

ness, the prior p.d.f. of W is smooth and very widely spread out over the line, then the statistician might find it convenient to assume a uniform, or constant, density over the whole line in order to represent this prior distribution. Such a uniform density is not the p.d.f. of any proper probability distribution on the real line. However, the statistician can often develop a posterior distribution which is a proper distribution by using this uniform density as a prior density and formally carrying out the calculations of Bayes' theorem with some observed values x_1, \dots, x_n . Thus, if $f_n(x_1, \dots, x_n|w)$ is the likelihood function for these observed values and if

$$0 < \int_{-\infty}^{\infty} f_n(x_1, \dots, x_n|w) dw < \infty, \quad (1)$$

then the posterior p.d.f. $\xi(\cdot | x_1, \dots, x_n)$ of W will be specified by the relation

$$\xi(w|x_1, \dots, x_n) \propto f(x_1, \dots, x_n|w). \quad (2)$$

As an illustration of the relation (2), suppose that X_1, \dots, X_n is a random sample from a normal distribution with an unknown value of the mean W and a specified value of the precision r . If the prior distribution of W is represented by a uniform density over the real line, then it follows from relation (4) of Sec. 9.5 that

$$\xi(w|x_1, \dots, x_n) \propto \exp \left[-\frac{nr}{2} (w - \bar{x})^2 \right]. \quad (3)$$

Therefore, the posterior distribution of W when $X_i = x_i$ ($i = 1, \dots, n$) is a normal distribution with mean \bar{x} and precision nr . Although the prior distribution is improper, the posterior distribution will be a proper normal distribution after just one observation has been made.

Another procedure when the prior knowledge is vague is to use a proper prior distribution of W from some appropriate conjugate family which is indexed by some parameter α , to compute the posterior distribution from the observed values x_1, \dots, x_n , and then to study the limiting posterior distribution as the parameter α approaches some limiting value. Often, this limiting posterior distribution will be a proper probability distribution for W , even though it did not result from any proper prior distribution.

As an illustration of the use of a limiting posterior distribution, consider again the problem of sampling from a normal distribution with an unknown value of the mean W and a specified value of the precision r . A conjugate family of distributions for this problem is described in Theorem 1 of Sec. 9.5. If we let $\tau \rightarrow 0$ in the posterior normal distribution of W , as described in that theorem, it can be seen that the limiting result will be

CHAPTER 10

Limiting posterior distributions

10.1 IMPROPER PRIOR DISTRIBUTIONS

In certain problems, the prior knowledge that a statistician has about some parameter W may be very slight and vague when compared with the information about W which he expects to acquire from available observations. In such a problem, even though the statistician may be able to find a suitable conjugate family of distributions of W , it may not be easy for him to select the appropriate prior distribution from this family. Because of the vagueness of his prior knowledge in comparison with the information that he will soon have from his observations, it would not be worthwhile for the statistician to expend a great deal of time or effort in determining a specific prior distribution. Instead, he would find it convenient to make use of a standard prior distribution that would be suitable in many situations in which it is desired to represent vague prior information.

Often, the standard prior distribution which is used in one of these problems is an improper distribution in the sense that it is represented by a nonnegative density function whose integral over the whole parameter space Ω is infinite; for any proper probability distribution, this integral must be unity. For example, if Ω is the real line and, because of vague-

a normal distribution with mean \bar{x} and precision $n\tau$. This distribution serves as a posterior distribution for W , but it is one that cannot be derived from any proper prior distribution. In this example, the result is the same as the posterior distribution (3), which was obtained by using an improper uniform prior density of W . This agreement could have been anticipated by noting that when the precision $\tau \rightarrow 0$ in the prior normal distribution of W , the variance becomes arbitrarily large and hence the distribution becomes spread more and more thinly over the line.

The posterior distribution specified by relation (3) has interesting properties. Let the random variable Z be defined by the equation

$$Z = (n\tau)^{1/2}(\bar{X} - W). \quad (4)$$

Then it follows from this posterior distribution of W that the conditional distribution of Z for any given value \bar{x} of \bar{X} is a standard normal distribution (i.e., a normal distribution with mean 0 and precision 1). Since this conditional distribution of Z is the same for any given value of \bar{X} , we can conclude that the random variables Z and \bar{X} are independent. Furthermore, it follows from the standard sampling theory for normal distributions, as described in Sec. 4.7, that the conditional distribution of Z for any given value w of W is also a standard normal distribution. Therefore, the random variables Z and W are also independent.

Strictly speaking, the two conclusions just stated are mutually incompatible. Under any proper bivariate distribution of \bar{X} and W for which the random variables Z and W are independent, it would not be possible for the random variables Z and \bar{X} also to be independent, unless the random variable Z is equal to a constant with probability 1 (see Exercise 1). The random variable Z is not constant here because of our use of an improper prior distribution. These considerations indicate why the statistician must take particular care when handling and interpreting improper distributions.

Confidence Intervals

We have noted that the conditional distribution of Z for any given value of \bar{X} is the same as the conditional distribution of Z for any given value of W . Because of this property, inferences about W that are based on its posterior distribution will, in general, agree with inferences about W that are based on the standard method of confidence intervals. This method may be described as follows:

Let $u_1(X_1, \dots, X_n)$ and $u_2(X_1, \dots, X_n)$ be random variables such that for any given value $w \in \Omega$, the conditional probability satisfies

the relation

$$\Pr\{u_1(X_1, \dots, X_n) \leq w \leq u_2(X_1, \dots, X_n) | W = w\} \geq \gamma, \quad (5)$$

where γ is a fixed number ($0 < \gamma < 1$). Then for any observed values x_1, \dots, x_n , the following interval can be specified for W :

$$u_1(x_1, \dots, x_n) \leq W \leq u_2(x_1, \dots, x_n). \quad (6)$$

This interval is called a *confidence interval* for W , and the number γ is called the *confidence coefficient* of the interval.

For any value of γ ($0 < \gamma < 1$), let k_γ denote the number such that

$$\Phi(k_\gamma) - \Phi(-k_\gamma) = \gamma, \quad (7)$$

where Φ is the standard normal d.f. It follows from the conditional distribution of Z specified earlier that a confidence interval for W with confidence coefficient γ can be defined by the relation

$$\bar{x} - k_\gamma(n\tau)^{-1/2} \leq W \leq \bar{x} + k_\gamma(n\tau)^{-1/2}. \quad (8)$$

But it also follows that under the posterior distribution of W , the probability that W lies in the interval in (8) is γ . Thus, according to the standard theory, that interval has confidence γ , and according to the Bayesian or subjective theory, that interval has probability γ . The operational interpretations of these two properties seem to be essentially the same. Therefore, as previously stated, inferences about W based on the two points of view will generally agree.

Further Remarks and References

Improper prior distributions play an important part in the statistical methods discussed by Jeffreys (1961) and Lindley (1965). In the next two sections of this chapter we shall consider posterior distributions which either result from improper prior distributions or are limiting posterior distributions in problems which involve samples from univariate and multivariate normal distributions. Problems which involve samples from other distributions are presented as exercises at the end of the chapter.

We shall show that in many problems in which the posterior distribution of a parameter is obtained in this way, confidence intervals—or, in higher dimensions, confidence sets—for the parameter which have a given confidence coefficient γ will agree with intervals or sets which have posterior probability γ .

Stein (1962a, 1965) has discussed the difficulty of representing vague prior knowledge about a vector \mathbf{W} of high dimension and other aspects of the use of improper distributions.

10.2 IMPROPER PRIOR DISTRIBUTIONS FOR SAMPLES FROM A NORMAL DISTRIBUTION

Consider the problem treated in Theorem 2 of Sec. 9.5. In this problem, a random sample X_1, \dots, X_n is to be taken from a normal distribution with a known value of the mean m and an unknown value of the precision W . In the posterior gamma distribution of W as given in that theorem, let $\alpha \rightarrow 0$ and $\beta \rightarrow 0$. Then the limiting posterior distribution of W is a gamma distribution for which the parameters are $n/2$ and $\frac{1}{2} \sum_{i=1}^n (x_i - m)^2$.

The same posterior distribution can be obtained from an improper prior distribution. Since the prior distribution of W is a gamma distribution with parameters α and β , then for $w > 0$ the prior p.d.f. ξ of W satisfies the relation

$$\xi(w) \propto w^{\alpha-1} e^{-\beta w}. \tag{1}$$

When we let $\alpha \rightarrow 0$ and $\beta \rightarrow 0$ in the relation (1), we obtain formally (but only formally, since in that relation we are ignoring the proportionality factor which involves α and β) the following result:

$$\xi(w) \propto \frac{1}{w}. \tag{2}$$

The right side of the relation (2) represents an improper prior density since its integral over Ω is not finite. However, it can be shown that if the improper prior density of W is specified by the relation (2), then the posterior distribution of W will be the proper gamma distribution already found.

Let the random variable Z be defined by the following equation:

$$Z = W \left[\sum_{i=1}^n (X_i - m)^2 \right]. \tag{3}$$

It follows from the preceding development that the conditional distribution of Z when $X_i = x_i$ ($i = 1, \dots, n$) is a χ^2 distribution with n degrees of freedom. Moreover, it is well known and can be easily verified that the conditional distribution of Z when $W = w$ is also a χ^2 distribution with n degrees of freedom. Hence, a confidence interval for W with confidence coefficient γ ($0 < \gamma < 1$) which is found from the conditional distribution of Z will also be an interval for W whose posterior probability is γ .

As another example involving sampling from a univariate normal distribution, consider the problem treated in Theorem 1 of Sec. 9.6. In this problem, a random sample X_1, \dots, X_n is to be taken from a distribution for which both the mean M and the precision R are unknown.

In the posterior joint distribution of M and R as given in Theorem 1 of Sec. 9.6, let $\tau \rightarrow 0$, let $\alpha \rightarrow -\frac{1}{2}$, and let $\beta \rightarrow 0$. Then, under the limiting posterior distribution, the conditional distribution of M when $R = r$ is a normal distribution with mean \bar{x} and precision nr , and the marginal distribution of R is a gamma distribution with parameters $(n - 1)/2$ and $\frac{1}{2} \sum_{i=1}^n (x_i - \bar{x})^2$. It is assumed that $n \geq 2$. Note that in order to obtain this posterior distribution, the parameter α must violate the condition that $\alpha > 0$, which was assumed in Theorem 1 of Sec. 9.6, and α must now approach the negative number $-\frac{1}{2}$. It also follows from the presentation in Sec. 9.6 that this same posterior joint distribution of M and R can be obtained from an improper prior joint density ξ which is specified over the parameter space Ω by the equation

$$\xi(m, r) = \frac{1}{r}. \tag{4}$$

The joint density specified by Eq. (4) is simply the product of a uniform density in m over the real line and the density $1/r$ over the values $r > 0$. In other words, this joint density is the product of the improper densities which were used when only the value of M or only the value of R was unknown.

Let the random variable T be defined by the equation

$$T = \frac{n^2(M - \bar{X})}{[\sum_{i=1}^n (X_i - \bar{X})^2 / (n - 1)]}. \tag{5}$$

We know from the discussion in Sec. 9.6 and from the limiting posterior joint distribution of M and R found above that the conditional distribution of T when $X_i = x_i$ ($i = 1, \dots, n$) is a t distribution with $n - 1$ degrees of freedom. But we also know from expression (3) of Sec. 4.12 that the conditional distribution of T when $M = m$ and $R = r$ is a t distribution with $n - 1$ degrees of freedom.

Furthermore, let the random variable U be defined by the equation

$$U = R \left[\sum_{i=1}^n (X_i - \bar{X})^2 \right]. \tag{6}$$

Then it follows from the limiting posterior distribution of R that the conditional distribution of U when $X_i = x_i$ ($i = 1, \dots, n$) is a χ^2 distribution with $n - 1$ degrees of freedom. As was pointed out in Sec. 4.8, the conditional distribution of U when $M = m$ and $R = r$ is also a χ^2 distribution with $n - 1$ degrees of freedom.

It follows from these remarks that confidence coefficients for the standard confidence intervals for M as found from the conditional distribution of T will agree with the probabilities of the same intervals computed under the limiting posterior distribution. A similar statement can

be made for confidence intervals for R as found from the conditional distribution of U . Stone (1963) has described other models which lead to conclusions like these.

Invariant Prior Distributions

Jeffreys (1961) has discussed the use of the prior density specified by Eq. (4), and he has presented arguments for the appropriateness of this density as a representation of vague prior knowledge in scientific work. His arguments are based partly on the following *invariance* properties of this density:

If the statistician has vague prior knowledge about the value of M , then he also has vague prior knowledge about the linear transform $M^* = aM + b$, where a and b are specified constants ($a \neq 0$). Therefore, if it is appropriate to represent the prior knowledge about M by a uniform density over the whole real line, it should also be appropriate to represent the prior knowledge about M^* by a uniform distribution over the whole real line. However, it follows from the theory given in Sec. 3.7 for transformations of random variables that if M has a uniform density, then so also does M^* . Therefore, the uniform density satisfies the desired invariance property.

The density of the precision R specified by Eq. (4) has a similar invariance property. If the statistician has vague prior knowledge about R , then he also has vague prior knowledge about the transform $R^* = R^\alpha$, where α is a specified constant ($\alpha \neq 0$). It can be shown (see Exercise 6a) that if the density of R is proportional to $1/r$ for $r > 0$, then the density of R^* will be proportional to $1/r^*$ for $r^* > 0$. It follows that regardless of whether R^* is the unknown precision, the unknown variance, or the unknown standard deviation, it is appropriate to represent vague prior knowledge about R^* by a density of the form $1/r^*$.

Finally, it should be noted (see Exercise 6b) that if R has a density of the form specified above for $r > 0$, then $S = \log R$ will have a uniform density over the whole real line. Therefore, the joint density of M and R described by Eq. (4) specifies that M and R are independent and both M and $\log R$ have uniform densities over the whole real line.

Invariant distributions are also discussed by Hartigan (1964).

10.3 IMPROPER PRIOR DISTRIBUTIONS FOR SAMPLES FROM A MULTIVARIATE NORMAL DISTRIBUTION

Consider the problem treated in Theorem 1 of Sec. 9.9. In this problem, a random sample $\mathbf{X}_1, \dots, \mathbf{X}_n$ is to be taken from a k -dimensional multivariate normal distribution with an unknown value of the mean vector \mathbf{M}

and a specified precision matrix \mathbf{r} . Just as for the univariate distribution, we shall let $\boldsymbol{\tau} \rightarrow \mathbf{0}$ in the posterior distribution of \mathbf{M} , where $\mathbf{0}$ is the matrix each of whose elements is 0; or, equivalently, we shall take the prior density of \mathbf{M} to be a uniform density over the whole space R^k . Then it can be seen that the posterior distribution of \mathbf{M} becomes a multivariate normal distribution with mean vector $\bar{\mathbf{x}}$ and precision matrix $n\mathbf{r}$.

Let the random vector \mathbf{Z} be defined by the equation $\mathbf{Z} = \bar{\mathbf{X}} - \mathbf{M}$. Then both the conditional distribution of \mathbf{Z} when $\mathbf{X}_i = \mathbf{x}_i$ ($i = 1, \dots, n$) and the conditional distribution of \mathbf{Z} when $\mathbf{M} = \mathbf{m}$ will be multivariate normal distributions for which the mean vector is $\mathbf{0}$ and the precision matrix is $n\mathbf{r}$. It follows that if a confidence set for \mathbf{M} constructed from the random vector \mathbf{Z} has confidence coefficient γ , then this set will also have probability γ under the posterior distribution of \mathbf{M} .

Now consider the problem treated in Theorem 1 of Sec. 9.10. In this problem, a random sample $\mathbf{X}_1, \dots, \mathbf{X}_n$ is to be taken from a k -dimensional multivariate normal distribution for which both the mean vector \mathbf{M} and the precision matrix \mathbf{R} are unknown. It is assumed that $n \geq k + 1$. In the posterior joint multivariate normal-Wishart distribution of \mathbf{M} and \mathbf{R} , as given in that theorem, let $\nu \rightarrow 0$, let $\alpha \rightarrow -1$, and let $\boldsymbol{\tau} \rightarrow \mathbf{0}$. The result is a joint distribution such that the conditional distribution of \mathbf{M} when $\mathbf{R} = \mathbf{r}$ is a multivariate normal distribution with mean vector $\bar{\mathbf{x}}$ and precision matrix $n\mathbf{r}$, and the marginal distribution of \mathbf{R} is a Wishart distribution with $n - 1$ degrees of freedom and precision matrix \mathbf{s} , where \mathbf{s} is defined by Eq. (1) of Sec. 9.10. Note that in order to obtain this posterior distribution, the parameter α must violate the condition that $\alpha > k - 1$, which was assumed in Theorem 1 of Sec. 9.10, and α must now approach the value -1 .

Furthermore, it can be seen from the relations (5) and (8) of Sec. 9.10 that this same posterior joint distribution will be obtained if we assume that the improper prior joint density ξ of \mathbf{M} and \mathbf{R} is as follows:

$$\xi(\mathbf{m}, \mathbf{r}) = \frac{1}{|\mathbf{r}|^{(k+1)/2}} \quad (1)$$

The density specified by Eq. (1) can be interpreted as the product of a uniform density in \mathbf{m} over the space R^k and a density in \mathbf{r} which is proportional to the function on the right side of Eq. (1).

It now follows from the results given in Sec. 9.11 that under the posterior distribution just developed, the marginal distribution of \mathbf{M} will be a multivariate t distribution with $n - k$ degrees of freedom, location vector $\bar{\mathbf{x}}$, and precision matrix $n(n - k)\mathbf{s}^{-1}$. Let the random matrix \mathbf{S} be defined by the equation

$$\mathbf{S} = \sum_{i=1}^n (\mathbf{X}_i - \bar{\mathbf{X}})(\mathbf{X}_i - \bar{\mathbf{X}})', \quad (2)$$

and let the random variable Q be defined by the equation

$$Q = \frac{n(n-k)}{k} (\bar{\mathbf{X}} - \mathbf{M})' \mathbf{S}^{-1} (\bar{\mathbf{X}} - \mathbf{M}). \tag{3}$$

Then [see expression (16) of Sec. 5.6] the conditional distribution of Q when $\mathbf{X}_i = \mathbf{x}_i$ ($i = 1, \dots, n$) will be an F distribution with k and $n - k$ degrees of freedom.

It is said that a random variable Y has *Hotelling's T^2 distribution* with k and q degrees of freedom ($q \geq k$) if the random variable $[(q - k + 1)/(kq)]Y$ has an F distribution with k and $q - k + 1$ degrees of freedom. If the random variable Q^* is defined by the equation

$$Q^* = n(n-1)(\bar{\mathbf{X}} - \mathbf{M})' \mathbf{S}^{-1} (\bar{\mathbf{X}} - \mathbf{M}), \tag{4}$$

then it follows from Eq. (3) that the conditional distribution of Q^* when $\mathbf{X}_i = \mathbf{x}_i$ ($i = 1, \dots, n$) will be Hotelling's T^2 distribution with k and $n - 1$ degrees of freedom. But it is known [see, e.g., Scheffé (1959), app. 5, or Anderson (1958), chap. 5] that the conditional distribution of Q^* when the values of \mathbf{M} and \mathbf{R} are given is also Hotelling's T^2 distribution with k and $n - 1$ degrees of freedom.

An ellipsoid \mathcal{E} can now be constructed in the space R^k such that \mathbf{M} will lie in this ellipsoid with a specified confidence coefficient γ or, equivalently, with a specified probability γ under the posterior distribution of \mathbf{M} . Let γ be a given number ($0 < \gamma < 1$), and let ϕ_γ be a constant such that if the random variable Q has an F distribution with k and $n - k$ degrees of freedom, then $\Pr(Q \leq \phi_\gamma) = \gamma$. Also, let $\bar{\mathbf{x}}$ and \mathbf{s} be the observed values of the sample mean vector $\bar{\mathbf{X}}$ and the sample matrix \mathbf{S} . Finally, let \mathcal{E} be the set of points \mathbf{m} ($\mathbf{m} \in R^k$) which satisfy the following relation:

$$(\mathbf{m} - \bar{\mathbf{x}})' \mathbf{s}^{-1} (\mathbf{m} - \bar{\mathbf{x}}) \leq \frac{k\phi_\gamma}{n(n-k)}. \tag{5}$$

Since the matrix \mathbf{s}^{-1} is positive definite, the set \mathcal{E} will be an ellipsoid in R^k . It follows from Eqs. (3) and (4) that \mathcal{E} will be a confidence ellipsoid for \mathbf{M} with confidence coefficient γ and also that the posterior probability that \mathbf{M} will lie in \mathcal{E} is γ .

10.4 PRECISE MEASUREMENT

In this section we shall consider further the consequences of using either a proper or an improper uniform prior density for a parameter W when observations which are expected to be very informative are to be made. We shall show that under quite general conditions the posterior distribution of W derived from a uniform prior density over the parameter space

Ω will be a close approximation to the posterior distribution derived from a more carefully specified proper prior distribution. The fact that posterior distributions derived from uniform prior densities are adequate approximations to the actual posterior distributions is a consequence of what Savage calls the *principle of stable estimation* [Edwards, Lindman, and Savage (1963)] or *precise measurement* [Savage (1961), Savage and others (1962)]. Each of these three references contains excellent discussions which pertain not only to the theory of precise measurement but also to the problems of Bayesian statistical inference in general.

One interesting feature of the theory of precise measurement is that specific quantitative results can be obtained which indicate how closely the approximate posterior distribution agrees with the actual posterior distribution. The theorem which will be presented here in this regard is due to Edwards, Lindman, and Savage (1963). This theorem will be valid regardless of whether W is a real-valued parameter or a vector and also regardless of whether W has a discrete prior p.f. on a countable set of values or a prior p.d.f. over R^k ($k \geq 1$). The presentation will be general enough to include all these possibilities.

Let W be a parameter whose possible values are in the parameter space Ω . Also, let X be either a random variable or a random vector whose conditional g.p.d.f. for any given value $W = w$ ($w \in \Omega$) is $f(\cdot | w)$. Let ξ denote the prior g.p.d.f. of W , and, as usual, let $\xi(\cdot | x)$ denote the posterior g.p.d.f. of W resulting from the observed value $X = x$. We shall assume that the prior g.p.d.f. ξ is a bounded function on the set Ω .

For the observed value $X = x$, we shall suppose that

$$0 < \int_{\Omega} f(x|w) d\nu(w) < \infty. \tag{1}$$

We can then define the function $\phi(\cdot | x)$ on Ω as follows:

$$\phi(w|x) = \frac{f(x|w)}{\int_{\Omega} f(x|w') d\nu(w')}. \tag{2}$$

The function $\phi(\cdot | x)$ is the posterior g.p.d.f. of W which would result from a uniform prior g.p.d.f., regardless of whether this prior g.p.d.f. is proper or improper. The next theorem provides bounds on the difference between the actual posterior g.p.d.f. $\xi(\cdot | x)$ and the g.p.d.f. $\phi(\cdot | x)$.

Theorem 1 *Let A be any subset of Ω such that*

$$m = \inf_{w \in A} \xi(w) > 0, \tag{3}$$

and let α, β , and γ be three numbers ($0 \leq \alpha < 1, \beta \geq 0$, and $\gamma \geq 0$) which satisfy the following three relations:

$$\int_A \phi(w|x) d\nu(w) \geq 1 - \alpha, \quad (4)$$

$$\sup_{w \in A} \xi(w) \leq (1 + \beta)m, \quad (5)$$

and

$$\sup_{w \in A^c} \xi(w) \leq (1 + \gamma)m. \quad (6)$$

Furthermore, let the number ϵ be defined by the equation

$$\epsilon = \max \left\{ \frac{\alpha + \beta}{1 - \alpha}, \frac{\alpha + \beta + \alpha\gamma}{1 + \alpha + \beta + \alpha\gamma} \right\} + \frac{\alpha(2 - \alpha + \gamma)}{1 - \alpha}. \quad (7)$$

Then

$$\int_{\Omega} |\xi(w|x) - \phi(w|x)| d\nu(w) \leq \epsilon. \quad (8)$$

Proof If $f(x|w) \neq 0$, then

$$\frac{\xi(w|x)}{\phi(w|x)} = \frac{\xi(w)}{\int_{\Omega} \xi(w') \phi(w'|x) d\nu(w')}. \quad (9)$$

If $f(x|w) = 0$, then the value of the ratio on the left side of Eq. (9) is indeterminate. We shall assign to it the value of the right side of Eq. (9) in order that this equation will be valid for every value of $w \in \Omega$.

From relations (3) to (6), we can obtain the following results:

$$\begin{aligned} \int_{\Omega} \xi(w) \phi(w|x) d\nu(w) & \\ & \leq (1 + \beta)m \int_A \phi(w|x) d\nu(w) + (1 + \gamma)m \int_{A^c} \phi(w|x) d\nu(w) \\ & \leq (1 + \beta)m + (1 + \gamma)m\alpha \end{aligned} \quad (10)$$

and

$$\int_{\Omega} \xi(w) \phi(w|x) d\nu(w) \geq m(1 - \alpha). \quad (11)$$

Furthermore, by combining the relations (9) to (11) with relations (3) to (6), we can obtain the following inequalities. For any value $w \in A$,

$$\frac{1}{1 + \alpha + \beta + \alpha\gamma} \leq \frac{\xi(w|x)}{\phi(w|x)} \leq \frac{1 + \beta}{1 - \alpha}. \quad (12)$$

Also, for any value $w \in A^c$,

$$\frac{\xi(w|x)}{\phi(w|x)} \leq \frac{1 + \gamma}{1 - \alpha}. \quad (13)$$

It now follows that

$$\begin{aligned} \int_{\Omega} |\xi(w|x) - \phi(w|x)| d\nu(w) &= \int_{\Omega} \left| \frac{\xi(w|x)}{\phi(w|x)} - 1 \right| \phi(w|x) d\nu(w) \\ &\leq \max \left\{ \frac{\alpha + \beta}{1 - \alpha}, \frac{\alpha + \beta + \alpha\gamma}{1 + \alpha + \beta + \alpha\gamma} \right\} \int_A \phi(w|x) d\nu(w) \\ &\quad + \left(\frac{1 + \gamma}{1 - \alpha} + 1 \right) \int_{A^c} \phi(w|x) d\nu(w) \leq \epsilon. \quad (14) \end{aligned}$$

One consequence of the relation (8) is that for any subset B of Ω , the difference between the probability of B computed from the g.p.d.f. $\phi(\cdot|x)$ and the probability of the same subset B computed from the actual posterior g.p.d.f. $\xi(\cdot|x)$ cannot be greater than ϵ . The value of ϵ which is obtained in Theorem 1 depends on the choice of the set A . An effective choice of the set A will be a set which yields a small value of ϵ , that is, a set A for which α, β , and γ will be small.

A great advantage of Theorem 1 is the simplicity of the three relations (4) to (6). For any set A , the number α specified in the relation (4) can be computed from the g.p.d.f.'s $f(\cdot|w)(w \in \Omega)$ and the observed value x . It is not necessary to consider the prior distribution. The numbers β and γ specified in the relations (5) and (6) can be computed by the statistician from two simple bounds on his prior g.p.d.f. In order that there may exist a set A for which α, β , and γ will be small, the observed value $X = x$ must convey enough information about W to permit the likelihood function $f(x|w)$ to be sharply peaked around some point in Ω . Under these conditions, the statistician knows, without any detailed analysis or specification of his prior g.p.d.f., that the g.p.d.f. $\phi(\cdot|x)$ is a good approximation to his posterior g.p.d.f.

For example, suppose that Ω is the real line and that the prior p.d.f. ξ of W is as sketched in Fig. 10.1. As indicated in that figure, the function ξ is constant on the interval from a to b . Suppose also that when the values x_1, \dots, x_n of a random sample are observed, the resulting p.d.f. $\phi(\cdot|x_1, \dots, x_n)$ is as sketched in Fig. 10.1. If the set A is chosen to be the interval from a to b , then the value of α will be equal to the total area of the two shaded regions in Fig. 10.1. Furthermore, it can be seen from the p.d.f. ξ that $\beta = 0$ and $\gamma = 0$. It now follows from Eq. (7) that $\epsilon \leq 3\alpha/(1 - \alpha)$. Therefore, the difference between the probability of any subset of Ω computed from the function $\phi(\cdot|x_1, \dots, x_n)$ sketched in Fig. 10.1 and the actual posterior probability of that subset cannot be greater than $3\alpha/(1 - \alpha)$.

10.5 CONVERGENCE OF POSTERIOR DISTRIBUTIONS

In the remainder of this chapter we shall consider the limiting properties of posterior distributions as the number of observations in a random

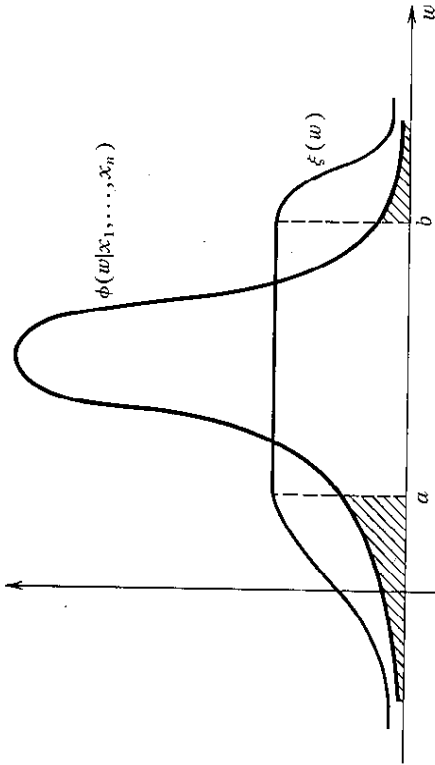


Fig. 10.1 An example of precise measurement.

sample becomes arbitrarily large. We shall show that if X_1, \dots, X_n is a random sample from a distribution for which the value of some parameter W is unknown and if the value of W is actually w_0 , then, under certain conditions, the posterior distribution of W will tend to become more and more concentrated around the value $W = w_0$ as we let $n \rightarrow \infty$.

A simple example is that in which the parameter W can have only k different values w_1, \dots, w_k . Suppose that $\Pr(W = w_i) = \xi_i > 0$ for $i = 1, \dots, k$, and suppose that for any given value $W = w_i$ ($i = 1, \dots, k$), the random variables X_1, \dots, X_n are a random sample from the g.p.d.f. f_i . We shall assume that the g.p.d.f.'s f_i are distinct in the following sense: If S denotes the sample space of any single observation, then

$$\int_S [f_i(x) - f_j(x)] d\mu(x) > 0 \quad \text{for } i \neq j. \tag{1}$$

For any observed values $X_j = x_j$ ($j = 1, \dots, n$), let $\xi_i(x_1, \dots, x_n)$ denote the posterior probability that $W = w_i$ ($i = 1, \dots, k$). Thus, this posterior probability is given by the equation

$$\xi_i(x_1, \dots, x_n) = \frac{\xi_i \prod_{j=1}^n f_i(x_j)}{\sum_{r=1}^k [\xi_r \prod_{j=1}^n f_r(x_j)]} \tag{2}$$

Suppose now that X_1, \dots, X_n is actually a random sample from the g.p.d.f. f , where t is one of the values $1, \dots, k$. We shall show that with probability 1, the following limiting values must be correct:

$$\lim_{n \rightarrow \infty} \xi_t(X_1, \dots, X_n) = 1 \tag{3}$$

and
$$\lim_{n \rightarrow \infty} \xi_i(X_1, \dots, X_n) = 0 \quad \text{for } i \neq t. \tag{4}$$

In other words, as $n \rightarrow \infty$, all of the posterior probability tends to be concentrated on the correct value w_t of W .

A basic result in problems involving large samples is the strong law of large numbers, which can be stated as follows:

Strong law of large numbers Let Y_1, Y_2, \dots be a sequence of independent and identically distributed random variables, for each of which the mean is μ . Then, with probability 1,

$$\lim_{n \rightarrow \infty} \frac{1}{n} \sum_{i=1}^n Y_i = \mu. \tag{5}$$

A proof of this result is given in such books as Feller (1957), chap. 10, and (1966), chap. 7; Loève (1965), sec. 16; and Rao (1965), chap. 2. This result will be used in establishing Eqs. (3) and (4).

For any fixed value of i such that $i \neq t$, let μ be defined by the equation

$$\mu = E \left[\log \frac{f_i(X)}{f_t(X)} \right]. \tag{6}$$

Here, X is any single observation and the expectation μ is computed under the assumption that $W = w_t$. The expectation μ is not necessarily finite. However, since f_t and f_i are distinct g.p.d.f.'s, it follows from Jensen's inequality that

$$\begin{aligned} \mu < \log E \left[\frac{f_i(X)}{f_t(X)} \right] &= \log \int_S \frac{f_i(x)}{f_t(x)} f_t(x) d\mu(x) \\ &= \log \int_S f_i(x) d\mu(x) = \log 1 = 0. \end{aligned} \tag{7}$$

Therefore, either μ is a finite negative number or μ can be assigned the value $-\infty$. In either case, it follows from the strong law of large numbers that with probability 1,

$$\lim_{n \rightarrow \infty} \frac{1}{n} \sum_{j=1}^n \log \frac{f_i(X_j)}{f_t(X_j)} = \mu < 0. \tag{8}$$

Hence,

$$\log \lim_{n \rightarrow \infty} \prod_{j=1}^n \frac{f_i(X_j)}{f_t(X_j)} = \lim_{n \rightarrow \infty} \sum_{j=1}^n \log \frac{f_i(X_j)}{f_t(X_j)} = -\infty. \tag{9}$$

However, Eq. (9) is equivalent to the following result: If $i \neq t$, then, with probability 1,

$$\lim_{n \rightarrow \infty} \prod_{j=1}^n \frac{f_i(X_j)}{f(X_j)} = 0. \quad (10)$$

Equations (3) and (4) now follow from Eq. (10) and from the expression given in Eq. (2) for the posterior probability $\xi_i(x_1, \dots, x_n)$.

A limiting result of this type is true for posterior distributions in many problems. Consider the problem treated in Theorem 1 of Sec. 9.5, in which a random sample X_1, \dots, X_n is taken from a normal distribution with an unknown value of the mean W and a specified value of the precision r . It was shown in that section that the mean μ' of the posterior distribution of W will be a weighted average of the sample mean \bar{X} and the mean of the prior distribution and that the precision of the posterior distribution of W will increase by r units for each observation which is taken.

Now suppose that X_1, \dots, X_n is, in fact, a random sample from a normal distribution with mean w_0 . If we let $n \rightarrow \infty$, the relative weight given to \bar{X} in the computation of μ' approaches the value 1. However, by the law of large numbers, $\bar{X} \rightarrow w_0$ with probability 1. Therefore, $\mu' \rightarrow w_0$ with probability 1. Furthermore, the variance of the posterior distribution of W approaches the value 0. It follows that the posterior distribution of W tends to become more and more concentrated around the value w_0 and that the posterior probability of any open interval containing w_0 must approach the value 1. Other examples are given in Exercises 19 to 21.

Further Remarks and References

The convergence of posterior distributions to the correct value of the parameter as $n \rightarrow \infty$ can be demonstrated for more general parameter spaces Ω . However, it becomes necessary to make special assumptions about the prior distribution of the parameter and the family of g.p.d.f.'s $\{f(\cdot | w), w \in \Omega\}$. This topic has a long history and was studied by Laplace. Some references on this topic and related questions are von Mises (1964), chap. 7; Berk (1966); and the more abstract work of Blackwell and Dubins (1962), Freedman (1963, 1965), Fabius (1964), and Schwartz (1965).

10.6 SUPERCONTINUITY

Now we shall consider a parameter W whose values must lie in an open interval Ω of the real line. The length of Ω may be either finite or infinite.

As usual, we assume that there is a given family of g.p.d.f.'s $\{f(\cdot | w), w \in \Omega\}$, but we shall now assume that the observations X_1, \dots, X_n are, in fact, a random sample from the g.p.d.f. $f(\cdot | w_0)$, where w_0 is a specific point in Ω . Under these conditions, we shall study the limiting behavior of the likelihood function computed from the observed values as $n \rightarrow \infty$. Specifically, we shall show that under certain regularity conditions, the likelihood function, and hence also the posterior p.d.f. of W , will converge in a special sense to the p.d.f. of a normal distribution.

For any point $w^* \in \Omega$ and any number $\alpha > 0$, we shall define $N(w^*; \alpha)$ as the interval around w^* containing every point in Ω whose distance from w^* is less than α .

Now let g be any real-valued function whose value $g(x, w)$ is specified at every point (x, w) of the product space $S \times \Omega$. Here, as usual, S denotes the sample space of a single observation X . It will be convenient in our development to use the following definition: The function g is *supercontinuous* at the value $w_0 \in \Omega$ if

$$\lim_{\alpha \rightarrow 0} E \left[\sup_{w \in N(w_0, \alpha)} |g(X, w) - g(X, w_0)| \right] = 0. \quad (1)$$

In Eq. (1) and in the remainder of this development, the expectation is computed with respect to the g.p.d.f. $f(\cdot | w_0)$ of each observation.

The meaning of Eq. (1) may be difficult to perceive, and so, before proceeding with the development of the limiting behavior of the likelihood function, we shall first discuss some properties of supercontinuous functions. For $x \in S$ and $\alpha > 0$, we shall let the value $h(x, \alpha)$ be defined as follows:

$$h(x, \alpha) = \sup_{w \in N(w_0, \alpha)} |g(x, w) - g(x, w_0)|. \quad (2)$$

With this definition, we can say that g is supercontinuous at w_0 if

$$\lim_{\alpha \rightarrow 0} E[h(X, \alpha)] = 0. \quad (3)$$

In the next two lemmas, we shall show that if the expectation $E[h(X, \alpha)]$ exists for all sufficiently small values of α , then supercontinuity of the function g at w_0 is equivalent to ordinary continuity of the function $g(x, \cdot)$ at w_0 for almost every value of $x \in S$. In regard to these lemmas, it should be kept in mind that when we say that a certain property is true for almost every value of $x \in S$, we mean that the property is true for each value of x with the possible exception of some values which form a subset of S whose probability is 0.

Lemma 1 *If g is supercontinuous at w_0 , then $g(x, \cdot)$ is continuous at w_0 for almost every value of $x \in S$.*