7

APPROXIMATE METHODS

The results obtained for normal distributions in the two previous chapters were exact for the specified prior distribution. In dealing with samples for distributions other than normal it is often necessary to resort to approximations to the posterior density even when the prior distribution is exact. In this chapter we shall mainly be concerned with binomial, multinomial and Poisson distributions, but begin by describing an approximate method of wide applicability.

7.1. The method of maximum likelihood

Let $\mathbf{x}$ be any observation with likelihood function $p(\mathbf{x}|\theta)$ depending on a single real parameter θ. The value of θ, denoted by $\hat{\theta}(\mathbf{x})$, or simply $\hat{\theta}$, for which the likelihood for that observation is a maximum is called the *maximum likelihood estimate* of θ. Notice that $\hat{\theta}$ is a function of the sample values only: it is an example of what we have previously called a statistic (§5.5). The definition generalizes to a likelihood depending on several parameters $p(\mathbf{x}|\theta_1, \theta_2, \ldots, \theta_s)$: the set $(\hat{\theta}_1, \hat{\theta}_2, \ldots, \hat{\theta}_s)$ for which the likelihood is a maximum form the set of maximum likelihood estimates, $\hat{\theta}_i$, of θ_i $(i = 1, 2, \ldots, s)$. The estimate is particularly important in the special case where $\mathbf{x} = (x_1, x_2, \ldots, x_n)$ is a random sample of size n from a distribution with density $f(x_i|\theta)$. Then

$$p(\mathbf{x}|\theta) = \prod_{i=1}^{n} f(x_i|\theta). \tag{1}$$

In this case the logarithm of the likelihood can be written as a sum:

$$L(\mathbf{x}|\theta) = \ln p(\mathbf{x}|\theta) = \sum_{i=1}^{n} \ln f(x_i|\theta). \tag{2}$$

Important properties of the log-likelihood, $L(\mathbf{x}|\theta)$, can be deduced from the strong law of large numbers (theorem 3.6.3) in the following way. In saying that $\mathbf{x}$ is a random sample we

imply that the sample values have the same density $f(x_i|\theta)$ for some θ fixed, for all i. Denote this value of θ by θ_0. We refer to it as the true value of θ. It is, of course, unknown. Then, for each θ, the quantities, $\ln f(x_i|\theta)$, are independent random variables with a common distribution, depending on θ_0, and, by the strong law, their mean converges strongly to their common expectation. By definition this expectation is

$$\mathscr{E}_0\{\ln f(x_i|\theta)\} = \int \ln f(x_i|\theta) . f(x_i|\theta_0)\,dx_i, \tag{3}$$

where the suffix has been added to the expectation sign to indicate the true value, θ_0, of θ. Hence the law says that with probability one

$$\lim_{n\to\infty} \{n^{-1} L(\mathbf{x}|\theta)\} = \mathscr{E}_0\{\ln f(x_i|\theta)\}. \tag{4}$$

Similarly, provided the derivatives and expectations exist,

$$\lim_{n\to\infty} \{n^{-1}\partial^r L(\mathbf{x}|\theta)/\partial\theta^r\} = \mathscr{E}_0\{\partial^r \ln f(x_i|\theta)/\partial\theta^r\}. \tag{5}$$

Equation (4) may be expressed in words by saying that, for large n, the log-likelihood behaves like a constant times n, where the constant is the expectation in (3). Similar results apply in the case of several parameters.

Theorem 1. *If a random sample of size n is taken from $f(x_i|\theta)$ then, provided the prior density, $\pi(\theta)$, nowhere vanishes, the posterior density of θ is, for large n, approximately normal with mean equal to the maximum likelihood estimate and variance, σ_n^2, given by*†

$$\sigma_n^{-2} = -\partial^2 L(\mathbf{x}|\hat{\theta})/\partial\theta^2. \tag{6}$$

It is not possible to give a rigorous proof of this theorem at the mathematical level of this book. The following 'proof' should convince most readers of the reasonableness of the result.

The posterior density is proportional to

$$\exp\{L(\mathbf{x}|\theta) + \ln \pi(\theta)\},$$

and since we have seen that $L(\mathbf{x}|\theta)$ increases like n, it will ultimately, as $n \to \infty$, dwarf $\ln \pi(\theta)$ which does not change with n. Hence the density is, apart from a constant, approximately

$$\exp\{L(\mathbf{x}|\theta)\} = \exp\{L(\mathbf{x}|\hat{\theta}) + \tfrac{1}{2}(\theta - \hat{\theta})^2 \partial^2 L(\mathbf{x}|\hat{\theta})/\partial\theta^2 + R\},$$

† $\partial^2 L(\mathbf{x}|\hat{\theta})/\partial\theta^2$ denotes the second derivative with respect to θ evaluated at $\hat{\theta}$.

on expanding $L(\mathbf{x}\,|\,\theta)$ by Taylor's theorem about $\hat{\theta}$, where R is a remainder term. Since the likelihood, and hence the log-likelihood, has a maximum at $\hat{\theta}$ the first derivative vanishes there. Also the second derivative will be negative there and may therefore be written $-\sigma_n^{-2}$. Furthermore, since it does not involve θ, the first term may be incorporated into the omitted constant of proportionality and we are left with

$$\exp\{-\tfrac{1}{2}(\theta-\hat{\theta})^2/\sigma_n^2+R\}. \tag{7}$$

From the discussions of the normal density in §2.5 it is clear that the term $\exp\{-\frac{1}{2}(\theta-\hat{\theta})^2/\sigma_n^2\}$ is negligible if $|\theta-\hat{\theta}|>3\sigma_n$; so that since σ_n^{-2} is, by (5), of order n, this term is only appreciable if θ differs from $\hat{\theta}$ by something of the order of $n^{-\frac{1}{2}}$. In that case the remainder term, R, which may be written

$$\tfrac{1}{6}(\theta-\hat{\theta})^3\,\partial^3L(\mathbf{x}\,|\,\theta_1)/\partial\theta^3$$

for some θ_1, is of order $n^{-\frac{3}{2}}$ times n (by (5)). Hence it is of order $n^{-\frac{1}{2}}$ and is negligible compared with the other term in (7). Hence, inserting the constant of proportionality, the posterior density is approximately

$$(2\pi\sigma_n)^{-\frac{1}{2}}\exp\{-\tfrac{1}{2}(\theta-\hat{\theta})^2/\sigma_n^2\}, \tag{8}$$

which establishes the result. Notice that σ_n^2 is, under the assumptions made here, of order n^{-1}.

Theorem 2. *If a random sample of size n is taken from*

$$f(x_i\,|\,\theta_1,\theta_2,\ldots,\theta_s)$$

then, provided the joint prior density, $\pi(\theta_1,\theta_2,\ldots,\theta_s)$ nowhere vanishes, the joint posterior density is, for large n, approximately multivariate normal with means equal to the maximum likelihood estimates $\hat{\theta}_i$ and a dispersion matrix whose inverse has typical element

$$-\partial^2L(\mathbf{x}\,|\,\hat{\theta}_1,\hat{\theta}_2,\ldots,\hat{\theta}_s)/\partial\theta_i\,\partial\theta_j. \tag{9}$$

This is the extension of theorem 1 to the case of several parameters. The proof proceeds as in that case. The important terms in the Taylor series expansion of $L(\mathbf{x}\,|\,\theta_1,\theta_2,\ldots,\theta_s)$ are

$$\frac{1}{2}\sum_{i,j=1}^{s}(\theta_i-\hat{\theta}_i)\,(\theta_j-\hat{\theta}_j)\,\partial^2L(\mathbf{x}\,|\,\hat{\theta}_1,\hat{\theta}_2,\ldots,\hat{\theta}_s)/\partial\theta_i\,\partial\theta_j, \tag{10}$$

and a comparison with the multivariate normal density (equation 3.5.17) establishes the result.

The matrix, whose typical element is given by (9), will be called the *information matrix*. It is the inverse of the dispersion matrix of the posterior density (compare the definition of precision in §5.1). Similarly (6) is called the *information*.

General remarks

Although known to earlier writers, the method of maximum likelihood has only become widely used through the work of R. A. Fisher, who obtained its principal properties. The main advantages of the method are that it produces a description of the posterior distribution which, because it is normal, is easy to handle, and which has a particularly simple mean and variance. (We shall see below that these are easy to compute.) Fisher used the method to provide a point estimate of θ. We shall not have much to say in this book about the problem of point estimation; by which is usually meant the problem of finding a single statistic which is, in some sense, near to the true value of a parameter (see §5.2); our reason for not doing so is that posterior distributions cannot be adequately described by one statistic.[†] But the problem is much discussed by some statisticians and, in large samples, is adequately solved by the maximum likelihood estimate, though other approximations are available. There is, as we shall see below, a close relationship between $\mathscr{D}^2(\hat{\theta})$ and σ_n^2 above: so that $\hat{\theta}$ and its variance do provide, because of the normality, an adequate description, in large samples, of the posterior density.

In the definition of $\hat{\theta}$ the word 'maximum' is used in the sense of 'largest value': that is, $L(\mathbf{x}\,|\,\hat{\theta}) \geqslant L(\mathbf{x}\,|\,\theta)$ for all θ. The estimate is therefore not necessarily determined by equating the first derivative to zero. This latter process will only yield the local maxima (or minima). An example where this process is inapplicable is provided by the uniform distribution discussed below. (In the 'proof' it has been assumed that the first derivative is zero at the maximum.)

[†] If we wished to use a single statistic we could take the mean of the posterior distribution. But this would not be helpful without, in addition, at least the variance.

Notice that the method has little or nothing to recommend it in small samples. There are two reasons for this. First, the posterior distribution is not necessarily normal. Secondly, the prior distribution is relevant in small samples because the information provided by it may be comparable with the information provided by the sample, and any method based on the likelihood alone may be misleading. We have mentioned the diminishing importance of the prior distribution as the sample size increases in connexion with the normal distribution (§§5.1, 5.2) but the point is quite general. Of course, if the prior knowledge is very vague (as in theorem 5.2.1), even a small sample may contain virtually all the information. Notice that, in the statement of the theorems, it has been necessary to assume that the prior density nowhere vanishes. If it did vanish near θ_1, say, then no amount of evidence would ever convince one that θ was near θ_1 (cf. §5.2) and the posterior density would vanish near θ_1 even if $\hat{\theta} = \theta_1$. (In the proof $\ln \pi(\theta)$ would be minus infinity, and certainly not negligible.)

Example: normal distribution

We begin with situations already studied, in order to see how the approximation compares with the exact result. In the case of the normal mean (§§5.1, 5.2) with known variance the likelihood is given by equation 5.1.9. The log-likelihood is therefore

$$L(\mathbf{x}\,|\,\theta) = C - \tfrac{1}{2}(\bar{x}-\theta)^2\,(n/\sigma^2),$$

where C is a constant. Differentiating and equating to zero gives the result $(\bar{x}-\hat{\theta})\,(n/\sigma^2) = 0$, so that $\hat{\theta} = \bar{x}$. A second differentiation gives

$$\sigma_n^{-2} = -\partial^2 L(\mathbf{x}\,|\,\hat{\theta})/\partial\theta^2 = (n/\sigma^2),$$

so that the posterior density is approximately $N(\bar{x}, \sigma^2/n)$, which agrees with theorem 5.2.1, and is exact (corollary to theorem 5.1.1) if the prior distribution of θ is uniform over the real line.

If the variance is also unknown (§5.4) then the logarithm of the likelihood is, from equation 5.4.3 rearranged in the same way as the posterior distribution was rearranged to obtain 5.4.4,

$$C - \tfrac{1}{2}n\ln\theta_2 - \{\nu s^2 + n(\bar{x}-\theta_1)^2\}/2\theta_2.$$

To obtain the maximum likelihood estimates we differentiate partially with respect to θ_1 and θ_2 and equate to zero. The results are

$$\left.\begin{aligned}\frac{\partial L}{\partial\theta_1} &= n(\bar{x}-\hat{\theta}_1)/\hat{\theta}_2 = 0,\\ \frac{\partial L}{\partial\theta_2} &= -\tfrac{1}{2}n/\hat{\theta}_2 + \{\nu s^2 + n(\bar{x}-\hat{\theta}_1)^2\}/2\hat{\theta}_2^2 = 0,\end{aligned}\right\} \tag{11}$$

so that $\hat{\theta}_1 = \bar{x}$ and $\hat{\theta}_2 = \nu s^2/n = \Sigma(x_i - \bar{x})^2/n$. (12)

The matrix whose elements are minus the second derivatives of the log-likelihood at the maximum (the information matrix of (9)) is easily seen to be

$$\begin{pmatrix} n\hat{\theta}_2^{-1} & 0 \\ 0 & \tfrac{1}{2}n/\hat{\theta}_2^2 \end{pmatrix}, \tag{13}$$

with inverse

$$\begin{pmatrix} \hat{\theta}_2/n & 0 \\ 0 & 2\hat{\theta}_2^2/n \end{pmatrix}. \tag{14}$$

The posterior distribution of θ_1 is thus approximately $N(\hat{\theta}_1, \hat{\theta}_2/n)$, or, from (12), $n^{\frac{1}{2}}(\theta_1-\bar{x})/s(\nu/n)^{\frac{1}{2}}$ is approximately $N(0, 1)$. This is in agreement with the exact result (theorem 5.4.1) that $n^{\frac{1}{2}}(\theta_1-\bar{x})/s$ has Student's distribution, since this distribution tends to normality as $n \to \infty$ (§5.4), and $\nu/n \to 1$. The distribution of θ_2 is approximately $N(\hat{\theta}_2, 2\hat{\theta}_2^2/n)$. This agrees with the exact result (theorem 5.4.2) that $\nu s^2/\theta_2$ is χ^2 with ν degrees of freedom, because the mean and variance of θ_2 are s^2 and $2s^4/\nu$ (equations 5.3.5 and 6) in large samples and the distribution of θ_2 tends to normality. This last result was proved in §5.3. Finally we note that the covariance between θ_1 and θ_2 is zero, which is in exact agreement with the result obtained in §5.4.

Example: binomial distribution

Consider a random sequence of n trials with constant probability θ of success. If r of the trials result in a success the likelihood is (cf. equation 5.5.9)

$$\theta^r(1-\theta)^{n-r}. \tag{15}$$

The derivative of the log-likelihood is therefore

$$r/\theta - (n-r)/(1-\theta),$$

so that $\hat{\theta} = r/n$, the proportion of successes in the n trials. The second derivative is

$$-r/\theta^2 - (n-r)/(1-\theta)^2,$$

which gives $\sigma_n^2 = r(n-r)/n^3$. These results agree with the exact results of §7.2.

Example: exponential family

The method of maximum likelihood is easily applied to any member of the exponential family. In the case of a single sufficient statistic for a single parameter the density is (equation 5.5.5)

$$f(x_i \mid \theta) = F(x_i)G(\theta)e^{u(x_i)\phi(\theta)}$$

and the log-likelihood is, apart from a constant,

$$ng(\theta) + t(\mathbf{x})\phi(\theta),$$

where $g(\theta) = \ln G(\theta)$ and $t(\mathbf{x}) = \sum_{i=1}^{n} u(x_i)$. The posterior density of θ is therefore approximately normal with mean equal to the root of

$$ng'(\theta) + t(\mathbf{x})\phi'(\theta) = 0 \qquad (16)$$

and variance equal to

$$\{-ng''(\theta) - t(\mathbf{x})\phi''(\theta)\}^{-1}$$

evaluated at that root. Similar results apply in the case of several sufficient statistics and parameters.

Solution of maximum likelihood equation

The equation for the maximum of the likelihood,

$$\partial L(\mathbf{x} \mid \theta)/\partial\theta = 0,$$

or, in the case of the exponential family, (16) above, may not be solvable in terms of elementary functions. However, there is an elegant numerical method of solving the equation, in the course of which σ_n^2 is also obtained. This is Newton's method of solving an equation. A reference to fig. 7.1.1 will explain the idea. On a graph of the derivative of the log-likelihood against θ a tangent to the graph is drawn at a first approximation, $\theta^{(1)}$, to the value of $\hat{\theta}$. The tangent intersects the θ-axis at a second approximation $\theta^{(2)}$ which is typically nearer to $\hat{\theta}$ than $\theta^{(1)}$ is, and, in any case, may be used in place of $\theta^{(1)}$ to repeat the process, obtaining $\theta^{(3)}$, and so on. The sequence $\{\theta^{(r)}\}$ usually converges to $\hat{\theta}$. Algebraically the method may be expressed by expanding in a Taylor series

$$\partial L(\mathbf{x} \mid \hat{\theta})/\partial\theta = 0 = \partial L(\mathbf{x} \mid \theta^{(1)})/\partial\theta + (\hat{\theta} - \theta^{(1)})\,\partial^2 L(\mathbf{x} \mid \theta^{(1)})/\partial\theta^2 + \dots$$

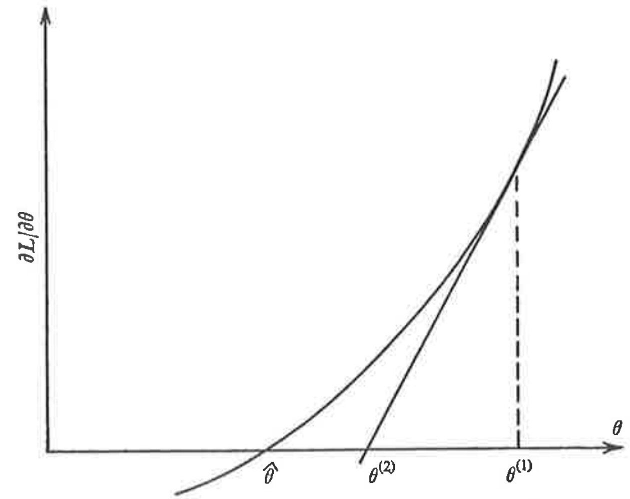

Fig. 7.1.1. Newton's method for solution of the maximum likelihood equation.

and retaining only the first two terms. The root of the equation for $\hat{\theta}$ so obtained is $\theta^{(2)}$, that is

$$\theta^{(2)} - \theta^{(1)} = \{\partial L(\mathbf{x} \mid \theta^{(1)})/\partial\theta\}/\{-\partial^2 L(\mathbf{x} \mid \theta^{(1)})/\partial\theta^2\}. \qquad (17)$$

It is not necessary to recalculate the second derivative at each approximation: the method will still work with a single value retained throughout. A final recalculation may be advisable at the termination of the process when $\hat{\theta}$ has been obtained to sufficient accuracy (that is when $\theta^{(r)} - \theta^{(r-1)}$ is negligible) in order to obtain a better value for $\sigma_n^2 = \{-\partial^2 L(\mathbf{x} \mid \theta^{(r)})/\partial\theta^2\}$.

The method is equally convenient when several parameters are involved. The Taylor series expansion gives

$$\partial L(\mathbf{x} \mid \boldsymbol{\theta}^{(1)})/\partial\theta_i = \sum_j (\theta_j^{(2)} - \theta_j^{(1)})\,\{-\partial^2 L(\mathbf{x} \mid \boldsymbol{\theta}^{(1)})/\partial\theta_i\partial\theta_j\}, \qquad (18)$$

where $\boldsymbol{\theta}^{(1)} = (\theta_1^{(1)}, \theta_2^{(1)}, \ldots, \theta_s^{(1)})$, a set of linear equations for $\boldsymbol{\theta}^{(2)}-\boldsymbol{\theta}^{(1)}$. The matrix which has to be inverted is the information matrix at argument $\boldsymbol{\theta}^{(1)}$ instead of at argument $\hat{\boldsymbol{\theta}}$. At the final approximation, $\boldsymbol{\theta}^{(r)}$, this has to be inverted in any case in order to obtain the dispersion matrix. Thus the method is well suited, not only to the evaluation of the means, but also to the evaluation of the dispersion matrix, of the posterior distribution. Numerical methods for the inversion of matrices are given in §8.4.

Example

As an illustration of Newton's method consider random samples from a Γ-distribution, or equivalently a χ^2-distribution, with both the index and the parameter unknown. The density for a single observation is (equation 2.3.7)

$$f(x_i \mid \theta_1, \theta_2) = \frac{\theta_1^{\theta_2}}{(\theta_2-1)!} e^{-x_i\theta_1} x_i^{\theta_2-1}, \tag{19}$$

where we have written θ_1 for λ and θ_2 for n. The likelihood for a random sample of size n is thus

$$\left\{\frac{\theta_1^{\theta_2}}{(\theta_2-1)!}\right\}^n \exp\left\{-\theta_1 \sum_{i=1}^n x_i + (\theta_2-1)\sum_{i=1}^n \ln x_i\right\}. \tag{20}$$

This shows that the distribution belongs to the exponential family and that $\bar{x} = \Sigma x_i/n$ and $\bar{y} = \Sigma \ln x_i/n$ are jointly sufficient statistics. Differentiation of the log-likelihood gives

$$\left.\begin{aligned} \frac{\partial}{\partial\theta_1}&: \quad n(-\bar{x}+\hat{\theta}_2/\hat{\theta}_1) = 0, \\ \frac{\partial}{\partial\theta_2}&: \quad n\left\{\ln\hat{\theta}_1+\bar{y}-\frac{d}{d\theta_2}\ln(\hat{\theta}_2-1)!\right\} = 0. \end{aligned}\right\} \tag{21}$$

The first of these equations is sufficiently simple to enable $\hat{\theta}_1$ to be eliminated and a single equation,

$$\ln\hat{\theta}_2 - \ln\bar{x} + \bar{y} - \frac{d}{d\theta_2}\ln(\hat{\theta}_2-1)! = 0, \tag{22}$$

for $\hat{\theta}_2$ to be solved by Newton's method. The derivative of the left-hand side of (22) is $\hat{\theta}_2^{-1}-(d^2/d\theta_2^2)\ln(\hat{\theta}_2-1)!$ and tables of the derivatives of the logarithm of the factorial function (see, for

example, Davis (1933)) enable the calculations to be carried out. It is necessary, however, to start with a first approximation. Here this is most easily obtained by using the approximation† to $d\ln(\hat{\theta}_2-1)!/d\theta_2$ of $\ln\hat{\theta}_2 - 1/2\hat{\theta}_2$, which, on insertion in (22), gives a value of θ_2 equal to $\{2(\ln\bar{x}-\bar{y})\}^{-1}$ to use as $\theta_2^{(1)}$ in the iteration. The approximation is remarkably good for all values of θ_2 except those near zero, so that except in that case, a single stage of Newton's procedure should be sufficient. We leave the reader to verify that the dispersion matrix is the matrix

$$\begin{pmatrix} \dfrac{d^2}{d\theta_2^2}\ln(\hat{\theta}_2-1)! & \hat{\theta}_1^{-1} \\ \hat{\theta}_1^{-1} & \hat{\theta}_2/\hat{\theta}_1^2 \end{pmatrix} \tag{23}$$

with each element divided by $n\hat{\theta}_1^{-2}\{\hat{\theta}_2 d^2\ln(\hat{\theta}_2-1)!/d\theta_2^2-1\}$.

These results might be of value if one wished to investigate whether observed incidents were occurring in a Poisson process (§2.3). It might be reasonable to suppose the intervals between successive incidents to be independent (for example if the incidents were failures of a component which was immediately replaced by a new one when it failed, §4.4), with a distribution of Γ-type. The Poisson process is the case $\theta_2 = 1$ (theorem 2.3.2), so one could perform an approximate significance test of the null hypothesis that $\theta_2 = 1$ by remarking that the posterior distribution of θ_2 is approximately normal with mean $\hat{\theta}_2$ and variance $\{n(d^2\ln(\hat{\theta}_2-1)!/d\theta_2^2-\hat{\theta}_2^{-1})\}^{-1} = \sigma_n^2$, say. The approximation to $d^2\ln(\theta_2-1)!/d\theta_2^2$ of $(1/\theta_2)+(1/2\theta_2^2)$, obtained from the above approximation to the first derivative by another differentiation of Stirling's formula, shows that σ_n^2 is approximately $2\hat{\theta}_2^2/n$. The result will therefore be significant at the 5 % level if $\hat{\theta}_2$ exceeds $1+2\sigma_n$. Notice that, in agreement with the general result, σ_n^2 is of the order n^{-1}.

Choice of parameter

In the method of maximum likelihood there is no distinction between the estimation of θ and the estimation of a function of θ, $\phi(\theta)$. We have the obvious relation that $\hat{\phi} = \phi(\hat{\theta})$. The

† This may be obtained by taking logarithms and differentiating both sides of Stirling's formula, equation 4.4.15.

variance of ϕ, $\{-\partial^2 L(\mathbf{x}|\hat{\phi})/\partial\phi^2\}^{-1}$, may also be related to the variance of θ in the following way. Write L for the log-likelihood in order to simplify the notation. Then

$$\frac{\partial L}{\partial \phi} = \frac{\partial L}{\partial \theta}\frac{d\theta}{d\phi} \quad \text{and} \quad \frac{\partial^2 L}{\partial \phi^2} = \frac{\partial^2 L}{\partial \theta^2}\left(\frac{d\theta}{d\phi}\right)^2 + \frac{\partial L}{\partial \theta}\frac{d^2\theta}{d\phi^2},$$

so that, since $\partial L/\partial\theta = 0$ at $\hat{\theta}$, the second equation gives

$$\mathscr{D}^2(\phi|\mathbf{x}) = \left(\frac{d\phi}{d\theta}\right)^2 \mathscr{D}^2(\theta|\mathbf{x}), \qquad (24)$$

where the derivative is evaluated at the maximum likelihood value. These results may also be obtained from theorem 3.4.1 since the variances are small, being of order n^{-1}. Thus in changing from θ to ϕ the means and variances change in the usual approximate way. Since the method does not distinguish between θ and ϕ, both parameters have an approximately normal distribution. At first glance this appears incorrect since if θ is normal then, in general, ϕ will not be normal. But it must be remembered that these results are only limiting ones as $n \to \infty$ and both the distributions of θ and ϕ can, and indeed do, tend to normality. What will distinguish θ from ϕ in this respect will be the rapidity of approach to normality: ϕ may be normal to an adequate approximation for smaller n than is the case with θ. It often pays, therefore, to consider whether some transform of θ is likely to be more nearly normal than θ itself and, if so, to work in terms of it. Of course, there is some transform of θ which is exactly normal since any (sufficiently respectable) distribution can be transformed into any other (compare the argument used in §3.5 for obtaining random samples from any distribution), but this transform will involve the exact posterior distribution and since the point of the approximation is to provide a simple result this is not useful. What is useful is to take a simple transformation which results in a more nearly normal distribution than is obtained with the untransformed parameter. No general results seem available here, but an example is provided by the variance θ_2 of a normal distribution just discussed. The distribution of θ_2, as we saw in §5.3, has a longer tail to the right (large θ_2) than to the left (small θ_2). This suggests considering $\ln\theta_2$ which might remove the effect of the long tail.

Detailed calculations show that the posterior distribution of $\ln\theta_2$ is more nearly normal than that of θ_2, though even better transformations are available. The approximate mean and variance of the distribution of $\ln\theta_2$ may be found either by maximum likelihood directly, equation (9), or from the results for θ_2, equation (14), combined with equation (24). Other examples will occur in later sections.

Distribution of the maximum likelihood estimate

We saw in §5.1 that when making inferences that involved using the mean of a normal distribution there were two distinct results that could be confused (statements (*a*) and (*b*) of that section). A similar situation obtains here because of the approximate normality of the posterior distribution. The two statements are:

(*a*) the maximum likelihood estimate, $\hat{\theta}$, is approximately normally distributed about θ_0 with variance the inverse of

$$I_n(\theta_0) = \mathscr{E}_0\{-\partial^2 L(\mathbf{x}|\theta_0)/\partial\theta^2\}; \qquad (25)$$

(*b*) the parameter θ is approximately normally distributed about $\hat{\theta}$ with variance σ_n^2.

(θ_0 is the true, unknown, fixed value of θ as explained before, equation (3).) Statement (*b*) is the result of theorem 1, statement (*a*) can be proved in essentially the same way as that theorem was proved. Statement (*a*) is a result, in frequency probability, about a statistic, $\hat{\theta}$: (*b*) is a result, in degrees of belief, about a parameter θ. In practice (*a*) and (*b*) can be confused, as explained in §5.1, without harm. Actually (*a*) is rarely used in the form given, since θ_0 is unknown and yet occurs in the variance (equation (25)). Consequently, (25) is usually replaced by $I_n(\hat{\theta})$. This still differs from σ_n^{-2} because of the expectation† used in (25) but not in the expression for σ_n^{-2}. It is interesting to note that the use of the expectation makes no difference in random samples of fixed size from an exponential family: there $I_n(\hat{\theta}) = \sigma_n^{-2}$. (See equation (16) and the one immediately following.)

† Those who have read the relevant paragraph in §5.6 will notice that the use of the expectation violates the likelihood principle, and is, on this score, unsatisfactory.

Exceptional cases

It is necessary to say a few words about the question of rigour in the proofs of the theorems. A complete proof with all the necessary details is only possible when certain assumptions are made about the likelihood: for example, assumptions about the existence and continuity of derivatives and their expectations. These assumptions are not always satisfied and the theorem is not always true; the most common difficulty arises when the range of possible values of x_i depends on θ. The difficulty is the same as that encountered in discussing sufficiency in §5.5 and the same example as was used there suffices to demonstrate the point here. If

$$f(x_i|\theta) = \theta^{-1} \quad (0 \leqslant x_i \leqslant \theta),$$

and is otherwise zero; then the likelihood is θ^{-n} provided $\theta \geqslant \max_i x_i = X$, say, and is otherwise zero. Hence the posterior density is approximately proportional to θ^{-n} and clearly this does not tend to normality as $n \to \infty$. Indeed, the maximum value is at $\theta = X$, so that $\hat{\theta} = X$, at the limit of the range of values of θ with non-zero probability. If the prior distribution of θ is uniform over the positive half of the real line, a simple evaluation of the constant shows that

$$\pi(\theta|\mathbf{x}) = (n-1)X^{n-1}\theta^{-n} \quad (\theta \geqslant X, n > 1),$$

with mean $(n-1)X/(n-2)$ and variance $(n-1)X^2/(n-3)(n-2)^2$ if $n > 3$. As $n \to \infty$ the variance is approximately X^2/n^2, whereas the theorem, if applicable here, would give a result that is of order n^{-1}, not n^{-2}. The estimation of θ is much more accurate than in the cases covered by the theorem. The practical reason for the great accuracy is essentially that any observation, x_i, immediately implies that $\theta < x_i$ has zero posterior probability; since, if $\theta < x_i$, x_i has zero probability. This is a much stronger result than can usually be obtained. The mathematical reason is the discontinuity of the density and its differential with respect to θ at the upper extreme of the range of x. Difficulties can also occur with estimates having smaller accuracy than suggested by the theorem, when dealing with several parameters. Examples of this phenomenon will not arise in this book.

7.2. Random sequences of trials

In this section we consider the simple probability situation of a random sequence of trials with constant probability, θ, of success, and discuss the inferences about θ that can be made. If a random variable, x, has a density

$$\frac{(a+b+1)!}{a!\,b!}x^a(1-x)^b, \tag{1}$$

for $0 \leqslant x \leqslant 1$ and $a, b > -1$, it is said to have a *Beta-distribution* with parameters a and b. We write it $B_0(a, b)$, the suffix distinguishing it from the binomial distribution $B(n, p)$ with index n and parameter p (§2.1).

Theorem 1. *If, with a random sequence of n trials with constant probability, θ, of success, the prior distribution of θ is $B_0(a, b)$, then the posterior distribution of θ is $B_0(a+r, b+n-r)$ where r is the number of successes.*

The proof is straightforward:

$$\pi(\theta) \propto \theta^a(1-\theta)^b, \quad \text{from (1)}, \tag{2}$$

the likelihood is $\quad p(\mathbf{x}|\theta) = \theta^r(1-\theta)^{n-r},$

so that
$$\pi(\theta|\mathbf{x}) \propto \theta^{a+r}(1-\theta)^{b+n-r}, \tag{3}$$

proving the result. (Here $\mathbf{x}$ denotes the results of the sequence of trials. Since r is sufficient $\pi(\theta|\mathbf{x})$ may be written $\pi(\theta|r)$.)

Corollary 1. *Under the conditions of the theorem the posterior distribution of*

$$F = \left(\frac{b+n-r+1}{a+r+1}\right)\left(\frac{\theta}{1-\theta}\right) \tag{4}$$

is $F[2(a+r+1), 2(b+n-r+1)]$.

From (4) $\quad dF/d\theta \propto (1-\theta)^{-2}$

and also $\quad \theta = a'F/(b'+a'F),$

where $a' = a+r+1$, $b' = b+n-r+1$. Substitution in (3), not forgetting the derivative (theorem 3.5.1), gives

$$\begin{aligned}\pi(F|\mathbf{x}) &\propto F^{a'-1}/(b'+a'F)^{a'+b'}\\ &\propto F^{a'-1}/(2b'+2a'F)^{a'+b'}.\end{aligned}$$