

MAS3301 Bayesian Statistics

M. Farrow
School of Mathematics and Statistics
Newcastle University

Semester 2, 2006-7

1 Introduction to Bayesian Inference

1.1 What is “Bayesian inference”?

1.1.1 Inference

Some may have expected the word “statistics” rather than “inference.” I have nothing against “statistics”, of course, but it may give too narrow an impression. It might suggest that some things, e.g. decision making under uncertainty, manipulation of uncertainty when there may appear to be few, if any, “data” (e.g. in probabilistic expert systems), are excluded. “Statistics” might also have connotations for some people which I would prefer to avoid. It is necessary to get rid of preconceptions.

In fact, the wider topic of “Bayesian analysis” deals with problems which involve one or both of the following.

- Inference. That is learning from data, or perhaps just the propagation of uncertainty or information among a collection of “unknowns.”
- Decision. Choosing an action. This falls within the scope of Bayesian analysis when the decision is made under uncertainty.

In this module we will be more concerned with inference than with decision but the two are closely linked.

1.1.2 “Bayesian”

Why do we use the word “Bayesian”? The truth is that “Bayesian” means slightly different things to different people. Even among people who call themselves “Bayesians” there are philosophical differences. (However these are generally small compared to the difference with non-Bayesians!) In this course we will take a fairly mainstream position, that is mainstream among Bayesians. We will adopt a subjectivist view of probability. (The word “Bayesian” is sometimes used to describe things which are not really Bayesian in the sense used here. Beware.)

So, what makes an analysis “Bayesian”?

- Full probability specification. The state of uncertainty over all of the unknown quantities (and statements) involved is described by a joint probability distribution.
- Probability represents “degree of belief.”
- Probability is subjective.

(Note that, at one end of the “Bayesian” spectrum there is subjectivist work which uses expectation directly, rather than probability, and, at the other end, there are attempts to use “Bayesian methods” in a non-subjective context).

Where does Bayes come in? Bayes’ theorem turns out to be crucial for many inference problems if we adopt this view. This is particularly true for the traditional “statistical” problems where Bayes’ theorem is more or less always required. Thomas Bayes (1702-1761) was a Presbyterian minister in Tunbridge Wells, Kent. In 1763 a paper by Bayes was published posthumously by his friend Richard Price. This paper gave us Bayes’ theorem. Bayes’ theorem tells us how to turn round conditional probability so that the conditioning is in the opposite direction. We shall see why this is important.

1.2 Motivational example 1

The university where I once worked has two main campuses, St. Peter’s and Chester Road. My office was at St. Peter’s but, from time to time, I needed to visit Chester Road. Very often I wanted to do the journey in such a way that as little time as possible was spent on it. (Sometimes my requirements were slightly different. It might be that I had to arrive before a certain time, for example). The university’s Campus Bus was scheduled to leave St. Peter’s at 15 and 45 minutes past each hour and arrive at Chester Road seven minutes later. Of course, in practice, these times

varied a little. I guessed it took me about two minutes to leave my office and go to the bus stop. On the other hand it took around 15 minutes to do the whole journey on foot. At what time should I leave the office on a journey to Chester Road, bearing in mind all these uncertainties and others, such as the possibility that my watch is wrong? If I left too early I would waste time waiting for the bus. If I missed the bus I either had to walk or return to my office and try again later, having wasted some time. If I went to the bus stop, arriving close to the scheduled departure time and there was nobody around, should I wait? Should I walk? Should I return to my office?

In order to try to answer the questions in the preceding paragraph, we need some way to deal with uncertainty. The approach taken in this course is *subjective*. That is, uncertainty is regarded as a feature of a person's beliefs (in the example, my beliefs) rather than of the world around us. I need to make the decision and I try to do it in a way which is rational and consistent with my beliefs. I need to be able to compare what I expect to happen under each of my possible decisions. I need a way to use the evidence available to me when I arrive at the bus stop to inform me about whether it is likely that the bus has already left. We might attempt to analyse the problem by constructing a model, for example a model for the arrivals and departures of the bus. Such a model would involve probability because the behaviour of the bus varies in a way which is not entirely predictable to us. The model would involve parameters. We may be uncertain about the values of these parameters. We may also be uncertain about the validity of assumptions in the model. We can use probability to deal with all of these uncertainties provided we regard probability as degree of belief.

The Campus Bus example may seem to be of little importance but it has features which may be found in many practical problems which may involve very important consequences.

1.2.1 Motivational Example 2

A company wants to know what kinds of people buy its product so that advertising may be aimed at the right people. Suppose that the company can observe various attributes of a sequence of people, including whether or not they buy the product. This could be by a survey or, perhaps, by monitoring visits to the company's web site.

Here we wish to learn about the association of attributes with the particular attribute "buys the product." We want to do this so that we will be able to make predictions about whether or not other people will buy the product, based on their attributes. So, we need to express our prior beliefs about these things, including our uncertainties about the associations and our beliefs about how the people we observe are related to the more general population of people who might buy the product. Is there, for example, anything unusual about the people we observe or are all people, including those whom we observe, "exchangeable"?

2 Beliefs and uncertainty

2.1 Subjective probability

We are dealing with what has been called "the logic of uncertainty." Some may ask, "Why should we use probability?" However this is really the wrong question. It probably betrays a misunderstanding about what Bayesians mean by "probability." Bruno de Finetti, a hero of Bayesianism, wrote in his classic book, *The Theory of Probability*, that "**Probability does not exist.**" It has no *objective* existence. It is not a feature of the world around us. It is a measure of degree of belief, your belief, my belief, someone else's belief, all of which could be different. So we do not ask, "Is probability a good model?" We start by thinking about how "degree of belief" should behave and derive the laws of probability from this. We say that, if we obey these rules, we are being *coherent*. (See below) Note then that this is a normative theory not a descriptive theory. We are talking about how a "rational" person "ought" to behave. People may behave differently in practice!

Students of MAS3301 will have come across expectation and probability earlier, of course, but here we will briefly present these concepts in Bayesian/subjectivist terms.

2.2 Expectation

2.2.1 Definition

Consider a quantity, X , whose value is currently unknown to us but which will, at least in principle, be revealed to us eventually. E.g. the amount of money in my coat pocket, the height in cm., of the next person to walk along the corridor.

Rough definition: Suppose you could have either $\pounds X$ or $\pounds c$ where c is a fixed number. We choose c so that you can not choose between $\pounds X$ and $\pounds c$. Then $c = E(X)$, your *expectation* or *expected value* of X . Notice that this is subjective.

Tighter definition: The rough definition is not rigorous because of difficulties with the values people assign to gambles. (This is the subject of the theory of *utility*). Imagine instead that you will lose $\pounds(c - X)^2$ and you can choose c . The value of c you choose is your expectation, $E(X)$, of X .

There are some quantities, e.g. the mean weight of eggs, which we will never observe. However we can express expectations about them because of the equivalence of these expectations with expectations about observable quantities, e.g. the weight of an egg.

We can define expectations for functions of X , e.g. X^2 . We can use this fact to indicate our uncertainty about the value of X .

2.2.2 Variance and Covariance

The *variance* of X is $E\{[X - E(X)]^2\} = \text{var}(X)$. The *standard deviation* of X is $\sqrt{\text{var}(X)}$. A large variance means that we are very uncertain about the value of X .

We call $E(X)$ the “*price*” of X in the sense that we would have no preference between $\pounds E(X)$ and $\pounds X$ (but recall comment on utility theory). Clearly the “price” of $aX + bY$, where a and b are constants and X and Y are both unknowns is $aE(X) + bE(Y)$. It follows that

$$\begin{aligned}\text{var}(X) &= E\{[X - E(X)]^2\} \\ &= E\{X^2 - 2XE(X) + [E(X)]^2\} \\ &= E(X^2) - 2[E(X)]^2 + [E(X)]^2 \\ &= E(X^2) - [E(X)]^2\end{aligned}$$

and

$$\begin{aligned}\text{var}(aX + bY) &= E\{[aX + bY - aE(X) - bE(Y)]^2\} \\ &= E\{a^2X^2 + b^2Y^2 + a^2[E(X)]^2 + b^2[E(Y)]^2 + 2abXY \\ &\quad - 2a^2XE(X) - 2abXE(Y) - 2abYE(X) - 2b^2YE(Y)\} \\ &\quad + 2abE(X)E(Y) \\ &= a^2\text{var}(X) + b^2\text{var}(Y) + 2ab\text{covar}(X, Y).\end{aligned}$$

where

$$\text{covar}(X, Y) = E(XY) - E(X)E(Y) = E\{[X - E(X)][Y - E(Y)]\}$$

is called the *covariance* of X and Y . We also talk about the *correlation* of X, Y which is defined as $\text{covar}(X, Y)/\sqrt{\text{var}(X)\text{var}(Y)}$.

2.3 Probability

For non-quantitative unknowns, such as the occurrence of a particular event or the truth of a particular statement, we can introduce indicator variables. For example $I_R = 1$ if it rains tomorrow, $I_R = 0$ otherwise. We call $E(I_R)$ the *probability* that it rains tomorrow, $\text{Pr}(\text{rain})$. For a discrete quantitative variable we can have, e.g., $I_3 = 1$ if $X = 3$, $I_3 = 0$ otherwise. For continuous quantitative variables we can have, e.g., $I_x = 1$ if $X \leq x$, $I_x = 0$ otherwise so $E(I_x) = \text{Pr}(X \leq x) = F_X(x)$. Thus we can evaluate a probability distribution for an unknown quantity and $\text{Pr}(X \leq x)$ becomes our degree of belief in the statement that $X \leq x$.

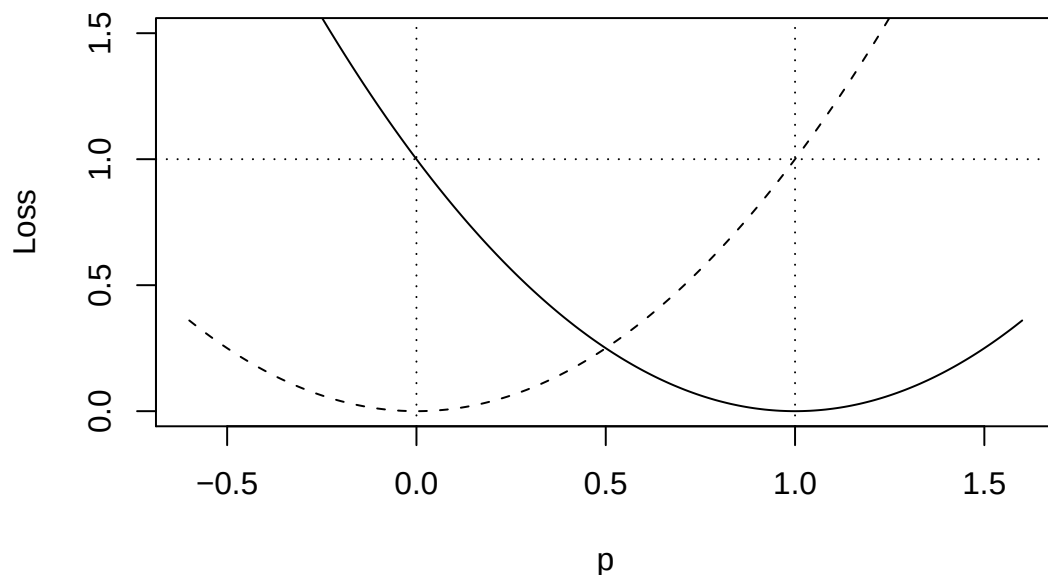


Figure 1: Losses if S is true (solid line) and if S is false (dashes) as a function of declared probability p .

2.4 Coherence

2.4.1 Concept

Can we assign our subjective probabilities to be whatever we like? No. We impose the conditions of “*coherence*”. These rule out beliefs which would appear to be irrational in the sense that they violate the “sure loser principle”. This principle states that we should not hold “beliefs” which would force us to accept a series of bets which was bound to make us lose. The usual rules of probability can all be derived from the idea of coherence. We omit the details, except for the following simple example (2.4.2).

2.4.2 Example

Consider the statement S : “It will rain tomorrow.”

Roughly: You pay me $\mathcal{L}\Pr(S)$ and, in return, I give you $\mathcal{L}1$ if S is true and $\mathcal{L}0$ if it is not.

More carefully: You pay me $\mathcal{L}k(1-p)^2$ if S true and $\mathcal{L}kp^2$ if not. You get to choose p . Clearly $0 \leq p \leq 1$.

If $p < 0$ then $p = 0$ would always be better,

If $p > 1$ then $p = 1$ would always be better,

whether or not S true.

This is illustrated in figure 1. this shows the curves p^2 and $(1-p)^2$ plotted against p . That is, it shows the losses if S is false and if it is true plotted against our choice of p . We can see that, to the left of 0 and to the right of 1, both curves increase. So, it is always better to have a value of p such that $0 \leq p \leq 1$.

2.5 A simple example

Let us consider a very simple illustrative example. Suppose we are interested in the proportion of animals in some population which carry a particular gene and we have a means of determining whether any given animal does. (Very similar considerations will apply, of course, in many kinds

		Second animal		
		YES	NO	
First Animal	YES	0.06	0.14	0.2
	NO	0.14	0.66	0.8
		0.2	0.8	1.0

Table 1: Probabilities for possession of a particular gene in two animals.

of repeatable experiments with two possible outcomes). Suppose that we are going to test just two animals. There are only four possible outcomes and these are represented in table 1 which also assigns a probability to each of the possible outcomes. For now just regard these as the subjective probabilities of someone who has thought about what might happen.

The table also gives the *marginal* probabilities for each animal individually. We see that these are the same. The animals are said to be *exchangeable*.

Let S_1 be the statement, or *proposition*, that the first animal has the gene and S_2 be the proposition that the second animal has the gene. We write $\Pr(A)$ for the probability of a proposition A . We see that $\Pr(S_1) = \Pr(S_2) = 0.2$. Some readers will have noticed, however, that S_1 and S_2 are not *independent* since the probability of S_1 and S_2 , $\Pr(S_1 \wedge S_2) = 0.06 > \Pr(S_1)\Pr(S_2) = 0.2 \times 0.2 = 0.04$. This is quite deliberate and reflects the fact that we do not know the underlying population proportion of animals with the gene. This is what makes it possible to learn from observation of one animal about what we are likely to see in another.

The *conditional probability* that the second animal has the gene given that the first does is

$$\Pr(S_2|S_1) = \frac{\Pr(S_1 \wedge S_2)}{\Pr(S_1)} = \frac{0.06}{0.2} = 0.3$$

Now, what happens to our beliefs about the second animal if we actually observe that the first has the gene? Well, unless something else causes us to change our minds, our probability that the second animal has the gene should become the conditional probability that the second animal has the gene given that the first does. That is 0.3. There are deep philosophical arguments about whether you should *have* to adjust your beliefs in this way. You could have coherent beliefs before you see the first animal and a different set of coherent beliefs afterwards so that, at any point in time you are coherent. This Bayesian updating of beliefs requires beliefs before and after the observation to cohere with each other. Suffice it to say that, if you do not adjust beliefs in this way, you need to ask yourself why. It was the rational thing to do *before* you saw the data. If you programmed a computer to take observations and make decisions for you, and, of course, scientists are increasingly doing just this sort of thing, then, in order to be coherent at the moment you switch it on and leave it to run, you must program it to update beliefs in this way.

This adjustment of probabilities from *prior* probabilities to *posterior* probabilities by conditioning on observed data is fundamental to Bayesian inference.

The uncertainties involved in this example can be thought of as falling into two categories. Some uncertainty arises because we do not know the underlying population proportion of animals with the gene. There is only one correct answer to this but we do not know it. Such uncertainty is described as *epistemic*. There is also uncertainty to do with the selection of animals to test. We might happen to choose one with the gene or one without. Uncertainty caused by such “randomness” is described as *aleatory*. Similarly, in identifying a protein from a mass spectrum there is epistemic uncertainty in that we do not know the true identity of the protein and aleatory uncertainty in the various kinds of “noise” which affect the observed spectrum.

2.6 Bayes’ Theorem

Let $S_1, \dots, S_n$ be events which form a *partition*. That is $\Pr(S_1 \vee \dots \vee S_n) = 1$ and $\Pr(S_i \wedge S_j) = 0$ for any $i \neq j$. In other words one and only one of the events $S_1, \dots, S_n$ must occur or one and only one of the statements $S_1, \dots, S_n$ must be true. Let D be some other event (or statement). Provided $\Pr(S_i) \neq 0$, the conditional probability of D given S_i is

$$\Pr(D | S_i) = \frac{\Pr(S_i \wedge D)}{\Pr(S_i)}.$$

Hence the joint probability of S_i and D can be written as

$$\Pr(S_i \wedge D) = \Pr(S_i) \Pr(D \mid S_i).$$

Then the *law of total probability* says that

$$\begin{aligned} \Pr(D) &= \sum_{i=1}^n \Pr(S_i \wedge D) \\ &= \sum_{i=1}^n \Pr(S_i) \Pr(D \mid S_i) \end{aligned}$$

The conditional probability of S_k given D is

$$\begin{aligned} \Pr(S_k \mid D) &= \frac{\Pr(S_k \wedge D)}{\Pr(D)} \\ &= \frac{\Pr(S_k) \Pr(D \mid S_k)}{\sum_{i=1}^n \Pr(S_i) \Pr(D \mid S_i)}. \end{aligned}$$

2.7 Problems 1

Answer all questions. Solutions are to be submitted in the Homework Box no later than 5.00pm on Wednesday February 14th.

1. Let E_1, E_2, E_3 be events. Let I_1, I_2, I_3 be the corresponding indicators so that $I_1 = 1$ if E_1 occurs and $I_1 = 0$ otherwise.

- (a) Let $I_A = 1 - (1 - I_1)(1 - I_2)$. Verify that I_A is the indicator for the event A where $A = (E_1 \vee E_2)$ (that is “ E_1 or E_2 ”) and show that

$$\Pr(A) = \Pr(E_1) + \Pr(E_2) - \Pr(E_1 \wedge E_2)$$

where $(E_1 \wedge E_2)$ is “ E_1 and E_2 ”.

- (b) Find a formula, in terms of I_1, I_2, I_3 for I_B , the indicator for the event B where $B = (E_1 \vee E_2 \vee E_3)$ and derive a formula for $\Pr(B)$ in terms of $\Pr(E_1), \Pr(E_2), \Pr(E_3), \Pr(E_1 \wedge E_2), \Pr(E_1 \wedge E_3), \Pr(E_2 \wedge E_3), \Pr(E_1 \wedge E_2 \wedge E_3)$.
2. In a certain place it rains on one third of the days. The local evening newspaper attempts to predict whether or not it will rain the following day. Three quarters of rainy days and three fifths of dry days are correctly predicted by the previous evening’s paper. Given that this evening’s paper predicts rain, what is the probability that it will actually rain tomorrow?
 3. A machine is built to make mass-produced items. Each item made by the machine has a probability p of being defective. Given the value of p , the items are independent of each other. Because of the way in which the machines are made, p could take one of several values. In fact $p = X/100$ where X has a discrete uniform distribution on the interval $[0, 5]$. The machine is tested by counting the number of items made before a defective is produced. Find the conditional probability distribution of X given that the first defective item is the thirteenth to be made.
 4. There are five machines in a factory. Of these machines, three are working properly and two are defective. Machines which are working properly produce articles each of which has independently a probability of 0.1 of being imperfect. For the defective machines this probability is 0.2.

A machine is chosen at random and five articles produced by the machine are examined. What is the probability that the machine chosen is defective given that, of the five articles examined, two are imperfect and three are perfect?

3 Parameters and likelihood

3.1 Introduction

Now, from where did I get the probabilities in table 1? Well, I could have thought of them directly. There are, after all, only four possible outcomes. However, in fact I thought less directly. Especially when the number of possible outcomes starts to increase it becomes helpful to structure our thoughts by introducing *parameters*. We also often make use of ideas such as that all individuals in a group appear, in some way, “the same” to us.

Introduce unknowns T_1, T_2 such that $T_i = 1$ if S_i and $T_i = 0$ otherwise (i.e. if $\bar{S}_i$, where this denotes “not S_i ”). Consider a quantity θ such that, *if we knew the value of θ* , then $\Pr(S_1) = \Pr(S_2) = \theta$ and T_1, T_2 are *independent* given θ . We can think of θ as the unknown proportion of animals that have the gene. We can represent this situation in a diagram or *graphical model*. There are various kinds of graphical model but the diagram shown here is a *directed acyclic graph* or DAG. The nodes are connected by directed arrows (called *arcs* or *edges*) and it is impossible to find a path along the directions of the arrows which returns to where you started. The term *influence diagram* is also sometimes used for such graphs. See figure 2.

In this case we would say that, in terms of possessing the gene, the animals are *exchangeable*. That is T_1 and T_2 are exchangeable.

Suppose that, before we observe either animal, we believe that θ can have only two values, 0.1 and 0.4, and we give these probabilities of 2/3 and 1/3 respectively. This is called a *prior* probability distribution for θ . (It is, of course, unrealistic to allow only two values but we will put this right very soon).

	$\theta = 0.1$	$\theta = 0.4$
Prior probs.	$\pi_1 = 2/3$	$\pi_2 = 1/3$
$\Pr(S \theta)$	0.1	0.4

We can now easily work out the joint probabilities for θ, T_1, T_2 .

$$\Pr(\theta, T_1, T_2) = \Pr(\theta) \Pr(T_1|\theta) \Pr(T_2|\theta)$$

Notice how this relates to the diagram.

We can find the probabilities in the original table by summing over the distribution of θ .

$$\begin{aligned} \Pr(S_1 \wedge S_2) &= (2/3) \times 0.1^2 + (1/3) \times 0.4^2 = 0.06 \\ \Pr(S_1 \wedge \bar{S}_2) &= (2/3) \times 0.1 \times 0.9 + (1/3) \times 0.4 \times 0.6 = 0.14 \\ \Pr(\bar{S}_1 \wedge S_2) &= (2/3) \times 0.9 \times 0.1 + (1/3) \times 0.6 \times 0.4 = 0.14 \\ \Pr(\bar{S}_1 \wedge \bar{S}_2) &= (2/3) \times 0.9^2 + (1/3) \times 0.6^2 = 0.66 \\ \Pr(S_1) &= (2/3) \times 0.1 + (1/3) \times 0.4 = 0.2 \end{aligned}$$

Now, if we observe the first animal, we can apply Bayes’ rule to find the *posterior* probabilities for the values of θ .

$$\begin{aligned} \Pr(\theta = 0.1|S_1) &= \frac{\pi_1 \times 0.1}{\pi_1 \times 0.1 + \pi_2 \times 0.4} \\ &= \frac{(2/3) \times 0.1}{(2/3) \times 0.1 + (1/3) \times 0.4} = \frac{1}{3} \end{aligned}$$

Similarly $\Pr(\theta = 0.4|S_1) = 2/3$. In fact we can deduce the following simple form of Bayes’ rule.

$$\text{“Posterior} \propto \text{Prior} \times \text{Likelihood”}$$

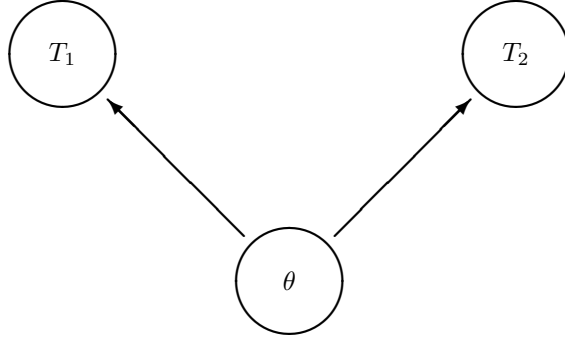


Figure 2: Graphical model for animals example

Here the “likelihood” is $\Pr(S_1|\theta)$.

Now the information propagates through θ to T_2 and the probability of S_2 becomes $(1/3) \times 0.1 + (2/3) \times 0.4 = 0.3$.

3.2 A slightly less simple example

We keep the same example except that this time there are twenty animals. We can use the same diagram as above except that there are twenty nodes $T_1, \dots, T_{20}$.

What happens if we observe 3 animals with the gene out of twenty observed?

The likelihood is proportional to $\theta^3(1-\theta)^{17}$. The likelihood function, scaled so that its maximum value is 1.0, is shown in figure 3. The maximum of this function occurs at $\theta = 3/20$.

The posterior probabilities are proportional to:

$$\begin{aligned} (2/3) \times 0.1^3 0.9^{17} &= 1.1118 \times 10^{-4} \\ (1/3) \times 0.4^3 0.6^{17} &= 3.61 \times 10^{-6} \end{aligned}$$

So the posterior probability that $\theta = 0.1$ is

$$\Pr(\theta = 0.1|D) = \frac{1.1118}{1.1118 + 0.0361} = 0.969$$

We can now find predictive probabilities for future animals. E.g. the probability that two new animals both have the gene:

$$0.969 \times 0.1^2 + 0.031 \times 0.4^2 = 0.01465.$$

In effect we are adding two extra nodes, T_{21}, T_{22} to our diagram but not observing these.

3.3 The example with a continuous prior distribution

More realistically we might have a continuous prior probability distribution for θ .

Prior p.d.f. $f_0(\theta)$.

Posterior p.d.f.:

$$\frac{f_0(\theta)\theta^3(1-\theta)^{17}}{\int_0^1 f_0(\theta)\theta^3(1-\theta)^{17}.d\theta}$$

The integration is straightforward in this case. In more complicated cases, especially when there are many unknowns, the integration has been the main computational problem. In the past this was a major obstacle to the use of Bayesian inference. In recent years huge progress has

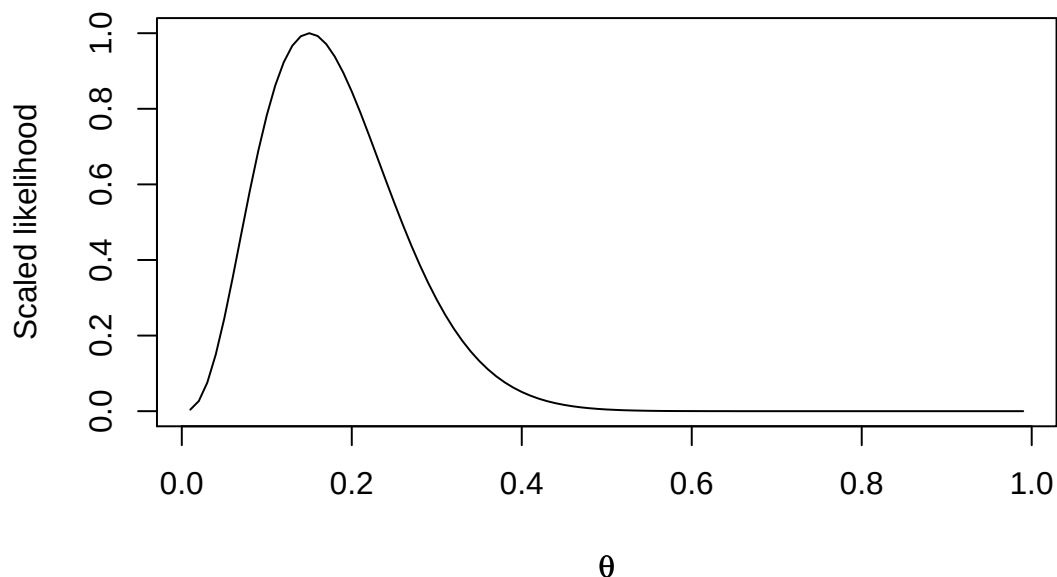


Figure 3: Scaled likelihood in the animals' gene example, with 3 animals out of twenty observed with the gene.

been made and we can now routinely handle complicated problems with very large numbers of unknowns.

In fact the calculation is particularly simple if we use a *conjugate* prior distribution. This simply means a prior distribution which “matches” the likelihood in the sense that the posterior distribution belongs to the same family as the prior distribution. In this example the conjugate family is the family of beta distributions. The prior p.d.f. is proportional to

$$\theta^{a-1}(1-\theta)^{b-1}$$

where we specify the values of a and b .

The posterior p.d.f. is then proportional to

$$\theta^{a+3-1}(1-\theta)^{b+17-1}.$$

The *prior mean* is

$$\frac{a}{a+b}$$

and the *posterior mean* is

$$\frac{a+3}{a+b+20}.$$

We might, of course, feel that our prior beliefs are not represented by a conjugate distribution. In this case we could use a different distribution and employ numerical methods of integration. We might however be able to use a *mixture* of conjugate prior distributions to approximate the shape we want. The posterior is then a mixture of the conjugate posterior distributions, although the weights are changed. This is the method used by the computer package *First Bayes*. We will consider this in more detail in a later lecture.

4 Probability in Models

4.1 Deterministic and stochastic models

We can apply the ideas of probability to “mathematical” models. We often distinguish between “deterministic” models, in which random variables do not play a part in the model itself, and “stochastic” models, in which they do. Clearly probability is involved in stochastic models but it may also be relevant in deterministic models.

In a deterministic model we suppose that, if we knew the values of all of the parameters, starting conditions etc. (and if the model is “correct”), we could predict the behaviour of the system being modelled with certainty. In a stochastic model, such as a model of a queueing system, we consider the behaviour of a system to be less than completely predictable even if we know the values of all model parameters. Thus probabilities are built in as part of the model. A second reason why we may be unable to predict behaviour with certainty is that we may be unsure of the values of parameters. Thirdly, we may be unsure about the validity of model assumptions.

In stochastic models the probabilities can often be thought of as conditional on the model and the values of the parameters. Given this condition the probabilities appear to be objective mathematical consequences. Subjectivity enters when we remove this condition and allow for our uncertainty about the model and the parameters. We can, however, interpret all probabilities in terms of degrees of belief and this provides a unified theory. The only distinction is that when we discuss a stochastic model we agree, for the purposes of the discussion, to consider the set of “beliefs” specified by the model. Strictly though, all of the probabilities do represent features of a set of beliefs.

We can distinguish three cases:

1. A purely deterministic model in which the variables are directly observable. Assuming validity of the model we can normally determine the values of the parameters by a finite number of observations.
2. A deterministic model in which observations on the variables contain “error”.
3. A stochastic model which has probabilities built into it.

4.2 Deterministic example

After material is extracted from an organism, the concentration of a certain compound decreases exponentially over time so that the concentration Z , in mmol/l, after t minutes is given by

$$Z = Z_0 e^{-ct} = e^{a-ct},$$

where $Z_0 = e^a$ is the initial concentration.

There are two parameters, a and c . We might describe our beliefs about a and c by a bivariate normal distribution where the mean and variance of a are μ_a and σ_a^2 , the mean and variance of c are μ_c and σ_c^2 and the covariance of a and c is γ . Then $a - ct$ is normal with mean

$$E(a - ct) = \mu_a - t\mu_c$$

and variance

$$\text{var}(a - ct) = \sigma_a^2 + t^2\sigma_c^2 - 2t\gamma.$$

Thus we can find probabilities for values of Z at time t .

Suppose our beliefs about a and c are such that a and c are independent so $\gamma = 0$. Suppose our median for Z_0 is 100 so $\mu_a = \ln(100) = 4.605$ and we are 95% sure that $50 < Z_0 < 200$ so $\sigma_a = \ln(2)/1.96 = 0.3536$ and $\sigma_a^2 = 0.1251$. Similarly suppose we choose $\mu_c = 0.01$, $\sigma_c^2 = 0.000025$ (so we are 95% sure that $0.0002 < c < 0.0198$). Then

$$\ln(Z) \sim N(4.605 - 0.01t, 0.1251 + 0.000025t^2).$$

For example, at $t = 100$, we have $\ln(Z) \sim N(3.605, 0.3751)$ and we are 95% sure that $2.405 < \ln(Z) < 4.805$, that is $11.07 < Z < 122.17$.

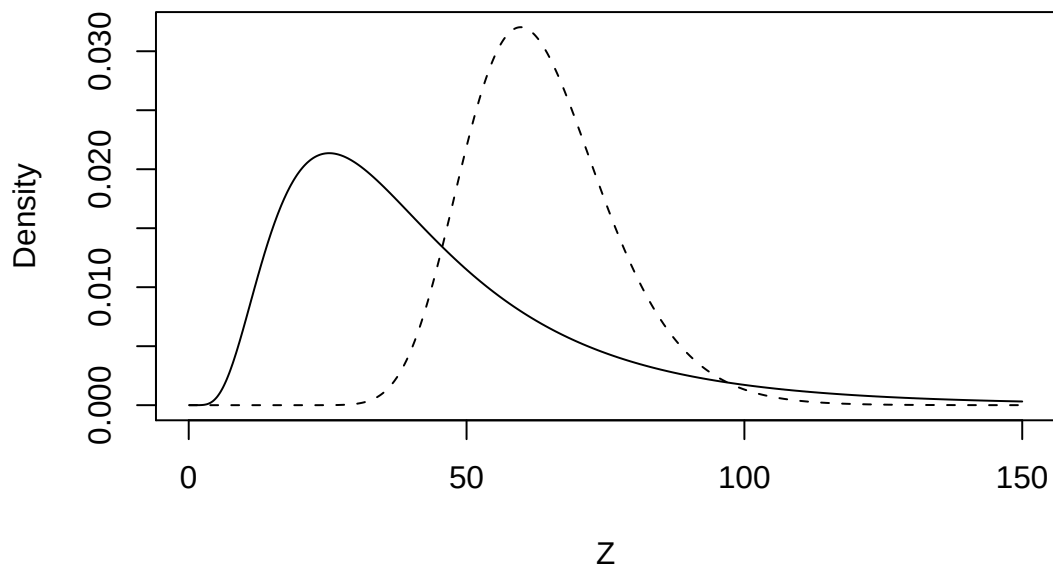


Figure 4: P.d.f. for Z at $t = 100$. Solid line: original. Dashed line: after observation at $t = 50$.

The density for Z at $t = 100$ is shown by the solid line in figure 4.

We may prefer to quote an interval such that the density is greater at every point inside than at any point outside. Such intervals are sometimes called “highest probability density” or “h.p.d.” regions (or intervals). This is more tricky in this example. An exact solution for a given probability requires numerical iteration. However we can easily find the probability for a given interval. The density is approximately equal at the points $Z = 10$ and $Z = 65$. Now

$$\begin{aligned}
 \Pr(10 < Z < 65) &= \Pr(2.3026 < \ln(Z) < 4.1744) \\
 &= \Pr\left(\frac{2.3026 - 3.605}{\sqrt{0.3751}} < z < \frac{4.1744 - 3.605}{\sqrt{0.3751}}\right) \\
 &= \Pr(-2.127 < z < 0.9297) \\
 &\simeq 0.807
 \end{aligned}$$

or 80.7% (where $z \sim N(0, 1)$).

(N.B. The earlier 95% interval was a h.p.d. interval for $\ln(Z)$).

We can use R to find a h.p.d. interval for Z . The distribution of Z is actually a lognormal distribution. In R the parameters to specify this are the mean and standard deviation of $\log(Z)$. In this case these are 3.605 and $\sqrt{0.3751} = 0.6125$ respectively. We can write a function which, if we guess a lower limit for an interval, will tell us the upper limit and the density values at the two limits. Here is such a function definition.

```

hpd<-function(lower,mean,var,prob)
{stddev<-sqrt(var)
 plower<-plnorm(lower,mean,stddev)
 pupper<-plower+prob
 upper<-qlnorm(pupper,mean,stddev)
 ldens<-dlnorm(lower,mean,stddev)
}

```

```

udens<-dlnorm(upper,mean,stddev)
result<-data.frame(lower,upper,ldens,udens)
result
}

```

Here is an example of its use.

```

> m<-3.605
> v<-0.3751
> p<-0.95
> hpd(9,m,v,p)
  lower    upper    ldens    udens
1     9 108.0604 0.005155869 0.001281801
> hpd(5,m,v,p)
  lower    upper    ldens    udens
1     5 101.0626 0.000644929 0.001651351
> hpd(7,m,v,p)
  lower    upper    ldens    udens
1     7 102.8224 0.002372777 0.001548699
> hpd(6,m,v,p)
  lower    upper    ldens    udens
1     6 101.6593 0.001356244 0.001615746
> hpd(6.5,m,v,p)
  lower    upper    ldens    udens
1    6.5 102.1545 0.001827710 0.00158683
> hpd(6.3,m,v,p)
  lower    upper    ldens    udens
1    6.3 101.9375 0.001630046 0.001599429
> hpd(6.2,m,v,p)
  lower    upper    ldens    udens
1    6.2 101.8388 0.001535725 0.001605202
> hpd(6.25,m,v,p)
  lower    upper    ldens    udens
1    6.25 101.8874 0.001582506 0.001602358

```

We see that $6.25 < Z < 101.89$ is approximately a 95% h.p.d. interval for Z . If we wished, we could write a more developed function which did not require us to use trial and error.

Clearly, if we observe Z at two time points we can determine the values of both parameters. If we only make one observation this reduces, but does not eliminate, our uncertainty. For example, suppose in our example that, at $t = 50$, we observe $Z = 90$. This means that $a - 50c = \ln(90) = 4.500$. It can be shown that, conditional on this information, a and c have a bivariate normal distribution in which the mean of a is $\tilde{\mu}_a = 4.8648$, the mean of c is $\tilde{\mu}_c = 0.007368$, the variance of a is $\tilde{\sigma}_a^2 = 4.168 \times 10^{-2}$, the variance of c is $\tilde{\sigma}_c^2 = 1.667 \times 10^{-5}$ and the covariance of a and c is $\tilde{\gamma} = 8.336 \times 10^{-4}$. The distribution for $\ln(Z)$ at $t = 100$ is now normal with mean 4.1316 and variance 0.04168. The density of the new distribution for Z at $t = 100$ is shown by the dashed line in figure 4.

4.3 Deterministic Example with Error

Suppose in our earlier deterministic example that we can not observe the actual value of Z at time t but instead observe

$$\tilde{Z}_t = \exp\{a + ct + \varepsilon_t\},$$

where $\varepsilon_t \sim N(0, \sigma_\varepsilon^2)$ and the value of σ_ε^2 is known. (This would often not usually be the case but we assume here that we know the properties of our measuring instrument sufficiently well). Thus our observation contains multiplicative errors. Assume further that the errors at different times are independent. Suppose we make observations $\tilde{Z}_1, \dots, \tilde{Z}_n$ on the population at times $t_1, \dots, t_n$

and let $Y_i = \ln \tilde{Z}_i$. If our prior probability distribution for a, c is bivariate normal as before then *a priori*, our joint distribution for $a, c, Y_1, \dots, Y_n$ is multivariate normal. It can be shown that our conditional distribution for a, c given $Y_1 = y_1, \dots, Y_n = y_n$ is bivariate normal with parameters given by the following formulae (proof omitted).

We specify a *multivariate* normal distribution, that is a normal distribution for several variables, by a mean vector $\underline{\mu}$ and a variance matrix V . When we have two variables we call it a *bivariate* normal distribution. Let

$$\underline{\beta} = \begin{pmatrix} a \\ c \end{pmatrix}.$$

In this way, *a priori*, $\underline{\beta} \sim N(\underline{\mu}_0, V_0)$, where

$$\underline{\mu}_0 = \begin{pmatrix} \mu_a \\ \mu_c \end{pmatrix}$$

and

$$V_0 = \begin{pmatrix} \sigma_a^2 & \gamma \\ \gamma & \sigma_c^2 \end{pmatrix}.$$

Let X be a $n \times 2$ matrix, the first column of which is $(1, 1, 1, \dots, 1)^T$ and the second column of which is $(-t_1, -t_2, \dots, -t_n)^T$. Let $\underline{y} = (y_1, y_2, \dots, y_n)^T$.

Let

$$\hat{\underline{\beta}} = (X^T X)^{-1} X^T \underline{y}.$$

Some will recognise this as the least squares estimate of $\underline{\beta}$. That is it is the “usual” estimates of a (intercept) and c (gradient) in a simple regression model.

Then the posterior distribution of $\underline{\beta}$ is bivariate normal with mean $\underline{\mu}_1$ and variance V_1 where

$$\begin{aligned} V_1 &= (V_0^{-1} + \sigma^{-2} X^T X)^{-1} \\ \underline{\mu}_1 &= V_1 (V_0^{-1} \underline{\mu}_0 + \sigma^{-2} X^T X \hat{\underline{\beta}}) = V_1 (V_0^{-1} \underline{\mu}_0 + \sigma^{-2} X^T \underline{y}) \end{aligned}$$

Notice that the posterior mean is a weighted average of the prior mean and the least squares estimate in this case.

Suppose that the data are as follows and $\sigma_\varepsilon^2 = 0.0025$.

Time t	25	50	75	100	125	150
Measured Concentration $\tilde{Z}$	113	81	74	52	43	36
Log Concentration Y	4.73	4.39	4.30	3.95	3.76	3.58

The posterior means of a and c are now 4.913 and 9.091×10^{-3} respectively and their posterior variances are 2.114×10^{-3} and 2.234×10^{-7} . The posterior covariance of a and c is 1.948×10^{-5} .

5 Bayes' Rule

5.1 Theory

In general we can use Bayes' rule to change our *prior probability distribution*, which expresses our beliefs about parameters before we see the data, to a *posterior probability distribution* representing beliefs about the parameters given the data.

Suppose we have a prior probability density function for a vector θ of parameters, $f_\theta(\theta)$. Suppose the p.d.f. for a vector $\underline{Y}$ of observations *given* $\underline{\theta}$ is $f_{Y|\theta}(\underline{y} | \underline{\theta})$. This latter p.d.f. is treated as a function of $\underline{\theta}$ once $\underline{y}$ is observed and is called the *likelihood*.

Then our *posterior p.d.f.* is

$$f_{\theta|y}(\underline{\theta} | \underline{y}) = \frac{f_\theta(\underline{\theta}) f_{Y|\theta}(\underline{y} | \underline{\theta})}{\int_{\theta} f_\theta(\underline{\theta}) f_{Y|\theta}(\underline{y} | \underline{\theta}) d\theta}.$$

We can think of this as

$$\Pr(\underline{\theta} | \underline{y}) = \frac{\Pr(\underline{\theta}) \Pr(\underline{y} | \underline{\theta})}{\Pr(\underline{y})}.$$

Often it is sufficient to write

$$f_{\theta|y}(\underline{\theta} | \underline{y}) \propto f_\theta(\underline{\theta}) f_{Y|\theta}(\underline{y} | \underline{\theta}).$$

That is **posterior** $\propto$ **prior** $\times$ **likelihood**.

When $\underline{Y}$ or $\underline{\theta}$ is discrete rather than continuous, the probability density functions are replaced by probabilities as appropriate.

5.2 Stochastic example: The rate of a Poisson process

We wish to model the occurrence of events in time as a Poisson process with rate λ . (This is a model for the times when events occur if the events occur “at random.” The rate is the average number per unit time).

Suppose our prior probability density function for λ is proportional to

$$\lambda^{\alpha-1} e^{-\beta\lambda}.$$

This means that our prior distribution for λ is a gamma distribution. The probability density function for a gamma(α, β) distribution is

$$f(x) = \begin{cases} 0 & (x < 0) \\ \frac{\beta^\alpha x^{\alpha-1} e^{-\beta x}}{\Gamma(\alpha)} & (x \geq 0) \end{cases}$$

where $\Gamma(y)$ is the *gamma function* which has the property that $\Gamma(y) = (y-1)\Gamma(y-1)$ and, if n is a positive integer, $\Gamma(n) = (n-1)!$.

We observe the process for τ time units. Events occur at times $t_1, t_2, \dots, t_n$. Writing $t_0 = 0$, the likelihood is

$$L = \prod_{i=1}^n \{\lambda e^{-\lambda(t_i - t_{i-1})}\} e^{-\lambda(\tau - t_n)}.$$

The last term is for the probability that no events occur in (t_n, τ) .

We can simplify L to

$$L = \lambda^n e^{-\lambda\tau}.$$

This is effectively the probability of n events in $(0, \tau)$, i.e.

$$(\lambda\tau)^n e^{-\lambda\tau} / n! \propto \lambda^n e^{-\lambda\tau}.$$

Hence the posterior p.d.f. is proportional to

12.40	3, 6, 9, 15, 24, 28, 30	12.50	4, 30, 34, 47
12.41	7, 12, 14, 16, 21, 24, 30, 50	12.51	0, 4, 18, 21, 52, 57, 59
12.42	9, 22, 28, 46, 53	12.52	4, 9, 29, 30, 31, 38, 59
12.43	22, 25, 35, 38, 58	12.53	2, 6, 31, 53
12.44	2, 5, 8, 10, 14, 17, 27, 30, 45	12.54	7, 9, 13, 24, 39, 47
12.45	3, 46	12.55	28, 46, 49, 59
12.46	13, 42, 51	12.56	6, 9, 14, 35, 41, 46
12.47	0, 9, 11, 18, 23, 26, 39, 51	12.57	8, 10, 15, 22, 25, 49, 52, 59
12.48	35, 39, 55	12.58	31, 34, 53, 55, 56
12.49	8, 10, 19, 20, 33, 45, 56, 58	12.59	2, 31, 34, 38, 54, 59

Table 2: Times of arrival of motor vehicles

3	3	3	6	9	4	2	37	5	2	2	5	3	6	20	19	13	6	18	7
29	3	10	3	20	4	3	3	2	4	3	10	3	15	18	43	27	29	9	9
9	2	7	5	3	13	12	44	4	16	13	2	9	1	13	12	11	2	6	26
4	13	13	4	14	3	31	5	2	5	5	20	1	1	7	21	3	4	25	22
14	2	4	11	15	8	41	18	3	10	7	3	5	21	6	5	22	2	5	7
3	24	3	7	32	3	19	2	1	6	29	3	4	16	5					

Table 3: Inter arrival times of motor vehicles

$$\lambda^{\alpha+n-1}e^{-(\beta+\tau)\lambda}.$$

This is clearly a gamma distribution so the posterior p.d.f. is

$$\frac{(\beta + \tau)^{\alpha+n} \lambda^{\alpha+n-1} e^{-(\beta+\tau)\lambda}}{\Gamma(\alpha + n)}.$$

The posterior mean is $(\alpha + n)/(\beta + \tau)$ and the posterior variance is $(\alpha + n)/(\beta + \tau)^2$.

5.3 Conjugate and non-conjugate priors

We are not, of course, restricted to using a prior of the form

$$\lambda^{\alpha-1}e^{-\beta\lambda}.$$

This particular form works out neatly because it is the *conjugate* form of prior for this likelihood. If our beliefs can not be adequately represented by a conjugate prior we may resort to numerical evaluation of the posterior distribution. The normalising constant, to make the posterior p.d.f. integrate to 1, may be found, if necessary, by numerical integration. In complicated models numerical approaches are commonly used. Monte Carlo integration has become particularly popular since about 1990.

Another approach, which is used in the software package “First Bayes,” is to form the prior density as a *mixture* of conjugate prior densities with different parameters. This keeps the calculation relatively straightforward while allowing considerable flexibility.

5.4 Chester Road traffic

At one time I worked in an office overlooking Chester Road in Sunderland. The times of arrivals of motor vehicles passing a point in Chester Road going East, from 12.40 till 13.00 on Wednesday 30th September 1987 are given in Table 2.

These can be converted to time intervals between vehicles. These intervals, in seconds, are given in Table 3. (The first value is the time till the first arrival).

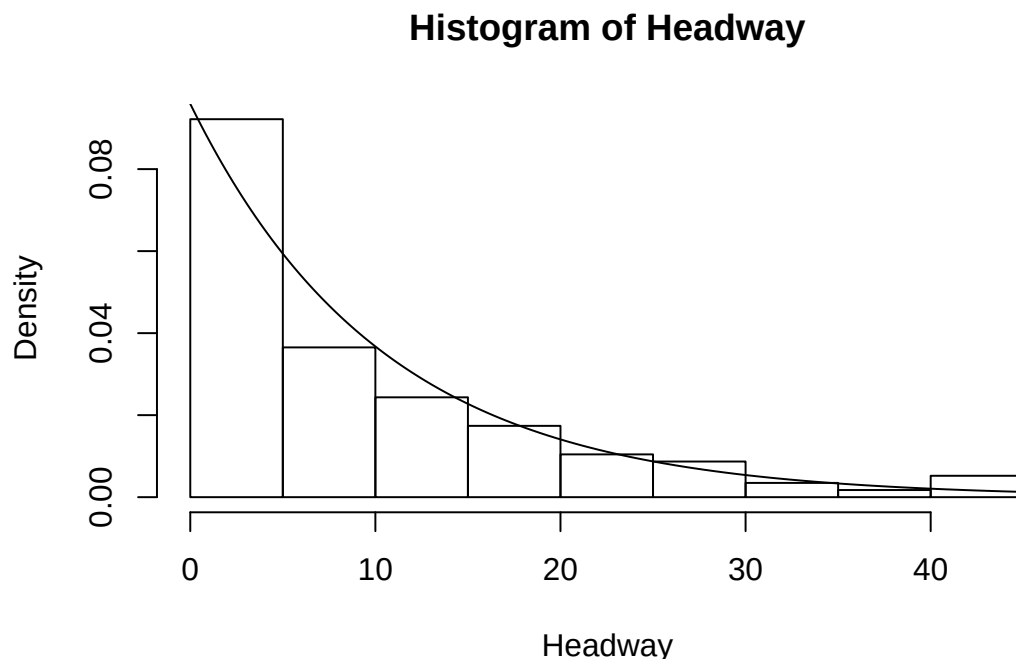


Figure 5: Inter arrival times (115 observations) with negative exponential p.d.f.

A histogram of these data is shown in Figure 5 together with a plot of the p.d.f. of a negative exponential distribution, the parameter of which was chosen so that the distribution had the same mean as the observed data. The fit appears satisfactory.

No evidence of nonzero correlations between successive inter-arrival times was found, either by plotting t_i against t_{i-1} or by estimating the correlation coefficients. Similarly a plot of t_i against i showed no obvious pattern.

Suppose the prior expectation for the rate of arrival was, say, 1 vehicle every 5 seconds, i.e. $\lambda_0 = 0.2$. Suppose we were almost certain that the rate would be less than 1 vehicle every 2 seconds. Let us say that $\Pr(\lambda < 0.5) = 0.99$. Suppose we use the conjugate (gamma) prior. Then $\alpha/\beta = 0.2$.

We require

$$\int_0^{0.5} f(\lambda).d\lambda = 0.99$$

where $f(\lambda)$ is the p.d.f. of a $\text{gamma}(\alpha, \alpha/0.2)$ distribution. We can evaluate this integral in R.

First we set up a vector of values for α and a corresponding vector for β . Then we evaluate the integral for each pair of values.

```
> alpha<-seq(1,10)
> beta<-alpha/0.2
> prob<-pgamma(0.5,alpha,beta)
> d<-data.frame(alpha,beta,prob)
> d
```

	alpha	beta	prob
1	1	5	0.9179150
2	2	10	0.9595723
3	3	15	0.9797433
4	4	20	0.9896639
5	5	25	0.9946545

6	6	30	0.9972076
7	7	35	0.9985300
8	8	40	0.9992214
9	9	45	0.9995856
10	10	50	0.9997785

We find that we get approximately what we require with $\alpha = 4$, $\beta = 20$.

Now $\tau = 1200$ and $n = 115$.

Thus the posterior p.d.f. is

$$\frac{1220^{119}}{118!} \lambda^{118} e^{-1220\lambda}.$$

The prior mean was 0.2. The posterior mean is

$$\frac{119}{1220} = 0.0975.$$

The prior variance was 0.01. The posterior variance is

$$\frac{119}{1220^2} = 0.00008.$$

We can evaluate the posterior pdf and distribution function using R. The posterior standard deviation is

$$\frac{\sqrt{119}}{1220} = 0.00894.$$

At a guess the important part of the posterior distribution is likely to be within ± 3 standard deviations of the posterior mean. That is roughly 0.07 to 0.13 so we create an array of λ values from 0.070 to 0.130 in steps of 0.001. We can check that this covers the important part of the distribution by looking at the distribution function. (In this case we find that very little of the probability is outside the range).

```
> lambda<-seq(0.07,0.13,0.001)
> pdf<-dgamma(lambda,119,1220)
> cdf<-pgamma(lambda,119,1220)
> plot(lambda,pdf,type="l")
> plot(lambda,cdf,type="l")
```

Figure 6 shows the prior density the (scaled) likelihood and the posterior density.

We can also use the distribution function to see, for example, that, in the posterior distribution,

$$\Pr(0.09 < \lambda < 0.11) = 0.91440 - 0.20176 = 0.71264.$$

What is the posterior probability now of observing j vehicles in the next t seconds?

The joint probability (density) of λ and j is

$$\frac{1220^{119}}{118!} \lambda^{118} e^{-1220\lambda} \frac{\lambda^j t^j e^{-\lambda t}}{j!} = \frac{1220^{119}}{118!} \frac{t^j}{j!} \lambda^{118+j} e^{-(1220+t)\lambda}.$$

We integrate out λ to get the marginal probability of j vehicles. By comparing the function with a gamma p.d.f. we see that

$$\begin{aligned} \int_0^\infty \lambda^{118+j} e^{-(1220+t)\lambda} .d\lambda &= \frac{\Gamma(119+j)}{(1220+t)^{119+j}} \\ &= \frac{(118+j)!}{(1220+t)^{119+j}} \end{aligned}$$

Hence the probability of observing j vehicles in the next t seconds is

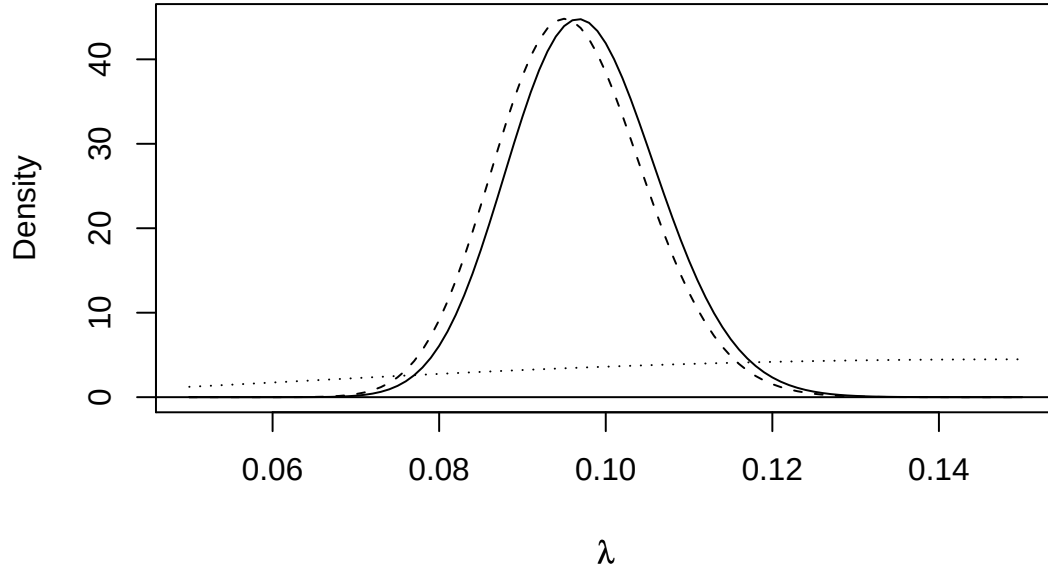


Figure 6: Chester road traffic arrival rate. Dots: prior pdf, Dashes: scaled likelihood, Solid line: posterior pdf.

$$\frac{(118+j)!}{118!j!} \frac{1220^{119}t^j}{(1220+t)^{119+j}} = \binom{118+j}{j} \left(\frac{1220}{1220+t} \right)^{119} \left(\frac{t}{1220+t} \right)^j.$$

This represents a negative binomial distribution. For large n and τ this would be approximately a Poisson distribution with mean equal to t times the posterior mean for λ .

What is the posterior p.d.f. now for the time T to the next arrival? (This is called a *predictive* distribution).

The joint p.d.f. of λ and T is

$$\frac{1220^{119}}{118!} \lambda^{118} e^{-1220\lambda} \lambda e^{-\lambda t} = \frac{1220^{119}}{118!} \lambda^{119} e^{-(1220+t)\lambda}.$$

We integrate out λ to get the marginal p.d.f. for T . By analogy with the gamma p.d.f.,

$$\int_0^\infty \lambda^{119} e^{-(1220+t)\lambda} d\lambda = \frac{119!}{(1220+t)^{120}}.$$

Hence the predictive p.d.f. for the waiting time is

$$\frac{119}{1220} \left(\frac{1220}{1220+t} \right)^{120}$$

which would be approximately a negative exponential distribution for large n .

6 More Than One Parameter

6.1 Theory

The principle is the same when we have more than one parameter. We simply obtain a joint posterior p.d.f. in the parameters. For example, if there are two parameters, we might produce a

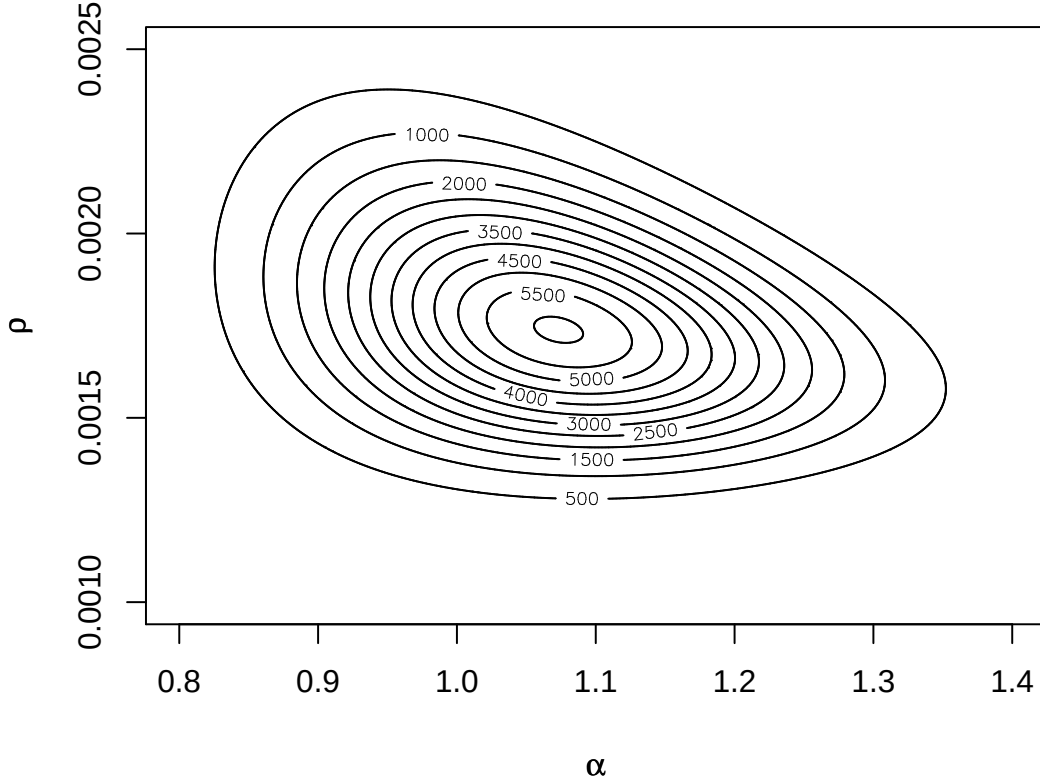


Figure 7: Posterior density of two unknowns: Contour plot

contour plot of the posterior p.d.f., as shown in figure 7, or a “3-d” plot, as shown in figure 8. If there are more than two parameters we need to “integrate out” some of the parameters in order to produce graphs like this.

In practice the integrations are often carried out numerically by computer. (Apart from being the only practical means in many cases, this removes the pressure to use a convenient conjugate prior). The basic rule is **posterior** $\propto$ **prior** $\times$ **likelihood**. If necessary, the normalising constant is found by integrating over all parameters. Posterior means, variances, marginal probability density functions, predictive distributions etc. can all be found by suitable integrations.

6.2 Example

Suppose we have observations on n independent and identically distributed lifetimes of cells. We wish to fit a Weibull distribution to these. The p.d.f. is

$$f(t) = \alpha \rho (\rho t)^{\alpha-1} \exp\{-(\rho t)^\alpha\}$$

for $0 \leq t < \infty$.

The likelihood is

$$L(\alpha, \rho) = \alpha^n \rho^{n\alpha} \left(\prod t_i \right)^{\alpha-1} \exp \left\{ -\rho^\alpha \sum t_i^\alpha \right\}.$$

Suppose that our prior for α, ρ is proportional to

$$\alpha^{k_1-1} e^{-k_2 \alpha} \rho^{k_3-1} e^{-k_4 \rho}.$$

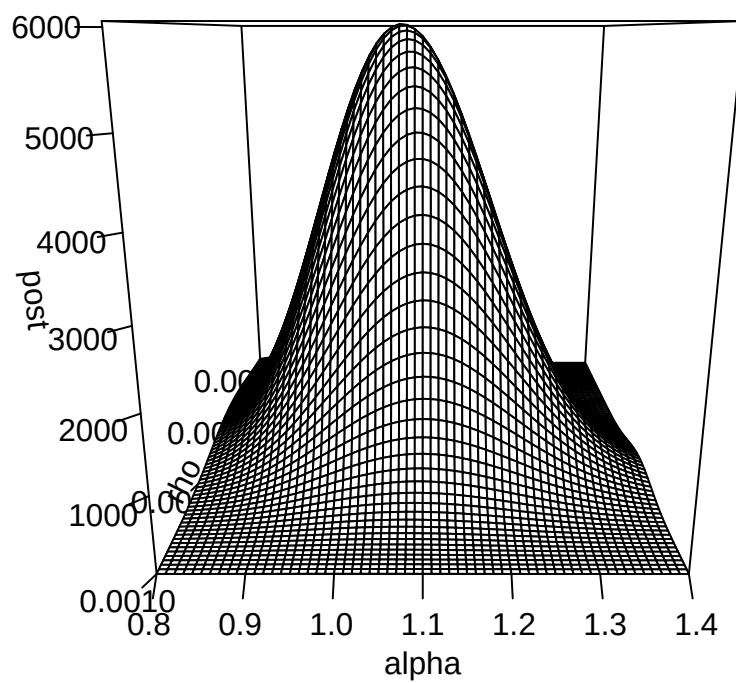


Figure 8: Posterior density of two unknowns: Wireframe plot

Note that, in this prior, ρ and α are independent. The mean is not independent of the shape, α , however. We may wish to make α and ρ non-independent in our prior.

The posterior pdf is proportional to

$$f_1(\alpha, \rho) = \alpha^{n+k_1-1} \rho^{n\alpha+k_3-1} \left(\prod t_i \right)^{\alpha-1} \exp \left\{ - \left[k_2 \alpha + k_4 \rho + \rho^\alpha \sum t_i^\alpha \right] \right\}.$$

To convert this into a p.d.f. we find

$$C_1 = \int_0^\infty \int_0^\infty f_1(\alpha, \rho). d\alpha. d\rho$$

numerically and divide $f_1(\alpha, \rho)$ by this.

Figures 7 and 8 show the posterior density of α and ρ when $n = 50$, $k_1 = 1$, $k_2 = 1$, $k_3 = 3$, $k_4 = 1000$ and the data are as follows.

67	313	1391	630	627	573	2093	28	492	482
206	1166	165	1088	496	313	437	815	436	17
32	131	340	939	247	1859	57	132	813	254
950	1615	463	258	2285	672	506	50	637	246
178	431	306	662	33	254	858	187	344	545

To find the posterior mean of ρ we evaluate

$$C_1^{-1} \int_0^\infty \int_0^\infty \rho f_1(\alpha, \rho). d\alpha. d\rho.$$

To find a 95 % h.p.d. region for α, ρ we can either choose a value k and evaluate $C_1^{-1} \int \int f_1(\alpha, \rho). d\alpha. d\rho$ over all points in a grid for which $f_1(\alpha, \rho) > k$ then adjust k and repeat until the value of 0.95 is obtained or rank all of the points in our grid in decreasing order of $f_1(\alpha, \rho)$ and cumulatively integrate over them until 0.95 is reached.

To find the marginal p.d.f. for α we evaluate

$$C_1^{-1} \int_0^\infty f_1(\alpha, \rho). d\rho.$$

It may help to use a transformation. For example we can transform the $0, \infty$ ranges of α, ρ to $0, 1$ by using

$$\beta = \frac{\alpha}{\alpha + 1}, \quad \gamma = \frac{\rho}{\rho + 1}.$$

The Jacobian is $(\alpha + 1)^{-2}(\rho + 1)^{-2}$ so we define

$$f_2(\beta, \gamma) = (\alpha + 1)^2(\rho + 1)^2 f_1(\alpha, \rho),$$

where

$$\alpha = \frac{\beta}{1 - \beta}, \quad \rho = \frac{\gamma}{1 - \gamma},$$

and

$$C_2 = \int_0^1 \int_0^1 f_2(\beta, \gamma). d\beta. d\gamma.$$

The posterior mean of ρ is then

$$C_2^{-1} \int_0^1 \int_0^1 \frac{\gamma}{1 - \gamma} f_2(\beta, \gamma). d\beta. d\gamma.$$

A h.p.d. region is found by integrating $C_2^{-1} f_2(\beta, \gamma)$ with respect to β, γ over the points with the greatest values of

$$f_1 \left(\frac{\beta}{1 - \beta}, \frac{\gamma}{1 - \gamma} \right) = f_1(\alpha, \rho).$$

6.3 Problems 2

Answer all questions. Solutions are to be submitted in the Homework Box no later than 5.00pm on Wednesday February 28th. You may use, for example, R to do numerical calculations in questions 1 and 2. You should show the commands which you have used.

1. In a small survey, a random sample of 50 people from a large population is selected. Each person is asked a question to which the answer is either “Yes” or “No.” Let the proportion in the population who would answer “Yes” be θ . Our prior distribution for θ is a beta(1.5, 1.5) distribution. In the survey, 37 people answer “Yes.”
 - (a) Find the prior mean and prior standard deviation of θ .
 - (b) Find the prior probability that $\theta < 0.6$.
 - (c) Find the likelihood.
 - (d) Find the posterior distribution of θ .
 - (e) Find the posterior mean and posterior standard deviation of θ .
 - (f) Plot a graph showing the prior and posterior probability density functions of θ on the same axes.
 - (g) Find the posterior probability that $\theta < 0.6$.

Notes:

The probability density function of a beta(a, b) distribution is $f(x) = kx^{a-1}(1-x)^{b-1}$ where k is a constant.

If $X \sim \text{beta}(a, b)$ then the mean of X is

$$E(X) = \frac{a}{a+b}$$

and the variance of X is

$$\text{var}(X) = \frac{ab}{(a+b+1)(a+b)^2}.$$

If $X \sim \text{beta}(a, b)$ then you can use a command such as the following in R to find $\Pr(X < c)$.

`pbeta(c,a,b)`

To plot the prior and posterior probability densities you may use R commands such as the following.

```
theta<-seq(0.01,0.99,0.01)
prior<-dbeta(theta,a,b)
posterior<-dbeta(theta,c,d)
plot(theta,posterior,xlab=expression(theta),ylab="Density",type="l")
lines(theta,prior,lty=2)
```

2. The populations, n_i , and the number of cases, x_i , of a disease in a year in each of six districts are given in the table below.

Population n	Cases x
120342	2
235967	5
243745	3
197452	5
276935	3
157222	1

We suppose that the number X_i in a district with population n_i is a Poisson random variable with mean $n_i\lambda/100000$. The number in each district is independent of the numbers in other districts, given the value of λ . Our prior distribution for λ is a gamma distribution with mean 3.0 and standard deviation 2.0.

- Find the parameters of the prior distribution.
- Find the prior probability that $\lambda < 2.0$.
- Find the likelihood.
- Find the posterior distribution of λ .
- Find the posterior mean and posterior standard deviation of λ .
- Plot a graph showing the prior and posterior probability density functions of λ on the same axes.
- Find the posterior probability that $\lambda < 2.0$.

Notes:

The probability density function of a $\text{gamma}(a, b)$ distribution is $f(x) = kx^{a-1}\exp(-bx)$ where k is a constant.

If $X \sim \text{gamma}(a, b)$ then the mean of X is $E(X) = a/b$ and the variance of X is $\text{var}(X) = a/(b^2)$.

If $X \sim \text{gamma}(a, b)$ then you can use a command such as the following in R to find $\Pr(X < c)$.

```
pgamma(c, a, b)
```

To plot the prior and posterior probability densities you may use R commands such as the following.

```
lambda<-seq(0.00,5.00,0.01)
prior<-dgamma(lambda,a,b)
posterior<-dgamma(lambda,c,d)
plot(lambda,posterior,xlab=expression(lambda),ylab="Density",type="l")
lines(lambda,prior,lty=2)
```

- Geologists note the type of rock at fixed vertical intervals of six inches up a quarry face. At this quarry there are four types of rock. The following model is adopted.

The conditional probability that the next rock type is j given that the present type is i and given whatever has gone before is p_{ij} . Clearly $\sum_{j=1}^4 p_{ij} = 1$ for all i .

The following table gives the observed (upwards) transition frequencies.

		To rock			
		1	2	3	4
From rock	1	56	13	24	4
	2	15	93	22	35
	3	20	25	153	11
	4	6	35	11	44

Our prior distribution for the transition probabilities is as follows. For each i we have a uniform distribution over the space of possible values of $p_{i1}, \dots, p_{i4}$. The prior distribution of $p_{i1}, \dots, p_{i4}$ is independent of that for $p_{k1}, \dots, p_{k4}$ for $i \neq k$.

Find the matrix of posterior expectations of the transition probabilities.

Note that the integral of $x_1^{n_1} x_2^{n_2} x_3^{n_3} x_4^{n_4}$ over the region such that $x_j > 0$ for $j = 1, \dots, 4$ and $\sum_{j=1}^4 x_j = 1$, where $n_1, \dots, n_4$ are positive is

$$\begin{aligned} & \int_0^1 x_1^{n_1} \int_0^{1-x_1} x_2^{n_2} \int_0^{1-x_1-x_2} x_3^{n_3} (1-x_1-x_2-x_3)^{n_4} .dx_3 .dx_2 .dx_1 \\ &= \frac{\Gamma(n_1+1)\Gamma(n_2+1)\Gamma(n_3+1)\Gamma(n_4+1)}{\Gamma(n_1+n_2+n_3+n_4+4)} \end{aligned}$$

7 Simple Numerical Methods

7.1 The need for integration

Consider a general Bayesian inference problem. We have a (scalar or vector) unknown θ . Usually θ will be a “parameter”, or a vector of parameters. We have observations Y . Let us assume that both θ and Y are continuous so that we can talk in terms of densities and integrals. If one is discrete then we replace the probability density with a probability. If θ is discrete then we replace the integrals with summations.

We have a prior probability density function $f_{\theta}^{(0)}(\theta)$ for θ . Let the range of possible values of θ be Θ .

We have a sampling distribution pdf $f_{Y|\theta}(y | \theta)$ for the observations Y . (We are treating Y as a vector here so this is the joint pdf for all of the elements of Y).

The joint density of θ and Y is then $f_{\theta}^{(0)}(\theta)f_{Y|\theta}(y | \theta)$.

The posterior density of θ is

$$f_{\theta}^{(1)}(\theta) = \frac{f_{\theta}^{(0)}(\theta)f_{Y|\theta}(y | \theta)}{\int_{\Theta} f_{\theta}^{(0)}(\theta)f_{Y|\theta}(y | \theta) d\theta}.$$

Let $g(\theta)$ be some scalar function of θ . Then the posterior mean of $g(\theta)$ is

$$E^{(1)}[g(\theta)] = \frac{\int_{\Theta} g(\theta)f_{\theta}^{(0)}(\theta)f_{Y|\theta}(y | \theta) d\theta}{\int_{\Theta} f_{\theta}^{(0)}(\theta)f_{Y|\theta}(y | \theta) d\theta}.$$

To find the posterior variance of $g(\theta)$ we can also find $E^{(1)}[\{g(\theta)\}^2]$ and then $\text{var}^{(1)}[g(\theta)] = E^{(1)}[\{g(\theta)\}^2] - \{E^{(1)}[g(\theta)]\}^2$.

Each integral which we need here is of the form

$$\int_{\Theta} g(\theta)f_{\theta}^{(0)}(\theta)f_{Y|\theta}(y | \theta) d\theta$$

(where we might have $g(\theta) = 1$).

Sometimes, when we have conjugate priors, these integrals work out analytically. Often they do not so we use numerical methods.

Note that θ is often a vector so we have multiple integrations.

7.2 Trapezium rule

Various more refined methods are available but a simple trapezium rule is often sufficient.

Consider a 1-dimensional case (i.e. θ a scalar).

Roughly speaking, the procedure is as follows. (We will discuss some details later).

1. Set up an array of values $\theta_1, \theta_2, \dots, \theta_m$ of values for θ . These values are separated by a *step size* of $\delta\theta$. We assume (for now) that the steps are all equal. That is $\theta_{j+1} = \theta_j + \delta\theta$.
2. Evaluate $h(\theta) = g(\theta)f_{\theta}^{(0)}(\theta)f_{Y|\theta}(y | \theta)$ at each value θ_j .
3. Approximate the area under the curve by the total area of a set of rectangular columns. Column j has height $h(\theta_j)$. Each column has width $\delta\theta$.

$$\int_{\Theta} h(\theta) d\theta \approx \sum_{j=1}^m h(\theta_j)\delta\theta$$

Note:

1. Usually the density becomes close to zero in the tails of the distribution so edge effects are not usually a problem. there can be exceptions though.
2. We often need some trial and error to find an appropriate range for the grid. Plotting $h(\theta)$ against θ can help.
3. Care needs to be taken with numerical issues. Values of functions can often be very small. It is usually best to evaluate logs first. See the examples below.

7.3 Simple one-dimensional example

Consider the Chester Road data describes in section 5.4. To illustrate the principles involved when we use a non-conjugate prior, we will try using a uniform(0.0, 0.5) prior. The calculations are quite simple in this case but we will do them using a general method. The likelihood is proportional to

$$\lambda^{115}e^{-1200\lambda}.$$

It is usually better to work in terms of logs, for numerical reasons. The log likelihood, apart from an additive constant, is

$$115\log(\lambda) - 1200\lambda.$$

We will also take logs of the prior density. Note that we could not do this where the prior density is zero but we really do not need to.

Having added the log prior to the log likelihood we subtract the maximum of this so that the maximum becomes zero. We then take exponentials and the maximum of this will be 1.0. The reason for doing this is to avoid very large or very small numbers. We then normalise by finding the integral and dividing by this. The integral is found numerically using a simple trapezium rule.

Suitable R commands are as follows.

```
lambda<-seq(0.05,0.15,0.001)
prior<-rep(2.0,100)
logprior<-log(prior)
```

```

loglik<-115*log(lambda)-1200*lambda
logpos<-logprior+loglik
logpos<-logpos-max(logpos)
posterior<-exp(logpos)
density<-posterior/(sum(posterior)*0.001)
plot(lambda,density,type="l")
lines(lambda,prior)

```

We can, of course, calculate the statistics we need for the likelihood from the raw data. In this case it is actually just the number of vehicles which passed. However we can also draw a histogram of the data, for example. You can obtain a copy of the file containing the raw data from the module Web page. The commands below read the data, count the number of vehicles and draw a histogram of the time intervals (or “headways”).

```

headway<-read.table("chester.dat")
headway<-headway[,1]
n<-length(headway)
n
hist(headway)

```

Two (or more) dimensions

If we have two unknown parameters θ_1 , θ_2 then we (often) need a two-dimensional grid of values, containing every combination of $\theta_{1,1}, \dots, \theta_{1,m_1}$ and $\theta_{2,1}, \dots, \theta_{2,m_2}$. We therefore have $m_1 m_2$ points and two step sizes, $\delta\theta_1$, $\delta\theta_2$. The approximate integral becomes

$$\sum_{j=1}^{m_1} \sum_{k=1}^{m_2} h(\theta_{1,j}, \theta_{2,k}) \delta\theta_1 \delta\theta_2.$$

Note:

1. We can extend this to three or more dimensions but it becomes impractical when the number of dimensions is large. In such cases we would usually use Markov chain Monte Carlo methods which are beyond the scope of this module.

2. It is sometimes possible to integrate analytically with respect to one parameter. This reduces the dimensionality of the numerical integral.

7.4 Two-dimensional example

We illustrate a two-parameter case analysed numerically so that we are not restricted to using the conjugate prior distribution.

Suppose our data are the following 30 observations from a normal distribution. We enter these into a vector `y`.

```
21.3 19.9 22.6 22.7 18.8 19.5
18.1 23.0 18.7 26.3 23.2 23.4
26.1 24.7 20.0 25.4 18.6 25.2
24.2 23.3 22.9 25.0 22.9 23.8
20.2 21.4 22.9 19.5 24.7 26.4
```

Suppose the prior distribution for population precision $p = 1/\sigma^2$, where σ^2 is the population variance, is negative exponential with mean 0.1.

Suppose the prior distribution for the population mean μ is normal with mean 20 and variance 16, independently of p .

For illustration we use a purely numerical procedure as follows.

Set up a grid of μ and p values, with μ between 21 and 24 in steps of 0.05 and p between 0 and 0.3 in steps of 0.005, as follows. (This gives a 61×61 grid).

```
muvals<-seq(21,24,0.05)
> muvals
[1] 21.00 21.05 21.10 21.15 21.20 21.25 21.30 21.35 21.40 21.45 21.50 21.55
[13] 21.60 21.65 21.70 21.75 21.80 21.85 21.90 21.95 22.00 22.05 22.10 22.15
[25] 22.20 22.25 22.30 22.35 22.40 22.45 22.50 22.55 22.60 22.65 22.70 22.75
[37] 22.80 22.85 22.90 22.95 23.00 23.05 23.10 23.15 23.20 23.25 23.30 23.35
[49] 23.40 23.45 23.50 23.55 23.60 23.65 23.70 23.75 23.80 23.85 23.90 23.95
[61] 24.00
> mu<-rep(muvals,61)
> x<-rep(61,61)
> pvals<-seq(0.0,0.3,0.005)
> pvals
[1] 0.000 0.005 0.010 0.015 0.020 0.025 0.030 0.035 0.040 0.045 0.050 0.055
[13] 0.060 0.065 0.070 0.075 0.080 0.085 0.090 0.095 0.100 0.105 0.110 0.115
[25] 0.120 0.125 0.130 0.135 0.140 0.145 0.150 0.155 0.160 0.165 0.170 0.175
[37] 0.180 0.185 0.190 0.195 0.200 0.205 0.210 0.215 0.220 0.225 0.230 0.235
[49] 0.240 0.245 0.250 0.255 0.260 0.265 0.270 0.275 0.280 0.285 0.290 0.295
[61] 0.300
> p<-rep(pvals,x)
```

Now calculate the natural logarithm of the joint prior density as follows.

```
> muprior<-dnorm(mu,20,4)
> pprior<-dexp(p,10)
> logprior<-log(muprior)+log(pprior)
```

Now the likelihood function is

$$\begin{aligned} L &= \prod_{i=1}^n (2\pi\sigma^2)^{-1/2} \exp \left\{ -\frac{1}{2\sigma^2} (y_i - \mu)^2 \right\} \\ &= (2\pi\sigma^2)^{-n/2} \exp \left\{ -\frac{1}{2\sigma^2} \sum_{i=1}^n (y_i - \mu)^2 \right\} \end{aligned}$$

and the log likelihood is

$$l = -(n/2) \ln(2\pi) + (n/2) \ln(p) - (1/2)p(\sum y_i^2 - 2\mu \sum y_i + n\mu^2).$$

So we calculate (up to a constant) the log posterior as follows.

```
> n<-length(y)
> sy<-sum(y)
> syy<-sum(y*y)
> loglik<-(n/2)*log(p) - (p/2)*(syy - 2*mu*sy + n*mu*mu)
> logpos<-logprior+loglik
```

Before exponentiating (i.e. taking antilogs) we add a constant to rescale so that the calculations do not overflow.

```
> logpos<-logpos-max(logpos)
> posterior<-exp(logpos)
```

We now integrate numerically to find the missing constant of proportionality and rescale the posterior density. This can now be displayed, for example, in a contour plot.

```
> k<-sum(posterior)*0.05*0.005
> posterior<-posterior/k
> dim(posterior)<-c(61,61)
> contour(muvals,pvals,posterior)
> persp(muvals,pvals,posterior)
> image(muvals,pvals,posterior)
```

To calculate the marginal posterior density of μ or p it is convenient to rearrange the array so that we can integrate along the rows or columns. In fact we already did this to make the 3-d plots.

```
> mupost<-rowSums(posterior)*0.005
> plot(muvals,mupost,type="l")
> ppost<-colSums(posterior)*0.05
> plot(pvals,ppost,type="l")
```

We can use the marginal density to calculate the posterior mean of μ .

```
> mumupost<-muvals*mupost
> pmmu<-sum(mumupost)*0.05
> pmmu
[1] 22.45547
```

Note: In this example we could, in fact, integrate with respect to μ analytically and so have only a one-dimensional numerical integration to do. We will look at this in a later lecture.

8 Conjugate Priors I: The beta distribution

8.1 Introduction

We have already come across the idea of a conjugate prior distribution in Lecture 5. Now we will look at this idea in more detail.

Suppose we wish to learn about a parameter θ . Suppose our prior distribution for θ has pdf $f_0(\theta)$. Suppose we observe data y and the likelihood function is $L(\theta; y)$. Now our posterior pdf for θ is

$$f_1(\theta) \propto f_0(\theta)L(\theta; y).$$

A *conjugate* prior distribution is one where the form of the posterior distribution is the same as the form of the prior distribution. That is $f_0(\theta)$ and $f_1(\theta)$ are pdfs belonging to the same family of (e.g. they are both normal or they are both gamma). Clearly the form of the conjugate prior depends on the form of the likelihood. So there is a different conjugate distribution for a different data (or “sampling”) distribution. In some cases there is no conjugate prior distribution.

We are not, of course, compelled to use a conjugate prior distribution. Our prior beliefs simply might not be of the appropriate form. However, very often, a conjugate prior will be able to represent our beliefs closely enough and the calculations then become much easier. When we feel that we can not use a conjugate prior, even though one exists, we may be able to use a *mixture* of conjugate priors to represent our beliefs. This topic will be discussed in a later lecture.

8.2 Beta distribution

In this lecture we will look at the *beta* distribution as a conjugate prior distribution.

If $\theta \sim \text{beta}(a, b)$ then it has pdf given by

$$f(\theta) = \begin{cases} 0 & (\theta \leq 0) \\ \frac{\Gamma(a+b)}{\Gamma(a)\Gamma(b)} \theta^{a-1} (1-\theta)^{b-1} & (0 < \theta < 1) \\ 0 & (\theta \geq 1) \end{cases}$$

where $a > 0$, $b > 0$ and $\Gamma(u)$ is the *gamma function*

$$\Gamma(u) = \int_0^\infty w^{u-1} e^{-w} dw = (u-1)\Gamma(u-1).$$

If $\theta \sim \text{beta}(a, b)$ then the mean of θ is

$$\begin{aligned} E(\theta) &= \int_0^1 \theta \frac{\Gamma(a+b)}{\Gamma(a)\Gamma(b)} \theta^{a-1} (1-\theta)^{b-1} d\theta \\ &= \frac{\Gamma(a+b)}{\Gamma(a)} \frac{\Gamma(a+1)}{\Gamma(a+b+1)} \int_0^1 \frac{\Gamma(a+b+1)}{\Gamma(a+1)\Gamma(b)} \theta^{a+1-1} (1-\theta)^{b-1} d\theta \\ &= \frac{\Gamma(a+b)}{\Gamma(a)} \frac{\Gamma(a+1)}{\Gamma(a+b+1)} \\ &= \frac{a}{a+b} \end{aligned}$$

Similarly

$$E(\theta^2) = \frac{\Gamma(a+b)}{\Gamma(a)} \frac{\Gamma(a+2)}{\Gamma(a+b+2)} = \frac{(a+1)a}{(a+b+1)(a+b)}.$$

So

$$\begin{aligned} \text{var}(\theta) &= E(\theta^2) - [E(\theta)]^2 \\ &= \frac{(a+1)a}{(a+b)(a+b+1)} - \frac{a^2}{(a+b)^2} \\ &= \frac{a(a+1)(a+b) - a^2(a+b+1)}{(a+b+1)(a+b)^2} \end{aligned}$$

$$\begin{aligned}
&= \frac{(a^2 + a)(a + b) - a^2(a + b) - a^2}{(a + b + 1)(a + b)^2} \\
&= \frac{ab}{(a + b + 1)(a + b)^2}
\end{aligned}$$

8.3 Binomial observations

8.3.1 Model

Suppose that we will observe $X_1, \dots, X_n$ where these are mutually independent given the value of θ and

$$X_i \sim \text{binomial}(N_i, \theta)$$

where $N_1, \dots, N_n$ are known or will be observed and $\sum_{i=1}^n N_i = N$.

Then the likelihood is

$$\begin{aligned}
L(\theta; x) &= \prod_{i=1}^n \binom{N_i}{x_i} \theta^{x_i} (1 - \theta)^{N_i - x_i} \\
&\propto \theta^S (1 - \theta)^{N - S}
\end{aligned}$$

where $S = \sum_{i=1}^n x_i$.

The conjugate prior is a *beta* distribution which has a pdf proportional to

$$\theta^{a-1} (1 - \theta)^{b-1}$$

for $0 < \theta < 1$.

The posterior pdf is proportional to

$$\theta^{a-1} (1 - \theta)^{b-1} \times \theta^S (1 - \theta)^{N-S} = \theta^{a+S-1} (1 - \theta)^{b+N-S-1}.$$

This is proportional to the pdf of a $\text{beta}(a + \sum x_i, b + N - \sum x_i)$ distribution.

8.3.2 Example

Consider the example in section 3.3 where each animal may have a particular gene or not. The prior pdf is proportional to

$$\theta^{a-1} (1 - \theta)^{b-1}$$

where we specify the values of a and b .

The posterior pdf is then proportional to

$$\theta^{a+3-1} (1 - \theta)^{b+17-1}.$$

This is also the pdf of a beta distribution.

The posterior distribution does depend on the choice of prior distribution, in this case represented by the parameters a and b . To illustrate this let us consider four quite different choices of prior distribution, as follows.

Case 1 With $a = b = 1$ we obtain a uniform prior distribution for θ . In this case the posterior pdf is simply proportional to the likelihood.

Case 2 With $a = 2$ and $b = 1$ we obtain a triangular prior distribution expressing a belief that larger values of θ are more likely.

Case 3 With $a = 1$ and $b = 2$ we obtain a triangular prior distribution expressing a belief that smaller values of θ are more likely.

Case 4 With $a = b = 2$ we obtain a quadratic prior distribution expressing the belief that values of θ close to 0.5 are more likely.

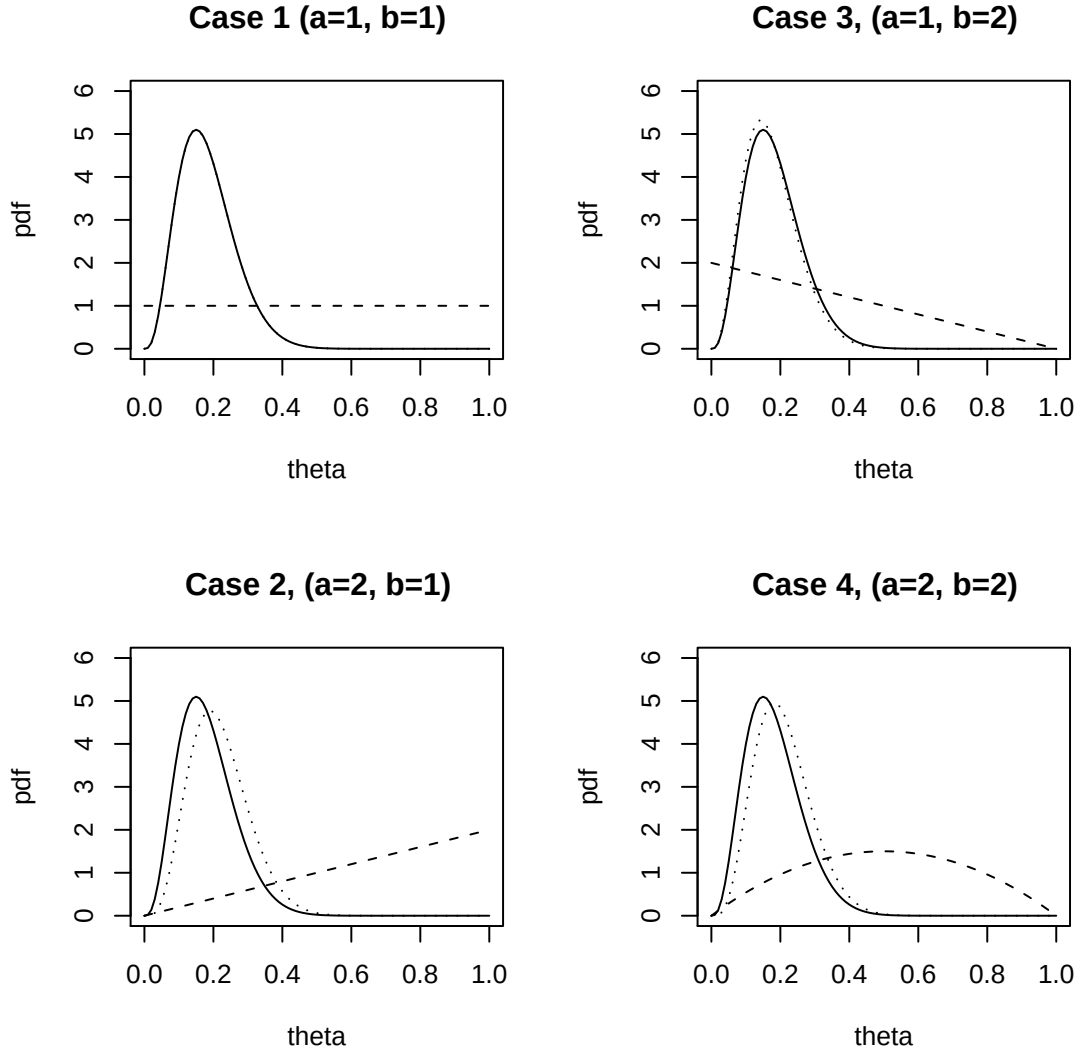


Figure 9: Likelihood (solid line), prior pdf (dashes) and posterior pdf (dots) in Cases 1, $\dots$, 4.

Figure 9 shows, for each of these cases, the (scaled) likelihood and the prior and posterior pdf. In this way we can examine the sensitivity of the posterior distribution to the prior specification. Remember that the likelihood depends only on the data, not the prior, and is the same in all four cases. In this example, with a fairly small sample size and strongly differing prior distributions, we can see the effect of changes in the prior, although the effect is perhaps smaller than some might expect. In the more general case where we observe x animals with the gene out of n animals, the parameters of the posterior distribution are $a + x$ and $b + n - x$. Clearly, as n becomes larger, the effect of the prior parameters a and b becomes less important. If we do find that the posterior distribution is very sensitive to the choice of prior, over a reasonable range of choices, then this simply tells us that the data are not very informative and so change our beliefs relatively little.

The inferential process really consists of calculating the posterior probability distribution. Unlike non-Bayesian statistics we do not need to have “estimators”, “confidence intervals” etc. However we might want to calculate particular posterior quantities for particular purposes, such as the posterior probability that the mean concentration of a compound in the bloodstream is increased after a particular treatment, or we may wish to summarise a posterior distribution simply for convenience.

For example, we might calculate the mean and variance of the posterior distribution. In this

Case	a	b	Prior			Posterior		
			Mean	Std.dev.	95% interval	Mean	Std.dev.	95% interval
1	1	1	0.500	0.289	$0.025 < \theta < 0.975$	0.182	0.080	$0.054 < \theta < 0.363$
2	2	1	0.667	0.236	$0.158 < \theta < 0.987$	0.217	0.084	$0.078 < \theta < 0.403$
3	1	2	0.333	0.236	$0.013 < \theta < 0.842$	0.174	0.077	$0.052 < \theta < 0.349$
4	2	2	0.500	0.224	$0.094 < \theta < 0.906$	0.208	0.081	$0.075 < \theta < 0.388$

Table 4: Numerical summaries for the four cases.

example, the *prior mean* is $a/(a+b)$ and the *prior variance* is $ab/[(a+b+1)(a+b)^2]$. The posterior mean and variance are $(a+3)/(a+b+20)$ and $(a+3)(b+17)/[(a+b+21)(a+b+20)^2]$. In the general case of observing x out of n , the posterior mean and variance are $(a+x)/(a+b+n)$ and $(a+x)(a+n-x)/[(a+b+n+1)(a+b+n)^2]$. Once again we see that the dependence on the prior will decrease as the sample size increases.

We can also calculate the probability that an unknown lies within a certain interval or range. We might have a particular interval in mind because it is of interest to us or we might simply want to do this as another way to summarise a posterior distribution. For example we might find an interval such that there is a posterior probability of 0.95 that an unknown θ lies in the interval. Such an interval is sometimes called a *95% credible interval*. Of course there are many such intervals which we could validly calculate. Perhaps the easiest, but not necessarily the most appropriate, is a symmetric interval such that the probability that θ is less than the lower limit is equal to the probability that it is greater than the upper limit.

Table 4 gives prior and posterior means, standard deviations and symmetric 95% credible intervals for θ in the four cases above.

8.4 Geometric observations

8.4.1 Model

Suppose that we will observe $X_1, \dots, X_n$ where these are mutually independent given the value of θ and

$$X_i \sim \text{geometric}(\theta)$$

where

$$\Pr(X_i = j \mid \theta) = (1 - \theta)^{j-1} \theta$$

for $j = 1, 2, \dots$.

Then the likelihood is

$$\begin{aligned} L(\theta; x) &= \prod_{i=1}^n \theta(1 - \theta)^{x_i-1} \\ &\propto \theta^n (1 - \theta)^{S-n} \end{aligned}$$

where $S = \sum_{i=1}^n x_i$.

The conjugate prior is a *beta* distribution which has a pdf proportional to

$$\theta^{a-1} (1 - \theta)^{b-1}$$

for $0 < \theta < 1$.

The posterior pdf is proportional to

$$\theta^{a-1} (1 - \theta)^{b-1} \times \theta^n (1 - \theta)^{S-n} = \theta^{a+n-1} (1 - \theta)^{b+S-n-1}.$$

This is proportional to the pdf of a $\text{beta}(a+n, b+\sum x_i - n)$ distribution.

8.4.2 Example

Consider the problem posed in Question 3 of Problems 1 (2.7) except that we will give the probability of a defective item a beta prior.

A machine is built to make mass-produced items. Each item made by the machine has a probability p of being defective. Given the value of p , the items are independent of each other. The prior distribution for p is a $\text{beta}(a, b)$ distribution. The machine is tested by counting the number of items made before a defective is produced. Find the posterior distribution of p given that the first defective item is the thirteenth to be made.

Here $n = 1$ and $x = 13$ so the posterior distribution is $\text{beta}(a + 1, b + 12)$. For example, suppose that $a = b = 1$ so that the prior distribution is uniform. Then the posterior distribution is $\text{beta}(2, 13)$. The prior mean is 0.5. The posterior mean is $2/(2 + 13) = 0.1333$. The prior variance is $1/12 = 0.0833333$ giving a standard deviation of 0.2887. The posterior variance is 0.007222 giving a standard deviation of 0.08498.

Note that, in the original question, we did not consider values of p greater than 0.05 to be possible. We could introduce such a constraint and still give p a continuous prior. We will return to this question later.

8.5 Negative binomial observations

8.5.1 Model

Suppose that we will observe $X_1, \dots, X_n$ where these are mutually independent given the value of θ and

$$X_i \sim \text{negative - binomial}(r_i, \theta)$$

where $r_1, \dots, r_n$ are known or will be observed and

$$\Pr(X_i = j \mid \theta) = \binom{j-1}{r_i-1} (1-\theta)^{j-r_i} \theta^{r_i}$$

for $j = r_i, r_i + 1, \dots$.

Then the likelihood is

$$\begin{aligned} L(\theta; x) &= \prod_{i=1}^n \binom{x_i-1}{r_i-1} \theta (1-\theta)^{S-R} \\ &\propto \theta^R (1-\theta)^{x_i-1} \end{aligned}$$

where $S = \sum_{i=1}^n x_i$ and $R = \sum_{i=1}^n r_i$.

The conjugate prior is a *beta* distribution which has a pdf proportional to

$$\theta^{a-1} (1-\theta)^{b-1}$$

for $0 < \theta < 1$.

The posterior pdf is proportional to

$$\theta^{a-1} (1-\theta)^{b-1} \times \theta^R (1-\theta)^{S-R} = \theta^{a+R-1} (1-\theta)^{b+S-R-1}.$$

This is proportional to the pdf of a $\text{beta}(a + \sum r_i, b + \sum x_i - \sum r_i)$ distribution.

8.5.2 Example

Consider the problem posed in Question 3 of Problems 1 (2.7) except that we will give the probability of a defective item a beta prior and we will continue observing items until we have found 5 defectives.

A machine is built to make mass-produced items. Each item made by the machine has a probability p of being defective. Given the value of p , the items are independent of each other. The prior distribution for p is a $\text{beta}(a, b)$ distribution. The machine is tested by counting the

number of items made before five defectives are produced. Find the posterior distribution of p given that the fifth defective item is the 73rd to be made.

Here $n = 1$, $r = 5$ and $x = 73$ so the posterior distribution is $\text{beta}(a + 5, b + 68)$. For example, suppose that $a = b = 1$ so that the prior distribution is uniform. Then the posterior distribution is $\text{beta}(6, 69)$. The prior mean is 0.5. The posterior mean is $6/(6 + 69) = 0.08$. The prior variance is $1/12 = 0.0833333$ giving a standard deviation of 0.2887. The posterior variance is 0.0009684 giving a standard deviation of 0.03112.

9 Conjugate Priors II: The gamma and Dirichlet distributions

9.1 Gamma distribution

The gamma distribution is the conjugate prior for a number of models, including Poisson and exponential data. In a later lecture we will also see that it has a role in the case of normal data.

If $\theta \sim \text{gamma}(a, b)$ then it has pdf given by

$$f(\theta) = \begin{cases} 0 & (\theta \leq 0) \\ \frac{b^a \theta^{a-1} e^{-b\theta}}{\Gamma(a)} & (0 < \theta < \infty) \end{cases}$$

where $a > 0$, $b > 0$ and $\Gamma(u)$ is the *gamma function*

$$\Gamma(u) = \int_0^\infty w^{u-1} e^{-w} dw = (u-1)\Gamma(u-1).$$

If $\theta \sim \text{gamma}(a, b)$ then the mean of θ is

$$\begin{aligned} E(\theta) &= \int_0^\infty \theta \frac{b^a \theta^{a-1} e^{-b\theta}}{\Gamma(a)} d\theta \\ &= \frac{\Gamma(a+1)}{b\Gamma(a)} \int_0^\infty \frac{b^{a+1} \theta^{a+1-1} e^{-b\theta}}{\Gamma(a+1)} d\theta \\ &= \frac{\Gamma(a+1)}{b\Gamma(a)} = \frac{a}{b}. \end{aligned}$$

Similarly

$$E(\theta^2) = \frac{\Gamma(a+2)}{b^2\Gamma(a)} = \frac{(a+1)a}{b^2}.$$

So

$$\begin{aligned} \text{var}(\theta) &= E(\theta^2) - [E(\theta)]^2 \\ &= \frac{(a+1)a}{b^2} - \frac{a^2}{b^2} = \frac{a}{b^2}. \end{aligned}$$

We call a the *shape parameter* or *index* and b the *scale parameter*. Sometimes people use $c = b^{-1}$ instead of b so the pdf becomes

$$\frac{c^{-1}(\theta/c)^{a-1} e^{-\theta/c}}{\Gamma(a)}.$$

Note that the coefficient of variation, that is the standard deviation divided by the mean, is $(\sqrt{a}/b^2)/(a/b) = 1/\sqrt{a}$.

For $a = 1$ the distribution is exponential. For large a it is more symmetric and closer to a normal distribution.

A χ^2_ν distribution is a $\text{gamma}(\nu/2, 1/2)$ distribution.

9.2 Poisson observations

9.2.1 Model

Suppose that we will observe $X_1, \dots, X_n$ where these are mutually independent given the value of θ and

$$X_i \sim \text{Poisson}(\theta).$$

Then the likelihood is

$$\begin{aligned} L(\theta; x) &= \prod_{i=1}^n \frac{e^{-\theta} \theta^{x_i}}{x_i!} \\ &\propto \theta^S e^{-n\theta} \end{aligned}$$

where $S = \sum_{i=1}^n x_i$.

The conjugate prior is a *gamma* distribution which has a pdf proportional to

$$\theta^{a-1} e^{-b\theta}$$

for $0 < \theta < \infty$.

The posterior pdf is proportional to

$$\theta^{a-1} e^{-b\theta} \times \theta^S e^{-n\theta} = \theta^{a+S-1} e^{-(b+n)\theta}.$$

This is proportional to the pdf of a $\text{gamma}(a + \sum x_i, b + n)$ distribution.

9.2.2 Example

An ecologist counts the numbers of centipedes in each of twenty one-metre-square quadrats. The numbers are as follows.

14 13 7 10 15 15 2 13 13 11 10 13 5 13 9 12 9 12 8 7

It is assumed that the numbers are independent and drawn from a Poisson distribution with mean θ . The prior distribution for θ is a gamma distribution with mean 20 and standard deviation 10. This gives $a/b = 20$ and $a/b^2 = 100$. Therefore $a = 4$ and $b = 0.2$. From the data $S = \sum x_i = 211$ and $n = 20$.

The posterior distribution for θ is $\text{gamma}(4 + 211, 0.2 + 20)$. That is $\text{gamma}(215, 20.2)$. The posterior mean is $215/20.2 = 10.64$, the posterior variance is $215/20.2^2 = 0.5269$ and the posterior standard deviation is $\sqrt{0.5269} = 0.726$.

9.3 Exponential observations

9.3.1 Model

Suppose that we will observe $X_1, \dots, X_n$ where these are mutually independent given the value of θ and

$$X_i \sim \text{exponential}(\theta).$$

Then the likelihood is

$$\begin{aligned} L(\theta; x) &= \prod_{i=1}^n \theta e^{-\theta x_i} \\ &= \theta^n e^{-S\theta} \end{aligned}$$

where $S = \sum_{i=1}^n x_i$.

The conjugate prior is a *gamma* distribution which has a pdf proportional to

$$\theta^{a-1} e^{-b\theta}$$

for $0 < \theta < \infty$.

The posterior pdf is proportional to

$$\theta^{a-1} e^{-b\theta} \times \theta^n e^{-S\theta} = \theta^{a+n-1} e^{-(b+S)\theta}.$$

This is proportional to the pdf of a $\text{gamma}(a + n, b + \sum x_i)$ distribution.

9.3.2 Example

A machine continuously produces nylon filament. From time to time the filament snaps. Suppose that the time intervals, in minutes, between snaps are random, independent and have an $\text{exponential}(\theta)$ distribution. Our prior distribution for θ is a $\text{gamma}(6, 1800)$ distribution. This gives a prior mean of $6/1800 = 0.0033$, a prior variance of $6/1800^2 = 1.85 \times 10^{-6}$ and a prior standard deviation of $\sqrt{1.85 \times 10^{-6}} = 0.00136$.

Note that the mean time between snaps is $1/\theta$. We say that this mean has an *inverse gamma* prior since its inverse has a gamma prior. The prior mean for $1/\theta$, when $\theta \sim \text{gamma}(a, b)$ with $a > 1$ is

$$\begin{aligned} E\left(\frac{1}{\theta}\right) &= \int_0^\infty \frac{\theta^{-1} b^a \theta^{a-1} e^{-b\theta}}{\Gamma(a)} d\theta \\ &= \frac{b^a \Gamma(a-1)}{b^{a-1} \Gamma(a)} \int_0^\infty \frac{b^{a-1} \theta^{a-1-1} e^{-b\theta}}{\Gamma(a-1)} d\theta \\ &= \frac{b}{a-1} \end{aligned}$$

Similarly, provided $a > 2$,

$$E\left(\frac{1}{\theta^2}\right) = \frac{b^2}{(a-1)(a-2)}$$

so

$$\text{var}\left(\frac{1}{\theta}\right) = \frac{b^2}{(a-1)(a-2)} - \left(\frac{b}{a-1}\right)^2 = \frac{b^2}{(a-1)^2(a-2)}.$$

In our case the prior mean of the time between snaps is $1800/5 = 360$ (i.e. 6 hours), the prior variance is $1800^2/(5^2 \times 4) = 32400$ and the prior standard deviation is $\sqrt{32400} = 180$ (i.e. 3 hours).

Forty intervals are observed with lengths as follows.

55	30	231	592	141	139	695	56	803	642	1890	208	246	183	38	486
264	1091	368	222	662	150	2	133	417	418	743	216	138	306	201	145
804	193	66	577	773	268	388	861								

So $S = \sum x_i = 15841$ and $n = 40$. Thus our posterior distribution for θ is $\text{gamma}(6 + 40, 1800 + 15841)$. That is it is $\text{gamma}(46, 17641)$. This gives a posterior mean of $46/17641 = 0.00261$, a posterior variance of $46/17641^2 = 1.478 \times 10^{-7}$ and a posterior standard deviation of $\sqrt{1.478 \times 10^{-7}} = 3.845 \times 10^{-4}$. The posterior mean for the time between snaps is $17641/45 = 392.0$, the posterior variance of the mean time between snaps is $17641^2/(45^2 \times 44) = 3492.76$ and the posterior standard deviation is $\sqrt{3492.76} = 59.1$.

9.4 Finding a hpd interval

Suppose that we want to find a hpd interval for a scalar parameter θ which has a unimodal posterior pdf $f^{(1)}(\theta)$ and we want this interval to have probability P . Then we require the following two conditions.

$$\begin{aligned} \int_{k_1}^{k_2} f^{(1)}(\theta) d\theta &= P \\ f^{(1)}(k_1) &= f^{(1)}(k_2) \end{aligned}$$

where k_1 is the lower limit and k_2 is the upper limit of the interval.

For example, consider the centipedes example in section 9.2.2. We can use a R function like the following.

```
function(p,a,b,delta=0.0001,maxiter=30)
{tail<-1-p
  delta<-delta^2
```

```

lower1<-0.0
lower2<-qgamma(tail,a,b)
i<-0
progress<-matrix(rep(0,(3*maxiter)),ncol=3)
write.table("Lower Upper Difference",file="")
while((diff^2>delta) & (i<maxiter))
{
  low<-(lower1+lower2)/2
  lprob<-pgamma(low,a,b)
  ldens<-dgamma(low,a,b)
  uprob<-lprob+p
  upp<-qgamma(uprob,a,b)
  udens<-dgamma(upp,a,b)
  diff<-(udens-ldens)/(udens+ldens)
  i<-i+1
  if (diff<0) lower2<-low else lower1<-low
  progress[i,1]<-low
  progress[i,2]<-upp
  progress[i,3]<-diff
}
progress<-signif(progress[1:i,],6)
write.table(progress,file="",col.names=FALSE)
result<-c(low,upp)
result
}

```

This function uses the fact that the lower limit of a $100\alpha\%$ interval has to be between 0.0 and the lower $100(1 - \alpha)\%$ point of the gamma distribution. It then uses “interval halving” to find the lower limit such that the difference between the densities at the upper and lower limits is zero.

Here we use it to find out hpd interval.

```

> hpdgamma(0.95,215,20.2)
" Lower Upper Difference"
1 4.73919 11.865 1
2 7.10879 11.865 0.999997
3 8.29359 11.8669 0.977787
4 8.88599 11.9082 0.664075
5 9.18219 12.0306 0.140152
6 9.33029 12.2034 -0.286223
7 9.25624 12.0986 -0.0541013
8 9.21921 12.0613 0.0470541
9 9.23773 12.079 -0.00244887
10 9.22847 12.07 0.0225619
11 9.2331 12.0745 0.0101224
12 9.23541 12.0767 0.00385341
13 9.23657 12.0779 0.000706445
14 9.23715 12.0785 -0.000870166
15 9.23686 12.0782 -8.15991e-05
[1] 9.23686 12.07817

```

So, we have found our 95% hpd interval to three decimal places as $9.237 < \theta < 12.078$.

9.5 The Dirichlet distribution

The Dirichlet distribution is a distribution for a set of quantities $\theta_1, \dots, \theta_m$ where $\theta_i \geq 0$ and $\sum_{i=1}^m \theta_i = 1$. An obvious application is to a set of probabilities for a partition (i.e. for an exhaustive set of mutually exclusive events).

The probability density function is

$$f(\theta_1, \dots, \theta_m) = \frac{\Gamma(A)}{\prod_{i=1}^m \Gamma(a_i)} \prod_{i=1}^m \theta_i^{a_i-1}$$

where $A = \sum_{i=1}^m a_i$ and $a_1, \dots, a_m$ are parameters with $a_i > 0$ for $i = 1, \dots, m$.

Clearly, if $m = 2$, we obtain a beta(a_1, a_2) distribution as a special case.

The mean of θ_j is

$$\begin{aligned} E(\theta_j) &= \int \dots \int \theta_j \frac{\Gamma(A)}{\prod_{i=1}^m \Gamma(a_i)} \prod_{i=1}^m \theta_i^{a_i-1} d\theta_1 \dots d\theta_m \\ &= \frac{\Gamma(A)}{\Gamma(A+1)} \frac{\Gamma(a_j+1)}{\Gamma(a_j)} \int \dots \int \frac{\Gamma(A+1)}{\prod_{i=1}^m \Gamma(a'_i)} \prod_{i=1}^m \theta_i^{a'_i-1} d\theta_1 \dots d\theta_m \\ &= \frac{\Gamma(A)}{\Gamma(A+1)} \frac{\Gamma(a_j+1)}{\Gamma(a_j)} \\ &= \frac{a_j}{A} \end{aligned}$$

where $a'_i = a_i$ when $i \neq j$ and $a'_j = a_j + 1$.

Similarly

$$E(\theta_j^2) = \frac{\Gamma(A)}{\Gamma(A+2)} \frac{\Gamma(a_j+2)}{\Gamma(a_j)} = \frac{(a_j+1)a_j}{(A+1)A}$$

so

$$\begin{aligned} \text{var}(\theta_j) &= \frac{(a_j+1)a_j}{(A+1)A} - \left(\frac{a_j}{A}\right)^2 \\ &= \frac{a_j}{A(A+1)} - \frac{a_j^2}{A^2(A+1)} \end{aligned}$$

Also

$$E(\theta_j \theta_k) = \frac{\Gamma(A)}{\Gamma(A+2)} \frac{\Gamma(a_j+1)}{\Gamma(a_j)} \frac{\Gamma(a_k+1)}{\Gamma(a_k)} = \frac{a_j a_k}{(A+1)A}$$

so

$$\begin{aligned} \text{covar}(\theta_j, \theta_k) &= \frac{a_j a_k}{(A+1)A} - \frac{a_j}{A} \frac{a_k}{A} \\ &= -\frac{a_j a_k}{A^2(A+1)} \end{aligned}$$

9.6 Multinomial observations

9.6.1 Model

Suppose that we will observe $X_1, \dots, X_m$ where these are the frequencies for categories $1, \dots, m$, the total $N = \sum_{i=1}^m X_i$ is fixed and the probabilities for these categories are $\theta_1, \dots, \theta_m$ where $\sum_{i=1}^m \theta_i = 1$. Then, given θ , where $\theta = (\theta_1, \dots, \theta_m)^T$, the distribution of $X_1, \dots, X_m$ is multinomial with

$$\Pr(X_1 = x_1, \dots, X_m = x_m) = \frac{N!}{\prod_{i=1}^m x_i!} \prod_{i=1}^m \theta_i^{x_i}.$$

Notice that, with $m = 2$, this is just a binomial(N, θ_1) distribution.

Then the likelihood is

$$\begin{aligned} L(\theta; x) &= \frac{N!}{\prod_{i=1}^m x_i!} \prod_{i=1}^m \theta_i^{x_i} \\ &\propto \prod_{i=1}^m \theta_i^{x_i}. \end{aligned}$$

The conjugate prior is a *Dirichlet* distribution which has a pdf proportional to

$$\prod_{i=1}^m \theta_i^{a_i-1}.$$

The posterior pdf is proportional to

$$\prod_{i=1}^m \theta_i^{a_i-1} \times \prod_{i=1}^m \theta_i^{x_i} = \prod_{i=1}^m \theta_i^{a_i+x_i-1}.$$

This is proportional to the pdf of a Dirichlet distribution with parameters $a_1 + x_1, a_2 + x_2, \dots, a_m + x_m$.

9.6.2 Example

In a survey 1000 English voters are asked to say for which party they would vote if there were a general election next week. The choices offered were 1: Labour, 2: Liberal, 3: Conservative, 4: Other, 5: None, 6: Undecided. We assume that the population is large enough so that the responses may be considered independent given the true underlying proportions. Let $\theta_1, \dots, \theta_6$ be the probabilities that a randomly selected voter would give each of the responses. Our prior distribution for $\theta_1, \dots, \theta_6$ is a Dirichlet() distribution.

This gives the following summary of the prior distribution.

Response	a_i	Prior mean	Prior var.	Prior sd.
Labour	5	0.25	0.008929	0.09449
Liberal	3	0.15	0.006071	0.07792
Conservative	5	0.25	0.008929	0.09449
Other	1	0.05	0.002262	0.04756
None	2	0.10	0.004286	0.06547
Undecided	4	0.20	0.007619	0.08729
Total	20	1.00		

Suppose our observed data are as follows.

Labour	Liberal	Conservative	Other	None	Undecided
256	131	266	38	114	195

Then we can summarise the posterior distribution as follows.

Response	$a_i + x_i$	Posterior mean	Posterior var.	Posterior sd.
Labour	261	0.2559	0.0001865	0.01366
Liberal	134	0.1314	0.0001118	0.01057
Conservative	271	0.2657	0.0001911	0.01382
Other	39	0.0382	0.0000360	0.00600
None	116	0.1137	0.0000987	0.00994
Undecided	199	0.1951	0.0001538	0.01240
Total	1020	1.0000		

10 More on Likelihood: Large sample approximations

10.1 Domination by the likelihood

We have an unknown parameter θ with prior pdf $f_{\theta}^{(0)}(\theta)$ and we will make observations $Y_1, \dots, Y_n$. Let us suppose that, given θ , $Y_1, \dots, Y_n$ are independent. (Similar results to the following can be proved in many cases where they are not). Let the pdf of Y_i given θ be $f_{Y_i|\theta}(y_i | \theta)$. The likelihood is then

$$L(\theta; y) = \prod_{i=1}^n f_{Y_i|\theta}(y_i | \theta) = \prod_{i=1}^n L_i(\theta; y_i).$$

The posterior density is

$$f_{\theta|y}^{(1)}(\theta | y) = C f_{\theta}^{(0)}(\theta) \prod_{i=1}^n L_i(\theta; y_i).$$

Assuming that neither the prior density or the likelihood is zero, the log of the posterior density is

$$\log\{f_{\theta|y}^{(1)}(\theta | y)\} = \log(C) + \log\{f_{\theta}^{(0)}(\theta)\} + \sum_{i=1}^n l_i(\theta; y_i) \quad (1)$$

where $l_i(\theta; y_i) = \log\{L_i(\theta; y_i)\}$. Clearly, as n increases, the second (prior) term on the right hand side of (1) stays constant but the third (likelihood) term becomes more important. In the limit as $n \rightarrow \infty$ the posterior is dominated by the likelihood, provided that the prior is nonzero everywhere. Thus in large samples the posterior density is approximately proportional to the likelihood and the *posterior mode* is approximately equal to the maximum likelihood estimate.

10.2 Approximate normality

10.2.1 Scalar case

Subject to certain conditions, when we have a large sample the shape of the likelihood is approximately the shape of a normal distribution.

Suppose we have a scalar parameter θ . Let the log likelihood be $l(\theta)$ and the maximum likelihood estimate be $\hat{\theta}$. We can expand the log likelihood in a Taylor series about $\hat{\theta}$.

$$l(\theta) = l(\hat{\theta}) + (\theta - \hat{\theta}) \left(\frac{dl(\theta)}{d\theta} \right)_{\hat{\theta}} + \frac{1}{2!} (\theta - \hat{\theta})^2 \left(\frac{d^2l(\theta)}{d\theta^2} \right)_{\hat{\theta}} + R(\theta, \hat{\theta})$$

where $R(\theta, \hat{\theta})$ represents the higher-order terms which are expected to be small when θ is close to $\hat{\theta}$. Now

$$\left(\frac{dl(\theta)}{d\theta} \right)_{\hat{\theta}} = 0$$

so

$$l(\theta) \approx C - \frac{1}{2} (\theta - \hat{\theta})^2 / V$$

where C is some constant and $V = -J^{-1}$ where

$$J = \left(\frac{d^2l(\theta)}{d\theta^2} \right)_{\hat{\theta}}$$

is the (observed) *Fisher information*.

Exponentiating we see that the posterior density of θ is approximately proportional to

$$\exp \left\{ -\frac{1}{2} (\theta - \hat{\theta})^2 / V \right\}$$

so the posterior distribution of θ is approximately a normal distribution with mean $\hat{\theta}$ and variance $V = -J^{-1}$.

Note that we can often approximate the shape of a posterior distribution with a normal distribution. The mode of a normal $N(M, V)$ distribution is M so, if we can find the posterior mode, for example by maximising the sum of the log prior and log likelihood, we can approximate the posterior mean by M . The normal probability density function is

$$(2\pi V)^{-1/2} \exp \left\{ -\frac{1}{2V} (\theta - M)^2 \right\}.$$

The log of this is

$$g(\theta) = -\frac{1}{2} \log(2\pi V) - \frac{1}{2V} (\theta - M)^2.$$

Differentiating this wrt θ we obtain

$$\frac{dg(\theta)}{d\theta} = -\frac{1}{V}(\theta - M).$$

Differentiating again we obtain

$$\frac{d^2g(\theta)}{d\theta^2} = -\frac{1}{V}.$$

So, if we differentiate the sum of the log prior and the log likelihood twice and evaluate this second derivative at the posterior mode we get an approximation to $-1/V$.

These results allow us, for example, to find approximate credible intervals.

Note that this is an *asymptotic* result for large samples. we could derive similar results for some transformation of θ , e.g. $\log(\theta)$. The normal approximation might work better if we transform θ in some way. For example, a parameter θ might be restricted to an interval, such as $(0, 1)$ and the approximation might be better if we transform so that θ can take any value on the real line, e.g. $\eta = \log[\theta/(1 - \theta)]$.

10.2.2 The vector case

Now θ is a vector parameter and its maximum likelihood estimate $\hat{\theta}$ is also a vector. The Taylor series expansion becomes

$$l(\theta) = l(\hat{\theta}) + (\theta - \hat{\theta})^T \left(\frac{\partial l(\theta)}{\partial \theta} \right)_{\hat{\theta}} + \frac{1}{2!} (\theta - \hat{\theta})^T \left(\frac{\partial^2 l(\theta)}{\partial \theta \partial \theta^T} \right)_{\hat{\theta}} (\theta - \hat{\theta}) + R(\theta, \hat{\theta})$$

This leads to an approximately multivariate normal posterior distribution with mean vector M and variance matrix V where M is the posterior mode (or the maximum likelihood estimate) and

$$V = -J^{-1}$$

where J is the (observed) Fisher information matrix, the matrix of partial second derivatives of the log likelihood (or the log of the prior density plus the log likelihood), evaluated at the mode.

10.2.3 The multivariate normal distribution

Suppose that a vector random variable $X = (X_1, \dots, X_n)^T$ has a multivariate normal $N(M, V)$ distribution. This distribution has a *mean vector* $M = (m_1, \dots, m_n)^T$ where m_i is the mean of X_i . In place of the scalar variance in the univariate case we have a *variance matrix* V . The diagonal elements of V are the variances of $X_1, \dots, X_n$ with the element in row and column i , v_{ii} being the variance of X_i . The covariance of X_i and X_j is v_{ij} , the element in row i and column j . Clearly $v_{ij} = v_{ji}$ and V is symmetric. It is also positive semi-definite.

The pdf is

$$f_X(x) = (2\pi)^{-n/2} |V|^{-1/2} \exp \left\{ -\frac{1}{2} (x - M)^T V^{-1} (x - M) \right\}.$$

The marginal distribution of X_i is $N(m_i, v_{ii})$.

10.3 Examples

1. The numbers of telephone calls arriving at a busy exchange in 200 short intervals are observed. Suppose that these are $X_1, \dots, X_n$, where $n = 200$, that $X_i \mid \lambda \sim \text{Poisson}(\lambda)$ and that X_i and X_j are independent, given λ , when $i \neq j$. The likelihood is therefore

$$L = \prod_{i=1}^n \frac{e^{-\lambda} \lambda^{x_i}}{x_i!} \propto e^{-n\lambda} \lambda^s$$

where $s = \sum_{i=1}^n x_i$.

Suppose that the prior for λ is a gamma(a, b) distribution. The posterior is then gamma($s + a, n + b$). This has mean $(s + a)/(n + b)$ and variance $(s + a)/(n + b)^2$ and we can approximate this with a normal distribution

$$N\left(\frac{s + a}{n + b}, \frac{s + a}{(n + b)^2}\right).$$

Suppose that $a = 2$, $b = 0.2$, $s = 1723$. Then the posterior distribution is approximately $N(8.6164, 0.2075^2)$ and an approximate 95% interval for λ is $8.6164 \pm 1.96 \times 0.2075$. This gives

$$8.210 < \lambda < 9.023.$$

Alternatively we can use the posterior mode to find the mean of the normal distribution. The posterior density is proportional to

$$\lambda^{s+a-1} e^{-(n+b)\lambda}.$$

The log of this is

$$g(\lambda) = (s + a - 1) \log(\lambda) - (n + b)\lambda.$$

Differentiating wrt λ we obtain

$$g'(\lambda) = \frac{s + a - 1}{\lambda} - (n + b).$$

This gives the posterior mode at $\lambda_m = (s + a - 1)/(n + b) = 8.6114$. Differentiating again we obtain

$$g''(\lambda) = -\frac{s + a - 1}{\lambda^2}.$$

This gives an approximate variance of

$$\frac{\lambda_m^2}{s + a - 1} = \frac{s + a - 1}{(n + b)^2} = 0.2075^2.$$

So our 95% interval becomes $8.6114 \pm 1.96 \times 0.2075$. This gives

$$8.205 < \lambda < 9.018.$$

Calculating the hpd interval without using the normal approximation (using `hpdgamma`) gives

$$8.211 < \lambda < 9.024$$

which is very close to the first of our approximations.

2. The proportion of people, given a drug to treat a medical condition, who contract a particular side effect depends on the dose of the drug. If p is the proportion suffering the side effect at dose x , then

$$\ln\left(\frac{p}{1-p}\right) = \alpha + \beta x$$

where α and β are parameters with unknown values.

At each of a number of doses, x_i , a number, n_i , of patients were given the drug and the number, r_i , with the side effect was recorded.

Dose	x_i	0.9	1.1	1.8	2.3	3.0	3.3	4.0
No. patients	n_i	46	72	118	96	84	53	38
No. with side effect	r_i	17	22	52	58	56	43	30

Our prior distribution for α and β is a bivariate normal distribution with mean vector $(m_\alpha, m_\beta)^T$ and variance matrix

$$V = \begin{pmatrix} v_{\alpha,\alpha} & v_{\alpha,\beta} \\ v_{\alpha,\beta} & v_{\beta,\beta} \end{pmatrix}.$$

The prior *precision matrix* is

$$W = V^{-1} = \begin{pmatrix} w_{\alpha,\alpha} & w_{\alpha,\beta} \\ w_{\alpha,\beta} & w_{\beta,\beta} \end{pmatrix}.$$

The likelihood is proportional to

$$L = \prod_{i=1}^n p_i^{r_i} (1 - p_i)^{n_i - r_i},$$

the prior density is proportional to

$$\exp \left\{ -\frac{1}{2} (\alpha - m_\alpha, \beta - m_\beta)^T W (\alpha - m_\alpha, \beta - m_\beta) \right\}$$

and the posterior density is proportional to the product of these.

The logarithm of the posterior density, apart from a constant, is

$$\begin{aligned} g(\alpha, \beta) &= -\frac{1}{2} [(\alpha - m_\alpha)^2 w_{\alpha,\alpha} + (\beta - m_\beta)^2 w_{\beta,\beta} + 2(\alpha - m_\alpha)(\beta - m_\beta)w_{\alpha,\beta}] \\ &\quad + \sum r_i \log(p_i) + (n_i - r_i) \log(1 - p_i) \end{aligned}$$

Now

$$\begin{aligned} \frac{\partial p_i}{\partial \alpha} &= \frac{(1 + e^{\alpha + \beta x_i})e^{\alpha + \beta x_i} - e^{\alpha + \beta x_i}e^{\alpha + \beta x_i}}{(1 + e^{\alpha + \beta x_i})^2} \\ &= \frac{e^{\alpha + \beta x_i}}{(1 + e^{\alpha + \beta x_i})^2} = p_i(1 - p_i) \\ \frac{\partial p_i}{\partial \beta} &= \frac{(1 + e^{\alpha + \beta x_i})x_i e^{\alpha + \beta x_i} - e^{\alpha + \beta x_i}x_i e^{\alpha + \beta x_i}}{(1 + e^{\alpha + \beta x_i})^2} \\ &= \frac{x_i e^{\alpha + \beta x_i}}{(1 + e^{\alpha + \beta x_i})^2} = x_i p_i(1 - p_i) \end{aligned}$$

So

$$\begin{aligned} \frac{\partial g(\alpha, \beta)}{\partial \alpha} &= -[(\alpha - m_\alpha)w_{\alpha,\alpha} + (\beta - m_\beta)w_{\alpha,\beta}] + \sum \frac{r_i}{p_i} \frac{\partial p_i}{\partial \alpha} - \sum \frac{(n_i - r_i)}{1 - p_i} \frac{\partial p_i}{\partial \alpha} \\ &= -[(\alpha - m_\alpha)w_{\alpha,\alpha} + (\beta - m_\beta)w_{\alpha,\beta}] + \sum r_i(1 - p_i) - \sum (n_i - r_i)p_i \\ &= -[(\alpha - m_\alpha)w_{\alpha,\alpha} + (\beta - m_\beta)w_{\alpha,\beta}] + \sum (r_i - n_i p_i) \end{aligned}$$

Similarly

$$\frac{\partial g(\alpha, \beta)}{\partial \beta} = -[(\beta - m_\beta)w_{\beta,\beta} + (\alpha - m_\alpha)w_{\alpha,\beta}] + \sum x_i(r_i - n_i p_i).$$

We set both of these partial derivatives equal to zero and solve for α and β to find the posterior mode but this requires an iterative numerical method. We can conveniently use Newton-Raphson iterations. For this we need the partial second derivatives.

$$\begin{aligned}
\frac{\partial^2 g(\alpha, \beta)}{\partial \alpha^2} &= -w_{\alpha, \alpha} - \sum n_i \frac{\partial p_i}{\partial \alpha} = -w_{\alpha, \alpha} - \sum n_i p_i (1 - p_i) \\
\frac{\partial^2 g(\alpha, \beta)}{\partial \beta^2} &= -w_{\beta, \beta} - \sum x_i n_i \frac{\partial p_i}{\partial \beta} = -w_{\beta, \beta} - \sum x_i^2 n_i p_i (1 - p_i) \\
\frac{\partial^2 g(\alpha, \beta)}{\partial \alpha \partial \beta} &= -w_{\alpha, \beta} - \sum n_i \frac{\partial p_i}{\partial \beta} = -w_{\alpha, \beta} - \sum x_i n_i p_i (1 - p_i)
\end{aligned}$$

We start with an initial guess for the posterior mode. For example we can find a starting value using a least squares regression of $\log[\hat{p}_i/(1 - \hat{p}_i)]$ on x_i where $\hat{p}_i = r_i/n_i$ with suitable adjustments where $r_i = 0$ or $r_i = n_i$.

Then let $\alpha^{(j)}, \beta^{(j)}$ be the values at iteration j . Let

$$\begin{aligned}
\theta^{(j)} &= \begin{pmatrix} \alpha^{(j)} \\ \beta^{(j)} \end{pmatrix} \\
h(\theta^{(j)}) &= \begin{pmatrix} \frac{\partial g(\alpha, \beta)}{\partial \alpha} \\ \frac{\partial g(\alpha, \beta)}{\partial \beta} \end{pmatrix}_{\alpha^{(j)}, \beta^{(j)}} \\
H(\theta^{(j)}) &= \begin{pmatrix} \frac{\partial^2 g(\alpha, \beta)}{\partial \alpha^2} & \frac{\partial^2 g(\alpha, \beta)}{\partial \alpha \partial \beta} \\ \frac{\partial^2 g(\alpha, \beta)}{\partial \alpha \partial \beta} & \frac{\partial^2 g(\alpha, \beta)}{\partial \beta^2} \end{pmatrix}_{\alpha^{(j)}, \beta^{(j)}}
\end{aligned}$$

Then each iteration is as follows.

$$\theta^{(j+1)} = \theta^{(j)} - H^{-1}(\theta^{(j)})h(\theta^{(j)}).$$

Here is a R function to do the calculations.

```

blogreg<- function(prior,x,n,r,iter=10)
{# prior is a vector containing :
# prior mean of alpha
# prior mean of beta
# prior variance of alpha
# prior variance of beta
# prior covariance of alpha, beta
det<-prior[3]*prior[4]-prior[5]^2
ppaa<-prior[4]/det
ppbb<-prior[3]/det
ppab<-prior[5]/det
p<-r/n
p<-pmax(p,0.01)
p<-pmin(p,0.99)
eta<-log(p/(1-p))
xbar<-mean(x)
ybar<-mean(eta)
sxx<-sum((x-xbar)^2)
sxy<-sum((x-xbar)*(eta-ybar))
beta<-sxy/sxx
alpha<-ybar-beta*xbar
theta<-c(alpha,beta)
d1<-c(0,0)
d2<-c(0,0,0,0)
dim(d2)<-c(2,2)
for (i in 1:iter)
  {d1[1]<-sum(r-n*p)-((theta[1]-prior[1])*ppaa+(theta[2]-prior[2])*ppab)

```

```

d1[2]<-sum(x*(r-n*p))-((theta[2]-prior[2])*ppbb+(theta[1]-prior[1])*ppab)
d2[1,1]<- -sum(n*p*(1-p))-ppaa
d2[2,2]<- -sum(n*p*(1-p)*x*x)-ppbb
d2[1,2]<- -sum(n*p*(1-p)*x)-ppab
d2[2,1]<-d2[1,2]
theta<-theta-solve(d2,d1)
write(theta,file="")
eta<-theta[1]+theta[2]*x
p<-exp(eta)/(1+exp(eta))
}
d2<- -solve(d2)
ans<-list(mean=theta,var=d2)
ans
}

```

Suppose our prior means for α and β are -2 and 1 respectively, our prior variances are both 1 and our prior covariance is 0 . Then using the function gives the following result.

```

> blogreg(prior,x,n,r)
-1.449914 0.758535
-1.554639 0.7945546
-1.555514 0.7948921
-1.555514 0.7948921
-1.555514 0.7948921
-1.555514 0.7948921
-1.555514 0.7948921
-1.555514 0.7948921
-1.555514 0.7948921
-1.555514 0.7948921
-1.555514 0.7948921
$mean
[1] -1.5555143 0.7948921

$var
      [,1]      [,2]
[1,] 0.06155529 -0.02465905
[2,] -0.02465905 0.01154082

```

Thus our posterior distribution for α , β is approximately a bivariate normal distribution with means -1.5555 and 0.7949 , variances of 0.06156 and 0.01154 and a covariance of -0.02466 .

10.4 Problems 3

Answer all questions. Solutions are to be submitted in the Homework Box no later than 5.00pm on Wednesday March 14th. You may use, for example, R to do numerical calculations where appropriate. You should show the commands which you have used.

1. I have records of the attendance of students at tutorials for this module. Suppose that we can, in some sense, regard the students as a sample from some population of students so that, for example, we can learn about the likely behaviour of next year's students by observing this year's. So far we have had tutorials in Week 2 and Week 4. Let the probability that a student attends in both weeks be θ_{11} , the probability that a student attends in week 2 but not Week 4 be θ_{10} and so on. The data are as follows.

Attendance	Probability	Observed frequency
Week 2 and Week 4	θ_{11}	$n_{11} = 25$
Week 2 but not Week 4	θ_{10}	$n_{10} = 7$
Week 4 but not Week 2	θ_{01}	$n_{01} = 6$
Neither week	θ_{00}	$n_{00} = 13$

Suppose that the prior distribution for $(\theta_{11}, \theta_{10}, \theta_{01}, \theta_{00})$ is a Dirichlet distribution with density proportional to

$$\theta_{11}^3 \theta_{10} \theta_{01} \theta_{00}^2.$$

- (a) Find the prior means and prior variances of θ_{11} , θ_{10} , θ_{01} , θ_{00} .
 - (b) Find the posterior distribution.
 - (c) Find the posterior means and posterior variances of θ_{11} , θ_{10} , θ_{01} , θ_{00} .
 - (d) Using R (or otherwise), find a 95% posterior hpd interval, based on the exact posterior distribution, for θ_{00} .
 - (e) Find an approximate 95% hpd interval for θ_{00} using a normal approximation based on the posterior mode and the partial second derivatives of the log posterior density. Compare this with the exact hpd interval.
Hint: To find the posterior mode you will need to introduce a Lagrange multiplier.
 - (f) The population mean number of attendances out of two is $\mu = 2\theta_{11} + \theta_{10} + \theta_{01}$. Find the posterior mean of μ and an approximation to the posterior standard deviation of μ .
2. The lifetimes of certain components are supposed to follow a Weibull distribution with known shape parameter $\alpha = 2$. The probability density function of the lifetime distribution is

$$f(t) = \alpha \rho^2 t \exp[-(\rho t)^2]$$

for $0 < t < \infty$.

We will observe a sample of n such lifetimes where n is large.

- (a) Assuming that the prior density is nonzero and reasonably flat so that it may be disregarded, find an approximation to the posterior distribution of ρ . Find an approximate 95% hpd interval for ρ when $n = 300$, $\sum \log(t) = 1305.165$ and $\sum t^2 = 3161776$.
- (b) Assuming that the prior distribution is a gamma(a, b) distribution, find an approximate 95% hpd interval for ρ , taking into account this prior, when $a = 2$, $b = 100$, $n = 300$, $\sum \log(t) = 1305.165$ and $\sum t^2 = 3161776$.