is, the first customer is willing to pay to be on TOD or seasonal rates probably more than the rates are actually worth in terms of reduced energy bills. Finally, this customer is willing to pay more than the average customer to stay with the local utility. From a marketing perspective, the local utility can easily retain and make extra profits from this customer by offering a long-term contract under time-of-day or seasonal rates.

The third customer dislikes seasonal and TOD rates, evaluating them as if all of his consumption were in the highest-priced periods. He dislikes long-term contracts far more than the average customer, and yet, unlike most customers, prefers to receive service from a known company that is not his local utility. This customer is a prime target for capture by a well-known company if the company offers him a fixed price without requiring a commitment.

The second customer is less clearly a marketing opportunity. A well-known company is on about an equal footing with the local utility in competing for this customer. This in itself might make the customer a target of well-known suppliers, since he is less tied to the local utility than most customers. However, beyond this information, there is little beyond low prices (which all customers value) that would seem to attract the customer. His evaluation of TOD and seasonal rates are sufficiently negative that it is unlikely that a supplier could attract and make a profit from the customer by offering these rates. The customer is willing to pay to avoid a long-term contract, and so a supplier could attract this customer by not requiring a contract if other suppliers were requiring contracts. However, if other suppliers were not requiring contracts either, there seems to be little leverage that any supplier would have over its competitors. This customer will apparently be won by the supplier that offers the lowest fixed price.

The discussion of these three customers illustrates the type of information that can be obtained by conditioning on customer's choices, and how the information translates readily into characterizing each customer and identifying profitable marketing opportunities.

Conditional probability for the last choice

Recall that the last choice situation faced by each customer was not included in the estimation. It can therefore be considered a new choice situation and used to assess the effect of conditioning on past choices.

We identified which alternative each customer chose in the new choice situation and calculated the probability of this alternative. The probability was first calculated without conditioning on previous choices. This calculation uses the mixed logit formula (11.5) with the population distribution of β_n and the point estimates of the population parameters. The average of this unconditional probability over customers is 0.353. The probability was then calculated conditioned on previous choices. Four different ways of calculating this probability were used:

1. based on formula (11.3) using the point estimates of the population parameters.

2. based on formula (11.3) along with the procedure in section (11.3) that takes account of the sampling variance of the estimates of the population parameters.

3-4 as the logit formula

$$\frac{e^{\beta_n' x_{niT+1}}}{\sum_j e^{\beta_n' x_{njT+1}}}$$

with the conditional mean $\bar{\beta}_n$ being used as β_n. This method is equivalent to using the customer's $\bar{\beta}_n$ as if it were an estimate of the customer's true coefficients, β_n. The two versions differ in whether $\bar{\beta}_n$ is calculated on the basis of the point estimate of the population parameters (method 3) or takes the sampling distribution into account (method 4).

Results are given in Table 11.6 for model 2. The most prominent result is that conditioning on each customer's previous choices improves the forecasts for the last choice situation considerably. The average probability of the chosen alternative increases from 0.35 without conditioning to over 0.50 with conditioning. For nearly three-quarters of the 361 sampled customers, the prediction of their last choice situation is better with conditioning than without, with the average probability rising by more than 0.25. For the other customers, the conditioning makes the prediction in the last choice situations less accurate, with the average probability for these customers dropping.

There are several reasons why the predicted probability after conditioning is not always greater. First, the choice experiments were constructed such that each situation would be fairly different from the

Table 11.6: Probability of chosen alternative in last choice situation

	Method 1	Method 2	Method 3	Method 4
Average probability	0.5213	0.5041	0.5565	0.5487
Number of customers whose probability rises with conditioning	266	260	268	264
Average rise in probability for customers with a rise	0.2725	0.2576	0.3240	0.3204
Number of customers whose probability drops with conditioning	95	101	93	97
Average fall in probability for customers with a drop	0.1235	0.1182	0.1436	0.1391

other situations so as to obtain as much variation as possible. If the last situation involves new tradeoffs, the previous choices will not be useful and may in fact be detrimental to predicting the last choice. A more appropriate test might be to design a series of choice situations that elicited information on the relevant tradeoffs and then design an extra "hold-out" situation that is within the range of tradeoffs of the previous ones. Second, we did not include in our model all of the attributes of the alternatives that were presented to customers. In particular, we omitted attributes that did not enter significantly in the estimation of the population parameters. Some customers might respond to these omitted attributes, even though they are insignificant for the population as a whole. Insofar as the last choice situation involves tradeoffs of these attributes, the conditional distributions of tastes would be misleading since the relevant tastes are excluded. This explanation might imply that, if a mixed logit is going to be used for obtaining conditional densities for each customer, the researcher might want include attributes that could be important for some individuals even though they are insignificant for the population as a whole. Third, regardless of how the survey and model are designed, some customers might respond to choice situations in a quixotic manner, such that the tastes that are evidenced in previous choices are not applied by the customer in the last choice situation. Last, random factors can cause the probability for some customers to drop with conditioning even when the first three reasons do not. While at least one of these factors may be contributing to the lower choice probabilities for some of the customers in our sample, the gain in predictive accuracy for the customers with an increase in probability after conditioning is over twice as great as the loss in accuracy for those with a decrease, and the number of customers with a gain is almost three times as great as the number with a loss.

The third and easiest method, which simply calculates the standard logit formula using the customers' $\bar{\beta}_n$ based on the point estimate of the population parameters, gives the highest probability. This procedure does not account for the distribution of β_n around $\bar{\beta}_n$ nor the sampling distribution of $\hat{\theta}$. Accounting for either variance reduces the average probability: using the conditional distribution of β_n rather than just the mean $\bar{\beta}_n$ (methods 1 and 2 compared with methods 3 and 4, respectively) reduces the average probability, and accounting for the sampling distribution of $\hat{\theta}$ rather than the point estimate (methods 2 and 4 compared with methods 1 and 3, respectively) also reduces the average probability. This result does not mean that method 3, which incorporates the least variance, is superior to the others. Methods 3 and 4 are consistent only if the number of choice situations is able to rise without bound, such that $\bar{\beta}_n$ can be considered to be an estimate of β_n. With fixed T, methods 1 and 2 are more appropriate since they incorporate the entire conditional density.

11.7 Discussion

This chapter demonstrates how the distribution of coefficients conditioned on the customer's observed choices are obtained for the distribution of coefficients in the population. While these conditional distributions can be useful in several ways, it is important to recognize the limitations of the concept. First, the use of conditional distributions in forecasting is limited to those customers whose previous choices are observed. Second, while the conditional distribution of each customer can be used in cluster analysis and for other identification purposes, the researcher will often want to relate preferences to observable demographics of the customers. Yet, these observable demographics of the customers could be entered directly into the model itself such that the population parameters vary with the observed characteristics of the customers in the population. In fact, entering demographics into the model is more direct and more accessible to hypothesis testing than estimating a model without these characteristics, calculating the conditional distribution for each customer, and then doing cluster and other analyses on the moments of the conditional distributions.

Given these issues, there are three main reasons that a researcher might benefit from calculating customers' conditional distributions. First, information on the past choices of customers are becoming more and more widely available. Examples include scanner data for customers with club cards at grocery stores, frequent flier programs for airlines, and purchases from internet retailers. In these situations, conditioning on previous choices allows for effective targeted marketing and the development of new products and services that match the revealed preferences of subgroups of customers.

Second, the demographic characteristics that differentiate customers with different preferences might be more evident through cluster analysis on the conditional distributions than through specification testing in the model itself. Cluster analysis has its own unique way of identifying patterns, which might in some cases be more effective than specification testing within a discrete choice model.

Third, examination of customers' conditional distributions can often identify patterns that cannot be related to observed characteristics of customers but are nevertheless useful to know. For instance, knowing that a product or marketing campaign will appeal to a share of the population because of their particular preferences is often sufficient, without needing to identify the people on the basis of their demographics. The conditional densities can greatly facilitate analyses that have these goals.

Chapter 12

Bayesian Procedures

12.1 Introduction

A powerful set of procedures for estimating discrete choice models has been developed within the Bayesian tradition. The breakthough concepts were introduced by Albert and Chib (1993) and McCulloch and Rossi (1994) in the context of probit and Allenby and Lenk (1994) and Allenby (1997) for mixed logits with normally distributed coefficients. These authors showed how the parameters of the model can be estimated without needing to calculate the choice probabilities. Their procedures provide an alternative to the classical estimation methods described in Chapter 10. Rossi et al. (1996), Allenby (1997), and Allenby and Rossi (1999) showed how the procedures can also be used to obtain information on individual-level parameters within a model with random taste variation. As such, they provide a Bayesian analog to the classical procedures that we describe in Chapter 11. Variations of these procedures to accommodate other aspects of behavior have been numerous. For example: Arora, Allenby and Ginter (1998) generalized the mixed logit procedure to account for the quantity of purchases as well as brand choice in each purchase occasion. Bradlow and Fader (2001) showed how similar methods can be used to examine rankings data at an aggregate level rather than choice data at the individual level. Chib and Greenberg (1998) and Wang, Bradlow and Wainer (2001) developed methods for interrelated discrete responses. Chiang *et al.* (1999) examined situations where the choice set that the decision-maker considers is unknown to the researcher. Train (2001) extended the Bayesian procedure for mixed logit to non-normal distributions of coefficients, including lognormal, uniform, and triangular distributions.

The Bayesian procedures avoid two of the most prominent difficulties associated with classical procedures. First, the Bayesian procedures do not require maximization of any function. With probit and some mixed logit models (especially those with lognormal distributions), maximization of the simulated likelihood function can be difficult numerically. Often the algorithm fails to converge for various reasons. The choice of starting values is often critical, with the algorithm converging from starting values that are close to the maximum but not from other starting values. The issue of local versus global maxima complicates the maximization further, since convergence does not guarantee that the global maximum has been attained. Second, desirable estimation properties, such as consistency and efficiency, can be attained under more relaxed conditions with Bayesian procedures than classical ones. As shown in Chapter 10, maximum simulated likelihood is consistent only if the number of draws used in simulation is considered to rise with sample size; and efficiency is attained only if the number of draws rises faster than the square root of sample size. In contrast, the Bayesian estimators that we describe are consistent for a fixed number of draws used in simulation and are efficient if the number of draws rises at any rate with sample size.

These advantages come at a price, of course. For researchers who are trained in a classical perspective, the learning curve can be steep. Numerous interrelated techniques and concepts must be assimilated before the power of them becomes clear. I can assure the reader, however, that the effort is worthwhile. Another cost of the Bayesian procedures is more fundamental. To simulate relevant statistics that are defined over a distribution, the Bayesian procedures use an iterative process that converges, with a sufficient number of iterations, to draws from that distribution. This convergence is different from the convergence to a maximum that is needed for classical procedures and involves its own set of difficulties. The researcher cannot easily determine whether convergence has actually been achieved. As such, the Bayesian procedures trade the difficulties of convergence to a maximum for the difficulties associated with this different kind of convergence. The researcher will need to decide, in a particular setting, which type of convergence is less burdensome.

For some behavioral models and distributional specifications, Bayesian procedures are far faster and, after the initial learning that a classicist might need, are more straightforward from a programming perspective than classical procedures. For other models, the classical procedures are easier. We will explore the relative speed of Bayesian and classical procedures in the sections to follow. The differences can be readily categorized, through an understanding of how the two sets of procedures operate. The researcher can use this understanding in deciding which procedure to use in a particular setting.

Two important notes are required before proceeding. First, the Bayesian procedures, and the term "hierarchical Bayes" that is often used in the context of discrete choice models, refer to an estimation method, not a behavioral model. Probit, mixed logit, or any other model that the researcher specifies can, in principle, be estimated by either classical or Bayesian procedures. Second, the Bayesian perspective from which these procedure arise provides a rich and intellectually satisfying paradigm for inference and decision-making. Nevertheless, a researcher who is uninterested in the Bayesian perspective can still benefit from Bayesian procedures: the use of Bayesian procedures does not necessitate that the researcher adopt a Bayesian perspective on statistics. As we will show, the Bayesian procedures provide an estimator whose properties can be examined and interpreted in purely classical ways. Under certain conditions, the estimator that results from the Bayesian procedures is asymptotically equivalent to the maximum likelihood estimator. The researcher can therefore use Bayesian procedures to obtain parameter estimates and then interpret them the same as if they were maximum likelihood estimates. A highlight of the Bayesian procedures is that the results can be interpreted from both perspectives simultaneously, drawing on the insights afforded by each tradition. This dual interpretation parallels that of the classical procedures, whose results can be transformed for Bayesian interpretation as described by Geweke (1989). In short, the researcher's statistical perspective need not dictate her choice of procedure.

In the sections below, we provide an overview of Bayesian concepts in general, introducing the prior and posterior distributions. We then show how the mean of the posterior distribution can be interpreted from a classical perspective as being asymptotically equivalent to the maximum of the likelihood function. Next we address the numerical issue of how to calculate the mean of the posterior distribution. Gibbs sampling and, more generally, the Metropolis-Hastings algorithm can be used to obtain draws from practically any posterior distribution, no matter how complex. The mean of these draws simulates the mean of the posterior and, as such, constitutes the parameter estimates. The standard deviation of the draws provides the classical standard errors of the estimates. We apply the method to a mixed logit model and compare the numerical difficulty and speed of the Bayesian and classical procedures under various specifications.

12.2 Overview of Bayesian concepts

Consider a model with parameters θ. The researcher has some initial ideas about the value of these parameters and collects data to improve this understanding. Under Bayesian analysis, the researcher's ideas about the parameters are represented by a probability distribution over all possible values that the parameters can take, where the probability represents how likely the researcher thinks it is for the parameters to take a particular value. Prior to collecting data, the researcher's ideas are based on logic, intuition or past analyses. These ideas are represented by a density on θ, called the prior distribution and denoted $k(\theta)$. The researcher collects data in order to improve her ideas about the value of θ. Suppose the researcher observes a sample of N independent decision-makers. Let y_n denote the observed choice (or choices) of decision-maker n, and let the set of observed choices for the entire sample be labeled collectively as $Y = \{y_1, \ldots, y_N\}$. Based on this sample information, the researcher changes, or updates, her ideas about θ. The updated ideas are represented by a new density on θ, labeled $K(\theta \mid Y)$ and called the posterior distribution. This posterior distribution depends on Y since it incorporates the information that is contained in the observed sample.

The question arises: how exactly do the researcher's ideas about θ change from observing Y? That is, how does the posterior distribution $K(\theta \mid Y)$ differ from the prior distribution $k(\theta)$? There is a precise relationship between the prior and posterior distribution, established by Bayes rule. Let $P(y_n \mid \theta)$ be the probability of outcome y_n for decision-maker n. This probability is the behavioral model that relates the explanatory variables and parameters to the outcome, though the notation for the explanatory variables is omitted for simplicity. The probability of observing the sample outcomes Y is

$$L(Y \mid \theta) = \prod_{n=1}^{N} P(y_n \mid \theta).$$

This is the likelihood function (not logged) of the observed choices. Note that it is a function of the parameters θ.

Bayes' rule provides the mechanism by which the researcher improves her ideas about θ. By the rules of conditioning,

$$K(\theta \mid Y)L(Y) = L(Y \mid \theta)k(\theta). \tag{12.1}$$

where $L(Y)$ is the marginal probability of Y, marginal over θ:

$$L(Y) = \int L(Y \mid \theta)k(\theta)d\theta.$$

Both sides of equation (12.1) represent the joint probability of Y and θ, with the conditioning in opposite direction. The left hand side is the probability of Y times the probability of θ given Y, while the right hand side is the probability of θ times the probability of Y given θ. Rearranging we have

$$K(\theta \mid Y) = \frac{L(Y \mid \theta)k(\theta)}{L(Y)}. \tag{12.2}$$

This equation is Bayes' rule applied to prior and posterior distributions. In general, Bayes rule links conditional and unconditional probabilities in any setting and does not imply a Bayesian perspective on statistics. Bayesian statistics arises when the unconditional probability is the prior distribution (which reflects the researcher's ideas about θ *not* conditioned on the sample information) and the conditional probability is the posterior distribution (which gives the researcher's ideas about θ conditioned on the sample information.)

We can express equation (12.2) in a more compact and convenient form. The marginal probability of Y, $L(Y)$, is constant with respect to θ and, more specifically, is the integral of the numerator of (12.2). As such, $L(Y)$ is simply the normalizing constant that assures that the posterior distribution integrates to 1, as required for any proper density. Using this fact, equation (12.2) can be stated more succinctly by saying simply that the posterior distribution is proportional to the prior distribution times the likelihood function:

$$K(\theta \mid Y) \propto L(Y \mid \theta)k(\theta)$$

Intuitively, the probability that the researcher ascribes to a given value for the parameters *after* seeing the sample, is the probability that she ascribes *before* seeing the sample times the probability (i.e., *likelihood*) that those parameter values would result in the observed choices.

The mean of the posterior distribution is:

$$\bar{\theta} = \int \theta K(\theta \mid Y) d\theta \tag{12.3}$$

This mean has importance from both a Bayesian and classical perspective. From a Bayesian perspective, $\bar{\theta}$ is the value of θ that minimizes the expected cost of the researcher being wrong about θ, if the cost of error is quadratic in the size of the error. From a classical perspective, $\bar{\theta}$ is an estimator that has the same asymptotic sampling distribution as the maximum likelihood estimator. We explain both of these concepts below.

12.2.1 Bayesian properties of $\bar{\theta}$

The researcher's views about θ are represented by the posterior $K(\theta \mid Y)$ after observing the sample. Suppose that the researcher was required to guess the true value of θ and would be levied a penalty for the extent to which her guess differed from the true value. More realistically, suppose that some action must be taken that depends on the value of θ, such as a manufacturer setting the price of a good when the revenues at any price depend on the price elasticity of demand. There is a cost to taking the wrong action, such as setting price based on the belief that the price elasticity is -.2 when the true elasticity is actually -.3. The question becomes: what value of θ should the researcher use in these decisions in order to minimize her expected cost of being wrong, given her beliefs about θ as represented in the posterior distribution?

If the cost of being wrong is quadratic in the distance between the θ that is used in the decision and the true θ, then the optimal value of θ to use in the decision is $\bar{\theta}$. This fact can be demonstrated as follows. If the researcher uses θ_0 in her decisions when the true value is θ^*, the cost of being wrong is:

$$C(\theta_0, \theta^*) = (\theta_0 - \theta^*)' B(\theta_0 - \theta^*)$$

where B is a matrix of constants. The researcher doesn't know the true value of θ but has beliefs about its value as represented in $K(\theta \mid Y)$.

The researcher can therefore calculate the expected cost of being wrong when using the value θ_0. This expected cost is:

$$\begin{aligned} EC(\theta_0) &= \int C(\theta_0, \theta) K(\theta \mid Y) d\theta \\ &= \int (\theta_0 - \theta)' B (\theta_0 - \theta) K(\theta \mid Y) d\theta. \end{aligned}$$

The value of θ_0 that minimizes this expected cost is determined by differentiating $EC(\theta^0)$, setting the derivative to zero, and solving for θ^0. The derivative is:

$$\begin{aligned} \frac{\partial EC(\theta_0)}{\partial \theta_0} &= \int \frac{(\theta_0 - \theta)' B (\theta_0 - \theta)}{\partial \theta_0} K(\theta \mid Y) d\theta \\ &= \int 2(\theta_0 - \theta)' B K(\theta \mid Y) d\theta \\ &= 2\theta_0' B \int K(\theta \mid Y) d\theta - 2 \left(\int \theta K(\theta \mid Y) d\theta \right)' B \\ &= 2\theta_0' B - 2\bar{\theta}' B. \end{aligned}$$

Setting this expression to zero and solving for θ_0, we have

$$\begin{aligned} 2\theta_0' B - 2\bar{\theta}' B &= 0 \\ \theta_0' B &= \bar{\theta}' B \\ \theta_0 &= \bar{\theta}. \end{aligned}$$

The mean of the posterior, $\bar{\theta}$, is the value of θ that a Bayesian researcher would optimally act upon if the cost of being wrong about θ rises quadratically with the distance to the true θ.

Zellner (1971) describes the optimal Bayesian estimator under other loss functions. While the loss function is usually assumed to be symmetric and unbounded, like the quadratic, it need not be either; see, for example, Wen and Levy (2001). Importantly, Bickel and Doksum (2000) show that the correspondence that we describe in the next section between the mean of the posterior and the maximum likelihood estimator also applies to Bayesian estimators that are optimal under many other loss functions.

12.2.2 Classical properties of $\bar{\theta}$: The Bernstein-von Mises theorem

Classical statistics is not concerned with the researcher's beliefs and contains no notion of prior and posterior distributions. The concern of classical statistics is to determine the sampling distribution of an estimator. This distribution reflects the fact that a different sample would produce a different point estimate. The sampling distribution is the distribution of point estimates that would be obtained if many different samples were taken. Usually, the sampling distribution for an estimator cannot be derived for small samples. However, the asymptotic sampling distribution can usually be derived, which approximates the actual sampling distribution when sample size is large enough. In classical statistics, the asymptotic sampling distribution determines the properties of the estimator, such as whether the estimator is consistent, asymptotically normal, and efficient. The variance of the asymptotic distribution provides the standard errors of the estimates and allows for hypothesis testing, the accuracy of which rises with sample size.

From a classical perspective, $\bar{\theta}$ is simply a statistic like any other statistic. Its formula, given in (12.3), exists and can be applied even if the researcher does not interpret the formula as representing the mean a posterior distribution. The researcher can consider $K(\theta \mid Y)$ to be a function defined by equation (12.2) for any arbitrarily defined $k(\theta)$ that meets the requirements of a density. The relevant question for the classical researcher is the same as with any statistic: what is the sampling distribution of $\bar{\theta}$?

The answer to this question is given by the Bernstein-von Mises theorem. This theorem has a long provenance and takes many forms. In the nineteenth century, LaPlace (1820) observed that posterior distributions start to look more and more like normal distributions as sample size increases. Over the years, numerous versions of the observation have been demonstrated under various conditions, and its implications have been more fully explicated. See Rao (1987), Cam and Yang (1990), Lehmann and Casella (1998), and Bickel and Doksum (2000) for modern treatments with historical notes. The theorem is named after Bernstein (1917) and von Mises (1931) because they seem to be the first to provide a formal proof of LaPlace's observation, though under restrictive assumptions that others later relaxed.

I describe the theorem as three related statements. In these statements, the information matrix, which we used extensively in Chapters 8 and 10, is important. Recall that the score of an observation is the gradient of that observation's log-likelihood with respect to the parameters: $s_n = \partial logP(y_n \mid \theta)/\partial\theta$ where $P(y_n \mid \theta)$ is the probability of decision-maker n's observed choices. The information matrix, -**H**, is the negative expected derivative of the score, evaluated at the true parameters:

$$-\mathbf{H} = -E\left(\frac{\partial^2 logP(y_n \mid \theta^*)}{\partial\theta\partial\theta'}\right).$$

where the expectation is over the population. (The negative is taken so that the information matrix can be positive definite, like a covariance matrix.) Recall also that the maximum likelihood estimator has an asymptotic variance equal to $(-\mathbf{H})^{-1}/N$. That is, $\sqrt{N}(\theta^* - \hat{\theta}) \xrightarrow{d} N(0, (-\mathbf{H})^{-1})$, such that $\hat{\theta} \overset{a}{\sim} N(\theta^*, (-\mathbf{H})^{-1}/N)$, where $\hat{\theta}$ is the maximum likelihood estimator.

We can now give the three statements that collectively constitute the Bernstein-von Mises theorem:

1. $\sqrt{N}(\theta - \bar{\theta}) \xrightarrow{d} N(0, (-\mathbf{H})^{-1})$.

   Stated intuitively, the posterior distribution of θ converges to a normal distribution with variance $(-\mathbf{H})^{-1}/N$ as sample size rises. In using the expression $\xrightarrow{d}$ in this context, it is important to note that the distribution that is converging is the posterior distribution of $\sqrt{N}(\theta - \bar{\theta})$ rather than the sampling distribution. In classical analysis of estimators, as in Chapter 10, the notation $\xrightarrow{d}$ is used to indicate that the sampling distribution is converging. Bayesian analysis examines the posterior rather than the sampling distribution and the notation indicates that the posterior distribution is converging.

   The important points to recognize in this first statement are that, as sample size rises, (i) the posterior becomes normal and (ii) the variance of the posterior becomes the same as the sampling variance of the maximum likelihood estimator. These two points are relevant for the next two statements.

2. $\sqrt{N}(\bar{\theta} - \hat{\theta}) \xrightarrow{p} 0$.

   The mean of the posterior converges to the maximum of the likelihood function. An even stronger statement is being made. The difference between the mean of the posterior and the maximum of the likelihood function disappears asymptotically, *even when* the difference is scaled up by $\sqrt{N}$.

   This result makes intuitive sense, given the first result. Since the posterior eventually becomes normal, and the mean and maximum are the same for a normal distribution, the mean of the posterior eventually becomes the same as the maximum of the posterior. Also, the effect of the prior distribution on the posterior disappears as sample size rises (provided of course that the prior is not zero in the neighborhood of the true value). The posterior is therefore proportional to the likelihood function for large enough sample sizes. The maximum of the likelihood function becomes the same as the maximum of the posterior, which, as stated, is also the mean. Stated succinctly: since the posterior is asymptotically normal such that its mean equals its maximum, and the posterior is proportional to the likelihood function asymptotically, the difference between $\bar{\theta}$ and $\hat{\theta}$ eventually disappears.

3. $\sqrt{N}(\bar{\theta} - \theta^*) \xrightarrow{d} N(0, (-\mathbf{H})^{-1})$.

   The mean of the posterior, considered as a classical estimator, is asymptotically equivalent to the maximum likelihood estimator. That is, $\bar{\theta} \overset{a}{\sim} N(\theta^*, (-\mathbf{H})^{-1}/N)$, just like the maximum likelihood estimator. Note that since we are now talking in classical terms, the notation refers to the sampling distribution of $\bar{\theta}$, the same as it would for any estimator.

   This third statement is an implication of the first two. The statistic $\sqrt{N}(\bar{\theta} - \theta^*)$ can be rewritten as:

   $$\sqrt{N}(\bar{\theta} - \theta^*) = \sqrt{N}(\hat{\theta} - \theta^*) + \sqrt{N}(\bar{\theta} - \hat{\theta})$$

   From statement 2, we know that $\sqrt{N}(\bar{\theta} - \hat{\theta}) \xrightarrow{p} 0$, such that the second term disappears asymptotically. Only the first term affects the asymptotic distribution. This first term is the defining statistic for the maximum likelihood estimator $\hat{\theta}$. We showed in Chapter 10 that $\sqrt{N}(\hat{\theta} - \theta^*) \xrightarrow{d} N(0, (-\mathbf{H})^{-1})$. The statistic $\sqrt{N}(\bar{\theta} - \theta^*)$ therefore follows the same distribution asymptotically. Essentially, since $\bar{\theta}$ and $\hat{\theta}$ converge, their asymptotic sampling distributions are the same.

The Bernstein-von Mises theorem establishes that $\bar{\theta}$ is on the same footing, in classical terms, as $\hat{\theta}$. Instead of maximizing the likelihood function, the researcher can calculate the mean of the posterior distribution and know that the resulting estimator is as good in classical terms as maximum likelihood.

The theorem also provides a procedure for obtaining the standard errors of the estimates. Statement 1 says that asymptotically the variance of the posterior distribution is $(-\mathbf{H})^{-1}/N$, which, by statement 3, is the asymptotic sampling variance of the estimator $\bar{\theta}$. The variance of the posterior is the asymptotic variance of the estimates. The researcher can perform estimation entirely by using moments of the posterior: The mean of the posterior provides the point estimates, and the standard deviation of the posterior provides the standard errors.

In applications, the posterior mean and the maximum of the likelihood function can differ when sample size is insufficient for the asymptotic convergence. Huber and Train (2001) found the two to be remarkably similar in their application, while Ainslie et al. (2001) found them to be sufficiently different to warrant consideration. When the two estimates are not similar, other grounds must be used to choose between them (if indeed a choice is necessary), since their asymptotic properties are the same.

12.3 Simulation of posterior mean

To calculate the mean of the posterior distribution, simulation procedures are generally required. As stated above, the mean is:

$$\bar{\theta} = \int \theta K(\theta \mid Y) d\theta.$$

A simulated approximaton of this integral is obtained by taking draws of θ from the posterior distribution and averaging the results. The simulated mean is:

$$\check{\theta} = \frac{1}{R}\sum_{r=1}^{R} \theta^r$$

where θ^r is the r-th draw from $K(\theta \mid Y)$. The standard deviation of the posterior, which serves as the standard errors of the estimates, is simulated by taking the standard deviation of the R draws.

As stated above, $\bar{\theta}$ has the same asymptotic properties as the maximum likelihood estimator $\hat{\theta}$. How does the use of simulation to approximate $\bar{\theta}$ affect its properties as an estimator? For maximum simulated likelihood (MSL), we found that the number of draws used in simulation must rise faster than the square root of sample size in order for the estimator to be asymptotically equivalent to maximum likelihood.

With a fixed number of draws, the MSL estimator is inconsistent. If the number of draws rises with sample size but at a slower rate than the square root of sample size, then MSL is consistent but not asymptotically normal or efficient. As we will see, desirable properties of the simulated mean of the posterior (SMP) are attained with more relaxed conditions on the number of draws. In particular, the SMP estimator is consistent and asymptotically normal for a fixed number of draws and becomes efficient and equivalent to maximum likelihood if the number of draws rises at any rate with sample size.

To demonstrate these properties, we examine the normalized statistic $\sqrt{N}(\check{\theta} - \theta^*)$. This statistic can be rewritten as:

$$\sqrt{N}(\check{\theta} - \theta^*) = \sqrt{N}(\bar{\theta} - \theta^*) + \sqrt{N}(\check{\theta} - \bar{\theta}).$$

From statement 3 of the Bernstein-von Mises theorem, we know the limiting distribution of the first term: $\sqrt{N}(\bar{\theta} - \theta^*) \xrightarrow{d} N(0, (-\mathbf{H})^{-1})$. The central limit theorem gives us the limiting distribution of the second term. $\check{\theta}$ is the average of R draws from a distribution with mean $\bar{\theta}$ and variance $(-\mathbf{H}^{-1})/N$. Assuming the draws are independent, the central limit theorem states that the average of these R draws is distributed with mean $\bar{\theta}$ and variance $(-\mathbf{H})^{-1}/RN$. Plugging this information into the second term, we have: $\sqrt{N}(\check{\theta} - \bar{\theta}) \xrightarrow{d} N(0, (-\mathbf{H})^{-1}/R)$. The two terms are independent by construction, and so

$$\sqrt{N}(\check{\theta} - \theta^*) \xrightarrow{d} N(0, (1 + (1/R))(-\mathbf{H})^{-1}).$$

The simulated mean of the posterior is consistent and asymptotically normal for fixed R. The covariance is inflated by a factor of $1/R$ due to the simulation; however, the covariance matrix can be calculated and so standard errors and hypothesis testing can be conducted that take into account the simulation noise.

If R rises at any rate with N, then the second term disappears asymptotically. We have

$$\sqrt{N}(\check{\theta} - \theta^*) \xrightarrow{d} N(0, (-\mathbf{H})^{-1})$$

which is the same as for the actual (unsimulated) mean $\bar{\theta}$ and the maximum likelihood estimator $\hat{\theta}$. When R rises with N, $\check{\theta}$ is asymptotically efficient and equivalent to maximum likelihood.

Two notes are required regarding this derivation. First, we have assumed that the draws from the posterior distribution are independent. In the sections to follow, we describe methods for drawing from the posterior that result in draws that exhibit a type of serial correlation. When draws of this type are used, the variance of the simulated mean is inflated by more than a factor of $1/R$. The estimator is still consistent and asymptotically normal with a fixed number of nonindependent draws; its covariance is simply greater. And, if R rises with N, the extra covariance due to simulation disappears asymptotically even with nonindependent draws, such that the simulated mean is asymptotically equivalent to maximum likelihood.

Second, we have assumed that draws from the posterior distribution can be taken without needing to simulate the choice probabilities. For some models, taking a draw from the posterior requires simulating the choice probabilities on which the posterior is based. In this case, the simulated mean of the posterior involves simulation within simulation, and the formula for its asymptotic distribution is more complex. As we will see, however, for most models, including all the models that we consider in this book, draws from the posterior can be taken without simulating the choice probabilities. One of the advantages of the Bayesian procedures is that they usually avoid the need to simulate choice probabilities.

12.4 Drawing from the posterior

Usually, the posterior distribution does not have a convenient form from which to take draws. For example, we know how to take draws easily from a joint untruncated normal distribution; however, it is rare that the posterior takes this form for the entire parameter vector. Importance sampling, which we describe in section 9.2.7 in relation to any density, can be useful for simulating statistics over the posterior. Geweke (1992, 1997) describes the approach with respect to posteriors and provides practical guidance on appropriate selection of a proposal density. Two other methods that we described in Chapter 9 are particularly useful for taking draws from a posterior distribution: Gibbs sampling and the Metropolis-Hasting algorithm. These methods are often called Monte Carlo Markov Chain, or MCMC, methods. Formally, Gibbs sampling is a special type of Metropolis-Hasting algorithm (Gelman, 1992). However, the case is so special, and so conceptually straightforward, that the term Metropolis-Hasting (M-H) is usually reserved for versions that are more complex than Gibbs sampling. That is, when the M-H algorithm is Gibbs sampling, it is referred to as Gibbs sampling, and when it is more complex than Gibbs sampling it is referred to as the M-H algorithm. I maintain this convention hereafter.

It would be useful for the reader to review sections (9.2.8) and (9.2.9), which describe Gibbs sampling and the M-H algorithm, since we will be using these procedures extensively in the remainder of this chapter. As stated above, the mean of the posterior is simulated by taking draws from the posterior and averaging the draws. Instead of taking draws from the multidimensional posterior for all the parameters, Gibbs sampling allows the researcher to take draws of one parameter at a time (or a subset of parameters), conditional on values of the other parameters (Casella and George, 1992). Drawing from the posterior for one parameter conditional on the others is usually much easier than drawing from the posterior for all parameters simultaneously. In some cases, the M-H algorithm is needed in conjunction with Gibbs sampling. Suppose, for example, that the posterior for one parameter conditional on the other parameters does not take a simple form. In this case, the M-H algorithm can utilized, since it is applicable to (practically) any distribution (Chib and Greenberg, 1995). The M-H algorithm is particularly useful in the context of posterior distributions because the normalizing constant for the posterior need not be calculated. Recall that the posterior is the prior times the likelihood function, divided by a normalizing constant that assures that the posterior integrates to one:

$$K(\theta \mid Y) = \frac{L(Y \mid \theta)k(\theta)}{L(Y)}.$$

where $L(Y)$ is the normalizing constant

$$L(Y) = \int L(Y \mid \theta)k(\theta)d\theta.$$

This constant can be difficult to calculate since it involves integration. As described in section (9.2.9), the M-H algorithm can be applied without knowing or calculating the normalizing constant of the posterior.

In summary: Gibbs sampling, combined if necessary with the M-H algorithm, allows draws to be taken from the posterior of a parameter vector for essentially any model. These procedures are applied to a mixed logit model in section (12.6) below. First, however, we will derive the posterior distribution for some very simple models. As we will see, these results often apply in complex models for a subset of the parameters. This fact facilitates the Gibbs sampling of these parameters.

12.5 Posterior for the mean and variance of a normal distribution

The posterior distribution takes a very convenient form for some simple inference processes. We describe two of these situations, which, as we will see, often arise within more complex models for a subset of the parameters. Both results relate to the normal distribution. We first consider the situation where the variance of a normal distribution is known, but the mean is not. We then turn the tables and consider the mean to be known but not the variance. Finally, combining these two situations with Gibbs sampling, we consider the situation where both the mean and variance are unknown.

12.5.1 Result A: Unknown mean, known variance

We discuss the one-dimensional case first, and then generalize to multiple dimensions. Consider a random variable β that is distributed normal with unknown mean b and known variance σ. The researcher observes a sample of N realizations of the random variable, labeled $\beta_n,\ n = 1, \ldots, N$. The sample mean is $\bar{\beta} = (1/N)\sum_n \beta_n$. Suppose the researcher's prior on b is $N(\beta_0, s_0)$; that is, the researcher's prior beliefs are represented by a normal distribution with mean b_0 and variance s_0. Note that we now have two normal distributions: the distribution of β which has mean b, and the prior distribution on this unknown mean, which has mean β_0. The prior indicates that the researcher thinks it is most likely that $b = \beta_0$ and also thinks there is a 95 percent chance that b is somewhere between $\beta_0 - 1.96\sqrt{s_0}$ and $\beta_0 + 1.96\sqrt{s_0}$. Under this prior, the posterior on b is $N(b_1, s_1)$ where

$$b_1 = \frac{\frac{1}{s_0}b_0 + \frac{N}{\sigma}\bar{\beta}}{\frac{1}{s_0} + \frac{N}{\sigma}}$$

and

$$s_1 = \frac{1}{\frac{1}{s_0} + \frac{N}{\sigma}}.$$

The posterior mean b_1 is the weighted average of the sample mean and the prior mean.

Proof: The prior is

$$k(b) = \frac{1}{\sqrt{2\pi s_0}} e^{-(b-b_0)^2/2s_0}$$

The probability of drawing β_n from $N(b, \sigma)$ is

$$\frac{1}{\sqrt{2\pi\sigma}} e^{-(b-\beta_n)^2/2\sigma}$$

and so the likelihood of the N draws is

$$\begin{aligned} L(\beta_n \ \forall n \mid b) &= \prod_n \frac{1}{\sqrt{2\pi\sigma}} e^{-(b-\beta_n)^2/2\sigma} \\ &= \frac{1}{(2\pi\sigma)^{N/2}} e^{-\sum(b-\beta_n)^2/2\sigma} \\ &= \frac{1}{(2\pi\sigma)^{N/2}} e^{(-N\bar{s}-N(b-\bar{\beta})^2)/2\sigma} \\ &= \frac{1}{(2\pi\sigma)^{N/2}} e^{-N\bar{s}/2\sigma} \cdot e^{-N(b-\bar{\beta})^2/2\sigma} \end{aligned}$$

where $\bar{s} = (1/N)\sum(\beta_n - \bar{\beta})^2$ is the sample variance of the β_n's. The posterior is therefore

$$\begin{aligned} K(b \mid \beta_n \ \forall n) &\propto L(\beta_n \ \forall n \mid b)k(b) \\ &= \frac{1}{(2\pi\sigma)^{N/2}} e^{-N\bar{s}/2\sigma} \cdot e^{-N(b-\bar{\beta})^2/2\sigma} \times \frac{1}{\sqrt{2\pi s_0}} e^{-(b-b_0)^2/2s_0} \\ &= m_1 e^{-[N(b-\bar{\beta})^2/2\sigma]-[(b-b_0)^2/2s_0]} \end{aligned}$$

where m_1 is a constant that contains all the multiplicative terms that do not depend on b. With some algebraic manipulation, we have

$$\begin{aligned} K(b \mid \beta_n \ \forall n) &\propto e^{-[N(b-\bar{\beta})^2/2\sigma]-[(b-b_0)^2/2s_0)]} \\ &\propto e^{(b^2-2b_1b)/2s_1} \\ &\propto e^{(b-b_1)^2/2s_1}. \end{aligned}$$

The second $\propto$ removes $\bar{\beta}^2$ and b_0^2 from the exponential, since they do not depend on b and thereby only affect the normalizing constant. (Recall that $exp(a+b) = exp(a)exp(b)$, such that adding and removing terms from the exponential has a multiplicative effect on $K(b \mid \beta_n \ \forall n)$.) The third $\propto$ adds $b_1\bar{\beta}$ to the exponential, which also does not depend on b. The posterior is therefore

$$K(b \mid \beta_n \ \forall n) = me^{(b-b_1)^2/2s_1},$$

where m is the normalizing constant. This formula is the normal density with mean b_1 and variance s_1.

As stated above, the mean of the posterior is a weighted average of the sample mean and the prior mean. The weight on the sample mean rises as sample size rises, such that for large enough N, the prior mean becomes irrelevant.

Often a researcher will want to specify a prior that represents very little knowledge about the parameters before taking the sample. In general, the researcher's uncertainty is reflected in the variance of the prior. A large variance means that the researcher has little idea about the value of the parameter. Stated equivalently, a prior that is nearly flat means that the researcher considers all possible values of the parameters to be equally likely. A prior that represents little information is called "diffuse."

We can examine the effect of a diffuse prior on the posterior of b. By raising the variance of the prior, s_0, the normal prior becomes more spread out and flat. As $s_0 \to \infty$, representing an increasingly diffuse prior, the posterior approaches $N(\bar{\beta}, \sigma/N)$.

The multivariate versions of this result is similar. Consider a K-dimensional random vector $\beta \sim N(b, W)$ with known W and unknown b. The researcher observes a sample $\beta_n, \ n = 1, \ldots, N$ whose sample mean is $\bar{\beta}$. If the researcher's prior on b is diffuse (normal with an unboundedly large variance), then the posterior is $N(\bar{\beta}, W/N)$.

Taking draws from this posterior is easy. Let L be the Choleski factor of W/N. Draw K iid standard normal deviates, $\eta_i, i = 1, \ldots, K$, and stack them into a vector $\eta = \langle \eta_1, \ldots, \eta_K \rangle'$. Calculate $\tilde{b} = \bar{\beta} + L\eta$. The resulting vector $\tilde{b}$ is a draw from $N(\bar{\beta}, W/N)$.

12.5.2 Result B: Unknown variance, known mean

Consider a (one-dimensional) random variable that is distributed normal with known mean b and unknown variance σ. The researcher observes a sample of N realizations, labeled β_n, $n = 1, \ldots, N$. The sample variance *around the known mean* is $\bar{s} = (1/N)\sum_n(\beta_n - b)^2$. Suppose the researcher's prior on σ is inverted gamma with degrees of freedom v_0 and scale s_0. This prior is denoted $IG(v_0, s_0)$. The density is zero for any negative value for σ, reflecting the fact that a variance must be positive. The mode of the inverted gamma prior is $s_0v_0/(1+v_0)$. Under the inverted gamma prior, the posterior on σ is also inverted gamma $IG(v_1, s_1)$, where

$$\begin{aligned} v_1 &= v_0 + N \\ s_1 &= \frac{v_0 s_0 + N\bar{s}}{v_0 + N}. \end{aligned}$$

Proof: An inverted gamma with v_0 degrees of freedom and scale s_0 has density

$$k(\sigma) = \frac{1}{m_0 \sigma^{(v_0+1)/2}} e^{-v_0 s_0/2\sigma}$$

where m_0 is the normalizing constant. The likelihood of the sample, treated as a function of σ, is

$$L(\beta_n \; \forall n \mid \sigma) = \frac{1}{(2\pi\sigma)^{N/2}} e^{-\sum(b-\beta_n)^2/2\sigma} = \frac{1}{(2\pi\sigma)^{N/2}} e^{-N\bar{s}/2\sigma}.$$

The posterior is then

$$\begin{aligned} K(\sigma \mid \beta_n \; \forall n) &\propto L(\beta_n \; \forall n \mid \sigma) k(\sigma) \\ &\propto \frac{1}{\sigma^{N/2}} e^{-N\bar{s}/2\sigma} \times \frac{1}{\sigma^{(v_0+1)/2}} e^{-v_0 s_0/2\sigma} \\ &= \frac{1}{\sigma^{(N+v_0+1)/2}} e^{-(N\bar{s}+v_0 s_0)/2\sigma} \\ &= \frac{1}{\sigma^{(v_1+1)/2}} e^{-v_1 s_1/2\sigma} \end{aligned}$$

which is the inverted gamma density with v_1 degrees of freedom and scale s_1.

The inverted gamma prior becomes more diffuse with lower v_0. For the density to integrate to one and have a mean, v_0 must exceed 1. It is customary to set $s_0 = 1$ when specifying $v_0 \to 1$. Under this diffuse prior, the posterior becomes $IG(1+N, (1+N\bar{s})/(1+N))$. The mode of this posterior is $(1+N\bar{s})/(2+N)$ which is approximately the sample variance $\bar{s}$ for large N.

The multivariate case is similar. The multivariate generalization of an inverted gamma distribution is the inverted Wishart distribution. The result in the multivariate case is the same as with one random variable except that the inverted gamma is replaced by the inverted Wishart.

A K-dimensional random vector $\beta \sim N(b, W)$ has known b but unknown W. A sample of size N from this distribution has variance around the known mean of $\bar{S} = (1/N)\sum_n(\beta_n - b)(\beta_n - b)'$. If the researcher's prior on W is inverted Wishart with v_0 degrees of freedom and scale matrix S_0, labeled $IW(v_0, S_0)$, then the posterior on W is $IW(v_1, S_1)$ where

$$\begin{aligned} v_1 &= v_0 + N \\ S_1 &= \frac{v_0 S_0 + N\bar{S}}{v_0 + N}. \end{aligned}$$

The prior becomes more diffuse with lower v_0, though v_0 must exceed K in order for the prior to integrate to one and have means. With $S_0 = I$, where I is the K-dimensional identity matrix, the posterior under a diffuse prior becomes $IW(K+N, (KI+N\bar{S})/(K+N))$. Conceptually, the prior is equivalent to the researcher having a previous sample of K observations whose sample variance was I. As N rises without bound, the influence of the prior on the posterior eventually disappears.

It is easy to take draws from inverted gamma and inverted Wishart distributions. Consider first an inverted gamma $IG(v_1, s_1)$. Draws are taken as follows:

1. Take v_1 draws from a standard normal and label the draws $\eta_i,\ i = 1, \ldots, v_1$.

2. Divide each draw by $\sqrt{s_1}$, square the result, and take the average. That is, calculate $r = (1/v_1)\sum_i(\sqrt{1/s_1}\eta_i)^2$, which is the sample variance of v_1 draws from a normal distribution whose variance is $1/s_1$.

3. Take the inverse of r: $\tilde{s} = 1/r$ is a draw from the inverted gamma.

Draws from a K-dimensional inverted Wishart $IW(v_1, S_1)$ are obtained as follows.

1. Take v_1 draws of K-dimensional vectors whose elements are independent standard normal deviates. Label these draws $\eta_i,\ i = 1, \ldots, v_1$.

2. Calculate the Choleski factor of the inverse of S_1, labeled L where $LL' = S_1^{-1}$.

3. Create $R = (1/v_1)\sum_i (L\eta_i)(L\eta_i)'$. Note that R is the variance of draws from a distribution with variance S_1^{-1}.

4. Take the inverse of R. The matrix $\tilde{S} = R^{-1}$ is a draw from $IW(v_1, S_1)$.

12.5.3 Unknown mean and variance

Suppose that both the mean b and variance W are unknown. The posterior for both parameters does not take a convenient form. However, draws can easily be obtained using Gibbs sampling and results A and B. A draw of b is taken conditional on W, and then a draw of W is taken conditional on b. Result A says that the posterior for b conditional on W is normal, which is easy to draw from. Result B says that the posterior for W conditional on b is inverted Wishart, which is also easy to draw from. Iterating through numerous cycles of draws from the conditional posteriors provides, eventually, draws from the joint posterior.

12.6 Hierarchical Bayes for mixed logit

In this section we show how the Bayesian procedures can be used to estimate the parameters of a mixed logit model. We utilize the approach developed by Allenby (1997), implemented by SawtoothSoftware (1999), and generalized by Train (2001). Let the utility that person n obtains from alternative j in time period t be

$$U_{njt} = \beta_n' x_{njt} + \varepsilon_{njt}$$

where ε_{njt} is iid extreme value and $\beta_n \sim N(b, W)$. Giving β_n a normal distribution allows us to use results A and B, which speeds estimation considerably. In the following section, we discuss the use of non-normal distributions.

The researcher has priors on b and W. Suppose the prior on b is normal with a unboundedly large variance. Suppose that the prior on W is inverted Wishart with K degrees of freedom and scale matrix I, the K-dimensional identity matrix. Note that these are the priors