Bayesian Inference: Introduction

Preamble: Probability and Uncertainty

In Bayesian statistics, we *use the calculus of probability to represent uncertainty.* Sometimes people (including possibly me!) will say that Bayesian statistics treats things about which we are uncertain as "random variables". This is because in mathematics, we usually use the calculus of probability for random variables. However, uncertainty and randomness are different: one can be uncertain about something that is not random, at least not in the traditional sense of the word random. For example, what is the population of the US to the nearest 1 million? Some of you may know this, and others may not. Those who do not know it are uncertain about it. The Bayesian view is that we can use probability to describe and communicate this uncertainty.

Another example: will it rain tomorrow in Chicago? Presumably you are all somewhat uncertain about this. But whether or not rain in Chicago is a "random" event seems like, perhaps, a tricky philosophical question. Fortunately, if we embrace the use of probability to describe uncertainty, we do not have to concern ourselves with whether or not it is actually random: we can use probability to represent uncertainty in either case. Indeed, you have probably all encountered probabilistic weather forecasts in your day-to-day life. Did you every worry about whether weather is truly random? Probably not. Did you find the use of probability to represent the uncertainty difficult to get to grips with? Again, probably not: even people unfamiliar with Bayesian statistics, or indeed random variables, find the use of probability to represent uncertainty in the weather both natural and useful. (However, we will later discuss in a little more detail what one might mean when one says that the probability of rain tomorrow is, say, 20%.)

In statistics we often come across the need to perform inference from data. Of course the need to perform statistical inference necessarily implies uncertainty about the answer to a question (at least before viewing the data). From the discussion above, you should not be suprised to learn that Bayesian statistics uses the calculus of probability to perform statistical inference, and to describe uncertainty about, for example, the values of parameters in a model.

Bayesian Inference in Outline

We now very briefly introduce the mechanics of Bayesian inference for a generic statistical inference problem. The main goal is simply to introduce some notation and terminology. Subsequent sections will illustrate these ideas.

Assume that we wish to perform inference about unknown quantities θ ("parameters"), from observed data x. Bayesian inference proceeds as follows:

1. Specify a *prior distribution* $p(\theta)$ for θ, which reflects our knowledge or uncertainty about θ prior to observing the data x;

2. Specify a *probability model* $p(x|\theta)$;

3. Compute a *posterior distribution* for θ, using Bayes Theorem:

$$p(\theta|x) = p(x|\theta)p(\theta)/p(x). \tag{1}$$

   The posterior distribution reflects our knowledge of θ after observing the data x. Note that $p(x) = \int p(x|\theta)p(\theta)\,d\theta$ so the posterior distribution is entirely determined by the prior and the likelihood.

Note that, when considered as a function of θ, $p(x|\theta)$ is the likelihood function for θ. This leads to a convenient way to verbalize Bayes theorem:

$$\text{posterior} \propto \text{likelihood} \times \text{prior}. \tag{2}$$

This verbalization also emphasizes that the posterior distribution results from combining the information in the data (likelihood) with external information (prior).

The following examples illustrate these basic ideas, terminologies and mechanics of inference.

Example: classification into two groups from binary data

The following classification problem arises quite generally, but for concreteness we consider a specific example from genetics, where the goal is to identify the origin of a sample based on its DNA.

There are two subspecies of African Elephant: savannah and forest elephants, which differ slightly in their genes. Consider measuring them at a single marker (point in their genome) where there are two alleles (types), A and a. We will assume that allele A occurs at frequency f_S in savannah elephants and at frequency f_F in forest elephants (and that the allele a occurs at frequencies $1-f_S$ and $1-f_F$). Now assume that Interpol have seized an illegally-smuggled tusk, and they measure this marker in DNA from the tusk and find that it carried the A allele. The question before us is: Which subspecies of elephant did the tusk come from, and how confident should we be in this conclusion? (This is simplified version of a real problem: Interpol and other authorities want to know the origin of poached tusks to help focus efforts on curbing this illegal activity; in practice they are interested in much finer-level discrimination, and measure many genetic markers to get more information. See Wasser et al. (2007) and Wasser et al. (2008) for more details.)

Solution

In this problem the main "unknown" of interest is the subspecies from which the tusk arose. We'll use θ to denote this unknown, so $\theta \in \{\text{savannah}, \text{forest}\}$. The "data" are that the tusk has allele A.

First we need to specify a prior distribution for θ, which reflects our knowledge about θ before collecting the genetic data. If we had no prior reason to believe that the tusk was more or less likely from a savannah elephant that a forest elephant then we would use the prior distribution $p(\theta = \text{savannah}) = p(\theta = \text{forest}) = 0.5$[1]. If we had prior experience that poaching tended to occur more often from one subspecies than another then we might modify this prior to reflect this. We will talk a lot more about prior specification later. For now we assume that the prior is specified, with $p(\theta = \text{savannah}) = \pi_S = 1 - \pi_F$.

Next we specify a model for the observed data, $p(x|\theta)$. If the tusk came from a savannah elephant then we know that allele A occurs at frequency f_S so $p(x|\theta = \text{savannah}) = f_S$. Similarly $p(x|\theta = \text{forest}) = f_F$. This completes the specification of the probability model, and hence the likelihood.

Now we apply Bayes Theorem to obtain the posterior distribution for θ. For example:

$$p(\theta = \text{savannah}|x) = p(x|\theta = \text{savannah})p(\theta = \text{savannah})/p(x) \quad (3)$$

$$= f_S\pi_S/(f_S\pi_S + f_F\pi_F), \quad (4)$$

and similarly

$$p(\theta = \text{forest}|x) = f_F\pi_F/(f_S\pi_S + f_F\pi_F). \quad (5)$$

To make this more concrete, let's try some numbers. Let's assume both subspecies are a priori equally likely, so $\pi_F = \pi_S = 0.5$. And assume that the A is twice as common in savannah elephants as forest elephants, say $f_S = 0.3$ and $f_F = 0.15$. Then $p(\theta = \text{savannah}|x) = 0.3/(0.15 + 0.3) = 2/3$.

An alternative solution, and the posterior odds

Here it is helpful to note an alternative (but equivalent) way of solving this kind of problem, which can simplify algebraic manipulation. First, to simplify notation we introduce H_S and H_F to denote the events "θ = savannah" and "θ = forest" respectively. Note that (3) gives an expression for the posterior probability $p(H_S|x)$. Now, the idea is to avoid computing the denominator $p(x)$ in (3) by considering

[1] Throughout I use $p(\cdot)$ to denote both probabilities and densities. Although such sloppiness can sometimes get you into trouble, usually there is no problem.

the ratio $p(H_S|x)/p(H_F|x)$, so that $p(x)$ cancels out. This ratio is known as the "posterior odds" for H_S vs H_F. Applying Bayes theorem to the numerator and denominator, and cancelling out the $p(x)$ that occurs in each, we get

$$\frac{p(H_S|x)}{p(H_F|x)} = \frac{p(x|H_S)}{p(x|H_F)} \frac{p(H_S)}{p(H_F)}. \tag{6}$$

This equation, which applies quite generally for any two events H_S and H_F (but is usually used for mutually exclusive events), also has a convenient verbalization:

$$\text{Posterior Odds} = \text{Bayes Factor} \times \text{Prior Odds} \tag{7}$$

where the Bayes Factor for H_S vs H_F is defined to be the ratio $p(x|H_S)/p(x|H_F)$, and quantifies the *information in the data regarding the relative plausibility of H_S and H_F*. Note that this equation really emphasizes the way that the prior information (prior odds) are combined with the information in the data (Bayes Factor) to give the posterior.

In our case, the Bayes factor is f_S/f_F and the prior odds is π_S/π_F, and we obtain

$$\frac{p(H_S|x)}{p(H_F|x)} = \frac{f_S \pi_S}{f_F \pi_F}. \tag{8}$$

Noting that $p(H_F|x) = 1 - p(H_S|x)$, and solving for $p(H_S|x)$ gives (4).

Could we do this any other way?

The above example serves to illustrate many of the fundamental aspects of Bayesian statistical inference, and in particular how data are combined with prior information to come to a conclusion, including measures of uncertainty in that conclusion. The example was chosen both because it is simple, and because I know of no other satisfactory way to tackle this problem. One could treat θ as a "parameter" and estimate it by "maximum likelihood" (choose θ to maximise $p(x|\theta)$), but this would not give an estimate of uncertainty in the conclusion. Getting a confidence interval for θ would be of little help: either the confidence interval would contain both savannah and forest, or just one of them, suggesting either complete confidence or no confidence, with no room for gradations of uncertainty. And if one tries to treat this as a hypthesis test, comparing H_S vs H_F, which is the null hypothesis?

Thus understanding this example is a key first step to appreciating the fundamental ideas behind Bayesian inference and why it is useful and important.

Alternative inference paradigms: Frequentist inference

The main alternative paradigm to Bayesian inference is Frequentist inference. In broad terms this paradigm proceeds as follows:

- Define a procedure for making some kind of inference about an unknown, θ, from data x. For example, we might define a "confidence interval" $[A(x), B(x)]$ that is designed to contain θ with high probability.
- Study, mathematically, how this procedure would perform *on average* for x randomly sampled from $p(x|\theta)$. (Note that this performance may, in general, depend on θ, necessitating consideration of performance characteristics for different values of θ.)
- For actual observed data x report the outcome of the procedure, and report its expected average performance. For example, we might report the observed value of the interval $[A(x), B(x)] = [1, 3]$ say, along with the fact that, on average, $[A(x), B(x)]$ is expected to contain the true value of θ 95% of the time.

Here are the important ideas I hope you will take away from this class regarding frequentist inference, and frequentist ideas more generally.

1. Studying the average performance of a procedure (across possible datasets x) can be a useful way to compare procedures with one another. For example, I often distribute software implementing methods for performing statistical inference. If lots of people use my software, across lots of different datasets, then I would certainly hope that the the average performance (by some appropriate metric) will be good. However, the different datasets will usually have different values of θ, so average performance across a range of values of θ is usually more relevant than average performance for a specific value of θ.
2. Bayesian methods usually have good average performance across different datasets, particularly when the average is taken over different values of θ (which makes sense because different datasets will typically have different θ!). Indeed, their average performance across different (θ, x) pairs is effectively "optimal" provided that the average is taken over i) a distribution for θ that matches the assumed prior distribution, and ii) a distribution for the data given θ that matches the assumed model. That is, when expectation is taken over $p(D, \theta) = p(\theta)p(D|\theta)$. You might summarize this by saying that Bayesian methods are optimal when the prior and model are both "correct".

3. Although average performance is a reasonable way to compare procedures, *reporting the average performance as a measure of confidence or uncertainty is fraught with theoretical and practical problems.* First the theoretical problem: just because a procedure performs well *on average*, doesn't mean it performs sensibly for that particular x. Indeed, one can easily construct examples of procedures that perform well *on average* but seem silly for particular values of x (see example below). In practice you have usually just observed a particular value of x: wouldn't you want to choose a procedure that performs sensibly for that x, rather than one that works well on average (for other x!) Second, the practical problem: what does it mean to say that $[1, 3]$ is a 95% confidence interval for θ? Does it mean that the true value of θ likely lies between 1 and 3? How likely?

4. Regarding p values specifically, which are one of the most widely used frequentist tools, it is difficult to decide what an appropriate p value threshold should be for rejecting a null hypothesis. This is sometimes referred to as the problem of *calibrating* p values. For example, if $p = 0.05$, does this represent strong or weak evidence against the null hypothesis? Should I reject at $p = 0.05$? or $p = 0.01$? or ...? In practice $p = 0.05$ has been widely adopted as a threshold, but this is an arbitrary convention, and certainly not appropriate for all settings.

Problems with p values

There are essentially two problems with p values that every statistician should be familiar with. The first is that they are easy to mis-define and mis-interpret. The problem of misdefinition is easily addressed in principle, by just avoiding this pitfall and getting the definition right. But in practice this turns out to be harder than expected. There are just so many ways to get the definition wrong! The second problem is that, as mentioned above, p values are hard to calibrate, and in particular it is difficult to say what values of p correspond to evidence against the null hypothesis. For example, in some cases $p = 0.01$ may correspond to evidence for the null hypothesis, and others may correspond to evidence against the null hypothesis. The following example demonstrates the former case ($p = 0.01$ corresponds to data that favor the null); modifying the example so that $p = 0.01$ corresponds to data against the null is left as an exercise.

Example (modified from Berger, p25). Suppose $x \in \{1, 2, 3\}$ and $\theta \in \{0, 1\}$ with

Let H_i denote the hypothesis $\theta = i$, so H_0 is the null hypothesis. Note that $x = 2$ corresponds to evidence against H_0 ($x = 2$ is more probable under H_1 than H_0) and

x	1	2	3
$p(x\|\theta = 0)$	0.005	0.005	0.99
$p(x\|\theta = 1)$	0.0049	0.9851	0.01

$x = 1, 3$ correspond to evidence for H_0 ($x = 1$ and $x = 3$ are both more probable under H_0 than H_1). However, $x = 1$ is only very weak evidence either way because it is almost equally probable under H_0 and H_1. If we order the possibe outcomes in terms of their strength of the evidence against H_0 we get $x = 2$ is the strongest, followed by $x = 1$ and then $x = 3$.

So now, compute the p value for the $x = 1$. Since the p value is the probability of seeing evidence "as or more extreme against H_0 if H_0 holds" the p value is the probability of seeing $x = 1$ or $x = 2$ if H_0 holds, which is $p = 0.01$. By convention $p = 0.01$ would usually be considered evidence against H_0, but in this case $x = 1$ represents evidence slightly in favor of the null hypothesis.

It is a simple exercise to modify this example so that $p = 0.01$ corresponds to evidence against the null. This demonstrates that sometimes $p = 0.01$ can correspond to evidence for the null, and other times it can correspond to evidence against the null.

Some statisticians prefer to work with type I error rates and type II error rates of test procedures, rather than with p values. The idea now is that we will use a test with good type I error rates and type II error rates. The problem now is that a test may have low type I and type II error rates, but sometimes produce a silly answer.

For example, consider the test procedure "Reject H_0 if $x \in \{1, 2\}$" and note that*on average* this procedure performs well, whatever the value of θ. Indeed the probability of making a mistake is 0.01, whatever the value of θ. The probability of a type I error and a type II error are both 0.01. (In other words, the size of the test is 0.01 and the power is 0.99. Furthermore, this is the most powerful test of size 0.01.) So a frequentist using this test, on observing $x = 1$, could reasonably say "Using a test that has (frequentist) error probability 0.01, I reject H_0". But as we have discussed $x = 1$ represents evidence *for* H_0 (albeit very weak), so this seems a silly procedure.

As Berger points out, the resolution here is that $x = 1$ is unlikely to occur. So the test does indeed do well "on average" - it just doesn't produce sensible results in the rare cases where $x = 1$. However, as Berger also points out, in the rare cases where $x = 1$ shouldn't we still provide sensible results?

For more on calibration of p values, including some useful practical guidelines, see Sellke et al. (2001).

Crazy Confidence Intervals

Confidence intervals are probably equally slippery as p values. Let us first review what a confidence interval is. Recall we observe x informing us about a parameter θ. Here is will be useful to explicitly consider that x is the realization of a random variable X whose distribution given by $p(x|\theta)$.

Let $A(\cdot), B(\cdot)$ represent functions, such that

$$\Pr(\theta \in [A(X), B(X)]) = 0.95. \tag{9}$$

Note that here the "randomness" in the probability on the left side refers to randomness in X, not in θ which is to be considered fixed. That is,

$$\Pr(\theta \in [A(X), B(X)]) := \int 1(A(x) \leq \theta \leq B(x)) p(x|\theta) dx. \tag{10}$$

Then $[A(X), B(X)]$ is said to be a 95% Confidence Interval (CI) for θ. Also, in practice on observing $X = x$, analysts will report that "$[A(x), B(x)]$ is a 95% CI for θ".

Example: Suppose $X|\theta \sim N(\theta, 1)$. Then $[X - 1.96, X + 1.96]$ is a 95% CI for θ. This is because

$$\Pr(\theta \in [X - 1.96, X + 1.96]) = \Pr(X \in [\theta - 1.96, \theta + 1.96]) = 0.95.$$

(Remember that all these probability statements are about randomness in X.)

Further, if we observed $X = 2$, we would report $[0.04, 3.96]$ as a 95% CI for θ.

Here is the problem: what does it actually mean to say that "$[0.04, 3.96]$ is a 95% CI for θ"? It sounds like it ought to mean that θ is "likely" to lie in the range $[0.04, 3.96]$. Perhaps even that

$$\Pr(\theta \in [0.04, 3.96] | X = 2) = 0.95. \tag{11}$$

That would be a nice thing to be able to say. But what would this statement mean? It would *have* to be a probabilty statement about θ (because X is no longer random), and to make such a statement we must either treat θ as random, or we have to take the view that we are using probability to represent uncertainty about θ: in either case, computing the conditional probability in (11) would first require specification of the distribution of θ unconditional on X (the "prior distribution"), after which the conditional probability could be computed using Bayes Theorem.

So if "$[0.04, 3.96]$ is a 95% CI for θ" does not mean (11) holds, what does it mean? I personally can provide no other useful meaning to the statement. If (11) does not hold, at least approximately, then I think that the statement "$[0.04, 3.96]$ is a 95%

CI for θ" has no useful meaning. However, confidence intervals are widely used, so people must *think* they are useful, right? The only explanation I can provide is that in fact people routinely "misinterpret" the statement "$[0.04, 3.96]$ is a 95% CI for θ" as meaning that (11) does in fact hold. So now let us consider: are there conditions under which this "misinterpretation" is in fact correct?

References

Sellke, T., M. Bayarri, and J. Berger (2001). Calibration of p values for testing precise null hypotheses. *The American Statistician 55*(1), 62–71.

Wasser, S. K., W. Joseph Clark, O. Drori, E. Stephen Kisamo, C. Mailand, B. Mutayoba, and M. Stephens (2008, Aug). Combating the illegal trade in african elephant ivory with dna forensics. *Conserv Biol 22*(4), 1065–1071.

Wasser, S. K., C. Mailand, R. Booth, B. Mutayoba, E. Kisamo, B. Clark, and M. Stephens (2007, Mar). Using DNA to track the origin of the largest ivory seizure since the 1989 trade ban. *Proc Natl Acad Sci U S A 104*(10), 4228–4233.