CHAPTER 5

MAXIMUM SUPPORT: THE METHOD OF MAXIMUM LIKELIHOOD

5.1. INTRODUCTION

In the past four chapters we have seen how to assess rival hypotheses by the Method of Support; how this entails a simple calculation in the case of two discrete hypotheses, and the drawing of a support curve in the case of a family of hypotheses specified by a parameter which may take any value in a particular range. With discrete hypotheses we can hope for no simpler treatment than the calculation of a single value of relative support, but with continua of hypotheses specified by one or more parameters, the quotation of the complete support function, or the drawing of its graph, is usually an unnecessary, and sometimes an impossible, labour. For the support function itself is not directly interpretable, and the drawing of its graph will be impossible with more than two parameters. It is therefore natural to seek ways of conveying most of the information in simpler numerical form, and to elaborate methods for obtaining and handling these numbers. In view of our dedication to the Method of Support, we shall pay particular attention to the parameter values which jointly (if there is more than one parameter) maximize the support.

In this chapter and the next we shall assume (unless otherwise indicated) that the support function is sufficiently regular for the suggested methods to be valid. In particular it will be assumed to be free of points of infinite support, to be unimodal, and to possess derivatives of all necessary orders. Situations leading to irregular support functions will be treated in chapter 8.

DEFINITION

The best-supported value of a parameter (that value for which, on the data, the support is a maximum) is called the *evaluate*. In the case of two or more parameters, the evaluates are those for which the support is a maximum over all parameters jointly.

The corresponding word in standard statistical theory is *estimate*; I shall introduce the words *evaluator* and *evaluation* as the substitutes for *estimator* and *estimation*.

With the emphasis placed on the evaluate, three questions call for attention: first, by what means shall we describe the shape of the support curve in the vicinity of the maximum; secondly, how shall we combine evaluates from different sets of data; and thirdly, how shall we locate the maximum when drawing the curve is impossible?

Before treating these questions, we must momentarily digress to consider *sufficient evaluates*. Sometimes (given the model and the sample size) the whole support function is uniquely determined by the position of its maximum, in which case the evaluate is itself a sufficient statistic and may be referred to as a *sufficient evaluate*. Indeed, it must be minimally sufficient in the sense of section 2.3. In the case of k parameters we may find *jointly sufficient evaluates*: k evaluates, one for each parameter, which jointly specify the support function. In general the individual evaluates will not be severally sufficient for the corresponding parameters.

The phenomenon of sufficient evaluates means that our aim of specifying the entire support function by a single number, given the sample size and the model, has already been achieved in those few but important situations where sufficiency occurs. Unfortunately the proviso 'given the model' is an important one, for without the model we do not know the functional form of the support. When the question of interpretation arises, therefore, the occurrence of sufficiency is not a great help, for with or without it we still have a support curve whose shape near the maximum we wish to communicate. In most of what follows, sufficiency will only be of marginal interest.

5.2. INDICES FOR THE SHAPE OF THE SUPPORT CURVE

In example 3.4.3 I suggested that the support curve for the binomial probability p given a sample of 33 successes and 47 failures (figure 2) could be summarized by quoting the evaluate, or best-supported value, of p and the two values for which the support is 2 units below the maximum, thus: $\hat{p} = 0.4125$ (0.3066, 0.5241).

DEFINITION

The *m-unit support limits* for a parameter are the two parameter values astride the evaluate at which the support is m units less than the maximum. The ***m*-unit support region** for a number of parameters is that region in the parameter space bounded by the curve on which the support is m units less than the maximum.

Support limits are perhaps the most natural way of numerically communicating information about a parameter, but they have disadvantages when one wishes to combine the results of different experiments, and they do not readily generalize to the multi-parameter case: a support region is better in theory than in practice. An alternative method, which readily generalizes to the case of many parameters, is to obtain the Taylor's series approximation to the support curve in the region of the maximum, and thus to use the second partial-differential coefficients.

In the case of a single parameter θ, suppose the support function be $S(\theta)$. Then the evaluate of θ is, in well-behaved situations, the solution of

$$\frac{dS}{d\theta} = 0.$$

DEFINITION

The first derivative of the support function is known as the *score*; taken at the evaluate, the score is zero.

Writing the evaluate $\hat{\theta}$, the support at any other value of θ is approximately given by the Taylor expansion

$$S(\theta) = S(\hat{\theta}) + (\theta - \hat{\theta})\frac{dS}{d\theta} + \tfrac{1}{2}(\theta - \hat{\theta})^2\frac{d^2S}{d\theta^2} + \ldots,$$

where the differential coefficients are evaluated at $\theta = \hat{\theta}$. But at this point $\frac{dS}{d\theta}$ is zero, so that approximately we have

$$S(\theta) = S(\hat{\theta}) + \tfrac{1}{2}(\theta - \hat{\theta})^2\frac{d^2S}{d\theta^2}. \qquad (5.2.1)$$

DEFINITION

Minus the second derivative of the support function is known as the *information*; when taken at the evaluate, it is known as the *observed information*.

The complete justification for the use of the word 'information' in this context will be postponed until chapter 7; at this stage we

may simply observe that the usage is intuitively satisfactory, for the difference in support between $\hat{\theta}$ and some other value θ, which we may expect to be a measure of the informativeness of the data about θ, is proportional to the observed information in the region near $\hat{\theta}$ (equation 5.2.1).

We note that the observed information may be interpreted geometrically as the spherical curvature of the support curve at its maximum, and hence that its reciprocal is the radius of curvature. This is precisely the kind of index that we need for communicating the *form* of the curve near the maximum, and it is useful to give it a special name:

DEFINITION[1]

The radius of curvature of the support curve at its maximum, being the reciprocal of the observed information, is called the *observed formation*. The word *formation* alone may be used for the reciprocal of the information at points other than the maximum.

EXAMPLE 5.2.1

We have seen that the likelihood for p, the parameter of a binomial distribution, is proportional to

$$p^a(1-p)^b,$$

given a successes and b failures. The support function is thus

$$S(p) = a \ln p + b \ln (1-p),$$

which is maximized for p at

$$\frac{dS}{dp} = \frac{a}{p} - \frac{b}{(1-p)} = 0,$$

whence

$$\hat{p} = \frac{a}{a+b},$$

the proportion of successes in the sample. We have already noticed this solution for the evaluate in earlier numerical examples.

$$-\frac{d^2S}{dp^2} = \frac{a}{p^2} + \frac{b}{(1-p)^2}$$

is the *information*, and writing $a = n\hat{p}$ and $b = n(1-\hat{p})$, where $n = a + b$ is the sample size, the *observed information* is

$$n\left(\frac{1}{\hat{p}} + \frac{1}{1-\hat{p}}\right) = \frac{n}{\hat{p}(1-\hat{p})}.$$

Hence the *observed formation* is $\hat{p}(1-\hat{p})/n$.

Readers who are familiar with standard statistical parlance will recognize the observed formation in the above example as equal to the variance of the proportion of successes observed in a sample of n when the probability of success is *known* to be $\hat{p}$. We must wait until the theorems of chapter 7 before we can see why, but now is an opportune moment to reflect on the inadequacy of statements of the form 'the estimate of p is $\hat{p} = a/n$, and the standard error of the estimate is $\sqrt{\{\hat{p}(1-\hat{p})/n\}}$'. This is objectionable first because it is not true, the standard error of $\hat{p}$ in fact being $\sqrt{\{p^*(1-p^*)/n\}}$, where p^* is the unknown 'true' value of p, the result therefore being only asymptotically correct as $\hat{p} \to p^*$; and secondly because, even in the possession of the complete form of the distribution of $\hat{p}$ given p^*, the Likelihood Axiom indicates that only the Method of Support allows us to make comparative statements about p^* given $\hat{p}$, and they will involve the observed formation, and not a variance. Nevertheless, it is encouraging to find the conventional methods so closely shadowed by the Method of Support, for in spite of their logical frailty experience has shown them not to be misleading in well-behaved situations.

It will not always be possible, as it was in the above example, to quote the observed formation in terms of the evaluate alone. For this to be done, it will be necessary for the support function to be expressible in terms of the evaluate, in which case it is then a sufficient evaluate, as defined in section 5.1.

The radius of curvature suffers one disadvantage as a measure of the shape of a curve near its maximum, for it is not a linear measure of the width of the curve some specified distance below the peak: in fact it is proportional to the square of such a width. A measure of the width, which is perhaps the most meaningful index intuitively, is thus afforded by the square-root of the radius of curvature.

DEFINITION

The square-root of the observed formation is called the *span*, and is a measure of the width of the support curve near the maximum.

In geometrical terms, a Taylor's series approximation corresponds to the replacement of the true support curve by a parabola passing though the maximum, with axis vertical and

having the same radius of curvature as the true curve at the maximum. An example is given in figure 11. If we denote the *span* by w, where

$$w^2 = -1\Big/\frac{d^2S}{d\theta^2}$$

taken at the maximum, the equation of the approximating parabola is

$$S(\theta) = S(\hat{\theta}) - (\theta - \hat{\theta})^2/2w^2. \tag{5.2.2}$$

For any particular value of $S(\theta)$ this is a quadratic equation in θ, with roots symmetrically placed about $\hat{\theta}$, and if the roots are chosen so that the distance between them, $2(\theta - \hat{\theta})$, is equal to the span

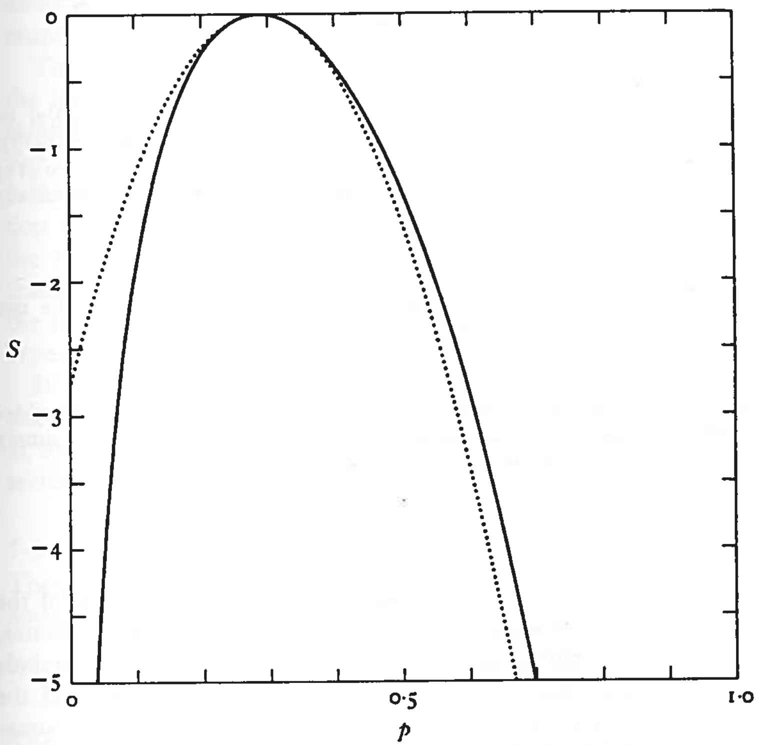

Figure 11. The support curve for the binomial parameter p for a sample of 4 successes and 10 failures (curve (*a*) of figure 2), together with its approximating parabola (dotted curve).

w, the support at the roots will be less than that at the maximum by an amount equal to $(\theta - \hat{\theta})^2/8(\theta - \hat{\theta})^2$, or $\frac{1}{8}$. Hence the span is the width of the approximating parabola to the support curve at $\frac{1}{8}$ of a unit of support below the maximum. Similarly, we see that it is the semi-width half a unit below the maximum. At one unit the width is $2\sqrt{2}$ times the span, and at m units $2\sqrt{(2m)}$ times the span. Since the half-width is therefore $\sqrt{(2m)}$ times the span, the m-unit support limits on the true curve are approximately given by the parameter values $\sqrt{(2m)}$ from the evaluate on either side of it. In particular, the 2-unit support limits are at $\pm 2w$. Since the span corresponds to the square-root of the variance, it is the analogue of the standard deviation.

EXAMPLE 5.2.2

The span of the support curve for a binomial parameter is

$$\sqrt{\{\hat{p}(1 - \hat{p})/n\}}.$$

The 2-unit support limits are thus approximately $2\sqrt{\{\hat{p}(1 - \hat{p})/n\}}$ on either side of the evaluate. In example 3.4.3, n was 80 and $\hat{p}$ 0.4125. The limits are therefore approximately 0.4125 ± 0.1101, or $\hat{p} = 0.4125$ (0.3024, 0.5226). It may be recalled that the exact limits were $\hat{p} = 0.4125$ (0.3066, 0.5241).

EXAMPLE 5.2.3

We may note the curious fact, given by Todhunter,[2] that the two points of inflection of the binomial likelihood curve

$$p^a(1 - p)^b$$

are equidistant from the maximum $\hat{p} = a/n$, where $n = a + b$. Furthermore the square of the distance h of each from the maximum is simply $\hat{p}(1 - \hat{p})/(n - 1)$. Thus h is given by

$$h^2 = \frac{n}{n - 1} w^2.$$

The extent to which the span is an accurate reflection of the width of the true support curve near the maximum is, of course, dependent on the goodness-of-fit of the approximating parabola and the excellence of the Taylor's series approximation. If the support curve is symmetrical about the evaluate, the approximation should be satisfactory, for then all the odd-order derivatives are zero at the maximum. It may, in some cases, be worthwhile to transform to a new variable for which the support curve is

more nearly symmetrical, but one may lose in ease of interpretation what one gains in accuracy. Examples involving transformations are given in later sections of this chapter.

It is interesting to note that if the support curve takes the parabolic form

$$S(\theta) = S(\hat{\theta}) - (\theta - \hat{\theta})^2/2w^2$$

the likelihood must have the form

$$e^{S(\theta)} = k\, e^{-(\theta-\hat{\theta})^2/2w^2},$$

that is, the form of a Normal curve with mean $\hat{\theta}$ and standard deviation equal to the span w. There is, however, in this correspondence, no implication of a Normal *probability distribution*. The *distribution* of evaluates will be a matter for consideration in chapter 7.

There is a theorem to the effect that, under suitable conditions, the likelihood becomes more and more Normal in form as the sample size increases, but it is a rather weak and useless theorem. For if, as it asserts, the likelihood on θ *and* the likelihood on ϕ *both* tend to the Normal form, where ϕ is a one-to-one transformation of θ, it is really only saying that with a large-enough sample the range of interest around the maximum is so small that the transformation in this range is practically linear, and that within the range the support function is well-approximated by a Taylor expansion up to the quadratic term.

It would only be surprising were it not true that as the sample size increases the support becomes more and more concentrated at the true value of the parameter, a result proved in the next section.

5.3. GENERAL FORMS FOR THE SCORE AND INFORMATION

The general form for the support function in the case of a multinomial sample was given (equation (3.4.4)) as

$$S(\theta) = \sum_{i=1}^{s} a_i \ln p_i(\theta). \tag{5.3.1}$$

The score may therefore be written

$$\frac{dS}{d\theta} = \sum_{i=1}^{s} \frac{a_i}{p_i}\frac{dp_i}{d\theta}, \tag{5.3.2}$$

and the information

$$-\frac{d^2S}{d\theta^2} = \sum_{i=1}^{s}\left\{\frac{a_i}{p_i^2}\left(\frac{dp_i}{d\theta}\right)^2 - \frac{a_i}{p_i}\frac{d^2p_i}{d\theta^2}\right\}. \tag{5.3.3}$$

It is generally easier to work from first principles than to remember these forms, but they will be used to derive the *expected score* and the *expected information* in section 7.2.

For a continuous distribution we had (equation (3.4.5))

$$S(\theta) = \sum_{i=1}^{n} \ln f(x_i, \theta), \tag{5.3.4}$$

whence

$$\frac{dS}{d\theta} = \sum_{i=1}^{n} \frac{1}{f}\frac{df}{d\theta} \tag{5.3.5}$$

and

$$-\frac{d^2S}{d\theta^2} = \sum_{i=1}^{n}\left\{\frac{1}{f^2}\left(\frac{df}{d\theta}\right)^2 - \frac{1}{f}\frac{d^2f}{d\theta^2}\right\}, \tag{5.3.6}$$

where f stands for $f(x_i, \theta)$.

The forms for discrete and continuous distributions are, of course, quite equivalent, but it is more convenient to sum over classes with the former and over individual observations with the latter.

Suppose θ^* is the 'true' value of θ in a multinomial situation. As the sample size increases, the observed class frequencies a_i will approach their expectations $np_i(\theta^*)$, and the score will approach

$$n\sum_{i=1}^{s}\left(\frac{p_i(\theta^*)}{p_i(\theta)}\frac{dp_i(\theta)}{d\theta}\right).$$

At $\theta = \theta^*$ this reduces to

$$n\sum_{i=1}^{s}\left(\frac{dp_i(\theta)}{d\theta}\right)_{\theta=\theta^*} = n\frac{d}{d\theta}\left(\sum_{i=1}^{s} p_i\right) = 0,$$

showing that as the sample size increases indefinitely, the true value of the parameter is approached by a turning point of the support curve. The information is then easily seen to be positive, indicating that the turning point is a maximum. But at any value of θ other than θ^* the score will increase proportionately with n, the sample size, and since the score represents the gradient of the support curve, it is evident that as the sample size increases, the support other than at the true value of the parameter decreases, by comparison, indefinitely. A similar argument may be applied to the continuous case.

This property of evaluates, that they approach the true value of the parameter (where such a concept is meaningful) as the sample size increases, is called *consistency*, and evaluates are said to be *consistent*. It is obviously then a desirable property. There has been some discussion of the proper definition of consistency, and of the behaviour of evaluates in irregular situations; the reader is referred to Rao[3] for an introduction to the subject.

5.4. TRANSFORMATION AND COMBINATION OF EVALUATES

In section 2.5 we noted that the likelihood function referred to a new parameter ϕ, where $\theta = f(\phi)$, θ being the old parameter and f a one-to-one transformation, is found by the direct substitution of $f(\phi)$ for θ; and hence that the maximizing values of θ and ϕ, $\hat\theta$ and $\hat\phi$, are related by the equation $\hat\theta = f(\hat\phi)$. In chapter 3 we saw how this conformed to our requirements for a measure of support. The analytic demonstration of the transformation property of evaluates is immediate, for at the evaluate we have

$$\frac{dS}{d\phi} = \frac{d\theta}{d\phi}\cdot\frac{dS}{d\theta} = 0.$$

THEOREM 5.4.1

The observed informations are related by the formula

$$\frac{d^2S}{d\phi^2} = \left(\frac{d\theta}{d\phi}\right)^2\frac{d^2S}{d\theta^2}, \tag{5.4.1}$$

where $d\theta/d\phi$ is taken at the evaluate.

Proof.

$$\frac{d^2S}{d\phi^2} = \frac{d^2\theta}{d\phi^2}\cdot\frac{dS}{d\theta} + \left(\frac{d\theta}{d\phi}\right)^2\cdot\frac{d^2S}{d\theta^2};$$

but at the evaluate $dS/d\theta$ is zero, and the theorem follows immediately.

If w_ϕ and w_θ are the spans of $\hat{\phi}$ and $\hat{\theta}$ respectively, then they are related by the equation

$$w_\phi^2 = w_\theta^2 \Big/ \left(\frac{\mathrm{d}\theta}{\mathrm{d}\phi}\right)^2. \qquad (5.4.2)$$

It must of course be remembered that any non-linear transformation of the parameter will render the support curve either more or less parabolic at its maximum, so that its representation by the evaluate and the span is rendered either more or less accurate. Support limits directly transformed will not correspond to the support limits found from the new span. As I have mentioned, we may sometimes take advantage of a transformation to improve the approximation of the support curve by a parabola. The support limits for the new parameter may then be directly transformed into limits for the old parameter, and any asymmetry about the evaluate which they then exhibit is an indication of the asymmetry of the support curve. An example (5.6.2) is given below in connection with Newton–Raphson iteration.

An appropriate transformation will be one which renders the third derivative of the support function, with respect to the new parameter, zero at the evaluate. We already have

$$\frac{\mathrm{d}^2S}{\mathrm{d}\phi^2} = \frac{\mathrm{d}^2\theta}{\mathrm{d}\phi^2}\cdot\frac{\mathrm{d}S}{\mathrm{d}\theta} + \left(\frac{\mathrm{d}\theta}{\mathrm{d}\phi}\right)^2\frac{\mathrm{d}^2S}{\mathrm{d}\theta^2},$$

whence, on differentiating again and setting $\mathrm{d}S/\mathrm{d}\theta = 0$,

$$\frac{\mathrm{d}^3S}{\mathrm{d}\phi^3} = \frac{\mathrm{d}\theta}{\mathrm{d}\phi}\left\{\frac{\mathrm{d}^3S}{\mathrm{d}\theta^3}\left(\frac{\mathrm{d}\theta}{\mathrm{d}\phi}\right)^2 + 3\frac{\mathrm{d}^2S}{\mathrm{d}\theta^2}\cdot\frac{\mathrm{d}^2\theta}{\mathrm{d}\phi^2}\right\},$$

taken at the evaluate. It follows that if the third derivative with respect to ϕ is to be zero,

$$\frac{\left(\dfrac{\mathrm{d}\theta}{\mathrm{d}\phi}\right)^2}{\dfrac{\mathrm{d}^2\theta}{\mathrm{d}\phi^2}} = -\frac{3\dfrac{\mathrm{d}^2S}{\mathrm{d}\theta^2}}{\dfrac{\mathrm{d}^3S}{\mathrm{d}\theta^3}}.$$

EXAMPLE 5.4.1[4]

In a Poisson distribution, the probability that the variate takes the value r ($r = 0, 1, 2, \ldots$) is

$$\frac{e^{-\lambda}\lambda^r}{r!}$$

In a sample of n, let the observed frequency in class r be a_r. Then the support function may be taken as

$$S(\lambda) = \sum_{r=0}^{\infty} a_r(-\lambda + r\ln\lambda)$$
$$= n(\bar{r}\ln\lambda - \lambda),$$

showing that $\bar{r}$ is a sufficient statistic for λ.

$$\frac{\mathrm{d}S}{\mathrm{d}\lambda} = n\left(\frac{\bar{r}}{\lambda} - 1\right),$$

from which we see that $\bar{r}$ is the evaluator of λ. Furthermore,

$$\frac{\mathrm{d}^2S}{\mathrm{d}\lambda^2} = -\frac{n\bar{r}}{\lambda^2};$$

the observed formation of the evaluate is therefore $\hat{\lambda}/n$.

Proceeding to the third derivative we find

$$\frac{\mathrm{d}^3S}{\mathrm{d}\lambda^3} = \frac{2n\bar{r}}{\lambda^3},$$

which is not zero at $\lambda = \bar{r}$.

We thus require a new parameter, say $\phi = \phi(\lambda)$, for which

$$\frac{\left(\dfrac{\mathrm{d}\lambda}{\mathrm{d}\phi}\right)^2}{\dfrac{\mathrm{d}^2\lambda}{\mathrm{d}\phi^2}} = \frac{3\lambda}{2}.$$

The simplest solution to this differential equation is $\phi = \lambda^{1/3}$, which is therefore a suitable transformation. The support function for λ when $n = 10$ and $\bar{r} = 0.8$ is shown in figure 12, and for $\lambda^{1/3}$ in figure 13.

One of our requirements for a measure of support was that it should be additive over independent sets of data, and we have seen how support, as here defined, and hence a support function, satisfies this requirement. It immediately follows that the score and the information are additive over independent sets.

Since we may wish to work in terms of evaluates and their spans, the question arises as to how these may be combined from independent sets of data, given that the support functions may be added. In the presence of sufficient evaluates, the addition of the support functions will correspond to some well-defined combining operation on the evaluates, and an exact solution will be possible. Thus we have already seen that the combination of binomial samples corresponds to the summing of the number of successes and of the number of failures; in terms of the sufficient evaluate a/n and the sample size n, the procedure amounts to finding the weighted sufficient evaluate, the weights being given by the sample sizes. In general, however, we will work with evaluates and their informations, and may anticipate that any such combination of values will be an approximate procedure, dependent on the excellence of the Taylor's series approximations.

THEOREM 5.4.2

Evaluates may be combined approximately by forming their weighted average, the weights being equal to their observed

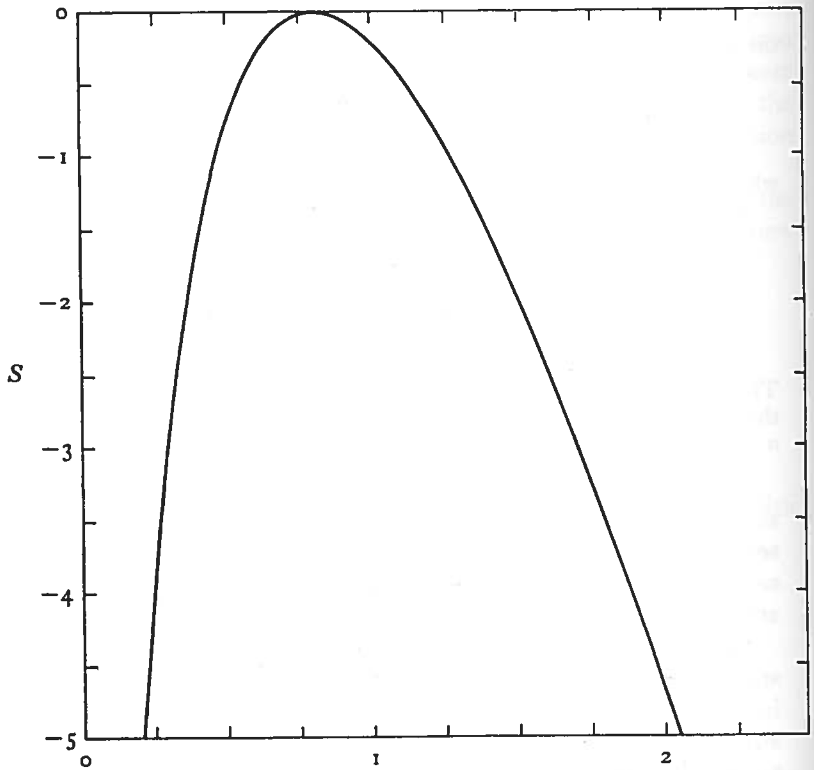

Figure 12. The support curve for the Poisson parameter λ for a sample of 10 with mean 0.8.

informations. The combined observed information is approximately equal to the sum of the individual observed informations.

We prove the theorem for the combination of two values, the extension to any number being immediate.

Proof. Let the two evaluates be $\hat{\theta}_1$ and $\hat{\theta}_2$, derived from support functions $S_1(\theta)$ and $S_2(\theta)$, and let them have observed formations w_1^2 and w_2^2. Then, to a quadratic approximation,

$$S_1(\theta) = S_1(\hat{\theta}_1) - (\theta - \hat{\theta}_1)^2/2w_1^2$$

and

$$S_2(\theta) = S_2(\hat{\theta}_2) - (\theta - \hat{\theta}_2)^2/2w_2^2.$$

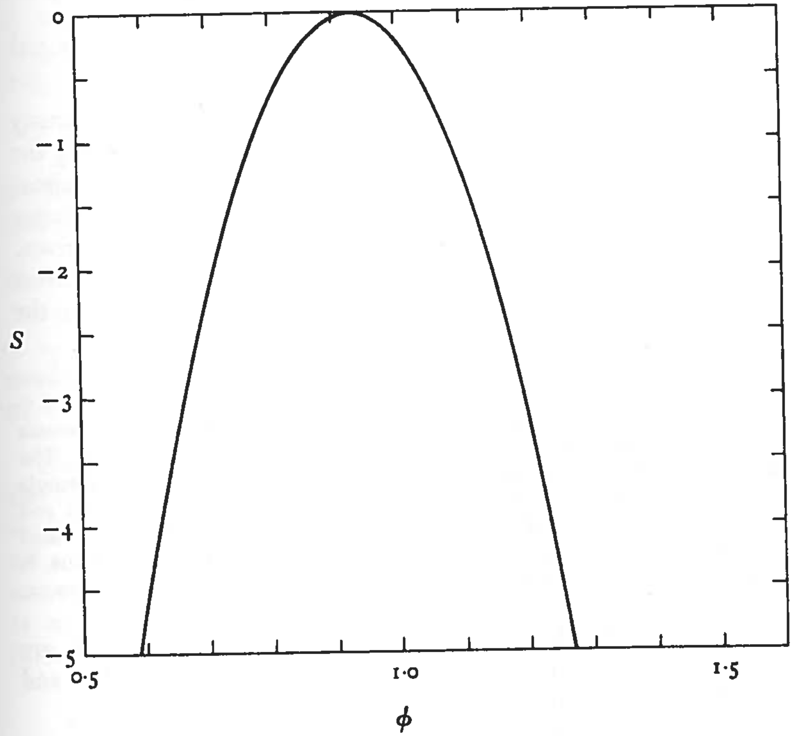

Figure 13. The support curve of figure 12 as a function of $\phi = \lambda^{1/3}$, demonstrating the efficacy of this transformation in rendering the peak more symmetrical.

For a combined quadratic support function we have, therefore,

$$S_3(\theta) = S_1(\theta) + S_2(\theta)$$
$$= S_1(\hat\theta_1) + S_2(\hat\theta_2) - \frac{(\theta-\hat\theta_1)^2}{2w_1^2} - \frac{(\theta-\hat\theta_2)^2}{2w_2^2},$$

which is maximized for variation in θ at

$$\frac{dS_3}{d\theta} = -\theta\left(\frac{1}{w_1^2}+\frac{1}{w_2^2}\right) + \frac{\hat\theta_1}{w_1^2} + \frac{\hat\theta_2}{w_2^2} = 0,$$

of which the solution is

$$\hat\theta_3 = \left(\frac{\hat\theta_1}{w_1^2}+\frac{\hat\theta_2}{w_2^2}\right)\Big/\left(\frac{1}{w_1^2}+\frac{1}{w_2^2}\right). \qquad (5.4.3)$$

Differentiating again we find that the formation of the combined evaluate is given by

$$\frac{1}{w_3^2} = -\frac{d^2S_3}{d\theta^2} = \frac{1}{w_1^2}+\frac{1}{w_2^2}. \qquad (5.4.4)$$

In practice it should be remembered that, even with many parameters, there will rarely be any excuse for not summing the true support functions rather than their Taylor's series approximations. The above theorem is really a last resort when only evaluates and their spans are quoted. If accurate support limits are given, but the form of the support function is unknown, the correct procedure would be to reconstruct the support curve by fitting the approximating cubic.

EXAMPLE 5.4.2

In section 2.3 we had two binomial samples, one with 4 successes and 10 failures, and another with 29 successes and 37 failures. The evaluates were $\hat p_1 = 0.2857$ and $\hat p_2 = 0.4394$. Application of the formula for the observed information (example 5.2.1) gives $1/w_1^2 = 68.60$ and $1/w_2^2 = 267.94$, whence $1/w_3^2 = 336.54$ by addition. The combined evaluate is found from (5.4.3) to be $\hat p_3 = 0.4081$. The correct value, it may be recalled, was 0.4125; the correct observed information is 330.11.

In probability theory there is a theorem that if θ_1 and θ_2 are independent random variables with means $E(\theta_1)$, $E(\theta_2)$ and variances $V(\theta_1)$, $V(\theta_2)$, then $\theta_1+\theta_2$ has mean

$$E(\theta_1+\theta_2) = E(\theta_1)+E(\theta_2)$$

and variance

$$V(\theta_1+\theta_2) = V(\theta_1)+V(\theta_2).$$

In most conventional schemes of statistical inference this theorem enables us to find the 'error' of the sum of two unknowns, given their separate 'errors'. Thus if we have made independent estimates of the lengths of two sticks, we may apparently find an estimate of the total length of the two. In denying the validity of the standard approach (though not, of course, of the above theorem in *probability*) are we preventing ourselves from taking such a step?

Let the support for θ_1 be

$$S_1(\theta_1) = S_1(\hat\theta_1) - (\theta_1-\hat\theta_1)^2/2w_1^2$$

and independently for θ_2 be

$$S_2(\theta_2) = S_2(\hat\theta_2) - (\theta_2-\hat\theta_2)^2/2w_2^2.$$

Now S_1 is what the support for the mean θ_1 of a Normal distribution of known variance w_1^2 would be, given a single observation $\hat\theta_1$. S_2 may be similarly interpreted. If $\hat\theta_1$ is $N(\theta_1, w_1^2)$ and $\hat\theta_2$ is $N(\theta_2, w_2^2)$ then (by the above theorem, in fact) $\hat\theta_1+\hat\theta_2$ is $N(\theta_1+\theta_2, w_1^2+w_2^2)$, and the support for $\theta_1+\theta_2$ must be

$$S_3(\theta_1+\theta_2) = S_3(\hat\theta_1+\hat\theta_2) - \frac{\{(\theta_1+\theta_2)-(\hat\theta_1+\hat\theta_2)\}^2}{2(w_1^2+w_2^2)},$$

having added the quantity $S_3(\hat\theta_1+\hat\theta_2)$ simply to make the support zero at the evaluate. We have now obtained the following theorem:

THEOREM 5.4.3

If the evaluate of θ_1 is $\hat\theta_1$ with formation w_1^2, and, independently, of θ_2 is $\hat\theta_2$ with formation w_2^2, and if the support functions are quadratic, the evaluate of $\theta_1+\theta_2$ is $\hat\theta_1+\hat\theta_2$ with formation $w_1^2+w_2^2$.

The derivation of this theorem relies on the fact that the sum of two Normal variates is also a Normal variate. It is not valid unless the support surfaces are quadratic, though it may be a valuable aid to interpretation when they are approximately so. It may, of course, be extended to any number of parameters. The extension to non-independent parameters is given in the next chapter (theorem 6.2.2).

5.5. ANALYTIC MAXIMIZATION: THE METHOD OF MAXIMUM LIKELIHOOD

Our third task was to find the evaluate when graphical methods are impossible. As in the last section, we shall deal with one parameter only, in order to establish the principles, even though in this case the support curve may always be drawn. The generalization to many parameters comes later.

Given the support function $S(\theta)$, it is open to us to find the evaluator of the parameter θ analytically by solving the equation $dS/d\theta = 0$. We did this in example 5.2.1 for a binomial parameter. Maximization of the support or likelihood function was placed on a sound footing as a method of *estimation* by Fisher in 1922, under the name of the *Method of Maximum Likelihood.* The Method of Support renders the concept of statistical estimation in scientific inference obsolete, but it is our good fortune that Fisher's maximum-likelihood method achieved widespread popularity because of its properties in the theory of estimation (which need not now concern us), so that its mathematical basis has been extensively studied. The present mathematical development is, therefore, standard, though its logical application is not. It would, however, be churlish to speak of the 'Method of Maximum Support' in our usage, particularly as Fisher was also largely responsible for the elaboration of likelihood as a measure in its own right. He writes 'The Method of Maximum Likelihood is indeed much used and widely appreciated in the statistical literature, without, I fancy, so much appreciation of the significance of the system of likelihood values at other possible values of the parameter.'[5] I shall continue to write of 'Maximum Likelihood', and we may recall that the method has its origins in the work which Daniel Bernoulli[6] published in 1777.

DEFINITION

The equation obtained by setting the score equal to zero is the *support equation.* When an explicit solution for the evaluate of the parameter is possible it is, in its algebraic form, known as the *evaluator* of the parameter.

EXAMPLE 5.5.1

Continuing the binomial example, the support equation is

$$\frac{dS}{dp} = \frac{a}{p} - \frac{b}{(1-p)} = 0,$$

and the evaluator of p thus $a/(a+b)$.

Frequently in practice it will not be possible to solve the support equation explicitly, and it will therefore be necessary to treat each individual case numerically.

EXAMPLE 5.5.2

The gamma-distribution provides a case in which there exists a minimal-sufficient statistic which is a single number, but it is not the observed arithmetic mean, nor can the evaluate be expressed explicitly in terms of it. In example 2.3.2 we saw that the geometric mean of the observations was sufficient for the parameter μ, itself the arithmetic mean of the distribution. Continuing with the same notation, the score is

$$\frac{dS}{d\mu} = \ln \prod_{i=1}^{n} x_i - n \frac{d}{d\mu} \ln (\mu - 1)! \qquad (5.5.1)$$

and hence the evaluate is the solution of

$$\frac{d}{d\mu} \ln (\mu - 1)! = \ln \tilde{x},$$

where $\tilde{x}$ is the geometric mean of the sample. The function of μ on the left is the digamma function, of which tables exist. The reader who is alert enough to ask 'What happens if any member of the sample is zero?' should (until chapter 8) console himself with the thought that the probability of this occurrence is *infinitesimal.*

The numerical solution of a support equation may be conveniently handled by Newton–Raphson iteration in most cases. Sometimes other methods will be more suitable, but since the problem of the numerical solution of an equation, or, what amounts to the same thing, the location of the maximum of a function, is extensively treated in numerous texts, I will limit detailed discussion to the Newton–Raphson method, indicating its limitations.

5.6. NEWTON–RAPHSON ITERATION

In passing, we note that a solution to $S'(\theta) = 0$ may be obtained graphically by plotting the score $dS/d\theta$ against θ and observing where the curve intersects the θ-axis. This simple method, however, does not generalize to many parameters.

Suppose, rather, that we make an initial guess θ' at the evaluate. Let $T(\theta) = dS/d\theta$ be the score at θ. Then, by Taylor's theorem,

$$T(\hat{\theta}) = 0 = T(\theta') + (\hat{\theta} - \theta')\frac{dT}{d\theta} + \ldots,$$

where $\mathrm{d}T/\mathrm{d}\theta = \mathrm{d}^2S/\mathrm{d}\theta^2$ is minus the information at θ'. Solving this equation involving only the first two terms of the Taylor series, we obtain an approximate value for $\hat{\theta}$, say θ'':

$$\theta'' = \theta' - T(\theta')\Big/\frac{\mathrm{d}T}{\mathrm{d}\theta} = \theta' - \frac{\mathrm{d}S}{\mathrm{d}\theta}\Big/\frac{\mathrm{d}^2S}{\mathrm{d}\theta^2}, \tag{5.6.1}$$

where the differentials are taken at $\theta = \theta'$. In words, a corrected value is obtained by adding to the first value the score divided by the information, both taken at the first value. Iterating according to this formula will lead, under suitable conditions, to the evaluate $\hat{\theta}$. The correction may also be thought of as the score multiplied by the formation, taken at the trial value.

The Newton–Raphson method has a direct geometrical interpretation. It amounts to fitting a parabola to the support curve at the trial value, with axis vertical and having the same gradient and curvature as the curve at that point, and then proceeding to the value of the parameter for which the parabola has its maximum. It follows that the closer the support curve is to a parabolic shape, the faster will be the rate of convergence of the Newton–Raphson process; and if the curve is a true parabola, the first iterate is exactly the evaluate. In that event, of course, analytic maximization would be possible, but it is of interest to note that the method is at its best when the use of evaluates is most justified. The method may also be viewed on the graph of the score against the parameter (see example 5.6.2 and figure 15), where each iterate is the point at which the tangent to the curve at the preceding iterate intersects the axis. If the support curve is a true parabola, then of course the score curve is a straight line, and convergence to the evaluate is immediate.

Most of the common single-parameter distributions, both continuous and discrete, have support equations which are easily solved. For a wide class of distributions, solution of the support equation is equivalent to equating the observed and expected means, the observed mean being a sufficient statistic. This is not true of all distributions (the gamma-distribution is an exception – example 5.5.2), and when it is true an explicit solution does not necessarily follow (see example 5.6.2 on Fisher's Logarithmic Series distribution). As an elementary example of Newton–Raphson iteration I shall therefore use the binomial distribution, even though it admits an explicit solution.

EXAMPLE 5.6.1

We have seen (example 5.2.1) that the score for the binomial distribution is

$$\frac{\mathrm{d}S}{\mathrm{d}p} = \frac{a}{p} - \frac{b}{(1-p)},$$

and that the information is

$$-\frac{\mathrm{d}^2S}{\mathrm{d}p^2} = \frac{a}{p^2} + \frac{b}{(1-p)^2}.$$

Suppose, as before, $a = 33$ and $b = 47$, and that our trial value for p is 0.5000. At this value, the score is -28, the information 320, and hence the correction to p is $-\frac{28}{320} = -0.0875$ exactly. The revised value is thus 0.4125, which is, as we have seen, the actual evaluate. In this instance Newton–Raphson iteration has led to the exact solution of the support equation in one step. This is most unusual, and always arises in connection with the binomial if the trial value for p is $\frac{1}{2}$, as may be verified algebraically. If, instead, we take 0.2500 as the trial value, the score is $\frac{208}{3}$ and the information $\frac{5504}{9}$, whence the correction to p is $\frac{39}{344} = 0.1134$. The corrected value for p is thus 0.3634. If this value, which is not very close to what we know to be the evaluate, is used to prime a further iteration, a new value $p = 0.4098$ is obtained. The third iterate is 0.4125, which is satisfactory.

Provided the support function is everywhere twice-differentiable, the major source of upsets to the Newton–Raphson method is the proximity of points of inflection. At a point of inflection the information is zero, and the correction therefore infinite. Near a point of inflection the information may be so small, and the correction so large, that the 'improved' value for the parameter may be wildly out – even outside the permitted range. Beyond points of inflection (that is, in regions of positive curvature), the Newton–Raphson method will in fact lead away from the maximum towards a minimum of the support curve. These possibilities are illustrated in example 5.6.2, and figure 15.

I have already suggested that, in order to calculate support limits, it may be worthwhile to transform to a new parameter on which the support curve is more nearly parabolic. It may also be advisable, and possibly necessary, to do this in order to achieve convergence of the Newton–Raphson process, as in the following example.

where $dT/d\theta = d^2S/d\theta^2$ is minus the information at θ'. Solving this equation involving only the first two terms of the Taylor series, we obtain an approximate value for $\hat{\theta}$, say θ'':

$$\theta'' = \theta' - T(\theta') \bigg/ \frac{dT}{d\theta} = \theta' - \frac{dS}{d\theta} \bigg/ \frac{d^2S}{d\theta^2}, \qquad (5.6.1)$$

where the differentials are taken at $\theta = \theta'$. In words, a corrected value is obtained by adding to the first value the score divided by the information, both taken at the first value. Iterating according to this formula will lead, under suitable conditions, to the evaluate $\hat{\theta}$. The correction may also be thought of as the score multiplied by the formation, taken at the trial value.

The Newton–Raphson method has a direct geometrical interpretation. It amounts to fitting a parabola to the support curve at the trial value, with axis vertical and having the same gradient and curvature as the curve at that point, and then proceeding to the value of the parameter for which the parabola has its maximum. It follows that the closer the support curve is to a parabolic shape, the faster will be the rate of convergence of the Newton–Raphson process; and if the curve is a true parabola, the first iterate is exactly the evaluate. In that event, of course, analytic maximization would be possible, but it is of interest to note that the method is at its best when the use of evaluates is most justified. The method may also be viewed on the graph of the score against the parameter (see example 5.6.2 and figure 15), where each iterate is the point at which the tangent to the curve at the preceding iterate intersects the axis. If the support curve is a true parabola, then of course the score curve is a straight line, and convergence to the evaluate is immediate.

Most of the common single-parameter distributions, both continuous and discrete, have support equations which are easily solved. For a wide class of distributions, solution of the support equation is equivalent to equating the observed and expected means, the observed mean being a sufficient statistic. This is not true of all distributions (the gamma-distribution is an exception – example 5.5.2), and when it is true an explicit solution does not necessarily follow (see example 5.6.2 on Fisher's Logarithmic Series distribution). As an elementary example of Newton–Raphson iteration I shall therefore use the binomial distribution, even though it admits an explicit solution.

EXAMPLE 5.6.1

We have seen (example 5.2.1) that the score for the binomial distribution is

$$\frac{dS}{dp} = \frac{a}{p} - \frac{b}{(1-p)},$$

and that the information is

$$-\frac{d^2S}{dp^2} = \frac{a}{p^2} + \frac{b}{(1-p)^2}.$$

Suppose, as before, $a = 33$ and $b = 47$, and that our trial value for p is 0.5000. At this value, the score is -28, the information 320, and hence the correction to p is $-\frac{28}{320} = -0.0875$ exactly. The revised value is thus 0.4125, which is, as we have seen, the actual evaluate. In this instance Newton–Raphson iteration has led to the exact solution of the support equation in one step. This is most unusual, and always arises in connection with the binomial if the trial value for p is $\frac{1}{2}$, as may be verified algebraically. If, instead, we take 0.2500 as the trial value, the score is $\frac{208}{3}$ and the information $\frac{5504}{9}$, whence the correction to p is $\frac{39}{344} = 0.1134$. The corrected value for p is thus 0.3634. If this value, which is not very close to what we know to be the evaluate, is used to prime a further iteration, a new value $p = 0.4098$ is obtained. The third iterate is 0.4125, which is satisfactory.

Provided the support function is everywhere twice-differentiable, the major source of upsets to the Newton–Raphson method is the proximity of points of inflection. At a point of inflection the information is zero, and the correction therefore infinite. Near a point of inflection the information may be so small, and the correction so large, that the 'improved' value for the parameter may be wildly out – even outside the permitted range. Beyond points of inflection (that is, in regions of positive curvature), the Newton–Raphson method will in fact lead away from the maximum towards a minimum of the support curve. These possibilities are illustrated in example 5.6.2, and figure 15.

I have already suggested that, in order to calculate support limits, it may be worthwhile to transform to a new parameter on which the support curve is more nearly parabolic. It may also be advisable, and possibly necessary, to do this in order to achieve convergence of the Newton–Raphson process, as in the following example.

EXAMPLE 5.6.2

Fisher's *Logarithmic Series* distribution is a discrete distribution for a random variable r $(1 \leqslant r \leqslant \infty)$ with probability density function

$$P(r) = \frac{\theta^r}{-r \ln(1-\theta)} \quad (0 < \theta < 1).$$

The mean of the distribution is $\theta/\{-(1-\theta)\ln(1-\theta)\}$. Let a_r be the observed frequency in the rth class, and $\bar{r}$ the observed mean. Σ will signify summation over $r = 1, 2, \ldots$; $\Sigma a_r = n$. The support function is

$$S(\theta) = \sum a_r \ln\left(\frac{\theta^r}{-r\ln(1-\theta)}\right)$$
$$= \sum a_r[r \ln\theta - \ln r - \ln\{-\ln(1-\theta)\}]. \qquad (5.6.2)$$

The support equation is

$$\frac{dS}{d\theta} = \frac{\Sigma r a_r}{\theta} + \frac{\Sigma a_r}{(1-\theta)\ln(1-\theta)} = 0, \qquad (5.6.3)$$

and may be written

$$\frac{\theta}{-(1-\theta)\ln(1-\theta)} = \frac{\Sigma r a_r}{\Sigma a_r} = \bar{r}, \qquad (5.6.4)$$

indicating that the evaluate for the parameter θ is to be found by equating the observed and expected means. $\bar{r}$ is a sufficient statistic; suppose that in a particular case it had the value 5.940, the sample size being $n = 50$. The support equation does not admit an explicit solution, so that iteration must be used. The information is readily found, by differentiating the score and changing the sign, but Newton–Raphson iteration fails, unless one is extremely lucky in the choice of a trial value for θ, because the corrected value is very likely to fall outside the permitted range for θ. That this must be so is obvious on an examination of the support curve (figure 14). A transformation is called for which will stretch out the steeply-turning part of the curve near $\theta = 1$, and the following suggests itself:

$$\phi = \frac{\theta}{(1-\theta)}, \quad \theta = \frac{\phi}{(1+\phi)} \quad (0 < \phi < \infty)$$

The tip of the support curve for ϕ, within five units of support of the maximum, is shown in figure 15. The transformation has evidently succeeded rather too well in its object.

The support equation for ϕ is found to be

$$\frac{dS}{d\phi} = n\left(\frac{\bar{r}}{\phi(1+\phi)} - \frac{1}{(1+\phi)\ln(1+\phi)}\right) = 0, \qquad (5.6.5)$$

which, as an equation, may be written

$$\phi = \bar{r}\ln(1+\phi).$$

It is interesting to note that in this form the equation is ripe for immediate iteration, a trial value of ϕ being inserted in the right-hand side to give

a corrected value. It is always worthwhile keeping an eye open for such possibilities, but for the purpose of the present example we may continue in the standard way:

$$-\frac{d^2S}{d\phi^2} = n\left(\bar{r}\frac{(1+2\phi)}{[\phi(1+\phi)]^2} - \frac{1+\ln(1+\phi)}{[(1+\phi)\ln(1+\phi)]^2}\right). \qquad (5.6.6)$$

Starting with a trial value $\phi = 10$, the successive iterates, together with their scores and informations, are given in table 2. The solution is $\phi = 17.2511$, correct to four places of decimals, at which point the information is 27.1118. The approximate 2-unit support limits are therefore 6.8373 and 27.6649, but from figure 15 we see that the actual limits are nearer 10.0 and 34.0, the difference being accounted for by the poorness of the quadratic approximation. The evaluate and the actual support limits may be transformed back into values of θ, giving $\hat{\theta} =$ 0.9452 (0.909, 0.971).

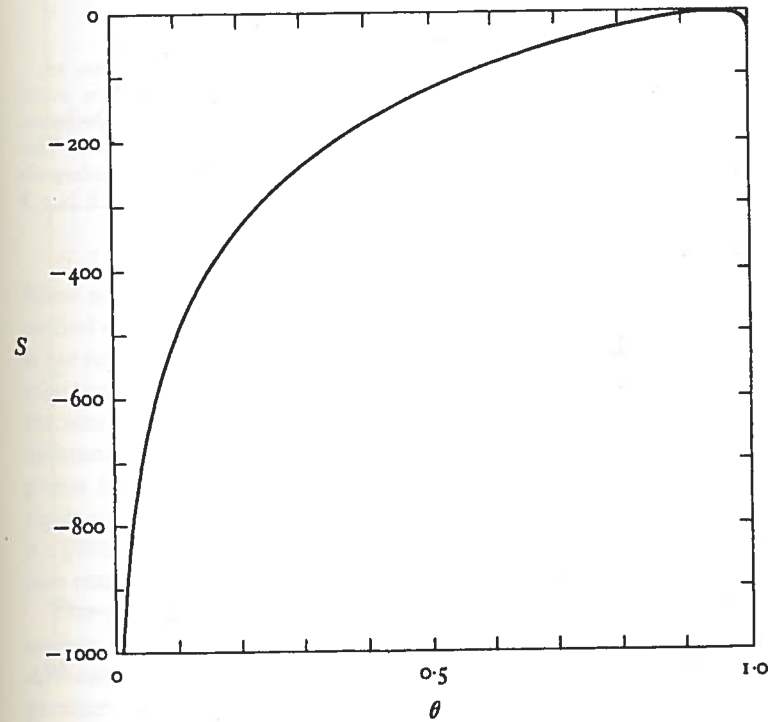

Figure 14. The support curve for the parameter θ of Fisher's logarithmic series distribution (example 5.6.2). Note that the scale of support is from 0 to −1000 rather than from 0 to −5 as in other figures.

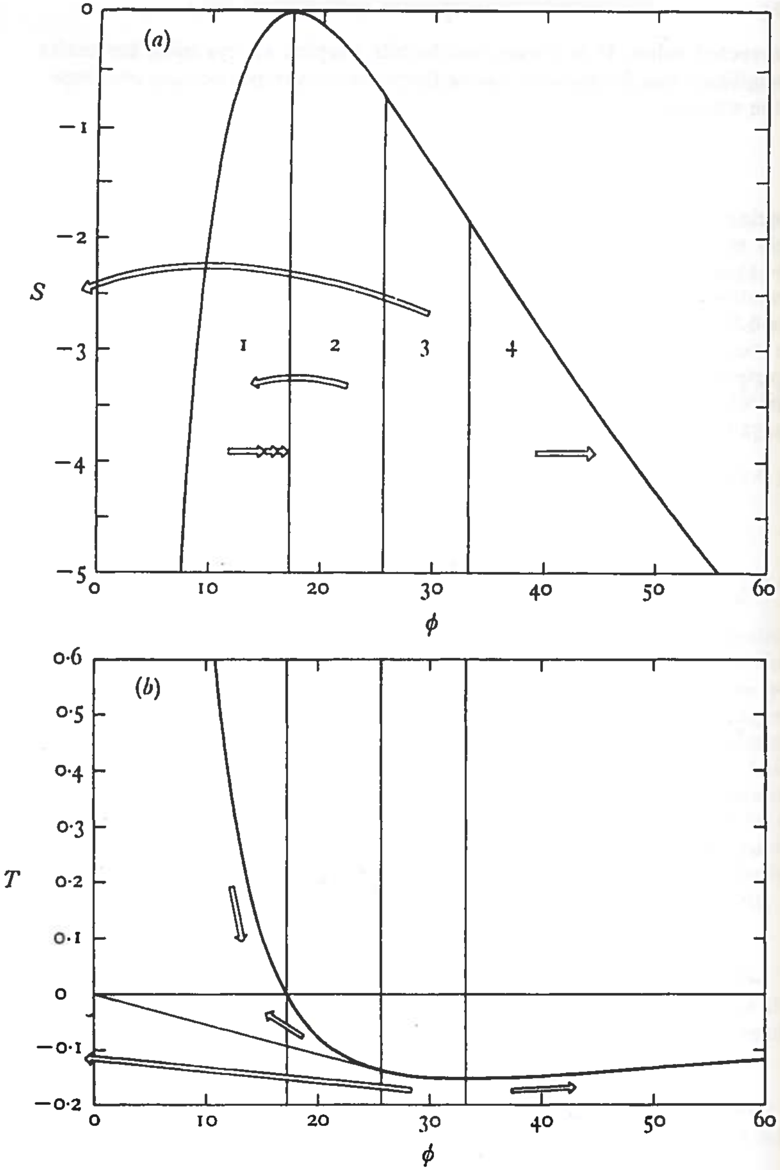

Figure 15. (*a*) Part of the support curve of figure 14 near the maximum, following the parameter transformation $\phi = \theta/(1-\theta)$. (*b*) The score curve for ϕ. The numbers indicate regions of convergence and divergence under Newton–Raphson iteration: (1) region of monotonic convergence to the maximum; (2) from this region the succeeding iterate is in (1); (3) from

TABLE 2. Iterates, scores and informations for the logarithmic distribution with $n = 50$ and $\bar{r} = 5.94$.

Iteration	Parameter ϕ	Score	Information
—	10.0000	0.804 398	0.271 261
1	12.9654	0.282 353	0.109 851
2	15.5357	0.078 329	0.055 921
3	16.9364	0.012 050	0.039 749
4	17.2396	0.000 425	0.036 985
5	17.2511	0.000 001	0.036 884
6	17.2511	—	—

As well as giving the support curve for ϕ, figure 15 gives the score curve and indicates the zone of convergence under Newton–Raphson iteration, illustrating how, from some starting values, the first iterate may be further still from the maximum, or outside the permitted range altogether. The zone of convergence is $0 < \phi < 25.5910$, or, in terms of θ, $0 < \theta < 0.9624$; the transformation has evidently been very successful.

5.7. GENERAL COMMENTS ON ITERATIVE METHODS

Since nowadays most iterative solutions to support equations are carried out on a computer, the rate of convergence to the evaluate is not normally an important matter. The most general programs in existence rely on numerical differentiation for the calculation of the score and information, thus obviating the need for the user to differentiate analytically. They are therefore slower than programs tailor-made for specific problems, which incorporate the algebraic forms for the score and information. A useful compromise is a general program into which the algebraic forms for a particular case can be inserted as subroutines.

The standard Newton–Raphson method described in the last section has two important variants. In one, known as the *fixed-derivative* method, the information evaluated at the first trial parameter value is used also in the remaining iterations, thus obviating its fresh calculation each time. This has little to commend it, and may even lead to cycling round the maximum.[7] The other variant, known as Fisher's *scoring for parameters* method,[8] replaces the observed information at each parameter value by what the expected information would be if that value were the true one (section 7.2). The motives for such a substitution are, first, that with discrete distributions it leads to a particularly simple way of performing the iterative calculations manually, and, secondly, that under the standard theory the new quantity is used to derive the approximate sampling variance of the maximum-likelihood estimate. But with the use of computers and the adoption of the Method of Support, these justifications are no longer relevant, and in some cases the method is less reliable than the standard method.

It is possible to invent further variants of the Newton–Raphson approach, which might be better for particular applications. The standard method involves solving for the three coefficients of a second-degree curve by equating its ordinate, gradient, and curvature, to the actual values for the support curve at the trial parameter value. But the same procedure could be applied to any approximating curve with three coefficients, and not merely to a second-degree polynomial.

EXAMPLE 5.7.1

Suppose it is felt that a circle would provide a better approximation than a parabola. Let the inclination of the support curve at the trial value θ' be α, and the radius of curvature ρ. Then we have

$$\tan \alpha = \frac{dS}{d\theta}$$

and

$$-\rho = \left\{1 + \left(\frac{dS}{d\theta}\right)^2\right\}^{\frac{3}{2}} \Big/ \frac{d^2S}{d\theta^2},$$

where the score and information are taken at $\theta = \theta'$. The adjusted value, θ'', is found by adding to θ' the correction $\rho \sin \alpha$ (figure 16). Substituting for α and ρ, the correction becomes

$$\theta'' - \theta' = -\frac{dS}{d\theta}\left\{1 + \left(\frac{dS}{d\theta}\right)^2\right\} \Big/ \frac{d^2S}{d\theta^2}, \qquad (5.7.1)$$

which differs from the usual Newton–Raphson correction by the factor

$$1 + \left(\frac{dS}{d\theta}\right)^2.$$

It is likely, however, that similar improvements can be obtained more easily by a parameter transformation which renders the

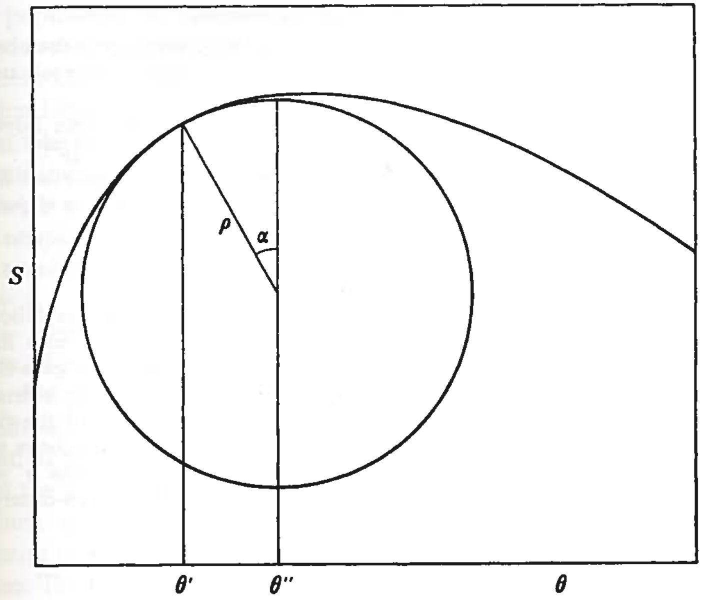

Figure 16. Approximating a curve by the osculating circle at a trial value θ' of the parameter θ.

support curve more nearly parabolic, as was done in example 5.6.2.

Amongst other iterative methods which are sometimes used is the *method of false position*, which approximates the score curve by a straight line joining two points, and finds the point at which the line crosses the axis $T = 0$. Naturally, it is at its best if the initial two points straddle the axis. A further method has already been indicated in example 5.6.2, and relies on throwing the support equation into the form $\theta = f(\theta)$. Finally, by means of an example, we shall illustrate the *counting method*,[9] which is applicable in any discrete case where an exact solution would be possible except for the fact that some of the classes are indistinguishable and therefore grouped together. It has been shown to lead to the maximum of the likelihood, and is of particular use in genetics. Unfortunately it was unknown in the early days of genetics, when much effort was expended on devising approximate solutions to likelihood equations, and now that computers are available the need for it is not so great.

EXAMPLE 5.7.2

Out of 100 families with three children, 48 have more girls than boys, and 52 more boys than girls. On a binomial model, what is the evaluate of p, the probability of a birth being male?

The expected proportion of families with more girls than boys is $(1-p)^3 + 3p(1-p)^2$, and with more boys than girls is $p^3 + 3p^2(1-p)$. With only one degree of freedom and a single parameter, an exact fit is possible, the evaluate of p being the solution in the interval $0 \leqslant p \leqslant 1$ of the equation

$$p^3 + 3p^2(1-p) = 0.52.$$

In order to solve this, we may use the counting method, as follows. Taking p' as a trial value, we divide the observed classes into their components according to the expected ratios, the class 'more girls than boys' being divided into families with no boys and families with one boy in the ratio $(1-p')^3 : 3p'(1-p')^2$, or $1-p' : 3p'$, and the class 'more boys than girls' being divided into families with three boys and families with two boys in the ratio $p' : 3(1-p')$. If we take $p' = \frac{1}{2}$ initially, the divisions will both be $1:3$, leading to the pseudo-observations

0 boys	12 families
1	36
2	39
3	13

Were these actual observations from a binomial distribution the evaluate of p would be given by the proportion of boys in the sample, $\frac{153}{300}$, and we use this as an improved value for p : $p'' = 0.51$. The process may now be repeated. The new divisions are $0.49 : 1.53$ and $0.51 : 1.47$, giving for the second round of pseudo-observations

0 boys	11.64 families
1	36.36
2	38.61
3	13.39

The resulting approximation to the evaluate is 0.5125, and succeeding iterates are 0.5131 and 0.5133, which is correct to four places of decimals.

5.8. INTEGER SOLUTIONS AND SOLUTION BY SIMULATION

It may sometimes happen that the parameter under estimation may only take discrete values. The situation is then formally identical to having a number of discrete hypotheses under consideration, but it may be simpler to think in terms of a continuous parameter in the first instance. As an example, let us consider the determination of the frequency of the Rh$-$ blood-group gene in

the population of Cambridge (example 3.4.4). As mentioned in chapter 1, the gene frequency, q, is in principle a discrete variable. If there are 100 000 people in Cambridge, there are 200 000 genes at this locus in the population, and the proportion that is Rh$-$ must take one of the 200 001 possible discrete values. In this case, a continuous approximation will find wide acceptance. But suppose we were to ask the question: How many Rh$-$ genes are there in the *sample*? If the sample consists of r Rh$-$ individuals out of n, the number of Rh$-$ genes must be one of the numbers $2r$, $2r+1$, $2r+2, \ldots 2r+(n-r)$, depending on whether the $(n-r)$ Rh+ phenotypes include none, $1, 2, \ldots$ or $(n-r)$ heterozygotes. If r and n are small, the continuous approximation may be regarded as unacceptable. Each one of the possibilities will have a likelihood, calculable by considering the probability of the arrangement. It may be necessary, depending on the model adopted, to treat the genes of maternal and paternal origin separately.

Such questions are becoming common in modern genetics because it is frequently possible to 'sample' an entire small population. The only meaningful question then involves the numbers of genes of each kind in the sample.

On occasion it may prove impossible to write down the likelihood owing to the complexity of the model. If it is practicable to simulate results in such a situation, having adopted a trial value for the parameter, such results may be compared with the actual data, and the parameter adjusted until the simulated and actual results are as close as possible. This proposal raises a number of interesting questions, such as how to formulate rules for determining when to stop a particular simulation and start another with a different parameter value. However, it seems unlikely to be required in single-parameter situations.

5.9. HISTORICAL NOTE ON POINT ESTIMATION

In section 3.1 I quoted Daniel Bernoulli as being the first person to use likelihood as a criterion for choosing the 'best' value of a parameter. I agree with Hacking[10] that the interpretation to be put on Bernoulli's writing is that he was simply choosing the 'best-supported' value, rather than that he was using the method because he thought it would provide the 'best' estimate in some other sense.

The circumstance that the Method of Maximum Likelihood is analytically identical to the method of inverse probability if a uniform prior distribution is adopted has obscured the origins of the former as a method of point estimation in its own right. Gauss originally developed his theory of least squares using inverse probability and a uniform prior, but later preferred a 'loss function' approach, arguing that a quadratic loss function, though arbitrary, was the simplest sensible one.[11] Laplace, of course, maximized the posterior probability. Haldane[12] claimed some priority for Karl Pearson in the matter, but in fact Pearson seems to have been doing no more than Laplace. In a paragraph headed 'On the best value of the correlation coefficient'[13] he wrote 'Thus, it appears that the observed result is the most probable, when r is given the value $S(xy)/(n\sigma_1\sigma_2)$. This value presents no practical difficulty in calculation, and therefore we shall adopt it.' No other justification is offered, but it is quite clear from the following paragraph, in which Pearson uses inverse probability to obtain the distribution of the parameter given the sample, that he is simply following Laplace's procedure. The same may be said of the subsequent work of Pearson and Filon,[14] in which the second differentials of the posterior probability are used to obtain, approximately, the variances and covariances of estimates, a procedure which analytically, though not logically, closely foreshadows the maximum-likelihood approach. Haldane also quotes Edgeworth as being one of the forerunners of Fisher in the use of maximum likelihood, but here there can be no argument: Edgeworth[15] was quite explicitly using inverse probability: 'I submit that very generally we are justified in assuming an equal distribution of *a priori* probabilities over that tract of the measurable with which we are concerned.' He quotes Gauss and Laplace.

The vital break with inverse probability seems to have been Fisher's alone. He advocated what he later called the Method of Maximum Likelihood in his very first paper,[16] as a means of point estimation. The break, though clear in retrospect, was not quite clean: having not yet adopted the word 'likelihood', he wrote of 'inverse probability', a fact he regretted in his 1922 paper. But he was clear that these 'probabilities' were only relative, and he specifically stated that it was 'illegitimate' to integrate them with respect to a parameter. He noted the inconsistency that would follow from parameter transformation if the differential elements were included in the likelihood, and wrote

We have now obtained an absolute criterion for finding the relative probabilities of different sets of values for the elements of a probability system of known form. It would now seem natural to obtain an expression for the probability that the true values of the elements should lie within any given range. Unfortunately we cannot do so . . . P is a relative probability only, suitable to compare point with point, but incapable of being interpreted as a probability distribution, or of giving any estimate of absolute probability.

In this first paper Fisher does not justify his 'absolute criterion'; he may have simply felt that it was intuitively reasonable. Subsequently he espoused likelihood as a measure of relative belief *and* advocated maximum likelihood as a means of point estimation justified by the repeated-sampling properties of the estimators. At first sight there may seem to be some inconsistency here: did he, or did he not, believe in the relevance of repeated-sampling characteristics?

I think the answer is to be found in the historical development of the non-Bayesian theory of estimation. Until Fisher's 1922 paper[17] the problem had never been clearly put. The method of least squares had, at least in astronomy, satisfied the demand for some criterion (albeit arbitrary) by which to choose estimators, and the method of moments had likewise offered a practical procedure.

Fisher's first great contribution to estimation was a careful specification of the problem, and his second was the development of criteria 'without reference to extraneous or ulterior considerations'[18] by which to judge estimators. The first point is so obvious to us that we tend to forget what an important advance it was at the time; and we seem to have forgotten about the second point altogether. For the subsequent development of estimation theory has been in terms of externally-imposed criteria, such as minimum-variance unbiassedness or optimum confidence sets, which are at best arbitrary and at worst comic in their effects. By contrast, Fisher's theory proceeds from the most general considerations; he observes that we require consistent estimators, that in large samples estimates will be Normally distributed, and thus that only their variance presents itself as a criterion. He shows how this variance cannot be less than a certain quantity, whose reciprocal he calls the information, and how of all the methods of estimation based on linear functions of the observations, the Method of Maximum Likelihood provides estimators which achieve this lower limit. Then, observing that the information is also defined for small samples, and that it has precisely those qualities we might expect a measure of information to possess, he finds that in some cases estimators are sufficient, preserving all the information, and that the Method of Maximum Likelihood leads to them where they exist. Where they do not exist, he then shows that the residual information not conveyed by the maximum of the likelihood is contained in other aspects of the likelihood curve.

Fisher has given us a descriptive account of these developments[19] which concludes: 'Thus, basing our theory entirely on considerations independent of the possible relevance of mathematical likelihood to inductive inferences in problems of estimation, we seem inevitably led to recognize in this quantity the medium by which all such information as we possess may be appropriately conveyed.' Basing his researches on the concept of repeated sampling, he is led inexorably to the likelihood function, that very function which 'supplies a natural order of preference among the possibilities under consideration'.[20] Having thus so strongly reinforced his intuition, in later writing he is inclined to let estimation theory play second fiddle in the arguments for the use of likelihood.[21]

I think it may fairly be said, therefore, that the inconsistency in Fisher's two approaches is more apparent than real, for in his hands the repeated-sampling approach led to precisely the same conclusion as the more intuitive direct likelihood approach. The gulf between the two came later, when others tried to impose external criteria on estimators, and judged maximum-likelihood estimators by such criteria, an exercise which is proving one of the largest red herrings in modern mathematics.

There is a very revealing comment by Fisher[22] to a paper of Jeffreys, published in 1938:

> Dr Jeffreys says that I am entitled to use maximum likelihood as a primitive postulate. In this I believe he is right. A worker with more intuitive insight than I might perhaps have recognized that likelihood must play in inductive reasoning a part analogous to that of probability in deductive problems. My own procedure has been more pedestrian.

Here is Fisher, at a time when his theory of estimation was at its zenith, wistfully suggesting that it might be better to regard likelihood as the more primitive concept, a position towards which he later moved.

Even today, thirty-five years after Fisher drew attention to the importance of the *whole* likelihood function in estimation, it is difficult to convey to a statistical audience the vital distinction between likelihood regarded as a basis for a theory of inference, and likelihood regarded as a commodity to be maximized in a method of point estimation. At one recent international conference at which I laboured for three-quarters of an hour to make clear the advantages of likelihood inference, the chairman thanked me for my lecture on the Method of Maximum Likelihood. The phrase 'Method of Support' has, indeed, been coined in order to emphasize the distinction.

Following Fisher, Barnard[23] has been almost the sole custodian of the likelihood function amongst statisticians, but one suspects that it has been flourishing independently in other fields. Thus: 'In radar problems, fortunately, it is generally possible to present $p_x(y)$ [the likelihood function of x] for all values of x and the question of point estimation need not arise.'[24]

I conclude this chapter with an extract from Fisher's 1935 paper,[25] in which he tells us quite clearly what to do next:

> To those who wish to explore for themselves how far the ideas so far developed on this subject will carry us, two types of problem may be suggested. First, how to utilize the whole of the information available in the likelihood function. Only two classes of cases have yet been solved. (*a*) Sufficient statistics, where the whole course of the function is determined by the value which maximizes it, and where consequently all the available information is contained in the maximum likelihood estimate, without the need of ancillary statistics. (*b*) In a second case, also of common occurrence, where there is no sufficient estimate, the whole of the ancillary information may be recognized in a set of simple relationships among the sample values, which I have called the configuration of the sample. With these two special cases as guides the treatment of the general problem might be judged, as far as one can judge these things, to be ripe for solution.
>
> Problems of the second class concern simultaneous estimation, and seem to me to turn on how we should classify and recognize the various special relationships which may exist among parameters estimated simultaneously.

In the Method of Support we solve the first problem by looking at the log-likelihood function itself, only resorting to evaluates in regular situations; the second problem we consider in the next chapter.

5.10. SUMMARY

The Method of Maximum Likelihood is introduced from the point of view of support. For the case of a single parameter, methods are given for summarizing the support curve near its maximum in terms of the evaluate and its span, for combining such values from different sets of data, and for finding approximations to the evaluate when an iterative solution is necessary. The Newton–Raphson method of iteration is discussed in detail, and examples are given both of its use and of cases in which it is inapplicable, or applicable only with modification. Formulae are given for the span of the evaluate following parameter transformation. The problems of integer solutions, and solution by simulation, are touched upon. In conclusion, an outline is given of the reasons which led Fisher to recognize the importance of the likelihood function from a repeated-sampling point of view.

CHAPTER 6

THE METHOD OF SUPPORT FOR SEVERAL PARAMETERS

6.1. INTRODUCTION

In chapter 5 several analytical methods for dealing with a single parameter were considered, but since in that case the support curve may always be drawn, they only become essential when there is more than one parameter. In this chapter, therefore, the methods of chapter 5 will be extended to the case of several parameters, and I shall then consider some of the short-cuts that can be employed, and difficulties that may be encountered. Many of the comments of the last chapter are also apposite to the multiparameter case, by analogy. Where the extension is obvious, they will not be repeated.

The fact that we are unable to appreciate a multiparameter *support surface* directly means that we will have to rely heavily on the Method of Maximum Likelihood. It follows that support surfaces which are not even approximately quadratic will be peculiarly difficult to handle; some cases will be presented in chapter 8. This difficulty is, of course, a reflection on the techniques which are available to us for the interpretation of multidimensional surfaces, rather than on the Method of Support itself. We will have to do the best we can.

6.2. INTERPRETATION OF EVALUATES

In dealing with more than one parameter it is important to be clear, at the outset, about the interpretation of evaluates. As defined in the previous chapter, the evaluates of the parameters of a model are those for which the support is a maximum over all the parameters jointly. It will generally be an over-simplification, taking for example two parameters θ_1 and θ_2, to speak of the evaluate $\hat{\theta}_1$ and the evaluate $\hat{\theta}_2$ as though each had an independent existence. Rather, we must speak of the pair $(\hat{\theta}_1, \hat{\theta}_2)$. Only in special circumstances, treated below, may we make separate statements, the most important one being where the support function is quadratic.

By the Same Author

Foundations of Mathematical Genetics, 1977

Pascal's Arithmetical Triangle, 1987

LIKELIHOOD

Expanded Edition

A. W. F. EDWARDS, Sc.D.

Fellow of Gonville and Caius College

Reader in Biometry, University of Cambridge

THE JOHNS HOPKINS UNIVERSITY PRESS