1 Introduction

In the last 30 years there has been an enormous increase in the importance of *Bayesian* methods of statistical inference. Whereas at one time it was relatively unusual to see statistical analyses in the literature that used these methods, it has now become commonplace. Indeed, in some areas such as genetics, image analysis and epidemiology, the Bayesian approach has become a standard approach. In this section we will seek to answer the following questions:

- What are the fundamental principles of the Bayesian approach?
- How does the Bayesian approach differ from other philosophies of statistical inference?

Therefore we begin with a brief summary of the essential elements of statistical inference and a brief description of the classical approach.

1.1 Statistical Inference

If you look in the literature you can find many different definitions for the process of statistical inference. For example, in 1962 Savage stated:

> *By [statistical] inference I mean how we find things out - whether with a view to using the new knowledge as a basis for explicit action or not - and how it comes to pass that we often acquire practically identical opinions in the light of evidence.*

Inference can be described more succinctly as the process whereby we alter our beliefs regarding the world on the basis of observations. With this in mind we can recognise 'problems' such as:

- classifying a target from observed radar echoes;
- diagnosing an illness from a battery of tests;
- estimating the mean lifetime of kind of component from a random sample;

as examples of inference problems. A key element of statistical inference is probability theory. Probability theory is used to construct the models from which observations (are hypothesised to) arise and also to give a measure of certainty in the conclusions we extract from observations. It is with respect to the latter use that the different philosophies of inference differ from each other. In the next section we describe the main features of the *classical (or frequentist)* approach to statistical inference.

1.2 Frequentist methods of inference

If you've ever calculated a confidence interval, or calculated a p-value for a hypothesis test, then you've already come across frequentist methods. To see what distinguishes this approach we have to look at the fundamental definition of a probability.

1.2.1 Frequentist versus subjective probability

To a pure *frequentist*, the only meaningful definition of probability of an event is the frequency with which it occurs over a *long sequence of independent trials.* Thus statements such as:

- the probability that a '6' is scored when a fair die is rolled is $\frac{1}{6}$
- the probability that a 'H' is uppermost when a fair coin is tossed is $\frac{1}{2}$

would be perfectly meaningful to a frequentist. They would be happy to ponder the probability of any event that corresponds to a subset of outcomes of an a experiment that can be repeated infinitely often. They would not recognize, for example:

- the 'probability' that I voted Labour in the last election
- the 'probability' that Jack the Ripper was the Duke of Clarence
- the 'probability' that the average height of Scotsmen is less than 1.7 metres

as true probabilities. These deal with propositions that must either be true or false, or quantities that have a particular (albeit unknown) value. Note that the exclusion of this latter form of probability renders the frequentist

philosophy somewhat restrictive. Most people would agree that these probabilities - which express the degree of belief in a proposition regarding the world and are known as *subjective probabilities* - *are* meaningful.

1.2.2 Frequentist (classical) inference

In this section we discuss how the frequentist goes about statistical inference. If they are to attach any probabilistic measures of uncertainty to their conclusions then these must be interpretable as frequencies over multiple repetitions of the experiment that is being analysed, and are calculated from the so-called *sampling distributions* of measured quantities. We illustrate this with the example of calculating a confidence interval for the mean of a normally distributed population.

Example 1 - Confidence intervals for the normal mean Suppose we wish to calculate a 95% confidence interval for the population mean μ based on a random sample of n observations x_1, x_2, ..., x_n. Suppose that we know the poplution variance σ^2. Standard recipes tell us that we can calculate a 95% CI for μ as

$$(\bar{x} - 1.96\frac{\sigma}{\sqrt{n}}, \bar{x} + 1.96\frac{\sigma}{\sqrt{n}})$$

where $\bar{x}$ denotes the sample mean.

What does this mean – in particular what does the 95% mean? The above interval can be 'derived' from the knowledge that over repeated sampling the distribution of the mean of a random sample X_1, X_2, ..., X_n is distributed as $\mathrm{N}(\mu, \frac{\sigma^2}{n})$. Appealing to properties of the normal distribution, it follows that the frequency with which the sample mean $\bar{X}$ lies in the interval

$$(\mu - 1.96\frac{\sigma}{\sqrt{n}}, \mu + 1.96\frac{\sigma}{\sqrt{n}})$$

over repeated sampling is 0.95. Equivalently we can say that the frequency with which μ is contained in the interval

$$(\bar{X} - 1.96\frac{\sigma}{\sqrt{n}}, \bar{X} + 1.96\frac{\sigma}{\sqrt{n}})$$

is 0.95. Therefore the 95%-CI that we report above for the *particular* random sample x_1, x_2, ..., x_n is simply a single random draw from a population of intervals, 95% of which contain μ. Put another way, it is an interval obtained via a process which yields an interval containing μ with frequency 0.95.

What a confidence interval is not.... The quantity 0.95 does not represent the probability that μ lies in $(\bar{x} - 1.96\frac{\sigma}{\sqrt{n}}, \bar{x} + 1.96\frac{\sigma}{\sqrt{n}})$ conditional on the observations x_1, x_2, ..., x_n. To see why this is the case, consider an alternative construction of a confidence interval. Recall that this can be done using the *sample variance*

$$s^2 = \frac{1}{n-1}(\sum x^2 - \frac{(\sum x)^2}{n})$$

instead of the variance, yielding a 95% CI

$$(\bar{x} - t_{n-1}(2.5)\frac{s}{\sqrt{n}}, \bar{x} + t_{n-1}(2.5)\frac{s}{\sqrt{n}})$$

where $t_{n-1}(2.5)$ denotes the 2.5% point of the t distribution on $n-1$ degrees of freedom. In a sequence of independent experiments, the CI constructed in this second manner will contain μ with frequency 0.95. However, on any given trial it is certain that the two intervals will not be equal. One will be contained in the the other. The conditional probabilty that μ lies in the larger interval (if it can be defined) should surely by larger than the probability that it lies in the smaller interval, even though both are 95% CIs.

It is therefore clear that the frequentist measure of 'confidence' is distinct from a conditional probability. The next example shows that the concept of a p-value is similarly hard to interpret.

Example 2. Testing the hypothesis that a distribution is binomial. Suppose that a ornithologist believes that for the sex of an offspring of a certain species of bird is equally likely to be male or female and is independent of the sex of any siblings. He selects a random sample of 100 broods with 4 offspring and counts the number of male offspring in each. The results are summarised in the following table.

Number of males x	0	1	2	3	4
Number of broods f_x	5	20	50	20	5

If X denotes the number of male offspring then the scientist's hypothesis is that $X \sim Bin(4, 1/2)$. This can be tested by calculating the so called χ^2 statistic - which measures how well the observations conform to the supposed distribution - as

$$\chi^2 = \sum_{x=0}^{4} \frac{(E_x - f_x)^2}{E_x}$$

where E_x is the expected number of broods in category x, under H_0. (Note that these numbers are 6.25, 25, 37.5, 25 and 6.25.) For the above data $\chi^2 = 6.67$. Under H_0, over repeated sampling the distribution of the χ^2 statistic follows a χ^2 distribution on 4 degrees of freedom. The scientist calculates a p-value ($\mathrm{P}(\chi^2_4 > 6.67) = 0.15$) and interprets this as a measure of the strength of the evidence in the data against H_0.

Consider now a second scientist who is more open-minded than his colleague. He hypotheses only that $X \sim Bin(4, p)$ where p is unknown. To test his H_0 he first calculates the maximum likelihood estimator of p, $\hat{p} = 0.5$, computes the values of E_x using this estimate and calculates χ^2 to be 6.67 like his colleague. However, because he is estimating p as part of the process of computing χ^2 over repeated sampling the distribution of his χ^2 is (approximately) a χ^2 distribution on 3 degrees of freedom. His p-value is therefore around 0.08.

We now have the somewhat confusing scenario whereby the second scientist finds that the evidence against his null hypothesis is stronger than that found by the first scientist against his null hypothesis, despite the second hypothesis being weaker (i.e. a logical consequence of the first). Thus, whatever a p-value represents, *it must never be interpreted as the probability that the null hypothesis is true conditional on the observed data*. In this example it simply tells us how 'extreme' the observed values of χ^2 are when compared to their distribution over many repetitions of the experiment.

However, the logical basis for interpreting a p-value as a measure of evidence against a hypothesis is also questionable. Most people would agree that any procedure for quantifying evidence should have the property that if Hypothesis A implies Hypothesis B, then any experiment should provide at least as much evidence against A as against B. Example 2 shows that a p-value does not have this desired property.

If your confused, don't worry. It is always difficult to interpret frequentist inferences as statements about the world in which we live. However it precisely interpretations about*this world* that are required if we wish to take a decision about which course of action to take on the basis of evidence. To illustrate this we consider a further example.

Example 3: The Monty Hall Problem This is a very famous problem that has been the subject of much discussion. It is based on an American game show in which a contestant is shown 3 doors (1, 2 and 3) behind which a star prize has been randomly placed with equal probability for each door. The contestant then selects a door after which the host (Monty Hall) opens one

of the two remaining doors to reveal no star prize behind it. The contestant is then invited to guess which of the two closed doors hides the star prize.

Suppose that a particular contestant selects door 1, after which 2 is opened to reveal no prize behind it. Should she stick with door 1 or switch to door 3 when making her final guess?

The frequentist solution says that it is best to switch since the contestant only loses by switching if she happens to choose the door hiding the prize at the first guess. The probability of choosing the correct door first time is 1/3. Thus the probability that she wins by switching is 2/3. More explicitly, this means that a contestant adopting a switching strategy over many independent repetitions of the game will win with frequency 2/3. One opting for a 'sticking' strategy will only win with frequency 1/3.

To our contestant, what happens over a long sequence of plays of the game is irrelevant. She only has one bite at the cherry and wishes to make judgements only about the particular set of circumstances in which she has found herself. Is the prize behind 1 or is it behind 3? Does the frequentist solution allow her to claim that the probability that the prize is behind 3, conditional on Monty's actions, is 2/3? As we shall see in the next section, the problem as described above *does not* allow her to formulate conditional probabilities of this kind, without making additional assumptions.

1.3 Conditional probability, Bayes theorem, and inference

1.3.1 Conditional probability

Let S be a sample space representing the set of outcomes of some experiment. We recall the definition of the conditional probability of an event. If A and B are 2 events (subsets) with $P(B) > 0$, then we say that the conditional probability of A given B, written $P(A|B)$, is

$$P(A|B) = \frac{P(A \bigcap B)}{P(B)}$$

The frequentist interpretation of $P(B|A)$ is that it represents the frequency with which A occurs over those experiments in which B occurs. The calculation of a conditional probability can also be seen to be the appropriate way of expressing our belief in what outcome has occured *given partial information on that outcome*. For example, suppose that we randomly draw a

single card from a pack from which the Ace of Spades is missing. Without any knowledge of what card has been drawn your subjective probability that it is an Ace is $P(A) = 3/51$. Suppose that you are now told the colour of the card - black in this case (event B). Then your subjective belief that the card is an Ace should change in the light of this evidence to become 1/25. This is of course the value that would be obtained by calculating

$$P(A|B) = \frac{P(A \bigcap B)}{P(B)} = \frac{P(Ace\ of\ clubs)}{P(B)} = \frac{1/51}{25/51}$$

Rearranging the above definition of conditional probability and noting the symmetry of the expression we obtain

$$P(A \bigcap B) = P(A)P(B|A) = P(B)P(A|B).$$

This can be rearranged to obtain *Bayes' Law*

$$P(A|B) = \frac{P(A)P(B|A)}{P(B)}$$

Actually, Bayes' law has a more general formulation relating to partitions of the sample space. Let A_1, A_2,... denote a partition of the sample space (i.e. they are mutually exclusive and their union is the whole sample space). Then since

$$P(B) = \sum_i P(A_i)P(B|A_i)$$

for any j, we can write

$$P(A_j|B) = \frac{P(A_j)P(B|A_j)}{\sum_i P(A_i)P(B|A_i)}$$

Now let us attempt to analyse the Monty Hall problem within this framework. Let A_1, A_2, and A_3 denote the events that the prize is placed behind doors 1, 2 and 3, respectively. Let B denote the event that the host opens door 2 (the contestant having selected door 1). Using the above formulation of Bayes' Law, the contestant should calculate the P($A_3|B$) as

$$P(A_3|B) = \frac{P(A_3)P(B|A_3)}{(P(A_1)P(B|A_1) + P(A_2)P(B|A_2) + P(A_3)P(B|A_3))}.$$

Now $\mathrm{P}(A_1) = \mathrm{P}(A_2) = \mathrm{P}(A_3) = 1/3$ since the prize is initially placed behind each door equiprobably. Furthermore, $\mathrm{P}(B|A_2) = 0$ and $\mathrm{P}(B|A_3) = 1$ (since the host must reveal a door with no prize). However, the quantity $\mathrm{P}(B|A_1)$ has not been specified anywhere in the problem. We have not said how the host selects a door when neither door not selected by the contestant hides the prize. Therefore we cannot compute the conditional probability $\mathrm{P}(A_3|B)$ without further assumptions. For example, we might propose that $\mathrm{P}(B|A_1) = 1/2$, in which case we would have $\mathrm{P}(A_3|B) = 2/3$. If we assumed that, when either door 2 or 3 can be opened, the host selects 2 with some probability p, $0 < p < 1$, then all we can say is that $1/2 < P(B|A_1) < 1$. Nevertheless, this tells the contestant that, at the very least, nothing will be lost by switching to door 3 when making the final choice.

This problem illustrates some of the essential elements of the Bayesian approach (described fully in the next section of the course). In particular, we highlight the following features.

- The inferences from a Bayesian analysis are expressed in the form of probabilities that are *conditional on the observed events* and represent our subjective beliefs about the world in the light of these events (referred to as *posterior probabilities*).

- In order to allow such conditional probabilities to be calculated we need to ensure that *all probabilities involved in the calculations are defined.* This may require additional assumptions to be made (e.g. $P(B|A_1) = 1/2$ in the Monty Hall problem.)

- When calculating Bayesian posterior probabilities of hypotheses or quantities of interest (parameters) e.g $P(A_i|B), i = 1, 2, 3$ in the above problem, we must specify the unconditional probabilities $P(A_i)$, representing our belief in these possibilities before observing any data and known as *prior probabilities.* In the case of the Monty Hall problem the parameter (the door hiding the prize) has been generated from a process with known frequencies of outcomes. It is therefore logical (*don't you think it is?*) that the our beliefs regarding which door hides the prize (given no other information) should correspond exactly with these frequencies. In this case, our posterior probabilities will also have a frequentist interpretation. (*What is it?*) However in many cases, there is no underlying sampling scheme by which the 'true scenario' has been selected from the various possibilities. In these cases, prior probabilities must still

be assigned but these can only have the interpretation as *a measure of prior belief in the possibilities.* This arises in forensic statistics where the two hypotheses may be the guilt or innocence of a defendant in a court case. It is the extension of Bayes' theorem to this kind of scenario that offends the pure frequentist.

- The Monty Hall problem involves calculating posterior probabilities of discrete quantities (the number of the door hiding the prize.) However, the Bayesian approach also extends to continuously valued quantities for which prior belief must be summarised in terms of a *density function.* The mathematics of continous probability densities will therefore play a prominent role in the rest of the course.

Exercises

1. Y has a pack of 4 cards (Ace and Queen of clubs, Ace and Queen of Hearts) from which he deals a random of selection 2 to player X. What is the probability that X receives both Aces conditional on them receiving at least 1 Ace. Suppose now that Y deals X two cards from the pack of 4, after which X says "*I have an Ace*".

   (a) Discuss whether the above information is sufficient to calculate the conditional probability P(X has 2 Aces | X says "I have an Ace").

   (b) If it is not, what other information would be required in order to calculate this conditional probability?

2. An urn is known to contain n differently coloured balls where n can be any integer in the set 1, 2, 3. Your prior information tells you that n is equally likely to be any of these values. A ball is drawn randomly from the urn - it is red. Alice argues that since the probability of the red ball being drawn conditional on there being n balls in the urn is $1/n$, then

$$P(n = 1|red\ ball\ drawn) = \frac{\frac{1}{3} \times \frac{1}{1}}{\frac{1}{3} \times \frac{1}{1} + \frac{1}{3} \times \frac{1}{2} + \frac{1}{3} \times \frac{1}{3}}$$

   and calculates the posterior probabilities of n being 1, 2, and 3 as 6/11, 3/11 and 2/11 respectively. She then expresses her surprise that the her beliefs regarding N have changed having observed only the colour of a single draw from the urn.

   (a) Explain the fallacy in her argument and why the above the information alone does not define a posterior probability for n.

   (b) Bertie assumes that the n balls placed in the urn are drawn uniformly at random from a large stock of differently coloured balls. Calculate Bertie's posterior probabilites for $n = 1, 2, 3$.

   (c) Under what circumstances would Alice's posterior probabilities be correct?

3. Suppose now in the situation of question 2, two balls are drawn from the urn *with replacement* and the event that both are the same colour

is observed. Calculate the posterior probabilities for $n = 1, 2, 3$ in this case.

4. (Yet more balls and urns) Five balls are drawn uniformly randomly from a very large population of black and white balls where the proportion of black balls is 1/3. You do not know the colours of the balls selected.

   (a) Give suitable prior probabilities for the number of black balls in the urn.

   (b) You now select two balls uniformly at random from the urn with replacement. They are both white. Calculate the posterior probabilities for the number of black balls in the urn.

   (c) Suppose that the two balls were selected from the urn without replacement and were both white. Calculate your posterior probabilities for the number of black balls in the urn for this case.

5. A fair coin is tossed n times where n can take the values 1, 2, ..., 5 with equal probability. Suppose that 2 heads result from the n tosses.

   Determine the posterior distribution (i.e. work out the probability function) of n and identify the value of n that is *a posteriori* most likely.

   Suppose now the coin is to be tossed repeatedly until m tails are obtained where the value m is first selected from a $Geometric(1/3)$ distribution. Suppose that 2 heads are obtained in the sequence. What is the posterior distribution of m given this information? (It is sufficient to write an expression involving infinite sums!)

6. Formulate a new problem (with solution) to illustrate the use of Bayes' theorem to calculate posterior probabilities. The best examples will be given (with appropriate acknowledgement) to the next class to take this course.

2 Bayesian Inference

2.1 Priors, Likelihoods and Posteriors

Throughout this section we will consider problems of the following kind. We carry out some kind of *experiment* (a procedure where the outcome is uncertain) for which we believe the outcome can be considered as a random draw from some probability model. This probability model is specified by an unknown parameter vector, which we denote by θ. Our aim in carrying out the experiment is to estimate the unknown θ.

Particular instances of this include:

- estimating the proportion of the population who will vote Labour in the next election from the results of a poll ($\theta = p$).
- determining whether or not a target is located in a range-cell based on radar echo (θ here would indicate presence (1) or absence (0) of a target)
- estimating the mean incubation period of an infectious disease based on observations of an epidemic

Suppose that our experiment produces observations, y, from which we must estimate θ. Then in the *Bayesian approach* we update our knowledge of θ, having obtained y by applying Bayes' theorem. Let us suppose for the moment that both θ and y are discrete quantities. Applying Bayes' Theorem we would seek to obtain an expression

$$P(\theta_j|y) = \frac{P(\theta_j)P(y|\theta_j)}{\sum_i P(\theta_i)P(y|\theta_i)}$$

Working backwards, we can see that this expression can only be made meaningful if we can define the quantities $P(y|\theta_j)$ (the probability of the observations for parameter value θ_j and $P(\theta_j)$ (the probability of θ_j) for each θ_j. These two functions of θ are the cornerstones of the Bayesian approach. They are the *likelihood function* and the *prior distribution* for θ respectively and must be specified in any Bayesian calculation. The conditional distribution specified by $P(\theta_j|y)$ is known as the *posterior distribution* of θ, and can be considered as the complete representation of our knowledge about θ in the light of y.

In many situations we will be interested in a continuously-valued parameter. For this case we obtain an expression for the *posterior probability density function* of θ,

$$\pi(\theta|y) = \frac{\pi(\theta)P(y|\theta)}{\int \pi(\theta')P(y|\theta')d\theta'} \tag{1}$$

where $\pi(\theta)$ is referred to as the *prior density* for θ. Students who have done a course in measure theory will realise that there is no essential conceptual difference between the expressions for the posterior distribution in discrete and continuous cases.

We now consider how the prior and the likelihood are specified in more detail.

2.2 The prior distribution

The major source of controversy in the Bayesian approach is the assignment of the prior distribution for the unknown parameter θ. Sometimes (if the value of θ arises through some random sampling procedure) there is no controversy. For example, if θ denotes the presence ($\theta = 1$) or absence ($\theta = 0$) of a gene that is known to occur in a certain proportion, p of the population, and y records the presence or absence of an associated characteristic in a randomly selected member of the population, it would be natural to define the prior for θ by $P(\theta = 1) = p$. However, in general there will be no such sampling procedure for θ to which we can appeal. Then we will require to set the prior for θ using our subjective judgement.

How should this be done? Later in the course we will return to this question in more detail. For the moment we will simply distinguish two cases. If we have some prior knowledge of the 'likely' values of θ then we might suggest a prior which reflects this and places more weight on the likely values and less weight on more improbable values. If on the other hand we have little genuine prior knowledge then we should opt for a prior distribution that 'supports' a broad range of values of θ. We can think of this as a distribution that spreads the available probability mass widely. Precisely how 'vague' priors of this kind should be selected is very contentious (see later).

2.3 The likelihood function

The *likelihood function* is fundamental to many statistical inferential procedures, not just Bayesian ones. Since we are thinking of $P(y|\theta)$ as a function of θ (with y fixed) we will write the likelihood as $L(\theta|y)$ to emphasise the dependence on the θ. The importance of the likelihood is summarised by the *likelihood principle*, which (in essence) says the following.

> Suppose that two experimental outcomes y_1 and y_2 define likelihoods $L(\theta|y_1)$ and $L(\theta|y_2)$ that are proportional to each other, that is $L(\theta|y_1) = C(y_1, y_2)L(\theta|y_2)$, for all θ. Then the conclusions about θ drawn from y_1 and y_2 should be identical.

The likelihood principle says that the all the information that the data tell you about the parameters is embodied in the likelihood. A fuller account of the likelihood principle - and how it is a logical consequence of two other principles (the *conditionality* and *sufficiency* principles) - can be found in texts such as Casella & Berger. Further support for the likelihood as the most appropriate expression of information regarding the plausibility of different values of θ can be seen from the Neyman-Pearson Lemma in classical hypothesis testing.

All in all, in any problem of statistical estimation or inference it's a good idea to see if you can write down a likelihood function for the data. This requires you to use the rules of probability theory in order to work out the 'probability' of the observations given the parameter θ. Depending on whether the observations y are discrete, continuous, precisely measured, or censored (known only to lie in certain intervals) constructing likelihoods will typically require use of probability mass or density functions and cumulative distribution functions. We will illustrate the process of constructing the likelihood in several practical examples of Bayesian inference.

2.4 Pros and cons of Bayesian inference

There is a long-standing debate within statistical inference with regard to the most appropriate philosophy of inference. Pure Bayesians will point out that their approach respects the likelihood principle since it is through the likelihood *only* that the data enter the calculations. Critics will point to the *subjectivity* in the choice of prior distributions for parameters and its influence on the conclusions. Bayesians may counter that any source of dubiety in their

analyses is clearly identified in this single issue (the choice of prior), while logical inconsistencies can be found throughout classical statistics (e.g. p-values as a measure of evidence).

A major selling point of the Bayesian approach is that it provides a very natural means of including prior knowledge of a system in experimental analyses, provided this knowledge can be expressed in terms of a probability distribution on the parameters. It also provides a natural means of combining information from several experiments. Since the posterior parameter density represents the sum total of knowledge of parameters in the light of experimental data, it provides us with the appropriate prior to use in the analysis of subsequent experiments.

2.5 Some practical examples of Bayesian inference

Example 1: Bayesian estimation of Binomial proportion, p. A geneticist wishes to estimate the proportion of the population carrrying a certain gene. They collect DNA from a random sample of 20 individuals, of whom 5 are found to carry the gene. Carry out an investigation of p using Bayesian techniques.

The first thing we need to do is construct the likelihood $L(p)$. If Y denotes the number of gene-carriers in the sample, then this is simply $P(Y = 5)$ expressed as a function of p. Now, since we are taking a random sample from a large population then this probability is given by the binomial probability:

$$L(p) = P(Y = 5|p) = \binom{20}{5} p^5 (1-p)^{15}.$$

Now, we need to specify a prior density for p. Let us suppose that the geneticist has little information on the value of p. To reflect this they choose for their prior density, $\pi(p) = 1$, that is a uniform distribution on the interval (0, 1). We will later discuss the extent to which a uniform prior can be considered to represent *prior ignorance*.

We can now identify the posterior distribution of p as

$$\pi(p|y) = \frac{\pi(p)L(p)}{\int_0^1 \pi(p')L(p')dp'}$$

Now the denominator being an integral over the range of p is independent of

p, therefore we can write

$$\pi(p|y) \propto \pi(p)L(p) = \binom{20}{5} p^5(1-p)^{15} \propto p^5(1-p)^{15}$$

This functional dependence on p identifies the $\pi(p|y)$ as a Beta distribution. In fact it is a Beta(6, 16) density.

Self-study exercise: Review the properties of the Beta distribution

We can therefore identify the posterior mean of p as $3/11$, and the mode as $1/4$.

Example 2. Inference for the Exp(λ) distribution. Suppose that the lifetime of a particular type of component is believed to be distributed as an Exp(λ) where λ is unknown. In order to estimate lambda, you select a random sample of 20 components and measure their lifetimes to be $t_1, t_2, t_3, ..., t_{20}$. Carry out a Bayesian analysis of these data to estimate λ.

Again let's assume that we have little prior knowledge about the value of λ. Therefore to select a prior we want a density that 'supports' a broad range of values of λ. One possibility would be to use an Exp(α) distribution where α is small in some sense. Suppose we set $\alpha = 0.1$.

Now suppose that we take the observations and note that $\sum t_i = 10.0$ time units. Since (we assume) the observed times are independent of each other, then we express the likelihood as the joint density $f(t_1, x_2, ..., t_{20}|\lambda)$. This is given by

$$L(\lambda) = \lambda^{20} e^{-\sum \lambda t_i} = \lambda^{20} e^{-\lambda \sum t_i}$$

We can now exhibit the posterior density as

$$\pi(\lambda|t) \propto \pi(\lambda)L(\lambda) = \lambda^{20} e^{-\lambda(\alpha + \sum t_i)}.$$

We can see that the posterior density of λ must be a $\Gamma(21, 10.1)$. We can immediately determine the posterior mean and variance to be 2.08 and 0.21 respectively.

Example 3. Consider the same set-up as Example 2 but suppose that now the data, rather than giving the precise times of failure of all components, only record the times of failure up to $t = 0.5$, at which time the experiment ceases. Suppose that 14 components fail in this period with $\sum_{i=1}^{i=14} t_i = 2.2$, the remaining 6 components being operational at $t = 0.5$. To construct the likelihood for this case we note that $P(T > 0.5|\lambda) = e^{-0.5\lambda}$, and insert this factor for each component whose lifetime exceeds 0.5. This yields a likelihood

$$L(\lambda) = \lambda^{14} e^{-\lambda(3.0 + \sum_0^{14} t_) i} = \lambda^{14} e^{-5.2\lambda}.$$

If we again use the Exp(0.1) prior for λ, then the posterior density of λ is $\Gamma(15, 5.3)$. In this case the posterior mean and variance of λ are 2.83 and 0.53 respectively.

We can use Bayesian methods whenever we can write down a likelihood function for the observations that we have.

2.6 Conjugate prior densities - prior elicitation

All the prior densities used in the previous section are examples of *conjugate prior densities*. That is, the posterior density derived is another member of the same family of distributions. In Example 1, the U(0, 1) prior density for p is a particular case of a Beta distribution - a Beta(1, 1) distribution. The analysis could have been repeated using any Beta prior distribution, yielding a Beta posterior. In general, if the prior for p were $Beta(\alpha, \beta)$, then the posterior would be $Beta(\alpha + 5, \beta + 15)$.

Suppose the geneticist in Example 1 wished to include his prior experience in the analysis by using a more informative (less vague) Beta prior than the Beta(1, 1) density. His choice of α and β could loosely be interpreted as quantifying the extent of this experience. Suppose that the geneticist had carried out n tests in the past of which r proved to be positive for the gene. Then he could represent this experience by using a Beta$(r + 1, n - r + 1)$ density for p. The larger n and r, the smaller the prior variance for p.

The values of α and β used to specify the prior have little effect on the posterior in the case where a very large sample is collected. If the sample size is m and the number of positives is q, then the posterior density of p is Beta$(\alpha + q, \beta + m - q)$. The mean and variance of this posterior density depend weakly on α and β when m and q are large. In this case we say that the data *swamp* the prior - a nice situation to be in if you're a statistician.

In Example 2, the choice of Exp(α) as the prior for λ is a particular case of the Gamma distribution - a $\Gamma(1, \alpha)$. It is straightforward to check that, if a $\Gamma(\alpha, \beta)$ had been used, then the resulting posterior density would have been $\Gamma(20 + \alpha, \beta + \sum t_i)$. If you wanted to select a more informative prior, perhaps contentrating weight around a particular value, then larger values of α and β should be chosen.

2.6.1 Selecting a prior to represent your knowledge

As mentioned before, it is through the selection of the prior (and the choice of the model, of course) that subjectivity enters into any Bayesian analyses. There are various ways in which a prior distribution might be elicited for any particular parameter in an analysis. Here we will focus on the 'cheap and cheerful' approach. Suppose in Example 2 your prior belief is summarised by the statements:

The probability that the mean lifetime of a component is greater than 2 days is 0.25

The probability that the mean lifetime is less than 0.25 days is 0.25.

Since the mean lifetime is given by $1/\lambda$, this is equivalent to specifying that the prior for λ should have its quartiles at $\lambda = 0.5$ and $\lambda = 4.0$ respectively. There is a Gamma distribution with this property but it is rather messy to obtain, since this requires solution of a system of simultaneous equations (involving incomplete gamma functions) for α and β.

An easier approach is to consider a Weibull distribution as the prior for λ. A Weibull(γ, β) density for λ is defined by

$$f(\lambda) = \gamma\beta\lambda^{\gamma-1}e^{-\beta\lambda^{\gamma}}, 0 \leq \lambda < \infty.$$

If $\lambda \sim Weibull(\gamma, \beta)$, then $\lambda^{\gamma} \sim Exp(\beta)$. We now need to solve the system of equations:

$$e^{-\beta 4.0^{\gamma}} = 0.25$$

$$e^{-\beta 0.5^{\gamma}} = 0.75$$

Now transforming these equations we obtain

$$-\beta 4.0^{\gamma} = \log 0.25$$

$$-\beta 0.5^{\gamma} = \log 0.75$$

It follows that

$$8^{\gamma} = \frac{\log 0.25}{\log 0.75}$$

so that

$$\gamma = \log(\frac{\log 0.25}{\log 0.75})/\log 8 = 0.756$$

It follows that $\beta = 0.486$. This implies that the prior mean and variance of λ are 3.07 and 16.97 respectively. Of course a problem with the Weibull

prior is that we lose the property of conjugacy - our posterior will not be a Weibull distribution! *What if we use a* $\Gamma(\alpha, \delta)$ *prior with the same mean and variance?* This implies that $\delta = 3.07/16.97 = 0.181$, while $\alpha = 3.07\delta = 0.556$. Now we can check the values of the *cdf* for the $\Gamma(0.556, 0.181)$ at 0.5 and 4.0 to see how well they match our desired prior specification. The corresponding values are 0.29 and 0.74, respectively. This is in 'reasonable' agreement with our prior specifications, and we could use this gamma prior (with its convenient conjugacy properties).

Although prior elicitation is an important process, it is more important to bear in mind that the choice of prior is *subjective*. It is the most questionable aspect of the Bayesian approach. In any Bayesian analysis it is sensible to investigate how the conclusions would be affected if a different prior were used and to experiment with a few priors with differing degrees of 'informativeness'.

2.7 Reporting conclusions from a Bayesian analysis

Having derived the posterior distribution of a parameter there are several ways in which we can express the results. For single parameters, a plot of the posterior density is very informative and shows clearly the range of values consistent with your posterior beliefs. We can also quote quantities such as the posterior mean of a parameter or the posterior variance. Indeed any summary of a distribution can be used.

A natural analogue of the frequentist confidence interval for a parameter is the *Bayesian credible interval*. For example, suppose that given data y you derive the posterior density of θ as $\pi(\theta|y))$. Then a 95% credible interval (a, b) is any interval whose posterior probability of containing θ is 0.95. Often we might quote an equal-tailed interval (obtained by selecting the 97.5% and 2.5% critical points of $\pi(\theta|y))$, or a minimum-width interval (by thresholding $\pi(\theta|y)))$ assuming this can be calculated.

To calculate credible intervals for the parameter we need the cumulative distribution function of the posterior. Where the posterior has a convenient form, such as a Beta or Gamma distribution, we can usually use standard functions from a computer package such as 'R' to do the calculations for us. For the particular case of the posterior density of θ being $\Gamma(n, \beta)$, where n is an integer, we can exploit the fact that

$$2\beta\theta \sim \Gamma(n, 1/2) \sim \chi^2_{2n}$$

and use statistical tables to work out the bounds of the credible region.

Depending on the circumstances, we may be interested in the posterior probability of a parameter being greater than or less than some threshold. For example, this might be the case where the experiment has been done for the purpose of quality control.

2.7.1 Predictive distributions

Usually when we carry out a Bayesian analysis to obtain $\pi(\theta|y)$ our interest lies in predicting some other quantity of practical importance, z, whose distribution is determined by θ. For example, in the case of Example 2, our interest may be in the lifetime of the next component that we select. Having obtained $\pi(\theta|y)$, what we really want to do is determine the distribution $\pi(z|y)$. This distribution, known as the *predictive distribution* of z, must be exhibited as a *mixture distribution* over the possible values of θ. We must write

$$\pi(z|y) = \int f(z|\theta)\pi(\theta|y)d\theta$$

where $f(z|\theta)$ denotes the density of z given θ. In this section we give some examples of calculating predictive distributions and highlight some of the mixture distributions that arise in standard problems.

Example 4 Following change in regulations, students are suppose to bring their own calculators to examinations. However a number invariably forget and invigilators bring a small number to exams for these individuals. Suppose that in the first exam after the rule change an invigilator finds that 2 students out of a class of 30 have forgotten their calculators. The next week she has to invigilate an exam with 25 students. How many calculators should she bring in order to be 95% certain that she will have enough? Assume she is a Bayesian who takes a pessimistic view of the organisational skills of students.

A Bayesian solution. First of all, we need to assume some statistical model for the number of students that forget to bring a calculator to an exam. A natural assumption is that this number follows a $Bin(n, p)$ distribution where n denotes the number of students taking the exam, and that p is the same for all exams.

Next she needs to identify a prior for p, representing her beliefs before having seen the data from the 1st exam. Being pessimistic she assumes a U(0, 1) prior for p. Now, as in Example 1 in the previous section she immediately calculates that *a posteriori* $p \sim Beta(3, 29)$.

Now she must consider *the predictive distribition* of Z, the number of students who forget their calculators in the next exam. Given p, then $Z \sim Bin(25, p)$. To get the predictive distribution $\pi(z|y)$ we need take the expectation of the binomial probability given p over the posterior distribution of p. The probability function of Z, given the data, is then

$$f(z) = \int_0^1 \pi(p|y)\binom{25}{z}p^z(1-p)^{1-z}dp, 0 \leq z \leq 25$$

This the pmf of the Beta-Binomial distribution. An explicit formula can be obtained (see lecture for a derivation of this) as

$$f(z) = \binom{25}{z}\frac{\Gamma(2+z+1)\Gamma(53-z+1)\Gamma(32)}{\Gamma(3)\Gamma(29)\Gamma(57)}$$

Now by examining the associated cumulative distribution function, we find that $P(Z \leq 5) = 0.931$ while $P(Z \leq 6) = 0.966$. Therefore she should bring 6 calculators to be 95% certain of having enough. In the event she brings 7. *Why?*

Exercise Verify the derivation and the calculations using a package like 'R'. What would happen if a more informative prior were used, giving more weight to smaller values of p?

Example 5. Consider the component experiment of Example 2 above in which the posterior density of λ was $\Gamma(\alpha, \beta)$. Let us suppose that this component is used in a space vehicle which has to perform a flight of duration 1 day? If the component fails during the flight then it is replaced immediately from a pool of identical components whose lifetimes are all independent of each other. How many components in total are needed to ensure that the vehicle completes the flight with at least 90% certainty?

To solve this we need to consider the predictive distribution of the number of components, Z, which fail during a 1-day flight. Now, given λ, the distribution of Z is Poisson(λ) (see lecture for a derivation). Therefore to obtain the predictive probability mass function $f(z)$ we must take the expectation of this Poisson (conditional) probability over λ. This gives

$$f(z) = \int_0^\infty \pi(\lambda|y)\frac{e^{-\lambda}\lambda^z}{z!}dz, 0 \leq z \leq \infty$$

$$= \frac{\Gamma(z+\alpha)}{\Gamma(\alpha)z!}p^\alpha(1-p)^z, 0 \leq z \leq \infty$$

where

$$p = \frac{\beta}{1+\beta}$$

This specifies the predictive distribution of Z as being a *negative binomial* distribution. By considering the cdf for the case $\alpha = 21$, $\beta = 10.1$ we note that $P(Z \leq 3) = 0.83$, while $P(Z \leq 4) = 0.93$. It follows that 5 components are required to give at last 90% certainty of completing the 1-day flight successfully.

These examples of calculating predictive distributions show how mixture distributions naturally arise. In the cases considered so far, there has been only a single parameter and the integrals have been analytically tractable. More generally, Bayesian inference and prediction can require calculation of integrals that may be multidimensional (in the case of more complex models), or may fail to be analytically tractable. One of the barriers to widespread implementation of Bayesian ideas in the past was the complexity of the integrations that naturally arose. As we see later in the course, this difficulty has been overcome to a major extent through the use of stochastic integration techniques, coupled with modern computer power. This allows such integrals to be estimated numerically.

Exercises

1. Let $\mathbf{X} = (X_1, X_2, \ldots, X_n)$ be a random sample from the Poisson distribution with mean μ. Show that the conjugate prior is Gamma. Using $\mu \sim Gamma(\alpha, \beta)$ as the prior, determine the posterior distribution of μ.

   Suppose you specify $\mu \sim Gamma(1, 0.5)$, $n = 5$ and observe $\sum x_i = 15.0$. Calculate an equal-tailed 95% credible region for μ in this case.

2. Let $\mathbf{X}$ be a random sample from the Exponential distribution $Exp(\lambda)$ with mean $1/\lambda$, i.e. $\Gamma(1, \lambda)$. Show that the conjugate prior is Gamma. In particular, if X is a single observation, show that the prior $\Gamma(\alpha, \beta)$ leads to a posterior density for λ being $\Gamma(\alpha + 1, \beta + X)$.

   An important consequence of the conjugacy property is that if observations arrive sequentially then updating the posterior distribution is simple. Suppose that the prior distribution is $G(\alpha, \beta)$ and that x_1 is observed. Obtain the posterior distribution. Now suppose that x_2 is observed. Find the new posterior distribution by updating the existing posterior. Finally, show that this posterior distribution is the same as that obtained from the original prior if we observe a random sample of size 2 consisting of (x_1, x_2).

3. The lifetime of a component, T, follows an $Exp(\lambda)$ distribution where *a priori* $\lambda \sim \Gamma(1, 2)$. You select a random sample of 5 components for which $\sum t_i = 3.0/days$. Find the posterior distribution of λ.

   A component of this kind forms part of certain system which is required to function continuously for a period of 6 hrs. What is the probability that the component fails before the end of 6 hrs? (You will have to work out the posterior predictive distribution of the lifetime of a component.)

4. Let $\mathbf{x} = (x_1, x_2, \ldots, x_n)$ be a random sample from the Pareto distribution with p.d.f. $f(x) = \theta(1 + x)^{-(\theta+1)}$, $0 < x < \infty$. Show that the Gamma distribution is the conjugate prior for this distribution by proving that if the prior distribution of θ is $G(\alpha, \beta)$, then the posterior is $G(\alpha + n, \beta + t(\mathbf{x}))$, where $t(\mathbf{x}) = \sum_i = \log(1 + x_i)$.

5. *[1997 Statistical Inference Exam, Q4]* In a raid on a coffee shop, Bayesian trading inspectors take a random sample of n packets of coffee,

each of nominal weight 125 g. They model these data as independent values $X_1, \ldots, X_n$ from a Normal $\mathrm{N}(\mu, \sigma^2)$ distribution. They take σ^2 to be known, while for μ they assume a prior distribution of $\mathrm{N}(\mu_0, \sigma_0^2)$, where μ_0 and σ_0^2 are specified values.

(a) Show that the inspectors' posterior distribution is also Normal, and find its mean and variance.

Show that the mean of this distribution is a weighted average of the prior mean μ_0 and the sample mean $\bar{x}$.

(b) The data they obtain are (weights in grams):

105.3, 113.3, 114.5, 121.2, 122.9, 123.7, 124.0, 124.6, 124.9, 124.9, 124.9, 125.1, 125.5, 125.9, 126.8, 127.7, 128.2, 128.3, 128.5, 130.2 ($\sum x_i = 2470.4$, $\sum x_i^2 = 305828.98$).

The parameter values they assume are $\mu_0 = 126$, $\sigma_0^2 = 1$, $\sigma^2 = 4$.

The inspectors can impose a fine if their 95% credible interval falls wholly below the claimed value of $\mu = 125$ g.

i. Show that the inspectors' 95% credible interval for μ for these data does lie wholly below 125 g; they therefore impose a fine on the owners of the coffee shop.

ii. Sketch the data (a dotplot or similar), and calculate their sample median and sample variance.

iii. Comment briefly as to whether the inspectors are justified in imposing a fine on the basis of this sample.

3 Bayesian inference continued

This section extends the ideas of the previous section to examine some deeper issues of Bayesian Inference.

3.1 Non-informative prior densities

For some researchers in Bayesian statistics, the concept of a prior density which represents *complete prior ignorance* of a parameter is something of a Holy Grail. In reality, there is no such thing as a truly non-informative prior. Any prior density assigned to a parameter embodies a lot of information in the probability statements that would follow from it. It is tempting to think of a uniform prior density as representing ignorance as in the case of Example 1 in the previous section. By assigning a uniform prior density, however, we are saying that the value of p is equally likely to fall in *e.g.* the intervals (0, 0.1), (0.1, 0.2),..., (0.9, 1.0). This says some very particular things about our prior beliefs and cannot represent total ignorance. That the uniform density does not really represent prior ignorance is more obvious when we consider transformations of a parameter. If we have no knowledge of p, then we have no knowledge of $q = p^2$. However, if the prior distribution of p is uniform, then the prior distribution of q is certainly not.

Self-study exercise Ensure you are familiar with the method for deriving the pdf of $\phi = h(\theta)$ from the pdf of θ for 1-1 differentiable mappings h. What is the pdf of q above?

What kind of prior distributions can be suggested when there is no available information? Sticking with the 1-dimensional situation we now look at the Jeffreys prior Suppose that we wish to estimate q based on some random observation Y. Now in classical estimation theory a quantity of interest is the *Fisher information* defined by

$$F(\theta) = -E_{Y|\theta}[\frac{\partial^2 logL(\theta|Y)}{\partial\theta^2}]$$

Roughly speaking, the reciprocal of this quantity determines the width of likelihood based confidence intervals (under large sample conditions). Now suppose that $\phi(\theta)$ is a transformation of θ, such that $F(\phi)$ is constant. Then we obtain Jeffreys' prior for θ by placing a uniform prior on ϕ and then examining what prior this 'induces' on θ. It turns out that the induced prior

on θ must satisfy

$$\pi(\theta) \propto F(\theta)^{1/2}$$

Now a Jeffreys prior has certain invariance properties. If we place a Jeffreys prior on θ and then consider $\phi(\theta)$, the prior induced on ϕ is automatically the Jeffreys prior for ϕ. We now look at some particular cases of Jeffreys' priors.

3.1.1 Jeffreys' prior for a binomial proportion

Suppose that we observe Y successes out of n independent trials where the probability of success is p. Now it is easily seen that the Fisher information is

$$F(p) = \frac{n}{p(1-p)}$$

From this we see that our Jeffreys prior for p should satisfy

$$\pi(p) \propto p^{-1/2}(1-p)^{-1/2}.$$

That is our Jeffreys' prior for p is Beta(1/2, 1/2).

3.1.2 Jeffreys' prior for μ in the N(μ, 1)

Suppose we observe random observations $Y = (X_1, ..., X_n)$ from a N(μ, 1) distribution. Then in this case we can show that the Fisher information is

$$F(\mu) = n.$$

This would imply that our Jeffreys' prior must be constant over the real number line. Note that this property does not define a 'proper' probability density function in that it could not possible integrate to unity over the real number line. Nevertheless, (most) Bayesians would admit this choice of prior because when we plug it into equation (1) in 2.1.1 we obtain a proper posterior density. In general, we can use an improper prior so long as the integral of (*prior* $\times$ *likelihood*) over the parameter space is finite. Some Bayesians, known as 'proper' Bayesians, do not approve of the use of improper prior densities.

Many statisticians choose Jeffreys' priors to represent lack of knowledge when they are available. However, cases can be made for other forms of non-informative priors. Very broadly speaking, so long as there is sufficient data,

any prior which is reasonably constant over the range of values for which the likelihood is large should produce similar posterior inferences. If data are scarce, then some dependence on the particular choice of prior should be expected. This is a fact of life with Bayesian inference. It is a very good idea to experiment with a few different priors in any given situation.

3.2 Multiparameter situations - things start getting more tricky

In many practical situations there will be more than a single unknown parameter to estimate. The Bayesian approach can again be applied except that our $1-$ *dimensional* integrals considered before now become multivariate integrals. We illustrate the approach for some simple cases.

3.3 Inference for normal mean and variance jointly

Suppose that we observe a random sample $x_1, x_2, ..., x_n$ from a population that we believe to be $N(\mu, \phi)$ where the mean μ and variance ϕ are unknown. Let us suppose that we propose independent noninformative, improper, uniform priors for μ and an improper prior for ϕ that is inversely proportional to ϕ. Then $\pi(\mu, \phi) \propto \frac{1}{\phi}, -\infty < \mu < \infty, 0 < \phi < \infty$. Carry out a Bayesian analysis in order to estimate μ and ϕ.

To solve this we simply mirror the 1-dimensional case and first derive the joint posterior density of $\theta = (\mu, \phi)$. First we construct the likelihood:

$$L(\mu, \phi) \propto \phi^{-n/2} exp(-\frac{\sum(x_i - \mu)^2}{2\phi})$$

Then our posterior density $\pi(\theta|\mathbf{x})$ can be written as

$$\pi(\mu, \phi|x) \propto \phi^{-(n+2)/2} exp(-\frac{\sum(x_i - \mu)^2}{2\phi})$$

If we wish to make inference on μ and ϕ separately then we can do this from the *marginal* densities of these parameters. This requires us to integrate the above posterior with respect to ϕ and μ respectively.

When we do this (see lectures for a derivation) we can show that

$$\pi(\mu|x) \propto (1 + \frac{1}{n-1}(\frac{\bar{x} - \mu}{V^{1/2}/n^{1/2}})^2)^{-n/2}$$

where V denotes the sample variance. Comparing this with the form of the density of the t-distribution we see that *a posteriori*, $\frac{\bar{x}-\mu}{V^{1/2}/n^{1/2}} \sim t_{n-1}$. Since μ is a linear function of this quantity, then we can make posterior inferences about it quite easily.

To get the posterior density of ϕ we have to integrate the joint posterior with respect to μ. (See lecture for a derivation of this.) 'Cutting to the chase' we see that the posterior density of ϕ satisfies

$$\pi(\phi|\mathbf{x}) \propto \psi^{\frac{\mathbf{n+1}}{\mathbf{2}}} \mathbf{e}^{-\frac{\mathbf{1}}{\mathbf{2}}\psi}$$

where $\psi = \frac{(n-1)V}{\phi}$. The posterior density of ψ itself is proportional to $\pi(\phi|\mathbf{x})$ multiplied by the modulus of $\frac{d\phi}{d\psi}$. Therefore we have

$$\pi(\psi|x) \propto \psi^{\frac{n-3}{2}} e^{-\frac{1}{2}\psi}$$

from which we see that $(n-1)V/\phi \sim \chi^2_{n-1}$.

For this choice of prior for (μ, ϕ) we see that the posterior credible intervals we would calculate for either parameter would correspond exactly with confidence intervals calculated using classical methods. See section 1 for details. Priors which lead to posterior inferences that 'match' the results of classical analyses are sometimes called *probability matching priors*.

The above example is a case where things can be tackled analytically. The poterior densities for both parameters can be indentified and their properties are well known, and are tabulated in statistical tables. However, in many other situations in Bayesian analyses the resulting integrals will not be so simple. For example, consider the case where we wish to make inference on the parameters of the $\Gamma(\alpha, \beta)$ given a random sample $\mathbf{x}$. Then the likelihood for this case looks like

$$L(\alpha, \beta) \propto \frac{\beta^\alpha}{\Gamma(\alpha)} z_1^{\alpha-1} e^{-\beta z_2}$$

where z_1 and z_2 are the product and sum of the x's respectively. If α is known and we are only ignorant about β, for which we assign a Gamma prior, then we can obtain the posterior density of β with little difficulty. However, if α is unknown, then calculation of marginal posterior densities for α and β will be much harder.

Frequently, in Bayesian analysis, we will be forced to resort to computational methods of working out posterior densities, or to investigate their properties by simulating directly from them. Simulation methods are at the heart of many developments in modern statistics.

Exercises

1. Suppose that $x_1, x_2, ..., x_n$ is a random sample of observations from an $Exp(\lambda)$ distribution where λ is unknown.

   (a) Show that the Jeffreys' prior for λ in this case satisfies $\pi(\lambda) \propto \lambda^{-1}$.

   (b) For $n = 5$ and $\sum x_i = 10$ calculate a 95% equal-tailed credible interval for λ using the Jeffreys prior.

2. Show that if $x_1, x_2, ..., x_n$ is a random sample from a Poisson(λ) distribution, then the Jeffrey's prior for λ is given by $\pi(\lambda) \propto \lambda^{-1/2}$. Comment on this in the light of the connection between the Exp(λ) and the Poisson(λ) distribution.

3. An educationalist is interested in the distribution of the number of exam attempts required by individuals to qualify in a certain profession. They believe that it follows a negative binomial distribution with p.m.f.

$$f_X(x) = \binom{x-1}{r-1} p^r (1-p)^{x-r}$$

   for $x = r, r+1, r+2, ...$, where r is a postive integer and p a probability between 0 and 1. Suppose that they place a Uniform(0, 1) prior on p, an improper prior on r proportional to $1/r$ and assume that the parameters are a priori independent of each other. They select a random sample of 5 qualified individuals and count their exam attempts. These are in order: 4, 4, 6, 8, 12

   (a) Construct the likelihood $L(r, p)$ for these data.

   (b) Show that the marginal posterior probability mass function of r satisfies

$$p(r|\mathbf{x}) \propto \frac{1}{r} \frac{(5r)!(5(\bar{x} - r))!}{((r-1)!)^5 \prod_i (x_i - r)!}$$

   for $r = 1, 2, 3, 4$ and is zero for larger values of r.

   (c) Calculate (using a computer) the r.h.s. of the above expression for $r = 1, 2, 3, 4$ and use the calculated values to make inference on the value of r.

4. Suppose that a random sample of size 8 from a normally distributed population of mean μ and variance ϕ results in the values

   3.1, 3.3, 3.6, 4.2, 4.3, 4.8, 5.4, 5.7.

   Assuming that you take independent priors, constant for μ and proportional to ϕ^{-1} for ϕ, calculate:

   (a) the posterior probability that μ exceeds 5.0;

   (b) the posterior probability that ϕ is less than 1.

4 Introduction to simulation techniques

4.1 Simulating continuous random variables by transforming U(0, 1) samples

By a *simulation algorithm* we mean any computation procedure for producing a random draw from some desired probability distribution. Most computer languages have built-in (pseudo-)random number generators that produce samples that come from a Uniform(0, 1) distribution. In this section we will look at some ways in which these can be transformed to a give sample from a desired or 'target' distribution.

1. Simulating from the Exp(λ) distribution. If $X \sim U(0,1)$ then $Y = -\log(X) \sim Exp(1)$. (*Exercise*: verify this!) It follows that $Z = Y/\lambda \sim Exp(\lambda)$. Simple!

2. Simulating from the Normal distribution

First we note that if $(X, Y) \sim BVN(\mathbf{0}, \mathbf{I}_2)$ i.e. X and Y are both $N(0,1)$ and independent, then $X^2 + Y^2 \sim \chi^2_2$. Now the joint density of (X, Y) has radial symmetry (its value depends only on $r^2 = x^2+y^2$), this we can simulate a sample (x, y) from the $BVN(\mathbf{0}, \mathbf{I}_2)$ by first simulating

$$R^2 \sim Exp(1/2)$$

(using the recipe described above) and taking its square root R to obtain a range. We then obtain two independent samples from the $N(0,1)$ by selecting a random angle uniformly from $(0, 2\pi)$, as $\Theta = 2\pi Z$, where $Z \sim U(0,1)$ and calculate $X = Rcos\Theta$ and $Y = Rsin\Theta$.

3. Simulating from the $\Gamma(n, \lambda)$ (Erlang) distribution

When the first parameter of the Gamma distribution is an integer, we can exploit the fact that $X \sim \Gamma(n, \lambda)$ can be 'decomposed' as a sum of n independent $Exp(\lambda)$ random variables. A straightforward adaption of the recipe in 1 above can be used.

4. Simulating by Inversion of the Cumulative Distribution Function

A general recipe that can be applied involves the inverse of the cumulative distribution function. Suppose that X is a random variable with cdf $F_X(x)$

and that we can obtain the inverse function F_X^{-1}. Then we can show that that if $Y \sim U(0,1)$, than the cdf of $X = F_X^{-1}(Y)$ is F_X. This means that we can generate from the target distribution by simulating from the $U(0,1)$ and transforming by F_X^{-1}. Actually, recipe (1) is (more or less) an example of this approach.

Although (4) apparently gives a general recipe for simulating random variables, in many cases is it not easy to compute F_X^{-1}. Even when F_X can be easily computed (not always the case), it may be necessary to use an interative scheme to evaluate the inverse function. In these cases, it may be more efficient to simulate from the desired distribution using a different approach.

4.2 Rejection sampling methods

Rejection sampling works in the following way. Suppose that $q(x)$ is a probability density function from which we can simulate samples relatively easily and that $p(x)$ is the target density from which we would like to simulate samples. Suppose in addition that there exists a constant, $c > 0$, such that

$$\frac{p(x)}{q(x)} \leq c$$

whenever $p(x) > 0$. Then we can formulate an algorithm for simulating from $p(x)$ as follows.

1. Simulate a value, y from the density $q(x)$.
2. Calculate $k(y) = \frac{p(y)}{cq(y)}$ and simulate a random number $U \sim U(0,1)$
3. If $U < k$ accept (and report) the value y. Otherwise return to step 1 and repeat the process.

Now if X is the random variable corresponding to the first value that is accepted, the we can show that the density of X is given by $f_X(x) = p(x)$. Essentially this follows from the fact that

$$f_X(x) \propto q(x)P(U < k(x)) = q(x)\frac{p(x)}{cq(x)} = \frac{p(x)}{c}.$$

This forces $f_X(x) = p(x)$ as required.

We can see that this recipe will be very efficient if the densities $q(x)$ and $p(x)$ are approximately equal and the constant c can be chosen close to 1. In this case, k will generally be close to 1 and the value sampled in the first step above will accepted most of the time. On the other hand, if $p(x)$ and $q(x)$ are very different, then many repetitions of the algorithm may be required before a value is finally accepted.

We illustrate the use of rejection sampling to sample from the Beta(2, 2) distribution. In this case $p(x) = 6x(1-x), 0 \leq x \leq 1$. Suppose that we select $q(x)$ to be the uniform density on (0, 1). Then we can see that $p/q \leq 3/2$ and the conditions for rejection sampling hold. However, we can see that the values of $k(x)$ that arise will often be small. For example if $y < 0.25$ or $y > 0.75$ (something that will happen 1/2 of the time) then $k(y) \leq 3/4$. This suggests that there will be a significant chance of rejection in the algorithm, and that we will probably have to generate considerably more random numbers than the number of samples from $p(x)$ that we wish to obtain. For the Beta(3, 3) distribution the rejection rate would be even higher.

Note that with rejection sampling, once we have accepted n values $x_1, .., x_n$ at step 3, then these form the realised values in a random sample from the distribution. (Recall that a *random sample* consists of random variables that are identically distributed and independent of each other (i.i.d.)). This is because in rejection sampling, successive draws from the density $q(x)$ are independent of all previous draws. In the next section we shall investigate a simulation method which generates a sequence of random values $X_1, X_2, ..., X_i, ...$ where these random variables are not independent of each other.

4.3 Markov chain Monte Carlo Methods

Modern Bayesian statistics has been revolutionised by these techniques which have enabled many Bayesian problems that were previously intractable to be solved. Essentially, they allow posterior distributions to be explored by simulation and conclusions and inferences to be drawn directly from the sampled values. In this section we introduce the basic ideas behind Markov chain Monte Carlo in the context of simulating from finite, discrete probability spaces. Later in the course we will cover the theory of Markov chains in more detail and look at MCMC as applied in continuous state spaces. We begin by motivating its use by a (somewhat artificial) example.

4.3.1 The argumentative party guests

Twelve people (with surnames from A-L) are on the guest list at a very large party. They arrive at the party in a random order (with all orderings of the 12 equally likely). On arriving at the party they are spotted coming in the door by each member of the 12 already there (independently of each other) with unknown probability p. Each of the 12 note the names of the people that they spotted arriving (but not the order in which they saw them). This results in the following data The symbol '>' means "spotted on arrival".

1 > 3 6 7 11 12
2 > 3 6 8
3 > 4
4 > 7 12
5 > 6 9 11 12
6 > 8 11
7 > 8 9 11
8 > 11
9 > 11
10 > 11
11
12

The next day individuals 1, 2 and 5 have an argument each claiming to have arrived first at the party. Meanwhile 11 and 12 argue - each claiming that the other was last. You wish to help to resolve these arguments by investigating the conditional distribution of the name of the 1st arrival and the last arrival.

How can you do this? Let's try and solve this using Bayes' Theorem. There are two unknown parameters here - the particular order in which individuals arrived and the fixed, 'detection' probabability p. Let y denote the data in the above table. Let s denote an ordering of the guests which does not contradict the data in y (i.e. does not have any guest arriving before a guest who claims to have spotted them). Then for any p, I claim,

$$P(y|s) = p^{23}(1-p)^{43}$$

and is zero for any s contradicting y. It follows that the posterior probability

function of the ordering satisfies

$$\pi(s|y) \propto \pi(s) \times P(y|s) \propto 1.$$

for all orderings satisfying y since the prior $\pi(s) = \frac{1}{12!}$ (all orderings a priori equiprobable). Hence the posterior distribution of s is uniform over the set of orderings that satisfy y.

Now if we can generate samples from $\pi(s|y)$ we could investigate the marginal distribution of 'first guest' and 'last guest'. One way to sample (essentially rejection sampling) would choose random orderings independently until one satisfies y. This could take a long time since the majority would be rejected.

What is really required is some algorithm that will restrict attention only to suitable permutations, and produce a random selection from this subset (with all suitable permutations equally likely). This is what Markov chain Monte Carlo will do for us.

4.3.2 Some basic properties of Markov chains

Here we give a very basic introduction to Markov chains and Markov chain Monte Carlo. Suppose $S = \{s_1, s_2, ..., s_n\}$ is a discrete finite space (the state space). Then we can think of a Markov chain on S as a stochastic process which evolves in discrete time and takes values in S and satisfies certain additional conditions. Specifically, if X_i denotes the sequence of random states, then the distribution of X_i conditional on $(X_1, ..., X_{i-1})$ is the same as the distribution of X_i conditional on X_{i-1}. More intuitively, this says that the probability law for generating the next state in the chain from the current state is governed entirely by the current state and is independent of earlier states, given the current state.

Our finite state Markov chain is defined by the *transition matrix*. This is a $n \times n$ matrix $P = p_{ij}$ where the entry p_{ij} is the probability that, given the current state is s_i, the next state is s_j. It follows that the i^{th} row of P contains the probability function for the next state of the chain given the current state is s_i. Now suppose that the current state is drawn from a distribution, specified by some vector of probabilities $\mathbf{q}$. Then the distribution of the next state will be a mixture distribution defined by a vector of probabilities, $\mathbf{q}'$, where

$$q'_j = \sum_i q_i p_{ij}$$

In matrix notation we have that $\mathbf{q}' = \mathbf{q}P$. Now this means that the probability function of the next state is obtained from the probability function of the current state simply by post-multiplying by the transition matrix P. We can also see that in order to work out the distribution of the state reached after k iterations of the chain we need only calculate $\mathbf{q}P^k$. *Does this remind you of anything to do with eigenvalues/eigenvectors?* From courses in linear algebra you may remember that repeated multiplication (followed by normalisation) is one method for determining the eigenvector corresponding to the maximum eigenvalue and associated eigenvector of a matrix.

Now, if a transition matrix, P has a unique eigenvalue of modulus 1 which takes the value 1, then there is a unique distribution,π, that satisfies

$$\pi P = \pi$$

We refer to π as the *stationary distribution* of the Markov chain. We will also have that for any initial distribution $\mathbf{q}$, then $\mathbf{q}^{(k)} = \mathbf{q}P^k$ tends to π as k becomes large. Thus, if we set our initial state to be s_j for any j, then the distribution of the state we reach after k iterations can be made arbirarily close to π, just by taking k to be sufficiently large. To generate a sample from the distribution π all we need to do is start our chain in any state and iterate a suitably large number of times, where an 'iteration' involves drawing a sample from the distribution $\mathbf{p_i}$ to obtain the next state from the current state s_i.

In order to check whether a given Markov chain has these properties (*i.e.* that $\mathbf{q}^{(k)} = \mathbf{q}P^k$ tends to π as k becomes large), we need only verify that the following conditions hold.

1. $\pi P = \pi$
2. The Markov chain is *irreducible*. This means that it is possible to reach any state s_j from any other state s_i within a finite number of transitions.
3. The Markov chain is *aperiodic*. That is, the state space cannot be partitioned into disjoint subsets $S_1, S_2, .., S_m$ for which the Markov chain is constrained to cycle through these subsets from iteration to iteration.

In the practical design of Markov chains it is usually easy to ensure the last condition of aperiodicity. The property of irreducibility is sometimes not

trivial to verify. As regards the first property, we can sometimes show this by demonstrating that a stronger property holds.

Definition Let π denote a distribution and P denote a transition matrix. Suppose that for all i, j,

$$\pi_i p_{ij} = \pi_j p_{ji}$$

Then we say that chain specified by P satisfies *detailed balance* with respect to π. Moreover, it follows that $\pi P = \pi$. This is easily seen since the j^{th} component of πP is $\sum_i \pi_i p_{ij} = \sum_i \pi_j p_{ji}$, if detailed balance holds. Now the j^{th} row of P, being a probability function, sums to unity and the j^{th} component of πP is therefore π_j as required.

In MCMC we are concerned with designing Markov chains whose stationary distributions coincide with the target distribution. One of the simplest recipes for doing this is the *Metropolis-Hastings algorithm.*

4.3.3 The Metropolis-Hastings Algorithm

We use the same notation as previous sections and consider a finite state-space on which we have a target distribution π from which we would like to simulate. We define a Markov chain with stationary distribution π, in the following way.

Suppose that our state after t iterations is $s^t = s_i$. Then we propose a new state $s' = s_k$ from a distribution $\mathbf{q}_i$ known as the *proposal distribution.* To get our state at time $t+1$ we must either accept the proposed state (next state is $s^{t+1} = s'$) or reject it (next state is $s^{t+1} = s_i$).

The proposed state is accepted with some probability

$$p_{acc} = min\{1, \frac{q_{ki}\pi_k}{q_{ik}\pi_i}\}$$

This 2-stage process defines a Markov chain with a transition matrix P where the elements of P can be derived from the proposal transition matrix $Q = (\mathbf{q}_1, \mathbf{q}_2, ...)$ and the acceptance probability above. Now we can show that P satisfies detailed balance with respect to π. If, moreover, we can show that the chain is irreducible and aperiodic then we are justified in using it to generate samples from π.

Example 1 Consider a clock with numbers 1, 2, 3, 4 arranged cyclically. Suppose that we wish to generate samples from a uniform target distribution

on 1, 2, 3, 4. A Metroplis-Hastings algorithm can be formulated as follows. If at state i then propose moving to state (i+1)(mod 4) or (i-1)(mod 4) with probability 1/2. Then accept the proposed state with the above acceptance probability which in this case is 1 since (in the above notation) $q_{ik} = q_{ki} = 1/2$ and $\pi_k = \pi_j = 1/4$. It is clear that this chian is irreducible, since we can reach any state from any other in a finite number of steps. However, it is not aperiodic, since if we start at position 1, for example, we are constrained to cycle between the sets 2, 4 and 1, 3 after odd and even numbers of iterations.

It is straightforward to remove this periodicity by using a proposal distribution whereby, with probability p we propose the current state (i.e. no move) and with probability $(1 - p)/2$ propose steps clockwise and anticlockwise.

Example 2 Another way of sampling uniformly from 1, 2, 3, 4 is to formulate a Markov chain that can only move to adjacent states with the states arranged in a line. This means that if we are in states 1 or 4 we can only propose moves to 2 and 3 respectively. Consider the following proposal transition matrix:

$$\begin{pmatrix} 0 & 1 & 0 & 0 \\ 1/2 & 0 & 1/2 & 0 \\ 0 & 1/2 & 0 & 1/2 \\ 0 & 0 & 1 & 0 \end{pmatrix}$$

where the i^{th} row gives the proposal distribution if the current state is i. Now in the corresponding Metropolis-Hastings sampler, we see that the the acceptance probabilities are not all 1. Proposed moves between 2 and 3 are always accepted, proposed moves to 1 and 4 are always accepted, but moves from 1 and 4 are only accepted with probability 1/2.

Note that this Metropolis-Hastings sampler is aperiodic since there are states from which the probability of moving at the next iteration is less than 1.

Example 3 We can construct a Markov chain whose state space is the set of orderings of n objects and whose stationary distribution is uniform on this set of orderings. If the current state is the ordering $\mathbf{s} = (i_1, i_2, i_3, ..., i_n)$ then we could propose the next state by selecting a position, j, uniformly from 1, 2, 3, ..., $n - 1$ and swapping the positions of i_j and i_{j+1}. we can check that the acceptance ratio is 1 here. Of course the Markov chain will not be

aperiodic, but this could be assured by allowing a non-zero probability of 'staying put'.

For other practical examples of Markov chain Monte Carlo for choosing random orderings see various methods of shuffling packs of cards.

Returning to the problem of selecting the ordering of party guests consistent with the 'inequalities', we see that it could be solved by constructing a Markov chain whose state space is the set of allowable orderings and whose stationary distribution is uniform on this set. In the assignment (see later) you are asked to design such a chain, implement it in a programming language and then use it to answer some questions on the target distribution.

The M-H algorithm is particularly useful in Bayesian statistics because the target density π only enters the calculations through the acceptance ratio. Now this means that you only need to know the relative values of the density (since any unknown normalising constant would cancel from the ratio). Knowing that a posterior density is $\pi(\theta|y) \propto \pi(\theta)L(\theta|y)$ is sufficient to formulate a M-H sampler to investigate it. For example, consider the second part of Qu 5 in Part 1. Here we observed a single sample, $y = 2$, from a $NBin(m, 1/2)$ distribution (i.e. the distribution of the number of heads accrued when a fair coin is tossed until m tails are achieved) where the prior for m was Geometric(1/3).

Exercise. Design and implement a M-H sampler to generate samples from $\pi(m|y = 2)$ for this example.

4.3.4 Using Markov chain output to draw conclusions

Having designed a Markov chain with the desired stationary distribution, the next step is usually to implement it in a computer programme, select a starting state and interate the chain by simulation. Suppose that the random sequence of states visited by the chain is $s^1, s^2, s^3, ...$ Now it is true that the *marginal distribution* of s^i can be made as close to the target distribution π as desired simply by choosing i to be sufficiently large. However the states are not independent of each other. Depending on the particular chain, it is likely that s^{i+1} will be 'close' to s^i in some sense. If we wish to obtain a random sample (i.e. a set of independent observations from π) then we

might wish to 'thin' the chain. This means outputing the state at every m^{th} iteration, where the m is chosen to be sufficiently large so that s^i and s^{i+m} are more or less independent.

The smaller the value of m required for this, the better the *mixing qualities* of the chain. In practice, time-series measures, such as autocorrelation, can be used to assess how 'independent' a sequence of samples appears to be.

If we wish to investigate properties of π by investigating histograms, it is not necessary to thin the chain. Even if successive samples are not independent, we can use the frequency with which any state occurs in the unthinned chain to estimate the probability of that state according to π. However, the standard error in such an estimate will, in reality, be larger than that predicted assuming independence of samples. In any case, it is normal to discard a number of iterations at the beginning of a simulation in case the initial state selected is 'atypical' of π. This initial period is sometimes referred to as the $burn - in$ period.

Exercises

1. *Simulating from the Cauchy distribution.* The Cauchy distribution is defined by the density

$$f_X(x) = \frac{1}{\pi}\frac{1}{1+x^2}, -\infty < x < \infty$$

   (a) By first deriving the c.d.f. of X and its inverse function, describe how samples from a U(0, 1) random number generator could be transformed to give samples from the Cauchy distribution.

   (b) By appealing to the circular symmetry of the standard bivariate normal distribution, show how samples from a Cauchy distribution could be generated from independent N(0, 1) samples.

   (*Comment.* The Cauchy distribution is the same as the t_1 distribution.)

2. *Simulating from the Beta distribution.*

   (a) Show how U(0, 1) random variates can be transformed by inversion of the c.d.f. to generate samples from a Beta(n, 1) distribution.

   (b) How would you generate samples from a Beta(2, 2) distribution by inversion of the c.d.f.?

   (c) Given that if $X \sim Gamma(n, 1)$ and $Y \sim Gamma(m, 1)$ are independent where n and m are positive integers then $\frac{X}{X+Y} \sim Beta(n, m)$, describe an algorithm for simulating samples with a Beta(m, n) distribution from independent samples from a U(0, 1) random number generator.

3. *Simulating from the Beta distribution using rejection sampling.* Design an algorithm to simulate samples from the $Beta(\alpha, \beta)$ distribution where $\alpha, \beta > 1$, using the U(0, 1) as the density $q(x)$,

   (a) Derive an expression for the probability that a value generated from the U(0, 1) is accepted for your algorithm.

   (b) How does this expression behave as α and β become large?

4. *Assignment* Consider the problem regarding the party guests presented in 4.3.1.

   (a) Prove carefully that $P(y|s) = p^{23}(1-p)^{43}$ and identify the posterior density $\pi(p|y)$ assuming a suitable non-informative prior for p (which is is independent of the unknown ordering s).

   (b) Design a Markov chain using Metropolis-Hastings (or any appropriate MCMC method):

   - whose state space, S, is the set of orderings of the 12 guests that are consistent with the constraints presented in 4.3.1.
   - whose stationary distribution, π, is uniform on this state space.

   (c) Verify that the chain you construct satisfies detailed balance, is aperiodic and irreducible.

   (d) Give a description of your algorithm in the form of a flow chart or pseudo-code and include a copy of source code used for the computations.

   (e) Use your chain to carry out the following computations.

   - For each guest, give an estimate of the marginal distribution of their position in a random ordering generated from π.
   - Estimate the distribution of the random variables 'number of first guest' and 'number of last guest'.

   (f) Comment on how well your chain 'mixes' e.g. by showing time-series plots of outputs, or through calculation of autorcorrelation functions of thinned outputs.

   (g) Suppose now that a guest's number indicate their ranking in order of height (with 1 the tallest) and, as a result, each has a different probability p_i of spotting any guest on arrival. Suppose further that these probabilities $p_1, \ldots, p_{12}$, are

   0.7, 0.65, 0.6, 0.55, 0.5, 0.45, 0.4, 0.35, 0.3, 0.25, 0.2, 0.15

   respectively. Adapt your Markov chain to investigate the distribution of first and last guest to arrive under this new assumption regarding detection probabilities. How do your conclusions change from the constant p case?

Your report should include a clear description of any mathematical arguments. You are encouraged to present results in graphical as well as tabular format. You should identify any statistical packages used. I am happy to give advice on any aspect of the assignment. Please contact me by email (gavin@ma.hw.ac.uk) or phone (0131-451-3205) if you have any queries.

The assignment should be handed in on or before Thursday 24th March.