Bayes factors

Bayes factors are somewhat essential to Bayesian statistics. **Tony O'Hagan** explains their basics.

Bayes factors are a fundamental part of the Bayesian approach to testing hypotheses. In frequentist statistics, hypothesis testing is a matter of choosing a test statistic and calculating the p-value. The p-value is regarded as a measure of the evidence in favour of (or against) a null hypothesis. In general, the smaller its value, the stronger the evidence against the null hypothesis, but more precise interpretation of p-values is a matter of convention. For instance, $p < 0.05$ is widely regarded as a yard-stick: any larger value of p is considered to be insufficiently strong evidence against the null hypothesis for it to be rejected.

In interpreting p-values, and in communicating their meaning to others, researchers frequently fall into the trap of saying that $p < 0.05$ means that there is a less than 5% chance that the null hypothesis is true. This is wrong. Unfortunately, p-values have no such direct interpretation as measures of evidence.

Bayesian hypothesis testing

It is only within the Bayesian framework that we can discuss the probability that a hypothesis is true. Indeed, the essence of a Bayesian hypothesis test is, quite simply, to compute and report this probability.

Given data x and a hypothesis H, we require the posterior probability $P(H \mid x)$ that H is true, and this is obtained using Bayes' theorem as:

$$P(H|x) = \frac{P(H)P(x|H)}{P(x)}$$

In this formula, $P(x \mid H)$ is the probability of observing the data x if H is true, while $P(x)$ is the unconditional probability of observing x. This is usually expressed as:

$$P(x) = P(H)P(x|H) + P(H^c)\,P(x|H^c)$$

where H^c denotes the complement of H (i.e., the hypothesis that is true if H is false). In the frequentist terminology, we could call H the null hypothesis, and H^c the alternative hypothesis.

In both these formulae, we have the prior probability $P(H)$ that H is true, and we also have $P(H^c) = 1 - P(H)$. Thus, the posterior probability $P(H \mid x)$ depends on three things: $P(x \mid H)$, $P(x \mid H^c)$ and $P(H)$, and the two formulae together seem to imply that it depends on these three things in complicated ways. The Bayes factor arises from a different way of writing these expressions, which clarifies how these three quantities influence the posterior probability.

Consider the posterior odds in favour of H, which we define to be

$$\frac{P(H|x)}{P(H^c|x)} = \frac{P(H|x)}{1 - P(H|x)}$$

As $P(H \mid x)$ increases, it is clear that this ratio increases (because the denominator decreases), so the odds value is another way of representing the strength of evidence in favour of H. In particular, if the posterior probability of H is $P(H \mid x) = 0.5$, then the odds in favour of H is 1, so odds of 1 corresponds to a situation where the evidence favours neither H nor H^c. Odds greater than 1 corresponds to the weight of posterior evidence being in favour of H, while odds less than 1 corresponds to the weight of evidence being against H. In simple terms, odds of z means that H is z times as likely as H^c.

There is a corresponding prior odds, $P(H)/P(H^c)$, and the key formula is:

$$\frac{P(H|x)}{P(H^c|x)} = \frac{P(H)}{P(H^c)} \times \frac{P(x|H)}{P(x|H^c)} \qquad (1)$$

The final factor on the right is the *Bayes factor*, $B_H(x)$. In words, this formula says that the posterior odds is equal to the prior odds multiplied by the Bayes factor.

If the Bayes factor is greater than 1, then the posterior odds will be larger than the prior odds, and so the posterior probability of H will be larger than its prior probability. Conversely, if $B_H(x) < 1$, then $P(H \mid x) < P(H)$. So, the Bayes factor says how the evidence in the data modifies the prior probability. The formula distinguishes two separate influences on the posterior probability of H: first, there is the prior probability, based on prior evidence; second, there is the Bayes factor, which measures the strength of the new evidence in the data, x.

Interpreting Bayes factors

The Bayes factor has a very clear interpretation as a measure of evidence in favour of the (null) hypothesis H. If $B_H(x) < 0.05$, then the posterior odds in favour of H will be less than a twentieth of the prior odds. This represents appreciable evidence against H. For instance, if you initially had negligible prior evidence for or against H, then the prior odds would be 1, and given a Bayes factor less than 0.05, the posterior odds will also be less than 0.05. We can easily convert odds back to probability by using the formula:

$$\text{Probability} = \frac{\text{Odds}}{1 + \text{Odds}}$$

Thus, posterior odds less than 0.05 converts to a posterior probability less than

$$\frac{0.05}{1.05} = 0.0476$$

With a Bayes factor below 0.05, then, there is only a small posterior probability that H is true, unless there was prior evidence in its favour.

To take another example, suppose that prior information suggested that H was twice as likely to be true as false, in which case the prior odds would be 2. Now suppose that we obtain data x such that the Bayes factor is $B_H(x) = 0.01$, which is clearly strong evidence against H. The posterior odds would now be only 0.02, and H is 50 times more likely to be false than to be true. This is the effect of the strong Bayes factor; the posterior probability is

$$\frac{0.02}{1.02}$$

which is also close to 0.02.

There are two points to be aware of in interpreting Bayes factors. The first is that the Bayes factor only measures evidence in the data, and must be interpreted relative to the prior evidence. A Bayes factor of, say, 0.01 does indeed represent strong evidence against H, but if the prior evidence was strongly in its favour,

The Bayes factor only measures evidence in the data, and must be interpreted relative to the prior evidence

for instance, by prior odds of 1000, then the posterior odds will still be in favour of H. This might be the case, for instance, if H were the null hypothesis that the subject in an experiment does not have extra-sensory perception (ESP). In this case, most people would demand extremely strong evidence in the data to overcome a strong prior belief that people do not have ESP.

The second point is that in practice the Bayes factor still makes use of prior information. To see why this is the case, we need to consider the distinction between simple and compound (or composite) hypotheses.

Deriving the Bayes factor

To calculate the Bayes factor, we need to find the probability of the data x under the hypothesis H and under its complement H^c. Consider the example of ESP. Suppose that the data consist of asking the subject to guess the colour (red or black) of a random sequence of 20 playing cards. The subject correctly guesses 15. The null hypothesis H of no ESP means that the probability of guessing correctly is 0.5 for each card. Hence, $P(H \mid x)$ is the probability that $x = 15$ when x has the binomial distribution Bi(20, 0.5), which is

$$\binom{20}{15} 0.5^{20} = 0.0418.$$

In frequentist theory, the p-value is the probability of getting 15 *or more* correct guesses, which is 0.0207 and is small enough, by convention, to imply appreciable evidence against H.

We have $P(x \mid H) = 0.0148$, which is the numerator of the Bayes factor, but what about the denominator, $P(x \mid H^c)$? To derive this, we need to know the probability of a correct guess under the alternative hypothesis, H^c, which is the hypothesis that the subject does have ESP. But that probability depends on how strong the ability is. If the subject had the power to correctly identify every card, then the probability of a correct guess would be 1. Any weaker ability would give a probability less than 1, but greater than 0.5. If we denote that probability by θ, then we can find

$$P(x|q) = \binom{20}{15} q^{15} (1 - q)^5$$

for any θ. But the denominator of the Bayes factor is not $P(x \mid \theta)$ for some specified θ, it is $P(x \mid H^c)$, where, in effect, we are only specifying that $\theta > 0.5$.

The denominator, $P(x \mid H^c)$, is an average of the values $P(x \mid \theta)$, averaged over the prior distribution of θ. Thus, to compute the Bayes factor we have to say how likely it would be, a priori, that θ would take any given value between 0.5 and 1, assuming that the subject does indeed have the power of ESP.

This is the second point mentioned above about Bayes factors. If H is a simple hypothesis, then it completely specifies the probability of x because it specifies the values of all the unknown parameters in the statistical model. However, in practice it is extremely rare for both H and H^c to be simple hypotheses. (In the example, H is simple but H^c is not.) When one or both of the hypotheses is compound (i.e., not simple), then the Bayes factor depends on the prior distributions for the unknown parameters.

The fact that Bayes factors depend on prior information is an important distinction between the Bayesian and frequentist approaches

Weak prior information

The fact that Bayes factors depend on prior information is an important distinction between the Bayesian and frequentist approaches to hypothesis testing, because p-values do not have this property. This is a standard feature of Bayesian statistics, and the requirement for prior distributions to be specified has been at the core of the debate between the two approaches for decades. One response by Bayesians has been to make use of prior distributions that represent, in some formal sense, very weak prior information. For instance, in the context of the ESP question, it can be argued that a uniform prior distribution (giving equal prior density to all values of θ between 0.5 and 1) represents an absence of prior information about how strong the subject's powers might be if H is false. See the box "Extra-sensory perception?" for the result of using the prior distribution.

There is some arbitrariness about how we try to represent weak information, but in most applications of Bayesian statistics, given enough data, the posterior conclusions will not be sensitive to which representation we use. Unfortunately, one area in which this argument does not work is in testing hypotheses. The posterior

Bayes factor

The Bayes factor in favour of H (and against its complement H^c), based on data x is

$$B_H(x) = \frac{P(x|H)}{P(x|H^c)}$$

Extra-sensory perception?

A subject is asked to guess the colour (red or black) of each of 20 randomly selected playing cards. She correctly guesses 15 colours. Does this level of success indicate that she has extra-sensory perception (ESP)?

Under the null hypothesis H that she does not have ESP, the probability of correctly identifying 15 out of 20 colours is $P(x \mid H) = 0.0148$. If she does have ESP (the alternative hypothesis H^c), then we suppose that she has a probability $\theta > 0.5$ of correctly guessing each card. Under the assumption that all values of θ from 0.5 to 1 are equally likely, the probability that she correctly identifies 15, given that she does have ESP, is $P(x \mid H^c) = 0.0940$. The Bayes factor in favour of the null hypothesis is therefore $B_H(x) = 0.0148/0.0940 = 0.16$. This is clearly not strong evidence against H, and would not be enough to persuade most people to believe in ESP. For instance, if you start with a prior belief that the subject does not have ESP is 100 times as likely as her having ESP, your prior odds is 100. Multiplying this by the Bayes factor of 0.16 gives a posterior odds of 16. So you still believe it is 16 times more likely that the subject does not have ESP than that she does.

If we had a different prior distribution for θ under the alternative hypothesis, we would get a different posterior odds. However, it can be shown that, based on the data of 15 correct answers out of 20 and the prior odds of 100 against ESP, then the posterior odds will always be at least 7.

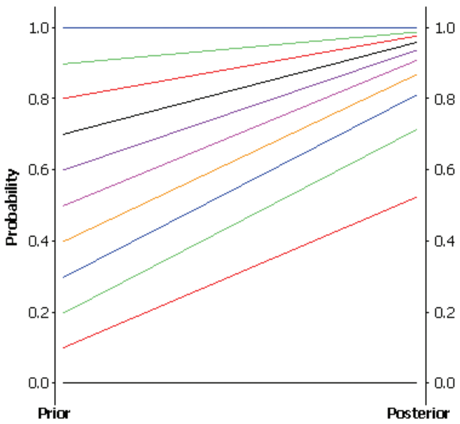

Figure 1. The effect of a Bayes factor of 10 in favour of *H*. Even with a prior probability of only 0.1, the posterior probability of *H* is better than 0.5

probability of *H* may remain sensitive to the choice of weak prior (and may even be formally undefined for most choices), regardless of how much data we have. A variety of more or less *ad hoc* variations on the Bayes factor have been proposed to address this difficulty, including the "intrinsic Bayes factor" and the "fractional Bayes factor". These are supposed to have the same interpretation as the full Bayes factor, but can be applied in situations of weak prior information; however, this is an area of on-going discussion within Bayesian statistics.

One advantage of the Bayesian approach is that it is possible to compare more than two hypotheses

Many hypotheses

The preceding discussion has been based entirely within the context of comparing two hypotheses, which we can characterise as a null hypothesis and an alternative hypothesis in line with frequentist terminology. However, one advantage of the Bayesian approach is that it is possible to compare more than two hypotheses. The key formula (**1**) generalises to

$$\frac{P(H_1|x)}{P(H_2|x)} = \frac{P(H_1)}{P(H_2)} \quad \frac{P(x|H_1)}{P(x|H_2)}$$

where H_1 and H_2 are any two hypotheses. We then call the second factor on the right-hand side of this formula, $P(x \mid H_1) = P(x \mid H_2)$ the Bayes factor for H_1 *versus* H_2 and denote it by $B_{12}(x)$. No matter what other hypotheses are under consideration, the ratio of posterior probabilities for H_1 and H_2 is their prior ratio multiplied by the Bayes factor $B_{12}(x)$.

Bayes factors are an important part of the applied Bayesian statistician's toolkit. Although it is not possible in an account such as this to go into technical detail about how they are derived or computed in practice, I have tried to explain what they mean, issues around their use and how they relate to frequentist *p*-values. As Bayesian methods are increasingly used in statistical applications, a basic understanding of how to interpret them should be a part of every scientist's toolkit.

Tony O'Hagan is Professor of Statistics at the University of Sheffield. His research is in the theory and application of Bayesian statistics.