the spread of the prior, then relatively little is learned from the data compared to what we already knew before, so the difference between $\hat{b}$ and b will be small. Conversely, if the data are relatively precise, such that the standard error of the sample is small when compared to the spread of the prior, then much will be learned about μ from the data and $\hat{b}$ will be much smaller than b.

To find $\hat{a}$, the mean of the posterior distribution for μ, we need to compute a weighted average of the prior mean and the sample mean. In the formula above, the weights attached to a and $\bar{x}$ sum to 1 and are determined by how much each component contributes to the total precision of the posterior distribution. Naturally, the best guess for the value of μ splits the difference between what we knew of μ before seeing the sample and the estimate of μ obtained from the sample; whether the posterior mean is closer to the prior mean or the sample mean depends on a comparison of their relative precision. If the data are noisy compared to the prior (i.e., the difference between prior variance b^2 and posterior variance $\hat{b}^2$ is small, meaning W^2 is near 1), then the posterior mean will stay relatively close to the prior mean. If the data are relatively precise (i.e., W^2 is near zero), the posterior mean will move to be closer to the sample mean. If the precision of the prior and the precision of the data are approximately equal then W^2 will be near $1/2$, so the posterior mean for μ will fall halfway between a and $\bar{x}$.

The above effect is often known as *shrinkage* because our sample estimates are pulled back toward prior estimates (i.e., shrunk). Shrinkage is generally a desirable effect, in that it will lead to more accurate parameter estimates and empirical predictions (see 9). Since Bayesian estimates are automatically shrunk according to the relative precision of the prior and the data, incorporating prior information simultaneously improves our parameter estimates and protects us from being otherwise misled by noisy estimates in small samples. Quoting Gelman (16, p. 163): "Bayesian inference is conservative in that it goes with what is already known, unless the new data force a change."

Another way to interpret these weights is to think of the prior density as representing some amount of information that is available from an unspecified number of previous hypothetical observations, which are then added to the information from the real observations in the sample. For example, if after collecting 20 data points the weights come to $W^2 = .5$ and $1 - W^2 = .5$, that implies that the prior density carried 20 data points' worth of information. In studies for which obtaining a large sample is difficult, the ability to inject outside information into the problem to come to more informed conclusions can be a valuable asset. A common source of outside information is estimates of effect sizes from previous studies in the literature. As the sample becomes more precise, usually through increasing sample size, W^2 will continually decrease, and eventually the amount of information added by the prior will become a negligible fraction of the total (see also the principle of stable estimation, described in 8).

Example 4: "Of Murtlaps and Muggles". According to *Fantastic Beasts and Where to Find Them* (55), a Murtlap is a "rat-like creature found in coastal areas of Britain" (p. 56). While typically not very aggressive, a startled Murtlap might bite a human, causing a mild rash, discomfort in the affected area, profuse sweating, and some more unusual symptoms.

Anecdotal reports dating back to the 1920s indicate that Muggles (non-magical folk) suffer a stronger immunohistological reaction to Murtlap bites. This example of physiological differences between wizards and Muggles caught the interest of famed magizoologist Newton ("Newt") Scamander, who decided to investigate the issue: When bitten by a Murtlap, do symptoms persist longer in the average Muggle than in the average wizard?

The Ministry of Magic keeps meticulous historical records of encounters between wizards and magical creatures that go back over a thousand years, so Scamander has a great deal of information on wizard reactions to Murtlap bites. Specifically, the average duration of the ensuing sweating episode is 42 hours, with a standard deviation of 2. Due to the large amount of data available, the standard error of measurement is negligible. Scamander's question can now be rephrased: What is the probability a Murtlap bite on a Muggle results in an average sweating episode longer than 42 hours?

Scamander has two parameters of interest: the population mean—episode duration μ—and its corresponding population standard deviation σ. He has no reason to believe there is a difference in dispersion between the magical and non-magical populations, so he will assume for convenience that σ is known and does not differ between Muggles and wizards (i.e., $\sigma = 2$; ideally, σ would be estimated as well, but for ease of exposition we will take the standard deviation as known).

Before collecting any data, Scamander must assign to μ a prior distribution that represents what he believes to be the range of plausible values for this parameter before collecting data. To characterize his background information about the population mean μ, Scamander uses a prior density represented by a normal distribution, $p(\mu|a, b) = N(\mu|a, b^2)$, where a represents the location of the mean of the prior and b represents its standard deviation (i.e., the amount of uncertainty we have regarding μ). From his informal observations, Scamander believes that the mean difference between wizards and Muggles will probably not be larger than 15 hours. To reflect this information, Scamander centers the prior distribution $p(\mu|a, b)$ at $a = 42$ hours (the average among wizards) with a standard deviation of $b = 6$ hours, so that prior to running his study there is a 95% probability μ lies between (approximately) 27 and 57 hours. Thus, $p(\mu|a, b) = N(\mu|42, 6^2)$.

With these prior distributions in hand, Scamander can compute the prior probability that μ is less than 42 hours by finding the area under the prior distribution to the left of the benchmark value via integration. Integration from negative infinity to some constant is most conveniently calculated with the *cumulative distribution function* Φ:

$$\begin{aligned} p(\mu < 42|a, b) &= \int_{-\infty}^{42} N(\mu|a, b^2) d\mu \\ &= \Phi\left(42|a, b^2\right), \end{aligned}$$

which in this case is exactly 0.5 since the benchmark value is exactly the mean of the prior density: Scamander centered his prior on 42 and specified that the Muggle sweating duration could be longer or shorter with equal probability.

Scamander covertly collects information on a representative sample of 30 Muggles by exposing them to an angry Murtlap.*** He finds a sample mean of $\bar{x} = 43$ and standard error of $s = \sigma/\sqrt{n} = 2/\sqrt{30} = 0.3651$. Scamander can now use his data and the above formulas to update what he knows about μ.

[***] In order to preserve the wizarding world's statutes of secrecy, Muggles who are exposed to magical creatures must be turned over to a team of specially-trained wizards called *Obliviators*, who will erase the Muggles' memories, return them to their homes, and gently steer them into the kitchen.

Since the spread of the prior for μ is large compared to the standard error of the sample ($b = 6$ versus $s = 0.3651$), Scamander has learned much from the data and his posterior density for μ is much less diffuse than his prior:

$$\hat{b} = \sqrt{\frac{1}{\frac{1}{s^2} + \frac{1}{b^2}}} = \sqrt{\frac{1}{\frac{1}{0.3651^2} + \frac{1}{6^2}}} = 0.3645.$$

With $\hat{b}$ in hand, Scamander can find the weights needed to average a and $\bar{x}$: $W^2 = (0.3645/6)^2 = 0.0037$ and $1 - W^2 = 0.9963$, thus $\hat{a} = 0.0037 \times 42 + 0.9963 \times 43 = 42.9963$ hours. In summary, Scamander's prior distribution for μ, $p(\mu|a, b) = N(\mu|42, 6^2)$, is updated into a much more informative posterior distribution, $p(\mu|\hat{a}, \hat{b}) = N(\mu|42.9963, 0.3645^2)$. This posterior distribution is shown in the left panel of Figure 5; note that the prior density looks nearly flat when compared to the much more peaked posterior density.

Now that the posterior distribution of μ is known, Scamander can revisit his original question: What is the probability that μ is greater than 42 hours? The answer is again obtained by finding the area under the posterior distribution to the right of the benchmark value via integration:

$$\begin{aligned} p(\mu > 42|\hat{a}, \hat{b}) &= \int_{42}^{\infty} N(\mu|\hat{a}, \hat{b}^2) d\mu \\ &= 1 - \int_{-\infty}^{42} N(\mu|\hat{a}, \hat{b}^2) d\mu \\ &= 1 - \Phi\left(42|\hat{a}, \hat{b}^2\right) \\ &= 1 - \Phi\left(42|42.9963, 0.3645^2\right) \approx 0.9970. \end{aligned}$$

In summary, the probability that the reaction to Murtlap bites in the average Muggle is greater than in the average wizard increases from exactly 50% to 99.70%.

Discussion The conclusion of a Bayesian estimation problem is the full posterior density for the parameter(s). That is, once the posterior density is obtained then the estimation problem is complete. However, researchers often choose to report summaries of the posterior distribution that represent its content in a meaningful way. One common summary of the posterior density is a *posterior (credible) interval*. Credible intervals have a unique property: as Edwards et al. (8) put it, "The Bayesian theory of interval estimation is simple. To name an interval that you feel 95% certain includes the true value of some parameter, simply inspect your posterior distribution of that parameter; any pair of points between which 95% of your posterior density lies defines such an interval" (p. 213). This property is made possible by the inclusion of a prior density in the statistical model (52). It is important not to confuse credible intervals with *confidence intervals*, which have no such property in general (40). Thus, when Scamander reports that there is a 99.70% probability that μ lies between 42 and positive infinity hours, he is reporting a 99.70% credible interval. It is important to note that there is no unique interval for summarizing the posterior distribution; the choice depends on the context of the research question.

4. Model comparison

> [M]ore attention [should] be paid to the precise statement of the alternatives involved in the questions asked. It is sometimes considered a paradox that the answer depends not only on the observations but on the question; it should be a platitude.
>
> H. Jeffreys (23)

Consider the following theoretical questions. Is participant performance different than chance? Does this gene affect IQ? Does stimulus orientation influence response latency? For each of these questions the researcher has a special interest in a particular parameter value and entertains it as a possibility. However, when we estimate a parameter using a continuous distribution the answers to each of these questions is necessarily "yes." To see why, recall that a probability density function specifies how much probability exists *near*—not *at*—a particular value of the parameter. That is, with a continuous probability distribution, probability only exists within a given *range* of the parameter space; the probability of any *single point* within the distribution is zero. This is inconsistent with our belief that a specified parameter value might hold true. Moreover, this poses a problem for any research question that focuses on a single value of a continuous parameter, because if its prior probability is zero then no amount of data can cause its posterior probability to become anything other than zero.

A simple but brilliant solution to this problem was first executed by Haldane (17) but is credited mostly to Jeffreys (1939; see (11)). The solution involves applying the sum and product rules *across multiple independent statistical models at once*. We can specify multiple separate models that have different implications about the parameter of interest, call it θ, and calculate the probability of each model after data are collected. One model, say $\mathcal{M}_0$, says θ is equal to a single special value denoted θ_0. A second model, say $\mathcal{M}_1$, says θ is unknown and assigns it a continuous prior density, implying θ is not equal to θ_0. After collecting data X, there are two main questions to answer: (1) What is $P(\mathcal{M}_0|X)$, the posterior probability that $\theta = \theta_0$? And (2) what is $p(\theta|X, \mathcal{M}_1)$, the posterior distribution[†††] of θ under $\mathcal{M}_1$ (i.e., considering the new data X, if $\theta \neq \theta_0$ then what might θ be)?

As before, this scenario can be approached with the product and sum rules of probability. The setup of the problem is captured by Figure 7 (focusing for now on the left half). We start at the initial fork with two potential models: $\mathcal{M}_0$ and $\mathcal{M}_1$. This layer of analysis is called the *model space*, since it deals with the probability of the models. Subsequently, each model implies some belief about the value of θ. This layer of analysis is called the *parameter space* since it specifies what is known about the parameters within a model, and it is important to note that each model has its own independent parameter space. Under $\mathcal{M}_0$ the value of θ is known to be equal to θ_0, so all of its probability is packed into a "spike" (a *point mass*) at precisely θ_0. Under $\mathcal{M}_1$ the value of θ is unknown and we place a probability distribution over the potential values of θ in the form of a *conditional prior density*. Each model also makes predictions about what data will occur in the experiment (i.e., the model's prior predictive distribution), information represented by

---

[†††] Note that we will now be using probabilities and probability densities side-by-side. In general, if the event to which the measure applies (i.e., what is to the left of the vertical bar) has a finite number of possible values, we will consider probabilities and use uppercase $P(\cdot)$ to indicate that. If the event has an infinite number of possible values in a continuum, we will consider probability densities and use lowercase $p(\cdot)$. In the case of a joint event in which at least one component has an infinite set of possibilities, the joint event will also have an infinite set of possibilities and we will use probability densities there also.

each model's respective *sample space*. We then condition on the data we observe, which allows us to update each layer of the analysis to account for the information gained. Below is a step-by-step account of how this is done, but we remind readers that they should feel free to skip this technical exposition and jump right into the next examples.

We answer our questions in reverse order, first deriving the posterior distribution of θ under $\mathcal{M}_1$, for a reason that will become clear in a moment. In this setup there are events that vary among three dimensions: X, θ, and $\mathcal{M}_1$. When joint events have more than two components, the product rule decomposes $p(X, \theta, \mathcal{M}_1)$ one component at a time to create a chain of conditional probabilities and densities (for this reason the product rule is also known as the *chain rule*). This was seen above in Example 2. These chains can be thought of as moving from one layer of Figure 7 to the next. Thus, since we could choose any one of the three events to be factored out first, the product rule creates three possible initial chains with two probabilities per chain,

$$\begin{aligned} p(X, \theta, \mathcal{M}_1) &= P(\mathcal{M}_1)p(X, \theta|\mathcal{M}_1) \\ &= P(X)p(\theta, \mathcal{M}_1|X) \\ &= p(\theta)p(X, \mathcal{M}_1|\theta). \end{aligned}$$

(where the use of $P(X)$ or $p(X)$ depends on whether the data are discrete or continuous; we assume they are discrete here).

A natural choice is to work with the first formulation, $p(X, \theta, \mathcal{M}_1) = P(\mathcal{M}_1)p(X, \theta|\mathcal{M}_1)$, since $P(\mathcal{M}_1)$, the prior probability of the model, is known to us (it corresponds to the probability we take the right fork at the start of Figure 7). The product rule can then be applied again to the remaining joint probability on the right hand side as follows:

$$P(\mathcal{M}_1) \times p(X, \theta|\mathcal{M}_1) = P(\mathcal{M}_1) \times P(X|\mathcal{M}_1)p(\theta|X, \mathcal{M}_1), \quad [16]$$

By symmetry of the product rule, we can also write

$$P(\mathcal{M}_1) \times P(X, \theta|\mathcal{M}_1) = P(\mathcal{M}_1) \times p(\theta|\mathcal{M}_1)P(X|\theta, \mathcal{M}_1). \quad [17]$$

If we now equate the right hand sides of Equations 16 and 17, we can divide out $P(\mathcal{M}_1)$ and $P(X|\mathcal{M}_1)$:

$$\begin{aligned} P(\mathcal{M}_1)P(X|\mathcal{M}_1)p(\theta|X, \mathcal{M}_1) &= P(\mathcal{M}_1)p(\theta|\mathcal{M}_1)P(X|\theta, \mathcal{M}_1) \\ p(\theta|X, \mathcal{M}_1) &= \frac{p(\theta|\mathcal{M}_1)P(X|\theta, \mathcal{M}_1)}{P(X|\mathcal{M}_1)} \end{aligned}$$

and by recognizing that $P(X|\mathcal{M}_1) = \int_\Theta p(\theta|\mathcal{M}_1)P(X|\mathcal{M}_1, \theta)d\theta$ by way of the sum rule, we are left with the following:

$$p(\theta|X, \mathcal{M}_1) = \frac{p(\theta|\mathcal{M}_1)P(X|\theta, \mathcal{M}_1)}{\int_\Theta p(\theta|\mathcal{M}_1)P(X|\theta, \mathcal{M}_1)d\theta}. \quad [18]$$

This last formula is identical to the continuous form of Bayes' Rule (Equation 10), where now each term is also conditional on $\mathcal{M}_1$.

The implication of this finding is that it is possible to perform inference using the distribution of θ under $\mathcal{M}_1$, $p(\theta|X, \mathcal{M}_1)$, *ignoring everything relating to other models*, since no other models (such as $\mathcal{M}_0$) feature in this calculation. As before, the denominator is known as the *marginal likelihood* for $\mathcal{M}_1$, and represents a predictive distribution for potential future data, $P(X|\mathcal{M}_1)$. This predictive distribution is shown in the sample space under $\mathcal{M}_1$ in Figure 7, and can be thought of as the average prediction made across all possible parameter values in the model (weighted by the conditional prior density). Once the data are collected and the result is known, we can condition on the outcome and use it to update $p(\theta|\mathcal{M}_1)$ to obtain $p(\theta|X, \mathcal{M}_1)$.

To answer our first question—what is $P(\mathcal{M}_0|X)$?—we need to find our way back to the discrete form of Bayes' Rule (Equation 7). Recall that for hypothesis testing the key terms to find are $P(X|\mathcal{M}_0)$ and $P(X|\mathcal{M}_1)$, which can be interpreted as how accurately each hypothesis predicts the observed data in relation to the other. Since the parameter space under $\mathcal{M}_0$ is simply $\theta = \theta_0$, we can write $P(X|\mathcal{M}_0) = P(X|\theta_0)$. However, since the parameter space under $\mathcal{M}_1$ includes a continuous distribution, we need to find $\mathcal{M}_1$'s average predictive success across the whole parameter space, $P(X|\mathcal{M}_1) = \int_\Theta p(\theta|\mathcal{M}_1)P(X|\mathcal{M}_1, \theta)d\theta$. Conveniently, as we just saw above in Equation 18, this is also the normalizing constant in the denominator of the posterior distribution of θ under $\mathcal{M}_1$. Hence, the discrete form of Bayes' Rule for hypothesis testing can be rewritten as

$$\begin{aligned} P(\mathcal{M}_1|X) &= \frac{P(\mathcal{M}_1)P(X|\mathcal{M}_1)}{P(\mathcal{M}_1)P(X|\mathcal{M}_1) + P(\mathcal{M}_0)P(X|\mathcal{M}_0)} \\ &= \frac{P(\mathcal{M}_1)\int_\Theta p(\theta|\mathcal{M}_1)P(X|\theta, \mathcal{M}_1)d\theta}{P(\mathcal{M}_1)\int_\Theta p(\theta|\mathcal{M}_1)P(X|\theta, \mathcal{M}_1)d\theta + P(\mathcal{M}_0)P(X|\theta_0)}. \end{aligned}$$

Furthermore, in cases of model comparison between a "point null" (i.e., an hypothesis that, like our $\mathcal{M}_0$, involves a prior point mass on some parameter) and an alternative with a continuous prior for the parameter, one can rewrite the odds form of Bayes' Rule from Equation 9 as follows:

$$\begin{aligned} \underbrace{\frac{P(\mathcal{M}_1|X)}{P(\mathcal{M}_0|X)}}_{\text{Posterior odds}} &= \frac{P(\mathcal{M}_1)}{P(\mathcal{M}_0)} \times \frac{P(X|\mathcal{M}_1)}{P(X|\mathcal{M}_0)} \\ &= \underbrace{\frac{P(\mathcal{M}_1)}{P(\mathcal{M}_0)}}_{\text{Prior odds}} \times \underbrace{\frac{\int_\Theta p(\theta|\mathcal{M}_1)P(X|\theta, \mathcal{M}_1)d\theta}{P(X|\theta_0)}}_{\text{Bayes factor } (BF_{10})}, \end{aligned}$$

where the Bayes factor is the ratio of the marginal likelihoods from the two models, and its subscript indicates which models are being compared (BF_{10} means $\mathcal{M}_1$ is in the numerator versus $\mathcal{M}_0$ in the denominator).

Finally, we point out one specific application of Bayes' rule that occurs when certain values of θ have a special theoretical status. For example, if θ represents the difference between two conditions in an experiment, then the case $\theta = 0$ will often be of special interest (see also (51)). Dividing each side of Equation 18 by $p(\theta|\mathcal{M}_1)$ allows one to quantify the change in the density at this point:

$$\frac{p(\theta = 0|X, \mathcal{M}_1)}{p(\theta = 0|\mathcal{M}_1)} = \frac{P(X|\theta = 0, \mathcal{M}_1)}{\int_\Theta p(\theta|\mathcal{M}_1)p(X|\theta, \mathcal{M}_1)d\theta} = BF_{01}.$$

This change in density is known as the Savage–Dickey density ratio or the Savage–Dickey representation of the Bayes factor ((6); see also (62), and (65); and see also (37), for some cautionary notes). When it applies, the Savage–Dickey ratio allows for an especially intuitive interpretation of the Bayes factor: If the point null value is lower on the alternative model's conditional posterior density than its prior density, the Bayes factor supports $\mathcal{M}_1$ over $\mathcal{M}_0$ by the ratio of their respective heights, and vice-versa.

The conditions under which the Savage–Dickey ratio applies are typically met in practice, since they correspond to the natural way one would build nested models for comparison (for a good discussion on the different types of nested models see (3), Section 2). Namely, that all facets of the models are the same except that the smaller model fixes θ to be θ_0. In our development above there is only one parameter so this condition is automatically satisfied. If, however, we have additional parameters common to both models, say ϕ, then the Savage-Dickey ratio is obtained using the marginal prior and posterior densities, $p(\theta=\theta_0|X,\mathcal{M}_1)/p(\theta=\theta_0|\mathcal{M}_1)$, where the marginal distribution is found using the sum rule, $p(\theta|X,\mathcal{M}_1)=\int_\Phi p(\phi,\theta|X,\mathcal{M}_1)d\phi$. For this to be a proper representation of the Bayes factor, we must ensure that the conditional prior for ϕ under $\mathcal{M}_1$, when $\theta=\theta_0$, equals the prior density for ϕ under $\mathcal{M}_0$. In other terms, the Savage-Dickey representation holds only if the parameters are statistically independent a priori: $p(\phi|\theta=\theta_0,\mathcal{M}_1)=p(\phi|\mathcal{M}_0)$.

Above, our motivation for model comparison was that we wanted to test the hypothesis that a parameter took a single specified value. However, model comparison is not limited to cases where point nulls are tested. The above formulation allows us to compare any number of different types of models by finding the appropriate $P(X|\mathcal{M})$. Models do not need to be nested or even have similar functional forms; in fact, the models need not be related in any other way than that they make quantitative predictions about the data that have been observed. For example, a non-nested comparison might pit a model with a mostly positive prior distribution for θ against a model where the support of the prior distribution for θ is restricted to negative values only. Or rather than a precise point null we can easily adapt the null model such that we instead compare $\mathcal{M}_1$ against model $\mathcal{M}_S$, which says θ is "small." Extending model comparison to the scenario where there are more than two (but finitely many) competing models $\mathcal{M}_k$ is similar to before, in that

$$P(\mathcal{M}_i|X)=\frac{P(\mathcal{M}_i)p(X|\mathcal{M}_i)}{\sum_k P(\mathcal{M}_k)p(X|\mathcal{M}_k)}. \quad [19]$$

In practice, Bayes factors can be difficult to compute for more complicated models because one must integrate over possibly very many parameters to obtain the marginal likelihood (27, 67). Recent computational developments have made the computation of Bayes factors more tractable, especially for common scenarios (64, 65). For uncommon or complex scenarios, one might resort to reporting a different model comparison metric that does not rely on the marginal likelihood, such as the various information criteria (AIC, BIC, DIC, WAIC) or leave-one-out cross validation (LOOCV; see 56, 59, 60). However, it should be emphasized that for the purposes of inference these alternative methods can be suboptimal.

Example 5: "The French correction". Proud of his work on Murtlap bite sensitivity, Newt Scamander (from Example 4) decides to present his results at a conference on magical zoology held in Carcassonne, France. As required by the 1694 International Decree on the Right of Access to Magical Research Results, he has made all his data and methods publicly available ahead of time and he is confident that his findings will withstand the review of the audience at this annual meeting. He delivers a flawless presentation that culminates in his conclusion that Muggles are, indeed, slightly more sensitive to Murtlap bites than magical folk are. The evidence, he claims, is right there in the data.

After his presentation, Scamander is approached by a member of the audience—the famously critical high-born wizard Jean-Marie le Cornichonesque—with a simple comment on the work: "*Monsieur*, you have not told us the evidence for your claim."

"In fact," continues le Cornichonesque, "given your prior distributions for the difference between Muggles and magical folk, you have not even *considered* the possibility that the true difference might be exactly zero, and your results merely noise. In other words, you are putting the cart before the horse because you estimate a population difference before establishing that evidence for one exists. If I have reservations about whether a basilisk even exists, it does not help for you to give me an estimate for the length of the creature's tail! Instead, if you please, let us ascertain how much more stock we should put in *your* claim over the *more parsimonious* claim of no difference between the respective population means."

Scamander is unfazed by the nobleman's challenge, and, with a flourish of his wand makes the following equations appear in the air between them:

$$\begin{aligned}\mathcal{M}_s: &\quad \mu\sim N(42,6)\\ \mathcal{M}_c: &\quad \mu=42\end{aligned}$$

"These," Scamander says, "are our respective hypotheses. I claim that Muggles have different symptom durations on average than wizards and witches. I have prior information that completes my model. Your claim is that the population means may be exactly equal. In order to quantify the relative support for each of these hypotheses, we need a Bayes factor. Luckily, in this case the Bayes factor is quite easy to calculate with the Savage-Dickey density ratio, like so. . .

$$\begin{aligned}\frac{p(\mu|X,\mathcal{M}_s)}{p(\mu|\mathcal{M}_s)} &= \frac{p(\mu|X,\mathcal{M}_s)}{p(\mu|\mathcal{M}_s)}\\ &= \frac{N(\mu|\hat{a},\hat{b}^2)}{N(\mu|a,b^2)}\end{aligned}$$

"Now that we have derived the ratio of posterior to prior density, all that remains is to plug in the values of the parameters and to compute the ratio of Gaussian densities at the specified points. . . "

$$\begin{aligned}BF_{cs} &= \frac{N(42\,|\,42.9963,0.3645^2)}{N(42\,|\,42,6^2)}\\ &= \frac{0.0261}{0.0665}=0.3925=\frac{1}{2.5475}\end{aligned}$$

"*Tant pis.* A Bayes factor of not even three favors your hypothesis. You have essentially *no* evidence for your claim," snorts le Cornichonesque, before turning his back and leaving Scamander alone in the conference room.

Discussion What has happened here? At first glance, it appears that at first Scamander had strong evidence that Muggles are more sensitive than magical folk to Murtlap bites, and now through some sleight of hand his evidence appears to have vanished. To resolve the paradox of le Cornichonesque, it is important to appreciate a few facts. First, in Example 4, Scamander indeed did not consider the hypothesis $\mathcal{M}_c$ that $\mu=42$. In fact, because a continuous prior density was assigned to μ, the prior probability of it taking on any particular value is zero.

The paradox of le Cornichonesque occurs in part because of a confusion between the hypotheses being considered. While in our example, le Cornichonesque wishes to compare an "existence" and a "nonexistence" hypothesis, Scamander started out from an existence assumption and arrives at conclusions about *directionality* (see also 38).

Implicitly, there are four different models being considered in all. There is $\mathcal{M}_c$, which specifies no effect, and $\mathcal{M}_s$, which specifies

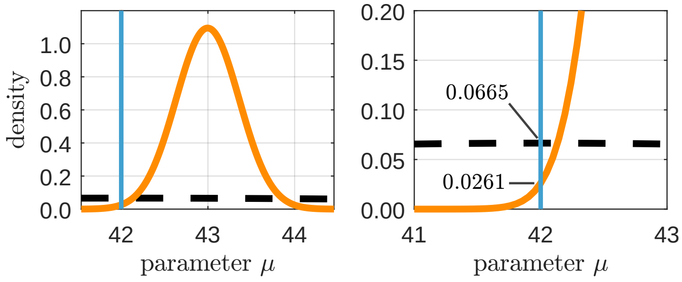

Fig. 5. A closer look at the prior (dashed) and posterior (solid) densities involved in Newt Scamander's study on the relative sensitivity of magical folk and Muggles to Murtlap bites. The left panel shows the location of the fixed value (42) in the body of the prior and posterior distributions. The right panel is zoomed in on the density in the area around the fixed value. Comparing the prior density to the posterior density at the fixed value reveals that very little was learned about this specific value: the density under the posterior is close to the density under the prior and amounts to a Bayes factor of approximately 3 supporting a deviation from the fixed value.

some effect, but also $\mathcal{M}_-$, which specifies an effect in the negative direction, and $\mathcal{M}_+$, which specifies an effect in the positive direction. These last two models are concealed by Scamander's original analysis, but his model specification implies a certain probability for the events $(\mu < 42)$ and $(\mu > 42)$. Indeed, because we know that the probability that Muggles are more (vs. less) sensitive than their magical counterparts increased from $P(\mu > 42) = 50\%$ to $P(\mu > 42|X) = 99.70\%$, we can compute Bayes factors for this case as well. In odds notation, the prior odds were increased from 1 to 333; the Bayes factor, found by taking the ratio of posterior to prior odds, is in this case equal to the posterior odds. Scamander's test for direction returns a much stronger result than le Cornichoneque's test of existence.

As a rule, inference must be limited to the hypotheses under consideration: No method of inference can make claims about theories not considered or ruled out a priori. Moreover, the answer we get naturally depends on the question we ask. The example that follows involves a very similar situation, but the risk of the paradox of le Cornichonesque is avoided by making explicit all hypotheses under consideration.

Example 6: "The measure of an elf". In the wizarding world, the Ministry of Magic distinguishes between two types of living creatures. *Beings*, such as witches, wizards, and vampires, are creatures who have the intelligence needed to understand laws and function in a peaceful society. By contrast, *Beasts* are creatures such as trolls, dragons, and grindylows, which do not have that capacity. Recently, the classification of house-elves has become a matter of contention. On one side of the debate is the populist wizard and radio personality Edward Runcorn, who claims that house-elves are so far beneath wizard intelligence that they should be classified as Beasts; on the other side is the famed elfish philosopher and acclaimed author Doc, who argues that elves are as intelligent as wizards and should be classified as Beings, with all the rights and responsibilities thereof. The Ministry of Magic decides to investigate and convene the *Wizengamot's Internal Subcommittee on House Elf Status* (W.I.S.H.E.S.), an ad-hoc expert committee. W.I.S.H.E.S. in turn calls on psychometrician Dr. Karin Bones of the Magical Testing Service to decide whether house-elves are indeed as intelligent as wizards.

Bones knows she will be asked to testify before W.I.S.H.E.S. and takes note of the composition of the three-member committee. The committee's chairperson is Griselda Marchbanks, a venerable and wise witch who is known for her impartiality and for being of open mind to all eventualities. However, the junior members of W.I.S.H.E.S. are not so impartial: one member is Edward Runcorn, the magical supremacist who believes that wizards and witches are more intelligent than house elves; the other is Hermione Granger, a strong egalitarian who believes that house elves are equal in intelligence to wizards and witches.

Bones begins her task by formalizing three basic hypotheses. She will call the population's average wizarding intelligence quotient (WIQ) μ_w for wizards and witches and μ_e for elves. She can now call the difference between the population means $\delta = \mu_w - \mu_e$ so that δ captures how much *more* intelligent magical folk are. If wizards and elves are equally intelligent, $\delta = 0$. If they are not, δ can take on nonzero values. We can restate this as an hypothesis of approximately no difference ($\mathcal{M}_0$), an hypothesis of substantial positive difference ($\mathcal{M}_+$; magical folk much more intelligent than elves), and an hypothesis of substantial negative difference ($\mathcal{M}_-$; elves much more intelligent than magical folk):

$$\begin{aligned}\mathcal{M}_0 :&\quad \delta \approx 0\\ \mathcal{M}_+ :&\quad \delta > 0\\ \mathcal{M}_- :&\quad \delta < 0.\end{aligned}$$

However, it is not enough to state simply that $\delta < 0$ because as a model for data, it is underspecified: no quantitative predictions follow (i.e., the likelihood for a specific data set cannot be calculated). In order to be more specific, Bones consults with W.I.S.H.E.S. and together they decide on three concrete models:[‡‡‡]

$$\begin{array}{rcll} p(\delta|\mathcal{M}_0) & = & I(-5 < \delta < 5)/10 & \text{if } -5 < \delta < 5\\ p(\delta|\mathcal{M}_+) & = & 2N(\delta|5,15)I(\delta > 5) & \text{if } \delta > 5\\ p(\delta|\mathcal{M}_-) & = & 2N(\delta|-5,15)I(\delta < -5) & \text{if } \delta < -5. \end{array}$$

$\mathcal{M}_0$ is the assumption that the true difference δ is somewhere between -5 and 5 with all values equally likely – a uniform distribution. This is based on a consensus among W.I.S.H.E.S. that differences of only five WIQ points are negligible for the Ministry's classification purposes: differences in this range are *practically equivalent to zero*. Under $\mathcal{M}_+$, it is assumed that wizards score at least 5 points higher than elves on average ($\delta > 5$) but differences of 20 are not unexpected and differences of 40 possible, if unlikely. Under $\mathcal{M}_-$, it is assumed that wizards score at least 5 points *lower* than elves ($\delta < -5$).

After having determined the three hypotheses that W.I.S.H.E.S. wishes to consider, Bones decides to collect one more piece of information: how strongly each member of the committee believes in each of the three options. She provides each member with 100 tokens and three cups, and gives them the following instructions:

> I would like you to distribute these 100 tokens over these three cups. The first cup represents $\mathcal{M}_-$, the second $\mathcal{M}_0$, and the third $\mathcal{M}_+$. You should distribute them proportionally to how strongly you believe in each hypothesis.

Marchbanks' inferred prior probabilities of each of the three hypotheses are $(25, 50, 25)$, Granger's are $(15, 70, 15)$, and Runcorn's are $(5, 15, 80)$. This type of procedure is known as *prior elicitation*; for more in-depth discussion on prior elicitation, see Garthwaite et al. (13) and Lee and Vanpaemel (30).

[‡‡‡] $I(\cdot)$ is the *indicator function*, which takes the value 1 if its argument is true and 0 otherwise; here it takes the role of a truncation. Since these distributions are truncated, they must be multiplied by a suitable constant such that they integrate to 1 (i.e., we *renormalize* them to be proper distributions).

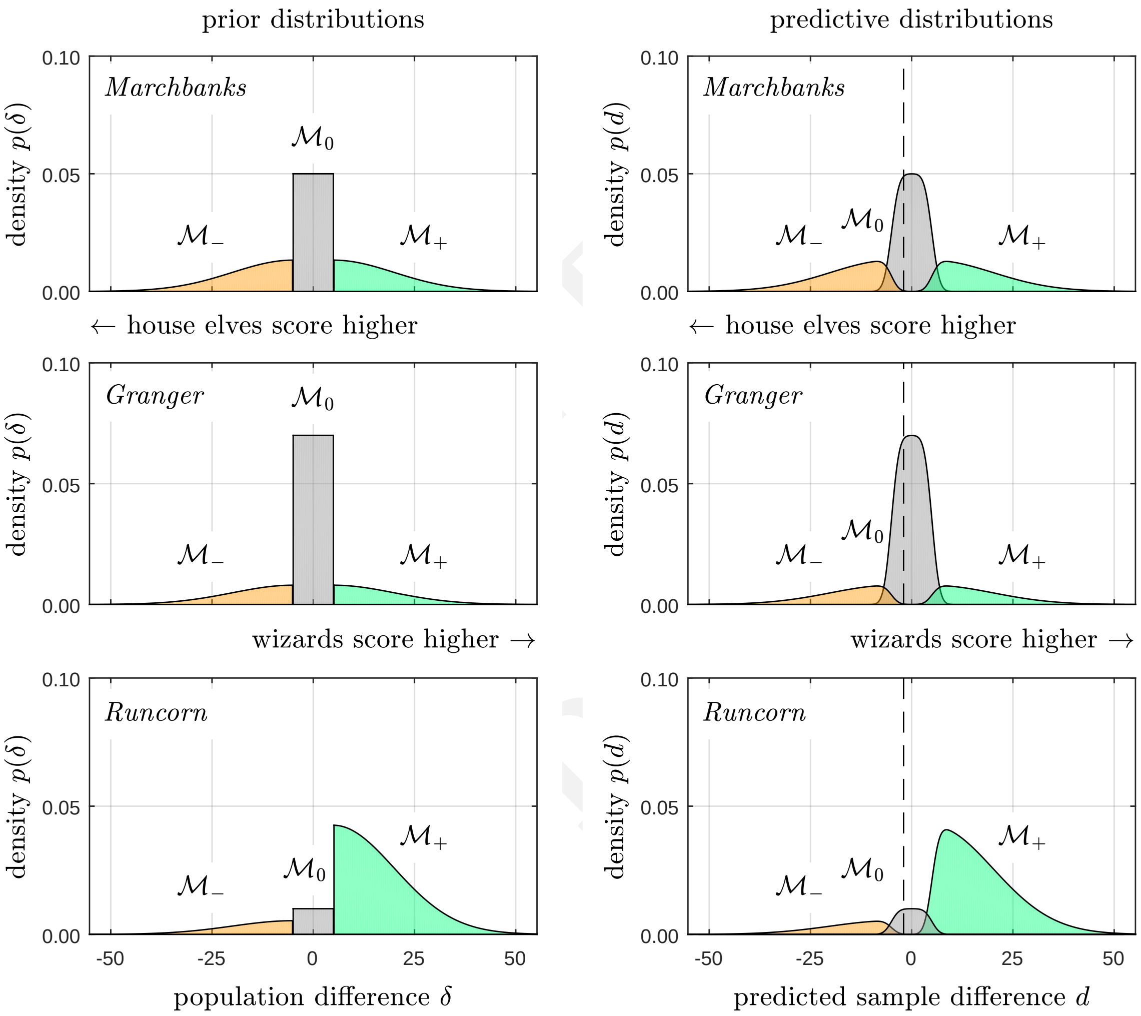

Fig. 6. Left: Each of the three panel members has their own prior probability on each of the three possible models $\mathcal{M}_-$, $\mathcal{M}_0$, and $\mathcal{M}_+$. In this scenario, the three models do not overlap in the parameter space: no parameter value is supported by more than one model. However, this is merely a convenient feature of this example and not a requirement of Bayesian model selection – it is entirely possible (and common) for two different models to support the same parameter value. **Right:** The predicted observed difference in a sample with a standard error of estimation of 1.5. Here, the predictive distribution for each model has been multiplied by the prior probability for that model. This representation has the interesting property that the posterior ratio between two models, given some observed difference, can be read from the figure as the ratio between the heights of the two corresponding densities. Note, for example, that at the dashed vertical line (where $d = 2$), the posterior probability for $\mathcal{M}_0$ is higher than that for $\mathcal{M}_-$ or $\mathcal{M}_+$ for every judge. If the distributions had not been scaled by the prior probability, these height ratios would give the Bayes factor.

To summarize the different prior expectations, Bones constructs a figure to display the marginal distribution of the effect size δ for each committee member. This marginal prior density is easily obtained with the sum rule:

$$\begin{aligned} p(\delta) &= \sum_{h \in (\mathcal{M}_-, \mathcal{M}_0, \mathcal{M}_+)} p(h)p(\delta|h) \\ &= p(\mathcal{M}_-)p(\delta|\mathcal{M}_-) + p(\mathcal{M}_0)p(\delta|\mathcal{M}_0) + p(\mathcal{M}_+)p(\delta|\mathcal{M}_+). \end{aligned}$$

Figure 6 shows the resulting distribution for each of the committee members. These graphs serve to illustrate the relative support each committee member's prior gives to each possible population difference.

Using a well-calibrated test, Bones sets out to gather a sample of $n_1 = 100$ magical folk and $n_2 = 100$ house-elves, and obtains WIQ scores of $M_w = 99.00$ for wizards and witches and $M_e = 101.00$ for elves, giving a sample difference of $d = -2.00$. The test is calibrated such that the standard deviation for magical folk and elves are both equal to 15: $\sigma_w = \sigma_e = 15.00$, which in turn gives a standard deviation for their difference δ of $\sigma_\delta = \sqrt{15^2 + 15^2} = 21.21$. Therefore, the standard error of measurement is $s_e = 21.21/\sqrt{n_1 + n_2} = 1.50$ and the likelihood function to use is now $N\left(d|\delta, s_e^2\right) = N\left(-2|\delta, 1.5^2\right)$.

To address the committee's question, Bones can now use Equation 19 to obtain the posterior probability of each model:

$$P(\mathcal{M}_i|d) = \frac{p(\mathcal{M}_i)p(d|\mathcal{M}_i)}{P(\mathcal{M}_0)p(d|\mathcal{M}_0) + P(\mathcal{M}_-)p(d|\mathcal{M}_-) + P(\mathcal{M}_+)p(d|\mathcal{M}_+)}$$

For this, she needs to compute the three marginal likelihoods $p(d|\mathcal{M}_0)$, $p(d|\mathcal{M}_-)$, and $p(d|\mathcal{M}_+)$, which are obtained with the continuous sum rule. For the case of $\mathcal{M}_0$, the marginal likelihood can be worked out by hand in a few steps:[§§§]

$$\begin{aligned} p(d|\mathcal{M}_0) &= \int_\Delta p(\delta|\mathcal{M}_0) \times p(d|\delta, \mathcal{M}_0) d\delta \\ &= \int_\Delta \frac{1}{10} I(-5 < \delta < 5) \times N(d|\delta, s_e^2) d\delta \\ &= \frac{1}{10} \int_{-5}^{5} N(d|\delta, s_e^2) d\delta \\ &= \frac{1}{10} \left[\Phi(2|-5, 1.5^2) - \Phi(2|5, 1.5^2)\right] \\ &= 9.772 \times 10^{-2} \end{aligned}$$

For the cases of $\mathcal{M}_+$ and $\mathcal{M}_-$, the derivation is much more tedious. It can be done by hand by making use of the fact that the product of two normal distributions has a closed-form solution. However, a numerical approximation can be very conveniently performed with standard computational software or—at the Ministry of Magic—a simple numerical integration spell.[¶¶¶] For this particular task, Dr. Bones arrives at $p(d|\mathcal{M}_+) = 8.139 \times 10^{-8}$ and $p(d|\mathcal{M}_-) = 1.209 \times 10^{-3}$.

Bones now has all that she needs to compute the posterior probabilities of each hypothesis and for each committee member. The prior and posterior probabilities are given in Table 3. As it turns out, the data that Bones has available should effectively overwhelm each of the three members' prior probabilities and put the bulk of the posterior probability on $\mathcal{M}_0$ for each member. Counting on the ability of each committee member to rationally update their beliefs, she prepares a concise presentation in which she lays out a confident case for elf equality and "Being" status.

Table 3. Prior and posterior probabilities for each hypothesis and each committee member. Probabilities are updated with Equation 19. The fourth row in each half of the table serves to emphasize that, for the purposes of the committee, $P(\mathcal{M}_-)$ and $P(\mathcal{M}_0)$ constitute a single category since they both lead to the classification of "Being" rather than "Beast." Thus, we consider $P(\text{"Being"}) = P(\mathcal{M}_-) + P(\mathcal{M}_0)$.

	Marchbanks	Granger	Runcorn
$P(\mathcal{M}_-)$	.250	.150	.050
$P(\mathcal{M}_0)$	.500	.700	.150
$P(\mathcal{M}_+)$	.250	.150	.850
P("Being")	.750	.850	.200
$P(\mathcal{M}_-\|d)$	.006	.003	.012
$P(\mathcal{M}_0\|d)$	.994	.997	.988
$P(\mathcal{M}_+\|d)$	.000	.000	.000
P("Being" $\|$ d)	1.000	1.000	1.000

Discussion Probability theory allows model comparison in a wide variety of scenarios. In this example the psychometrician deals with a set of three distinct models, each of which was constructed ad hoc – custom-built to capture the psychological intuition of the researcher (and a review panel). Once the models were built, the researcher had only to "turn the crank" of probabilistic inference and posterior probabilities are obtained through standard mechanisms that rely on little other than the sum and product rules of probability. As this example illustrates, the practical computation of posterior probabilities will often rely on calculus or numerical integration methods; several papers in this special issue deal with computational software that is available (39, 58, 64, 65).

An interesting aspect to this example is the fact that the analyst is asked to communicate to a diverse audience: three judges who hold different prior notions about the crucial hypotheses. That is, they hold different notions on the *prior probability that each hypothesis is true*. They happen to agree on the *prior distribution of the δ parameter* under each hypothesis (but we made that simplification only for ease of exposition; it is not a requirement of the method). This is comparable to the situation in which most researchers find themselves: there is one data set that brings evidence, but there are many—possibly diverse—prior notions. Given that prior probabilities must be subjective, how can researchers hope to reasonably communicate their results if they can only report their own subjective knowledge?

One potential strategy is the one employed by the psychometrician in the example. The strategy relies on the realization that we can compute posterior probabilities for *any* rational person as soon as we know their prior probabilities. Because the psychometrician had access to the prior probabilities held by each judge, she was able to determine whether her evidence would be compelling to this particular audience.

Social scientists who present evidence to a broad audience can take a similar approach by *formulating multiple prior distributions* – for example, some informative priors motivated by theory, some priors that are uninformative or indifferent in some ways, and some priors that might be held by a skeptic. Such a practice would be a form of *sensitivity analysis* or *robustness analysis*. If the data available are sufficiently strong that skeptics of all camps must

---

[§§§] Bones' derivation makes use of the fact that the identity function $I(\cdot)$ can be factored out of the integrand if the integration bounds are accordingly limited to the region where the argument is true. This fact is used in moving from the second step to the third.

[¶¶¶] Some popular non-magical options include MATLAB (57) and R (44), or readers can use www.wolframalpha.com. MATLAB and R code for this example is available on the OSF repository (https://osf.io/wskex/) and in the Appendix.

rationally come to the same conclusion, then concerns regarding the choice of priors are largely alleviated. This was the case above, where Marchbanks, Granger, and Runcorn all were left with a greater than 98% posterior probability for the model specifying elf equality despite their wide-ranging prior probabilities.

Of course, data is often noisy and the evidence may in many cases not be sufficient to convince the strongest skeptics. In such cases, collecting further data may be useful. Otherwise, the researcher can transparently acknowledge that reasonable people could reasonably come to different conclusions.

An alternative option is to report the evidence in isolation. Especially when the ultimate claim is binary—a discrimination between two models—one might report only the amount of discriminating evidence for or against a model. By reporting only the amount of evidence, in the form of a Bayes factor, every individual reader can combine that evidence with their own prior and form their own conclusions. This is now a widely-recommended approach (e.g., (65); but see (47), for words of caution; and see (28), for a discussion of scenarios in which the Bayes factor should not be the final step of an analysis) that is taken in the final example.

Example 7: "Luck of the Irish". Every four years, the wizarding world organizes the most exhilarating sporting event on earth: the Quidditch World Cup. However, the Cup is often a source of controversy. In a recent edition, aspersions were cast on the uncommonly strong showing by the Irish team: An accusation was brought that the Irish players were dosed with a curious potion called *felix felicis*, which gives an individual an extraordinary amount of "dumb luck."

At the Ministry of Magic's Department for International Magical Cooperation—who oversee the event and have decided to investigate the doping claims—junior statistician Angelina Johnson noticed that the Irish team had *another* striking piece of good luck: in each of the four games, the Irish team captain won the coin toss that allows them to choose in which direction to play. From these data, Johnson reasons as follows.

If the coin is fair, and there is no cheating, then the Irish team captain should win the toss with 50% probability on each occasion ($\mathcal{M}_0 : \theta = \theta_0 = 0.5$). However, if the captain has taken *felix felicis*, they should win with a higher, but unknown probability ($\mathcal{M}_J : \theta > 0.5$). Johnson then sets out to determine whether this small amount of data ($k = 4$ wins in $N = 4$ games) contains enough evidence to warrant strong suspicions.

The discriminating evidence is given by the Bayes factor, $BF_{J0} = P(k|\mathcal{M}_J)/P(k|\mathcal{M}_0)$, where the marginal likelihoods (with capital $P(\cdot)$ since number of wins are discrete) can be calculated one model at a time. Since the outcomes of the four coin tosses are assumed independent given θ, the probability of k successes in any sequence of length N is given by the binomial distribution: $\binom{N}{k}\theta^k(1-\theta)^{N-k}$, where the binomial coefficient $\binom{N}{k}$ is the number of ways N items can arrange themselves in groups of size k (e.g., 4 items can be arranged into a group of 4 exactly 1 way). Thus, for $\mathcal{M}_0$,

$$\begin{aligned} P(k|\mathcal{M}_0) &= \binom{4}{4} 0.5^4 \times 0.5^0 \\ &= \frac{1}{2^4} = \frac{1}{16}. \end{aligned}$$

For $\mathcal{M}_J$, Johnson needs to express her prior knowledge of the parameter θ. Since she knows very little about the potion *felix felicis*, she takes all values between 0.5 and 1.0 to be equally plausible, so that $P(\theta|\mathcal{M}_J) = 2I(0.5 < \theta < 1.0)$. The shape of this prior density is depicted in the left half of Figure 7. Hence,

$$\begin{aligned} P(k|\mathcal{M}_J) &= \int_\Theta p(\theta|\mathcal{M}_J) \times P(k|\theta, \mathcal{M}_J) d\theta \\ &= \int_\Theta 2I(0.5 < \theta < 1.0) \times \binom{4}{4} \theta^4 (1-\theta)^0 d\theta \\ &= 2\int_{0.5}^{1.0} \theta^4 d\theta \\ &= 2\left[\frac{\theta^5}{5}\right]_{0.5}^{1.0} = \frac{2}{5}\left(1^5 - 0.5^5\right) = \frac{31}{80}. \end{aligned}$$

Thus, the data are implied $(31/80) / (1/16) = 6.2$ times more strongly by $\mathcal{M}_J$ than by $\mathcal{M}_0$ (i.e., $BF_{J0} = 6.2$). Johnson concludes that these data afford only a modest amount of evidence—certainly not enough evidence to support a controversial and consequential recommendation—and decides to return to tallying quidditch-related nose fractures instead.

Example 7b: "Luck of the Irish — Part 2". As might be expected, the Irish quidditch controversy did not fail to pique interest throughout the wizarding world. Independently of the Ministry statistician, Barnabas Cuffe, Editor-in-Chief of the *Daily Prophet*—England's premier magical newspaper—had noticed the same peculiar luck in the Irish team's pregame coin tosses. In the editor's case, however, attention to the coin tosses was not a coincidence – in fact, "liquid luck" had helped him win a few career-saving coin tosses in a mildly embarrassing part of his journalistic past.

Cuffe's experience with *felix felicis* is straightforward: on eleven different occasions did he sip the potion just before a coin toss would decide which of two journalistic leads he would pursue that day – his colleague would pursue the other. He recalls clearly that on each of the eleven occasions, his leads carried him in the thick of dramatic, newsworthy events while his colleague's leads turned out dead ends. Cuffe was promoted; his colleague dismissed.

As it happens, Cuffe is an accomplished statistician, and he reasons in much the same way as Angelina Johnson (the junior statistician at the Ministry). If there is no cheating the winning probability should be 50% each time ($\mathcal{M}_0 : \theta = 0.5$). If there *is* cheating, the winning probability should be higher. In contrast to Johnson, however, Cuffe has a good idea how much higher the winning probability θ will be with *felix felicis*: before evaluating the Irish captain's luck he can estimate θ from additional information y that only he possesses.

Cuffe starts by writing down Equation 10 and filling in the quantities on the right hand side. Among these is the prior density $p(\theta)$, which gives the density at each possible value of θ *before considering his own eleven winning coin tosses* y. A reasonable place to start (as before) is that all values between 0.5 and 1.0 are equally plausible: $p(\theta) = 2I(0.5 < \theta < 1.0) = 2I_\theta$ (where we introduce I_θ as a shorthand for $I(0.5 < \theta < 1.0)$, the appropriate indicator function). He also uses the same binomial likelihood function as Johnson, hence,

$$\begin{aligned} p(\theta|y) &= \frac{p(\theta) \times p(y|\theta)}{\int_\Theta p(\theta) \times p(y|\theta) d\theta} \\ &= \frac{2I_\theta \times \binom{11}{11}\theta^{11}(1-\theta)^0}{\int_\Theta 2I_\theta \times \binom{11}{11}\theta^{11}(1-\theta)^0 d\theta} = \frac{2I_\theta \times \theta^{11}}{2\int_{0.5}^{1.0}\theta^{11} d\theta} \\ &= \frac{I_\theta \times \theta^{11}}{\left[\frac{\theta^{12}}{12}\right]_{0.5}^{1.0}} = \frac{I_\theta \times \theta^{11}}{\frac{1}{12}\left(1.0^{12} - 0.5^{12}\right)} \approx 12\theta^{11} I_\theta. \end{aligned}$$

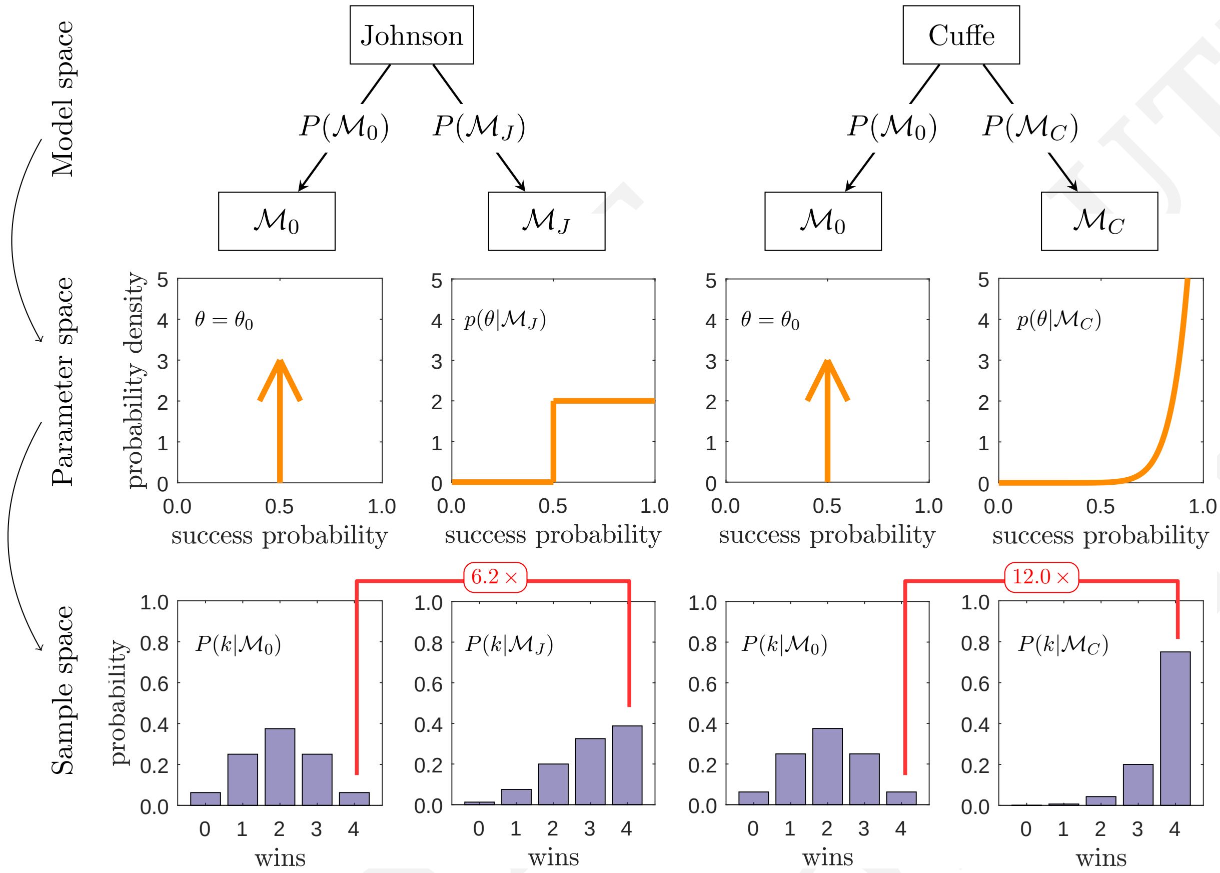

Fig. 7. The structure of Johnson and Cuffe's models, which can be viewed as more complex (rotated) versions of earlier path diagrams. **Top:** The model space shows the contending models. In this case, both Johnson and Cuffe are comparing two models. The prior probabilities for the models are left unspecified. **Middle:** The parameter space shows what each model predicts about the true value of θ (i.e., each model's *conditional prior distribution*). Johnson and Cuffe both use a point null model, which packs all of its mass into a single point (shown as the arrow spiking at $\theta = .5$). However, they have different background knowledge about *felix felicis*, so their prior distributions for θ under their respective alternative model differ. Note that $p(\theta|M_C)$ is obtained from updating $p(\theta|M_J)$ with 11 additional *felix felicis* successes. **Bottom:** The sample space shows what each model predicts about the data to be observed (i.e., each model's prior predictive distribution). The Bayes factor is formed by taking the ratio of the probability each model attached to the observed data, which was four wins in four coin tosses. Since the predictions from the null model are identical for Cuffe and Johnson, the difference in their Bayes factors is due to the higher marginal likelihood Cuffe's alternative model placed on the Irish captain winning all four coin tosses.

This calculation[17] yields Cuffe's *posterior density* of the winning probability θ, which captures his knowledge and uncertainty of the value of θ under luck doping. The shape of this density function is depicted in the right half of Figure 7. Crucially, Cuffe can use this knowledge to perform the same analysis as the Ministry statistician with only one difference: **yesterday's posterior** $p(\theta|y)$ **is today's prior** $p(\theta|\mathcal{M}_C)$. The fact that the latter notation of the prior does not include mention of y serves to illustrate that densities and probabilities are often implicitly conditional on (sometimes informal) background knowledge. Note, for instance, that the entire calculation above assumes that *felix felicis* was taken, but this is not made explicit in the mathematical notation.

Unknowingly repeating Johnson's calculation, Cuffe finds that the probability of the Irish team captain's $k = 4$ winning coin tosses assuming no luck doping is again $p(k|\mathcal{M}_0) = 1/16$. His calculation for the probability of the $k = 4$ wins assuming luck doping is

$$\begin{aligned} P(k|\mathcal{M}_C) &= \int_\Theta p(\theta|\mathcal{M}_C) \times p(k|\theta, \mathcal{M}_C) d\theta \\ &\approx \int_{0.5}^{1.0} 12\theta^{11} I_\theta \times \binom{4}{4} \theta^4 (1-\theta)^0 \, d\theta \\ &= 12 \left[\frac{\theta^{16}}{16}\right]_{0.5}^{1.0} = \frac{12}{16}\left(1^{16} - 0.5^{16}\right) \approx \frac{12}{16}. \end{aligned}$$

To complete his analysis, Cuffe takes the ratio of marginal likelihoods, $BF_{C0} = P(k|\mathcal{M}_C)/P(k|\mathcal{M}_0) \approx 12$, which is strong—but not very strong—evidence in favor of Cuffe's luck doping model.

Inspired partly by the evidence and partly by the recklessness that follows from years of *felix felicis* abuse, editor Cuffe decides to publish an elaborate exposé condemning both the Irish quidditch team for cheating and the Ministry of Magic for failing to act on strong evidence of misconduct.

Discussion This final, two-part example served mostly to illustrate the effects of prior knowledge on inference. This is somewhat in contrast to Example 6, where the prior information was overwhelmed by the data. In the two scenarios here, the Ministry junior statistician and the *Prophet* editor are both evaluating evidence that discriminates between two models. Both consider a "nil model" in which all parameters are known (the fairness of a coin implies that the parameter θ must be 0.5), but they critically differ in their definition of the alternative model. The Ministry statistician, having no particular knowledge of the luck doping potion, considers all better-than-chance values equally plausible, whereas the *Prophet* editor can quantify and insert relevant prior information that specifies the expected effects of the drug in question to greater precision.

As illustrated in the bottom row of Figure 7, these three models (the chance model $\mathcal{M}_0$, the Ministry model $\mathcal{M}_J$, and the *Prophet* model $\mathcal{M}_C$) make distinct predictions: $\mathcal{M}_0$ predicts a distribution of Irish coin toss wins that is symmetric about $k = 2$; $\mathcal{M}_J$ predicts a right-leaning distribution with a greater probability of four Irish wins; and $\mathcal{M}_C$ predicts an even greater such probability. More specifically, the marginal likelihoods are $P(k|\mathcal{M}_0) = 5/80$, $P(k|\mathcal{M}_J) = 31/80$, and $P(k|\mathcal{M}_C) \approx 60/80$, and the Bayes factor between any two of these models is given by forming the appropriate ratio.

This example illustrates a general property in Bayesian model comparison: A model that makes precise predictions can be confirmed to a much stronger extent than a model that makes vague predictions, while at the same time the precision of its predictions makes it easier to disconfirm. The reason Cuffe was able to obtain a higher Bayes factor than Johnson is because his alternative model made much more precise predictions; $\mathcal{M}_C$ packed three-quarters of its prior predictive distribution into $k = 4$, whereas $\mathcal{M}_J$ spread its probability more broadly among the potential outcomes. Since Cuffe's precise prediction was correct, he was rewarded with a larger Bayes factor. However, Cuffe's prediction was risky: if the Irish captain had won any fewer than all four coin tosses, $\mathcal{M}_0$ would have been supported over $\mathcal{M}_C$. In contrast, the Bayes factor would still favor $\mathcal{M}_J$ when $k = 3$ because Johnson's model is more conservative in its predictions. In sum, the ability to incorporate meaningful theoretical information in the form of a prior distribution allows for more informed predictions and hence more efficient inferences (30).

5. Broader appeal and advantages of Bayesian inference

> The Bayesian approach is a common sense approach. It is simply a set of techniques for orderly expression and revision of your opinions with due regard for internal consistency among their various aspects and for the data.
>
> W. Edwards et al. (8)

In our opinion, the greatest theoretical advantage of Bayesian inference is that it unifies all statistical practices within the consistent formal system of probability theory. Indeed, the unifying framework of Bayesian inference is so uniquely well-suited for scientific inference that these authors see the two as synonymous. Inference is the process of combining multiple sources of information into one, and the rules for formally combining information derive from two simple rules of probability. Inference can be as straightforward as determining the event of interest (in our notation, usually $\mathcal{M}$ or θ) and the relevant data and then exploring what the sum and product rules tell us about their relationship.

As we have illustrated, common statistical applications such as parameter estimation and hypothesis testing naturally emerge from the sum and product rules. However, these rules allow us to do much more, such as make precise quantitative predictions about future data. This intuitive way of making predictions can be particularly informative in discussions about what one should expect in future studies – it is perhaps especially useful for predicting and evaluating the outcome of a replication attempt, since we can derive a set of new predictions after accounting for the results of the original study (e.g., 61, 63).

The practical advantages of using probability theory as the basis of scientific and statistical inference are legion. One of the most appealing in our opinion is it allows us to make probabilistic statements about the quantities of actual interest, such as "There is a 90% probability the participants are guessing," or "The probability is .5 that the population mean is negative." It also allows us to construct hierarchical models that more accurately capture the structure of our data, which often includes modeling theoretically-meaningful variability at the participant, task, item, or stimulus level (14, 31, 53).

Bayesian inference also gracefully handles so-called *nuisance parameters.* In most of our present examples there has been only a single quantity of interest – in order to help keep the examples

---

[17] Note that here and below, we make use of a convenient approximation: $0.5^k \approx 0$ for large values of k. Making the calculation exact is not difficult but requires a rather unpleasant amount of space. Also note that the indicator function from the prior density carries over to the posterior density.

simple and easy to follow. In real applications, however, there are typically many parameters in a statistical model, some of which we care about and some of which we do not. The latter are called nuisance parameters because we have little interest in them: we only estimate them out of necessity. For example, if we were estimating the mean of a normal distribution (as in Example 4) and did not know the population standard deviation, then we would have to assign it a prior density, such that the overall prior density would be of the form $p(\mu, \sigma)$; after collecting data X, the posterior density would be of the form $p(\mu, \sigma|X)$. Since we are generally only interested in the parameter μ, estimating σ out of necessity, σ is considered a nuisance parameter. To make inferences about μ we merely integrate out σ from the posterior density using the sum rule: $p(\mu|X) = \int_\Sigma p(\mu, \sigma|X)d\sigma$, from which we can do inference about μ. Similarly, in Examples 7 and 7b, the exact win rate from a luck-doped coin toss is not of primary interest, only whether the coin tossed in the four games was plausibly fair or not. Here, the bias parameter of the coin can be seen as a nuisance parameter. Dealing with nuisance parameters in a principled way is a unique advantage of the Bayesian framework: except for certain special cases, frequentist inference can become paralyzed by nuisance parameters.

The ability of Bayesian inference to deal with nuisance parameters also allows it to flexibly handle one of the biggest statistical challenges for data analysts: situations in which the assumptions of the statistical model regarding the data are badly violated. For example, one of the most common assumptions violated is that of normality (e.g., due to the presence of many outliers). In technical terms, this means that we may not think the normal likelihood function adequately characterizes the data-generating mechanism for the inference problem at hand. In Bayesian inference the choice of likelihood is important because, as we have seen in the estimation examples above, with even moderate samples sizes the likelihood quickly begins to dominate the prior densities. To resolve this issue a Bayesian can construct two models: one that uses a normal likelihood function (model $\mathcal{M}_N$), and one that uses a likelihood function with wider tails (model $\mathcal{M}_W$), such as a t distribution with few degrees of freedom. After collecting data we then have a posterior distribution for the parameters of interest for each model, $p(\theta|X, \mathcal{M}_N)$ and $p(\theta|X, \mathcal{M}_W)$. If we assign prior probabilities to these two models (we emphasize that a "model" consists of both a prior distribution for the parameters and a likelihood function for the data), $P(\mathcal{M}_N)$ and $P(\mathcal{M}_W)$, we can calculate their posterior probabilities $P(\mathcal{M}_N|X)$ and $P(\mathcal{M}_W|X)$. We are then in a position to use the sum rule to marginalize over the different models (as Dr. Bones did with the various prior densities in Example 6), allowing us to find the *model-averaged* posterior density for θ,

$$p(\theta|X) = P(\mathcal{M}_N|X)p(\theta|X, \mathcal{M}_N) + P(\mathcal{M}_W|X)p(\theta|X, \mathcal{M}_W).$$

Note that model averaging is in a sense the flip-side of model selection: In model selection, the identity of the model is central while the *model parameters* are sometimes seen as nuisance variables to be integrated away. By contrast, in the previous equation the *model identities* are treated as nuisance variables while the shared model parameters remain central (see 11, 48). The flexibility to perform model averaging across any variable we care to name (e.g. 19, 35) is a unique advantage of Bayesian inference.

Finally, Bayesian analysis allows for immense freedom in data collection because it respects the *likelihood principle* (2). The likelihood principle states that the likelihood function of the data contains all of the information relevant to the evaluation of statistical evidence. What this implies is that other properties of the data or experiment that do not factor into the likelihood function are *irrelevant* to the statistical inference based on the data (33, 54). Adherence to the likelihood principle means that one is free to do analyses without needing to adhere to rigid sampling plans, or even have any plan at all (49). Note that we did not consider the sampling plan in any of our examples above, and none of the inferences we made would have changed if we had. Within a Bayesian analysis, "It is entirely appropriate to collect data until a point has been proven or disproven, or until the data collector runs out of time, money, or patience" (8, p. 193).

6. Conclusion

> [W]e believe that Bayes' theorem is not only useful, but in fact leads to the *only* correct formulas for solving a large number of our cryptanalytic problems.
>
> F. T. Leahy (29) [emphasis original]

The goal of this introduction has been to familiarize the reader with the fundamental principles of Bayesian inference. Other contributions in this special issue (7, 28) focus on why and how Bayesian methods are preferable to the methods proposed in the *New Statistics* (4). The Bayesian approach to all inferential problems follows from two simple formal laws: the sum and product rules of probability. Taken together and in their various forms, these two rules make up the entirety of Bayesian inference – from testing simple hypotheses and estimating parameters, to comparing complex models and producing quantitative predictions.

The Bayesian method is unmatched in its flexibility, is rooted in relatively straightforward calculus, and uniquely allows researchers to make statements about the relative probability of theories and parameters – and to update those statements with more data. That is, the laws of probability show us how our scientific opinions can evolve to cohere with the results of our empirical investigations. For these reasons, we recommend that social scientists adopt Bayesian methods rather than the *New Statistics*, and we hope that the present introduction will contribute to deterring the field from taking an evolutionary step in the wrong direction.

1 D. J. Bem. Feeling the future: Experimental evidence for anomalous retroactive influences on cognition and affect. *Journal of Personality and Social Psychology*, 100:407–425, 2011.

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

3 G. Consonni and P. Veronese. Compatibility of prior specifications across linear models. *Statistical Science*, 23:332–353, 2008.

4 G. Cumming. The new statistics: Why and how. *Psychological Science*, 25:7–?29, 2014.

5 B. de Finetti. *Theory of Probability, Vol. 1*. John Wiley & Sons, New York, 1974.

6 J. M. Dickey. The weighted likelihood ratio, linear hypotheses on normal location parameters. *The Annals of Mathematical Statistics*, 42:204–223, 1971.

7 Z. Dienes and N. McLatchie. Four reasons to prefer Bayesian over orthodox statistical analyses. *Psychonomic Bulletin and Review*, this issue.

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

9 B. Efron and C. Morris. Stein's paradox in statistics. *Scientific American*, 236:119–127, 1977.

10 A. Etz and Joachim Vandekerckhove. A Bayesian perspective on the reproducibility project. *PLOS ONE*, 11:e0149794, 2016. .

11 A. Etz and E.-J. Wagenmakers. J. B. S. Haldane's contribution to the Bayes factor hypothesis test. *Statistical Science*, in press. URL http://arxiv.org/abs/1511.08180.

12 Michael Evans. Discussion of "On the Birnbaum Argument for the Strong Likelihood Principle". *Statistical Science*, 29(2):242–246, 2014.

13 Paul H Garthwaite, Joseph B Kadane, and Anthony O'Hagan. Statistical methods for eliciting probability distributions. *Journal of the American Statistical Association*, 100(470):680–701, 2005.

14 A. Gelman and J. Hill. *Data Analysis Using Regression and Multilevel/Hierarchical Models*. Cambridge University Press, Cambridge, 2007.

15 A. Gelman, J. B. Carlin, H. S. Stern, and D. B. Rubin. *Bayesian Data Analysis (2nd ed.)*. Chapman & Hall/CRC, Boca Raton (FL), 2004.

16 Andrew Gelman. Bayesian statistics then and now. *Statistical Science*, 25(2):162–165, 2010.

17 J. B. S. Haldane. A note on inverse probability. *Mathematical Proceedings of the Cambridge Philosophical Society*, 28:55–61, 1932.

18 B. M. Hill. On coherence, inadmissibility and inference about many parameters in the theory of least squares. In Stephen E Fienberg and Arnold Zellner, editors, *Studies in Bayesian econometrics and statistics*, pages 555–584. North-Holland Amsterdam, 1974.

19 J. A. Hoeting, D. Madigan, A. E. Raftery, and C. T. Volinsky. Bayesian model averaging: A tutorial. *Statistical Science*, 14:382–417, 1999.

20 E. T. Jaynes. Bayesian methods: General background. In J. H. Justice, editor, *Maximum Entropy and Bayesian Methods in Applied Statistics*, pages 1–25. Cambridge University Press, 1986.

21 Edwin T Jaynes. The intuitive inadequacy of classical statistics. *Epistemologia*, 7(43):43–74, 1984.

22 H. Jeffreys. Some tests of significance, treated by the theory of probability. *Mathematical Proceedings of the Cambridge Philosophy Society*, 31:203–222, 1935.

23 H. Jeffreys. *Theory of Probability*. Oxford University Press, Oxford, UK, 1 edition, 1939.

24 H. Jeffreys. *Theory of Probability*. Oxford University Press, Oxford, UK, 3 edition, 1961.

25 H. Jeffreys. *Scientific Inference*. Cambridge University Press, Cambridge, UK, 3 edition, 1973.

26 Alan Jern, Kai-Min K Chang, and Charles Kemp. Belief polarization is not always irrational. *Psychological review*, 121(2):206, 2014.

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

28 J. Kruschke and T. Liddell. The Bayesian New Statistics: Hypothesis testing, estimation, meta-analysis, and planning from a Bayesian perspective. *Psychonomic Bulletin and Review*, this issue.

29 FT Leahy. Bayes marches on. *National Security Agency Technical Journal*, 5(1):49–61, 1960.

30 M. D. Lee and W. Vanpaemel. Determining informative priors for cognitive models. *Psychonomic Bulletin and Review*, this issue.

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

32 D. V. Lindley. *Making Decisions*. Wiley, London, 2 edition, 1985.

33 D. V. Lindley. The analysis of experimental data: The appreciation of tea and wine. *Teaching Statistics*, 15:22–25, 1993.

34 D. V. Lindley. The philosophy of statistics. *The Statistician*, 49:293–337, 2000.

35 William A Link and Richard J Barker. Bayes factors and multimodel inference. In *Modeling Demographic Processes In Marked Populations*, pages 595–615. Springer, 2009.

36 A. Ly, A. J. Verhagen, and E.-J. Wagenmakers. Harold Jeffreys's default Bayes factor hypothesis tests: Explanation, extension, and application in psychology. *Journal of Mathematical Psychology*, 72:19 – 32, 2016.

37 Jean-Michel Marin and Christian P Robert. On resolving the Savage–Dickey paradox. *Electronic Journal of Statistics*, 4:643–654, 2010. .

38 M Marsman and E.-J. Wagenmakers. Three insights from a Bayesian interpretation of the one-sided P value. *Educational and Psychological Measurement*, 2016.

39 D. Matzke, U. Boehm, and J. Vandekerckhove. Bayesian inference for psychology, part III: Parameter estimation in nonstandard models. *Psychonomic Bulletin and Review*, this issue.

40 R. D. Morey, R. Hoekstra, J. N. Rouder, M. D. Lee, and E.-J. Wagenmakers. The fallacy of placing confidence in confidence intervals. *Psychonomic Bulletin & Review*, 23(1):103–123, 2016.

41 R. D. Morey, J. W. Romeijn, and J. N. Rouder. The philosophy of Bayes factors and the quantification of statistical evidence. *Journal of Mathematical Psychology*, 72:6–18, 2016.

42 J. Neyman. Frequentist probability and frequentist statistics. *Synthese*, 36:97–131, 1977.

43 The Open Science Collaboration. Estimating the reproducibility of psychological science. *Science*, 349:aac4716, 2015.

44 R Development Core Team. *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria, 2004. URL http://www.R-project.org.

45 H. Raiffa and R. Schlaifer. *Applied Statistical Decision Theory*. The MIT Press, Cambridge (MA), 1961.

46 C. P. Robert. *The Bayesian choice: from decision-theoretic foundations to computational implementation*. Springer Science & Business Media, 2007.

47 C. P. Robert. The expected demise of the Bayes factor. *Journal of Mathematical Psychology*, 72:33–37, 2016.

48 Harry V Roberts. Probabilistic prediction. *Journal of the American Statistical Association*, 60 (309):50–62, 1965.

49 J. N. Rouder. Optional stopping: No problem for Bayesians. *Psychonomic Bulletin & Review*, 21:301–308, 2014.

50 J. N. Rouder and R. D. Morey. A Bayes–factor meta analysis of Bem's ESP claim. *Psychonomic Bulletin & Review*, 18:682–689, 2011.

51 J. N. Rouder and J. Vandekerckhove. Bayesian inference for psychology, part IV: Parameter estimation and Bayes factors. *Psychonomic Bulletin and Review*, this issue.

52 Jeffrey N. Rouder, Richard D. Morey, Josine Verhagen, Jordan M. Province, and Eric-Jan Wagenmakers. Is there a free lunch in inference? *Topics in Cognitive Science*, 8:520–547, 2016.

53 Jeffrey N. Rouder, Richard D. Morey, and Michael S. Pratte. *Bayesian Hierarchical Models*. Cambridge University Press, In Press.

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

55 Newton Artemis Fido Scamander. *Fantastic Beasts and Where to Find Them*. Harry Potter. Obscurus Books, London, UK, 2001. ISBN 9781408835050.

56 D. J. Spiegelhalter, N. G. Best, B. P. Carlin, and A. van der Linde. Bayesian measures of model complexity and fit. *Journal of the Royal Statistical Society B*, 64:583–639, 2002.

57 The Mathworks, Inc. *MATLAB version R2015a*. Natick, MA, 2015.

58 D. van Ravenzwaaij, P. Cassey, and S. Brown. A simple introduction to Markov chain Monte-Carlo sampling. *Psychonomic Bulletin and Review*, this issue.

59 J. Vandekerckhove, D. Matzke, and E.-J. Wagenmakers. Model comparison and the principle of parsimony. In J. Busemeyer, J. Townsend, Z. J. Wang, and A. Eidels, editors, *Oxford Handbook of Computational and Mathematical Psychology*, pages 300–319. Oxford University Press, 2015.

60 A. Vehtari and J. Ojanen. A survey of Bayesian predictive methods for model assessment, selection and comparison. *Statistics Surveys*, 6:142–228, 2012.

61 A. J. Verhagen and E.-J. Wagenmakers. Bayesian tests to quantify the result of a replication attempt. *Journal of Experimental Psychology: General*, 143:1457–1475, 2014.

62 E.-J. Wagenmakers, T. Lodewyckx, H. Kuriyal, and R. Grasman. Bayesian hypothesis testing for psychologists: A tutorial on the Savage–Dickey method. *Cognitive Psychology*, 60:158–189, 2010.

63 E.-J. Wagenmakers, A. J. Verhagen, and A. Ly. How to quantify the evidence for the absence of a correlation. *Behavior Research Methods*, 48:413–426, 2016.

64 Eric-Jan Wagenmakers, Jonathon Love, Maarten Marsman, Tahira Jamil, Alexander Ly, Josine Verhagen, Ravi Selker, Quentin F. Gronau, Damian Dropmann, Bruno Boutin, Frans Meerhoff, Patrick Knight, Akash Raj, Erik-Jan van Kesteren, Johnny van Doorn, Martin Šmíra, Sacha Epskamp, Alexander Etz, Dora Matzke, Jeffrey N. Rouder, and Richard D. Morey. Bayesian inference for psychology, part II: Example applications with JASP. *Psychonomic Bulletin and Review*, this issue.

65 Eric-Jan Wagenmakers, Maarten Marsman, Tahira Jamil, Alexander Ly, Josine Verhagen, Jonathon Love, Ravi Selker, Quentin Gronau, Martin Smira, Sacha Epskamp, Dora Matzke, Jeffrey Rouder, and Richard Morey. Bayesian inference for psychology, part I: Theoretical advantages and practical ramifications. *Psychonomic Bulletin and Review*, this issue.

66 Eric-Jan Wagenmakers, Richard D Morey, and Michael D Lee. Bayesian benefits for the pragmatic researcher. *Perspectives on Psychological Science*, in press.

67 L. Wasserman. Bayesian model selection and model averaging. *Journal of Mathematical Psychology*, 44:92–107, 2000.

68 Robert L Winkler. *An introduction to Bayesian inference and decision*. Holt, Rinehart and Winston New York, 1972.

69 D. Wrinch and H. Jeffreys. On some aspects of the theory of probability. *Philosophical Magazine*, 38:715–731, 1919.

70 D. Wrinch and H. Jeffreys. On certain fundamental principles of scientific inquiry. *Philosophical Magazine*, 42:369–390, 1921.

A. Computer code for “The measure of an elf”

```matlab
 1  % MATLAB/Octave code for Example 6
 2
 3  % Define the three models
 4  e =  5;
 5  s = 15;
 6
 7  h0 = @(x) (x < e & x > -e) / (2 * e);
 8  hn = @(x) normpdf(x, -e, s) * 2 .* (x < -e);
 9  hp = @(x) normpdf(x,  e, s) * 2 .* (x >  e);
10
11  % Define the data and likelihood
12  d   =  -2;
13  n   = 100;
14  sem = sqrt((s^2 + s^2) / (2 * n));
15
16  likelihood = @(x) normpdf(d, x, sem);
17
18  % Define the integrands and integrate
19  fn = @(x)likelihood(x) .* hn(x);
20  mn = quadgk(fn, -inf,  -e, 'waypoints', [-e, e]);
21
22  f0 = @(x)likelihood(x) .* h0(x);
23  m0 = quadgk(f0,   -e,   e, 'waypoints', [-e, e]);
24
25  fp = @(x)likelihood(x) .* hp(x);
26  mp = quadgk(fp,    e, inf, 'waypoints', [-e, e]);
27
28  ev = [mn, m0, mp];
29
30  % Apply Bayes' rule
31  eq19 = @(p,m) p .* m ./ sum(p .* m);
32
33  marchbanks = [.25, .50, .25];
34  granger    = [.15, .70, .15];
35  runcorn    = [.45, .10, .45];
36
37  eq19(marchbanks, ev)
38  % ans =    0.0061    0.9939    0.0000
39  eq19(granger, ev)
40  % ans =    0.0026    0.9974    0.0000
41  eq19(runcorn, ev)
42  % ans =    0.0122    0.9878    0.0000
```

```r
 1  # R code for Example 6
 2
 3  # Define the three models
 4  e <-  5
 5  s <- 15
 6
 7  h0 <- function(x) (x < e & x > -e) / (2 * e)
 8  hn <- function(x) dnorm(x, -e, s) * 2 * (x < -e)
 9  hp <- function(x) dnorm(x,  e, s) * 2 * (x >  e)
10
11  # Define the data and likelihood
12  d   <-  -2
13  n   <- 100
14  sem <- sqrt((s^2 + s^2) / (2 * n))
15
16  like <- function(x) dnorm(d, x, sem)
17
18  # Define the integrands and integrate
19  fn <- function(x) like(x) * hn(x)
20  mn <- integrate(fn, -Inf, -e)$value
21
22  f0 <- function(x) like(x) * h0(x)
23  m0 <- integrate(f0, -e, e)$value
24
25  fp <- function(x) like(x) * hp(x)
26  mp <- integrate(fp, e, Inf)$value
27
28  ev <- c(mn, m0, mp)
29
30  # Apply Bayes' rule
31  eq19 <- function(p,m) p*m / sum(p*m)
32
33  marchbanks <- c(.25, .50, .25)
34  granger    <- c(.15, .70, .15)
35  runcorn    <- c(.10, .10, .80)
36
37  eq19(marchbanks, ev)
38  # [1] 6.145693e-03 9.938539e-01 4.138483e-07
39  eq19(granger, ev)
40  # [1] 2.643151e-03 9.973567e-01 1.779886e-07
41  eq19(runcorn, ev)
42  # [1] 1.221623e-02 9.877772e-01 6.581086e-06
```

MATLAB/Octave users who do not have access to the Statistics Toolbox can add on line 6:

```matlab
normpdf = @(x,m,s) exp(-((x-m)./s).^2/2)./sqrt(2.*s.^2.*pi);
```