3

Bayesian statistics: principles and benefits

Anthony O'Hagan[#]

Abstract

The purpose of this article is to present the basic principles of the Bayesian approach to statistics and to contrast it with the frequentist approach. In particular, I hope to demonstrate the advantages that the Bayesian approach has, of providing more intuitive and meaningful inferences, of answering complex questions cleanly and exactly, of making use of all available information, and of being particularly well suited for decision-making.

Introduction

Bayesian statistics is a hot topic today in numerous fields in which statistics is applied. The Frontis workshop at Wageningen entitled 'Bayesian Statistics' is part of this wider phenomenon. As someone who has been researching in Bayesian statistics for 30 years, and as a committed proponent of the Bayesian approach, it is a wonderful thing to see these methods so much in demand. For half of those 30 years, I had no opportunity to apply the Bayesian approach, partly because in those days we lacked the necessary computational tools to do proper applications, and partly because almost nobody outside the academic statistics community was aware of Bayesian statistics. Now I am engaged in serious practical applications of Bayesian methods in cost-effectiveness of medicines, terrestrial carbon dynamics, auditing, radiocarbon dating, setting water quality standards, monitoring environmental pollution, to name just my most active application areas. Colleagues in the Bayesian community engage in applications in a huge variety of fields.

Food production and food technology are no exceptions to this widespread phenomenon, as the Wageningen Frontis workshop clearly testifies. However, I must confess that I have no special expertise in these fields. This article will concentrate on general principles, while other authors will demonstrate the power of Bayesian methods in the food industry.

Bayesian statistics is named after the 18th century minister Thomas Bayes, whose paper presented to the Royal Society in 1763 first used the kind of arguments that we now call Bayesian. Although the Bayesian approach can be traced back to Thomas Bayes, its modern incarnation began only in the 1950s and 1960s. In the first half of the 20th century the dominant philosophy of statistics was the approach we now call frequentist. That dominance has been gradually eroded during the second half of the 20th century. Although the vast majority of statistical analysis in practice is still frequentist, Bayesian methods are now the tools of choice in many application areas.

[#] University of Sheffield, Department of Probability and Statistics, Hicks Building, Sheffield, S3 7RH, UK. E-mail: a.ohagan@sheffield.ac.uk

This article is in two parts. The first will describe the Bayesian paradigm for statistics, and the second will contrast this with the frequentist paradigm. A final section summarizes the benefits of the Bayesian approach.

Bayesian statistics

This section will describe how Bayesian statistics works in a relatively non-technical way. The reader desiring more detail should find (in increasing order of mathematical complexity) the books by Lee (1997), Congdon (2001), O'Hagan (1994) and Bernardo and Smith (1994) useful. Lee is written as an undergraduate text, and provides a basic introduction to Bayesian statistics without getting into a lot of technical detail. Congdon is primarily about applied Bayesian modelling, and concentrates on how to structure and solve applied statistical problems using Bayesian methods. O'Hagan is a postgraduate text, intended for doctoral students who already have a good background in mathematics and statistics, and also as a work of reference; a revised and much extended second edition will appear in 2004. Bernardo and Smith is more like a research monograph; one of its strengths is the extremely comprehensive references to the Bayesian statistics literature up to 1993.

The Bayesian method

In Bayesian statistics we

- create a statistical model to link data to parameters
- formulate prior information about parameters
- combine the two sources of information using Bayes' theorem
- use the resulting posterior distribution to derive inferences about parameters.

The first step is common to all formulations of statistics, including the frequentist approach. The remainder are uniquely Bayesian.

A key feature of Bayesian statistics is the third step, in which Bayes' theorem synthesizes the two separate sources of information - see Figure 1 for a schematic representation of this process. The result of combining the prior information and data in this way is the posterior distribution.

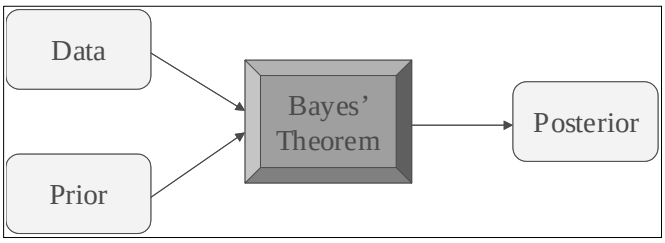

Figure 1. Synthesis of information by Bayes' theorem

A better illustration of how Bayes' theorem works is Figure 2. This is an example of a 'triplot', in which the prior distribution, likelihood and posterior distribution are all plotted on the same graph. In this example, the prior information (dashed line) tells us that the parameter is almost certain to lie between −4 and +4, that it is most likely to be between −2 and +2, and that our best estimate of it would be 0.

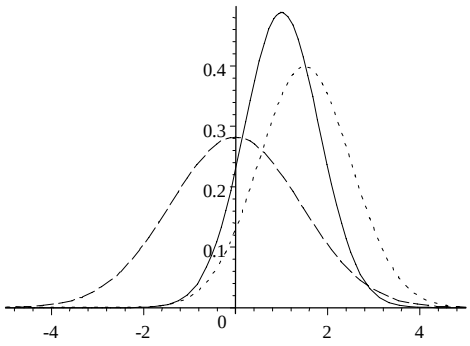

Figure 2. Triplot. Prior density (dashed), likelihood (dotted) and posterior density (solid)

The data (dotted line) favour values of the parameter between 0 and 3, and strongly argue against any value below −2 or above +4.

The posterior (solid line) puts these two sources of information together. So, for values below −2 the posterior density is tiny because the data are saying that these values are highly implausible. Values above +4 are ruled out by the prior; again, the posterior agrees. The data favour values around 1.5, while the prior prefers values around 0. The posterior listens to both and the synthesis is a compromise. After seeing the data, we now think the parameter is most likely to be around 1.

The strength of each source of information is indicated by the narrowness of its curve ─ a narrower curve rules out more parameter values, and so represents stronger information. In Figure 2 we see that the data (dotted) are a little more informative than the prior (dashed). Since Bayes' theorem recognizes the strength of each source, the posterior is influenced a little more by the data than by the prior. For instance, the posterior peaks at 1.0, a little closer to the peak of the data curve than to the prior peak. Notice also that the posterior is narrower than either the prior or the data curve, reflecting the way that the posterior has drawn strength from both of the other information sources.

The data curve is technically called the likelihood and is also important in frequentist inference. It derives from the statistical model that we have said is common to the two methods of inference. Its role in both inference paradigms is to describe the strength of support from the data for the various possible values of the parameter. The most obvious difference between frequentist and Bayesian methods is that frequentist statistics uses only the likelihood, whereas Bayesian statistics uses both the likelihood and the prior information.

In this example, the standard frequentist estimate of the parameter would be 1.5, which is the maximum likelihood estimate. The corresponding Bayesian estimate is 1.0. The difference is because the Bayesian makes use of additional information (the prior information) that suggests the parameter is more likely to be near 0 (or even negative).

Mathematical formulation

We can denote the prior distribution for the parameter θ by the function $p(\theta)$. Strictly, if θ is a discrete-valued parameter $p(\theta)$ is the probability that the parameter takes any particular value θ, while in the more usual case where θ is a continuous variable $p(\theta)$ is a probability-density function. Figure 2 represents such a situation for a continuous parameter, and so the dashed line is the prior density function.

The likelihood appears in both Bayesian and frequentist inference. It is the conditional probability distribution $f(x|\theta)$ of the data x when we suppose θ to be known. However, it is called the likelihood when we view $f(x|\theta)$ as a function of θ for fixed (observed) data x. Then we usually write it as $L(\theta;x)$. When we think of it as a function of θ in this way, it does not have to integrate (or sum) to 1 in the way that a probability distribution must. In fact, we are allowed to scale $f(x|\theta)$ by an arbitrary constant multiplier in order to define $L(\theta;x)$. We denote this by using the proportionality symbol: $L(\theta;x) \propto f(x|\theta)$. In the triplot, we scale $L(\theta;x)$ to integrate to 1, because this aids comparison with the other two curves (which, being probability distributions, must integrate to 1). This is the dotted line in Figure 2.

Bayes' theorem synthesizes the two sources of information by the simple process of multiplying. At every value of θ we multiply $p(\theta)$ by $L(\theta;x)$. However, the posterior distribution must integrate (or sum, if θ is discrete) to 1, so after multiplying in this way we scale the product so that it does integrate to 1. The result is the posterior distribution, which we will denote by $p^*(\theta|x)$. Thus,

$$p^*(\theta|x) \propto p(\theta)L(\theta;x) \qquad (1)$$

In words,

the posterior is proportional to prior times likelihood

It is easy to see in Figure 2 how the solid line results from multiplying together the other two lines and rescaling in this way. (It should also be clear that no matter how we scale the likelihood, which is arbitrary, the process of scaling the posterior distribution at the end produces a unique result)

Example

The following example concerns a new treatment protocol for a certain form of cancer. This is a severe form of cancer. With current treatment, just 40% of patients survive 6 months. Prior information suggests the new protocol might be expected to improve survival slightly. The doctor who is leading this new treatment gives a prior estimate of 45% for the 6-month survival rate of the new treatment protocol. She expresses her uncertainty in a standard deviation of 7%. After one year of using the new treatment protocol, 15 patients have been observed for 6 months, and 6 have survived (i.e. 40%).

Figure 3 shows the triplot for this analysis. The doctor's posterior estimate (the mean of her posterior distribution) is 0.438 (43.8%), which is a compromise between the prior estimate of 45% and the data estimate of 40%. It is closer to the prior in this case because the prior information is more informative than the small sample of just 15 patients.

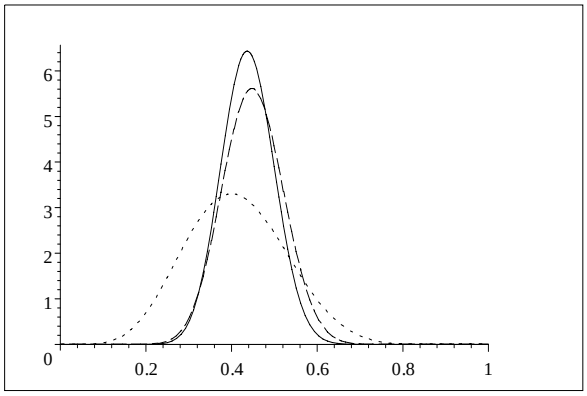

Figure 3. Triplot for cancer example, first year's data

The doctor will have been disappointed with these early results, since the data suggest only the same success rate as the current treatment. However, prior information leads to a higher estimate than the frequentist analysis would have done on the basis of the data alone. The doctor should not be too discouraged by this small sample, and the Bayesian analysis still gives a 0.73 probability that the new treatment is superior, almost the same as the prior probability.

After two years, a total of 70 patients have been followed up for 6 months, and we now have 34 survivors (i.e. 48.6%). This is a better result for the doctor; see Figure 4. Her posterior estimate is now 0.471 (47.1%). This figure still represents a compromise between the prior mean (45%) and data estimate (48.6%), but is now closer to the data because they are now more informative than the prior.

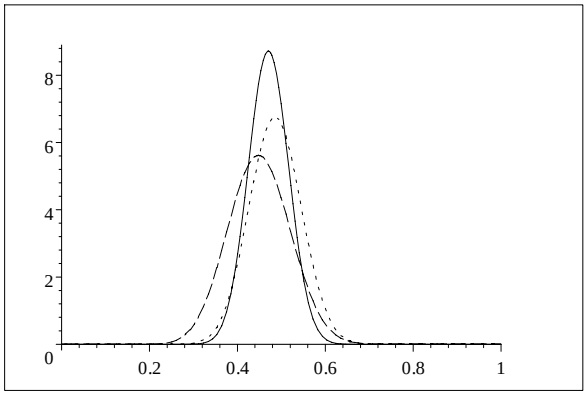

Figure 4. Triplot for cancer example, two years' data

Notice that the prior is now exerting a moderating influence on the data. The sample is still not large, and the Bayesian analysis is realistically cautious about an unexpectedly good outcome from the trial. However, the posterior probability that the new treatment is more effective than the old is 0.941, so the doctor can now be quite confident of achieving some improvement.

We have seen in this example that the prior distribution moderates the information in the data, leading to less pessimism when the data are unexpectedly bad and less optimism when they are unexpectedly good. Both influences are beneficial. In medical research, for instance, a large number of chemicals are investigated in the

hope of finding some effects. Some perform very well in early trials, and it is easy to become excited by these based on the frequentist analysis which uses the data alone. Bayesian analysis starts from the prior expectation that any given chemical is very unlikely to be effective, and so moderates this enthusiasm. Experience justifies this analysis, since many chemicals that perform well at an early stage do not fulfil that promise when subsequently tested in larger confirmatory trials.

On the other hand, pharmaceutical companies are regularly forced to abandon drugs that have just failed to demonstrate beneficial effects in frequentist terms, when a Bayesian analysis suggests that it would be worth persevering.

Subjectivity

As I have said several times, the main distinction between frequentist and Bayesian methods is that the Bayesian approach uses prior information about the parameters. This is a source of strength and also a source of controversy.

Controversy arises because prior information is in principle subjective. My prior information in any problem is different from yours, so we will have different prior distributions, and end up with different posterior distributions. In this sense, it is claimed that the whole Bayesian approach is subjective. Indeed, as we shall see in the next section, Bayesian methods are based on a subjective interpretation of probability. For many scientists trained to reject subjectivity whenever possible in their work, this is too high a price to pay for the benefits of Bayesian methods. To many of its critics, subjectivity is the key drawback of the Bayesian approach.

I believe that this objection is unwarranted, both in principle and in practice. It is an unwarranted objection in principle because science cannot be truly objective. In practice, the Bayesian method actually reflects very closely the real nature of the scientific method, in the following respects:

- Subjectivity in the prior distribution is minimized through basing prior information on defensible evidence and reasoning.
- Through the accumulation of data, differences in prior positions are resolved and consensus is reached.

Taking the second of these points first, Bayes' theorem weights the prior information and data according to their relative strengths, in order to derive the posterior distribution. If prior information is vague and insubstantial, then it will get negligible weight in the synthesis with the data, and the posterior will in effect be based entirely on data information (as expressed in the likelihood function). Similarly, as we acquire more and more data, the weight that Bayes' theorem attaches to the newly acquired data relative to the prior increases, and again the posterior is effectively based entirely on the information in the data. This feature of Bayes' theorem mirrors the process of science, where the accumulation of objective evidence is the primary process whereby differences of opinion are resolved. Once the data provide conclusive evidence, there is essentially no room left for subjective opinion.

In principle, subjectivity is actually a strength of the Bayesian method because it allows us to examine the range of posterior distributions that might be held by different informed observers. If individual i holds a prior distribution $p_i(\theta)$ and through observing the data x obtains the posterior distribution $p_i^*(\theta|x)$, we can consider the range of posterior distributions for all interested individuals. The second bullet point above is a statement of some rather deep mathematical theory which says that (subject to reasonable conditions, like all the individuals agreeing on the

likelihood, none of them being so bigoted that their prior distribution $p_i(\theta)$ gives zero prior probability to genuinely possible values of θ, and the data not being deficient in some important way) as we get more data the differences in their posterior distributions will disappear.

The point is that in a Bayesian framework we can actually test whether sufficient convergence of opinions has occurred, by trying out different prior distributions that might, in some sense, represent the extremes of legitimate prior opinion for well-informed individuals. If important differences in posterior distributions remain, then the data are not strong enough to resolve the individuals' differences in opinion, and this is an important conclusion in itself.

Formulating prior distributions

The first bullet point above says that where genuine, substantial prior information exists it needs to be based on defensible evidence and reasoning. This is clearly important when the new data are not so extensive as to overwhelm the prior information, so that Bayes' theorem will give the prior a non-negligible weight in its synthesis with the data. Prior information of this kind exists routinely in all kinds of applications. The most important consideration in the use of prior information is to ensure that the prior distribution honestly reflects genuine information, not personal bias, prejudice, superstition or other such factors that are justly condemned in science as 'subjectivity'.

Prior information should be based on sound evidence and reasoned judgements. A good way to think of this is to parody a familiar quotation: the prior distribution should be 'the evidence, the whole evidence and nothing but the evidence':

- 'the evidence' – genuine information legitimately interpreted;
- 'the whole evidence' – not omitting relevant information (preferably a consensus that pools the knowledge of a range of experts);
- 'nothing but the evidence' – not contaminated by bias or prejudice.

Of course, this may be good advice but it does not tell the reader how to formulate a prior distribution in practice. To give more than a brief indication of some of the main considerations is beyond the scope of this article. I will simply outline some of the more important approaches.

Conventional weak prior distributions

In response to the criticism of subjectivity, some have proposed conventional solutions that are supposed to represent a situation where there is no prior information. For instance, suppose that the parameter θ is a proportion that must lie in the range 0 to 1. Then an obvious way to express ignorance about θ is through the uniform prior density $p(\theta) = 1$ (for all θ in $[0,1]$), which gives equal prior density to all possible values of θ. Uniform prior belief is attractive when we consider the operation of Bayes' theorem through Equation (1), since then the posterior distribution will just be the likelihood function scaled to integrate to 1. So we can view this as an analysis in which we use only the data 'uncontaminated' by prior opinions.

Unfortunately there is a fatal flaw in this approach. If I am ignorant about θ then I am also ignorant about $\phi = \theta^2$. Now ϕ also lies in $[0,1]$, and so it seems obvious to put a uniform prior distribution on it, yet this uniform distribution is theoretically

incompatible with the uniform distribution for θ. If ϕ has a uniform distribution, then this implies that θ has the density $p(\theta) = 2\theta$, which is not uniform.

There have been numerous attempts to find a formula for representing prior ignorance, but without any consensus. Indeed, it is almost certainly an impossible quest. Nevertheless, the various representations that have been derived can be useful, at least for representing relatively weak prior information.

When the new data are strong (relative to the prior information), the prior information is not expected to make any appreciable contribution to the posterior. In this situation, it is pointless to spend much effort on carefully eliciting the available prior information. Instead, it is common in such a case to apply some conventional 'non-informative', 'default', 'reference', 'improper', 'vague', 'weak' or 'ignorance' prior. These terms are used more or less interchangeably in Bayesian statistics to denote a prior distribution representing very weak prior information. The use of the term 'improper' is because technically most of these distributions do not actually exist, in the sense that a normal distribution with an infinite variance does not exist.

Genuinely informative prior distributions

To gain the full benefit of the Bayesian approach, genuine prior information should not be ignored. However, the process of converting that information into its formal expression as a prior distribution is not straightforward. A particular difficulty arises when, as is often the case, the person who has useful prior information about the parameters in some problem is not a statistician but a specialist in the application area. Formulating a prior distribution then typically involves a dialogue between the expert and a statistically trained facilitator. This process is called 'elicitation', and has been studied both by statisticians and psychologists. Some useful references are Kadane and Wolfson (1996) and O'Hagan (1998). A fascinating collection of essays ranging from the discursive to the technical will be found in Wright and Ayton (1994).

Expressions of prior belief in relationships between parameters

In practice, most serious applications involve many parameters. Typically, genuine prior information relates only to a few parameters, and it is normal then to employ conventional weak prior formulations for the others. Often, the most useful information concerns relationships between parameters. For instance, when considering an experiment to compare the toxicity of two formulations of a pesticide, we may not have much prior idea about the toxicity of either formulation, but we would expect the two formulations to have similar toxicities. We can express this as genuine prior information about their difference or their ratio.

For example, if the two toxicity measures are θ_1 and θ_2 (on a suitable scale), and if we believe that the difference $\theta_1 - \theta_2$ will be normally distributed with mean 0 and variance 1, then the joint prior distribution is

$$p(\theta_1, \theta_2) \propto \exp\{-0.5(\theta_1 - \theta_2)^2\} \qquad (2)$$

Notice two things in this formula. The first is that the prior density is expressed using proportionality. This allows us to simplify the mathematics by, for instance, dropping the constant $\frac{1}{\sqrt{2\pi}}$ that should go with the normal density – any arbitrariness in scaling the prior will wash out when the posterior is scaled, in exactly the same way as arbitrariness in scaling the likelihood disappears. But the second point is that this is now asserted to be the joint density of θ_1 and θ_2, and as such it should integrate to 1 when we integrate with respect to both θ_1 and θ_2 from $-\infty$ to ∞. If we first integrate

with respect to θ_1 the resulting marginal distribution of θ_2 is seen to be uniform, but then when we try to integrate with respect to θ_2 the result is infinite, and we cannot scale this back to 1.

The point is that Equation (2), as a distribution for both θ_1 and θ_2 , is improper. In effect, we have only supplied information about $\theta_1 - \theta_2$, and this is not enough to constitute proper prior information about (θ_1, θ_2). This illustrates how we often formulate prior information about only a subset of the parameters, or on relationships between some parameters, and leave other aspects of the prior distribution 'vague' or 'uninformative'.

If the two pesticide formulations differed mainly in the concentration of the active ingredient, then we would have prior information to suggest that the higher concentration is likely to be the more toxic. This might be expressed through a non-zero mean for the difference, but we can still leave other aspects of prior information 'uninformative'

Such 'structural' prior information is often formulated in a hierarchical model. More details about hierarchical modelling can be found in any of the textbooks listed at the beginning of this section.

Computation

A major reason for the surge in interest in Bayesian statistics is the availability of powerful computational tools. Software is essential for any but the simplest of statistical techniques, and Bayesian methods are no exception. In Bayesian statistics, the key operations are to implement Bayes' theorem and then to derive relevant inferences or decisions from the posterior distribution. In very simple problems, these tasks can be done algebraically, but this is not possible in even moderately complex problems.

Until the 1990s, Bayesian methods might have been interesting, but they found little practical application because the necessary computational tools and software had not been developed. Anyone who wanted to do any serious statistical analysis had no alternative but to use frequentist methods. In little over a decade that position has been turned round dramatically. Computing tools were developed specifically for Bayesian analysis that are more powerful than anything available for frequentist methods, in the sense that Bayesians can now tackle enormously intricate problems that frequentist methods cannot begin to address. It is still true that Bayesian methods are more complex and that, although the computational techniques are well understood in academic circles, there is still a dearth of user-friendly software for the general practitioner.

The transformation is continuing, and computational developments are shifting the balance consistently in favour of Bayesian methods. The main tool is a simulation technique called Markov Chain Monte Carlo, universally abbreviated to MCMC. The idea of MCMC is in a sense to by-pass the mathematical operations rather than to implement them. Bayesian inference is solved by randomly drawing a very large simulated sample from the posterior distribution. The point is that if we have a sufficiently large sample from any distribution then we effectively have that whole distribution in front of us, and anything we want to know about the distribution we can calculate from the sample. So, for instance, if we wish to know the posterior mean we just calculate the mean of this 'inferential sample'. If the sample is big enough, the

sample mean is an extremely accurate approximation to the true distribution mean, such that we can ignore any discrepancy between the two.

The availability of computational techniques like MCMC makes exact Bayesian inferences possible even in very complex models. Generalized linear models, for example, can be analysed exactly by Bayesian methods, whereas frequentist methods rely on approximations. In fact, Bayesian modelling in seriously complex problems freely combines components of different sorts of modelling approaches with structural prior information, unconstrained by whether such model combinations have ever been studied or analysed before. The statistician is free to model the data and other available information in whatever way seems most realistic. No matter how messy the resulting model, the posterior inferences can be computed (in principle, at least) by MCMC.

Bayesian methods have become the only feasible tools in several fields, such as image analysis, spatial epidemiology and genetic pedigree analysis.

Although there is a growing range of software available to assist with Bayesian analysis, much of it is still quite specialized and not very useful for the average analyst. Unfortunately, there is nothing yet that is both powerful and user-friendly in the way that most people expect of statistical packages. Two particular software packages that are in general use, freely available and worth mentioning are First Bayes and WinBUGS.

First Bayes is a very simple programme that is aimed at helping the beginner to learn and understand how Bayesian methods work. It is not intended for serious analysis of data, nor does it claim to teach Bayesian statistics in itself, but it is in use in several universities worldwide to support courses in Bayesian statistics; it can be very useful in conjunction with a textbook such as those recommended at the beginning of this section. First Bayes can be freely downloaded from http://www.shef.ac.uk/~st1ao/.

WinBUGS is a powerful programme for carrying out MCMC computations and is in widespread use for serious Bayesian analysis. WinBUGS has been a major contributory factor to the growth of Bayesian applications and can be freely downloaded from http://www.mrc-bsu.cam.ac.uk/bugs/. It is worth noting, however, that WinBUGS is currently not very user-friendly and sometimes crashes with inexplicable error messages.

Given the growing popularity of Bayesian methods, it is likely that more powerful, robust and friendly commercial software will emerge in the coming years.

Contrasting Bayesian and frequentist inference

The basic differences between Bayesian and frequentist inference are fundamental, and concern the nature of probability, the nature of parameters and the nature of inferences.

The nature of probability

Frequentist and Bayesian methods are founded on different notions of probability. According to frequentist theory, only repeatable events have probabilities, whereas in the Bayesian framework probability simply describes uncertainty. The term “uncertainty” is to be interpreted very widely. An event can be uncertain by virtue of being intrinsically unpredictable, because it is subject to random variability, for example the concentration of mercury in a (randomly selected) tin of tuna. It can also

be uncertain simply because we have imperfect knowledge of it, for example the mean mercury concentration over all tins of tuna of a given brand. Only the first kind of uncertainty is acknowledged in frequentist statistics, whereas the Bayesian approach encompasses both kinds of uncertainty equally well.

Example. *Suppose that Mary has tossed a coin and that she knows the outcome, Head or Tail, but has not revealed it to Jamal. What probability should Jamal give to it being Head? When asked this question, most people say that the chances are 50-50, i.e. that the probability is one-half. This accords with the Bayesian view of probability, in which the outcome of the toss is uncertain for Jamal and so he can legitimately express that uncertainty by a probability. From the frequentist perspective, however, the coin is either Head or Tail and this is not a random event. For the frequentist it is no more meaningful for Jamal to give the event a probability than for Mary, who knows the outcome and is not uncertain. The Bayesian approach clearly distinguishes between Mary's knowledge and Jamal's.*

In frequentist statistics, a probability can only be a long-run limiting relative frequency. This is the familiar definition used in elementary courses, often motivated by ideas like tossing coins a very large number of times and looking at the long-run limiting frequency of 'heads'. It is because it is based firmly on this frequency definition of probability that we call the traditional statistical theory 'frequentist'. Bayesian statistics, in contrast, rests on an interpretation of probability as a personal degree of belief. Although to some this may seem 'woolly' and unscientific, it is important to recognize that Bayesian statisticians have developed analyses based on this interpretation very widely and successfully. As discussed above in the context of prior information, it does not lead to unbridled subjectivity and unscientific practices.

The frequency definition can be applied to measure probabilities where the uncertainty is due to random variation, whereas the Bayesian notion of personal probability is applicable to both kinds of uncertainty.

The nature of parameters

Statistical methods are generally formulated as making inferences about unknown parameters. The parameters represent things that are unknown, and can usually be thought of as properties of the population from which the data arise. An example is the mean mercury concentration in tins of tuna, or the probability of transmission of foot-and-mouth disease between sheep over a distance of 100m. Any question of interest can then be expressed as a question about the unknown values of these parameters. The reason why the difference between the frequentist and Bayesian notions of probability is so important is that it has a fundamental implication for how we think about parameters. Parameters are specific to the problem and are not generally subject to random variability. They are uncertain only because of lack of knowledge. Therefore frequentist statistics does not recognize parameters as being random, and so it does not regard probability statements about them as meaningful. In contrast, from the Bayesian perspective it is perfectly legitimate to make probability statements about parameters, simply because they are unknown.

Note that in Bayesian statistics, as a matter of convenient terminology, we refer to any uncertain quantity as a random variable, even when its uncertainty is not due to randomness but to imperfect knowledge.

Example. *Consider the proposition that a genetically modified form of some crop will produce larger yields than the standard variety in some region. This proposition concerns unknown parameters, such as each variety's mean yield if planted in that region. From the Bayesian perspective, since we are uncertain about whether this proposition is true, the uncertainty is described by a probability. Indeed, the result of a Bayesian analysis of the question can be simply to calculate the probability that the GM variety produces a larger mean yield. From the frequentist perspective, however, whether the GM crop produces a larger mean yield is a one-off proposition referring to two specific varieties in a specific context. It is not repeatable and so we cannot talk about its probability.*

The nature of inferences

In this last example the frequentist can conduct a significance test of the null hypothesis that the GM variety gives a higher mean yield, and thereby obtain a P-value. Suppose that the null hypothesis is rejected at the 5% level, i.e. P = 0.05. What does this mean? At this point, the reader should examine carefully the statements below, and decide which ones are correct.

1. Only 5% of fields would give a larger yield if planted with the GM crop.
2. If we were to repeat the analysis many times, using new data each time, and if the null hypothesis were really true, then on only 5% of those occasions would we (falsely) reject it.
3. There is only a 5% chance that the null hypothesis is true.

Statement 3 is how a P-value is commonly interpreted; yet this interpretation is not correct because it makes a probability statement about the hypothesis, which is a Bayesian not a frequentist concept. Hypotheses do not have probabilities in a frequentist analysis any more than parameters do. The correct interpretation of the P-value is much more tortuous, and is given by Statement 2. (Statement 1 is another fairly common misinterpretation. Since the hypothesis is about mean yield, it says nothing at all about individual fields.)

Probabilities in the frequentist approach must be based on repetition. Statement 2 is the correct interpretation of the P-value because it refers to repetition of the experiment.

To interpret a P-value as the probability that the null hypothesis is true is not only wrong but also dangerously wrong. The danger arises because this interpretation ignores how plausible the hypothesis might have been in the first place. Here are two examples.

Example 1*: screening. Consider a screening test for a rare disease. The test is very accurate, with false-positive and false-negative rates of 0.1% (i.e. only one person in a thousand who does not have the disease will give a positive result, and only one person in a thousand with the disease will give a negative result). You take the screen and your result is positive – what should you think? Since the screen only makes one mistake in a thousand, doesn't this mean you are 99.9% certain to have the disease? In hypothesis-testing terms, the positive result would allow you to reject the null hypothesis that you don't have the disease at the 0.1% level of significance, a highly significant result agreeing with that 99.9% diagnosis. But the disease is rare, and in practice we know that most positives reporting for further tests will be false positives. If only one person in 50,000 has this disease, your probability of having it after a positive screening test is less than 1 in 50.*

Although this example may not be obviously concerned with hypothesis testing, in fact there is a direct analogy. We can consider using the observation of a positive screening outcome as data with which to test the null hypothesis that you do not have the disease. If the null hypothesis is true, then the observation is extremely unlikely, and we could formally reject the null hypothesis with a P-value of 0.001. Yet the actual probability of the null hypothesis is more than 0.98. This is a dramatic example of the probability of the hypothesis (> 0.98) being completely different from the P-value (0.001). The difference clearly arises because the null hypothesis of you having the disease begins with such a low prior probability. Nobody who is familiar with the nature of screening tests would be likely to make the mistake of interpreting the false positive rate as the probability of having the disease (but it is important to make the distinction clear to patients!). By the same token, it is wrong to interpret a P-value as the probability of the null hypothesis, because this fails to take account of the prior probability in exactly the same way.

Example 2*: thought transference. An experiment is conducted to see whether thoughts can be transmitted from one subject to another. Subject A is presented with a shuffled deck of cards, and tries to communicate to Subject B by thought alone whether each card is red or black. In the experiment, subject B correctly gives the colour of 33 cards. The null hypothesis is that no thought-transference takes place and Subject B is just guessing randomly. The observation of 33 correct is significant with a (one-sided) P-value of 3.5%. Should we now believe that it is 96.5% certain that Subject A can transmit her thoughts to subject B?*

Most scientists would regard thought-transference as highly implausible, and in no way would be persuaded by a single, rather small, experiment of this kind. After seeing this experimental result, most would still believe quite strongly in the null hypothesis, and would regard the outcome as due to chance.

Both of these examples concern situations where the null hypothesis is highly plausible before we obtain the data. In practice, frequentist statisticians recognize that much stronger evidence would be required to reject a highly plausible null hypothesis such as these, than to reject a more doubtful null hypothesis. Conversely, if we expect the null hypothesis to be false, then it should not take much evidence to confirm this suspicion. This makes it clear that the P-value cannot mean the same thing in all situations, and to interpret it as the probability of the null hypothesis is not only wrong but could be seriously wrong when the hypothesis is a priori highly plausible (or highly implausible).

To many practitioners, and even to many practising statisticians, it is perplexing that one cannot interpret a P-value as the probability that the null hypothesis is true, and similarly that one cannot interpret a 95% confidence interval for a treatment difference as saying that the true difference has a 95% chance of lying in this interval. Nevertheless, these are wrong interpretations, and can be seriously wrong.

Analogous arguments apply to confidence intervals. The statement that [3.5, 11.6] is a 95% confidence interval for a parameter μ says that if we repeated this experiment a great many times, and if we calculated an interval each time using the rule we used this time to get the interval [3.5, 11.6], then 95% of those intervals would contain μ. The 95% probability is a property of the rule that was used to create the interval, not of the interval itself. It is simply not allowed, and would be wrong, to attribute that probability to the actual interval [3.5, 11.6]. This is again a very unintuitive argument and, just as with P-values, confidence interval statements are widely misinterpreted,

even by trained statisticians. The erroneous interpretation of the confidence interval, that there is a 95% chance of the parameter μ lying in the particular interval [3.5, 11.6], is almost universally made by statistician and non-statistician alike, but it is nevertheless incorrect.

In contrast, Bayesian inferences have exactly the desired interpretations. A Bayesian analysis of a hypothesis results precisely in the probability that it is true, and a Bayesian 95% interval for a parameter means precisely that there is a 95% probability that the parameter lies in that interval.

Benefits of the Bayesian approach

It is now possible to identify clearly the benefits of the Bayesian approach to statistics.

- Bayesian statistics provides more intuitive and meaningful inferences. For example, as discussed above, a frequentist P-value has a convoluted interpretation that does not actually say how likely the null hypothesis is on the basis of the evidence. A Bayesian analysis gives the more direct, intuitive and meaningful statement of the probability that the hypothesis is true.
- Bayesian methods can answer complex questions cleanly and exactly. As explained in the discussion of computation, the computing tools now available for Bayesian statistics allow us to tackle enormously more complex problems. Bayesian inferences can actually answer the investigator's questions, and the answers are computed essentially exactly.
- Bayesian methods make use of all available information. This is simply a reference to the fact that the Bayesian approach includes the prior information. Since the prior information should represent all the available knowledge apart from the data themselves, no relevant information is omitted in a Bayesian analysis.
- Bayesian techniques are particularly well suited for decision-making. I have not discussed this directly in the preceding sections, but it follows from the nature of probability and parameters. What makes decisions hard is uncertainty. There is uncertainty about the consequences of any given decision, due to lack of knowledge about some relevant facts or parameters. Bayesian methods can quantify those uncertainties using personal probability. Indeed, very often we explicitly derive a posterior distribution for unknown parameters based on the available evidence. This quantification of the uncertainties in a decision is a crucial component of rational, evidence-based decision-making.

Of course, the arguments in this long-running philosophical dispute about the foundations of statistics are not entirely one-sided. Bayesian methods are widely criticized for involving an element of subjectivity that is not overtly present in frequentist methods. I have argued that this criticism is misguided, but it is very persistent. On a more practical note, it is true that the prior distribution is difficult to specify reliably, that Bayesian methods are more complex than frequentist methods, and software to implement them is scarce or non-existent. These are genuine criticisms, but they are being addressed by more research into specification of prior distributions and by the development of more user-friendly tutorial and computing resources.

I look forward to Bayesian statistics being adopted strongly in the field of food science, as it is being in so many other areas, to the great benefit of research in that field.

References

Bernardo, J.M. and Smith, A.F.M., 1994. *Bayesian theory*. Wiley, New York.

Congdon, P., 2001. *Bayesian statistical modelling*. Wiley, Chichester. Wiley series in probability and statistics.

Kadane, J.B. and Wolfson, L.J., 1996. Experiences in elicitation. *The Statistician,* 47, 1-20.

Lee, P.M., 1997. *Bayesian statistics: an introduction*. 2nd edn. Arnold, London.

O'Hagan, A., 1994. *Kendall's advanced theory of statistics. Vol. 2B. Bayesian inference*. Arnold, London.

O'Hagan, A., 1998. Eliciting expert beliefs in substantial practical applications. *The Statistician,* 47, 21-35.

Wright, G. and Ayton, P. (eds.), 1994. *Subjective probability*. Wiley, London.