CHAPTER II

Principles of Bayesian Analysis with Selected Applications

In this chapter we view some basic principles and concepts of Bayesian analysis and provide analyses of some relatively simple but important models and problems.

2.1 BAYES'S THEOREM

An essential element of the Bayesian approach is Bayes's theorem, also referred to in the literature as the *principle of inverse probability*.[1] Here we state the theorem for continuous random variables. Let $p(\mathbf{y}, \boldsymbol{\theta})$ denote the joint probability density function (pdf)[2] for a random observation vector $\mathbf{y}$ and a parameter vector $\boldsymbol{\theta}$, also considered random. The parameter vector $\boldsymbol{\theta}$ may have as its elements coefficients of a model, variances and covariances of disturbance terms, and so on. Then, according to usual operations with pdf's, we have

$$p(\mathbf{y},\boldsymbol{\theta}) = p(\mathbf{y}|\boldsymbol{\theta})\, p(\boldsymbol{\theta}) = p(\boldsymbol{\theta}|\mathbf{y})\, p(\mathbf{y}) \tag{2.1}$$

and thus

$$p(\boldsymbol{\theta}|\mathbf{y}) = \frac{p(\boldsymbol{\theta})\, p(\mathbf{y}|\boldsymbol{\theta})}{p(\mathbf{y})}, \tag{2.2}$$

[1] In problems involving "inverse probability" we have given data and from the information in the data try to infer what random process generated them. On the other hand, in problems of "direct probability" we know the random process, including values of its parameters, and from this knowledge make probability statements about outcomes or data produced by the known random process. Problems of statistical estimation are thus seen to be problems in "inverse probability," whereas many gambling problems are problems in "direct probability."

[2] Here and below we use the symbol p to denote pdf's generally and not one specific pdf. The argument of the function p as well as the context in which it is used will identify the particular pdf being considered.

with $p(\mathbf{y}) \neq 0$.[3] We can write this last expression as follows:

$$p(\boldsymbol{\theta}|\mathbf{y}) \propto p(\boldsymbol{\theta})\, p(\mathbf{y}|\boldsymbol{\theta}) \propto \text{prior pdf} \times \text{likelihood function}, \tag{2.3}$$

where $\propto$ denotes proportionality, $p(\boldsymbol{\theta}|\mathbf{y})$ is the *posterior pdf* for the parameter vector $\boldsymbol{\theta}$, given the sample information $\mathbf{y}$, $p(\boldsymbol{\theta})$ is the *prior pdf*[4] for the parameter vector $\boldsymbol{\theta}$, and $p(\mathbf{y}|\boldsymbol{\theta})$, viewed as a function of $\boldsymbol{\theta}$, is the well-known *likelihood function*.[5] Equation 2.3 is a statement of Bayes's theorem, a simple mathematical result in the theory of probability. Note that the joint posterior pdf, $p(\boldsymbol{\theta}|\mathbf{y})$, has all the prior and sample information incorporated in it. The prior information enters the posterior pdf via the prior pdf, whereas all the sample information enters via the likelihood function. In this latter connection the "likelihood principle" states that $p(\mathbf{y}|\boldsymbol{\theta})$, considered as a function of $\boldsymbol{\theta}$ ". . . constitutes the entire evidence of the experiment, that is, it tells all that the experiment has to tell."[6] The posterior pdf is employed in the Bayesian approach to make inferences about parameters.

Example 2.1. Assume that we have n independent observations, $\mathbf{y}' = (y_1, y_2, \ldots, y_n)$, drawn from a normal population with unknown mean μ and *known* variance $\sigma^2 = \sigma_0^2$. We wish to obtain the posterior pdf for μ. Applying (2.3) to this particular problem, we have

$$p(\mu|\mathbf{y}, \sigma_0^2) \propto p(\mu)\, p(\mathbf{y}|\mu, \sigma_0^2), \tag{2.4}$$

where $p(\mu|\mathbf{y}, \sigma_0^2)$ is the posterior pdf for the parameter μ, given the sample information $\mathbf{y}$ and the assumed known value σ_0^2, $p(\mu)$ is the prior pdf for μ, and $p(\mathbf{y}|\mu, \sigma_0^2)$, viewed as a function of the unknown parameter μ is the likelihood function. The likelihood function is given by $\prod_{i=1}^{n} p(y_i|\mu, \sigma_0^2)$, or

$$\begin{aligned} p(\mathbf{y}|\mu, \sigma_0^2) &= (2\pi\sigma_0^2)^{-n/2} \exp\left[-\frac{1}{2\sigma_0^2}\sum_{i=1}^{n}(y_i - \mu)^2\right] \\ &= (2\pi\sigma_0^2)^{-n/2} \exp\left[-\frac{1}{2\sigma_0^2}[\nu s^2 + n(\mu - \hat{\mu})^2]\right], \end{aligned} \tag{2.5}$$

[3] The quantity $p(\mathbf{y})$, the reciprocal of the normalizing constant for the pdf in (2.2), can be written as $p(\mathbf{y}) = \int p(\boldsymbol{\theta})\, p(\mathbf{y}|\boldsymbol{\theta})\, d\boldsymbol{\theta}$.

[4] As noted in Chapter 1, the prior pdf depends on the state of our initial information denoted by I_0. Here, to simplify the notation, we do not show this dependence explicitly; that is, we write $p(\boldsymbol{\theta})$ rather than $p(\boldsymbol{\theta}|I_0)$.

[5] The likelihood function is often written as $l(\boldsymbol{\theta}|\mathbf{y})$ to emphasize that it is *not* a pdf, whereas $p(\mathbf{y}|\boldsymbol{\theta})$ is a pdf for the observations given the parameters.

[6] L. J. Savage, "Subjective Probability and Statistical Practice," in L. J. Savage et al., *The Foundations of Statistical Inference*. London and New York: Methuen and Wiley, 1962, pp. 9–35, p. 17. Savage presents a discussion of the likelihood principle and provides references to earlier literature.

where $\nu = n - 1$, $\hat{\mu} = (1/n)\sum_{i=1}^{n} y_i$, the sample mean, and

$$s^2 = (1/\nu)\sum_{i=1}^{n}(y_i - \hat{\mu})^2,$$

the sample variance.[7]

As regards a prior pdf for μ, we assume that our prior information regarding this parameter can be represented by the following univariate normal pdf:

$$p(\mu) = \frac{1}{\sqrt{2\pi}\,\sigma_a} \exp\left[-\frac{1}{2\sigma_a^2}(\mu - \mu_a)^2\right], \tag{2.6}$$

where μ_a is the prior mean and σ_a^2 is the prior variance, parameters whose values are assigned by the investigator on the basis of his initial information. Then, on using Bayes's theorem to combine the likelihood function in (2.5) and the prior pdf in (2.6), we obtain the following posterior pdf for μ:

$$\begin{aligned} p(\mu|\mathbf{y}, \sigma_0^2) &\propto p(\mu)\, p(\mathbf{y}|\mu, \sigma_0^2) \\ &\propto \exp\left\{-\frac{1}{2}\left[\frac{(\mu - \mu_a)^2}{\sigma_a^2} + \frac{n}{\sigma_0^2}(\mu - \hat{\mu})^2\right]\right\} \\ &\propto \exp\left[-\left(\frac{\sigma_a^2 + \sigma_0^2/n}{2\sigma_a^2\sigma_0^2/n}\right)\left(\mu - \frac{\hat{\mu}\sigma_a^2 + \mu_a\sigma_0^2/n}{\sigma_a^2 + \sigma_0^2/n}\right)^2\right], \end{aligned} \tag{2.7}$$

from which it is seen that μ is normally distributed, a posteriori, with mean

$$E\mu = \frac{\hat{\mu}\sigma_a^2 + \mu_a\sigma_0^2/n}{\sigma_a^2 + \sigma_0^2/n} = \frac{\hat{\mu}(\sigma_0^2/n)^{-1} + \mu_a(\sigma_a^2)^{-1}}{(\sigma_0^2/n)^{-1} + (\sigma_a^2)^{-1}} \tag{2.8}$$

and variance given by

$$\text{Var}(\mu) = \frac{\sigma_a^2\sigma_0^2/n}{\sigma_a^2 + \sigma_0^2/n} = \frac{1}{(\sigma_0^2/n)^{-1} + (\sigma_a^2)^{-1}}. \tag{2.9}$$

Note that the posterior mean in (2.8) is a weighted average of the sample mean $\hat{\mu}$ and the prior mean μ_a, with the weights being the reciprocals of σ_0^2/n and σ_a^2. If we let $h_0 = (\sigma_0^2/n)^{-1}$ and $h_a = (\sigma_a^2)^{-1}$, then $E\mu = (\hat{\mu}h_0 + \mu_a h_a)/(h_0 + h_a)$, where the h's are often referred to as "precision" parameters. Also we have $\text{Var}(\mu) = 1/(h_0 + h_a)$ from (2.9), and thus the precision parameter associated with the posterior mean is just $[\text{Var}(\mu)]^{-1} = h_0 + h_a$, the sum of the sample and prior precision parameters.

To provide some illustrative numerical results suppose that in Example 2.1 our sample of $n = 10$ observations is

[7] The expression in the exponent in the second line of (2.5) is obtained by using the following result: $\sum_{i=1}^{n}(y_i - \mu)^2 = \sum_{i=1}^{n}[(y_i - \hat{\mu}) - (\mu - \hat{\mu})]^2 = \sum_{i=1}^{n}(y_i - \hat{\mu})^2 + n(\mu - \hat{\mu})^2$ with the cross product term $\sum_{i=1}^{n}(y_i - \hat{\mu})(\mu - \hat{\mu})$ disappearing, since $\sum_{i=1}^{n}(y_i - \hat{\mu}) = 0$.

Observation Number	Observations y_i
1	0.699
2	0.320
3	−0.799
4	−0.927
5	0.373
6	−0.648
7	1.572
8	−0.319
9	2.049
10	−3.077

$$\text{Sample mean:}\quad \hat{\mu} = \tfrac{1}{10}\sum_{i=1}^{10} y_i = -0.0757$$

where the y_i's are independently drawn from a normal population with unknown mean μ and known variance $\sigma^2 = \sigma_0{}^2 = 1.00$. Assume that our prior information is suitably represented by a normal pdf with prior mean $\mu_a = -0.0200$ and prior variance, $\sigma_a{}^2 = 2.00$. This prior pdf, which is plotted in Figure 2.1, represents our initial beliefs about the unknown parameter μ. On combining this prior pdf with the likelihood function, the posterior pdf is given by the expression in (2.7). For the particular sample shown above, with mean $\hat{\mu} = -0.0757$ and the values of the prior parameters $\mu_a = -0.02$ and $\sigma_a{}^2 = 2.00$, the mean of the posterior pdf from (2.8) is

$$E\mu = \frac{-0.0757/0.100 - 0.0200/2.00}{1/0.100 + 1/2.00} = -0.0730$$

and its variance from (2.9) is

$$\text{Var}\,(\mu) = \frac{1}{1/0.100 + 1/2.00} = 0.0952.$$

For comparison with the prior pdf the posterior pdf is plotted in Figure 2.1. It is seen that combining the information contained in just 10 independent observations with our prior information has resulted in a considerable reduction in our uncertainty about the parameter μ; that is, our prior variance is $\sigma_a{}^2 = 2.00$, whereas the variance of our posterior pdf is 0.0952. In addition, our posterior mean $E\mu = -0.0730$ is not very different from $\hat{\mu} = -0.0757$, the sample mean, but is quite a bit larger in absolute value than our prior mean, $\mu_a = -0.0200$. Note, however, that our prior pdf has a substantial variance, $\sigma_a{}^2 = 2.00$, and thus initially there is substantial

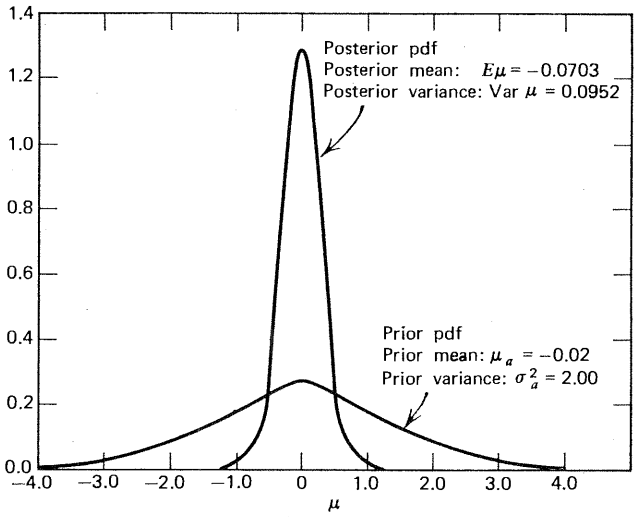

Figure 2.1 Plots of prior and posterior pdf's for μ. The prior and posterior pdf's are shown in (2.6) and (2.7), respectively.

probability density in the vicinity of −0.0730; that is, in this case our prior information is somewhat "vague" or "diffuse" in relation to the information in the sample.

2.2 BAYES'S THEOREM AND SEVERAL SETS OF DATA

If initially our prior pdf for a parameter vector $\boldsymbol{\theta}$ is $p(\boldsymbol{\theta})$ and we obtain a set of data $\mathbf{y}_1$ with pdf $p(\mathbf{y}_1|\boldsymbol{\theta})$, then from (2.3) the posterior pdf is

$$p(\boldsymbol{\theta}|\mathbf{y}_1) \propto p(\boldsymbol{\theta})\, p(\mathbf{y}_1|\boldsymbol{\theta}). \tag{2.10}$$

If we now obtain a new set of data, $\mathbf{y}_2$, generated independently of the first set, with pdf $p(\mathbf{y}_2|\boldsymbol{\theta})$, we can form the posterior pdf for $\boldsymbol{\theta}$ as follows. Use the posterior pdf in (2.10) as the prior pdf in the analysis of the new set of data $\mathbf{y}_2$ to obtain by means of Bayes's theorem

$$p(\boldsymbol{\theta}|\mathbf{y}_1, \mathbf{y}_2) \propto p(\boldsymbol{\theta}|\mathbf{y}_1)\, p(\mathbf{y}_2|\boldsymbol{\theta}), \tag{2.11}$$

where $p(\boldsymbol{\theta}|\mathbf{y}_1, \mathbf{y}_2)$ is the posterior pdf based on the information in $p(\boldsymbol{\theta})$ and the two samples of data $\mathbf{y}_1$ and $\mathbf{y}_2$. It is interesting to note that, since $p(\boldsymbol{\theta}|\mathbf{y}_1) \propto p(\boldsymbol{\theta})\, p(\mathbf{y}_1|\boldsymbol{\theta})$ from (2.10), (2.11) may be written as

$$p(\boldsymbol{\theta}|\mathbf{y}_1, \mathbf{y}_2) \propto p(\boldsymbol{\theta})\, p(\mathbf{y}_1|\boldsymbol{\theta})\, p(\mathbf{y}_2|\boldsymbol{\theta}). \tag{2.12}$$

In (2.12) $p(\mathbf{y}_1|\boldsymbol{\theta})\,p(\mathbf{y}_2|\boldsymbol{\theta})$ is the likelihood function for $\boldsymbol{\theta}$ based on the combined samples $\mathbf{y}_1$ and $\mathbf{y}_2$. Therefore it is the case that we obtain the same posterior pdf for $\boldsymbol{\theta}$, whether we proceed sequentially from $p(\boldsymbol{\theta})$ to $p(\boldsymbol{\theta}|\mathbf{y}_1)$ and then to $p(\boldsymbol{\theta}|\mathbf{y}_1, \mathbf{y}_2)$ or whether we use the likelihood function for the combined samples $p(\mathbf{y}_1, \mathbf{y}_2|\boldsymbol{\theta})$ in conjunction with the prior pdf $p(\boldsymbol{\theta})$. This general feature of the process of combining information in a prior pdf with information in successive samples can easily be shown to hold for cases involving more than two independent samples of data.

2.3 PRIOR PROBABILITY DENSITY FUNCTIONS

The prior pdf, denoted $p(\boldsymbol{\theta})$ in (2.3),[8] represents our prior information about parameters of a model; that is, in the Bayesian approach prior information about parameters of models is usually represented by an appropriately chosen pdf. In Example 2.1, for example, prior information about a mean μ is represented in (2.6) by a normal pdf with prior mean μ_a and variance σ_a^2. The prior mean and variance μ_a and σ_a^2 are assigned values by the investigator in accord with his prior information about the parameter μ. If this normal prior pdf is judged an adequate representation of the available prior information, it can be used, as demonstrated above, to obtain the posterior pdf for μ. On the other hand, if the prior information is not adequately represented by a normal prior pdf, another prior pdf that does so will be used by the investigator. To take a specific example, if we have a scalar parameter θ, say a proportion, that by its very nature is limited to the closed interval 0 to 1, it would not be appropriate to employ a normal prior pdf for θ, since a normal pdf does not limit the range of θ to the closed interval 0 to 1. The pdf chosen for θ should be one, say possibly a beta pdf, that can incorporate the available information on the range of θ. Considerations of this sort point up the importance of exercising care and thought in choosing a prior pdf to represent prior information.

As regards the nature of prior information, we recognize that it may be information contained in samples of past data which have been generated in a reasonably scientific manner and which are available for further analysis. When a prior pdf represents information of this kind, we shall term such a prior pdf a "data-based" (DB) prior. In other cases prior information may arise from introspection, casual observation, or theoretical considerations; that is, from sources other than currently available samples of past data of the kind described above. When a prior pdf represents information of this kind, we refer to it as a "nondata-based" (NDB) prior. Although in many

[8] In general, the pdf $p(\boldsymbol{\theta})$ will involve some prior parameters that we have not shown explicitly in order to simplify the notation.

situations prior pdf's represent both DB and NDB information, we think that the distinction between these two kinds of information is worth making, since they obviously have somewhat different characteristics.

It is extremely difficult to formulate general precepts regarding the appropriate uses of the two kinds of prior information mentioned above, since much depends on the objectives of analyses; for example, if an individual wishes to determine how new sample information modifies his own beliefs about parameters of a model and his initial information is NDB, he will, of course, use a NDB prior pdf in conjunction with a likelihood function to obtain a posterior pdf. Then, on comparing his posterior pdf with his NDB prior pdf, he can determine how the information in his sample data has modified his initial NDB beliefs, a fundamental operation in much scientific work. Again, if an economist is carrying through an analysis of sample data in order to make a policy decision, he may indeed incorporate NDB as well as DB prior information in his analysis to ensure that his final decision will be based on all his available information, prior and sample.

Although the above uses of NDB prior information are extremely valuable, it must be noted that one person's NDB prior information can differ from that of another's. In a research situation this is just another way of stating that different investigators may have different views, a not unusual state of affairs; for example, in the early days of Keynesian employment theory there were some old line quantity theorists who argued that the investment multiplier could be negative, zero, or positive. These views conflicted with those of the Keynesians, who argued, on the basis of theoretical considerations and casual observation, that the multiplier is strictly positive. Given a model for observations involving the multiplier and data, it is possible to compute a posterior pdf for the investment multiplier to determine what the information in the data has to say about the value of the multiplier. An analysis of this sort might yield the conclusion that the probability that the multiplier will be negative is negligibly small. Thus information in data can be employed to make comparisons of alternative prior beliefs or hypotheses. Specific techniques for making such comparisons are provided in Chapter 10. In addition, a framework for making a choice among alternative conflicting beliefs or hypotheses which utilizes information in sample data is described and applied.

It is possible that two investigators working with the same model and DB prior information can arrive at different posterior beliefs if they base their prior information on different bodies of past data. These investigators can be brought into agreement by pooling their past samples of data and thereby providing them with the same DB prior information.

Whether prior information is DB or NDB, it is conceivable that there is little prior information; for example, there may be no past sample data

information available. A situation involving NDB prior information may be one in which an investigator has vague ideas about the phenomenon under study and in which case we refer to our prior information as "vague" or "diffuse." If our prior information relates to parameters of a model and is vague or diffuse, we employ a "diffuse" prior pdf in the analysis of our data. Various considerations and principles used in obtaining "diffuse" prior pdf's are discussed in the appendix to this chapter. To illustrate the use of a diffuse prior pdf consider the following example.

Example 2.2. Consider n independent observations $\mathbf{y}' = (y_1, y_2, \ldots, y_n)$ drawn from a normal population with unknown mean μ and known standard deviation $\sigma = \sigma_0$. Assume that our prior information regarding the value of μ is vague or diffuse. To represent knowing little about the value of μ, we follow Jeffreys (see the appendix to this chapter) by taking

$$p(\mu) \propto \text{constant} \qquad -\infty < \mu < \infty \tag{2.13}$$

as our prior pdf.[9] Then the posterior pdf for μ, $p(\mu|\mathbf{y}, \sigma = \sigma_0)$ is given by

$$\begin{aligned} p(\mu|\mathbf{y}, \sigma = \sigma_0) &\propto p(\mu) l(\mu|\mathbf{y}, \sigma = \sigma_0) \qquad -\infty < \mu < \infty \\ &\propto \exp\left[-\frac{n}{2\sigma_0^2}(\mu - \hat{\mu})^2\right], \end{aligned} \tag{2.14}$$

where $l(\mu|\mathbf{y}, \sigma = \sigma_0)$ is the likelihood function and $\hat{\mu} = \sum_{i=1}^n y_i/n$, the sample mean. It is seen that the posterior pdf is normal with mean $\hat{\mu}$ and variance σ_0^2/n. This same result would be obtained in Example 2.1 if there we spread out the normal prior pdf for μ (i.e., allowed $\sigma_a \to \infty$).

When we have NDB prior information that we wish to incorporate in an analysis, the problem of choosing a prior pdf to represent the available prior information must be faced. Ideally, we should like to have a prior pdf represent our prior information as accurately as possible and yet be relatively simple so that mathematical operations can be performed conveniently; for example, in Example 2.1 our prior information about a mean was assumed to be adequately represented by the normal pdf in (2.6), which is relatively simple and mathematically convenient. We shall see from what follows that

[9] This prior pdf is improper; that is $\int_{-\infty}^{\infty} p(\mu)\, d\mu$ is not finite. Jeffreys and others make extensive use of improper prior pdf's to represent "knowing little." Jeffreys remarks in his *Theory of Probability* (3rd ed.). Oxford: Clarendon, 1961, p. 119, that use of improper pdf's poses no difficulty, and in fact Renyi's axioms and his accompanying definition of conditional probability can be used to state Bayes' theorem when improper prior pdf's are employed. See D. V. Lindley, *Introduction to Probabiiity and Statistics from a Bayesian Viewpoint. Part 1. Probability*. Cambridge: Cambridge University Press, 1965, pp. 11, 13, for a brief discussion of this point.

(2.6) is an example of a "natural conjugate" prior pdf.[10] Such prior pdf's are often useful in representing prior information, relatively simple, and mathematically tractable.

We now explain the definition of a natural conjugate prior pdf. Let $p(\mathbf{y}|\boldsymbol{\theta}, n)$ be the pdf for an $n \times 1$ vector of observations $\mathbf{y}$, where $\boldsymbol{\theta}$ is a parameter vector. If $p(\mathbf{y}|\boldsymbol{\theta}, n) = p_1(\mathbf{t}|\boldsymbol{\theta}, n)\, p_2(\mathbf{y})$, where $\mathbf{t}' = (t_1, t_2, \ldots, t_k)$, with $t_i = t_i(\mathbf{y})$ a function of the observations and $p_2(\mathbf{y})$ does not depend on $\boldsymbol{\theta}$, then the t_i's are defined as sufficient statistics.[11] A natural conjugate prior pdf for $\boldsymbol{\theta}$, say $f(\boldsymbol{\theta}|\cdot)$, is given by $f(\boldsymbol{\theta}|\cdot) \propto p_1(\mathbf{t}|\boldsymbol{\theta}, n)$, with the factor of proportionality depending on $\mathbf{t}$ and n but not $\boldsymbol{\theta}$. It is seen that $f(\boldsymbol{\theta}|\cdot)$, defined in this way, has a functional form precisely the same as that as $p_1(\mathbf{t}|\boldsymbol{\theta}, n)$; however, the argument of f is $\boldsymbol{\theta}$ and its $k + 1$ parameters are the elements of $\mathbf{t}$ and n. To represent prior information an investigator assigns values to $\mathbf{t}$ and n, say $\mathbf{t}_0$ and n_0, to obtain $f(\boldsymbol{\theta}|\mathbf{t}_0, n_0) \propto p_1(\mathbf{t}_0|\boldsymbol{\theta}, n_0)$ as his informative prior pdf.[12]

As an example of a natural conjugate prior pdf, consider the data in Example 2.2 with $\sigma = 1$. We have $p(\mathbf{y}|\mu, n) = (\sqrt{2\pi})^{-n} \exp\left[-\frac{1}{2}\sum_{i=1}^n (y_i - \mu)^2\right] = (\sqrt{2\pi})^{-n} \exp\left\{-\frac{1}{2}[(n-1)s^2 + n(\mu - \hat{\mu})^2]\right\}$, where $\hat{\mu} = \sum_{i=1}^n y_i/n$ and $(n-1)s^2 = \sum_{i=1}^n (y_i - \hat{\mu})^2$. Then we can write $p(\mathbf{y}|\mu, n) = p_1(\hat{\mu}|\mu, n)\, p_2(\mathbf{y})$, with $p_1(\hat{\mu}|\mu, n) = \exp\left[-(n/2)(\mu - \hat{\mu})^2\right]$ and $p_2(\mathbf{y}) = (\sqrt{2\pi})^{-n} \exp\left[-(n-1)s^2/2\right]$. Clearly $\hat{\mu}$ is a sufficient statistic for μ, and the natural conjugate prior pdf for μ, $f(\mu|\cdot)$, is: $f(\mu|\hat{\mu}_0, n_0) = c \exp\left[-(n_0/2)(\mu - \hat{\mu}_0)^2\right]$, with $c = \sqrt{n_0/2\pi}$, a normal pdf with prior mean $\hat{\mu}_0$ and prior variance $1/n_0$. To use this prior pdf an investigator should check that it adequately represents his prior information and, if it does, supply values for its parameters $\hat{\mu}_0$ and n_0.

2.4 MARGINAL AND CONDITIONAL POSTERIOR DISTRIBUTIONS FOR PARAMETERS

As with usual joint pdf's, marginal and conditional pdf's can be obtained from a joint posterior pdf; for example, let $\boldsymbol{\theta}$ be partitioned as $\boldsymbol{\theta}' = (\boldsymbol{\theta}_1' : \boldsymbol{\theta}_2')$ and suppose that we want the marginal posterior pdf for $\boldsymbol{\theta}_1$ which may

[10] See H. Raiffa and R. Schlaifer, *Applied Statistical Decision Theory*. Boston: Graduate School of Business Administration, Harvard University, 1961, Chapter 3, for a detailed discussion of natural conjugate prior pdf's.

[11] See, for example, Lindley, *Introduction to Probability and Statistics from a Bayesian Viewpoint. Part 2. Inference*. Cambridge: Cambridge University Press, 1965, pp. 46 ff. for further discussion of sufficient statistics.

[12] Note that when $p(\mathbf{y}|\boldsymbol{\theta}, n) = p_1(\mathbf{t}|\boldsymbol{\theta}, n)\, p_2(\mathbf{y})$ and $p(\boldsymbol{\theta})$ is a prior pdf for $\boldsymbol{\theta}$, the posterior pdf, $p(\boldsymbol{\theta}|\mathbf{y}, n)$ is $p(\boldsymbol{\theta}|\mathbf{y}, n) \propto p(\boldsymbol{\theta})\, p(\mathbf{y}|\boldsymbol{\theta}, n) \propto p(\boldsymbol{\theta})\, p_1(\mathbf{t}|\boldsymbol{\theta}, n)$. As explained in Section 2.11, for large n, $p(\boldsymbol{\theta}|\mathbf{y}, n)$ is approximately proportional to $p_1(\mathbf{t}|\boldsymbol{\theta}, n)$ under rather general conditions. Thus in large samples the posterior pdf assumes the form of $p_1(\mathbf{t}|\boldsymbol{\theta}, n)$, which is also the form of the natural conjugate prior pdf.

contain one or several elements of $\boldsymbol{\theta}$. This marginal posterior pdf, $p(\boldsymbol{\theta}_1|\mathbf{y})$, is readily obtained as follows:

$$p(\boldsymbol{\theta}_1|\mathbf{y}) = \int_{R_{\theta_2}} p(\boldsymbol{\theta}_1, \boldsymbol{\theta}_2|\mathbf{y})\, d\boldsymbol{\theta}_2 \tag{2.15a}$$

$$= \int_{R_{\theta_2}} p(\boldsymbol{\theta}_1|\boldsymbol{\theta}_2, \mathbf{y})\, p(\boldsymbol{\theta}_2|\mathbf{y})\, d\boldsymbol{\theta}_2, \tag{2.15b}$$

where R_{θ_2} denotes the region of $\boldsymbol{\theta}_2$ and $p(\boldsymbol{\theta}_1|\boldsymbol{\theta}_2, \mathbf{y})$ is the conditional posterior pdf for $\boldsymbol{\theta}_1$, given $\boldsymbol{\theta}_2$ and the sample information $\mathbf{y}$. Equation 2.15*b* illustrates the fact that the marginal posterior pdf for $\boldsymbol{\theta}_1$ may be viewed as an averaging of conditional posterior pdf's, $p(\boldsymbol{\theta}_1|\boldsymbol{\theta}_2, \mathbf{y})$, with the weight function being the marginal posterior pdf for $\boldsymbol{\theta}_2$, $p(\boldsymbol{\theta}_2|\mathbf{y})$. The integration shown in (2.15) provides an extremely useful way of getting rid of "nuisance" parameters, that is parameters that are not of special interest.

Example 2.3. Assume that we have n independent observations, $\mathbf{y}' = (y_1, y_2, \ldots, y_n)$, from a normal population with unknown mean μ and unknown standard deviation σ. If our prior information about values of the mean and standard deviation is vague or diffuse, we can represent this state of our initial information by taking our prior pdf as

$$p(\mu, \sigma)\, d\mu\, d\sigma \propto \frac{1}{\sigma}\, d\mu\, d\sigma, \qquad \begin{matrix} -\infty < \mu < \infty. \\ 0 < \sigma < \infty. \end{matrix} \tag{2.16}$$

In (2.16) we have assumed μ and σ to be independently distributed, a priori, with μ and $\log \sigma$ each uniformly distributed [see the appendix at the end of this chapter for further discussion of (2.16)]. Then the joint posterior pdf for μ and σ is

$$\begin{aligned} p(\mu, \sigma|\mathbf{y}) &\propto p(\mu, \sigma) l(\mu, \sigma|\mathbf{y}), \qquad \begin{matrix} -\infty < \mu < \infty, \\ 0 < \sigma < \infty. \end{matrix} \\ &\propto \sigma^{-(n+1)} \exp\left\{-\frac{1}{2\sigma^2}[\nu s^2 + n(\mu - \hat{\mu})^2]\right\}, \end{aligned} \tag{2.17}$$

where $l(\mu, \sigma|\mathbf{y}) \propto \sigma^{-n} \exp[-1/2\sigma^2 \sum_{i=1}^n (y_i - \mu)^2]$ is the likelihood function, $\nu = n - 1$, $\hat{\mu} = \sum_{i=1}^n y_i/n$, and $\nu s^2 = \sum_{i=1}^n (y_i - \hat{\mu})^2$. From the form of (2.17) it is clear that the conditional posterior pdf for μ given σ and the sample information, that is, $p(\mu|\sigma, \mathbf{y})$, is in the univariate normal form with conditional posterior mean and variance $E(\mu|\sigma, \mathbf{y}) = \hat{\mu}$ and $\mathrm{Var}(\mu|\sigma, \mathbf{y}) = \sigma^2/n$, respectively. Although these conditional results are of interest, it is clear that the conditional pdf for μ given σ and $\mathbf{y}$ depends critically on σ whose value is unknown. If we are mainly interested in μ, σ is a nuisance parameter

and, as stated above, such a parameter can generally be integrated out of the posterior pdf. In the present instance we have[13]

$$\begin{aligned} p(\mu|\mathbf{y}) &= \int_0^\infty p(\mu, \sigma|\mathbf{y})\, d\sigma \\ &\propto \int_0^\infty \sigma^{-(n+1)} \exp\left\{-\frac{1}{2\sigma^2}[\nu s^2 + n(\mu - \hat{\mu})^2]\right\} d\sigma \\ &\propto \{\nu s^2 + n(\mu - \hat{\mu})^2\}^{-(\nu+1)/2}. \end{aligned} \tag{2.18}$$

From (2.18) it is seen that the marginal posterior pdf for μ is in the form of a univariate Student t pdf[14] with mean $\hat{\mu}$; that is, the random variable

$$t = \frac{\mu - \hat{\mu}}{s/\sqrt{n}}$$

has a Student t pdf with $\nu = n - 1$ degrees of freedom.

If the parameter σ is of interest, we can integrate $p(\mu, \sigma|\mathbf{y})$ in (2.17) with respect to μ to obtain the marginal posterior pdf for σ, namely,[15]

$$\begin{aligned} p(\sigma|\mathbf{y}) &= \int_{-\infty}^\infty p(\mu, \sigma|\mathbf{y})\, d\mu \\ &\propto \sigma^{-(\nu+1)} \exp\left(-\frac{\nu s^2}{2\sigma^2}\right), \qquad 0 < \sigma < \infty. \end{aligned} \tag{2.19}$$

This posterior pdf for σ is in the "inverted gamma" form (see Appendix A) and will be proper for $\nu > 0$. Further, from properties of (2.19) we have

$$E(\sigma|\mathbf{y}) = \frac{\sqrt{\nu/2}\, \Gamma[(\nu - 1)/2]}{\Gamma(\nu/2)} s \qquad \text{for } \nu > 1$$

and

$$\mathrm{Var}(\sigma|\mathbf{y}) = \frac{\nu s^2}{\nu - 2} - [E(\sigma|\mathbf{y})]^2 \qquad \text{for } \nu > 2.$$

The modal value of the posterior pdf in (2.19) is $s\sqrt{\nu/(\nu + 1)}$.

[13] Note that $\int_0^\infty \sigma^{-(n+1)} \exp(-a/2\sigma^2)\, d\sigma = 2^{(n-2)/2}\Gamma(n/2)/a^{n/2}$. This result is easily obtained by letting $x = a/2\sigma^2$. Then the integral becomes $(2^{(n-2)/2}/a^{n/2})\int_0^\infty x^{(n-2)/2} e^{-x}\, dx = 2^{(n-2)/2}\Gamma(n/2)/a^{n/2}$, where Γ denotes the gamma function. In using this result in (2.18), $a = \nu s^2 + n(\mu - \hat{\mu})^2$ and the factor $2^{(n-2)/2}\Gamma(n/2)$ is absorbed in the factor of proportionality.

[14] See Appendix A at the end of the book for properties of this pdf.

[15] In integrating (2.17) with respect to μ, note that, for given σ, (2.17) is in the univariate normal form. Letting $z = \sqrt{n}(\mu - \hat{\mu})/\sigma$, $dz \propto d\mu/\sigma$, and thus (2.17) becomes $1/\sigma^n \exp(-\nu s^2/2\sigma^2) \exp(-z^2/2)\, d\sigma\, dz$ from which (2.19) follows.

2.5 POINT ESTIMATES FOR PARAMETERS

From Section 2.3 above it is seen that the Bayesian approach yields the complete posterior pdf for the parameter vector $\boldsymbol{\theta}$. If we wish, we may characterize this distribution in terms of a small number of measures, say measures of central tendency, dispersion, and skewness, with a measure of central tendency to serve as a point estimate. The problem of choosing a single measure of central tendency is a well-known problem in descriptive statistics. In some circumstances we may have a loss function, say $L = L(\boldsymbol{\theta}, \hat{\boldsymbol{\theta}})$, where $\hat{\boldsymbol{\theta}} = \hat{\boldsymbol{\theta}}(\mathbf{y})$ is a point estimate depending on the given sample observations $\mathbf{y}' = (y_1, \ldots, y_n)$. Since $\boldsymbol{\theta}$ is considered random, L is random. One generally used principle, which generates point estimates and which is in accord with the expected utility hypothesis, is to find the value of $\hat{\boldsymbol{\theta}}$ that minimizes the mathematical expectation of the loss function; that is

$$\min_{\hat{\boldsymbol{\theta}}} EL(\boldsymbol{\theta}, \hat{\boldsymbol{\theta}}) = \min_{\hat{\boldsymbol{\theta}}} \int_{R_{\boldsymbol{\theta}}} L(\boldsymbol{\theta}, \hat{\boldsymbol{\theta}})\, p(\boldsymbol{\theta}|\mathbf{y})\, d\boldsymbol{\theta}, \tag{2.20}$$

which assumes that $EL(\boldsymbol{\theta}, \hat{\boldsymbol{\theta}})$ is finite and that a minimum exists.

As an important illustration of (2.20), consider the case of a quadratic loss function, $L = (\boldsymbol{\theta} - \hat{\boldsymbol{\theta}})'C(\boldsymbol{\theta} - \hat{\boldsymbol{\theta}})$, where C is a known nonstochastic positive definite symmetric matrix. Then the posterior expectation of the quadratic loss function is[16]

$$\begin{aligned} EL &= E(\boldsymbol{\theta} - \hat{\boldsymbol{\theta}})'C(\boldsymbol{\theta} - \hat{\boldsymbol{\theta}}) \\ &= E[(\boldsymbol{\theta} - E\boldsymbol{\theta}) - (\hat{\boldsymbol{\theta}} - E\boldsymbol{\theta})]'C[(\boldsymbol{\theta} - E\boldsymbol{\theta}) - (\hat{\boldsymbol{\theta}} - E\boldsymbol{\theta})] \\ &= E(\boldsymbol{\theta} - E\boldsymbol{\theta})'C(\boldsymbol{\theta} - E\boldsymbol{\theta}) + (\hat{\boldsymbol{\theta}} - E\boldsymbol{\theta})'C(\hat{\boldsymbol{\theta}} - E\boldsymbol{\theta}). \end{aligned} \tag{2.21}$$

The first term of this last expression does not involve $\hat{\boldsymbol{\theta}}$. The remaining term $(\hat{\boldsymbol{\theta}} - E\boldsymbol{\theta})'C(\hat{\boldsymbol{\theta}} - E\boldsymbol{\theta})$ is nonstochastic and will be minimized if we take $\hat{\boldsymbol{\theta}} = E\boldsymbol{\theta}$, given that C is positive definite. Thus for positive definite quadratic loss functions the mean $E\boldsymbol{\theta}$ of the posterior pdf $p(\boldsymbol{\theta}|\mathbf{y})$, if it exists, is an optimal point estimate. For other loss functions similar analysis can be performed to yield optimal point estimates.

Example 2.4. Consider Example 2.1 when our loss function is $L(\mu, \breve{\mu}) = c(\mu - \breve{\mu})^2$, where $\breve{\mu}$ is a point estimate, and c is a positive constant. Then, taking $\breve{\mu} = E\mu = (h_o\hat{\mu} + h_a\mu_a)/(h_o + h_a)$, the mean of the posterior pdf for μ will minimize $EL = cE(\mu - \breve{\mu})^2$.

[16] In the second line of (2.21) the posterior mean $E\boldsymbol{\theta}$ has been subtracted from $\boldsymbol{\theta}$ and added to $\hat{\boldsymbol{\theta}}$ which does not affect the value of EL. In going from the second line of (2.21) to the third, the cross terms disappear; that is $E(\boldsymbol{\theta} - E\boldsymbol{\theta})'C(\hat{\boldsymbol{\theta}} - E\boldsymbol{\theta}) = 0$ since $E(\boldsymbol{\theta} - E\boldsymbol{\theta}) = \mathbf{0}$.

Example 2.5. Suppose that our loss function is $L = |\theta - \hat{\theta}|$ and the posterior pdf for θ is a proper continuous pdf, $p(\theta|\mathbf{y})$, with $a \leq \theta \leq b$ where a and b are known. Then the point estimate $\hat{\theta}$ which minimizes expected loss can be found as follows:

$$\begin{aligned} EL &= \int_a^b |\theta - \hat{\theta}|\, p(\theta|\mathbf{y})\, d\theta \\ &= \int_a^{\hat{\theta}} (\hat{\theta} - \theta)\, p(\theta|\mathbf{y})\, d\theta + \int_{\hat{\theta}}^b (\theta - \hat{\theta})\, p(\theta|\mathbf{y})\, d\theta \\ &= \hat{\theta}\, P(\hat{\theta}|\mathbf{y}) - \int_a^{\hat{\theta}} \theta\, p(\theta|\mathbf{y})\, d\theta + \int_{\hat{\theta}}^b \theta p\, (\theta|\mathbf{y})\, d\theta - \hat{\theta}[1 - P(\hat{\theta}|\mathbf{y})], \end{aligned}$$

where $P(\hat{\theta}|\mathbf{y}) = \int_a^{\hat{\theta}} p(\theta|\mathbf{y})\, d\theta$ is the cumulative posterior distribution function. Then, on differentiation[17] with respect to $\hat{\theta}$ and setting the derivative equal to zero, we have

$$\frac{dEL}{d\hat{\theta}} = P(\hat{\theta}|\mathbf{y}) - 1 + P(\hat{\theta}|\mathbf{y}) = 0$$

or

$$P(\hat{\theta}|\mathbf{y}) = \tfrac{1}{2}.$$

The $\hat{\theta}$ which satisfies this necessary condition for a minimum is the median of the posterior pdf. That this value for $\hat{\theta}$ produces a minimum of EL can be established by noting that $d^2EL/d\hat{\theta}^2$ is strictly positive for $\hat{\theta}$ = median of the posterior pdf. Thus for the absolute error function $L = |\theta - \hat{\theta}|$ the median of the posterior pdf is an optimal point estimate.

Next, we review a relationship between Bayesian and sampling theory approaches to point estimation. Let $\tilde{\boldsymbol{\theta}} = \tilde{\boldsymbol{\theta}}(\mathbf{y})$ be a sampling theory estimator.[18] The risk function associated with the estimator $\tilde{\boldsymbol{\theta}}$ is given by

$$r(\boldsymbol{\theta}) = \int_{R_{\mathbf{y}}} L(\boldsymbol{\theta}, \tilde{\boldsymbol{\theta}})\, p(\mathbf{y}|\boldsymbol{\theta})\, d\mathbf{y}, \tag{2.22}$$

where $L(\boldsymbol{\theta}, \tilde{\boldsymbol{\theta}})$ is a loss function, $p(\mathbf{y}|\boldsymbol{\theta})$ is a proper pdf for $\mathbf{y}$, given $\boldsymbol{\theta}$, and the integral in (2.22) is assumed to converge. As indicated explicitly in (2.22), the risk function depends on the value of the unknown parameter vector $\boldsymbol{\theta}$. Since it is impossible to find a $\tilde{\boldsymbol{\theta}}$ which minimizes $r(\boldsymbol{\theta})$ for all possible values of $\boldsymbol{\theta}$,[19]

[17] It is assumed that the needed derivatives exist for $a \leq \theta \leq b$.

[18] As is well known, the term "estimator" indicates that $\tilde{\boldsymbol{\theta}} = \tilde{\boldsymbol{\theta}}(\mathbf{y})$ is regarded as a random quantity.

[19] For example, if we take $\tilde{\boldsymbol{\theta}} = \mathbf{b}$, a vector of constants, this "estimator" will have smaller risk when $\boldsymbol{\theta} = \mathbf{b}$ than any other estimator and thus no single estimator can minimize $r(\boldsymbol{\theta})$ for all $\boldsymbol{\theta}$.

we shall seek the estimator that minimizes average risk when average risk is defined by

$$Er(\tilde{\boldsymbol{\theta}}) = \int_{R_\theta} p(\boldsymbol{\theta})\, r(\boldsymbol{\theta})\, d\boldsymbol{\theta}. \tag{2.23}$$

In (2.23) $p(\boldsymbol{\theta})$ is a "weighting function" used to weight the performance of $\tilde{\boldsymbol{\theta}}$, an estimator, in regions of the parameter space. Then our problem is to find the estimator that minimizes average risk, that is, that solves the following problem:

$$\min_{\tilde{\boldsymbol{\theta}}} Er(\tilde{\boldsymbol{\theta}}) = \min_{\tilde{\boldsymbol{\theta}}} \int_{R_\theta} \int_{R_\mathbf{y}} p(\boldsymbol{\theta})\, L(\boldsymbol{\theta}, \tilde{\boldsymbol{\theta}})\, p(\mathbf{y}|\boldsymbol{\theta})\, d\mathbf{y}\, d\boldsymbol{\theta}. \tag{2.24}$$

Given that the integrand of (2.24) is non-negative, we can interchange the order of integration and, using $p(\boldsymbol{\theta})\, p(\mathbf{y}|\boldsymbol{\theta}) = p(\mathbf{y})\, p(\boldsymbol{\theta}|\mathbf{y})$, write (2.24) as

$$\min_{\tilde{\boldsymbol{\theta}}} Er(\tilde{\boldsymbol{\theta}}) = \min_{\tilde{\boldsymbol{\theta}}} \int_{R_\mathbf{y}} \left[\int_{R_\theta} L(\boldsymbol{\theta}, \tilde{\boldsymbol{\theta}})\, p(\boldsymbol{\theta}|\mathbf{y})\, d\boldsymbol{\theta} \right] p(\mathbf{y})\, d\mathbf{y}. \tag{2.25}$$

The $\tilde{\boldsymbol{\theta}}$ that minimizes the expression in square brackets will minimize expected risk, provided that $Er(\boldsymbol{\theta})$ is finite, and this estimator is, by definition, the Bayes estimator.[20] Therefore, if a specification is made for the seriousness of estimation errors in the form of a loss function, $L(\boldsymbol{\theta}, \tilde{\boldsymbol{\theta}})$, and for the weighting of parameter values over which good performance is sought by a choice of $p(\boldsymbol{\theta})$, then on an average risk criterion the Bayesian estimator gives the best performance in repeated sampling.[21]

[20] When the double integral in (2.25) converges, and thus $Er(\boldsymbol{\theta})$ is finite, the $\tilde{\boldsymbol{\theta}}$ solving the minimization problem in (2.25) will also be a solution to the minimization problem in (2.20). If the double integral in (2.25) diverges, however, the minimization problem in (2.25) will have no solution, but still a solution to the problem in (2.20) often exists. When this is the situation, the solution to the minimization problem in (2.20) has been called a quasi-Bayesian estimator. Quasi-Bayesian estimators often arise when improper diffuse prior pdf's are employed along with usual loss functions, for example, quadratic loss functions. For further discussion of this point see H. Thornber, "Applications of Decision Theory to Econometrics," unpublished doctoral dissertation, University of Chicago, 1966, and M. Stone, "Generalized Bayes Decision Functions, Admissibility and the Exponential Family," *Ann. Math. Statist.*, **38**, 818–822 (1967).

[21] The relevance of the criterion of performance in *repeated samples* is questioned by some. They want an estimate that is appropriate for the given sample data and thus will solve the problem in (2.20) which involves no averaging over the sample space $R_\mathbf{y}$. When the solution to (2.20) is identical to the solution of (2.24), as it often is, this consideration makes no practical difference. On the other hand, many sampling theorists object to the introduction of the "weighting function" (prior pdf) $p(\boldsymbol{\theta})$ and therefore do not attach much importance to the minimal average risk property of Bayesian estimators.

2.6 BAYESIAN INTERVALS AND REGIONS FOR PARAMETERS

Given that the posterior pdf $p(\boldsymbol{\theta}|\mathbf{y})$ has been obtained, it is generally possible to compute the probability that the parameter vector $\boldsymbol{\theta}$ lies in a particular subregion, $\bar{R}$, of the parameter space as follows:

$$\Pr(\boldsymbol{\theta} \in \bar{R}|\mathbf{y}) = \int_{\bar{R}} p(\boldsymbol{\theta}|\mathbf{y})\, d\boldsymbol{\theta}. \tag{2.26}$$

The probability in (2.26) measures the degree of belief that $\boldsymbol{\theta} \in \bar{R}$ given the sample and prior information.

If we fix the probability in (2.26), say at 0.95, it is generally possible to find a region (or interval) $\bar{R}$, not necessarily unique, such that (2.26) holds. In many important problems with unimodal posterior pdf's, it is possible to obtain a unique region (or interval) $\bar{R}$ by imposing the conditions that its probability content be β, say $\beta = 0.95$, and that the posterior pdf's values over the region or interval be not less than those relating to any other region with the same probability content; for example, for unimodal symmetric posterior pdf's the region or interval with given probability content β, which is centered at the modal value of the posterior pdf is the Bayesian "highest posterior density" region or interval.[22]

Example 2.6. Consider Example 2.3 in which it was found that the posterior pdf of $(\mu - \hat{\mu})/s'$, where $s' = s/\sqrt{n}$ is a Student t pdf with $\nu = n - 1$ degrees of freedom. Thus the probability that μ will lie in a particular interval, say $\hat{\mu} \pm ks'$, with k given, can easily be evaluated by using tables of the Student t distribution.[23] Alternatively, k can be determined so that the posterior probability that $\hat{\mu} - ks' < \mu < \hat{\mu} + ks'$ is a given value, say $\beta = 0.90$. The interval so obtained, $\hat{\mu} \pm ks'$, is numerically exactly the same as a sampling theory confidence interval but is given an entirely different interpretation in

[22] See G. E. P. Box and G. C. Tiao, "Multiparameter Problems from a Bayesian Point of View," *Ann. Math. Statist.*, **36**, 1468–1482 (1965), for further discussion of "highest posterior density" Bayesian regions. In general, if we seek a "highest" interval with probability content β for a unimodal pdf, $p(x)$, it can be obtained by solving the following problem: minimize $(b - a)$ subject to $\int_a^b p(x)\, dx = \beta$. On differentiating $b - a + \lambda[\int_a^b p(x)\, dx - \beta]$, where λ is a Lagrange multiplier, partially with respect to a and b and setting these derivatives equal to zero, yields $1 + \lambda p(a) = 0$ and $1 + \lambda p(b) = 0$, and thus a and b must be such that $p(a) = p(b)$ for these necessary conditions to be satisfied. Determining a and b such that $\int_a^b p(x)\, dx = \beta$ with $p(a) = p(b)$ leads to a shortest interval with probability content β, and this interval will be a "highest" interval given that $p(x)$ is unimodal. In the example above in which z is a standardized normal variable $p(z)$ is unimodal and symmetric about zero. Thus taking $a = -z_\beta$ and $b = z_\beta$ satisfies the condition $p(a) = p(b)$.

[23] See, for example, N. V. Smirnov, *Tables for the Distribution and Density Function of t-Distribution.* New York: Pergamon, 1961.

the Bayesian approach. As is well known, the sampling theorist regards his interval as random and having probability $\beta = 0.90$ of covering the true value of the parameter. For the Bayesian whose work is conditional on the sample observations the interval $\hat{\mu} \pm ks'$ is regarded as given and his statement is that the posterior probability that μ will lie in the interval is $\beta = 0.90$. Note that the probability statements being made by the sampling theorist and the Bayesian are not identical.

2.7 MARGINAL DISTRIBUTION OF THE OBSERVATIONS

In certain instances it is of interest to obtain the marginal pdf for the observations, denoted by $p(\mathbf{y})$. This pdf can be obtained as follows:

$$p(\mathbf{y}) = \int_{R_\theta} p(\boldsymbol{\theta}, \mathbf{y})\, d\boldsymbol{\theta} = \int_{R_\theta} p(\mathbf{y}|\boldsymbol{\theta})\, p(\boldsymbol{\theta})\, d\boldsymbol{\theta}. \tag{2.27}$$

The second line of (2.27) indicates that the marginal pdf of the observations is an average of the conditional pdf $p(\mathbf{y}|\boldsymbol{\theta})$ with the prior pdf $p(\boldsymbol{\theta})$ serving as the weighting function.

Example 2.7. Let y_1 be an observation from a normal distribution with unknown mean μ and known standard deviation $\sigma = \sigma_0$. Then

$$p(y_1|\mu, \sigma = \sigma_0) = \frac{1}{\sqrt{2\pi}\,\sigma_0} \exp\left[-\frac{1}{2\sigma_0^2}(y_1 - \mu)^2\right].$$

If the prior pdf for μ is $p(\mu) = (\sqrt{2\pi}\,\sigma_a)^{-1} \exp\,[-(2\sigma_a^2)^{-1}(\mu - \mu_a)^2]$, $-\infty < \mu < \infty$, where μ_a and σ_a are the prior mean and standard deviation, respectively, the marginal pdf for y_1 is

$$p(y_1) = \int_{-\infty}^{\infty} p(y_1|\mu, \sigma = \sigma_0)\, p(\mu)\, d\mu = (2\pi\sigma_0\sigma_a)^{-1} \int_{-\infty}^{\infty} \exp\left\{-\frac{1}{2}\left[\frac{(y_1 - \mu)^2}{\sigma_0^2} + \frac{(\mu - \mu_a)^2}{\sigma_a^2}\right]\right\} d\mu.$$

On completing the square for μ in the exponent and performing the integration,[24] the result is

$$p(y_1) = \frac{1}{\sqrt{2\pi(\sigma_a^2 + \sigma_0^2)}} \exp\left[-\frac{(y_1 - \mu_a)^2}{2(\sigma_a^2 + \sigma_0^2)}\right].$$

[24] A less tedious way to derive $p(y_1)$ is to write $y_1 = \mu + \epsilon$ with ϵ, a scalar random variable normally distributed and independent of μ, with zero mean and variance σ_0^2. Then the mean of y_1 is μ_a, the mean of μ, and the variance of y_1 is $\sigma_a^2 + \sigma_0^2$. Since y_1 is linearly related to μ and ϵ, it will have a normal pdf.

Thus the marginal pdf for y_1 is normal with mean μ_a, the prior mean for μ, and variance $\sigma_a^2 + \sigma_0^2$. Since μ_a, σ_a^2, and σ_0^2 are assumed known, it is possible to use $p(y_1)$ to make probability statements about y_1, a fact that is often useful before y_1 is observed.

2.8 PREDICTIVE PROBABILITY DENSITY FUNCTIONS

On many occasions, given our sample information $\mathbf{y}$, we are interested in making inferences about other observations that are still unobserved, one part of the problem of prediction. In the Bayesian approach the pdf for the as yet unobserved observations, given our sample information, can be obtained and is known as the predictive pdf; for example, let $\tilde{\mathbf{y}}$ represent a vector of as yet unobserved observations. Write

$$p(\tilde{\mathbf{y}}, \boldsymbol{\theta}|\mathbf{y}) = p(\tilde{\mathbf{y}}|\boldsymbol{\theta}, \mathbf{y})\, p(\boldsymbol{\theta}|\mathbf{y}) \tag{2.28}$$

as the joint pdf for $\tilde{\mathbf{y}}$ and a parameter vector $\boldsymbol{\theta}$, given the sample information $\mathbf{y}$. On the right of (2.28) $p(\tilde{\mathbf{y}}|\boldsymbol{\theta}, \mathbf{y})$ is the conditional pdf for $\tilde{\mathbf{y}}$, given $\boldsymbol{\theta}$ and $\mathbf{y}$, whereas $p(\boldsymbol{\theta}|\mathbf{y})$ is the conditional pdf for $\boldsymbol{\theta}$ given $\mathbf{y}$, that is, the posterior pdf for $\boldsymbol{\theta}$. To obtain the predictive pdf, $p(\tilde{\mathbf{y}}|\mathbf{y})$, we merely integrate (2.28) with respect to $\boldsymbol{\theta}$; that is

$$p(\tilde{\mathbf{y}}|\mathbf{y}) = \int_{R_\theta} p(\tilde{\mathbf{y}}, \boldsymbol{\theta}|\mathbf{y})\, d\boldsymbol{\theta} = \int_{R_\theta} p(\tilde{\mathbf{y}}|\boldsymbol{\theta}, \mathbf{y})\, p(\boldsymbol{\theta}|\mathbf{y})\, d\boldsymbol{\theta}. \tag{2.29}$$

The second line of (2.29) indicates that the predictive pdf can be viewed as an average of conditional predictive pdf's, $p(\tilde{\mathbf{y}}|\boldsymbol{\theta}, \mathbf{y})$, with the posterior pdf for $\boldsymbol{\theta}$, $p(\boldsymbol{\theta}|\mathbf{y})$ serving as the weighting function.

Example 2.8. In Example 2.2 we had n independent observations $\mathbf{y}' = (y_1, y_2, \ldots, y_n)$ from a normal population with unknown mean μ and known standard deviation $\sigma = \sigma_0$. With diffuse prior information about μ, the posterior pdf [see (2.14)] was found to be normal with mean $\hat{\mu}$, the sample mean, and variance σ_0^2/n. We now wish to obtain the predictive pdf for a new observation, say $\tilde{y}_{n+1}$ which has not yet been observed. The two factors in the integrand of the second line of (2.29) are

$$p(\tilde{y}_{n+1}|\mu, \sigma = \sigma_0, \mathbf{y}) \propto \exp\left[-\frac{1}{2\sigma_0^2}(\tilde{y}_{n+1} - \mu)^2\right]$$

and from (2.14)

$$p(\mu|\sigma = \sigma_0, \mathbf{y}) \propto \exp\left[-\frac{n}{2\sigma_0^2}(\mu - \hat{\mu})^2\right], \qquad -\infty < \mu < \infty.$$

Then from (2.29)

$$p(\tilde{y}_{n+1}|\mathbf{y}) = \int_{-\infty}^{\infty} p(\tilde{y}_{n+1}|\mu, \sigma = \sigma_0, \mathbf{y})\, p(\mu|\sigma = \sigma_0, \mathbf{y})\, d\mu$$

$$\propto \int_{-\infty}^{\infty} \exp\left[-\frac{1}{2\sigma_0^2}\left[(\tilde{y}_{n+1} - \mu)^2 + n(\mu - \hat{\mu})^2\right]\right] d\mu. \tag{2.30}$$

On completing the square on μ in this last expression[25] and integrating (2.30) with respect to μ, the predictive pdf for $\tilde{y}_{n+1}$ is

$$p(\tilde{y}_{n+1}|\mathbf{y}) \propto \exp\left[-\frac{n}{2(n+1)\sigma_0^2}(\tilde{y}_{n+1} - \hat{\mu})^2\right]. \tag{2.31}$$

It is seen that $\tilde{y}_{n+1}$ is normally distributed with mean $E(\tilde{y}_{n+1}|\mathbf{y}) = \hat{\mu}$, the sample mean, and variance $\mathrm{Var}(\tilde{y}_{n+1}|\mathbf{y}) = \sigma_0^2(n+1)/n$. The pdf in (2.31) can, of course, be employed to make probability statements about $\tilde{y}_{n+1}$ given $\mathbf{y}$.

2.9 POINT PREDICTION

The predictive pdf, $p(\tilde{\mathbf{y}}|\mathbf{y})$, can be used to obtain a point prediction; for example, we can use a measure of central tendency, say the mean or modal value, as a point prediction, or, if we have a loss function $L = L(\tilde{\mathbf{y}}, \hat{\mathbf{y}})$, where $\hat{\mathbf{y}}$ is a point prediction for $\tilde{\mathbf{y}}$, we can seek the vector $\hat{\mathbf{y}}$ that minimizes the mathematical expectation of the loss function; that is

$$\min_{\hat{\mathbf{y}}} \int_{R_{\tilde{\mathbf{y}}}} L(\tilde{\mathbf{y}}, \hat{\mathbf{y}})\, p(\tilde{\mathbf{y}}|\mathbf{y})\, d\tilde{\mathbf{y}}. \tag{2.32}$$

If a solution to the problem in (2.32) exists, it is an optimal point prediction in the sense of minimizing expected loss. Analysis similar to that presented in Section 2.5 on point estimation provides the result that the mean of the predictive pdf is optimal if our loss function is quadratic; that is, if $L(\tilde{\mathbf{y}}, \hat{\mathbf{y}}) = (\tilde{\mathbf{y}} - \hat{\mathbf{y}})'Q(\tilde{\mathbf{y}} - \hat{\mathbf{y}})$, with Q a positive definite symmetric matrix, then taking $\hat{\mathbf{y}} = E(\tilde{\mathbf{y}}|\mathbf{y})$ as our point prediction provides minimal expected loss; for example, in Example 2.8 the mean of the predictive pdf is the sample mean $\hat{\mu}$, and this is an optimal point prediction for $\tilde{y}_{n+1}$, given that our loss function is of the form $L(\tilde{y}_{n+1}, \hat{y}_{n+1}) = c(\tilde{y}_{n+1} - \hat{y}_{n+1})^2$, $c > 0$. For other loss functions

[25] That is, $(\tilde{y}_{n+1} - \mu)^2 + n(\mu - \hat{\mu})^2 = \tilde{y}_{n+1}^2 + (n+1)\mu^2 - 2\mu(\tilde{y}_{n+1} + n\hat{\mu}) + n\hat{\mu}^2 = (n+1)[\mu^2 - 2\mu(\tilde{y}_{n+1} + n\hat{\mu})/(n+1)] + n\hat{\mu}^2 + \tilde{y}_{n+1}^2 = (n+1)[\mu - (\tilde{y}_{n+1} + n\hat{\mu})/(n+1)]^2 + n(\tilde{y}_{n+1} - \hat{\mu})^2/(n+1)$. On substituting this last expression in the second line of (2.30), the integration with respect to μ can be done readily to yield (2.31). An alternative, simpler derivation of (2.31) is obtained from noting that $\tilde{y}_{n+1} = \mu + \epsilon_{n+1}$ with the normal random error ϵ_{n+1} independent of μ, given $\mathbf{y}$, with mean zero and known variance σ_0^2. Since both $\mu|\mathbf{y}$ and ϵ_{n+1} are normal, $\tilde{y}_{n+1}$ has a normal pdf with mean $E\tilde{y}_{n+1}|\mathbf{y} = E\mu|\mathbf{y} = \hat{\mu}$, since $E\mu|\mathbf{y} = \hat{\mu}$ from (2.14), and $\mathrm{Var}(\tilde{y}_{n+1}|\mathbf{y}) = \mathrm{Var}(\mu|\mathbf{y}) + \mathrm{Var}\,\epsilon_{n+1} = \sigma_0^2/n + \sigma_0^2 = \sigma_0^2(n+1)/n$.

similar analysis can be performed to obtain optimal point predictions, of course under the assumption that a solution to the problem in (2.32) exists.

2.10 PREDICTION REGIONS AND INTERVALS

Given that we have the predictive pdf, $p(\tilde{\mathbf{y}}|\mathbf{y})$, we can, for a given region (or interval) $\bar{R}$, generally evaluate

$$\Pr(\tilde{\mathbf{y}} \in \bar{R}|\mathbf{y}) = \int_{\bar{R}} p(\tilde{\mathbf{y}}|\mathbf{y})\, d\tilde{\mathbf{y}}, \tag{2.33}$$

where $\bar{R}$ is a subspace of $R_{\tilde{\mathbf{y}}}$, the space of the elements of $\tilde{\mathbf{y}}$. In (2.33) we have the probability that the future observation vector $\tilde{\mathbf{y}}$ will lie in the region $\bar{R}$. Alternatively, given a stated probability in (2.33), we can seek a region $\bar{R}$ such that (2.33) is satisfied. As with regions for parameters in Section 2.6, this region can be made unique for unimodal pdf's if we require it to be a "highest predictive density" region; that is, a region with the given probability content and such that the predictive pdf's values over the region are not less than those relating to any other region with the same probability content.

Example 2.9. In Example 2.8 the predictive pdf for $\tilde{y}_{n+1}$ in (2.31) is normal with mean $\hat{\mu}$ and variance $\sigma_0^2(n+1)/n$. Then $z = (\tilde{y}_{n+1} - \hat{\mu})/\bar{\sigma}_0$, with $\bar{\sigma}_0 = \sigma_0\sqrt{(n+1)/n}$, has a normal pdf with zero mean and unit variance. From tables of the standardized normal distribution we can find the $\Pr\{a < z < b\}$, where a and b are given constants. The statement $a < z < b$ is equivalent to $\hat{\mu} + a\bar{\sigma}_0 < \tilde{y}_{n+1} < \hat{\mu} + b\bar{\sigma}_0$ and thus the probability that $\tilde{y}_{n+1}$ will satisfy these inequalities is the same as $\Pr\{a < z < b\}$. On the other hand, if we are required to find a and b such that $\Pr\{a < z < b\} = \beta$, where β is given, it is clear that there are many possible values for a and b such that $\Pr\{a < z < b\} = \beta$. The requirement that the interval be a "highest" interval leads to a unique a and b, namely, $a = -z_\beta$ and $b = z_\beta$, where the area over the interval $-z_\beta$ to z_β is just β.

2.11 SOME LARGE SAMPLE PROPERTIES OF BAYESIAN POSTERIOR PDF'S

In this section we discuss briefly some large sample properties of posterior pdf's.[26] First, let us consider the posterior pdf for a scalar parameter θ:

$$p(\theta|\mathbf{y}) \propto p(\theta)l(\theta|\mathbf{y})$$

$$\propto p(\theta)e^{\log l(\theta|\mathbf{y})}, \tag{2.34}$$

[26] For other discussions of this topic see Jeffreys, *op. cit.*, p. 193 ff.; Lindley, *op. cit.*, p. 128 ff., and "The Use of Prior Probability Distributions in Statistical Inference and Decisions," in J. Neyman (Ed.) *Proc. Fourth Berkeley Symp. Math. Statist. Probab.* Berkeley: University of California Press, **1**, 453–468 (1961); L. LeCam, "On Some

where $p(\theta)$ is our prior pdf and $l(\theta|\mathbf{y})$ denotes the likelihood function based on n independent sample observations, $\mathbf{y}' = (y_1, y_2, \ldots, y_n)$. We assume that both $p(\theta)$ and $l(\theta|\mathbf{y})$ are nonzero in the parameter space and have continuous derivatives and that $l(\theta|\mathbf{y})$ has a unique maximum at $\theta = \hat{\theta}$, the maximum likelihood estimate.

In general, as Jeffreys points out, $\log l(\theta|\mathbf{y})$ will be of order n, whereas $p(\theta)$ does not depend on n, the sample size. Thus, heuristically, in large-sized samples the likelihood factor in (2.34) will dominate the posterior pdf. Since under general conditions the likelihood function assumes a normal shape as n gets large, with center at the maximum likelihood estimate $\hat{\theta}$, the posterior pdf will be normal in large samples with mean equal to the maximum likelihood estimate $\hat{\theta}$.

To put these considerations in more explicit terms, we can expand both factors of (2.34) around the maximum likelihood estimate $\hat{\theta}$ as follows:

$$\begin{aligned} p(\theta) &= p(\hat{\theta}) + (\theta - \hat{\theta})p'(\hat{\theta}) + \tfrac{1}{2}(\theta - \hat{\theta})^2 p''(\hat{\theta}) + \cdots \\ &= p(\hat{\theta})\left[1 + \frac{(\theta - \hat{\theta})p'(\hat{\theta})}{p(\hat{\theta})} + \frac{\frac{1}{2}(\theta - \hat{\theta})^2 p''(\hat{\theta})}{p(\hat{\theta})} + \cdots\right] \end{aligned} \tag{2.35}$$

and, with $g(\theta) = \log l(\theta|\mathbf{y})$,

$$\begin{aligned} \exp\{g(\theta)\} &= \exp\{g(\hat{\theta}) + \tfrac{1}{2}(\theta - \hat{\theta})^2 g''(\hat{\theta}) + \tfrac{1}{6}(\theta - \hat{\theta})^3 g'''(\hat{\theta}) + \cdots\} \\ &\propto \exp\{\tfrac{1}{2}(\theta - \hat{\theta})^2 g''(\hat{\theta})\}[\,1 + \tfrac{1}{6}(\theta - \hat{\theta})^3 g'''(\hat{\theta}) + \cdots], \end{aligned} \tag{2.36}$$

where the fact that $g'(\hat{\theta}) = 0$ (since $\hat{\theta}$ is the maximum likelihood estimate) has been employed and where the expansion $e^x = 1 + x + \cdots$ has been utilized. Then, on multiplying (2.35) and (2.36), we have

$$p(\theta|\mathbf{y}) \propto e^{\frac{1}{2}(\theta-\hat{\theta})^2 g''(\hat{\theta})}\left[1 + \frac{(\theta - \hat{\theta})p'(\hat{\theta})}{p(\hat{\theta})} + \frac{\frac{1}{2}(\theta - \hat{\theta})^2 p''(\hat{\theta})}{p(\hat{\theta})} + \tfrac{1}{6}(\theta - \hat{\theta})^3 g'''(\hat{\theta}) + \cdots\right]. \tag{2.37}$$

The leading term in (2.37), $e^{\frac{1}{2}(\theta-\hat{\theta})^2 g''(\hat{\theta})}$, is in the normal form, centered at the maximum likelihood estimate $\hat{\theta}$ with variance[27]

$$\mathrm{Var}(\theta|\mathbf{y}) \doteq [-g''(\hat{\theta})]^{-1} = \left[-\frac{d^2 \log l(\theta|\mathbf{y})}{d\theta^2}\right]^{-1}_{\theta=\hat{\theta}}.$$

Thus, if we use just the leading term of (2.37), the approximate large sample posterior pdf for θ is

Asymptotic Properties of Maximum Likelihood and Related Bayes Estimates," *Univ. Calif. Publ. Statist.*, **1**, 277–330 (1953); and "Les Propriétés Asymptotiques des Solutions de Bayes," *Publ. Inst. Statist.*, University of Paris, Vol. 7, 1958, pp. 17–35; R. A. Johnson, "An Asymptotic Expansion for Posterior Distributions," *Ann. Math. Statist.* **38**, 1899–1906 (1967).

[27] Note that, since $g(\theta)$ has a maximum at $\theta = \hat{\theta}$, $g''(\hat{\theta}) < 0$.

$$p(\theta|\mathbf{y}) \doteq \frac{|g''(\hat{\theta})|^{\frac{1}{2}}}{\sqrt{2\pi}} e^{-\frac{1}{2}(\theta-\hat{\theta})^2|g''(\hat{\theta})|}. \tag{2.38}$$

Since $|g''(\hat{\theta})|$ is usually of order n, as n gets large the posterior pdf becomes sharply centered around $\hat{\theta}$, that is, $|g''(\hat{\theta})|^{-1}$, the variance, becomes smaller as n grows larger.

With respect to the quality of the approximation in (2.38), Jeffreys points out that $\theta - \hat{\theta}$ is of order $n^{-\frac{1}{2}}$, and thus in (2.37) the terms $(\theta - \hat{\theta})p'(\hat{\theta})/p(\hat{\theta})$ and $\frac{1}{6}(\theta - \hat{\theta})^3 g'''(\hat{\theta})$ are of order $n^{-\frac{1}{2}}$,[28] whereas $\frac{1}{2}(\theta - \hat{\theta})^2 p''(\hat{\theta})/p(\hat{\theta})$ is of order n^{-1}. Thus the approximation in (2.38)[29] involves an error of order $n^{-\frac{1}{2}}$.

Example 2.10. Assume that we have n independent observations from a normal population with unknown mean μ and known standard deviation $\sigma = \sigma_0$. It is well known that the sample mean $\hat{\mu} = \sum_{i=1}^n y_i/n$ is the maximum likelihood estimate for μ. Then, employing (2.38) for *any* prior pdf satisfying the assumptions set forth above, the posterior pdf, $p(\mu|\mathbf{y}, \sigma^2)$, can be approximated as follows in large samples:

$$\begin{aligned} p(\mu|\mathbf{y}, \sigma^2) &\doteq \frac{|g''(\hat{\mu})|^{\frac{1}{2}}}{\sqrt{2\pi}} e^{-\frac{1}{2}(\mu-\hat{\mu})^2|g''(\hat{\mu})|} \\ &\doteq \frac{\sqrt{n}}{\sqrt{2\pi}\,\sigma_0} e^{-(n/2\sigma_0^2)(\mu-\hat{\mu})^2} \end{aligned}$$

where

$$g(\mu) = \log l(\mu|\mathbf{y}, \sigma_0) = -\log\sqrt{2\pi} - n\log\sigma_0 - \frac{1}{2\sigma_0^2}\sum_{i=1}^n (y_i - \mu)^2$$

and

$$g''(\hat{\mu}) = \frac{-n}{\sigma_0^2}.$$

Thus the large sample posterior pdf for μ is a normal pdf with mean $\hat{\mu}$ and variance $|g''(\hat{\mu})|^{-1} = \sigma_0^2/n$.

The above argument generalizes easily to the case in which we have a vector of parameters, say $\boldsymbol{\theta}$, rather than a scalar parameter; that is, in large samples the posterior pdf for $\boldsymbol{\theta}$ will be approximately normal with mean $\hat{\boldsymbol{\theta}}$, the maximum likelihood estimate, and covariance matrix

$$\left[-\frac{\partial^2 \log l(\boldsymbol{\theta}|\mathbf{y})}{\partial\theta_i\,\partial\theta_j}\right]^{-1}_{\boldsymbol{\theta}=\hat{\boldsymbol{\theta}}}. \tag{2.39}$$

[28] $g'''(\hat{\theta})$ is usually of order n if it is nonzero.

[29] It is possible to improve the approximation in (2.38) by retaining additional terms appearing in the square brackets in (2.37). See, for example, Lindley, "The Use of Prior Probability Distributions in Statistical Inference and Decisions," *op. cit.*, p. 457 ff.

In this case, proceeding as above, we can expand the two factors in the posterior pdf for $\boldsymbol{\theta}$, $p(\boldsymbol{\theta}|y) \propto p(\boldsymbol{\theta})\, l(\boldsymbol{\theta}|\mathbf{y}) = p(\boldsymbol{\theta})e^{g(\boldsymbol{\theta}|y)}$, where $g(\boldsymbol{\theta}|\mathbf{y}) = \log l(\boldsymbol{\theta}|\mathbf{y})$. Then, if we retain just the leading term in the expansion, we have, with $\dot{\propto}$ denoting "approximately proportional to,"

$$p(\boldsymbol{\theta}|\mathbf{y}) \dot{\propto} \exp\left[-\tfrac{1}{2}(\boldsymbol{\theta} - \hat{\boldsymbol{\theta}})'C(\boldsymbol{\theta} - \hat{\boldsymbol{\theta}})\right], \tag{2.40}$$

which is in the multivariate normal form with mean $\hat{\boldsymbol{\theta}}$, the maximum likelihood estimate, and covariance matrix C^{-1}, which is just the matrix in (2.39).[30]

It is indeed interesting to observe the close agreement of Bayesian results in large samples with those flowing from the maximum likelihood approach. Of course, a moot problem is how large a sample size is required for these large sample approximate results to be reasonably accurate. Fortunately there is usually no need to rely on large-sample approximate results, since finite sample posterior pdf's are available, given the elements appearing in Bayes' theorem. In certain instances, however, in which computational problems arise in the analysis of complicated posterior pdf's the above large-sample results are useful.

2.14 REPORTING THE RESULTS OF BAYESIAN ANALYSES

In reporting the results of Bayesian analyses involving estimation of parameters in scientific journals, it is important to provide at least (a) a detailed discussion of the stochastic model assumed to generate the observations, (b) a full discussion of prior assumptions about parameter values, (c) the sample information, and (d) information about posterior pdf's for parameters of interest.

With respect to the stochastic model for the observations, subject matter considerations should be reviewed to justify its form and stochastic assumptions. Given that this has been done satisfactorily, the likelihood function $p(\mathbf{y}|\boldsymbol{\theta})$ should be shown explicitly, where $\mathbf{y}$ is an observation vector and $\boldsymbol{\theta}$ is a parameter vector.

[42] See, for example, Jeffreys, *Theory of Probability*, *op. cit.*, pp. 123–125.

[43] Lindley, *op. cit.*, p. 145.

[44] That is, from $v = \log \eta = \log \theta/(1 - \theta)$, $dv = d \log \theta - d \log (1 - \theta) = d\theta/\theta(1 - \theta)$ and thus $p(v)\, dv \propto dv$ implies $p(\theta)\, d\theta \propto [\theta(1 - \theta)]^{-1}\, d\theta$.

[45] If, instead of (2.54), we had used the Bayes-Laplace uniform prior, $p(\theta) \propto$ constant, $0 \leq \theta \leq 1$, the exponents in (2.55) would each be changed by one, the equivalent of two sample observations, which will not be important in moderate-sized samples.

As regards prior assumptions about $\boldsymbol{\theta}$, that is, a choice of prior pdf for $\boldsymbol{\theta}$, all information used to make such a choice should be explicitly stated. If data-based prior information is being employed, this fact should be noted and references provided to the sources of such prior information. If nondata-based prior information is employed, it should be carefully examined and explicated. In this way the reader will understand what information is being added to the sample information in performing an analysis. Of course, if little prior information is available or if the investigator wishes to show what results from an analysis assuming little prior information, he will use a vague or diffuse prior pdf.

With respect to reporting the data employed in an analysis, it is good procedure to describe in detail how they were obtained and have them available for any interested party by including them in the report or by making it known that they can be obtained on request. By having the data available other parties can perform analyses using whatever prior pdf's they choose to use. Also, should there be any controversy about the form of the likelihood function, the data can be employed to explore alternative formulations.[46]

With respect to reporting information about posterior pdf's for parameters of interest, it is good practice to report the complete posterior pdf and to provide summary characteristics, say measures of central tendency and dispersion. Also posterior intervals (or regions) often help readers to appreciate what the prior and sample information implies about the values of parameters.

By paying special attention to the above points, readers will understand how the reporting investigator learned from the information in his sample[47]; that is, he will have information regarding the investigator's initial beliefs about the parameters and model and can then see how they are altered by the data. This change in beliefs is indeed an essential part of the process of learning from experience.