Chapter 1
THE BAYESIAN ALGORITHM

An algorithm is a set of rules for doing a calculation. The Bayesian algorithm is a set of rules for using evidence (data) to change your beliefs. In this chapter we shall try to explain this algorithm. If this explanation is successful the reader may then put down the book and start doing Bayesian econometrics, for the rest of the book is little more than illustrative examples and technical details. Thus, chapter 1 and, to a lesser extent, chapter 2 are the most important parts of the book.

We begin by explaining how we view an econometric analysis and by drawing a distinction between this and statistical analyses.

1.1 ECONOMETRIC ANALYSIS

An econometric analysis is the confrontation of an economic model with evidence. An **economic model** usually asserts a relation between economic variables. For example, a model might assert that consumption, C, is linearly related to income, Y, so that $C = \alpha + \beta Y$ for some pair of numbers α, β. Such a model is typically intended to provide a causal explanation of how some variables, for example C, are determined by the values of others, for example Y. Typically, any model contains both potentially observable quantities, such as consumption and income, called (**potential**) **data**; and it involves quantities, like α and β, that are not directly observable. Variables of this latter type are called **parameters** and will be denoted generically by the symbol θ. They are usually constrained to lie in a set to be denoted by Θ. In our example $\theta = (\alpha, \beta)$ and the set Θ would normally be taken as two-dimensional euclidean space. Any value of θ, for example $\alpha = 10$, $\beta = 0.9$, defines a particular **structure**, in this case $C = 10 + 0.9Y$, and the set of structures under consideration is said to be **indexed** by a parameter, θ.

Evidence is provided by data on the operation of an economy. In the consumption/income example relevant data would be provided by pairs of values for C and Y. There are usually many types of data that are relevant to any particular model. For example, we might have data on the consumption and income of different households, or on the same household observed repeatedly, or on the aggregate income and consumption data of collections of households forming a region or a nation.

The objective of an econometric analysis is to answer two questions. The first question is whether the model is consistent with the evidence: this is called model criticism. This means asking whether any of the structures defined by the model are consistent with the evidence. In our example this would mean asking whether there is any parameter $\theta = (\alpha, \beta)$, lying in Θ, such that, in our data, $C = \alpha + \beta Y$. The second question presumes that the answer to the first is “yes” and it asks what are the probabilities of the different structures defined by the model. Once this question has been answered the model can then be used for purposes of economic decision making, perhaps by a policy maker, perhaps by an individual economic agent. Such use will typically involve predicting the value of the variables for households or regions that are not included in the data. For example, given the structure $\theta = (10, 0.9)$ and told that $Y = 100$ then the economist would predict that $C = 10 + 0.9 \times 100 = 100$.

The practice of econometrics is, in fact, to ask these questions in reverse order. We begin by presuming that our model is consistent with the data and ask for the most likely structure in the light of the evidence. In traditional econometrics this involves forming a good estimate of $\theta_0 \in \Theta$, the particular structure that is presumed to be, in some sense, true. In a Bayesian analysis this step involves using the data to form a probability distribution over the structures in Θ. An estimate, if one is required, might then be provided by reporting, for example, the most probable structure in the light of the evidence provided by the data.

How then do we go about answering these questions in practice? In this chapter we shall focus on the second question in which we presume the consistency of the model with the data and ask how we determine the probabilities of the structures of which the model is composed. The method of doing this is to apply a theorem of probability, Bayes’ theorem, and here we shall describe in some detail how Bayes’ theorem is used to construct probabilities over alternative structures.

In chapter 2 we shall describe some methods of answering the first question in which the investigator tries to decide whether the model is consistent with the evidence and if it is not, what to do next.

1.2 STATISTICAL ANALYSIS

Statistical analysis deals with the study of numerical data. This is a largely descriptive activity in which the primary aim is to find effective and economical representations or summaries of such data. The point of the activity is to reduce the complexity of a set of numbers to a form which can be more easily comprehended.[1]

[1] For instance, “. . . the object of statistical methods is the reduction of data. A quantity of data, which usually by its mere bulk is incapable of entering the mind, is to be replaced by relatively few quantities which shall adequately represent the whole, or which, in other words, shall contain as much as possible, ideally the whole, of the relevant information contained in the original data.” R. A. Fisher, On the mathematical foundations of theoretical statistics, *Phil. Trans. Royal Soc.*, A222, 1922, p. 309, quoted in T. C. Koopmans, *Linear Regression Analysis of Economic Time Series*, Netherlands Economic Institute, Haarlem, 1937.

The statistician summarizes data by calculating means, standard deviations, trends or regression lines; he represents data graphically by scatter diagrams, histograms, kernel smoothers and many other devices. He typically proposes and applies statistical models as simplified accounts of possible ways in which his data could have occurred. The application of such models involves estimating the parameters of such models and testing hypotheses about them.

Statistical analysis is in many ways very close to econometrics, a subject which, to a statistician, can appear like a branch of applied statistics. Econometric technique is largely drawn from statistics and much of the content of this book will be familiar to a statistician. Indeed, in writing it I have drawn extensively on statistical books and articles. But there are profound differences between econometrics and statistics. The econometrician is primarily concerned with the analysis of the behavior of economic agents and their interactions in markets and the analysis of data is secondary to that concern. But markets can be in, or near, equilibrium; economic agents are presumed to be maximizing or minimizing some objective function; economic agents are often presumed to know relevant things that the econometrician does not. All these considerations tend to be fundamental to an econometric analysis and to dictate the class of models that are worth considering. They make the results of an econometric analysis interpretable to the economist and give parameters solid meaning.

Of course there is not and should not be a sharp line between econometrics and statistics; there is nothing at all wrong with an economist parsimoniously describing data or with a statistician trying to relate the parameters of his model to some underlying theory. But the distinction between the disciplines exists, in my view, and should be kept in mind.

1.3 BAYES' THEOREM

Bayesian econometrics is the systematic use of a result from elementary probability, Bayes' theorem. Indeed, from one angle, that's all it is. There are not multiple methods of using numerical evidence to revise beliefs – there is only one – so this theorem is fundamental.

What is Bayes' theorem?

When A and B are two events defined on a sample space the **conditional probability** that A occurs given that B has occurred is defined as

$$P(A|B) = \frac{P(B \cap A)}{P(B)} \tag{1.1}$$

as long as $P(B) \neq 0$. Here $P(B \cap A)$ is the probability that both A and B occur and $P(A|B)$ is the probability that A occurs given the knowledge that B has occurred. Equation (1.1) is true, of course, with A and B interchanged so that we also have $P(B \cap A) = P(B|A)P(A)$. Substituting this expression into (1.1) then gives

$$P(A|B) = \frac{P(B|A)P(A)}{P(B)}. \tag{1.2}$$

When written in this form the definition is called **Bayes' theorem**. It is a universally accepted mathematical proposition. But there is disagreement about its applicability to econometrics.

Two related questions arise about Bayes' theorem: one is about its interpretation and the other is about its use.

Interpretations of probability

The function $P(.)$ has no interpretation in the mathematical theory of probability; all the theory does is define its properties. When probability theory is applied, as it is in econometrics, we need to decide how to interpret "the probability of an event." In this book we shall take $P(A)$ to measure the strength of belief in the proposition that A is true.[2] Thus the larger $P(A)$ the stronger your belief in the proposition that it represents. This is called a **subjective** view of probability. This interpretation of mathematical probabilities is close to the way we use the idea of probability in everyday language where we say that propositions are "very probable" or "highly unlikely." This closeness to ordinary usage is part of the attraction of Bayesian inference for many people. It allows us to conclude an econometric analysis by saying things such as "in the light of the evidence, theory A is very unlikely whereas theory B is quite probable." Oddly enough, statements like these are impermissible in traditional econometrics where theories are only true or false, not more or less probable.

Similarly, a probability density function for a random variable, $p_X(x)$, will describe degrees of belief in the occurrence of the various possible values of X. Degrees of belief, like utility functions to which they are closely related, are personal to each economic agent, so when you do applied economics using (1.2) you are in fact manipulating your beliefs. On this interpretation Bayes' theorem shows how one belief about A, measured by $P(A)$, is changed into another belief about A, measured by $P(A|B)$.

Range of application of the theorem

The second issue is the range of application of Bayes' theorem. Some people[3] choose to restrict the application of probability theory, including Bayes' theorem, to situations in which there is a series of repetitions in some of which A occurs and in the others it does not. Consequently $P(A)$ is understood as referring to the relative frequency

[2] **Degrees of belief** may be more precisely defined in terms of willingness to bet on the occurrence of A. We shall not pursue this line of thought but rather take "degree of belief" as a primitive one that most can understand intuitively. Further references on this subjective view of probability are provided in the bibliographical notes.

[3] Including virtually all authors of econometric textbooks.

with which A occurs during such repetitions. Econometric work that is done solely using this conception of probability as hypothetical relative frequency is often called "frequentist." We shall adopt a much more general range of application in which it will be meaningful to use the expression $P(A)$ to refer to all events, whether they are part of a repeatable sequence or not. Thus we can talk about the probability that the United States was in recession in 1983, that the moon is made of green cheese, or that the economy is in general equilibrium, all of which are events about which you may be sure or quite unsure but about which I assume you have beliefs and those beliefs are capable of being changed by evidence. Unless you have the imagination of a Jules Verne[4] there is no sequence of occasions in some of which the economy was in recession in 1983, or the moon was made of green cheese, and in others of which it was not. Narrower, frequency, interpretations of probability rule out such uses of $P(A)$.

Use of Bayes' theorem to make inferences

How the theorem expressed by (1.2) may be used as the basis for econometric inference from evidence may be shown by considering particular types of events A and B. Suppose that you have a model containing just two structures θ_1 and θ_2, so Θ contains just two elements, and you also have some data E.

EXAMPLE 1.1 *Take the consumption and income relation as an example and let structure* 1 *assert that* $C = 10 + 0.9Y$ *and let structure* 2 *assert that* $C = Y$. *Thus*

$$\theta_1 = (10, 0.9); \quad \theta_2 = (0, 1).$$

We now interpret the A of Bayes' theorem as referring to particular structures – you can think of these as alternative theories if you like – and we interpret B as describing the evidence E. So the event A is either "θ_1 is true" or it is "θ_2 is true." Suppose you think each structure equally probable so that $P(\theta_1) = P(\theta_2) = 0.5$. (There may be many other structures that you can think of but, for the moment, you are content to consider only these two.) You also, from the content of these two theories, form beliefs about the probabilities of the evidence given that either one is true. This means that you can place a number on both $P(E|\theta_1)$ and $P(E|\theta_2)$. For example, you might think that E, the data, is quite unlikely if θ_1 is true but fairly probable if θ_2 is true, say $P(E|\theta_1) = 0.1$ and $P(E|\theta_2) = 0.6$. Now note that these numbers and the rules of probability imply that

$$P(E) = P(E|\theta_1)P(\theta_1) + P(E|\theta_2)P(\theta_2) = 0.1 \times 0.5 + 0.6 \times 0.5 = 0.35,$$

[4] Pioneering French science fiction writer.

so E, the event that we have observed, was not particularly probable. We are now in a position to make an inference about the two structures θ_1 and θ_2. It follows from Bayes' theorem (1.2) that

$$P(\theta_1|E) = \frac{P(E|\theta_1)P(\theta_1)}{P(E)} = \frac{0.05}{0.35} = \frac{1}{7},$$

$$P(\theta_2|E) = \frac{P(E|\theta_2)P(\theta_2)}{P(E)} = \frac{0.30}{0.35} = \frac{6}{7},$$

We interpret these numbers as saying that although the theories θ_1 and θ_2 were believed to be equally probable before seeing the evidence represented by E, after the evidence has been seen theory θ_2 is six times more probable than theory θ_1. You have changed your beliefs. This is not an arbitrary change of opinion. It follows as an arithmetical consequence of the beliefs represented by $P(\theta_1) = P(\theta_2) = 1/2$ and $P(E|\theta_1) = 0.1$, $P(E|\theta_2) = 0.6$. Moreover this change of opinion will occur (almost) whatever $P(\theta_1)$ and $P(\theta_2)$ you had. They need not have been equally probable, you might have thought θ_2 very unlikely, yet still the arithmetic of Bayes' theorem will mean that you change your mind on seeing the evidence and, in the present example, you will come to think that θ_2 is more likely than before.

There are two important exceptions to the proposition that evidence will change your beliefs. Suppose that you had assigned probability zero to the theory represented by θ_1, so $P(\theta_1) = 0$ and hence $P(\theta_2) = 1$, then a glance at the arithmetic above shows that $P(\theta_1|E) = 0$ and therefore $P(\theta_2|E) = 1$. This means that if you gave no credence at all to a theory you will never learn that it is right. The other exception is when $P(E|\theta_1) = P(E|\theta_2)$ so that the data are equally probable on both hypotheses. In this case the data carry no information at all about the merits of the two theories. Again, a glance at the arithmetic of Bayes' theorem shows the truth of this remark.

Another way of expressing the change of beliefs uses **odds**. The odds on an event A are the probability of A divided by the probability of its complement. Thus,

$$\text{odds on } A = \frac{P(A)}{1 - P(A)}.$$

So the odds on θ_2 were 1 before you saw E, it was an "even money bet" in gambling jargon, but the odds on θ_2 after you have seen E have become 6, or six to one on. Thus the evidence has swung your beliefs fairly sharply towards theory θ_2 – you have made an inference from the evidence about the plausibility of one of the two theories.

When used in econometrics the A of Bayes' theorem typically is a statement about a parameter of an economic model and the event B is a statement about some data or evidence that bears on the truth of A. We then think of the movement from the right hand side of (1.2) to the left as occurring sequentially. $P(A)$ is the probability assigned to the truth of A *before* the data have been seen and $P(A|B)$ is its probability *after* the evidence is in. When thought of in this way we call $P(A)$ the **prior**

probability of A and $P(A|B)$ the **posterior probability** of A after the Latin phrases "a priori" and "a posteriori." Bayes' theorem can then be interpreted as showing how to revise beliefs in the light of the evidence – how $P(A)$ is changed by the evidence into $P(A|B)$. Notice in particular that the formula does *not* dictate what your beliefs should be, it only tells you how they should change.[5]

Bayes' theorem for random variables

The more usual form of the theorem is in terms of random variables. Suppose that X, Y are a pair of random variables defined on a sample space Ω and assigned, by you, joint probability density $p_{X,Y}(x, y)$ with marginal densities $p_X(x)$, $p_Y(y)$ and conditional densities $p_{X|Y}(x|y)$ and $p_{Y|X}(y|x)$. Then the theorem is

$$p_{X|Y}(x|y) = \frac{p_{Y|X}(y|x)p_X(x)}{p_Y(y)}.$$

In this notation the subscripts indicate the random variables and the arguments indicate particular values[6] of them. Thus, dropping the subscripts, (1.2) becomes

$$p(x|y) = \frac{p(y|x)p(x)}{p(y)}$$

and when used for inference about parameters given data it is conventional to write the parameters with a Greek symbol so we write

$$p(\theta|y) = \frac{p(y|\theta)p(\theta)}{p(y)}. \tag{1.3}$$

Notice that parameter and data are treated symmetrically before the data have been observed and are assigned a joint probability distribution. Notice also that we are now using the symbol y to denote what we previously called E. This is because economic data is almost always in the form of numerical data and y is a conventional symbol in this case.

[5] Actually there is one constraint on your beliefs: they should satisfy the laws of probability. For example, if A and B are two mutually exclusive and exhaustive events then your beliefs must satisfy $P(A) + P(B) = 1$. Beliefs that satisfy these laws are coherent. If your beliefs are incoherent and you bet according to them then a Dutch book can be made against you. This is a set of bets such that, whatever the outcome, you are sure to lose money. Economists in particular are likely to find such a constraint compelling. See, for example, J. M. Bernardo and A. F. M. Smith, *Bayesian Theory*, Wiley, 1994.

[6] We shall in future drop the subscripts unless they are needed for clarity so that random variables will be interpreted by the arguments of their probability function. We also try to follow the convention of using capital letters to denote random variables and lower case letters to denote their realizations but it is not always sensible to do this.

The econometric model

The numerator on the right hand side of (1.3) is the joint probability distribution of the data that are to be observed and the parameter, $p(y, \theta)$. We shall refer to this joint distribution as (your) **econometric model**. It has two components. The first, $p(y|\theta)$, is called the **likelihood** and it describes what you expect to see for every particular value of the parameter $\theta \in \Theta$. It gives your predictions as to what the data should look like if the parameter takes the particular value given by θ. For our little consumption example if $\theta = (\alpha, \beta) = (10, 0.9)$ then you expect to see that C and Y satisfy the relation $C = 10 + 0.9Y$. Formally this is a distribution that assigns probability one to pairs satisfying this relationship and zero to all other possible C, Y pairs. This, incidentally, is an example of a deterministic likelihood in that once the parameter is set there is no uncertainty about what the data should look like. Economic (as opposed to econometric) models often lead to deterministic likelihoods.

The second component $p(\theta)$ is a probability distribution over the parameter space Θ. It is called the **prior distribution** and it gives your beliefs about the possible values of θ. From the point of view taken in this book an econometric model is complete only when it specifies both the likelihood and the prior. Both are required in order to reach probabilistic conclusions either about θ or about the consistency of the model with the evidence.

Digression ***Objectivity*** *Bayesian inference is not "objective." Some people, believing that science must be objective and its methods objectively justifiable, find this a devastating criticism. Whatever the merit of this position it does not seem to be the way applied econometrics is practiced. The typical seminar in our subject appears to be an exercise in persuasion in which the speaker announces her beliefs in the form of a model containing and accompanied by a set of assumptions, these being additional (tentative) beliefs. She attempts to persuade her audience of the reasonableness of these beliefs by showing that some, at least, embody "rational" behavior by the agents she is discussing and promising that other beliefs will, in fact, be shown by the evidence to be not inconsistent with the data. She then presents her results and shows how some of her beliefs seem to be true and others false and in need of change. The entire process appears to be subjective and personal. All that a Bayesian can contribute to this is to ensure that the way in which she revises her beliefs conforms to the laws of probability and, in particular, uses Bayes' theorem.*

1.3.1 Parameters and data

The material in this section is essential to understanding the point of view taken in this book.

The most useful way to think about the difference between parameters θ and data y is that a parameter is a quantity that is unknown (to you) both before and after the data have been gathered although, of course, your beliefs about it will generally (but not necessarily) have been changed by the evidence; data are unknown before they have been gathered but known afterwards. The word parameter covers several meanings. It may refer to a property of the external world such as the distance from one point on the earth to another or the number of rotten apples in a barrel. Or it may refer to an object appearing in a theory such as "the elasticity of substitution" or "the coefficient of risk aversion." In the latter cases the parameter may well be defined only as a component of a theory and have no existence independent of that theory. And "parameter" does not only mean a constant appearing in a particular economic theory, it may be an index indicating different sub-theories of some larger scheme, as in the introduction to this section where θ_j indicated theory j. And parameters may refer to functions as well as constants as in a setting where it is proposed that $y = g(x)$ where y and x are two economic variables. If $g(.)$ is not given and known (to you), and thus data, this function is a parameter.

Digression ***Randomness*** *In the traditional literature we often find phrases such as "x is random" or "we shall treat x as random" or even "we shall treat x as fixed, i.e. as not random" where "random" means that the object in question will be assigned a probability distribution. In the Bayesian approach all objects appearing in a model are assigned probability distributions and are random in this sense. The only distinction between objects is whether they will become known for sure when the data are in, in which case they are data(!); or whether they will not become known for sure, in which case they are parameters. Generally, the words "random" and "fixed" do not figure in a Bayesian analysis and should be avoided.*[7]

1.3.2 The Bayesian algorithm

We can formulate the Bayesian method as an algorithm.

ALGORITHM 1.1 *BAYES*

1. *Formulate your economic model as a collection of probability distributions conditional on different values for a model parameter $\theta \in \Theta$.*
2. *Organize your beliefs about θ into a (prior) probability distribution over Θ.*
3. *Collect the data and insert them into the family of distributions given in step 1.*
4. *Use Bayes' theorem to calculate your new beliefs about θ.*
5. *Criticize your model.*

[7] Though in chapter 7 we defer to conventional usage and talk, through gritted teeth, about random and fixed effects models.

This book could end at this point, though the publisher might object. All that remains is to offer further explanation and illustration of these steps, not all of which are easy.

1.4 THE COMPONENTS OF BAYES' THEOREM

Let us examine the components of Bayes' theorem as expressed by (1.3), reproduced here for convenience as

$$p(\theta|y) = \frac{p(y|\theta)p(\theta)}{p(y)}, \tag{1.4}$$

using several simple examples. We shall choose examples that are potentially of econometric interest or lie at the heart of models of economic interest. But we shall initially restrict ourselves to cases in which the parameter θ is scalar and not vector valued and we shall consider only situations where each observation is also scalar. This will be rather artificial since in almost all econometric applications the parameter has several, possibly many, dimensions – even in our consumption income example the parameter $\theta = (\alpha, \beta)$ had two dimensions and, as we remarked before, most economic models involve relations between several variables. Moreover the examples use rather simple functional forms and these do not do justice to the full flexibility of modern Bayesian methods. But these restrictions have the great expositional advantage that they avoid computational complexity and enable us to show the workings of Bayes' theorem graphically.

The components of Bayes' theorem are the objects appearing in (1.4). The object on the left, $p(\theta|y)$, is the posterior distribution; the numerator on the right contains the likelihood, $p(y|\theta)$, and the prior $p(\theta)$. The denominator on the right, $p(y)$, is called the marginal distribution of the data or, depending on the context, the predictive distribution of the data. It can be seen that it does not involve θ and so for purposes of inference about θ it can be neglected and Bayes' theorem is often written as

$$p(\theta|y) \propto p(y|\theta)p(\theta) \tag{1.5}$$

where the symbol $\propto$ means "is proportional to." This last relation can be translated into words as "the posterior distribution is proportional to the likelihood times the prior." We shall focus here on the elements of (1.5).

1.4.1 The likelihood $p(y|\theta)$

The expression for the distribution of the data to be observed given the parameter, $p(y|\theta)$,[8] has two names. When thought of as the probability density or mass function of Y evaluated at the point y, conditional on the parameter taking the value θ,

[8] This is $P(B|A)$ in (1.2).

it is called just that, the pdf of Y given θ. But when y is thought of as the actual data that you have gathered, often denoted by the symbol y^{obs} for clarity, it is called the **likelihood function** (of θ). In this case it is often denoted by a different symbol as $\ell(\theta; y)$ and sometimes even more explicitly as $\ell(\theta; y^{obs})$. The likelihood function is not, in general, a probability distribution for θ given data y, nor even proportional to one, though it often is. This is why we separate y and θ by a semicolon and not a conditioning symbol, $|$, when we think of this object as a likelihood function. The likelihood is a function of θ with the data values serving as parameters of that function, hence the semicolon. Many statisticians and some econometricians base their inferences on likelihood functions following the work of the English statistician R. A. Fisher in 1925 and after. People who follow this approach will typically choose as their estimate of θ the value that provides the maximum (strictly the supremum) of the likelihood over Θ. This is called the maximum likelihood (ml) estimator. This tradition is why $p(y|\theta)$ has a special name and symbol when we think of it as a function of θ.

Choice of a likelihood function amounts to choice of a family of probability distributions, one for each $\theta \in \Theta$. The theory of probability offers many such distributions. These range from simple distributions, appropriate to data that can be regarded as conditionally independent realizations from elementary probability distributions with a small number of parameters, to probability models for high dimensional random variables involving many parameters and complex patterns of dependence. Choice among these distributions is an art but this choice must be constrained by a number of considerations. Most importantly, the choice must express the economic model that lies at the center of an econometric investigation. It must allow you to determine from the evidence a probability distribution over the parameter space from which you can calculate which parameter values are probable and which are not. And it must, looking ahead to chapter 2, allow you to conclude that the model itself is wrong or, more precisely, inconsistent with the evidence. But, given this fundamental requirement, many choices remain.

We now give three examples of econometric models and the likelihoods to which they lead. All three involve the dependence of one economic variable upon another. Though simplified here for expository reasons, in more complex and richer forms they lie at the heart of many standard econometric analyses. We preface the examples with a definition.

DEFINITION 1.1 REGRESSION *A regression function is a property of the joint distribution of a pair of random variables. Specifically, it is the expected value in the conditional distribution of one given the other. If the variates are X and Y it is $EX|Y = y$ as a function of y or $EY|X = x$ as a function of x. The term originates with Francis Galton, a nineteenth century English scientist and cousin of Charles Darwin. Galton collected data on heights of parents and their children and calculated the average height of children, $E(Y|X = x)$, of parents of specified height, $X = x$. Plotting these points on a graph he found that the mean height of children increased linearly with their parents' height. He also found that although tall parents tended to have tall children,* *on average they were not as tall as their parents. Children of short parents were also short but tended to be taller than their parents. He called this phenomenon regression (to mediocrity). The term now applies to any conditional mean function, linear or not, and regardless of the numerical character of the relationship. It also applies to any collection of random variables, not just two.*

EXAMPLE 1.2 ***LINEAR REGRESSION*** *Suppose a theorist reasons that one variable, for example consumption, c, should be proportional to another, for example income, y, so that $c = \beta y$, where the theory does not specify the numerical value of β. This deterministic model will be inconsistent with any collection of real economic data on c and y. So let us embed this idea in a less rigid econometric model that states that for any collection of c, y data we shall find that each value of c behaves like a realization of a normal[9] random variable with mean – conditional on y and β – equal to βy. This is called a* ***regression model*** *because the model specifies the regression function of one random variable given another. It is also a linear regression model because βy is linear in y. If the* ***precision***[10] *of these conditional distributions is denoted by the symbol τ and assumed to be the same regardless of the value taken by y, and if distinct c, y pairs are taken to be independent then the joint probability distribution of, say, n realizations of c given their corresponding y's is*

$$\begin{aligned} p(c|y, \beta) &= \Pi_{i=1}^{n}(\tau/2\pi)^{1/2} \exp\{-(\tau/2)(c_i - \beta y_i)^2\} \\ &\propto \exp\{-(\tau/2)\Sigma_{i=1}^{n}(c_i - \beta y_i)^2\}. \end{aligned} \tag{1.6}$$

Each component of the product in the first line is a normal density function of a random variable whose realization is c, whose mean is βy, and whose variance is equal to $1/\tau$. The product arises from the assumption that different realizations are independent, and so their probabilities multiply. In the second line we have collected terms together and then dropped all multiplicative terms that do not involve the (scalar) parameter β. To indicate this we have replaced the = sign by the symbol $\propto$ which, as we have noted, means "is proportional to." The expression that remains when we write a probability density without irrelevant multiplicative terms is called the ***kernel*** *of the distribution.*

In all Bayesian work and throughout this book we shall systematically retain only the kernels of the distributions we work with. Once you get used to it this makes for much easier reading, manipulation, and typing.

[9] The appendix to this chapter gives a brief review of univariate normal distributions.

[10] The precision of a normal distribution is the reciprocal of the variance. It is more convenient in Bayesian work to define a normal distribution by its mean and precision rather than the more customary mean and variance.

When we enter into the expression (1.4) the observed values of our c, y pairs and think of it as a function of β this object becomes the likelihood. (Recall that we are working only with scalar parameters so that we are taking τ as a known number.) If we rewrite the expression $\sum(c_i - \beta y_i)^2$ and again drop multiplicative terms not involving β we find[11] that the likelihood emerges as

$$\begin{aligned} \ell(\beta; c, y) &\propto \exp\{-(\tau \textstyle\sum y_i^2/2)(\beta - b)^2\} \qquad (1.7) \\ \text{for} \quad b &= \textstyle\sum_{i=1}^n c_i y_i / \sum_{i=1}^n y_i^2. \end{aligned}$$

The expression b is the **least squares** estimate of the slope of the regression line, β. Inspection of the likelihood function shows that it has the form of a normal distribution with a mean equal to b and a precision equal to $\tau\sum_{i=1}^n y_i^2$.

A valuable thing to do when learning the theory of econometrics is to study the formulae numerically and graphically. Some people prefer to do this using real economic data but in this book we shall often use simulated data in which you, the reader, specify the numerical values of the parameters and then use a computer to generate data. This artificial data can then be inspected numerically and, in particular, you can see what the likelihood looks like. To see this in action with example 1.2 you can proceed as follows (possible S code is provided in brackets).

ALGORITHM 1.2 *SIMULATING DATA FOR A REGRESSION MODEL*

1. *Choose values for n, β and τ.* `(n <- 50; beta <- 0.9; tau <- 1)`
2. *Select n values for $y = (y_1, y_2, ..., y_n)$.* `(y <- runif(n, 10, 20))`
3. *On your computer generate n independent realizations of normal variates*[12] *with means βy_i and variances equal to $1/\tau$.* `(consump <- rnorm(n, beta*y, 1/sqrt(tau)))`

ALGORITHM 1.3 *PLOTTING THE LIKELIHOOD FUNCTION*

To plot (1.7) you must calculate b; then choose a range of β values over which the plot will be built; then issue the plot command. Possible S code is

1. `b <- sum(consump*y)/sum(y*y)` *# least squares estimate*
2. `betavalues <- seq(0.86, 0.94, length=100)` *# trial and error needed to choose the plot interval*
3. `plot(betavalues, dnorm(betavalues, b, 1/sqrt(tau * sum(y*y))), type="l")` *# dnorm(x, m, s) is the normal density function of mean m, standard deviation s, evaluated at x.*

CALCULATION 1.1 For the following we chose $n = 50$, $\beta = 0.9$, $\tau = 1$ and drew the values of y from a uniform distribution over the interval ten to twenty. Panel 1 of figure 1.1 shows a plot of the data with y on the horizontal axis

[11] The algebraic manipulation involved here will be explained in more detail in chapter 3.

[12] Variate is a shorter version of the phrase "random variable."

and c on the vertical. The straight line is $c = by$ where b is the least squares estimate which turns out to be 0.899. The second panel shows a plot of the likelihood function. As we saw from its mathematical form, it has the shape of a normal curve centered at b and with a standard deviation equal to $1/\surd(\tau\Sigma y_i^2) = 0.0096$. Note that the likelihood is centered close to the value of β, 0.9, that was used to generate the data, and that the curve is effectively zero along the entire real line except for the rather short interval shown in panel 2.

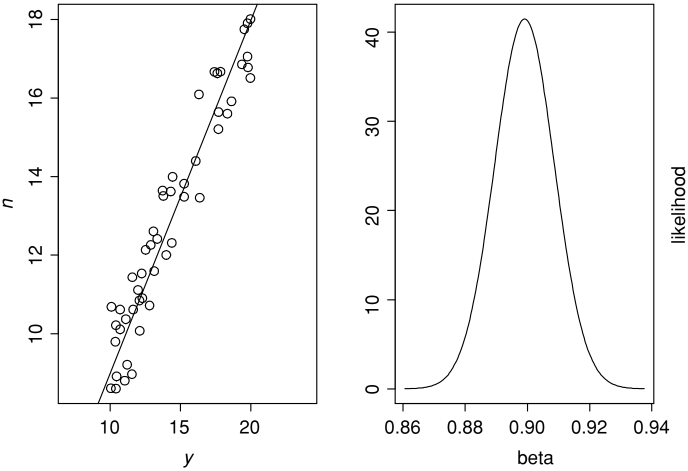

Figure 1.1 Plot of the data and the likelihood for calculation 1.1

For a second example of likelihood we take data that are to be observed in a temporal sequence, i.e. a time series.

EXAMPLE 1.3 ***AN AUTOREGRESSION*** *Suppose that your theory describes the way in which a sequence of values of some economic variable depends on earlier values in the sequence. Let the variable be denoted by y and suppose that you are to observe the sequence at successive time points labeled* 1 *to T. Thus the data are to be the vector* $y = (y_1, y_2, ..., y_T)$. *A simple theory might be that successive values of* y_t *follow the law* $y_t = y_{t-1}$. *As an empirical matter economic data do not follow a (deterministic) law like this just as in the last example data do not follow the deterministic law* $c = \beta y$. *A simple relaxation of this law to allow for some departure from strict equality would be to write* $y_t = y_{t-1} + u_t$ *where the sequence* $\{u_t\}$, $t = 2, 3, ..., T$ *is independently normally distributed random variables with mean zero and precision (reciprocal of the variance)* τ. *This is called a* ***random walk****. One way of setting up a likelihood that enables you to test this theory and, if necessary, to reject it is to embed the model* *in the following framework. First note that the theory asserts nothing about the initial observation, y_1, so it seems appropriate to write our likelihood as a probability distribution for $y_2, y_3, \ldots, y_T$ conditional on the value taken by y_1. Next note that if we enlarge the model, by introducing an additional parameter, to be $y_t = \rho y_{t-1} + u_t$ then the random walk model emerges as the special case in which $\rho = 1$ so we can declare the model inconsistent with the evidence if, after having seen the data, $\rho = 1$ seems to be improbable.*

To complete the argument let us take the parameter of the model as the scalar ρ, taking τ as known for simplicity, and form the likelihood as the joint probability density of $y = (y_2, y_3, \ldots, y_T)$ conditional on y_1 and, of course, the parameter ρ. This may be derived by first considering the joint density function of $u = (u_2, u_3, \ldots, u_T)$ given y_1 and ρ. Since the u's form a sequence of independent random variables we may take this distribution as

$$\begin{aligned} p(u|y_1, \rho) &= \prod_{t=2}^{T}(\tau/2\pi)^{1/2}e^{-(\tau/2)u_t^2} \\ &\propto \exp\{-(\tau/2)\textstyle\sum_{t=2}^{T}u_t^2\}. \end{aligned} \tag{1.8}$$

Now note that $y = (y_2, y_3, \ldots, y_T)$ is a linear function of $u = (u_2, u_3, \ldots, u_T)$ because $y_t = \rho y_{t-1} + u_t$ for $t = 2, 3, \ldots, T$. This function is one to one and its jacobian is readily verified to be unity. Thus the joint density of the data to be observed given the parameter is found by replacing u by y in (1.8) which gives

$$p(y|y_1, \rho) \propto \exp\{-(\tau/2)\textstyle\sum_{t=2}^{T}(y_t - \rho y_{t-1})^2\}.$$

Rearranging the sum of squares in exactly the same way as in example 1.2 and then regarding the whole expression as a function of ρ gives the likelihood kernel as

$$\begin{aligned} \ell(\rho; y, y_1, \tau) &\propto \exp\{-(\tau\textstyle\sum_{t=2}^{T}y_{t-1}^2/2)(\rho - r)^2\} \\ \textit{for} \quad r &= \textstyle\sum_{t=2}^{T}y_t y_{t-1}/\sum_{t=2}^{T}y_{t-1}^2. \end{aligned} \tag{1.9}$$

This likelihood is again of normal shape centered at r, the least squares estimate of ρ and with precision $\tau\sum_{t=2}^{T}y_{t-1}^2$. The reason for this similarity to example 1.2 is that we are again dealing with a regression model though this time the model is not that of one variable against another but of one variable against its own previous value. To see how this works we can simulate some data using an algorithm like:

ALGORITHM 1.4 *SIMULATING AUTOREGRESSIVE DATA*

1. *Choose values for T (which we shall here call n), ρ and τ.* (`n <- 51; rho <- 0.9; tau <- 1`)
2. *Set up an empty vector to hold the values of y.* (`y <- rep(0, n)`)
3. *Select the first value for the time series, y_1.* (`y[1] <- 0`)
4. *Generate, in sequence, the values of $y_2, \ldots, y_T$.* (`for(i in 2:n){y[i] <- rho*y[i-1]+rnorm(1,0,1/sqrt(tau))}`)

Notice that we have simulated data in which $\rho = 0.9$, not 1, so these data will not provide a realization of a random walk.

CALCULATION 1.2 The panels in figure 1.2 show some simulated data and the likelihood to which they lead. The length of the time series was 51, including the initial observation which was zero, and τ was chosen to be one. The first panel[13] shows a plot of the data against time. The second panel shows the likelihood (1.9) where the least squares estimate is $r = 0.747$ and the quantity $\sqrt{(\tau\sum_{t=1}^{T} y_{t-1}^2)}$ was 9.222. The time series graph showing the actual data seems hard to interpret but the likelihood is much clearer.[14] In particular, the second picture shows that the likelihood is essentially zero over the entire real line except for a very narrow band running from about 0.4 to a little over 1. We shall show later that this likelihood graph can be interpreted as giving the relative probabilities of the different values of ρ in the light of the evidence.

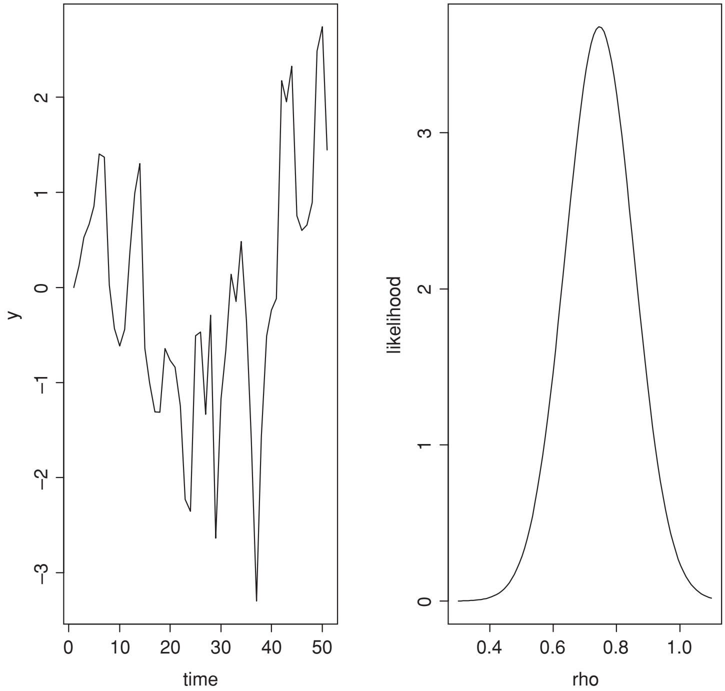

Figure 1.2 Time series data and its likelihood

[13] Plotting the likelihood follows the same pattern as algorithm 1.3.

[14] Some people like to speak of the likelihood graph as providing a window through which some features of the confusing picture on the left can be more clearly seen. But note that many possible windows can be devised.

Both these examples involve a model which began by positing that one random variable was normally distributed given another, and both lead to likelihoods that have the shape of a normal density function – symmetric and bell shaped. But models do not necessarily involve normal distributions nor are likelihood functions invariably shaped like normal distributions. Here is an example of a model whose structure is not normal and whose likelihood function may, or may not, be bell shaped.

EXAMPLE 1.4 ***BINARY CHOICE*** *Many important economic variables are binary. You either do or do not find work; the economy either grows or declines; the couple do or do not marry. A theorist reasons that such a binary outcome, which we shall denote by y, depends on the value of another variate x. We shall take the sample space for y as zero and one, and maybe the theorist thinks, for example, that y = 1, finding work, growing, marrying, is more likely to occur when x is large than when it is small. A variable like y whose sample space has only two points cannot possibly be normally distributed so this model is quite inappropriate here. Instead a binary variate has a distribution that is specified simply by stating the probability that y = 1. So a way of specifying an econometric model to capture the idea that y is more likely to be one when x is large is to write a probability model as $P(Y = 1|x) = p(x)$ and so the probability distribution of Y given x is*

$$p_{Y|X}(y|x) = p(x)^y(1 - p(x))^{1-y}, \quad y \in \{0, 1\}. \tag{1.10}$$

All that remains is to specify the form of the function $p(x)$. Note that the expected value of Y given x is just the probability, given x, that $Y = 1$, hence this is again, like examples 1.2 and 1.3, a regression model. If $p(x)$ is linear in x we have a linear regression model, but if $p(x)$ is non-linear in x we have a **non-linear regression model**.

A common choice in econometrics is to set $p(x) = \Phi(\beta x)$ where $\Phi(.)$ is the standard normal distribution function so Y has a non-linear regression on x. This choice has the advantage that its value always lies between zero and one as a probability must do; if β is positive it captures the theorists' idea that when x is large then y is more likely to be one; and it enables the discrediting of that idea when the parameter β appears in the light of the evidence to be negative or zero. Since $\Phi(\beta x)$ is a non-linear function of x this is a non-linear regression model.

Given this model – a **probit model** – the likelihood, when n observations on y and x can be presumed to be independent, with probabilities that multiply, is

$$\ell(\beta; y, x) = \prod_{i=1}^{n}\Phi(\beta x_i)^{y_i}(1 - \Phi(\beta x_i))^{1-y_i}. \tag{1.11}$$

No manipulations can simplify this expression further, nonetheless the likelihood can readily be drawn. Here is an example. First we show how to simulate binary data.

ALGORITHM 1.5 *SIMULATING BINARY DATA*

```
1. Choose n and β. (n <- 50; beta <- 0)
2. Simulate x values. (x <- runif(n, 10, 20))
3. Simulate y values. (y <- rbinom(n,1,pnorm(beta*x)))[15]
```

CALCULATION 1.3 We choose $n = 50$ and let the x values lie approximately uniformly between 10 and 20. For the first example we choose $\beta = 0$. The resulting simulated data has 28 ones and 22 zeros. The likelihood is plotted[16] in the first panel of figure 1.3. In the second example we choose $\beta = 0.1$ and the resulting simulated data has 48 ones and 2 zeros. The likeli-hood is plotted in the second panel. The first likelihood points to values around 0 and the second to values around 0.1. In the second graph the value $\beta = 0$ has effectively zero likelihood. For both likelihoods the function is essentially zero everywhere else on the real line!

Notice that both likelihoods are still approximately bell shaped although the second, with strongly unequal numbers of ones and zeros in the data, is slightly

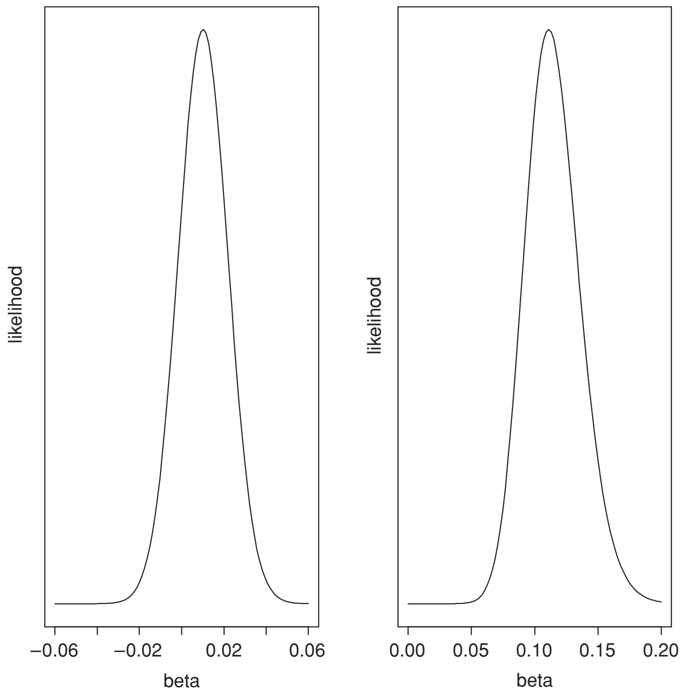

Figure 1.3 Two probit likelihoods

[15] The function `pnorm(x)` provides the value of the standard normal distribution function, $\Phi(x)$, at x.

[16] Plotting is as in algorithm 1.3 except that you would use the S command `pnorm( . . . )`.

asymmetric. This example suggests that even though a model does not involve an assumption of normality likelihoods can nonetheless appear to have the shape of a normal density function. There is a theorem that explains why this is so and we shall describe it later in this chapter.

Before concluding our discussion of likelihood it will be useful to give one final example, one which is mathematically simpler than the first three but which is both fundamental and a convenient vehicle with which to illustrate Bayesian ideas. It does not involve relations between variables and for that reason is of less intrinsic econometric interest, but it is important nonetheless.

EXAMPLE 1.5 ***BERNOULLI TRIALS*** *There is a very simple model for binary data that is valuable for illustrating some theoretical points in a very simple way and this is to consider a sequence of Bernoulli trials. Suppose the variable of interest is binary and takes the values zero or one. We dealt earlier with such a model in which the probability that y is one depended on the value of a covariate x. Now let us look at a simpler setup in which this probability, say* θ*, is the same for all agents. A model in which the data are represented as independent with the same probability of a "success" is called (a sequence of) Bernoulli trials. The random variables* $Y_1, Y_2, \ldots, Y_n$ *are now taken to be independent and identically distributed (iid) conditional on* θ *and n.*

The probability mass function of any element in this collection of random variables is $p(y|\theta) = \theta^y(1-\theta)^{1-y}$, *for* $0 \leq \theta \leq 1$ *and* $y \in \{0, 1\}$. *Because probabilities multiply when random variables are independent the mass function for n such variates is*

$$p(y|\theta, n) = \theta^s(1-\theta)^{n-s}. \tag{1.12}$$

Here, y is now the vector $(y_1, y_2, \ldots, y_n)$ *and* $s = \sum_{i=1}^{n} y_i$ *which is the total number of successes (ones) in the n trials. When the number s is replaced by a particular realization the likelihood is*

$$\ell(\theta; y) = \theta^s(1-\theta)^{n-s}, \quad 0 \leq \theta \leq 1. \tag{1.13}$$

This Bernoulli likelihood has the mathematical form of the kernel of the beta family[17] *of probability density functions (for* θ*). This will turn out to be a useful fact when it comes to drawing and simulating such likelihoods and their generalizations.*

[17] See the appendix to this chapter.

CALCULATION 1.4 ***BERNOULLI TRIAL LIKELIHOODS*** To study (1.13) numerically you can generate some data by choosing n and θ and then using the command `y <- rbinom(n,1,θ)` which will put a sequence of ones and zeros into y. The value of s can then be found as `s <- sum(y)`. The likelihood can be plotted using the fact that $\theta^s(1-\theta)^{n-s}$ is the kernel of a beta density with parameters $s+1$ and $n-s+1$. Thus, choose a sequence of theta values as, say, `thetaval <- seq(0,1,length=100)`, and plot with `plot(thetaval, dbeta(thetaval,s+1,n-s+1,type=″l″)`. Some plots are shown in figure 1.4.

The first row shows the two possible likelihood functions that can arise when only one trial is made. In this case either $s = 1$ or $s = 0$. The likelihood is linear in both cases and not at all bell shaped. The second row examines the case in which $n = 50$ so that s has 51 possible values. We draw the likelihood for two of these. When only one success is recorded the likelihood is concentrated near zero with an (interior) maximum located at $\theta = 0.02$ as can be quickly verified by differentiating $\theta(1-\theta)^{49}$. On the other hand when there are equal numbers of successes and failures the likelihood looks like a normal curve symmetrical about $\theta = 1/2$.

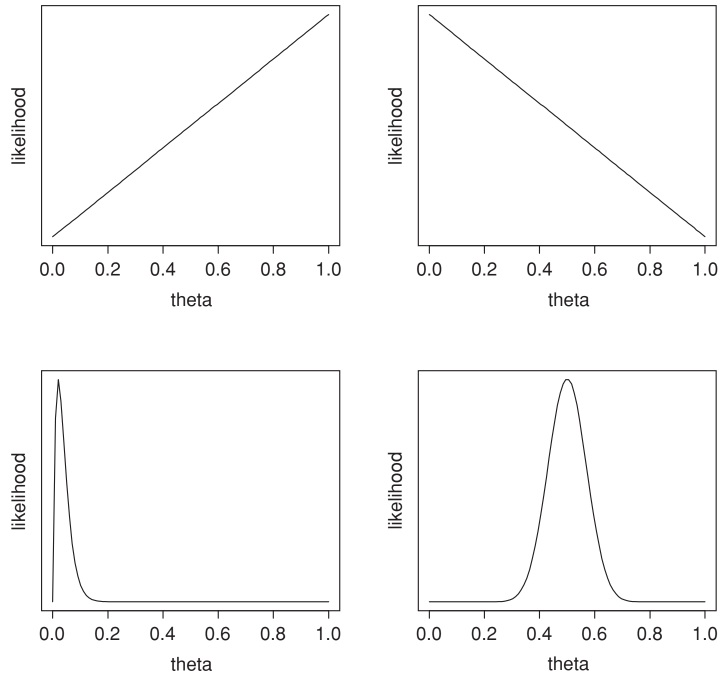

Figure 1.4 Some Bernoulli likelihoods

Parameters of interest

The Bernoulli trials example can be used to make the point that there can be many different parameters of interest, for any of which a likelihood can be constructed. We have taken θ as the parameter of interest but it could have been otherwise. Suppose someone told you that he had carried out n Bernoulli trials with a parameter θ that you and he agree is equal to 0.5 and that he had recorded $s = 7$, say, successes. But he declined to tell you the value of n, so now n is the parameter of interest and θ is data. The probability of s successes in n Bernoulli trials is the binomial expression

$$P(S = s|n, \theta) = \binom{n}{s} \theta^s (1 - \theta)^{n-s} \quad s = 0, 1, 2, ..., n, 0 \leq \theta \leq 1, \tag{1.14}$$

and on inserting the known data $s = 7$, $\theta = 1/2$ we get the likelihood for the parameter n

$$\ell(n; s, \theta) \propto \frac{n!}{(n-7)!}\left(\frac{1}{2}\right)^n \quad n \geq 7.$$

This is drawn in figure 1.5 for $n = 7, 8, ..., 30$.

The parameter here, n, is discrete and the evidence constrains the support of the likelihood – the set of points on which it is positive – to be the integers greater than or equal to 7. After all, if you observed 7 successes you could not possibly have had fewer than 7 trials! The picture clearly shows that only a small set of possible

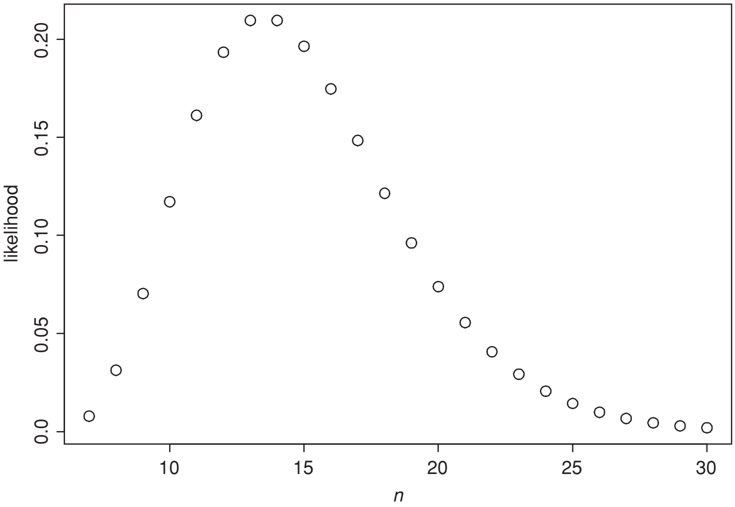

Figure 1.5 Likelihood for n

values of n has much likelihood, and it points to values of n that are close to 14, a number which is equal to $s/\theta = 7/0.5$ which would be most people's guess at the number of trials done to get 7 heads when throwing a fair coin.

The likelihood principle

We remarked earlier that in deducing the posterior we need only consider the kernel of the likelihood (and the same will be true of the prior which we consider in the next section) in order to deduce the posterior distribution of θ. After you have the kernel of the posterior it only takes an integration – $\int p(y|\theta)p(\theta)\,d\theta$ – to find the multiplicative constant that ensures that your posterior density integrates to one.

The fact that you need only the kernel of the posterior to complete your inference has an interesting, and deep, consequence, namely that different likelihoods can lead to the same posterior density and hence to the same inferences. To see this we can again use the Bernoulli trials model with parameter θ to make the point. Consider two investigators with the same prior beliefs about θ but who carry out quite different experiments. The first decides to make $n = 20$ trials and he happens to observe $s = 7$ successes. The second decides to observe Bernoulli trials until seven successes have occurred, and then stop. When he does this he finds that the seventh success occurs on the twentieth trial. The likelihood for the first investigator is the distribution of the number of successes in 20 Bernoulli trials (1.14), as we have seen, and at the observed data this is

$$\ell_1(\theta;\ n = 20,\ s = 7) = \binom{20}{7}\theta^7(1-\theta)^{13} \qquad (1.15)$$

For the second investigator, the probability distribution governing the observations he is about to make is that of the total number of trials, n, necessary to achieve seven successes. This is the negative binomial distribution

$$p(n|s,\theta) = \binom{n-1}{s-1}\theta^s(1-\theta)^{y-s}, \qquad y = s,\ s+1,\ s+2,\ \ldots. \qquad (1.16)$$

and at the observed data this becomes the likelihood

$$\ell_2(\theta;\ n = 20,\ s = 7) = \binom{19}{6}\theta^7(1-\theta)^{13}. \qquad (1.17)$$

Notice that both (1.15) and (1.17) have the same kernel and so, with the same prior, they lead to exactly the same inferences about θ. The likelihoods (1.17) and (1.15) are, as functions of θ, proportional to each other. This is an illustration of the fact that Bayesian inference satisfies the *likelihood principle*.

DEFINITION 1.2 THE LIKELIHOOD PRINCIPLE *This states that likelihoods that are proportional should lead to the same inferences (given the same prior). Notice that the data that might have been observed by the two investigators are quite different. What matters for Bayesian inference are the data that were observed; the data that might have been seen but were not are irrelevant.*[18]

The application of the likelihood principle just described is often referred to as "irrelevance of the stopping rule." That is, it doesn't matter whether you chose in advance to do 20 trials or you chose in advance to do trials until you observed 7 successes. It doesn't, in fact, matter whether you had anything in your head at all before you began doing trials. Maybe you just got bored, or felt ill, after 20 trials. Maybe you can't remember what you planned to do. Most people meeting this implication of the likelihood principle feel shocked, even outraged, and some continue to do so. A common complaint is that stopping trials when you have enough successes sounds like cheating, and that if you find this out your inferences about θ would be different than if you were sure that 20 trials had been decided on in advance. The likelihood principle is a radical idea. Think about it carefully. Given that she was conducting Bernoulli trials, would *your* inference about θ depend on what you knew of the trialer's intentions? Would you throw away as worthless the, possibly unique, even priceless, data if you couldn't find out what she planned to do when she started the trials?

The point of this section has been to define the likelihood principle and to point out that Bayesian inference adheres to it. It has not been to argue for or against the likelihood principle as a fundamental principle of statistical inference. Such arguments can be made and there is a rich literature to consult – pointers to this are given at the end of this chapter. Professional opinion is divided on whether inference should adhere to the likelihood principle.

After these examples we now return to the general issue of the art of likelihood construction to embody and test economic theories.

Populations and samples

There is an important branch of statistics called **survey sampling**. In this field there exists a well defined collection of agents – people or households, for example – and

[18] This is in sharp contrast to standard econometric inference in which the data that might have been observed but were not play a key role through the idea of distributions of estimators in repeated samples. Distributions of statistics over hypothetical repeated samples play no role in Bayesian inference. (See appendix 1.)

Frequentist inferences about θ in this model vary according to whether the data are binomial or negative binomial. That is, they will differ according to the content of the trialer's head when he stopped making his trials. If the frequentist doesn't know this mental state he can make no inferences about θ from the fact that 20 Bernoulli trials produced 7 successes.

each member of this population has a unique vector of characteristics, for example his wage, job, employment status, intention to vote and so on. Call this vector *y*. The object of the calculation is to learn about this collection of vectors, for example the average wage in the population or the fraction of the population without a job. This feature of the population distribution of *y* is *defined* to be the parameter of interest and we learn about it by taking a sample from the population. Often samples are assumed to be random, though there are many other ways of sampling a population. A **random sample** is one in which every time a member of the population is to be selected each member has exactly the same chance as every other of being picked. The data or evidence is provided by the sample and, in frequentist statistics, the parameter is estimated from the sample by choosing a formula, or **estimator**, and applying it to the sample. For example, if the parameter is the average wage in the population an estimator might be "calculate the average wage in the sample." Obeying this instruction with a particular (sample) set of data obtained by selecting individuals from the population then yields an **estimate** of the parameter of interest.

Many applied economists use this story as an aid to thinking about how to embed their economic theory within a probability model, that is, to construct a likelihood. That is, they think of what they are doing as analogous to survey sampling. This point of view holds particular attractions for people whose specialism is microeconomics, since this field typically deals with individual agents and it is often plausible for them to think of their data sets as if they had arisen by some act of randomly sampling a population. Sometimes this story has elements of truth when data really are gathered by sampling a particular collection of individuals, for example the poorly paid or the retired or old. And to think of the object of interest as a characteristic of a real population sounds more practical and concrete than to think of it as a parameter defined within an economic model. In addition it holds particular conviction for those who base their inferences not on Bayes' theorem, as we do in this book, but on imagined sequences of repetitions. The economist can imagine repeatedly drawing a sample from "the population" and then think about the properties of his estimator in such a sequence of repetitions.[19]

In fact economists are rarely, if ever, concerned solely with the properties of some particular, historical population. They wish to generalize and the basis for this generalization is economic theory which points to relationships between variables that are intended to be relatively deep and stable features of the workings of the economy. It is the unspecified constants that appear in this theory that are the ultimate objects of interest.[20]

Nonetheless, thinking about one's data as if they were a sample of some type drawn from a population can be a vivid metaphor and helpful to the applied worker in setting up an econometric model. Moreover, a story about a population and a sample

[19] We shall not, in the body of this book, undertake a criticism of the frequentist approach, but some comments are given in appendix 1 to the book entitled A Conversion Manual.

[20] Economists whose applied work is claimed to rest on "estimating some feature of a population" seem very rarely to define that population precisely.

from it can help the particular model chosen by an investigator seem plausible to his audience and readers, and this is an important consideration in scientific communication which is, in part, an exercise in persuasion. There is no reason why a Bayesian econometrician should not think of his data as a sample from some, possibly hypothetical, population if it assists him in drawing up what appears to be a defensible econometric model. This is as long as he remembers that this is, usually, only a metaphor.

Identification

It is perfectly possible for a likelihood function to point not to one particular value of θ, as in the illustration that we gave earlier, but to be such that all elements of a set of points in Θ give equal values for the likelihood. We saw an example of this at the start of section 1.3 where we remarked that if $P(E|\theta_1) = P(E|\theta_2)$ then no change of opinion about θ would take place. There is nothing problematic about this, though it may be disappointing. You will have to depend on the prior distribution to distinguish among such θ values. And if the prior distribution assigns equal probability to such values they will be equally probable a posteriori.

A deeper phenomenon occurs if such flat spots in the likelihood occur *for any possible data set* that could be observed. Specifically, we can imagine a likelihood derived from the conditional distribution $p(y|\theta)$, $y \in \Omega$, $\theta \in \Theta$ such that for a set of values of $\theta \in \Theta$ we have $p(y|\theta)$ = constant for all $y \in \Omega$. This means that whatever Y realizations occur there is a set of values of θ that all give the same value for the likelihood. In this case we say that θ is not (likelihood) identified. Note that this definition makes no reference to the size of the set of θ values that have equal likelihood. This may be just two points or it may be, for example, the entire real line.

EXAMPLE 1.6 ***NON-IDENTIFIABILITY*** *Consider the likelihood for n independent observations of a normal random variable with mean μ and precision = 1. Multiplying n such normal densities together and dropping irrelevant multiplicative constants gives*

$$\begin{aligned} \ell &\propto \exp\{-(1/2)\textstyle\sum_{i=1}^{n}(y_i - \mu)^2\} \\ &\propto \exp\{-(n/2)(\mu - \bar{y})^2\} \end{aligned}$$

after rearranging the sum in the first line and dropping still more terms not involving μ. Now suppose that your theory leads you to assert that μ is the sum of two theoretically quite distinct effects, one represented by a number α and the other represented by a number β. Thus $\mu = \alpha + \beta$ and we can write the likelihood for α, β as

$$\ell(\alpha, \beta; y) \propto \exp\{-(n/2)(\alpha + \beta - \bar{y})^2\}.$$

For any particular data set providing a value for $\bar{y}$, say 1.415, *we can see from the above expression that all values of α and β whose sum is* 1.415 *yield exactly the same value for the likelihood. But, more importantly, this will be true for any and every data set that you obtain. It will always be true that there is a collection of points in the α, β parameter space that yield equal values for the likelihood. The parameters α and β are not identified.*

Technically, we define[21] identification as

DEFINITION 1.3 IDENTIFICATION *A value θ_a of a parameter is identified if there is no other value θ_b such that $p(y|\theta_a) = p(y|\theta_b)\ \forall\ y \in \Omega$. The model is identified if all the parameter values are identified, in which case the parameter θ is said to be identified.*

If $p(y|\theta_a) = p(y|\theta_b)$ for all y then θ_a and θ_b are said to be observationally equivalent.

Historically, identification has been a major issue in econometrics and the early discovery of potential non-identifiability in even a simple market demand and supply model was a major event in the evolution of our subject. To find that the economic model you have devised and the likelihood to which it leads does not permit the discovery of a single numerical value for a parameter, whatever the data, can be an important insight. The discovery of non-identifiability has prompted the search for credible prior information that can help to distinguish among non-identified values of θ. Traditionally these have either taken the form of exact restrictions on the parameter space – dogmatic priors – or the discovery of further data.

Flat spots at the top of the likelihood pose a problem for maximum likelihood inference since there will never be unique maxima and second derivative matrices will typically be singular at non-identified points. It is of no special significance from the Bayesian point of view because Bayesians do not maximize likelihoods – they combine them with priors and integrate them. A qualification to this is that if all values of a parameter on, say, the real line are unidentified then an (improper) flat prior distribution on that line would lead to a flat posterior and this is not allowed. We shall illustrate non-identifiability in specific contexts later in the book.

Exchangeability

It is almost impossible to construct an econometric model without, at some stage, invoking a proposition of the form "$(Y_1, Y_2, \ldots, Y_n)$ are independently and

[21] Following Bauwens, Lubrano and Richard (1999). This definition refers to parametric identifiability. For a more general definition see Manski (1988).

identically distributed random variables." But this seems to imply that somewhere, out there, is a machine that is similar to the random number generator on your computer and capable of producing a stream of numbers that appear as if they were independent draws from the marginal distributions of any of the $\{Y_i\}$. From the subjective point of view, in which probabilities are private and personal the phrase just cited doesn't look meaningful – it appears to give probability an objective existence.

Because of this some writers prefer, following de Finetti, to derive their likelihoods or probability models via the deeper idea of exchangeability.

DEFINITION 1.4 EXCHANGEABILITY *A sequence of random variables $Y_1, Y_2, ..., Y_n$ is called exchangeable if its joint probability distribution is unchanged by permutation of the subscripts. For example when $n = 3$, $p(Y_1, Y_2, Y_3) = p(Y_2, Y_3, Y_1)$ etc.*

Exchangeability implies that the random variables $\{Y_i\}$ all have the same means and variances, if they exist, and that the correlations between every pair Y_i, Y_j must be the same as for every other pair. Note that exchangeable sequences are not necessarily sequences of independently and identically distributed (iid) random variables, though sequences of iid random variables are exchangeable. Whether you think a sequence is exchangeable is a matter of judgement. Consider, for example, a sequence of, say, 3 tosses of a coin with Y_i denoting the occurrence of heads on the ith throw. You form a judgement about $p(Y_1, Y_2, Y_3)$ and then you are asked to form a judgement about $p(Y_2, Y_3, Y_1)$: would you give a different answer? If you would not and the same was true for all the six possible permutations of the subscripts then your beliefs about Y_1, Y_2, Y_3 are exchangeable.

The relevance of this idea to the question of the choice of prior is a famous result of de Finetti. We give it for binary random variables though more general versions are available. This states that if a sequence of n binary random variables is exchangeable *for every n* then the joint probability distribution of $Y_1, Y_2, ..., Y_n$ *must* take the form

$$p(y_1, y_2, ..., y_n) = \int \theta^s (1 - \theta)^{n-s} dF(\theta).$$

This has the form of a Bayesian marginal data distribution derived from a likelihood equal to $\theta^s(1 - \theta)^{n-s}$ and a prior distribution function equal to $F(\theta)$. So exchangeability implies the existence of a likelihood and a prior. It is an amazingly powerful idea. It means that you have no need to start your modelling with the assertion that a collection of random variables are independent and identically distributed. You can instead merely state that your beliefs about them are exchangeable and this will automatically imply that the model takes the form of a likelihood and a prior.

Having said this, it is the case in almost all practice by Bayesian econometricians that they begin modeling in the conventional way without any deeper justification. We shall mostly follow that path in this book.

Concluding remarks about likelihood

The likelihood (together with the prior which we shall describe next) is a framework within which to confront an economic model with evidence about the economy. Both are probability distributions and in particular the likelihood is, before the data are seen, the joint probability distribution, conditional on a parameter, of all the random variables that will be observed. To construct a likelihood you choose a family of distributions by drawing on the vast collection of such models available within the theory of probability. The likelihood that you choose must be appropriate to the type of data that are to be observed; it must make it possible for you to represent the economic model within it; and it should make it possible for you to discredit that model when it is clearly inconsistent with the evidence.

Your likelihood is not sacrosanct. After all it carries with it restrictions, for example normality, that are not themselves part of the economic model and such restrictions may be inconsistent with the evidence and, if falsely imposed, distort your conclusions about the economic model. In example 1.2 the theorist who proposed that c is proportional to y did not add “and to the extent that it is not, variations about a line through the origin will have a normal distribution.” His theory does not refer to data at all; it exists on a different plane of discourse. This does not imply that you must make no restrictions in constructing your likelihood other than those implied by the theorist. But it does imply that you should explore variations in your inferences over a set of likelihoods each of which embodies the theory. And it also suggests that it is better if your likelihood is relatively unrestricted. Thus, for example, you might want either to assume normality, if this makes sense, and then test whether normality was in fact a restriction consistent with the evidence. Or you might want to begin by assuming not normal variation but some more general distributional family that includes normality as a special case. Both strategies are sensible. The models described in this introductory chapter (of an introductory book) are necessarily very simple and do not represent the full range of probability structures that are available and computionally feasible. In later chapters we shall describe some richer models that are available to the econometrician.

From the subjectivist perspective adopted in this book, a likelihood represents your beliefs about the values of the data conditional on θ. It is *your* likelihood, in the same way that the marginal distribution for θ, the prior $p(\theta)$, will represent *your* beliefs about that parameter. But if your aim is to persuade others of the interest of your results you will be well advised to choose a likelihood that is not clearly inconsistent with the beliefs of your audience and your readers. Assuming that your audience is fellow economists your likelihood should embody both a defensible and coherent economic model and a probability structure that is not obviously inappropriate.

1.4.2 The prior $p(\theta)$

The prior is the other component of Bayes' theorem and together with the likelihood it provides the basis for inference from the evidence. On the subjective view the prior represents *your* beliefs about θ in the form of a probability distribution.[22] You may choose whatever distribution you like in the same way that you can choose whatever likelihood function you like. But a number of points might usefully be made. Some of these points are relatively technical and some are present largely for historical reasons. It may be that a reader willing to accept the simple idea of a prior as a personal probability distribution over the parameter space and anxious to get on with doing Bayesian econometrics would wish to skip over the rest of this section at first and move directly to section 1.4.3 on posterior distributions, or even to chapters 3 and 4 on regression models and markov chain monte carlo methods respectively.

Tentative priors

The first point is that although $p(\theta)$ represents your beliefs you don't need to believe it! You may and indeed should examine the impact of alternative beliefs – alternative priors – on your subsequent, posterior, conclusions. This is done in the spirit of "what if?" You ask "if I had believed this . . . before seeing the data, what would I now believe?" This is called sensitivity analysis and it applies to the likelihood function just as much as to the prior. You may, for example, consider changing the prior from $p(\theta)$ to $q(\theta)$ and you would then recalculate the posterior to study how beliefs about θ have changed. Similarly, you may consider changing the likelihood from $p(y|\theta)$ to $q(y|\theta)$ and seeing how the posterior distribution of θ has changed. The idea here is to explore how sensitive your main conclusions are to alterations in the model, i.e. in the prior and likelihood. We shall illustrate this idea in chapter 2 once we have completed our survey of the main components of Bayesian inference.

In the same spirit, although we usually interpret Bayes' theorem as operating in temporal order, prior beliefs → data → posterior beliefs, this is not a necessary interpretation and it is formally quite legitimate to allow your "prior" beliefs to be influenced by inspection of the data. This is in fact the practice of most applied workers who act in the spirit of Sherlock Holmes' dictum "It is a capital mistake to theorize before one has data."[23] The legitimacy of such data-dependent priors follows from the

[22] It's also possible to take an objective view of a prior distribution over θ values. On this view there exists a population of agents with different values of θ so there is an objectively existing collection of θ values and you can think of $p(\theta)d\theta$ as referring to the proportion of such agents with θ values in the short interval $d\theta$. This gives a relative frequency interpretation to the prior. Some people feel more comfortable with this interpretation of the prior and there is nothing in what follows to preclude this point of view. It is entirely consistent with the mathematics of Bayesian inference and the reader who prefers such an **objective Bayesian** perspective can use all the techniques described in this book to carry out his econometric analysis.

[23] *A Scandal in Bohemia* by Arthur Conan Doyle.

fact that Bayes' theorem does not restrict the choice of prior, it only prescribes how beliefs change.

Encompassing priors

The second point is that it is necessary to take account of the beliefs of your audience and your readers, if any. Prior beliefs[24] that conflict sharply with those of your readers will make your work of little interest to them. You will be saying "If you believed A before seeing the data you should now believe B." But this will be met with the response "So what, I don't believe A." It is therefore a good idea, for public scientific work, to use priors that are not sharply or dogmatically inconsistent with any reasonable belief. In low dimensions this requirement can sometimes be met by using a uniform or flat distribution on some reasonable function of the parameter. In the Bernoulli trials example a uniform distribution for θ on the interval zero to one will not be inconsistent with any belief. It will not *represent* any belief, for example it would not represent the belief of someone who is quite convinced that θ lies between 0.4 and 0.6, but it wouldn't be inconsistent with such a belief. In a sense such a prior encompasses all reasonable beliefs. Such priors are often called **vague**.[25]

As a particular case of this it would be wise to avoid using priors that assign zero probability to parts of the parameter space. Because the posterior density is formed by multiplication of the prior and likelihood – see (1.3) – a prior that assigns probability zero to a set will necessarily assign zero posterior probability to that set. Such a prior is very dogmatic and this is to be avoided wherever possible in scientific enquiry. On the other hand, any model involves some dogmatic assertions since without them the theory would be vacuous. So the recommendation to avoid dogmatic priors can never be strictly fulfilled.

Natural conjugate priors

The posterior density function $p(\theta|y)$ is formed, apart from the multiplicative constant $1/p(y)$, by multiplying the likelihood and the prior. There is some merit in choosing a prior from a family of density functions that, after multiplication by the likelihood, produce a posterior distribution in the same family. Such a prior is called *natural conjugate*. In this case only the parameters of the prior change with the accumulation of data, not its mathematical form. Such priors also have the advantage that they can be interpreted as posterior distributions arising from some earlier, possibly fictional, evidence. Thus we might try to form our prior for θ in the

[24] And likelihoods, for that matter.

[25] There is a history of efforts to find priors that are "uninformative" in some sense, compared to the likelihood. These efforts do not seem to have been very fruitful particularly in the case of models with parameters of several dimensions.

Bernoulli trials example by trying to imagine what our beliefs would currently be had we seen the evidence of some earlier trials prior to which our beliefs were vague.

Before illustrating natural conjugacy it will be helpful to reintroduce the idea of a kernel.

DEFINITION 1.5 A KERNEL *A probability density or mass function of a random variable X typically has the form $kg(x)$ where k is a numerical constant whose role is to ensure that $kg(x)$ integrates to one. The remaining portion, $g(x)$, which does involve x, is called the* ***kernel*** *of the function.*

For the beta family of probability density functions the kernel is $x^{a-1}(1-x)^{b-1}$ while k is the ratio of gamma functions given in the appendix to this chapter. What constitutes the kernel of a density or mass function depends on what you think the argument is. For example if x is of interest the kernel of an $n(\mu, \tau)$ density function is $\exp\{-\tau(x-\mu)^2/2\}$ while the constant is $\tau^{1/2}/\surd(2\pi)$. On the other hand if one is thinking about the normal density for given x as a function of μ and τ then the kernel would be $\tau^{1/2}\exp\{-\tau(x-\mu)^2/2\}$. In neither case is the numerical factor $1/\surd(2\pi)$ of any relevance.

The purpose of k is to make the density or mass function integrate to one. Once you know the kernel the constant can be found by integration but it is usually of little interest in itself. Since a family of distributions can be recognized from its kernel it is usually convenient to omit constants when we manipulate probability distributions and we shall follow this convention in this book. It makes for algebra that is much easier to follow.

Indeed, as we remarked earlier, Bayes' theorem itself is often stated up to a missing constant as

$$p(\theta|y) \propto p(y|\theta)p(\theta) \tag{1.18}$$

or, in words, the posterior is proportional to the product of the likelihood and the prior.

EXAMPLE 1.7 *NATURAL CONJUGACY FOR THE BERNOULLI TRIALS PARAMETER* *To illustrate natural conjugacy consider the likelihood of θ in the Bernoulli trials example which has, up to a multiplicative constant, the general form $\theta^s(1-\theta)^{n-s}$. Since the posterior density of θ is formed by multiplying the prior and the likelihood it is clear, by contemplating the multiplication of such functions, that any prior that is proportional to $\theta^{a-1}(1-\theta)^{b-1}$ will lead to a posterior density of the same mathematical form. It follows that the natural conjugate family of prior distributions for this problem is the beta family.*

Notice that this argument never needed to mention the constants multiplying these kernels. The property of natural conjugacy was of more importance in the days when posterior distributions were computed analytically and not, as now, numerically.

Improper priors

A "probability distribution" for θ is called improper if its integral over the sample space Θ does not converge. A simple example is the expression

$$p(\theta) \propto 1, \quad -\infty < \theta < \infty \tag{1.19}$$

which is called a uniform distribution on the real line and can be thought of as a rectangle on an infinitely long base. Its integral, the area under the line, does not converge, it is infinite and so (1.19) is not, in fact, a probability distribution. Nevertheless such improper distributions are frequently used in applied Bayesian inference and there are several reasons for this.

One reason is that often it does not matter, at least mathematically, if the prior is improper. Because the object of ultimate interest is the posterior distribution of θ and this is formed by multiplying the likelihood and the prior it is perfectly possible for the posterior distribution to be proper even though the prior is not. To see this consider the following.

EXAMPLE 1.8 *PROPER POSTERIOR FROM IMPROPER PRIOR* *Let the likelihood be formed as the distribution of n independent normal variates with mean θ and precision one. Thus*

$$\begin{aligned}\ell(\theta; y) &\propto \textstyle\prod_{i=1}^{n}\exp\{-(1/2)(y_i-\theta)^2\}\\ &= \exp\{-(1/2)\textstyle\sum_{i=1}^{n}(y_i-\theta)^2\},\end{aligned} \tag{1.20}$$

and using the fact that

$$\begin{aligned}\textstyle\sum_{i=1}^{n}(y_i-\theta)^2 &= \textstyle\sum_{i=1}^{n}(y_i-\bar{y}+\bar{y}-\theta)^2\\ &= \textstyle\sum_{i=1}^{n}(y_i-\bar{y})^2 + \textstyle\sum_{i=1}^{n}(\theta-\bar{y})^2,\end{aligned}$$

we find that

$$\ell(\theta; y) \propto \exp\{-(n/2)(\theta-\bar{y})^2\}. \tag{1.21}$$

This is (the kernel of) a normal distribution with mean $\bar{y}$ and precision n. It is a perfectly proper probability density function whatever the values of $n > 0$ and $\bar{y}$. So if you multiply the likelihood (1.21) by the improper prior (1.19) the resulting posterior distribution is proper.

Thus, at least mathematically and for at least some models,[26] it is unnecessary for a prior to be proper. Improper priors can lead to proper posteriors.

Another reason is that an improper prior can often be thought of as an approximation to a proper prior that is intended to represent very imprecise or vague beliefs. To see this consider the last example again.

EXAMPLE 1.9 *Suppose we multiply the likelihood (1.21) by a prior density for θ that is normal with precision equal to τ and mean zero. This gives a posterior density*

$$\begin{aligned} p(\theta|y) &\propto \exp\{-(n/2)(\theta-\bar{y})^2\}\exp\{-(\tau/2)\theta^2\} \\ &\propto \exp\{-(n+\tau)(\theta-\bar{\theta})^2/2\} \end{aligned} \tag{1.22}$$

after a little rearrangement of the exponent and the dropping of irrelevant multiplicative constants. This expression, (1.22), is the kernel of a normal distribution with mean $\bar{\theta}=n\bar{y}/(n+\tau)$ and precision $n+\tau$. Now let the positive number τ approach zero. The prior density $e^{-\tau\theta^2/2}$ approaches a constant, the posterior mean $n\bar{y}/(n+\tau)$ approaches $\bar{y}$ and the posterior precision approaches n and these are the values that correspond to the improper uniform prior underlying (1.21). A proper prior with τ sufficiently small will produce much the same posterior as an improper, uniform, prior.

It follows from this example that when your prior beliefs are very vague you can (sometimes) act as if your prior was uniform and find a numerically very accurate approximation to what your real posterior beliefs would be. The uniform prior is a labor saving device in that it saves you the trouble of specifying your exact beliefs. It's often used in this way in practice, during preliminary analyses. Many of the standard calculations in econometrics, for example the use of least squares regression and all maximum likelihood methods, can be thought of as flat prior Bayes, as we shall see.

A third reason goes by the name of the **principle of precise measurement**.

Precise measurement

Recall from Bayes' theorem (1.3) that the posterior distribution, which provides our inferences about θ, is formed as the *product* of the likelihood and the prior. This simple fact is of enormous consequence. We have already remarked that because of it you should never assign zero prior probability to a set in Θ since, because zero times any number always gives zero, this action necessarily assigns zero posterior

[26] In other models improper priors can lead to improper posteriors, as we shall see.

probability to that set so you can never learn that in fact, in the light of the evidence, that set is quite probable. We now use this fact again by remarking that almost always – the first two panels in figure 1.4 are an exception – the likelihood is effectively zero over most of the parameter space. To see this look at examples 1.2, 1.3, and 1.4 where the likelihood is negligible everywhere on the real line except in the region we have plotted. Thus the prior is multiplied by (almost) zero almost everywhere in the parameter space and it does not matter what your prior beliefs were in that region. Whatever they were they will not change the posterior density in regions where the likelihood is negligible. This implies that a prior that is intended to be roughly neutral as between different values of θ need only be so in the region where the likelihood is non-negligible – how the prior behaves outside that region is of no consequence. A conclusion that could be drawn from these remarks is that a prior that is, formally, uniform on the real line is practically equivalent to one which is uniform where the likelihood is non-negligible but behaves in any other (bounded) way outside that region.

We now turn to look at the possibilities of finding objective and default priors.

Objective and default priors

Much ink has been spilt in the search for a rule that would produce a prior distribution for any model and one that would, in some sense, be minimally informative. The search could be said to begin with the Marquis de Laplace in the eighteenth century but in its modern form it could be said to begin with Harold Jeffreys in 1938 and it still continues. In his book Jeffreys proposed a rule that possesses an apparently persuasive property, that of invariance.

JEFFREYS' INVARIANT PRIORS

We can parametrize a model in an infinite number of ways and the parametrization we choose is important in Bayesian inference. For example we can parametrize a zero mean normal distribution in terms of its standard deviation σ, its variance σ^2, its precision $\tau = 1/\sigma^2$ and generally we can use any one-to-one function of σ. Suppose that we choose any particular parametrization, for example σ, and apply a rule for constructing a prior distribution for that parameter and then, using the prior that results from following that rule we construct the posterior distribution of σ. Now suppose that you re-analyze the data but work in terms of a different parametrization, say σ^2, but you apply *the same rule* to form your prior for σ^2. Jeffreys then argued that the beliefs about the first parameter σ that can be deduced from the posterior distribution for σ^2 should be identical to those reached in the first analysis; in Jeffreys' words "equivalent propositions should have the same probability." Posterior beliefs about the same quantity should be *invariant* to the parametrization used. Jeffreys showed that there exists a rule, now named after him, that does satisfy this invariance condition.

His rule is to choose the prior proportional to the square root of the *information*,

$$I_\theta = -E\left(\frac{\partial^2 \log \ell(\theta;y)}{2\theta^2}\right) \tag{1.23}$$

where the expectation is taken with respect to $p(y|\theta)$. This is (the negative) second derivative of the logarithm of the likelihood function averaged over repeated realizations of y. It, and its matrix version, plays a major role in both likelihood and Bayesian inference. Here is the argument that shows that Jeffreys' rule is invariant to reparametrization. Suppose a second parametrization is in terms of $h(\theta)$, for example θ might be σ and $\gamma = h(\theta)$ might be $1/\sigma^2$. Now note that

$$\frac{\partial \log \ell}{\partial \gamma} = \frac{\partial \log \ell}{\partial \theta}\frac{\partial \theta}{\partial \gamma},$$

$$\frac{\partial^2 \log \ell}{\partial \gamma^2} = \frac{\partial^2 \log \ell}{\partial \theta^2}\left(\frac{\partial \theta}{\partial \gamma}\right)^2 + \frac{\partial \log \ell}{\partial \theta}\frac{\partial^2 \theta}{\partial \gamma^2},$$

$$\text{so} \quad I_\gamma = I_\theta \left(\frac{\partial \theta}{\partial \gamma}\right)^2 \tag{1.24}$$

where the last line follows because $E(\partial \log \ell/\partial\theta) = 0$.[27] Here, I_γ is the information about γ and I_θ is the information about θ. Note that (1.24) implies that $I_\gamma^{1/2} = I_\theta^{1/2} |\partial\theta/\partial\gamma|$. Now the posterior distribution of θ for someone who works in terms of θ and follows Jeffreys' rule will be $\ell(\theta)I_\theta^{1/2}$. For someone who works in terms of γ, his posterior distribution for γ will be $\ell(h(\theta))I_\gamma^{1/2}$. From this we can deduce what the second person's beliefs about θ will be by following the standard rule for deducing the distribution of a function of a random variable. This gives the second person's beliefs about θ as $\ell(\theta)I_\gamma^{1/2}|\partial\gamma/\partial\theta| = \ell(\theta)I_\theta^{1/2}|\partial\theta/\partial\gamma||\partial\gamma/\partial\theta| = \ell(\theta)I_\theta^{1/2}$ which are precisely the same as the first person's beliefs about θ. To illustrate this potentially confusing little argument consider the following.

EXAMPLE 1.10 *JEFFREYS' PRIOR FOR A NORMAL PRECISION* *If $p(y|\sigma)$ is the density of n independent normal variates of mean zero and standard deviation σ then* $\log \ell(\sigma) = -n \log \sigma - \sum y_i^2/2\sigma^2$. *So the hessian is* $\partial^2 \log \ell(\sigma)/\partial\sigma^2 = (n/\sigma^2) - 3\sum y_i^2/\sigma^4$. *Since the expected value of y^2 is σ^2 the information about σ, I_σ, is $2n/\sigma^2$ and Jeffreys' prior for σ will be $\propto 1/\sigma$. An alternative parametrization is $\tau = 1/\sigma^2$ and the log likelihood in terms of τ is* $\log \ell(\tau) = (n/2)\log\tau - \tau\sum y_i^2/2$. *Differentiating twice, taking expectations and changing sign then gives the information*

[27] Take the identity $\int p(y|\theta)dy = 1$; differentiate with respect to θ; then rearrange using $\partial \log p/\partial\theta = (1/p)\partial p/\partial\theta$.

for τ as $I_\tau = n/2\tau^2$ implying the Jeffreys' prior for τ is $\propto 1/\tau$. The posterior beliefs about σ for the person working in terms of σ will be

$$p(\sigma|y) \propto \sigma^{-(n+1)}\exp\{-\Sigma y_i^2/2\sigma^2\}. \tag{1.25}$$

The beliefs about τ for the person working in terms of τ will be

$$p(\tau|y) \propto \tau^{n/2-1}\exp\{-\tau\Sigma y_i^2/2\}.$$

Finally, the beliefs about σ held by the latter person are found by the change of variable from τ to σ and are

$$p(\sigma|y) \propto \sigma^{-n+2}\exp\{-\Sigma y_i^2/2\sigma^2\}\left|-2/\sigma^3\right| \propto \sigma^{-(n+1)}\exp\{-\Sigma y_i^2/2\sigma^2\}.$$

This is identical to (1.25) which confirms invariance in this case.

Prior beliefs formed using Jeffreys' rule are often improper as the preceding example illustrates – $1/\tau$ is an improper prior over $0 < \tau < \infty$ since $\int_0^\infty \tau^{-1}d\tau$ diverges. The invariance argument generalizes straightforwardly to the case in which θ is a vector parameter. Informations are replaced by information matrices and Jeffreys' takes the form $|I_\theta|^{1/2}$ – the square root of the determinant of the information matrix.

EXAMPLE 1.11 *JEFFREYS' PRIOR FOR BERNOULLI TRIALS*

With n Bernoulli trials the likelihood for θ is $\ell(\theta; y) \propto \theta^s(1-\theta)^{n-s}$. To calculate Jeffreys' prior we need to differentiate the log likelihood twice and take expectations. The calculation, with L denoting $\log \ell$, is as follows.

$$L(\theta) = s\log\theta + (n-s)\log(1-\theta)$$

$$\frac{\partial L}{\partial\theta} = \frac{s}{\theta} - \frac{n-s}{1-\theta},$$

$$\frac{\partial^2 L}{\partial\theta^2} = \frac{-s}{\theta^2} - \frac{n-s}{(1-\theta)^2},$$

$$\textit{since } E(s|\theta, n) = n\theta, \quad I_\theta = \frac{n}{\theta} + \frac{n}{1-\theta} = \frac{n}{\theta(1-\theta)}.$$

It follows that Jeffreys' prior is

$$p(\theta) \propto \frac{1}{\sqrt{\theta(1-\theta)}}.$$

This is a beta(1/2, 1/2) density which is proper, but U shaped. According to this prior the least likely value of θ is 1/2. Notice that Jeffreys' prior is not uniform, as one might, perhaps, have anticipated.

There remains much debate about the value of Jeffreys' rule. It often doesn't seem to give very appealing results particularly when θ is vector valued but even in a simple model such as an autoregression a Jeffreys' prior on the autoregressive coefficient may seem strange to many. Objective rules in general are not very appealing to those who prefer beliefs to represent subjective opinion based on an informed appreciation of the economic meaning of θ. Jeffreys' prior involves taking expectations with respect to y which is a repeated sampling calculation, and many writers take the view that such calculations are, in general, not well defined and they certainly violate the likelihood principle.

It's also not very clear in what sense Jeffreys' rule produces prior distributions that are minimally informative. There is another strand in the literature, due to Bernardo and Berger,[28] which starts with a precise measure of the amount of information in a probability distribution based on information theory, and asks for the prior distribution whose contribution to the total information in the posterior distribution is minimal. This leads to the class of **reference priors**. Unfortunately these do not always exist, even for econometrically simple models, but where they do exist they typically take Jeffreys' form. So in this sense Jeffreys' priors can be justified as minimally informative.

But even if one doesn't like general rules for forming prior distributions there exists a need for **default priors** to use in standard situations when an investigator, at least initially, doesn't wish to spend much time thinking about the details of his prior beliefs about θ. So just as there are default likelihoods for automatic use in standard models there are default priors in general use. These are typically uniform (and therefore improper) distributions of functions of the parameter concerned. For example linear regression coefficients are usually taken to be uniform on the real line and normal precisions to be such that the log precision is uniform on the real line, so that the precision itself has "density" $1/\tau$ on the positive axis.[29] We shall use such default priors quite often in this book. Application of such default uniform distributions to high dimensional parameters must however be done with great caution[30] and we shall see several illustrations of this caution in the chapters on panel data and on time series.

Hierarchical priors

When dealing with vector valued parameters it is often persuasive to think about your prior distribution hierarchically. Suppose you are dealing with a parameter with, say, n elements and these elements are similar in the sense that they have the same dimension (units of measurement) and play similar roles in the model. An example might be the coefficient of the same variable in a regression model where each agent is allowed to have his own response. Another example might be the set of precisions

[28] See the bibliographic notes at the end of this chapter.

[29] Using a change of variable with jacobian $\partial \log \tau/\partial\tau = 1/\tau$.

[30] BUGS, the software package recommended for this book, *requires* the user to employ proper priors.

in a model where each agent is allowed to have his own precision. If the parameter is $\theta = (\theta_1, \theta_2, ..., \theta_n)$ one might construct a prior that expresses the similarity among the elements of θ, by taking these elements to be an independent set of realizations of some appropriate parent distribution, say $h(\theta|\lambda)$, where the parameter of the parent, λ, is of much smaller dimension than θ. This parent or second stage parameter is then assigned a prior distribution, typically one of the default choices. Formally, we want $p(\theta)$ and we get this by stating firstly $p(\theta|\lambda)$ and then forming $p(\lambda)$. This forms $p(\theta)$ implicitly as $p(\theta) = \int p(\theta|\lambda)p(\lambda)d\lambda$. The parameters θ represent the first stage in the hierarchical structure; the parameters λ represent the second stage, and so on. There is no limit to the number of stages in a hierarchical model though in practice two or three is usual. Here is an example.

EXAMPLE 1.12 ***A HIERARCHICAL PRIOR*** *Let $\tau = (\tau_1, \tau_2, ..., \tau_n)$ be a set of precisions that are thought to be similar, but not identical, and have the same dimension. Since they are non-negative an obvious choice for a hierarchy is to let them be realizations of a gamma variate with parameter $\lambda = (\alpha, \beta)$. Thus (see the appendix to this chapter)*

$$p(\tau|\lambda) \propto \prod_{i=1}^n \tau_i^{\alpha-1} e^{-\beta\tau_i}. \tag{1.26}$$

As an application, consider a collection of independent normal variates of mean zero and precisions τ_i. Then the likelihood is

$$\ell(y;\ \tau, \lambda) \propto \prod_{i=1}^n \tau_i^{1/2} \exp\{-y_i^2\tau_i/2\}, \tag{1.27}$$

and the whole model is

$$\begin{aligned} p(\tau, \lambda|y) &= \ell(y;\ \tau, \lambda)p(\tau|\lambda)p(\lambda) \\ &= \prod_{i=1}^n \tau_i^{1/2} \exp\{-y_i^2\tau_i/2\} \prod_{i=1}^n \tau_i^{\alpha-1} e^{-\beta\tau_i} p(\lambda), \\ &= \prod_{i=1}^n \tau_i^{\alpha+1/2-1} \exp\{-\tau_i(\beta + y_i^2/2)\} p(\lambda). \end{aligned} \tag{1.28}$$

This prior structure is often used as the basis *for* robust *Bayesian analysis in the sense that it relaxes the somewhat dogmatic restriction that all y's have the same precision. We shall return to this model in later chapters.*

An interesting feature of hierarchical priors is that they reveal the somewhat arbitrary nature of the distinction between the likelihood and the prior. Take a model written with parameter θ; let θ have a prior that depends upon a hyperparameter ψ; and let ψ have a prior $p(\psi)$ involving no unknown parameters. Then one way of presenting the model is as

$$\ell(\theta;\ y)p(\theta|\psi)p(\psi)$$

where the prior is $p(\theta|\psi)p(\psi) = p(\theta, \psi)$. Another way is to integrate out θ and write the model as

$$\ell(\psi; y)p(\psi), \quad \text{where } \ell(\psi; y) = \int \ell(\theta; y)p(\theta|\psi)\,d\theta.$$

Which is the likelihood, $\ell(\psi; y)$ or $\ell(\theta; y)$? The answer is that it doesn't really matter; all that does matter is the product of prior and likelihood which can be taken as $p(y, \theta, \psi)$ or as the y, ψ marginal, $p(y, \psi) = \int p(y, \theta, \psi)\,d\theta$.

Priors for multidimensional parameters

While it doesn't matter in principle to the Bayesian method whether θ is scalar or multidimensional it matters quite a lot in practice. Default choice of prior for scalar parameters has been rather thoroughly studied. For example, many reference priors (mentioned above) for scalar parameters have been produced, but the situation for vector parameters is notably less clear or, indeed, simple. Jeffreys' priors for scalars are known to provide acceptable results in most cases but, as Jeffreys himself observed, the position is less satisfactory when I_θ is matrix valued.

One promising line of work has been to try to reduce the situation to one involving many scalar parameters. This can be done if you can separate the likelihood into a product form, each term of which involves only a single element of θ. Then if you can reasonably assume independence of the elements of θ in the prior, the posterior distribution will also factor and you have, at least in a numerical sense, k separate analyses. Separating the likelihood in this way typically will involve finding a different parametrization of the model from the one in which you originally wrote it. That is, working in terms of some one-to-one function $g(\theta)$ instead of θ.

EXAMPLE 1.13 *PARAMETER SEPARATION IN REGRESSION* *As a fairly simple example of parameter separation consider a version of example 1.2 in which there are two parameters, α and β, so that $\theta = (\alpha, \beta)$. Let the relation between consumption and income be*

$$c_i = \alpha + \beta y_i + \varepsilon_i, \quad \varepsilon_i \sim n(0, 1), \tag{1.29}$$

for $i = 1, 2, \ldots, n$ with observations independent given the parameters and the y's. By the argument leading to (1.6) the likelihood is

$$\ell(\alpha, \beta) \propto \exp\{-(1/2)\Sigma_{i=1}^{n}(c_i - \alpha - \beta y_i)^2\}$$

and this, after a little tedious algebra, can be written as a generalization of (1.7)

$$\ell(\alpha, \beta) \propto \exp\{-(1/2)(\theta - \hat{\theta})'X'X(\theta - \hat{\theta})\} \tag{1.30}$$

$$\textit{where} \quad X'X = \begin{bmatrix} n & \Sigma y_i \\ \Sigma y_i & \Sigma y_i^2 \end{bmatrix}, \tag{1.31}$$

$$\textit{and} \quad \theta = \begin{bmatrix} n & \Sigma y_i \\ \Sigma y_i & \Sigma y_i^2 \end{bmatrix}^{-1} \begin{bmatrix} \Sigma c_i \\ \Sigma c_i y_i \end{bmatrix}.$$

Inspection of (1.30) shows that it will not *break apart into a product of a term involving only α and a term involving only β unless $X'X$ is a diagonal matrix and this requires that $\sum_{i=1}^{n} y_i$ is zero, which will not usually be true.* But we can make it true.

To see this let us rewrite the model as

$$\begin{aligned} c_i &= (\alpha + \beta\bar{y}) + \beta(y_i - \bar{y}) + \varepsilon_i \\ &= \alpha^* + \beta y_i^* + \varepsilon_i \end{aligned}$$

with the same distributional assumptions. This is exactly the same model as (1.29) but with a different parametrization: instead of $\theta = (\alpha, \beta)$ it is now parametrized in terms of $g(\theta) = (\alpha^*, \beta)$, a one-to-one function of θ. Now the new $X'X$ matrix has the form

$$\begin{bmatrix} n & 0 \\ 0 & \Sigma(y_i - \bar{y})^2 \end{bmatrix} \tag{1.32}$$

where the zeros appear because the sum of observations measured from their mean is identically zero. It follows from this diagonality that the likelihood in terms of $g(\theta)$ takes the form

$$\ell(g(\theta); y) \propto e^{-(n/2)(\alpha^* - \widehat{\alpha^*})^2} e^{-(\Sigma y_i^{*2}/2)(\beta - \hat{\beta})^2}$$
$$\textit{where} \quad \widehat{\alpha^*} = \bar{c}, \textit{ and } \hat{\beta} = \Sigma(c_i - \bar{c})(y_i - \bar{y})/\Sigma(y_i - \bar{y})^2.$$

So the first component of the reparametrized likelihood has the shape of a normal curve centered at mean consumption, and the second component has the shape of a normal curve centered at $\hat{\beta}$, the least squares estimate.

One feature of this example that is particularly important is the effect of the parameter transformation on the information matrix. It's obvious that if a likelihood is multiplicatively separable then the log likelihood is additively separable, and it follows from this that the cross partial second derivatives of the log likelihood will be identically zero. For this model the information matrix for θ is given by (1.31) but the information matrix for $g(\theta)$ is given by (1.32), which is diagonal.

This remark suggests that we can search for separable reparametrization by looking for functions $g(\theta)$ that diagonalize the information matrix. In later chapters we shall show that such new parametrizations, called **information orthogonal**, can often be found and that they tend to simplify the search for default priors in models with multidimensional parameters.

1.4.3 The posterior $p(\theta|y)$

The posterior density represents your beliefs about θ given your prior beliefs and the beliefs embodied in the likelihood. In many applications the posterior is the culmination of an empirical analysis.[31] To report your results you will display the posterior distributions to which your model and data have led. Let us look at examples of posterior distributions before making some general comments.

EXAMPLE 1.14 *BERNOULLI TRIALS* *Suppose that your prior beliefs are described by a member of the (natural conjugate) beta family. Formally,* $p(\theta) \propto \theta^{a-1}(1-\theta)^{b-1}$, $0 \leq \theta \leq 1$. *With a model in which n Bernoulli trials are undertaken, with outcomes which are conditionally independent, with common expectation θ, the likelihood was given by (1.13). Hence, by Bayes' theorem the posterior density of θ has the form*

$$p(\theta|y) \propto \theta^{s+a-1}(1-\theta)^{n-s+b-1} \tag{1.33}$$

which we recognize as the kernel of a beta density with mean and variance

$$E(\theta|y) = \frac{s+a}{n+a+b}, \quad V(\theta|y) = \frac{(s+a)(n-s+b)}{(n+a+b)^2(n+a+b+1)}. \tag{1.34}$$

(Note that if both s and n are large and in the ratio r then these moments are approximately

$$E(\theta|y) = r, \quad V(\theta|y) = \frac{r(1-r)}{n}.$$

When n is large and r = s/n is fixed, the posterior variance becomes small and almost all its probability mass is confined near s/n, the fraction of successes. It is easy to see that this is true whatever member of the beta family was used as the prior and we conclude that for this model, as evidence accumulates, the posterior becomes dominated by the likelihood, virtually independent of the shape of the prior, and ultimately converges to a point.)

[31] Its counterpart in frequentist econometrics is a table of estimated values of θ with their estimated standard errors.

CALCULATION 1.5 As a numerical example of a posterior distribution take the case of the likelihood plotted in the first panel of figure 1.4 which arose when one success was observed in one trial. If our prior was uniform – a beta(1, 1) density – then the posterior distribution is just the likelihood and is

$$p(\theta|y) \propto \theta$$

which is the 45 degree line plotted in that graph. This is a proper posterior density, its normalizing constant is 2, and, for example, the posterior expectation of θ after one success in one trial is

$$E(\theta|s = 1,\ n = 1) = \int_0^1 2\theta^2\, d\theta = 2/3.$$

This contrasts rather sharply with the maximum likelihood estimate, which either doesn't exist or is 1, depending on how you define the parameter space, Θ.

In this example, with a uniform prior, $p(\theta) \propto 1$, the posterior distribution, without its normalizing constant, is identical to the likelihood. This is clearly generally true. So if you now look back to likelihoods plotted earlier in this chapter, for example in figures 1.1, 1.2, and 1.3, you are, in effect, looking at posterior distributions under uniform priors. So you can read these figures as if they told you your beliefs about θ from that model and that data. For example, figure 1.1 tells you that the most probable value of β is about 0.9; that values in excess of 0.93 or less than 0.86 seem very unlikely; that $\theta = 0.89$ is about five times as probable as, say, $\theta = 0.88$ and so on. Similarly, under a prior for n which is uniform on the positive integers figure 1.5 shows the posterior distribution which points to values of n of about 14 and indicates that values of n greater than 25 are very improbable.

Reporting the posterior distribution

How might you report your posterior distribution?

DRAW IT

In the case in which θ is scalar the best way of conveying to readers the content of the posterior distribution is by drawing it. This is also true when θ is vector valued but the parameter of interest is a one-dimensional function of θ, as it often is in econometrics. For example economists are often interested in $\partial y/\partial x$, the marginal effect of x on y. Outside the linear model this may well be a function of many or all parameters of the model as in the probit model of example 1.4. With x and β vectors of k elements the general version of that model is

$$P(Y = 1|x, \beta) = \Phi(x\beta)$$

and $\partial P(Y = 1|x, \beta)/\partial x_j$ where x_j is the j'th element of x is given by $\beta_j\Phi(x\beta)$ which involves every element of the k dimensional parameter β. To report this object at some chosen value of x you would compute its posterior distribution from that of β and draw it.

REPORT ITS MOMENTS

Traditional econometric practice is to report an estimate of θ together with an estimate of the standard deviation of its repeated sampling distribution. If you wish to conform to this practice you might want to report the mean (or median) of the posterior distribution together with the standard deviation of that distribution.

REPORT A HIGHEST POSTERIOR DENSITY REGION

Similarly, traditional practice often reports a confidence interval for (scalar) θ. This is a numerical interval with the somewhat arcane interpretation that if you calculated your interval in the same way over many hypothetical repeated samples of the same size and using the same model then, say, 95% of such intervals would contain within them the "true" value of θ. The Bayesian analogue is to find, from the posterior distribution of θ, an interval[32] in Θ such that with probability 0.95 θ lies within it. It's as simple as that. Of course there are many ways of capturing 95% of the probability in a distribution and standard Bayesian practice is to construct the interval in such a way that no point in Θ has smaller probability density than any point outside it. This is called a (95%) *highest posterior density* – hpd – interval. Here is a numerical example. We take the autoregressive model of example 1.3 and artificially generate 51 observations, starting at $y_1 = 0$, with $\rho = 1$ which is called the **random walk** model, and with unit precision. The likelihood was shown to be of the normal form with mean $r = \sum_{t=2}^{51} y_t y_{t-1}/\sum_{t=2}^{51} y_{t-1}^2$ and standard deviation equal to $s = 1/\surd(\sum_{t=2}^{51} y_{t-1}^2)$. From the data we generated we find that $r = 1.011$ and $s = 0.037$. Now if we take the prior for ρ to be uniform on the real line, $p(\rho) \propto 1$, $-\infty < \rho < \infty$, the posterior density of ρ is equal to the likelihood and so is itself normal (r, s). Then from well known properties of the normal curve we know that 95% of the distribution will lie within 1.96 standard deviations of the mean and 99% will lie within 2.58 standard deviations of the mean. Further, the intervals $r \pm 1.96s$ and $r \pm 2.58s$ are such that all points within them have higher probability density than any point outside them. Thus they are an hpd interval. For our data we find a 95% hpd interval to be $0.939 < \rho < 1.084$ and a 99% interval is $0.916 < \rho < 1.107$. The interpretation of such intervals is very simple: for example, "the probability that ρ lies within the interval 0.939 to 1.084, given the model and data, is 0.95."

[32] More generally, a set.

CALCULATE THE MARGINALS

The calculation involved in forming the posterior distribution of the object of interest may well be mathematically challenging, to say the least. To work out the distribution of $\beta_j \phi(x\beta)$ in a probit model is very hard. Similarly, if the object of interest is, say, the third element θ_3 in a model parameter of k elements, to find its marginal density will involve doing the sum

$$p(\theta_3|y) = \int_{\theta_1} \int_{\theta_2} \int_{\theta_4} \dots \int_{\theta_k} p(\theta|y)\, d\theta_1\, d\theta_2\, d\theta_4 \dots d\theta_k, \tag{1.35}$$

a $k - 1$ dimensional integration. This is, in general, a hard problem.[33] Fortunately there are two solutions, one fairly old and of wide though not universal applicability, the second new, rather easy and of what is apparently universal application. The first is the use of approximations to posterior distributions and the second is the method of (computer) assisted sampling, which we shall treat in chapter 4.

Approximate properties of posterior distributions

If your posterior distribution is mathematically complicated or the dimension of θ is such that the integration (1.35) is hard to do it seems natural to look for a useful approximation. Clues to such an approximation are the likelihood graphs that we have drawn earlier in this chapter. These seem to suggest that likelihoods tend to look roughly normal, at least when the number of observations is not very small. Now if we could prove a theorem that states that when the number of observations in large posterior distributions are approximately normal, then integrals such as (1.35) are easily done. This is because if $p(\theta|y)$ is multivariate normal then all its marginal distributions are themselves normal so we would know immediately that, say, $p(\theta_3|y)$ is just a normal distribution. All that would then remain is to deduce its mean and precision.

The relevant theorem states the following proposition.

THEOREM 1.1 LARGE SAMPLE APPROXIMATE POSTERIOR

Let θ be the parameter, possibly vector valued, and let $p(\theta|y)$ be the posterior distribution, then for sufficiently large n, θ is approximately normally distributed with mean equal to $\hat{\theta}$ and precision (matrix) equal to $-H(\hat{\theta})$ where $\hat{\theta}$ is the posterior mode and H, the hessian, is the matrix of second derivatives of the logarithm of the posterior density function. Under a uniform prior for θ the posterior distribution is equal to the likelihood and so $-H(\hat{\theta})$ is equal to the negative second derivative of the log likelihood evaluated at $\hat{\theta}$. The expected value of the negative hessian of the log likelihood with respect to the

[33] It's the sort of problem that, as I described in the preface, defeated my efforts many years ago.

distribution of y given θ is the information (matrix), I_θ, mentioned earlier. In practice, $I_\theta(\hat{\theta})$ – called the ***observed information*** *– and $-H(\hat{\theta})$ will be close except when the number of observations is relatively small or the prior is far from flat near $\hat{\theta}$.*

Proof *For further discussion and references to proofs see Bernardo and Smith (1994).*

It should be noted that this multivariate normal approximation to the posterior distribution applies to *any* parametrization of the model. Since for θ the result states that approximately $\theta \sim n(\hat{\theta}, -H(\hat{\theta}))$,[34] it also implies that $g(\theta) \sim n(g(\hat{\theta}), -J(\hat{\theta}))$, approximately, where

$$J(\hat{\theta}) = -\left(\frac{\partial^2 p(\theta|y)}{\partial g^2}\right)_\theta = -\frac{\partial^2 p(\theta|y)}{\partial \theta^2}\left(\frac{\partial \theta}{\partial g}\right)_\theta^2 = H(\hat{\theta})\left(\frac{\partial \theta}{\partial g}\right)_\theta^2$$

and $g(\theta)$ is any differentiable one-to-one function of θ. A potentially important warning should be made here, that for any given data set and model, the normal approximation for θ can be very accurate, but the corresponding normal approximation for $g(\theta)$ can be very inaccurate, especially if $g(.)$ is a markedly non-linear function. This works in reverse in that $g(\theta)$ can be nearly normal but θ far from normal. Also, since a multivariate normal distribution has a single mode this theorem can't provide a useful approximation when the posterior density has several modes.

Another important warning is that although the theorem is stated as a "large *n*" result, it is almost always not the sample size that determines when the sample is large but some other function of the data. For example, in a non-linear regression model, which we shall study in chapter 5, it is objects such as $\sum_{i=1}^n (x_i - \bar{x})^2$ that determine whether approximate normality of the posterior distribution is, or is not, a good approximation. This sum of squares generally increases with the sample size, *n*, yet it may be very small, even zero, even though *n* is very large, and it can be very large even though *n* is very small. Just looking at the number of observations generally gives a misleading answer to the question of whether approximate normality of the posterior is reasonable.[35]

EXAMPLE 1.15 ***PROBIT COEFFICIENT POSTERIOR*** *For an example of a normal approximation take the probit model of example 1.4 where, under a uniform prior for β, the posterior density is equal to (1.11) with logarithm equal to*

[34] The symbol ~ means "is distributed as."

[35] We shall see a striking example of this in chapter 8 where we find that 36,000 observations in a model with seven parameters is a very "small" sample indeed.

$$\log p(\beta|y) = \Sigma_{i=1}^{n} y_i \log \Phi(\beta x_i) + \Sigma_{i=1}^{n}(1 - y_i)\log(1 - \Phi(\beta x_i)).$$

The derivative of this expression with respect to β is

$$\begin{aligned}\frac{\partial \log p(\beta|y)}{\partial \beta} &= \Sigma_{i=1}^{n} y_i x_i \frac{\phi(\beta x_i)}{\Phi(\beta x_i)} - \Sigma_{i=1}^{n}(1 - y_i)x_i \frac{\phi(\beta x_i)}{1 - \Phi(\beta x_i)} \\ &= \Sigma_{i=1}^{n} \frac{x_i \phi(\beta x_i)(y_i - \Phi(\beta x_i))}{\Phi(\beta x_i)(1 - \Phi(\beta x_i))},\end{aligned}$$

and the posterior mode $\hat{\beta}$ *equates this derivative to zero. The solution exists and is unique as long as the y's and x's vary but it must be found numerically. Differentiating this expression to find the hessian results in a rather complicated expression though one which is readily evaluated on the computer. The negative hessian is, however, when the number of observations is not too small, often well approximated by the information. The information matrix is usually a simpler expression than the hessian itself and it is in this case where it is*

$$-E\left(\frac{\partial^2 \log p(\beta|y)}{\partial \beta^2}|\beta\right) = \Sigma_{i=1}^{n} \frac{x_i^2 \phi(\beta x_i)^2}{\Phi(\beta x_i)(1 - \Phi(\beta x_i)} = I(\beta).$$

A normal approximation to the joint posterior density of β would then be

$$p(\beta|y) \approx n(\hat{\beta}, I(\hat{\beta})). \tag{1.36}$$

CALCULATION 1.6 For a numerical comparison we generate some data with $n = 50$ and $\beta = 0$ and plot the posterior density under a uniform prior for β – this is the dotted line in figure 1.6. On this we superimpose as the solid line the normal approximation (1.36). This is the normal density with mean $\hat{\beta}$ and precision $I(\hat{\beta})$ where $\hat{\beta}$ is the maximum likelihood estimate of β – the posterior mode under a uniform prior. The two curves are indistinguishable.

Likelihood dominance

Another important feature of posterior distributions in general is that they typically depend very little on the prior when the number of observations is large relative to the number of parameters and the prior does not assign probability zero to the relevant parts of Θ. To see generally why this is likely, consider the logarithm of

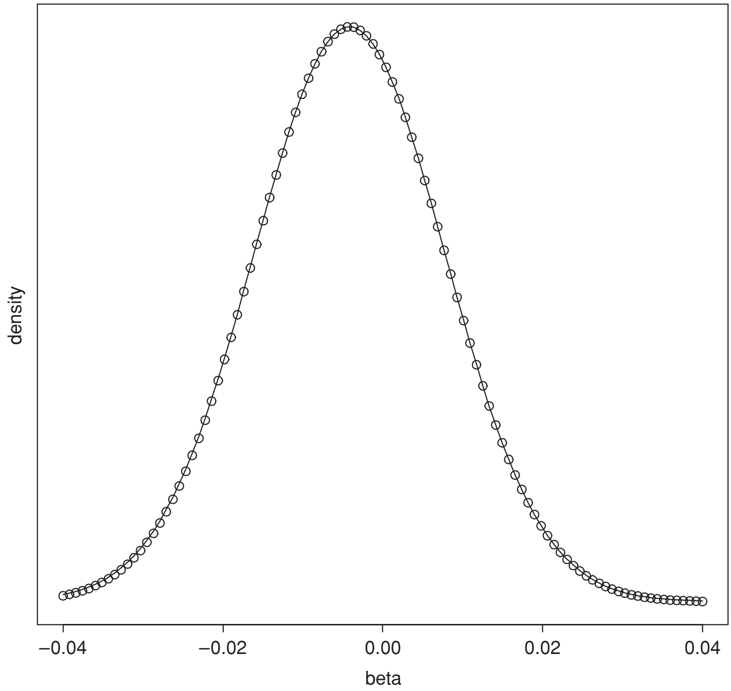

Figure 1.6 Probit posterior and its asymptotic normal approximation

the posterior density of θ using a sample of size n. It has two components,[36] the log likelihood and the log prior,

$$\log p(\theta|y_1, ..., y_n) = \log \ell(\theta; y_1, ..., y_n) + \log p(\theta).$$

Now as data accumulate and n increases the likelihood changes and tends to increase in modulus, but the prior stays the same. To see an example of this increase, consider the likelihood for n independent normal (μ) variates which is $\exp\{-(1/2)\Sigma(y_i - \mu)^2\}$ with logarithm $-(1/2)\Sigma_{i=1}^{n}(y_i - \mu)^2$. The increment in the likelihood when you add an extra observation is therefore $-(1/2)(y_n - \mu)^2$ which is either negative or zero (which happens with probability zero) and so, for almost all μ, the log likelihood becomes a larger and larger negative number as observations accrue. Hence, in large samples the likelihood will be the numerically dominant term as long, of course, as $p(\theta) > 0$. This is true rather generally and it also works in the case of dependent or non-identically distributed data. This argument will fail if $p(\theta)$ is zero over the region in Θ where $\ell(\theta|y)$ tends to concentrate since $\log p(\theta) = -\infty$ over that region. But if $p(\theta)$ is not dogmatic and assigns some probability to all relevant parts of Θ then it will indeed be eventually dominated. Here is an example of dominance of the posterior by the likelihood.

[36] Apart from an irrelevant additive constant.

EXAMPLE 1.16 ***WEAK DEPENDENCE OF THE POSTERIOR ON THE PRIOR*** *Consider the Bernoulli trials example with likelihood $\propto \theta^s(1-\theta)^{n-s}$ and consider the effect of varying the prior within the natural conjugate beta family, $p(\theta) \propto \theta^{a-1}(1-\theta)^{b-1}$. Suppose that $n = 20$ with $s = 7$. Figure 1.7 shows three quite different beta prior distributions. The horizontal line is a uniform prior with $a = b = 1$; the line composed of circles is beta density with $a = b = 3$; and finally the solid line is the Jeffreys' prior with $a = b = 1/2$. These priors show quite different initial beliefs. Figures 1.8 and 1.9 show posteriors using these priors. In figure 1.8 we had $n = 5$ trials with $s = 2$ successes, while in figure 1.9 we had $n = 20$ trials with $s = 8$ successes. The solid line is the posterior with Jeffreys' prior; the starred line corresponds to the uniform prior; and the remaining line to the beta (3, 3) prior.*

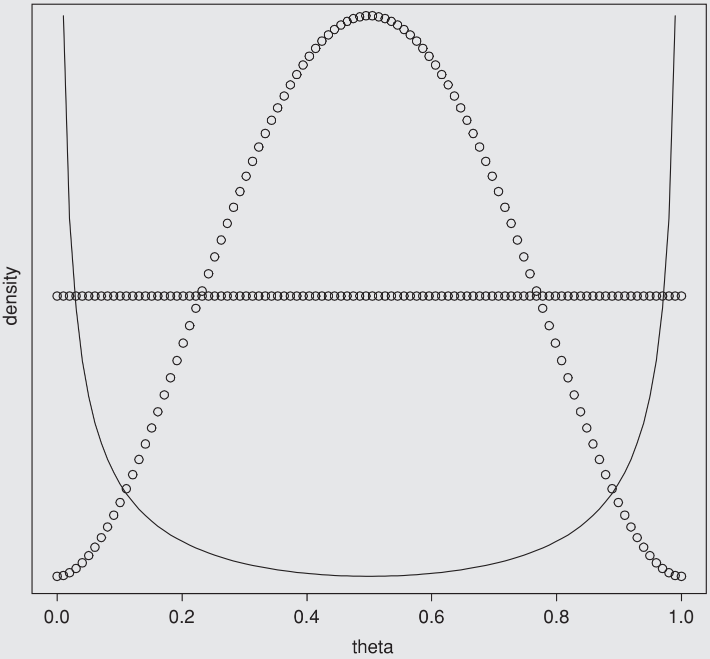

Figure 1.7 Three priors for a Bernoulli parameter

The message of the figures is that divergent prior beliefs can be brought rapidly into rough agreement in the face of quite limited amounts of evidence, and that the agreement is more complete the more data are available.

This argument for the large sample dominance of the likelihood over the prior is sometimes said to mimic a process of rational scientific enquiry in that two individuals with quite different, but non-dogmatic, prior beliefs will be brought into agreement

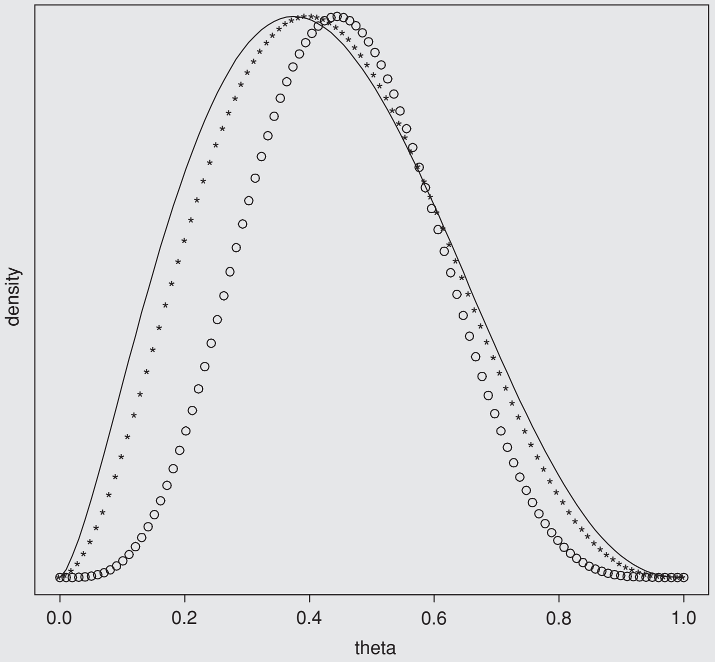

Figure 1.8 Three posteriors: $n = 5$, $s = 2$

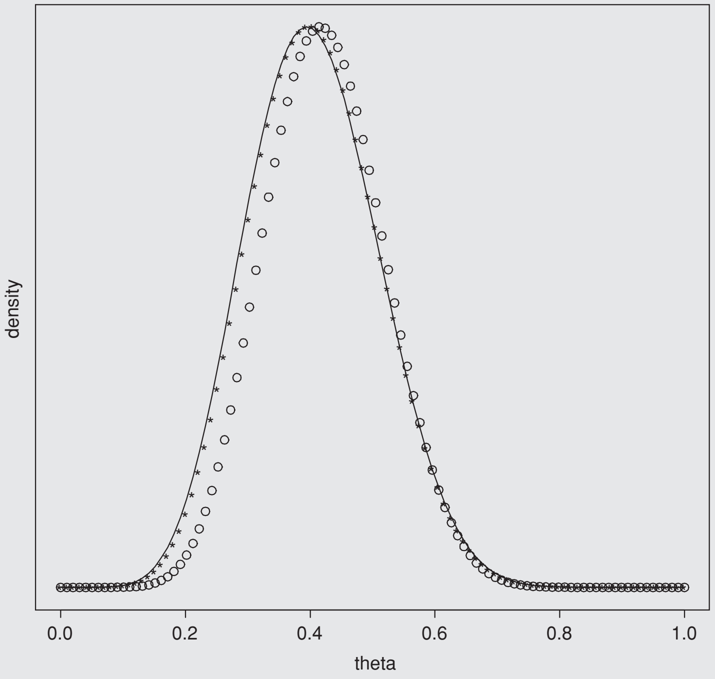

Figure 1.9 Three posteriors: $n = 20$, $s = 8$

by the accumulation of evidence. Note that such individuals must agree on the likelihood even if they disagree in their prior beliefs.

Convergence of the posterior distribution

We have just shown, with an example, that people with quite diverse prior beliefs can be brought into agreement if the sample size is large enough. But a different question also arises. What happens to the posterior distribution when the number of observations becomes large? Do your beliefs tend to concentrate on some element of Θ? And if so, on what? Here's an argument that shows what happens in a special but important case.

THEOREM 1.2 CONVERGENCE OF POSTERIOR DISTRIBUTIONS

Suppose the parameter space Θ is discrete with elements θ_1, θ_2, ... possibly infinite in number. Let the observations be iid conditional on θ with densities $p(x_i|\theta)$ and suppose that there exists in Θ a true parameter labeled θ_t, which is distinguishable from all the other elements of Θ by the condition that

$$\int p(x|\theta_t)\log\left[\frac{p(x|\theta_s)}{p(x|\theta_t)}\right]dx < 0 \quad \textit{for all } s \neq t. \tag{1.37}$$

The integral in this expression is the Kullback–Leibler measure of the divergence between the two probability distributions $p(x|\theta_t)$ and $p(x|\theta_s)$ and what the condition says is that all the possible data distributions (likelihoods) arising from values of θ different from θ_t are different from $p(x|\theta_t)$.[37] *Some such identification condition is clearly necessary to prove convergence. After all, suppose that there existed a θ, say θ_s, such that $p(x|\theta_t) = p(x|\theta_s)$ for all x, then there would be no way of deciding whether an observed sample x had been provided by $p(x|\theta_t)$ or by $p(x|\theta_s)$. (1.37) is called an* ***identification condition****.*

Then the theorem is

$$\lim_{n\to\infty} p(\theta_t|x) = 1$$

This means that all the mass in the posterior distribution comes eventually to concentrate on a single point in the parameter space.

Proof *Taking the prior as $p_s > 0$ for each $\theta_s \in \Theta$ the posterior density is*

[37] Cf. definition 1.3 above.

$$
\begin{aligned}
p(\theta_s|x) &= p_s\frac{p(x|\theta_s)}{p(x)}, \quad \textit{for } p(x\,|\,\theta_s) = \Pi_{i=1}^{n} p(x_i\,|\,\theta_s) \\
&= \frac{p_s\{p(x|\theta_s)/p(x|\theta_t)\}}{\Sigma_j p_j\{p(x|\theta_j)/p(x|\theta_t)\}} \\
&= \frac{\exp\{\log p_s + S_s\}}{\Sigma_j \exp\{\log p_j + S_j\}}
\end{aligned}
$$

where

$$S_j = \log\frac{p(x|\theta_j)}{p(x|\theta_t)} = \Sigma_{i=1}^{n}\log\frac{p(x_i|\theta_j)}{p(x_i|\theta_t)}.$$

But the right hand expression shows that S_j is the sum of n independent and identically distributed random variables so that, by a strong law of large numbers,

$$\lim_{n\to\infty}\frac{S_j}{n} = E\left(\frac{S_j}{n}\right) = \int p(x|\theta_t)\log\left[\frac{p(x|\theta_j)}{p(x|\theta_t)}\right]dx = \begin{cases} 0 & \textit{if } j = t \\ <0 & \textit{if } j \neq t \end{cases}.$$

If we then apply this result to find the limiting behavior of $p(\theta_s|x)$ as $n \to \infty$ we see that terms like $\exp\{\log p_j + S_j\}$ *converge to zero because S_j becomes a larger and larger negative number, except when $j = t$, from which the theorem follows.*

This type of theorem can be generalized to continuous parameter spaces and to observations that are neither independent nor identically distributed under suitable further conditions.

This theorem forms a precise statement of how individuals with quite different initial beliefs can be brought into ultimate agreement by the accumulation of evidence.

Sampling the posterior

The material in this section is essential to understanding the point of view taken in this book.

The difficulty with (asymptotic) approximations like the one sketched in the last section is that one can never be sure of their accuracy. Indeed the only way of finding out the accuracy of an approximation to your posterior distribution is to calculate the exact distribution which is what you wanted to avoid doing! This is one reason why approximations, though important, take a second place in modern Bayesian calculations to simulation methods developed during the last ten years or so. These methods depend upon the following remarks.

Suppose that you take a posterior distribution and draw from it a collection of realizations of θ. If you program your machine to produce `nrep` realizations from $p(\theta_1, \theta_2|y)$ your output will be a matrix with `nrep` rows and as many columns as there are elements of θ. Thus, when θ has two elements, it will look like

$$\begin{matrix} \theta_{11} & \theta_{21} \\ \theta_{12} & \theta_{22} \\ \theta_{13} & \theta_{23} \\ . & . \\ \theta_{1,nrep} & \theta_{2,nrep} \end{matrix}$$

Each row of this matrix contains a realization of a random variable whose distribution is $p(\theta_1, \theta_2|y)$. The whole matrix contains nrep realizations from the *joint* distribution of θ_1 and θ_2 while the j'th column contains nrep realizations from the *marginal distribution* of θ_j. To study the distribution of, say, θ_1 given the data, $p(\theta_1|y)$, just ignore the second column, it's as simple as that. To study the distribution of some function of θ, say $g(\theta)$, just apply this function to every row of your output matrix and the result will be a set of realizations of the random variable $g(\theta)$. (It is desirable, but not essential, that the rows of your output matrix be independent realizations of θ.) Whether they are independent or not, a law of large numbers will generally apply and it can be proved that moments of $g(\theta)$ from a sample of nrep realizations of θ will converge in probability to the moments of the distribution of $g(\theta)$ as nrep $\rightarrow \infty$. Since *you choose nrep* you can make arbitrarily accurate estimates of any aspect of the posterior distribution including, for example, the mean, precision, distribution and density functions. *It follows that if you can sample the distribution in question you can know it with arbitrary accuracy.* Computer assisted sampling to avoid integration is the key feature of *modern* Bayesian econometrics[38] and the approach described in this paragraph is critical to understanding this subject. Increasingly, difficult mathematics is being abandoned in favor of computer power.

Computer assisted sampling requires not only computer power but also effective algorithms that can be proved to sample the distribution in question. We have already in this book made extensive use of computer routines to provide artificial data sets and to sample likelihoods and posterior distributions. These calculations rely on computer languages like S or Matlab that have built in commands, like `rnorm` or `rexp`, to sample most of the standard distributions of elementary probability theory. But where the distribution to be sampled is not standard, researchers either have to have put together their own program or, increasingly, use specialized sampling software. In the rest of this book we shall use one of the most widely used pieces of sampling software, a program called **BUGS**. In chapter 4 we shall give an account of the theory behind this program and in appendix 2 we shall provide some instruction on

[38] It was only in about 1990 that computer assisted sampling started to become widespread in many areas of applied statistics. This is because it was about that time that powerful computers became readily available to researchers. This development has radically altered applied statistical practice.

its use. We conclude these introductory remarks on the sampling study of posterior distributions with an example of the use of BUGS to solve a complicated problem.

EXAMPLE 1.17 ***PROBIT REVISITED*** *Consider binary data y whose mean depends on two variables* x_1 *and* x_2 *according to*

$$E(Y|x, \beta) = \Phi(x\beta),$$

where $x\beta = \beta_0 + \beta_1 x_1 + \beta_2 x_2$*. Thus* $\theta = (\beta_0, \beta_1, \beta_2)$ *is a three-dimensional parameter. If interest centers on a scalar function of* θ *then a two-dimensional integration will be needed to find its marginal distribution. For example, you might want to know the posterior distribution of the derivative of the probability that* $Y = 1$ *with respect to* x_1 *evaluated at alternative choices of x. This derivative may be of considerable economic interest and so you need to know its most likely value or its expected value or the chance that it is negative. In this case the parameter of interest is*

$$\gamma = \frac{\partial\Phi(x\beta)}{\partial x_1} = \beta_1\phi(x\beta)$$

and it is its posterior distribution that you require. The modern way of finding this distribution is to sample the joint posterior distribution of $(\beta_0, \beta_1, \beta_2)$ *then, for each realization of these three numbers, compute* γ *for some, perhaps typical, x vector of interest.*

A first look at BUGS

BUGS calculation To illustrate the method we generated some artificial data and used the BUGS program to generate a sample of 10,000 realizations of $(\beta_0, \beta_1, \beta_2)$. We then substituted these values into the expression for γ at the x vector, say, $x_1 = 1$, $x_2 = 1$ and this gives 10,000 realizations from the marginal posterior distribution of γ. These can then be studied in whatever way you find helpful.

Data were generated with $n = 50$, $\beta_0 = 0$, $\beta_1 = 0.5$, $\beta_2 = -0.5$ The BUGS program follows exactly the Bayesian algorithm and so it requires you to tell it your likelihood and then to tell it your prior. The likelihood is (1.11) and it is written for the program as

```
model
{for(i in 1:n){
y[i]~dbin(p[i],1)
mu[i]<-beta0+beta1*x1[i]+beta2*x2[i]
p[i]<-phi(mu[i])}
```

The third line specifies that the *i*'th observation is a realization of a binomial (1, p_i) variate, the notation `~dbin` meaning "is distributed binomially." That is, Y_i is 1 with probability p_i and zero with probability $1 - p_i$. The next two lines state the probit model in which $p_i = \Phi(\beta_0 + \beta_1 x_{1i} + \beta_2 x_{2i})$. The statement `phi(x)` means evaluate the standard normal distribution at `x`, that is, calculate $\Phi(x)$. Lines two through five together provide the likelihood.

We then give the second component of the model which is the prior for β which in this case is specified as

```
beta0~dnorm(0,0.001)
beta1~dnorm(0,0.001)
beta2~dnorm(0,0.001)}
```

These statements imply that the three elements of β are distributed independently normally with means zero and very low precisions implying standard deviations of $1/\surd(0.001) = 31$. This means that we are saying that, before seeing the data, we think that each element of β is very likely to lie within –93 and +93 which is plus and minus three standard deviations of zero. This is meant to be a vague prior. If you wish to put in more precise information, including dependence among the elements of β, or even less precise information, you may, of course, do so. Notice that the whole model, the likelihood and the prior, is enclosed by {}.

After supplying the data matrix containing as columns the values of y, x_1 and x_2 and some further details including the number of realizations required which in the present case was chosen to be nrep = 10,000, the program then produces an output matrix containing 10,000 realizations of the three elements of β. Each element of this matrix contains, to a close approximation, realizations from the joint posterior distribution of β corresponding to the likelihood, prior and data that we supplied.

To illustrate the procedure, figure 1.10 gives the smoothed histogram of the realizations of the marginal distribution of β_2. The value of β_2 that generated the data used here was $\beta_2 = -0.5$; the plot, as can be seen, is centered about −0.22. The mean and median of β_2 were both about −0.21. Finally, for comparison, the maximum likelihood estimate of β_2 was also −0.21. The "true" value of β_2 is one of the less probable values, though it would still lie within a 95% highest posterior density region.

To conclude this example we calculate the value of γ for $x_1 = 1$, $x_2 = 1$ for each of our 10,000 realizations and its smoothed histogram is given in figure 1.11. It can be seen that the effect of x_1 on the success probability is certainly positive and most probably about 0.17.

We shall explain in chapter 4 the methods used by BUGS to produce the realizations described here, and we shall make frequent use of this program throughout the book. Appendices 2 and 3 describe the BUGS language and give BUGS programs for many standard econometric models.

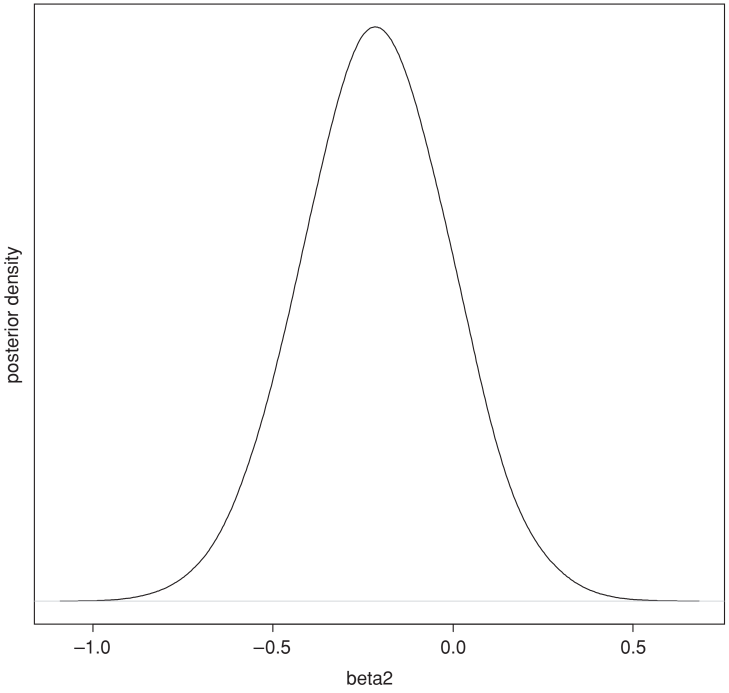

Figure 1.10 Marginal posterior density of β_2

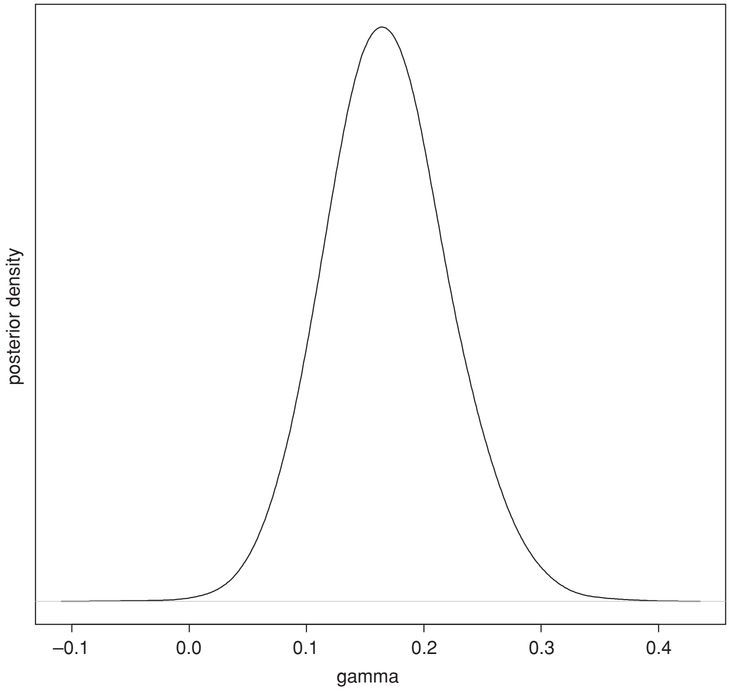

Figure 1.11 Posterior density of $\beta_1\phi(x\beta)$ at $x_1 = x_2 = 1$

1.4.4 Decisions

Many writers prefer to view the problem of inference from data or evidence as that of making a decision or taking an action. An agent is viewed as setting up a model, including both $p(y|\theta)$ and $p(\theta)$; observing the data, y^{obs}; and then, in the light of the posterior distribution of θ, taking a decision. The decision might be to invest some money, to announce a forecast of Y, to report a single numerical value as an estimate of θ, etc. This is a point of view that is attractive to economists, who want to consider the agents whose behavior and interaction they model to be rationally coping with the uncertainties they face. This book does not take a decision theoretic perspective, though it is not inconsistent with one. This is because the problem faced by most economists or intending economists does not seem well described as one of decision. It seems more like that of sensibly and concisely reporting their findings, and for this the recommended procedure is to draw the marginal(s) of the object of interest. This leaves it up to others, for example policy makers, to use your report as a basis for decision making.

For the sake of completeness we give a very brief review of one version of the decision problem, that of choosing a point estimate of θ. Suppose that you have a model for potential data y involving a parameter θ. After having seen the data you will have a posterior distribution for it, $p(\theta|y)$. You are required to reach a single numerical decision, d, about θ. This decision will depend on the data y so $d = d(y)$. The decision theory approach assumes the existence of a **loss function** $L(d, \theta)$ that gives the loss to the decision maker, you perhaps, of making decision d when the parameter, about which you are uncertain, takes the value θ. A **Bayes decision** minimizes the expected loss

$$\hat{d} = \text{arg.min} \int_{\Theta} L(d, \theta)p(\theta|y)\, d\theta \quad \text{for } \hat{d} \in \Theta.$$

EXAMPLE 1.18 ***SQUARED ERROR LOSS*** *Suppose the loss function takes the symmetric form $L(d, \theta) = (d - \theta)^2$ – squared error loss – then $\hat{d}$ is the posterior mean, $E(\theta|y)$. To prove this note that the expected loss is*

$$\int_{\Theta} L(d, \theta)p(\theta|y)\, d\theta = \int_{\Theta} (d - \theta)^2 p(\theta|y)\, d\theta$$

and a simple differentiation with respect to d provides the result.

A nice application of this is:

EXAMPLE 1.19 *DATA UNIFORM ON 0 TO θ* *Let Y be uniformly distributed between 0 and θ so its density function is*

$$f_Y(y) = \frac{1}{\theta}, \quad 0 \leq y \leq \theta,$$

and zero elsewhere. Under the default (improper) prior density $p(\theta) \propto 1/\theta$ the posterior density from n independent realizations is

$$p(\theta | y) \propto \frac{1}{\theta^{n+1}}, \quad \theta \geq y_{\max} \tag{1.38}$$

where $y_{\max}$ is the largest of the n sample realizations. (This comes about because θ is, by definition, not less than any observation, so it is certainly not less than the largest observation.) The kernel (1.38) is that of a proper density, for $n \geq 1$, and after supplying the normalizing constant it can be written as

$$p(\theta | y) = \frac{n y_{\max}^n}{\theta^{n+1}} \quad \theta \geq y_{\max}. \tag{1.39}$$

Under squared error loss the Bayes decision is the posterior mean. Carrying out the integration we find that the mean exists for $n > 1$ and is

$$d(y) = \frac{n}{n-1} y_{\max}.$$

So your Bayes decision under squared error loss is to take the largest observation in your data, multiply it by $n/(n-1)$, i.e. slightly increase it, and report the resulting number. (The maximum likelihood estimator of θ for this problem is $\hat{\theta} = y_{\max}$ which will underestimate θ for essentially any data set – it will always *be too low.)*

Decision making and reporting are not necessarily alternatives. In many situations an economic model will envisage agents taking decisions under uncertainty. An analysis of data using a model that incorporates agents making decisions under uncertainty will lead to you – the uncertain investigator – reporting your analysis of uncertain agents who are presumed to be taking decisions under uncertainty in an optimal (Bayesian) way. So really both the decision making and reporting perspectives on Bayesian inference should receive emphasis in a text on Bayesian econometrics not because econometricians take decisions but because agents do.

1.5 CONCLUSION AND SUMMARY

The Bayesian approach to econometrics is conceptually simple and, following recent developments, computationally straightforward. Following the algorithm given in section 1.3.2 you must formulate your theory as a conditional probability statement for the data that you are about to see and a prior distribution over the parameters of that statement. This is equivalent to making a simple statement about what you think the data should look like on your theory since $\int p(y|\theta)p(\theta)d\theta = p(y)$, a probability distribution for the data. You then study the data and determine whether the model is, at least roughly, consistent with the evidence and, if it is, you proceed to revise your views about the model parameters. Whether the data are consistent with the model or not, you will have learned something. In view of this conceptual and computational simplicity the rest of this book is little more than a series of examples with some account of recently developed computational algorithms.

1.6 EXERCISES AND COMPLEMENTS

In this section we give some further worked examples on priors, likelihoods and posterior distributions and ways to study them and we suggest some exercises.

(1) Simulation

Study of likelihoods and posterior or prior distributions is, as we have seen, often aided by computer simulation and graphics. In some of the examples in this chapter we simulated some data satisfying the model and plotted both the data and the likelihood they imply. Many computer languages make this easy and you may, of course, use whatever language you like but my own favorite is the language S. An appendix to the book describes the elements of this language and suggests reading. Here are some examples in which data are simulated and graphics used. If you have access to a copy of S – there is a shareware version, called *R*, on the web at *http://www.r-project.org/* – you should try these commands.

To simulate a regression model as in example 1.2 in which y is normal given x with mean βx and precision τ you can use

`beta <- 0.3` . . . specifies the value of β.

`tau <- 1` . . . specifies the value of τ.

`n <- 50` . . . chooses the sample size.

`x <- runif(n,10,20)` . . . produces x values uniformly distributed from 10 to 20.

`y <- rnorm(n,beta*x,1/sqrt(tau))` . . . produces y values with mean βx and standard deviations $1/\sqrt{\tau}$.

`plot(x,y)` . . . plots the data on a scatter diagram.

`b <- sum(x*y)/sum(x^2)` . . . calculates the least squares estimate.

`abline(0,b)` . . . draws the least squares line $y = bx$ on the scatter diagram, with intercept zero and slope b.

`sdb <- 1/sqrt(tau*sum(x^2))` . . . finds the standard deviation of the normal curve that defines the likelihood, (1.7)

`bval <- seq(b-4sdb, b+4*sdb, length=200)` . . . chooses 200 points at which to evaluate the likelihood. These points cover the range over which the likelihood will be non-negligible.

`plot(bval,dnorm(bval,b,sdb),type="l")` . . . draws the likelihood exploiting the fact that for this model it has the shape of a normal curve with mean `b` and standard deviation `sdb`. The plot command evaluates the function at the 200 points specified in `bval` and then the command `type = "l"` joins the points to form a continuous curve.

EXERCISE Generate your own data using this normal regression model and plot the data and the likelihood.

(2) A regression model for counts

Theory suggests that y should depend on x but the data will be counts of how often some events occurred. Econometric applications might be the numbers of strikes occurring in particular industries in a year or the numbers of patent applications filed by different firms over a year. Because the y values will be positive integers or zero such data cannot be normally distributed. The standard model for count data is the poisson with probability mass function

$$p_Y(y) = \frac{\mu^y e^{-\mu}}{y!}, \qquad y = 0, 1, 2, 3, \ldots, \qquad \mu > 0.$$

A (non-linear) regression model then takes Y as poisson with mean $\mu = \exp\{\beta x\}$ given x. The exponential function is chosen because it guarantees that the mean is always positive. This implies that for n independent observations on y and x the likelihood is

$$\begin{aligned} \ell(\beta; y) &\propto \textstyle\prod_{i=1}^{n} \exp\{\beta x_i y_i\} \exp\{-e^{\beta x_i}\} \\ &= \exp\{\beta \textstyle\sum_{i=1}^{n} x_i y_i\} \exp\{-\textstyle\sum_{i=1}^{n} e^{\beta x_i}\}. \end{aligned} \qquad (1.40)$$

This does not have the shape of a normal curve but, nonetheless, if you simulate some data and draw the function you will find, for almost all data sets, that the curve is approximately bell shaped.

To simulate some data choose n, β and x as in exercise 1 and then use the command `y <- rpois(n,exp(beta*x))`. To draw the likelihood define a set of

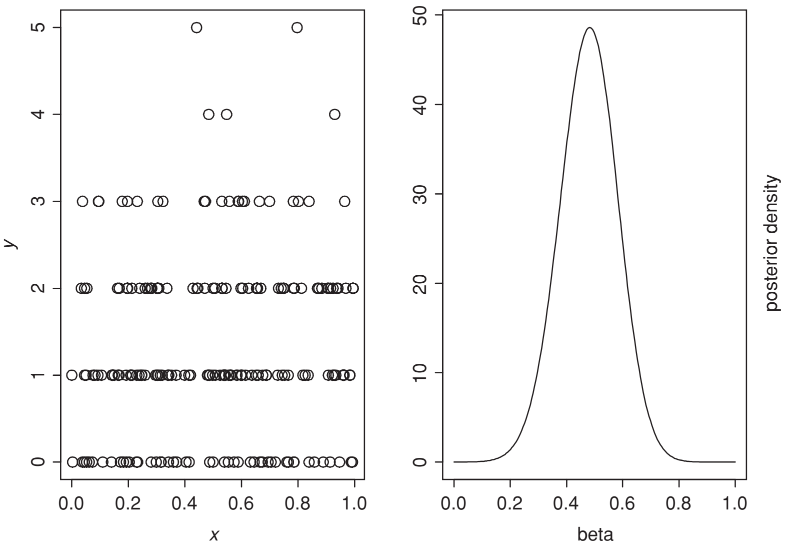

Figure 1.12 Data and likelihood for exercise 2

β values at which to evaluate it and store these in a vector `bval` as before, then define the logarithm of the likelihood function (1.40) by, say,

```
poissonlogl <- function(b){b*sum(x*y)-sum(exp(b*x))}
```

Finally use a "do loop" to evaluate the function at the points in bval, store these values in, say, `val` and plot the elements of exp(val) against bval. This could be done, if bval has `nval` elements, by

```
for(i in 1:nval){val[i] <- poissonlogl(bval[i])}
plot(bval,exp(val),type="l").
```

Figure 1.12 shows the results of such a simulation with $n = 200$, $\beta = 0.5$ and the x's uniform from zero to one. The first panel shows a scatter plot of the data and the second shows the likelihood. The scatter plot, as often with a discrete dependent variable, is quite hard to interpret. The likelihood is much easier and points clearly to a beta value in the neighbourhood of the value, 0.5, that produced the data.

There are two slight difficulties in doing this calculation. One is that, unlike the normal case of exercise 1, it is not evident where the likelihood will take its largest values. One solution is to make a preliminary calculation using a maximum likelihood routine to find the β value that gives maximum likelihood. Another solution is to calculate the likelihood for a wide range of β values to find out where the function is large. Another slight difficulty is that likelihood values can be very large numbers which may be hard to plot. The solution here is to calculate the mean value of the log likelihood and subtract it from `val` before issuing the plot command.

EXERCISE Generate your own count data satisfying a regression model and calculate the likelihood. Try increasing the value of n to observe the convergence of the posterior towards a point.

(3) Exponential likelihoods

The fundamental probability model for the duration of an event – how long it lasts – is the exponential. If a theorist reasons that agents with large values of x tend to have longer events than those with smaller x's a natural econometric model within which to embed this idea is to let the duration y be exponentially distributed with mean μ depending on x as for example $e^{\beta x}$. This would be a non-linear regression model with regression function $E(Y|x) = e^{\beta x}$. The exponential density of mean μ is

$$p_Y(y) = (1/\mu)e^{y/\mu}, \quad \mu, y > 0,$$

so the likelihood for n independent realizations corresponding to different x's is

$$\ell(\beta; y, x) = \exp\{-\beta\textstyle\sum_{i=1}^{n}x_i\}\exp\{-\sum_{i=1}^{n}y_i e^{-\beta x_i}\}.$$

EXERCISE Generate some durations using the command `rexp(n,exp(b*x))`, if you are using S, and plot the data and posterior density of β assuming a uniform prior.

(4) A double exponential model

A probability model that is in some ways a useful alternative to the normal distribution is the double exponential or Laplace distribution. In its simplest form this has density function

$$p(y|\theta) = \exp\{-|y - \theta|\}, \qquad -\infty < y, \theta < \infty.$$

This function is symmetrical about $y = \theta$ and on each side of θ it declines exponentially, hence the name. The mean, median and mode of Y are θ and the standard deviation is $\sqrt{2}$. This distribution is less dogmatic than the normal in that its tails decline like $e^{-|y|}$ which is slower than the normal rate e^{-y^2} so it allows for greater uncertainty about where y is located. Figure 1.13 plots the double exponential density function for the case $\theta = 1$.

Then n independent realizations of Y will have the joint probability density $p(y|\theta) \propto \exp\{-\sum_{i=1}^{n}|y_i - \theta|\}$ and this is also the posterior density under a uniform prior for θ. Figure 1.14 is a Laplace likelihood with $n = 3$.

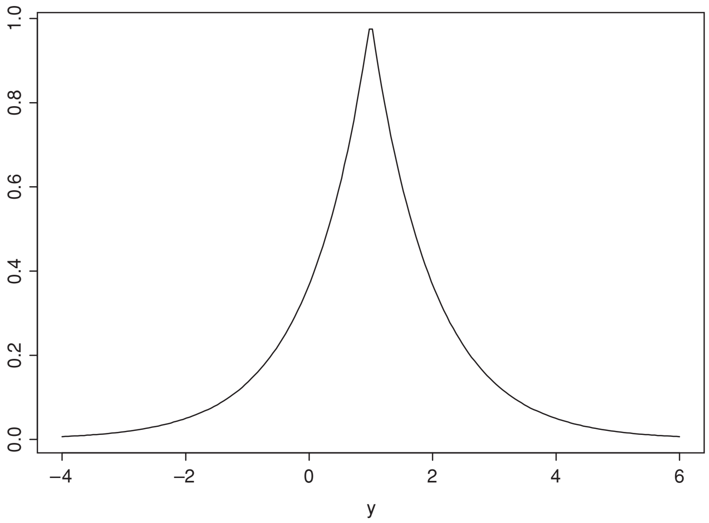

Figure 1.13 A double exponential density

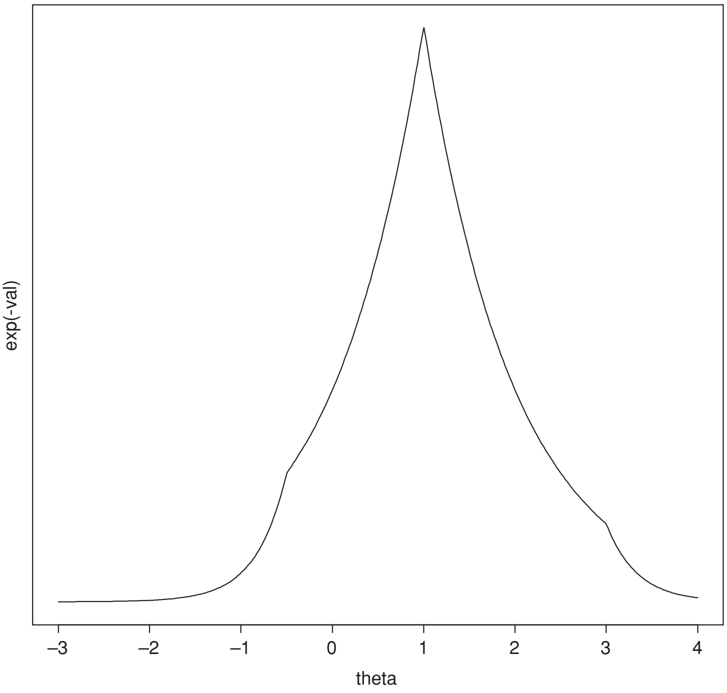

Figure 1.14 The likelihood for 3 observations of a Laplace variate

As can be seen from figure 1.14, the likelihood is kinked at each of the observations, which were $y = (-0.5, 1, 3)$. There are always as many kinks – points of non-differentiability – in the Laplace likelihood as there are distinct observations. (Nonetheless, the likelihood still approaches normality!)

EXERCISE Choose a value for θ and n and generate some data from the double exponential model. This can be done by generating n observations from an exponential distribution with mean = 1, changing the sign of these numbers with probability 0.5, and then adding θ. The first two steps here generate data from a double exponential centered at zero, and the final step centers the distribution at θ. The S command

```
  y<-rexp(n)*(-1+2*(runif(n)>0.5))+theta
```

will do this. The statement `runif(n)>0.5` produces n numbers equal to 1 if a uniform variate on 0 to 1 exceeds 0.5, which has probability 0.5, and 0 otherwise; multiplying these numbers by 2 and subtracting 1 turns them into a sequence of plus and minus ones; and these in turn randomly change the sign of the elements of `rexp(n)`.

(1) Choose a small, odd value of n and generate some data.
(2) Sketch the posterior density – by hand – and show that it is continuous but not everywhere differentiable.
(3) Show that the most probable value of θ is the median observation.

EXERCISE Generalize the previous model by setting $\theta = \beta x_i$ for $i = 1, 2, ..., n$, so that each y_i is Laplace distributed about its mean. This is an alternative to the normal regression model. Write down the likelihood for β; generate some data and plot the posterior density of β under a uniform prior for this parameter. Note that the most probable value of β minimizes the expression $\sum_{i=1}^{n} |y_i - \beta x_i|$. This is sometimes called a **median regression** model.

Further exercises

EXERCISE For an example of normal approximation take n independent normal variates with means zero and common precision τ. The likelihood is $\ell(\tau) \propto \tau^{n/2} \exp\{-\tau \sum y_i^2/2\}$ and the Jeffreys' prior is $\propto 1/\tau$. Find the log posterior density, calculate the posterior mode of τ and find $-H(\tau)$ and $I_\tau(\hat{\tau})$ at this mode. Hence find the normal approximation to $p(\tau|y)$. Now take $\sigma = \tau^{-1/2}$ as the parameter and find the normal approximation to the posterior density of σ.

EXERCISE The poisson distribution has mass function

$$p(y|\theta) = \frac{\theta^y e^{-\theta}}{y!}, \quad y = 0, 1, 2 \ldots, \quad \theta > 0.$$

The mean and variance of Y given θ are both equal to θ. Write down the likelihood for n independent realizations of Y and then the posterior density of θ under the conventional vague prior $p(\theta) \propto 1/\theta$. Work out the hessian of the log posterior and the posterior mode and hence construct an asymptotic normal approximation to the posterior. Simulate some data for $n = 5, 10, 20$ and compare the exact (gamma) posterior density of θ to its normal approximation.

1.7 APPENDIX TO CHAPTER 1: SOME PROBABILITY DISTRIBUTIONS

In this appendix we review some of the elementary probability distributions that have been used in the body of this chapter.

THE UNIVARIATE NORMAL FAMILY

The univariate normal family has two parameters, the mean μ and the precision τ. The kernel is $\exp\{-(\tau/2)(y - \mu)^2\}$ and the full density is

$$p(y) = \sqrt{\frac{\tau}{2\pi}} \exp\{-(\tau/2)(y - \mu)^2\}.$$

We refer to such a distribution by writing $Y \sim n(\mu, \tau)$. The standard normal has $\mu = 0$, $\tau = 1$, and kernel $e^{-y^2/2}$ with distribution function $\int_{-\infty}^{y} e^{-u^2/2}\, du/\surd(2\pi)$. Its density and distribution functions at the point y are denoted by $\phi(y)$ and $\Phi(y)$. The moment generating of a normal (μ, τ) variate is

$$M(t) = E(e^{tY}) = \int_{-\infty}^{\infty} e^{ty} p(y)\, dy = \exp\{t\mu + t^2/2\tau\}$$

from which the mean and variance are

$$E(Y) = \mu; \quad V(Y) = \frac{1}{\tau}.$$

The variance is denoted by σ^2.

Relevant S commands are as follows.

`rnorm(n, m, s)`................ *n* independent realizations from a normal distribution of mean *m* and standard deviation (not variance and not precision) *s*.

`dnorm(y, m, s)`................ the value of the density function at *y* of a normal (*m*, *s*) variate.

`pnorm(y, m, s)`................ the value of the (cumulative) distribution function at *y* of a normal (*m*, *s*) variate

`qnorm(p, m, s)`................ the quantile function at *p* of a normal (*m*, *s*) variate. This will produce the number *y* which is exceeded with probability 1 − *p* with such a distribution.

`qqnorm(y)`................. this will plot the quantiles of the data vector *y* against the quantiles of the standard normal distribution. This provides a graphical test of normality. The plot will be linear if *y* comes from a normal distribution but not otherwise. Non-linearity indicates non-normality.

In S the default values for *m* and *s* are zero and one.

THE GAMMA FAMILY

The gamma family of probability distributions has kernel $p(y) \propto y^{\alpha-1}e^{-\beta y}$; $y > 0$; $\alpha, \beta > 0$. The full density function is

$$p(y) = \frac{y^{\alpha-1}e^{-\beta y}}{\Gamma(\alpha)\beta^{-\alpha}} \tag{1.41}$$

where $\Gamma(\alpha)$ is the complete gamma function defined by

$$\Gamma(\alpha) = \int_0^\infty x^{\alpha-1}e^{-x}\,dx, \quad \alpha > 0.$$

The family can be thought of as generalizing the exponential family $\beta e^{-\beta y}$ which is a gamma distribution with $\alpha = 1$. The unit exponential has $\beta = 1$ and has mean and variance equal to one.

The mean and variance of a gamma(α, β) variate are

$$E(Y) = \frac{\alpha}{\beta}; \quad V(Y) = \frac{\alpha}{\beta^2}.$$

A one parameter sub-family of the gamma distributions is the χ^2 distributions which have $\alpha = \nu/2$ and $\beta = 1/2$ where ν is positive and usually an integer. Their density functions are

$$p(y) = \frac{y^{\nu/2-1}e^{-y/2}}{\Gamma(\nu/2)2^{\nu/2}}, \quad y > 0, \ \nu > 0.$$

Relevant S commands follow the same format as for the normal, for example `rgamma(n, a, b)` produces n independent realizations from a gamma distribution with $\alpha = a$ and $\beta = b$. Chi-squared (χ^2) results can be got from `rchisq(n, ν)` etc.

THE BETA FAMILY

The beta densities are continuous on the unit interval and depend on two parameters. Their form is

$$p(x|\alpha, \beta) = \frac{\Gamma(\alpha + \beta)}{\Gamma(\alpha)\Gamma(\beta)} x^{\alpha-1}(1 - x)^{\beta-1}, \qquad \alpha, \beta > 0,\ 0 \le x \le 1. \tag{1.42}$$

When α is a positive integer then $\Gamma(\alpha) = (\alpha - 1)!$. In particular, $\Gamma(1) = 0! = 1$. Since probability density functions integrate to one over the sample space it follows from (1.42) that

$$\int_0^1 x^{\alpha-1}(1 - x)^{\beta-1} dx = \frac{\Gamma(\alpha)\Gamma(\beta)}{\Gamma(\alpha+\beta)}. \tag{1.43}$$

The means and variances of the densities (1.42) are

$$E(X) = \frac{\alpha}{\alpha + \beta} \qquad V(X) = \frac{\alpha\beta}{(\alpha+\beta)^2(\alpha+\beta+1)}. \tag{1.44}$$

The density is symmetrical about $X = 1/2$ if $\alpha = \beta$. If $\alpha = \beta = 1$ it reduces to the uniform distribution.

S commands are again in standard format in which, for example, `rbeta(n, a, b)` produces n independent realizations from a beta distribution with $\alpha = a$ and $\beta = b$.

THE MULTINOMIAL FAMILY

Suppose a vector discrete random variable $Y = (Y_0, Y_1, ..., Y_L)$ is such that Y_j measures the number of occasions that the j'th of $L + 1$ mutually exclusive and exhaustive events occurs in n trials. Thus each of the $\{Y_j\}$ takes values in the set $Y_j = \{0, 1, 2, ..., n\}$ subject to the condition that $\Sigma_{j=0}^{L} Y_j = n$. Then if $p = \{p_0, p_1, ..., p_L\}$ and p_j is the probability that at any one trial the j'th event occurs,

$$p_Y(y|p) = \frac{n!}{y_0!y_1! \ldots y_L!} p_0^{y_0} p_1^{y_1} \times \ldots \times p_L^{y_L} \tag{1.45}$$

$$\text{where} \quad \Sigma_{j=0}^{L} y_j = n;\ y_j \in \{0, 1, 2, ..., n\};\ \Sigma_{j=0}^{L} p_j = 1. \tag{1.46}$$

This is best thought of as the distribution of L random variables since the last one is determined by the condition that they sum to n. The means, variances and covariances of the $\{Y_j\}$ are

$$E(Y_j) = np_j; \quad V(Y_j) = np_j(1 - p_j); \quad C(Y_iY_j) = -np_ip_j.$$

A particular case of the multinomial is the binomial which arises when $L = 1$ so there are two categories and the probability mass function takes the form

$$p_Y(y) = \frac{n!}{(n - y)!y!}p^y(1 - p)^{n-y}, \quad y = 0, 1, ..., n, \quad 0 \le p \le 1. \tag{1.47}$$

And a particular case of the binomial is the Bernoulli family which arises when $n = 1$,

$$p_Y(y) = p^y(1 - p)^{1-y}, \quad y \in \{0, 1\}, \quad 0 \le p \le 1. \tag{1.48}$$

THE DIRICHLET FAMILY

This family generalizes the beta family to a vector $p = (p_0, p_1, ..., p_L)$ in which $\sum_{i=0}^L p_i = 1$ and the $\{p_i\}$ are non-negative. If $\alpha = \sum_{l=0}^L \alpha_l$ the density function takes the form

$$f_P(p) = \frac{\Gamma(\alpha)}{\Gamma(\alpha_0)\Gamma(\alpha_1)\ldots\Gamma(\alpha_L)}p_0^{\alpha_0-1} \times \ldots \times p_L^{\alpha_L-1}, \tag{1.49}$$

$$\text{where} \quad \{p_i\} \ge 0; \quad \textstyle\sum_{i=1}^L p_i = 1; \quad \{\alpha_i\} \ge 0; \quad \sum_{i=0}^L \alpha_i = \alpha.$$

The means, variances and covariances are

$$E(P_j) = \frac{\alpha_j}{\alpha}; \quad V(P_j) = \frac{\alpha_j(\alpha - a_j)}{\alpha^2(\alpha + 1)}; \quad C(P_iP_j) = -\frac{\alpha_i\alpha_j}{\alpha^2(\alpha + 1)}.$$

The beta family emerges as the special case in which $L = 1$. Comparison of (1.49) and (1.45) shows that if the prior for p is dirichlet then the posterior is of the same form. This means that the dirichlet is the natural conjugate prior family for the multinomial.

1.8 BIBLIOGRAPHIC NOTES

For a perspective on econometric analysis not dissimilar to that taken in this book the reader might wish to look at Heckman (2001: 673–748).

Almost all books on Bayesian inference are written by statisticians. The main work specifically devoted to Bayesian inference in econometrics is Zellner (1971). This is a useful book and gives a systematic and detailed account of the main analytical results and so it is necessarily confined to the simpler, mathematically tractable, models. Written well before the computer revolution it is naturally dated, but remains a valuable reference. *Bayesian Analysis in Econometrics and Statistics: The Zellner View and Papers*, edited by Zellner (1997), provides convenient access to more recent work by Zellner and others. A stimulating early book by Leamer (1978) is well worth reading but is currently out of print. Bauwens, Lubrano and Richard (1999) is a useful text on econometrics from a Bayesian perspective with a particular emphasis on time series models, but again it shows traces of being written before the computer revolution.

Though statistics and econometrics differ radically in their point of view they have a great deal in common, particularly computational techniques. Useful sources in increasing order of difficulty are Gelman, Carlin, Stern and Rubin (2003) and, particularly, even though it shows the effect of being written before the coming of fast simulation methods, Berger (1993). Bernardo and Smith (1994) is a major study of the principles of Bayesian inference written from a perspective strongly influenced by de Finetti. It contains an account of reference priors, proposed originally by Bernardo. The proof of convergence of posterior mass functions, theorem 1.2 in this chapter, is based on theirs. Another recent statistics text is Press (2003).

The source text for a subjective view of inference is de Finetti's two volumes, *Theory of Probability* (1974 and 1975), though it is not easy reading. See also Cox (1961). Harold Jeffreys' classic *Theory of Probability* (1966) is available in paperback from the Clarendon Press, Oxford. It is not easy reading, partly for notational reasons. Readers curious about foundational issues might like to read Keynes (1920), particularly chapter 24 which gives a subtle discussion of "objective probability". Frank Ramsey replied, arguing for the subjective view, in Ramsey (1931: 156–98). Another valuable source on foundational issues and the subjective versus objective view of probability is Kyburg and Smokler (1964) which contains lengthy excerpts by many of the major contributors to the debate, including de Finetti. An introductory exposition of Bayesian method from a subjective point is by Edwards, Lindman and Savage (1963: 193–242). This article, although published in a psychological journal, is not especially "psychological" and it is not particularly technical. It is recommended to all those interested in Bayesian methods. It is reprinted in *The Writings of Leonard Jimmie Savage – A Memorial Selection*, published by the American Statistical Association and the Institute of Mathematical Statistics (1981).

There is a persistent interest in "objective" Bayesian methods that involve priors or likelihoods that are in some sense minimally informative. A good source, which is strongly recommended even if you aren't interested in objective methods, is the work of E. T. Jaynes, much of it unpublished and which is best accessed through http:\\omega.math.albany.edu/JaynesBook.html.

Zellner (1986) proposes a variant of the natural conjugate prior that has proved popular.

Lindley and Smith's classic article (1972: 1–14) is about hierarchical prior structures as applied to linear regression models.

Lindley and Novick (1981: 45–58) provide an enlightening discussion both of exchangeability and its relation to similar ideas of R. A. Fisher.

Decision theory is dealt with in many books. A recent and clear introductory exposition may be found in Leonard and Hsu (1999). A classic monograph on Bayesian decision theory is DeGroot (1970). Berger (1993), mentioned above, is another recommended source on the decision theoretic approach. A recent paper by Chamberlain (2000: 255–84) is worth study.

Berger and Wolpert (1988) is a very readable study of fundamental principles of inference and is the main source for this book on the likelihood principle.

On the Bayesian view of identifiability see Kadane (1974: 175–91).

On separating models with multidimensional parameters Lancaster (2000: 391–413) and (2002: 647–60) contain a number of examples.

The paper introducing regression is Galton (1886: 246–63).

Full text copies of all but the most recent *Handbook of Econometrics* chapters can be downloaded from www.elsevier.com/hes/books/02/menu02.htm. This source includes the important Leamer (1983).