1

Introduction to Bayesian Statistics

In the last few years the use of Bayesian methods in the practice of applied statistics has greatly increased. In this book we will show how the development of computational Bayesian statistics is the key to this major change in statistics. For most of the twentieth century, frequentist statistical methods dominated the practice of applied statistics. This is despite the fact that statisticians have long known that the Bayesian approach to statistics offered clear cut advantages over the frequentist approach. We will see that Bayesian solutions are easy in theory, but were difficult in practice. It is easy to find a formula giving the shape of the posterior. It is often more difficult to find the formula of the exact posterior density. Computational Bayesian statistics changed all this. These methods use algorithms to draw samples from the incompletely known posterior and use these random samples as the basis for inference. In Section 1.1 we will look briefly at the the ideas of the frequentist approach to statistics. In Section 1.2 we will introduce the ideas of Bayesian statistics. In Section 1.3 we show the similarities and differences between the likelihood approach to inference and Bayesian inference. We will see that the different interpretations of the parameters and probabilities lead to the advantages of Bayesian statistics.

1.1 THE FREQUENTIST APPROACH TO STATISTICS

In frequentist statistics, the parameter is considered a fixed but unknown value. The sample space is the set of all possible observation values. Probability is interpreted as long-run relative frequency over all values in the sample space given the unknown parameter. The performance of any statistical procedure is determined by averaging

Understanding Computational Bayesian Statistics. By William M. Bolstad

over the sample space. This can be done prior to the experiment and does not depend on the data.

There were two main sources of frequentist ideas. R. A. Fisher developed a theory of statistical inference based on the likelihood function. It has the same formula as the joint density of the sample, however, the observations are held fixed at the values that occurred and the parameter(s) are allowed to vary over all possible values. He reduced the complexity of the data through the use of sufficient statistics which contain all the relevant information about the parameter(s). He developed the theory of maximum likelihood estimators (MLE) and found their asymptotic distributions. He measured the efficiency of an estimator using the Fisher information, which gives the amount of information available in a single observation. His theory dealt with nuisance parameters by conditioning on an ancillary statistic when one is available. Other topics associated with him include analysis of variance, randomization, significance tests, permutation tests, and fiducial intervals. Fisher himself was a scientist as well as a statistician, making great contributions to genetics as well as to the design of experiments and statistical inference. As a scientist, his views on inference are in tune with science. Occam's razor requires that the simplest explanation (chance) must be ruled out before an alternative explanation is sought. Significance testing where implausibility of the chance model is required before accepting the alternative closely matches this view.

Jerzy Neyman and Egon Pearson developed decision theory, and embedded statistical inference in it. Their theory was essentially deductive, unlike Fisher's. They would determine criteria, and try to find the optimum solution in the allowed class. If necessary, they would restrict the class until they could find a solution. For instance, in estimation, they would decide on a criterion such as minimizing squared error. Finding that no uniformly minimum squared error estimator exists, they would then restrict the allowed class of estimators to unbiased ones, and find uniformly minimum variance unbiased estimators (UMVUE). Wald extended these ideas by defining a loss function, and then defining the risk as the expected value of the loss function averaged over the sample space. He then defined as inadmissible any decision rule that is dominated by another for all values of the parameter. Any rule that is not inadmissible is admissible. Unexpectedly, since he was using frequentist criteria, he found that the class of admissible rules is the class of Bayesian rules. Other topics in this school include confidence intervals, uniformly most powerful tests of hypothesis, uniformly most powerful unbiased tests, and James-Stein estimation.

The disputes Fisher had with the Neyman are legendary (Savage, 1976). Fisher strongly opposed the submerging of inference into decision theory and Neyman's denial that inference uses inductive logic. His specific criticisms about the Neyman-Pearson methods include:

- Unbiased estimators are not invariant under one-to-one reparameterizations.
- Unbiased estimators are not compatible with the likelihood principle.
- Unbiased estimates are not efficient. He scathingly criticized this waste of information as equivalent to throwing away observations.

Nevertheless, what currently passes for frequentist parametric statistics includes a collection of techniques, concepts, and methods from each of these two schools, despite the disagreements between the founders. Perhaps this is because, for the very important cases of the normal distribution and the binomial distribution, the MLE and the UMVUE coincided. Efron (1986) suggested that the emotionally loaded terms (unbiased, most powerful, admissible, etc.) contributed by Neyman, Pearson, and Wald reinforced the view that inference should be based on likelihood and this reinforced the frequentist dominance. Frequentist methods work well in the situations for which they were developed, namely for exponential families where there are minimal sufficient statistics. Nevertheless, they have fundamental drawbacks including:

- Frequentist statistics have problems dealing with nuisance parameters, unless an ancillary statistic exists.
- Frequentist statistics gives prior measures of precision, calculated by sample space averaging. These may have no relevance in the post-data setting.

Inference based on the likelihood function using Fisher's ideas is essentially constructive. That means algorithms can be found to construct the solutions. Efron (1986) refers to the MLE as the "original jackknife" because it is a tool that can easily be adapted to many situations. The maximum likelihood estimator is invariant under a one-to-one reparameterization. Maximum likelihood estimators are compatible with the likelihood principle. Frequentist inference based on the likelihood function has some similarities with Bayesian inference as well as some differences. These similarities and differences will be explored in Section 3.3.

1.2 THE BAYESIAN APPROACH TO STATISTICS

Bayesian statistics is based on the theorem first discovered by Reverend Thomas Bayes and published after his death in the paper *An Essay Towards Solving a Problem in the Doctrine of Chances* by his friend Richard Price in *Philosophical Transactions of the Royal Society*. Bayes' theorem is a very clever restatement of the conditional probability formula. It gives a method for updating the probabilities of unobserved events, given that another related event has occurred. This means that we have a prior probability for the unobserved event, and we update this to get its posterior probability, given the occurrence of the related event. In Bayesian statistics, Bayes' theorem is used as the basis for inference about the unknown parameters of a statistical distribution. Key ideas forming the basis of this approach include:

- Since we are uncertain about the true values of the parameters, in Bayesian statistics we will consider them to be random variables. This contrasts with the frequentist idea that the parameters are fixed but unknown constants. Bayes' theorem is an updating algorithm, so we must have a prior probability distribution that measures how plausible we consider each possible parameter value before looking at the data. Our prior distribution must be subjective, because

somebody else can have his/her own prior belief about the unknown values of the parameters.

- Any probability statement about the parameters must be interpreted as "degree of belief."
- We will use the rules of probability directly to make inferences about the parameters, given the observed data. Bayes' theorem gives our posterior distribution, which measures how plausible we consider each possible value after observing the data.
- Bayes' theorem combines the two sources of information about the unknown parameter value: the prior density and the observed data. The prior density gives our relative belief weights of every possible parameter value before we observe the data. The *likelihood function* gives the relative weights to every possible parameter value that comes from the observed data. Bayes' theorem combines these into the posterior density, which gives our relative belief weights of the parameter value after observing the data.

Bayes' theorem is the only consistent way to modify our belief about the parameters given the data that actually occurred. A Bayesian inference depends only on the data that occurred, not on the data that could have occurred but did not. Thus, Bayesian inference is consistent with the *likelihood principle*, which states that if two outcomes have proportional likelihoods, then the inferences based on the two outcomes should be identical. For a discussion of the likelihood principle see Bernardo and Smith (1994) or Pawitan (2001). In the next section we compare Bayesian inference with likelihood inference, a frequentist method of inference that is based solely on the likelihood function. As its name implies, it also satisfies the likelihood principle.

A huge advantage of Bayesian statistics is that the posterior is always found by a single method: Bayes' theorem. Bayes' theorem combines the information about the parameters from our prior density with the information about the parameters from the observed data contained in the likelihood function into the posterior density. It summarizes our knowledge about the parameter given the data we observed.

Finding the posterior: easy in theory, hard in practice

Bayes' theorem is usually expressed very simply in the unscaled form, *posterior proprotional to prior times likelihood*:

$$g(\theta_1, \ldots, \theta_p | y_1, \ldots, y_n) \propto g(\theta_1, \ldots, \theta_p) \times f(y_1, \ldots, y_n | \theta_1, \ldots, \theta_p). \quad (1.1)$$

This formula does not give the posterior density $g(\theta_1, \ldots, \theta_p | y_1, \ldots, y_n)$ exactly, but it does give its shape. In other words, we can find where the modes are, and relative heights at any two locations. However, it cannot be used to find probabilities or to find moments since it is not a density. We can't use it for inferences. The actual posterior density is found by scaling it so it integrates to 1:

$$g(\theta_1, \ldots, \theta_p | y_1, \ldots, y_n) = \frac{g(\theta_1, \ldots, \theta_p) \times f(y_1, \ldots, y_n | \theta_1, \ldots, \theta_p)}{K} \quad (1.2)$$

where the divisor needed to make this a density is

$$K = \int \dots \int g(\theta_1, \dots, \theta_p) \times f(y_1, \dots, y_n | \theta_1, \dots, \theta_p)\, d\theta_1 \dots d\theta_p\,. \qquad (1.3)$$

A closed form for the p-dimensional integral only exists for some particular cases.[1] For other cases the integration has to be done numerically. This may be very difficult, particularly when p, the number of parameters, is large. When this is true, we say there is a high dimensional parameter space.

Finding the posterior using Bayes' theorem is easy in theory. That is, we can easily find the unscaled posterior by Equation 1.1. This gives us all the information about the shape of the posterior. The exact posterior is found by scaling this to make it a density and is given in Equation 1.2. However, in practice, the integral given in Equation 1.3 can be very difficult to evaluate, even numerically. This is particularly difficult when the parameter space is high dimensional. Thus we cannot always find the divisor needed to get the exact posterior density. In general, the incompletely known posterior given by Equation 1.1 is all we have.

In Bayesian statistics we do not have to assume that the observations come from an easily analyzed distribution such as the normal. Also, we can use any shape prior density. The posterior would be found the same way. Only the details of evaluating the integral would be different.

Example 1 *A random sample $y_1, \dots, y_n$ is drawn from a distribution having an unknown mean and known standard deviation. Usually, it is assumed the draws come from a normal (μ, σ^2) distribution. However, the statistician may think that for this data, the observation distribution is not normal, but from another distribution that is also symmetric and has heavier tails. For example, the statistician might decide to use the Laplace(a, b) distribution with the same mean μ and variance σ^2. The mean and variance of the Laplace(a, b) distribution are given by a and $2b^2$, respectively. The observation density is given by*

$$f(y|\mu, \sigma) = \frac{1}{\sqrt{2}\sigma} e^{-\frac{|y-\mu|}{\sigma/\sqrt{2}}}\,.$$

The posterior distribution is found by the same formula

$$g(\mu|y_1, \dots, y_n) = \frac{g(\mu) \times f(y_1, \dots, y_n|\mu)}{\int g(\mu) \times f(y_1, \dots, y_n|\mu) d\mu}\,.$$

The only difference would be the details of the integration. In most cases, it would have to be done numerically.

[1]Where the observation distribution comes from an exponential family, and the prior comes from the family that is conjugate to the observation distribution.

1.3 COMPARING LIKELIHOOD AND BAYESIAN APPROACHES TO STATISTICS

In this section we graphically illustrate the similarities and differences between the likelihood and Bayesian approaches to inference; specifically, how a parameter is estimated using each of the approaches. We will see that:

1. The likelihood and the posterior density are found in a similar manner, by cutting a surface with the same vertical hyperplane. However, the surfaces used in the two approaches have different interpretations and in most cases they will have different shapes.

2. Even when the surfaces are the same (when flat priors are used) the estimators are chosen to satisfy different criteria.

3. The two approaches have different ways of dealing with nuisance parameters.

The observation(s) come from the observation density $f(y|\theta)$ where θ is the fixed parameter value. It gives the probability density over all possible observation values for the given value of the parameter. The parameter space, Θ, is the set of all possible parameter values. The parameter space ordinarily has the same dimension as the total number of parameters, p. The sample space, S, is the set of all possible values of the observation(s). The dimension of the sample space is the number of observations n. Many of the commonly used observation distributions come from the one-dimensional exponential family of distributions. When we are in the one-dimensional exponential family, the sample space may be reduced to a single dimension due to the single sufficient statistic.

The Inference Universe

We define the inference universe of the problem to be the Cartesian product of the parameter space and the sample space. It is the $p + n$ dimensional space where the first p dimensions are the parameter space, and the remaining n dimensions are the sample space. We do not ever observe the parameter, so the position in those coordinates is always unknown. However, we do observe the sample, so we know the last n coordinates.

We will let the dimensions be $p = 1$ and $n = 1$ for illustrative purposes. This is the case when we have a single parameter and a single observation (or we have a random sample of observations from a one-dimensional exponential family). The inference universe has two dimensions. The vertical dimension is the parameter space and is unobservable. The horizontal dimension is the sample space and is observable. We wish to make inference about where we are in the vertical dimension given that we know where we are in the horizontal dimension.

Let $f(y|\theta)$ be the observation density. For each value of the parameter θ, it gives the probability density of the observation y for that parameter value. Actually, this formula is a function of both the value of the observation and the parameter value. It

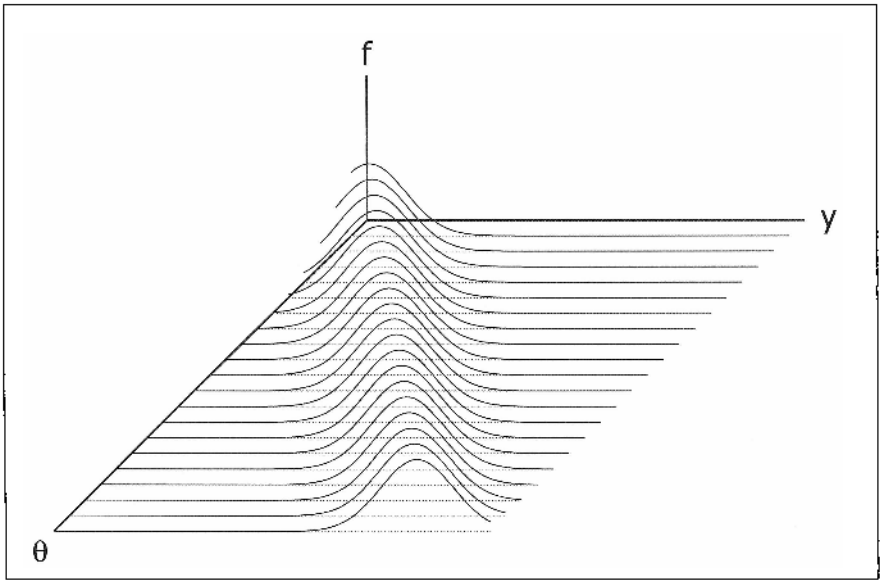

Figure 1.1 The observation density surface in 3D perspective.

is defined for all points in the inference universe, thus it forms a surface defined on the inference universe. It forms a probability density in the observation dimension for each particular value in the parameter dimension. However in general, it is not a probability distribution in the parameter dimension. Figure 1.1 shows the observation density surface in 3D perspective.

The likelihood function, first defined by R. A. Fisher (1922), has the same functional form as the observation density, only y is held at the observed value, and θ is allowed to vary over all possible values. Thus, it is a function of the parameter θ. It is found by cutting the observation density surface with a vertical plane parallel to the θ axis through the observed value. This is shown in Figure 1.2. Likelihood inference is based entirely on the likelihood function.

Maximum Likelihood Estimation

We are trying to choose an estimator (function of the observations) to represent the unknown value of the parameter. In likelihood inference, the likelihood function cannot be considered to be a probability density in general. Because of this, Fisher (1922) decided that the best way to estimate the parameter is to choose the parameter value that has the highest value of the likelihood function, i.e., its mode. This is the parameter value that gives the observed data the highest probability. He named this the *maximum likelihood estimator* (MLE). The mode will be invariant under any

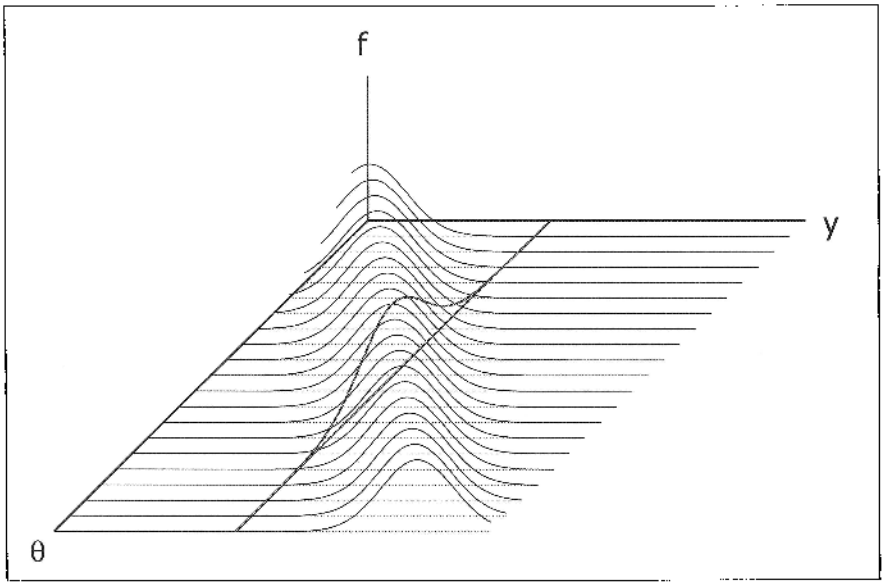

Figure 1.2 Observation density surface with the likelihood function shown in 3D perspective.

one-to-one transformation of the parameter space. Hence, the MLE will be invariant under any reparameterization of the problem.[2]

Bayesian Estimation

Bayesian estimation requires that we have a probability distribution defined on the parameter space before we look at the data. It is called the *prior* density because it gives our belief weights for each of the possible parameter values before we see the data. This requires that we allow a different interpretation of probability on the parameter space than on the sample space. It is measuring our belief, and thus is subjective. The probability on the sample space has the usual long-run relative frequency interpretation. The prior density of the parameter is shown with the observation density surface in Figure 1.3. The joint density of the parameter and the observation is found by multiplying each value of the observation density surface by the corresponding height of the prior density. This is shown in Figure 1.4. Bayesians call joint density of the parameters and the observation "the full Bayesian model." It is clear that the full Bayesian model surface will not be the same shape as the sampling surface unless we use a flat prior that gives all possible parameter values equal weight. To find the posterior density of the parameter given the observed value we cut the

[2]Fisher was well aware of Bayes' theorem, and wanted his method to work on the same type of problems. He viewed Bayes' use of flat prior to be very arbitrary, and realized that the Bayesian estimator would not be invariant under the reparameterization.

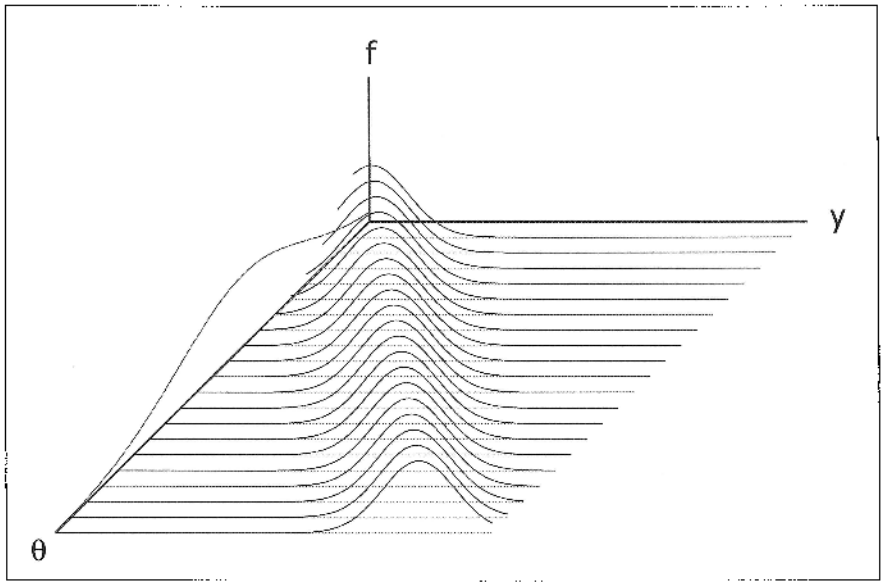

Figure 1.3 Prior density and observation density surface in 3D perspective.

joint density of the parameter and the observation with a vertical plane parallel to the parameter axis through the observed value of y. Thus, the likelihood and the posterior density are found by cutting different surfaces with the same hyperplane. The posterior density is shown in 3D perspective in Figure 1.5. The posterior density is the complete inference in the Bayesian approach. It summarizes the belief we can have about all possible parameter values, given the observed data. The posterior will always be a probability density, conditional on the observed data. Because of this, we can use the mean of the posterior distribution as the estimate of the parameter. The mean of a distribution is the value that minimizes the mean-squared deviation. Hence, the Bayesian posterior mean is the estimator that minimizes the mean-squared deviation of the posterior distribution.[3]

The Likelihood Function Can Be a Posterior

If we decide to use a flat prior density that gives equal weight to all values of the parameter, the joint density on the inference universe will be the same as the observation density surface. This is shown in Figure 1.6. Note that this prior density will be improper (the integral over the whole range will be infinite) unless the parameter values have finite lower and upper bounds. When the prior is improper, we do not have a joint probability density for the full Bayesian model. However,

[3]In decision theory, this means the posterior mean is the optimal estimator when we are using a *squared error* loss function. We can find the optimal Bayesian estimator for any particular loss function. For example, the posterior median is the optimal estimator when we are using an *absolute value* loss function.

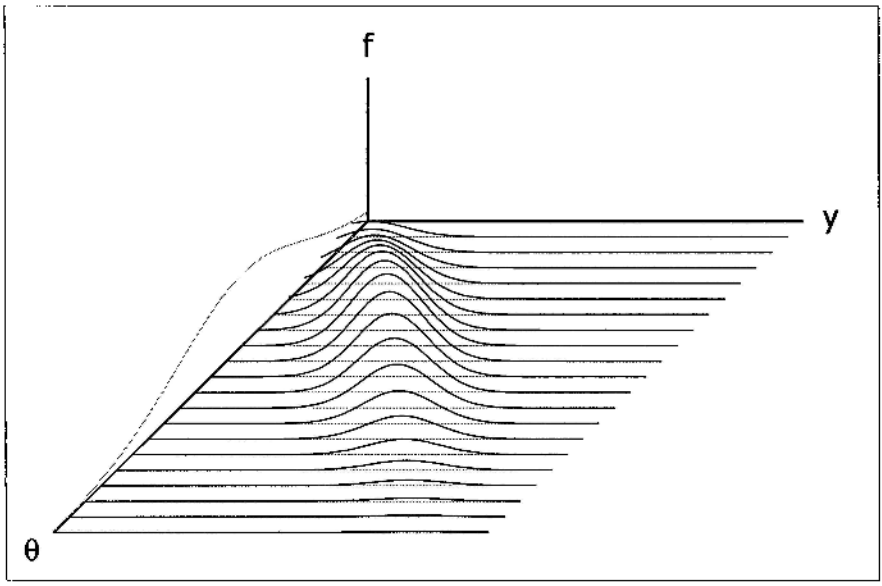

Figure 1.4 The joint density of parameter and observation in 3D perspective. The prior density of the parameter is shown on the margin.

the normed likelihood (likelihood function divided by its integral over the whole range of parameter values) will usually be a probability density. Thus, the likelihood function will have the same shape as the posterior density in this case. The Bayesian posterior mean estimator would be the mean value (balance point) of the likelihood function. This would not generally be the same value as the maximum likelihood estimator, unless the likelihood function is symmetric and unimodal such as in the *normal* likelihood. Figure 1.7 illustrates the difference between these estimators on a nonsymmetric likelihood function that could also be considered a Bayesian posterior density with a flat prior density. The maximum likelihood estimator is the mode of this curve, while the Bayesian posterior estimate is its mean, the balance point. This shows the two estimators are based on different ideas, even when the likelihood function and the posterior density have the same shape.[4]

Note; we are not advocating always using flat priors. We only want to illustrate that when we do, the posterior will be the same shape as the likelihood. Hence, the likelihood can be thought of as an unscaled posterior when we have used flat priors. When the integral of the flat prior over its whole range is infinite, the flat prior will be improper. Despite this, the resulting posterior which is the same shape as the likelihood will usually be proper. For many models, such as the regression-type models that we will discuss in Chapters 8 and 9, it is ok to use improper flat priors.

[4]Jaynes and Bretthorst (2000) show that the maximum likelihood estimator implies that we are using a 0:1 loss function, 0 at the true value of θ, and 1 at every other value. This means getting it exactly right is everything, and getting it close is of no value.

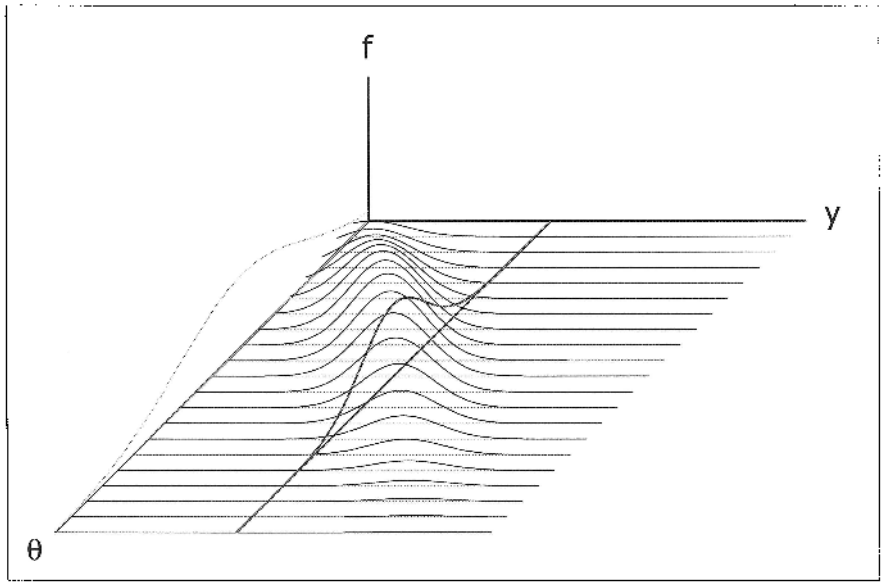

Figure 1.5 Posterior density of the parameter in the inference universe. The prior density is shown in the margin.

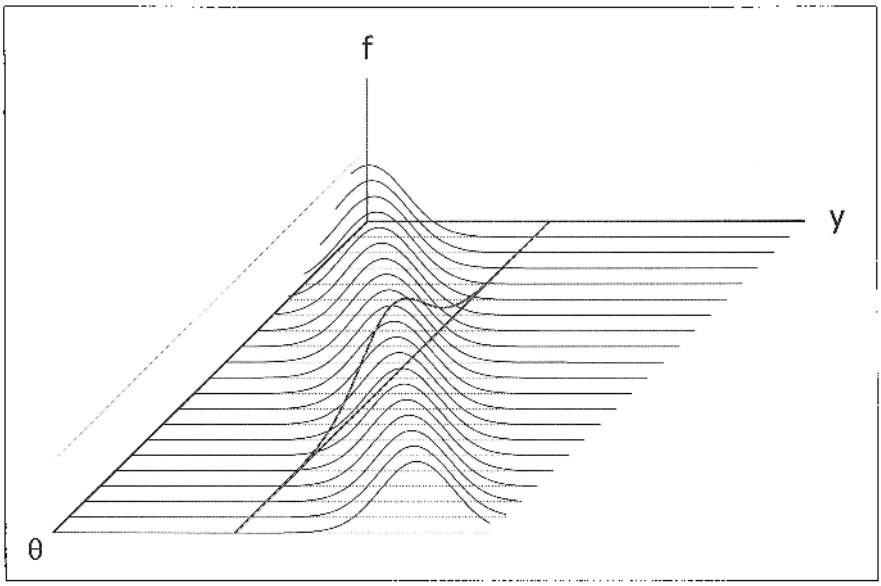

Figure 1.6 Posterior density in the inference universe using a flat prior. It has the same shape as the likelihood function.

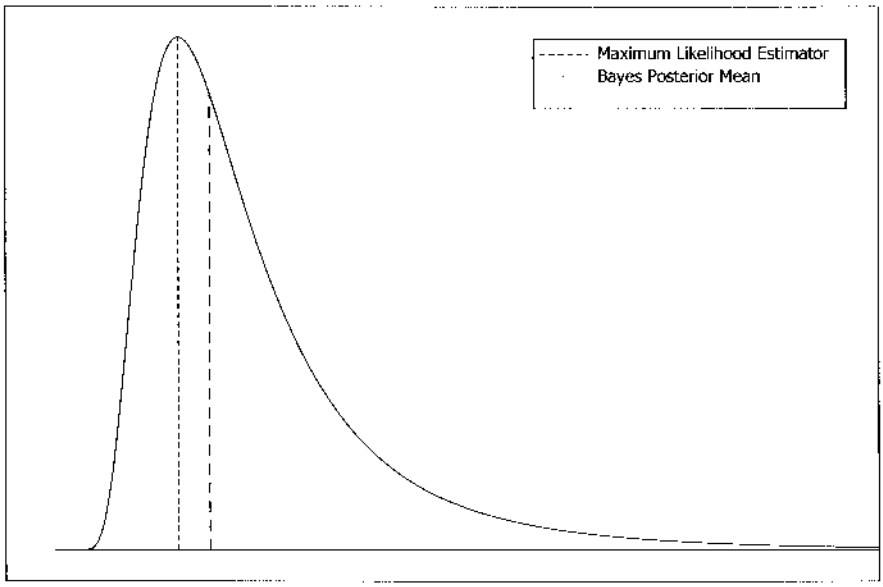

Figure 1.7 Maximum likelihood estimator and Bayesian posterior estimator for a non-symmetric likelihood (posterior with flat prior).

However, there are situations such as when we have a hierarchical normal model where improper priors should not be used for variance components. This will be discussed more fully in Chapter 10.

Multiple Parameters

When we have $p \geq 2$ the same ideas hold. However, we cannot project the surface defined on the inference universe down to a two-dimensional graph. With multiple parameters, Figures 1.1, 1.2, 1.3, 1.4, 1.5, and 1.6 can be considered to be schematic diagrams that represent the ideas rather than exact representations.

We will use the two-parameter case to show what happens when there are multiple parameters. The inference universe has at least four dimensions, so we cannot graph the surface on it. The likelihood function is still found by cutting through the surface with a hyperplane parallel to the parameter space passing through the observed values. The likelihood function will be defined on the the two parameter dimensions as the observations are fixed at the observed values and do not vary. We show the bivariate likelihood function in 3D perspective in Figure 1.8. In this example, we have the likelihood function where θ_1 is the mean and θ_2 is the variance for a random sample from a normal distribution. We will also use this same curve to illustrate the Bayesian posterior since it would be the joint posterior if we use independent flat priors for the two parameters.

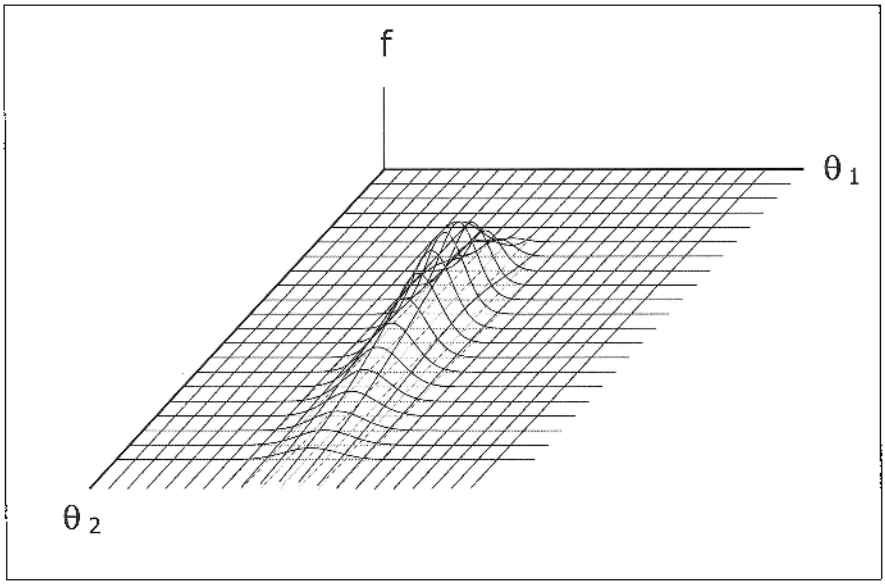

Figure 1.8 Joint Likelihood Function. Note: this can also be considered the joint posterior density when independent flat priors are used for θ_1 and θ_2.

Inference in the Presence of Nuisance Parameters

Sometimes, only one of the parameters is of interest to us. We don't want to estimate the other parameters and call them "nuisance" parameters. All we want to do is make sure the nuisance parameters don't interfere with our inference on the parameter of interest. Because using the Bayesian approach the joint posterior density is a probability density, and using the likelihood approach the joint likelihood function is not a probability density, the two approaches have different ways of dealing with the nuisance parameters. This is true even if we use independent flat priors so that the posterior density and likelihood function have the same shape.

Likelihood Inference in the Presence of Nuisance Parameters

Suppose that θ_1 is the parameter of interest, and θ_2 is a nuisance parameter. If there is an ancillary[5] sufficient statistic, conditioning on it will give a likelihood that only depends on θ_1, the parameter of interest, and inference can be based on that conditional likelihood. This can only be true in certain exponential families, so is of limited general use when nuisance parameters are present. Instead, likelihood

[5]Function of the data that is independent of the parameter of interest. Fisher developed ancillary statistics as a way to make inferences when nuisance parameters are present. However, it only works in the exponential family of densities so it cannot be used in the general case. See Cox and Hinkley (1974).

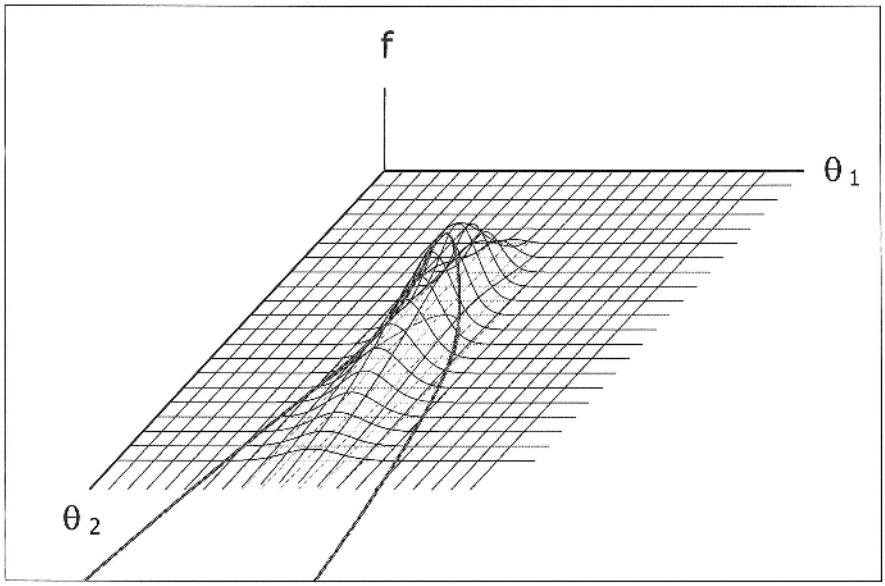

Figure 1.9 Profile likelihood of θ_1 in 3D.

inference on θ_1 is often based on the *profile likelihood* function given by

$$L_p(\theta_1; data) = \sup_{\theta_2|\theta_1} L(\theta_1, \theta_2; data)$$

where $L(\theta_1, \theta_2; data)$ is the joint likelihood function. Essentially, the nuisance parameter has been eliminated by plugging $\hat{\theta}_2|\theta_1$, the conditional maximum likelihood value of θ_2 given θ_1, into the joint likelihood. Hence

$$L_p(\theta_1; data) = L(\theta_1, \hat{\theta}_2|\theta_1; data)\,.$$

The profile likelihood function of θ_1 is shown in three-dimensional space in Figure 1.9. The two-dimensional profile likelihood function is found by projecting it back to the $f \times \theta_1$ plane and is shown in Figure 1.10. (It is like the "shadow" the curve $L(\theta_1, \hat{\theta}_2|\theta_1, data)$ would project on the $f \times \theta_1$ plane from a light source infinitely far away in the θ_2 direction.) The profile likelihood function may lose some information about θ_1 compared to the joint likelihood function. Note that the maximum profile likelihood value of θ_1 will be the same as its maximum likelihood value. However, confidence intervals based on profile likelihood may not be the same as those based on the joint likelihood.

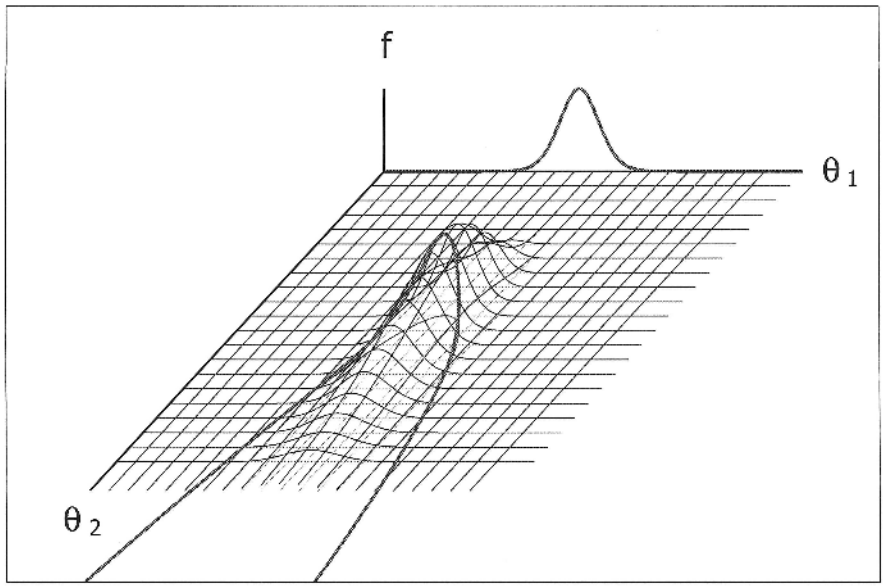

Figure 1.10 Profile likelihood of θ_1 projected onto $f \times \theta_1$ plane.

Bayesian Inference in the Presence of Nuisance Parameters

Bayesian statistics has a single way of dealing with nuisance parameters. Because the joint posterior is a probability density in all dimensions, we can find the marginal densities by integration. Inference about the parameter of interest θ_1 is based on the marginal posterior $g(\theta_1|data)$, which is found by integrating the nuisance parameter θ_2 out of the joint posterior, a process referred to as *marginalization*:

$$g(\theta_1|data) = \int g(\theta_1, \theta_2|data)\, d\theta_2 .$$

Note: we are using independent flat priors for both θ_1 and θ_2, so the joint posterior is the same shape as the joint likelihood in this example. The marginal posterior density of θ_1 is shown on the $f \times \theta_1$ plane in Figure 1.11. It is found by integrating θ_2 out of the joint posterior density. (This is like sweeping the probability in the joint posterior in a direction parallel to the θ_2 axis into a vertical pile on the $f \times \theta_1$ plane.) The marginal posterior has all the information about θ_1 that was in the joint posterior.

The Bayesian posterior estimator for θ_1 found from the marginal posterior will be the same as that found from the joint posterior when we are using the posterior mean as our estimator. For this example, the Bayesian posterior density of θ_1 found by marginalizing θ_2 out of the joint posterior density, and the profile likelihood function of θ_1 turn out to have the same shape. This will not always be the case. For instance, suppose we wanted to do inference on θ_2, and regarded θ_1 as the nuisance parameter.

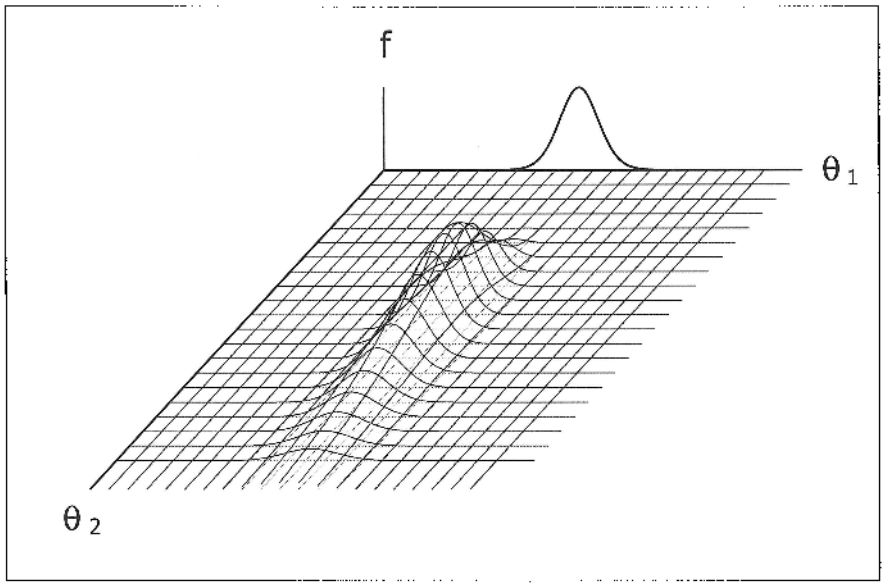

Figure 1.11 Marginal posterior of θ_1.

We have used independent flat priors for both parameters, so the joint posterior has the same shape as the joint likelihood. The profile likelihood of θ_2 is shown in 3D perspective in Figure 1.12 and projected onto the $f \times \theta_2$ plane in Figure 1.13. The marginal posterior of θ_2 is shown in Figure 1.14.

Figure 1.15 shows both the profile likelihood function and the marginal posterior density in 2D for θ_2 for this case. Clearly they have different shapes despite coming from the same two-dimensional function.

Conclusion

We have shown that both the likelihood and Bayesian approach arise from surfaces defined on the inference universe, the observation density surface and the joint probability density respectively. The sampling surface is a probability density only in the observation dimensions, while the joint probability density is a probability density in the parameter dimensions as well (when proper priors are used). Cutting these two surfaces with a vertical hyperplane that goes through the observed value of the data yields the likelihood function and the posterior density that are used for likelihood inference and Bayesian inference, respectively.

In likelihood inference, the likelihood function is not considered a probability density, while in Bayesian inference the posterior always is. The main differences between these two approaches stem from this interpretation difference; certain ideas arise naturally when dealing with a probability density. There is no reason to use the

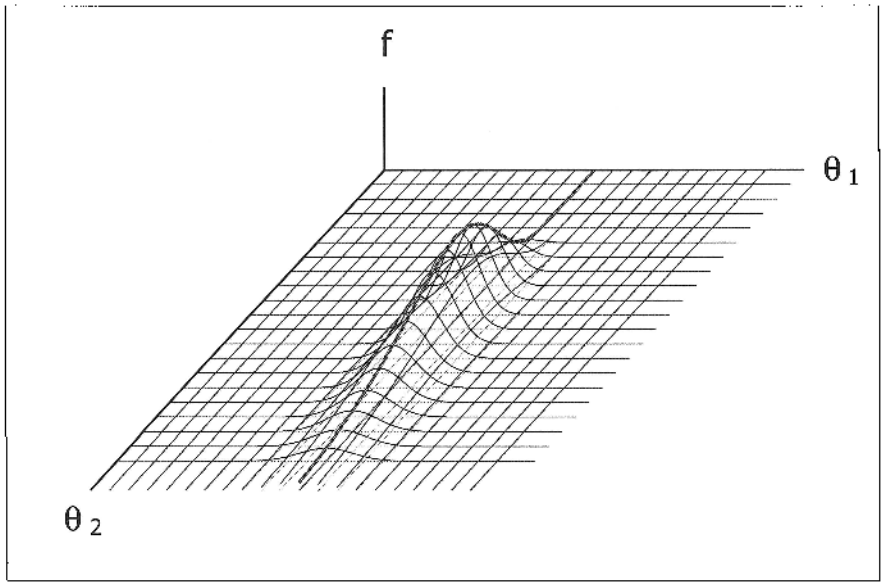

Figure 1.12 Profile likelihood function of θ_2 in 3D.

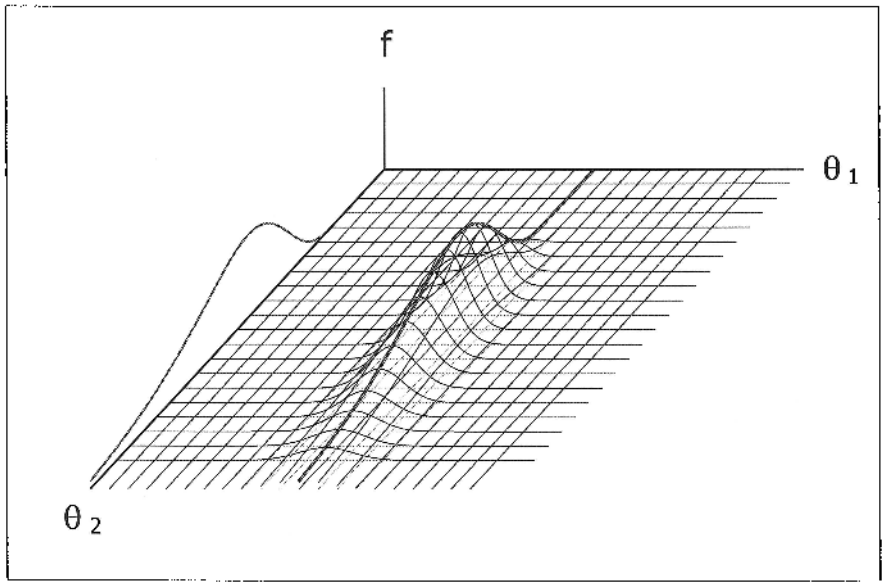

Figure 1.13 Profile likelihood of θ_2 projected onto $f \times \theta_2$ plane.

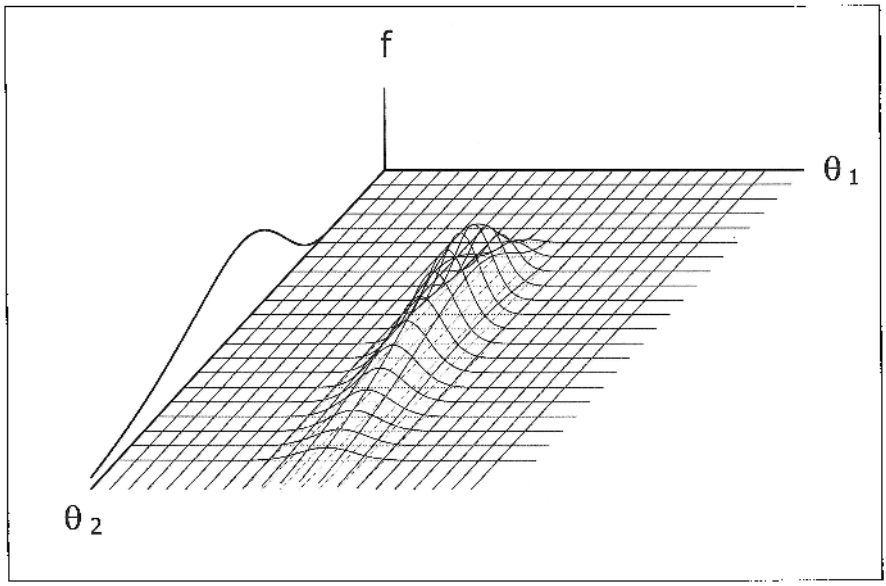

Figure 1.14 Marginal posterior of θ_2.

first moment of the likelihood function without the probability interpretation. Instead, the maximum likelihood estimator is the value that gives the highest value on the likelihood function. When a flat prior is used, the posterior density has the same shape as the likelihood function. Under the Bayesian approach it has a probability interpretation, so the posterior mean will be the estimator since it minimizes the mean squared deviation.

When there are nuisance parameters, there is no reason why they could not be integrated out of the joint likelihood function, and the inference be based on the marginal likelihood. However, without the probability interpretation on the joint likelihood, there is no compelling reason to do so. Instead, likelihood inference is commonly based on the profile likelihood function, where the maximum conditional likelihood values of the nuisance parameters given the parameters of interest are plugged into the joint likelihood. This plug-in approach does not allow for all the uncertainty about the nuisance parameters. It treats them as if it were known to have their conditional maximum likelihood values, rather than treating them like unknown parameters. This may lead to confidence intervals that are too short to have the claimed coverage probability. Under the Bayesian approach the joint posterior density is clearly a probability density. Hence Bayesian inference about the parameter of interest will be based on the marginal posterior where the nuisance parameters have been integrated out. The Bayesian approach has allowed for all the uncertainty about the nuisance parameters.

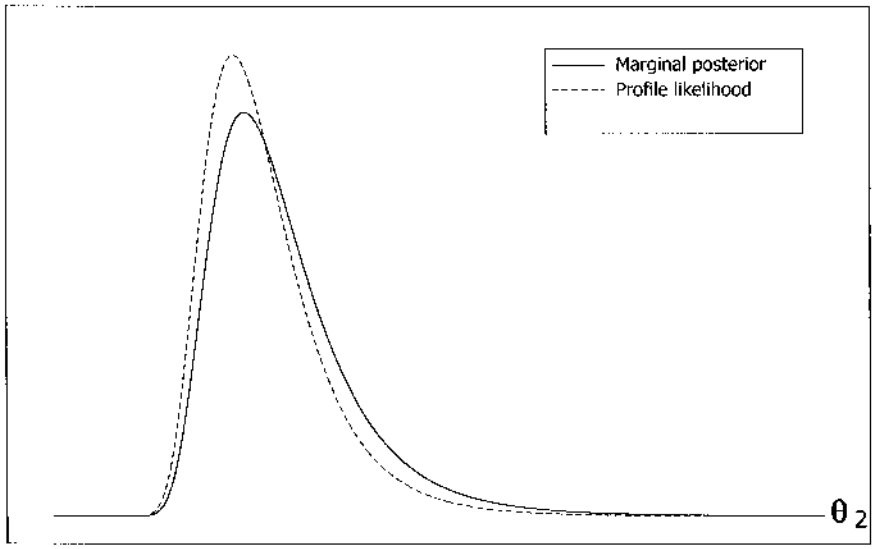

Figure 1.15 Profile likelihood and marginal posterior density of θ_2.

1.4 COMPUTATIONAL BAYESIAN STATISTICS

In this section we introduce the main ideas of computational Bayesian statistics. We show how basing our inferences on a random sample from the posterior distribution has overcome the main impediments to using Bayesian methods. The first impediment is that the exact posterior cannot be found analytically except for a few special cases. The second is that finding the numerical posterior requires a difficult numerical integration, particularly when there is a large number of parameters.

Finding the posterior using Bayes' theorem is easy in theory. That is, we can easily find the unscaled posterior by Equation 1.1. This gives us all the information about the shape of the posterior. The exact posterior is found by scaling this to make it a density and is given in Equation 1.2. However, in practice, the integral given in Equation 1.3 can be very difficult to evaluate, even numerically. This is particularly difficult when the parameter space is high dimensional. Thus we cannot always find the divisor needed to get the exact posterior density. In general, the incompletely known posterior given by Equation 1.1 is all we have.

Computational Bayesian statistics is based on developing algorithms that we can use to draw samples from the true posterior, even when we only know the unscaled version. There are two types of algorithms we can use to draw a sample from the true posterior, even when we only know it in the unscaled form. The first type are direct methods, where we draw a random sample from an easily sampled density, and reshape this sample by only accepting some of the values into the final sample, in such a way that the accepted values constitute a random sample from the posterior. These methods quickly become inefficient as the number of parameters increase.

The second type is where we set up a Markov chain that has the posterior as its long-run distribution, and letting the chain run long enough so a random draw from the Markov chain is a random draw from the posterior. These are known as Markov chain Monte Carlo (MCMC) methods. The Metropolis-Hastings algorithm and the Gibbs sampling algorithm are the two main Markov chain Monte Carlo methods. The Markov chain Monte Carlo samples will not be independent. There will be serial dependence due to the Markov property. Different chains have different mixing properties. That means they move around the parameter space at different rates. We show how to determine how much we must thin the sample to obtain a sample that well approximates a random sample from the posterior to be used for inference.

Inference from the Posterior Sample

The overall goal of Bayesian inference is knowing the posterior. The fundamental idea behind nearly all statistical methods is that as the sample size increases, the distribution of a random sample from a population approaches the distribution of the population. Thus, the distribution of the random sample from the posterior will approach the true posterior distribution. Other inferences such as point and interval estimates of the parameters can be constructed from the posterior sample. For example, if we had a random sample from the posterior, any parameter could be estimated by the corresponding statistic calculated from that random sample. We could achieve any required level of accuracy for our estimates by making sure our random sample from the posterior is large enough. Existing exploratory data analysis (EDA) techniques can be used on the sample from the posterior to explore the relationships between parameters in the posterior.

1.5 PURPOSE AND ORGANIZATION OF THIS BOOK

The development and implementation of computational methods for drawing random samples from the incompletely known posterior has revolutionized Bayesian statistics. Computational Bayesian statistics breaks free from the limited class of models where the posterior can be found analytically. Statisticians can use observation models, and choose prior distributions that are more realistic, and calculate estimates of the parameters from the Monte Carlo samples from the posterior. Computational Bayesian methods can easily deal with complicated models that have many parameters. This makes the advantages that the Bayesian approach offers accessible to a much wider class of models.

This book aims to introduce the ideas of computational Bayesian statistics to advanced undergraduate and first-year graduate students. Students should enter the course with some knowledge in Bayesian statistics at the level of Bolstad (2007). This book builds on that background. It aims to give the reader a big-picture overview of the methods of computational Bayesian statistics, and to demonstrate them for some common statistical applications.

In Chapter 2, we look at methods which allow us to draw a random sample directly from an incompletely known posterior distribution. We do this by drawing a random sample from an easily sampled distribution, and only accepting some draws into the final sample. This reshapes the accepted sample so it becomes a random sample from the incompletely known posterior. These methods work very well for a single parameter. However, they can be seen to become very inefficient as the number of parameters increases. The main use of these methods is as a small step that samples a single parameter as part of a Gibbs sampler.

Chapter 3 compares Bayesian inferences drawn from a numerical posterior with Bayesian inferences from the posterior random sample.

Chapter 4 reviews Bayesian statistics using conjugate priors. These are the classical models that have analytic solutions to the posterior. In computational Bayesian statics, these are useful tools for drawing an individual parameter as steps of the Markov chain Monte Carlo algorithms.

In Chapter 5, we introduce Markov chains, a type of process that evolves as time passes. They move through a set of possible values called *states* according to a probabilistic law. The set of all possible values is called the *state space* of the chain. Markov chains are a particular type of stochastic process where the probability of the next state given the past history of the process up to and including the current state only depends on the current state, that is, it is memoryless. The future state depends only on the current state and the past states can be forgotten. We study the relationship between the probabilistic law of the process and the long-run distribution of the chain. Then we see how to solve the inverse problem. In other words, we find a probabilistic law that will give a desired long-run distribution.

In Chapter 6, we introduce the Metropolis-Hastings algorithm as the fundamental method for finding a Markov chain that has the long-run distribution that is the same shape as the posterior. This achieves the fundamental goal in computational Bayesian statistics. We can easily find the shape of the posterior by the proportional form of Bayes' theorem. The Metropolis-Hastings algorithm gives us a way to find a Markov chain that has that shape long-run distribution. Then by starting the chain, and letting it run long enough, a draw from the chain is equivalent to a draw from the posterior. Drawing samples from the posterior distribution this way is known as Markov chain Monte Carlo sampling. We have lots of choices in setting up the algorithm. It can be implemented using either random-walk or independent candidate densities. We can either draw a multivariate candidate for all parameters at once, or draw the candidates blockwise, each conditional on the parameters in the other blocks. We will see that the Gibbs sampling algorithm is a special case of the Metropolis-Hastings algorithm. These algorithms replace very difficult numerical calculations with the easier process of drawing random variables using the computer. Sometimes, particularly for high-dimensional cases, they are the only feasible method.

In Chapter 7 we develop a method for finding a Markov chain that has good mixing properties. We will use the Metropolis-Hastings algorithm with heavy-tailed independent candidate density. We then discuss the problem of statistical inference on the sample from the Markov chain. We want to base our inferences on an approximately random sample from the posterior. This requires that we determine

the burn-in time and the amount of thinning we need to do so the thinned sample will be approximately random.

In Chapter 8 we show how to do computational Bayesian inference on the logistic regression model. Here we have independent observations from the *binomial* distribution where each observation has its own probability of success. We want to relate the probability of success for an observation to known values of the predictor variables taken for that observation. Probability is always between 0 and 1, and a linear function of predictor variables will take on all possible values from $-\infty$ to ∞. Thus we need a link function of the probability of success so that it covers the same range as the linear function of the predictors. The logarithm of the odds ratio

$$\log_e\left(\frac{\pi}{1-\pi}\right)$$

gives values between $-\infty$ to ∞ so it is a satisfactory link function. It is called the *logit* link function. The logistic regression model is a generalized linear model, so we can find the vector of maximum likelihood estimates, along with their matched curvature covariance matrix, by using iteratively reweighted least squares. We approximate the likelihood by a multivariate normal having the maximum likelihood vector as its mean vector, and the matched curvature covariance. We can find the approximate normal posterior by the simple normal updating rules for normal linear regression. We develop a heavy-tailed candidate density from the approximate normal posterior that we use in the Metropolis-Hastings algorithm to find a sample from the exact posterior. The computational Bayesian approach has a significant advantage over the likelihood approach. The computational Bayesian approach gets a random sample from the true posterior, so credible intervals will have the correct coverage probabilities. The covariance matrix found by the likelihood approach does not actually relate to the spread of the likelihood, but rather to its curvature so coverage probabilities of confidence intervals may not be correct.

In Chapter 9 we develop the Poisson regression model, and the proportional hazards model. We follow the same approach we used for the logistic regression model. We find the maximum likelihood vector and matched curvature covariance matrix. Then we find the normal approximation to the posterior, and modify it to have heavy tails so it can be used as the candidate density. The Metropolis-Hastings algorithm is used to draw a sample from the exact posterior. We find that when we have censored survival data, and we relate the linear predictor, the censoring variable has the same shape as the Poisson, so we can use the same algorithm for the proportional hazards model. Again we will find that the computational Bayesian approach has the same advantages over the likelihood approach since the sample is from the true posterior.

In Chapter 10, we show how the Gibbs sampling algorithm cycles through each block of parameters, drawing from the conditional distribution of that block given all the parameters in other blocks at their most recent value. In general these conditional distributions are complicated. However, they will be quite simple when the parameters have a hierarchical structure. That means we can draw a graph where each node stands for a block of parameters or block of data. We connect the nodes

with arrows showing the direction of dependence. When we look at the resulting graph, we will find there are no connected loops. All nodes that lead into a specific node are called its parent nodes. All nodes that lead out of a specific node are called its child nodes. The other parent nodes of a child node are called coparent nodes. For this model, the conditional distribution of a specified node, given all the other nodes, will be proportional to its distribution given its parent nodes (the prior) times the joint distribution of all its child nodes given it and the coparent nodes of the child nodes (the likelihood). They will be particularly simple if the likelihood distributions are from the exponential family and the priors are from the conjugate family.

The biggest advantage of Markov chain Monte Carlo methods is that they allow the applied statistician to use more realistic models because he/she is not constrained by analytic or numerical tractability. Models that are based on the underlying situation can be used instead of models based on mathematical convenience. This allows the statistician to focus on the statistical aspects of the model without worrying about calculability.

Main Points

- Bayesian statistics does inference using the rules of probability directly.
- Bayesian statistics is based on a single tool, Bayes' theorem, which finds the posterior density of the parameters, given the data. It combines both the prior information we have given in the *prior* $g(\theta_1, \ldots, \theta_p)$ and the information about the parameters contained in the observed data given in the *likelihood* $f(y_1, \ldots, y_n|\theta_1, \ldots, \theta_p)$.
- It is easy to find the unscaled posterior by *posterior* proportional to *prior* times *likelihood*

$$g(\theta_1, \ldots, \theta_p|y_1, \ldots, y_n) \propto g(\theta_1, \ldots, \theta_p) \times f(y_1, \ldots, y_n|\theta_1, \ldots, \theta_p)\,.$$

  The unscaled posterior has all the shape information. However, it is not the exact posterior density. It must be divided by its integral to make it exact.
- Evaluating the integral may be very difficult, particularly if there are lots of parameters. It is hard to find the exact posterior except in a few special cases.
- The *Likelihood principle* states that if two experiments have proportional likelihoods, then they should lead to the same inference.
- The *Likelihood* approach to statistics does inference solely using the likelihood function, which is found by cutting the sampling surface with a vertical hyperplane through the observed value of the data. It is not considered to be a probability density.
- The *maximum likelihood estimate* is the mode of the likelihood function.

- The complete inference in Bayesian statistics is the posterior density. It is found by cutting the joint density of parameters and observations with the same vertical hyperplane through the observed values of the data.

- The usual Bayesian estimate is the mean of the posterior, as this minimizes mean-squared error of the posterior. Note: this will be different from the MLE even when flat priors are used and the posterior is proportional to the likelihood!

- The two approaches have different ways of dealing with nuisance parameters. The likelihood approach often uses the *profile likelihood* where the maximum conditional likelihood value of the nuisance parameter is plugged into the joint likelihood. The Bayesian approach is to integrate the nuisance parameter out of the joint posterior.

- Computational Bayesian statistics is based on drawing a Monte Carlo random sample from the unscaled posterior. This replaces very difficult numerical calculations with the easier process of drawing random variables. Sometimes, particularly for high dimensional cases, this is the only feasible way to find the posterior.

- The distribution of a random sample from the posterior approaches the exact posterior distribution. Estimates of parameters can be calculated from statistics calculated from the random sample.