8 Bayesian inference

With $\mathbb{P}(A|B)$ defined as the degree to which logical proposition B would, assuming it were true, imply that logical proposition A is true (Section 4.1.1), the main requirement for a method of inference is to establish a relationship between $\mathbb{P}(A|B)$ and $\mathbb{P}(B|A)$. This enables the shift from deductive reasoning, in which implications of an assumed model of reality are explored, to inductive reasoning, where it is possible to assess the implications of available data for such a model. The necessary probabilistic relationship is given by Bayes's theorem.

8.1 Bayes's theorem

One of the various laws of probability asserted or derived in Section 4 is the chain rule (Equation 33),

$$\mathbb{P}(A, B|C) = \mathbb{P}(B|C)\,\mathbb{P}(A|B, C), \tag{79}$$

valid for any propositions A, B and C, given only that C provides sufficient information that $\mathbb{P}(A|C)$ and $\mathbb{P}(B|C)$ are defined. The symmetry of the "and" operation implies that "A, B" and "B, A" are identical (compound) propositions, so it is also true that

$$\mathbb{P}(A, B|C) = \mathbb{P}(B, A|C) = \mathbb{P}(A|C)\,\mathbb{P}(B|A, C). \tag{80}$$

Equating the right-hand sides of the above two equations then gives

$$\mathbb{P}(B|C)\,\mathbb{P}(A|B, C) = \mathbb{P}(A|C)\,\mathbb{P}(B|A, C). \tag{81}$$

Dividing both sides of Equation 81 by $\mathbb{P}(B|C)$ (assuming it is positive) then gives

$$\mathbb{P}(A|B, C) = \frac{\mathbb{P}(A|C)\,\mathbb{P}(B|A, C)}{\mathbb{P}(B|C)}, \tag{82}$$

known universally as Bayes's theorem.

The key point of this result is that it provides a way to "invert" probabilities, allowing the evaluation of $\mathbb{P}(A|B, C)$ in terms of $\mathbb{P}(B|A, C)$. However, $\mathbb{P}(A|B, C)$ is not given *just* in terms of $\mathbb{P}(B|A, C)$ – the self consistency conditions from which this result follows necessitate the inclusion of $\mathbb{P}(A|C)$. All four probabilities that make up Bayes's theorem are conditioned on C, which means that it represents knowledge that is not being questioned, at least in the context of this calculation (altough there is nothing to stop Bayes's theorem being applied with C as the subject, provided some presumably more fundamental background information is assumed).

8.1.1 Infinitessimal probabilities

If both A and B concern quantities which could take on a continuous range of values (Section 4.1.2), some care must be taken in formulating Bayes's theorem correctly. If A is "The value of X is between x and $x + \mathrm{d}x$." and B is "The value of Y is between y and $y + \mathrm{d}y$." then Bayes's theorem becomes

$$\begin{aligned}&\mathbb{P}(x \leq X \leq x + \mathrm{d}x | y \leq Y \leq y + \mathrm{d}y, C)\\ &= \frac{\mathbb{P}(x \leq X \leq x + \mathrm{d}x|C)\,\mathbb{P}(y \leq Y \leq y + \mathrm{d}y | x \leq X \leq x + \mathrm{d}x, C)}{\mathbb{P}(y \leq Y \leq y + \mathrm{d}y|C)}.\end{aligned} \tag{83}$$

However, the statements about the value of X and Y that are being conditioned on do not need the range to be specified, as $\mathbb{P}(A|x) = \mathbb{P}(A|x + \mathrm{d}x)$ for any A, given infinitessimal $\mathrm{d}x$ and $\mathrm{d}y$. These values can be set exactly, leading to the slightly more compact

$$\mathbb{P}(x \leq X \leq x + \mathrm{d}x|Y = y, C) = \frac{\mathbb{P}(x \leq X \leq x + \mathrm{d}x|C)\,\mathbb{P}(y \leq Y \leq y + \mathrm{d}y|X = x, C)}{\mathbb{P}(y \leq Y \leq y + \mathrm{d}y|C)}. \tag{84}$$

Further using the relationship between infinitessimal probabilities and probability densities from Equation 12 then gives the standard result that

$$\mathbb{P}(X = x|Y = y, C) = \frac{\mathbb{P}(X = x|C)\,\mathbb{P}(Y = y|X = x, C)}{\mathbb{P}(Y = y|C)} \tag{85}$$

or, in the more concise notation, just

$$\mathbb{P}(x|y, C) = \frac{\mathbb{P}(x|C)\,\mathbb{P}(y|x, C)}{\mathbb{P}(y|C)}. \tag{86}$$

In most situations the meaning of this more compact form is sufficiently clear to warrant its use; but, as always, if there is any ambiguity it is important to be able to rephrase any such condensed expressions in terms of full logical propositions.

8.2 Bayesian data analysis

The most important application of Bayes's theorem is to use information gathered about the world, *i.e.*, the data, d, in combination with some assumed knowledge, $\mathcal{K}$, assess the likely truth of some hypothesis, H. All of d, $\mathcal{K}$ and H could be compound statements: the data d is often results of millions or billions of measurements; the background information $\mathcal{K}$ consists of a huge range of assumptions and the measurement process that yielded the data; and H could include statements about nuisance parameters, the true properties of population members, *etc.*The more complicated the d, $\mathcal{K}$ and H are, the more mathematically challenging it is to evaluate Bayes's theorem, but conceptual structure of Bayesian inference is always the same.

In this setting, Bayes's theorem (Equation 82) becomes

$$\underbrace{\mathbb{P}(\text{hypothesis} \mid \text{data},\ \text{bkg. info.})}_{\text{posterior}} = \frac{\overbrace{\mathbb{P}(\text{hypothesis} \mid \text{bkg. info.})}^{\text{prior}}\ \overbrace{\mathbb{P}(\text{data} \mid \text{hypothesis}, \text{bkg. info.})}^{\text{likelihood}}}{\underbrace{\mathbb{P}(\text{data} \mid \text{bkg. info.})}_{\text{marginal likelihood}}} \tag{87}$$

or, in more compact notation,

$$\underbrace{\mathbb{P}(H|d, \mathcal{K})}_{\text{posterior}} = \frac{\overbrace{\mathbb{P}(H|\mathcal{K})}^{\text{prior}}\ \overbrace{\mathbb{P}(d|H, \mathcal{K})}^{\text{likelihood}}}{\underbrace{\mathbb{P}(d|\mathcal{K})}_{\text{marginal likelihood}}}, \tag{88}$$

or, more qualitatively,

$$\text{posterior} \propto \text{prior} \times \text{likelihood}. \tag{89}$$

Writing Bayes's theorem in this context-specific form, the four constituent probabilities all have distinct interpretations: the prior probability (Section 8.2.1) expresses what was known before

considering the (new) data; the likelihood (Section 8.2.2) encodes what has been learned; the posterior (Section 8.2.3) expresses the totality of what is now known about they hypothesis; and the marginal likelihood (Section 8.2.4) provides information about the data is consistent with information that is being assumed.

The distinction between prior and posterior is a very useful device for framing simple problems, but it also has its limitations, and breaks down entirely in complicated situations in which several inference steps are required (*e.g.*, hierarchical models). Another potential difficulty arises in cases where the analysis is sequential, in which case what is a posterior probability at one step is a prior probability at a later step. In such situations it can actually be confusing to refer to prior and posterior probabilities, it being instead better to rely on explicit mathematical notation.

As written in Equation 88, Bayes's theorem has a clear asymmetry to it: the probabilities on the right-hand side are inputs; the probability on the left-hand side is the output. The product of the prior and the likelihood, is a mathematical rule for how new data can be combined into an existing understanding, *i.e.*, learning. As such, Equation 88 is – or at least should be – at the heart of all attempts to synthesize incomplete information into a model of reality. The parallels between this formal statistical formula and learning is strengthened further by looking separately at each of the four terms in Equation 88.

8.2.1 The prior probability

$\mathbb{P}(\text{hypothesis} \,|\, \text{bkg. info.}) = \mathbb{P}(H|\mathcal{K})$ is the probability that the hypothesis is true conditioned only on the background information. As such, it is referred to as the prior probability, as it the probability before the data has been considered.

The presence of the prior probability in Bayes's theorem can be a source of debate or concern, as it does not involve the data being considered; but is an inescapable consequence of the self-consistency conditions that lead to Bayes's theorem. It also reinforces the point, made in Section 4, that defining probability as a degree of implication necessarily means that all probabilities are conditional on some relevant information, and that Bayesian inference is a formalism for expanding on current (*i.e.*, prior) knowledge, rather than creating knowledge from scratch.

In general there is no way of calculating or deriving the prior probability for a hypothesis. In some cases symmetry arguments (*e.g.*, due to the arbitrary labelling of the six sides a die or the 52 cards in a pack) provide an argument for assigning prior probabilities as an expression of ignorance, an idea that is formalised most powerfully as the principle of maximum entropy. But in other common situations (*e.g.*, assessing whether a drug has an effect or an accused person is guilty of a crime) no such general argument is available.

There is, however, nothing to prevent Bayes's theorem being evaluated for several different values of the prior probability; if the data is sufficiently informative then Bayes's theorem might lead to the same broad conclusion (*e.g.*, the the drug has an effect or that the accused person is guilty) for all plausible values of $\mathbb{P}(H|\mathcal{K})$. Such investigations of prior sensitivity are often an important practical aspect of Bayesian inference.

8.2.2 The likelihood

$\mathbb{P}(\text{data} \,|\, \text{hypothesis}, \text{bkg. info.}) = \mathbb{P}(d|H, \mathcal{K})$ is the the probability that this data would have been obtained if the hypothesis (and the assumed background information) were true. This probability

is universally referred to as a the likelihood, although it is often treated as a function primarily of H as the data is fixed, whereas different potential hypotheses might be investigated. The likelihood provides the only link between the data and the hypothesis on the right-hand side of Equation 88, a notion which is sometimes referred to as the likelihood principle. The only way a hypothesis can be favoured by the data is if it predicts it well.

Conversely, if the likelihood is independent of H, so that $\mathbb{P}(d|H,\mathcal{K}) = \mathbb{P}(d|\mathcal{K})$, the implication is that this data provides no information (either way) about whether H is true. The posterior probability from Equation 88 hence simplifies to

$$\mathbb{P}(H|d,\mathcal{K}) = \frac{\mathbb{P}(H|\mathcal{K})\,\mathbb{P}(d|H,\mathcal{K})}{\mathbb{P}(d|\mathcal{K})} = \frac{\mathbb{P}(H|\mathcal{K})\,\mathbb{P}(d|\mathcal{K})}{\mathbb{P}(d|\mathcal{K})} = \mathbb{P}(H|\mathcal{K}), \tag{90}$$

which is equal to the prior probability. This is an important self-consistency check: if the data reveal nothing of relevance about H then one's state of knowledge is unchanged by including said data into the analysis.

In a practical data analysis setting the likelihood is a probabilistic model of the data-generation process, and so must incorporate any measurement noise, truncation, incomplete sampling, missing data, censoring, *etc.* The likelihood is often quite complicated mathematically, but it is also typically the least ambiguous of the three inputs into Equation 88 as the experimental model is the focus of the analysis. The likelihood can take a wide range of forms and learning to formulate the likelihood is best done by working through specific data analysis problems.

When considering the range of possible data that might have arisen from an equivalent measurement it useful to think of $\mathbb{P}(d|H,\mathcal{K})$ as the sampling distribution for hypothetical data. It is this distribution from which draws would be made if the experiment or measurement was being simulated.

8.2.3 The posterior probability

$\mathbb{P}(\text{hypothesis}\,|\,\text{data, bkg. info.}) = \mathbb{P}(H|d,\mathcal{K})$ is the termed the posterior because it the probability calculated after the data has been considered (along with the assumed background information). This term is in obvious contrast to the prior, and it is reasonable to think of Bayes's theorem as a way from going from the prior probability (*i.e.*, before the data have been considered) to the posterior probability (*i.e.*, after whatever information in the data has been incoroporated).

The interpretation of the posterior probability sometimes results in semantic debate, as it is usually equivalent to something like "There was an announcement of delays on the train, so I think there's a 90% chance I'll be late for my appointment." or "There's never been an instance of a person coming back from the dead, so the chance of a zombie apocalypse is tiny." But, really, these are no different from the probabilities discussed in Section 4: they are just statements of (partial) implication that are conditional on some assumed information. Indeed, perhaps the most important aspect of Equation 88 is that this conditioning information is made explicit: if someone disputes a posterior probability you have calculated – or even if you doubt your own result – the obvious explanation is that there is some extra information that was not included when applying Equation 88. (In the above examples: someone might have talked to a member of staff on the platform who had more up-to-date information about the trains, and so discount the automated announcement; a Christian who believes in the literal resurrection of Jesus might argue that there has been an instance of sombody returning to life after death.)

Formally, the posterior probability is the end point of a Bayesian calculation: all the relevant information has been combined according to the laws of probability that lead to Equation 88, and there is nothing more to be said. But in practice there is often more to be done, especially in an everyday setting where, even when reasoning in a broadly Bayesian way, it is implausible to have one's world model be an inordinate set of posterior probabilities. In many cases it is reasonable to approximate a high (posterior) probability as 1; this is what happens when, *e.g.*, someone goes from considering a proposition to de facto belief. Similarly, it is often practical to approximate a low (posterior) probability as 0; this is what happens when someone decides something (*e.g.*, the zombie apocalypse) is impossible.

While such internal assessments could always be revisted (if *e.g.*, a train unexpectedly runs as an express, or the dead rise up from their graves in search of blood), the same is not true once a decision has been made to act on the basis of a probability. The threshold for action cannot be calculated from probabilities alone – it depends on the cost of being wrong – and this is the subject of decision theory which can be seen as a practically-motivated extension of Bayesian inference.

8.2.4 The marginal likelihood

$\mathbb{P}(\text{data}|\text{bkg. info.}) = \mathbb{P}(d|\mathcal{K})$ is the probability that the observed data would have been obtained given only the assumed background information. It is referred to here as the marginal likelihood[12] because it can be re-written as a weighted sum of the likelihoods under two distinct hypotheses: H and $\bar{H}$. Applying the laws of probability derived in Section 5 gives

$$\begin{aligned}
\mathbb{P}(d|\mathcal{K}) &= \mathbb{P}(d, H \text{ or } \bar{H}|\mathcal{K}) \\
&= \mathbb{P}(d, H|\mathcal{K}) + \mathbb{P}(d, \bar{H}|\mathcal{K}) \\
&= \mathbb{P}(H|\mathcal{K})\,\mathbb{P}(d|H, \mathcal{K}) + \mathbb{P}(\bar{H}|\mathcal{K})\,\mathbb{P}(d|\bar{H}, \mathcal{K}) \\
&= \mathbb{P}(H|\mathcal{K})\,\mathbb{P}(d|H, \mathcal{K}) + [1 - \mathbb{P}(H|\mathcal{K})]\,\mathbb{P}(d|\bar{H}, \mathcal{K}),
\end{aligned} \tag{91}$$

where $\mathbb{P}(d|\bar{H}, \mathcal{K})$ is the likelihood under in the case that the hypothesis H is false. Thus $\mathbb{P}(d|\mathcal{K})$ can be seen as the summed probability of getting the data by the two possible routes (*i.e.*, that H is true or H is false). The reason for the terminology is that this is hence the likelihood marginalised over these two possibilities, a concept that is central to the manipulation of probability distributions (Section 6).

As is case with the likelihood (Section 8.2.2), this gives a second way of thinking about $\mathbb{P}(d|\mathcal{K})$, as the sampling distribution for the data under the assumed background information. The fact that this term appears in Equation 88 means that a requirement for performing Bayesian data analysis is a fully-specified algorithm for generating potential data conditional only on the background information that is not being questioned; if this is not present then there is a danger that the resultant inference is invalid.

While it may seem strange to deliberately replace a compact expression, in most cases this more explicit form is a necessary step to calculating $\mathbb{P}(d|\mathcal{K})$ and hence $\mathbb{P}(H|d, \mathcal{K})$. The posterior probability can now be re-written without explicit reference to $\mathbb{P}(d|\mathcal{K})$ as

$$\mathbb{P}(H|d, \mathcal{K}) = \frac{\mathbb{P}(H|\mathcal{K})\,\mathbb{P}(d|H, \mathcal{K})}{\mathbb{P}(H|\mathcal{K})\,\mathbb{P}(d|H, \mathcal{K}) + [1 - \mathbb{P}(H|\mathcal{K})]\,\mathbb{P}(d|\bar{H}, \mathcal{K})}, \tag{92}$$

[12]The marginal likelihood is also known as the "model-averaged likelihood" or, particularly in physics and astronomy, as the "(Bayesian) evidence".

which contains only one prior probability, $\mathbb{P}(H|\mathcal{K})$, along with two likelihoods $\mathbb{P}(d|H,\mathcal{K})$ and $\mathbb{P}(d|\bar{H},\mathcal{K})$ under what are, effectively, competing hypotheses.

Even though $\bar{H}$ is just the negation of H, the presence of $\mathbb{P}(d|\bar{H},\mathcal{K})$ in Equation 92 is of fundamental significance, as it is the reason that there can be no pure Bayesian hypothesis test: the posterior probability of an hypothesis H depends in part on the probability of obtaining the data in the situation that H is not true, which in turn implies the requirement of an alternative model that makes specific predictions for the data. Very often this is not available, even if H itself is well defined. For instance, if H is the hypothesis that a coin is fair then $\bar{H}$ is the the hypothesis that it is biased in some way; but this is a possibility that could take a wide range of different forms, from the coin being slightly-weighted to it having two heads or two tails.

8.3 Odds ratios and the Bayes factor

Bayes's theorem holds separately for different hypotheses, say H_1 and H_2, and so taking the ratio of Equation 88 evaluated for each implies that

$$\frac{\mathbb{P}(H_2|d,\mathcal{K})}{\mathbb{P}(H_1|d,\mathcal{K})} = \frac{\mathbb{P}(H_2|\mathcal{K})\,\mathbb{P}(d|H_2,\mathcal{K})}{\mathbb{P}(H_1|\mathcal{K})\,\mathbb{P}(d|H_1,\mathcal{K})}. \tag{93}$$

Mathematically, this equation is convenient as the marginal likelihood, $\mathbb{P}(d|\mathcal{K})$, cancels out, leaving only terms involving the two hypotheses. But, if only H_1 and H_2 are under consideration, it is also a useful relationship at a more conceptual level because of the link to (betting) odds. From the correspondence between odds and probabilities derived in Section 4.3, the (posterior) odds associated with the hypothesis H_1 can be written as

$$O = \frac{1}{\mathbb{P}(H_1|d,\mathcal{K})} - 1 = \frac{1-\mathbb{P}(H_1|d,\mathcal{K})}{\mathbb{P}(H_1|d,\mathcal{K})} = \frac{\mathbb{P}(H_2|d,\mathcal{K})}{\mathbb{P}(H_1|d,\mathcal{K})}, \tag{94}$$

where the final step invokes the fact that there are no other models. Comparing this to Equation 93, leads to a restatement of Bayes's theorem entirely in terms of odds as

$$\underbrace{\frac{\mathbb{P}(H_2|d,\mathcal{K})}{\mathbb{P}(H_1|d,\mathcal{K})}}_{\text{posterior odds}} = \underbrace{\frac{\mathbb{P}(H_2|\mathcal{K})}{\mathbb{P}(H_1|\mathcal{K})}}_{\text{prior odds}}\ \underbrace{\frac{\mathbb{P}(d|H_2,\mathcal{K})}{\mathbb{P}(d|H_1,\mathcal{K})}}_{\text{likelihood ratio}} = \underbrace{\frac{\mathbb{P}(H|\mathcal{K})}{\mathbb{P}(H|\mathcal{K})}}_{\text{prior odds}}\ \underbrace{B_{2,1}}_{\text{Bayes factor}}, \tag{95}$$

or, more qualitatively, as

$$\text{posterior odds} = \text{prior odds} \times \text{Bayes factor}. \tag{96}$$

The above equations implicitly define the Bayes factor, which is just the ratio of the likelihoods under any pair of models, so

$$B_{2,1} = \frac{\mathbb{P}(d|H_2,\mathcal{K})}{\mathbb{P}(d|H_1,\mathcal{K})}. \tag{97}$$

The utility of the Bayes factor is that it is a single number which encodes the full implications of the data for two competing hypotheses – but if there are more possible explanations being considered it is not so useful. The notational compactness does come at a price, however: the indexing can potentially be a little ambiguous if more than two hypotheses are under consideration – and even then it is important to be careful about the index ordering.

If there is no strong reason to prefer one hypothesis to the other a priori then it is reasonable to set the prior odds ratio to $\mathbb{P}(H_2|\mathcal{K})/\mathbb{P}(H_1|\mathcal{K}) = 1$, which then means that Equation 95 simplifies to

$$\frac{\mathbb{P}(H_2|d,\mathcal{K})}{\mathbb{P}(H_1|d,\mathcal{K})} = B_{2,1}. \tag{98}$$

So in the absence of any discriminating prior information the Bayes factor is a full summary of the inference calculation in the context of a pair of competing hypotheses.

Learning objectives

After studying this section of the module you should be able to:

- Understand Bayes's theorem as a consequence of basic self-consistency requirements for an inference scheme.
- Know Bayes's theorem and the interpretation of the terms (prior, likelihood, posterior, *etc.*) in a simple data analysis setting.
- Use Bayes's theorem to analyse data and calculate the probability that a hypothesis or model is correct.
- Separate out the relevant information into prior and data if appropriate.