Biometrika (1966), **53**, 1 *and* 2, *p.* 151
With 5 text-figures
Printed in Great Britain

Evaluation of the maximum-likelihood estimator where the likelihood equation has multiple roots

BY V. D. BARNETT
University of Birmingham

1. INTRODUCTION

Even under the usual regularity conditions, where a unique consistent root of the likelihood equation is known to exist, it is often not possible to obtain an explicit solution for the maximum-likelihood estimate (M.L.E.) of a parameter as a function of the sample. In such cases it is necessary to use numerical methods to evaluate the M.L.E. by successive iteration. This was first discussed in the statistical literature by Fisher (1925) who advocated the use of a method known as 'scoring for parameters'. He suggested that '...starting with an inefficient statistic, a single process of approximation will in ordinary cases give an efficient statistic differing from the maximum-likelihood solution, by a quantity which with increasing samples decreases as n^{-1}', where n is the sample size; and concluded that one iteration is therefore sufficient in a practical sense in such 'ordinary cases'. Norton (1956) has used a particular example in genetics described by Fisher (1950, chapter 9) to show that this is not true, but that several iterations may be necessary for reasonable convergence. A variety of numerical methods are available for locating the root of an equation, of which the method of 'scoring for parameters' is but a single example of the Newtonian approach to the problem. Kale (1961) has discussed several of these methods (the fixed-derivative Newton, Newton–Raphson and 'scoring for parameters' methods) for obtaining the M.L.E. of a single parameter under the usual regularity conditions, from the point of view of whether or not they satisfy certain desirable probabilistic properties as $n \to \infty$. He states: 'In effect, it is shown that the iteration processes usually applied in practice are justifiable, in large samples at least'. In a subsequent paper, Kale (1962) makes a similar study for the multi-parameter case. The relative merits of the various methods in terms of their order of convergence have been extensively discussed by the numerical analysts, together with the need to effect a compromise between the order of convergence and the practical effort required to apply the different methods (see e.g. Hamming, 1962).

In any practical problem, however, we are not necessarily concerned with relative orders of convergence, utility of application or asymptotic probabilistic properties of the different methods. Given a single sample of observations $x_1, x_2, \ldots, x_n$, of fixed finite size n, from a distribution with parameter θ, we wish to evaluate the M.L.E. of *θ for that sample*. Regularity conditions and the associated existence of a unique consistent root are no guarantee that a *single* root of the likelihood equation will exist for this sample. In fact there will often exist multiple roots, corresponding to multiple relative maxima of the likelihood function, even if the regularity conditions are satisfied. The results described above do not consider this effect specifically, either because (as in the case of Kale, 1961) the author is not basically concerned with finite samples, or (as Norton, 1956) particular examples discussed quite fortuitously have a unique root for the likelihood equation.

In general, then, we have a more fundamental problem of whether or not a particular

method will actually locate a root of the likelihood equation, rather than how well it will do so in terms of the various criteria described above. Further, we require a method which can be applied systematically to locate all roots of the likelihood equation, enabling us to choose as the M.L.E. that one which corresponds to the absolute maximum of the likelihood function. From this standpoint the various methods have quite distinct properties, and this paper is concerned with a study of five particular methods, which are commonly used, in an attempt to determine which of them are viable in the practical situation.

The methods discussed are the fixed-derivative Newton, 'scoring for parameters' and Newton–Raphson methods together with the method of false positions (*regula falsi*, see Whittaker & Robinson, 1944, p. 92) and a method used by Cohen (1957) in estimating the parameters of a truncated normal distribution. This is another standard method and is a special case of a method described by Whittaker & Robinson (1944, p. 81). In terms of order of convergence all these methods are of first order except the Newton–Raphson method which is second order. But if we are to ensure convergence when the likelihood equation has multiple roots we find that only one of these methods is suitable, namely the method of false positions. Its form also makes it a natural choice from the point of view of a systematic scanning of the whole likelihood function, which is necessary when this function has multiple relative maxima.

To illustrate the problems involved in estimating the M.L.E. where multiple roots of the likelihood equation exist, and to compare the five methods mentioned above, it is adequate to consider in detail a single example. The example chosen is that of the location parameter of the Cauchy distribution (which satisfies the regularity conditions), but any observations and conclusions about the shortcomings and relative merits of the different methods apply equally to any distribution whose likelihood function exhibits multiple relative maxima (or even points of inflexion or too sharp a maximum, see § 3).

The various effects described are best illustrated by considering particular random samples from the Cauchy distribution together with their corresponding likelihood functions. A very large number of such samples were simulated and analysed on a digital computer for sample sizes from 3 to 19. In the course of this study a great deal of quantitative information was obtained specifically concerning the location parameter of the Cauchy distribution. This information is of interest in its own right and is discussed in the later sections of the paper. In particular, empirical results are given concerning the distribution of the number of roots of the likelihood equation for samples size 3 to 19, and estimates are obtained for the variance and small sample efficiency of the M.L.E. of the location parameter for these sample sizes. These latter quantities are particularly useful since their theoretical evaluation is intractable and asymptotic expressions are grossly inaccurate over this range of sample sizes.

2. THE M.L.E. OF THE LOCATION PARAMETER OF THE CAUCHY DISTRIBUTION

Suppose a random variable X has a Cauchy distribution, with density function

$$f(x) = \frac{1}{\pi(1+(x-\theta)^2)} \quad (-\infty < x < \infty). \tag{1}$$

A sample of n independent observations, $x_1, x_2, \ldots x_n$, of X are available, on the basis of which we wish to obtain the maximum-likelihood estimator of θ. The log-likelihood of the sample is

$$L = -n\log\pi - \sum_{i=1}^{n}\log(1+(x_i-\theta)^2), \tag{2}$$

with derivative

$$\frac{\partial L}{\partial \theta} = \sum_{i=1}^{n} \frac{2(x_i - \theta)}{1 + (x_i - \theta)^2}. \tag{3}$$

For a given sample we need to determine the values of θ corresponding to relative maxima of L, and if more than one occurs, choose as M.L.E. that value which yields the absolute maximum of L.

Apart from the possibility of multiple relative maxima we have, in this situation, the added inconvenience that no explicit solution of

$$\partial L/\partial \theta = 0 \tag{4}$$

is possible and we have to resort to iterative methods to locate the turning points of L. The usual procedure is to take as starting value some consistent estimate of θ (the requirement of consistency having advantages theoretically), and to apply numerical techniques to obtain a relative maximum of L. Kendall & Stuart (1961) discuss the estimation of θ in (1) specifically and suggest the use of a fixed-derivative Newton method taking the sample median as starting value. In the event of multiple relative maxima of L, it is necessary to establish that any solution of (4) obtained by this approach corresponds to a maximum rather than a minimum (this must be so for the particular fixed derivative suggested by Kendall & Stuart), and further that we have obtained the absolute maximum rather than a relative one. This latter point is mentioned by Kendall & Stuart in their general discussion of M.L.E., but several practical problems arise in trying to use the particular approach they suggest for the Cauchy distribution.

For a large number of samples $x_1, x_2, \ldots, x_n$, from a Cauchy distribution L was found to have multiple relative maxima. But apart from empirical results, the fact that multiple relative maxima may occur is obvious from the form of $\partial L/\partial \theta$ given by (3); $\partial L/\partial \theta$ tends to zero, from below when $\theta \to +\infty$ and from above when $\theta \to -\infty$, whilst any 'relatively isolated' observation x_i must have the effect of making $\partial L/\partial \theta$ pass through zero from above, implying the presence of a relative maximum. Thus we may expect to encounter relative maxima up to n in extent, where n is the sample size. The occurrence of relatively isolated observations is quite likely in this situation on account of the extreme flatness of the Cauchy distribution. So it becomes a very real problem to determine which of these is the M.L.E. or indeed to locate them. The method suggested by Kendall & Stuart may locate a relative maximum, but on occasions even fails to converge. Should it locate a maximum there is still no guarantee that this will be the required absolute maximum. Nor is it obvious what other starting values should be taken to yield any other relative maxima, let alone ensure that we have located all of them.

In an attempt to avoid the trouble encountered with the fixed-derivative Newton method due to failure to converge, a Newton–Raphson approach was tried where the derivative was re-evaluated at each iteration. This proved even less successful due to the iterate often going off to infinity after encountering a turning point in $\partial L/\partial \theta$ (corresponding to a point of inflexion of L). Furthermore, there was no simple way of ensuring with this method that any zero of $\partial L/\partial \theta$ which was obtained, corresponded to a maximum rather than a minimum of L.

Another Newtonian method commonly used is the method of 'scoring for parameters'. In the particular example under discussion this is precisely the same as Kendall & Stuart's fixed-derivative Newton method for their particular choice of derivative. (Kendall & Stuart presumably applied the method of 'scoring for parameters' since this has been almost

exclusively advocated in the statistical literature.) Thus this method is seen to be prone to the same disadvantage as the fixed-derivative method in this case. (The two methods will not necessarily be equivalent in other situations, but this single example is sufficient to illustrate the potential disadvantage of 'scoring for parameters'.)

The failure of the above three methods to converge for some samples makes it undesirable to use any of them in general. What is required is a method that will guarantee to locate relative maxima of L for any sample of observations without degeneracy (as well as locating all relative maxima for each sample rather than just one; but this is a further problem and is discussed later). This requirement is met by the so-called 'method of false positions', subject to a minor modification to ensure that what is located is necessarily a maximum and not a minimum. Ostensibly this method has two drawbacks in that it is a lower order process than Newton–Raphson and needs the specification of two, rather than one, starting values. However, the mere fact that it is practicable (whereas the other three methods are not, for the reasons described above) outweighs the disadvantage of its lower order; and the need to specify two starting values presents no difficulties for the more general requirement of locating all relative maxima and choosing as M.L.E. that value of θ corresponding to the absolute maximum.

A fifth method is mentioned in the introduction and used by Cohen (1957). This is immediately seen to be another special case of the fixed-derivative Newton method, with its attendant disadvantages, so again cannot be recommended for general use. In fact all four methods other than the method of false positions are basically similar in form, stemming from a Taylor expansion up to the first derivative. They differ only in the value assigned to this first derivative.

These five methods are described in more detail in the next section, their interrelationships demonstrated and illustrative examples given to show the disadvantages of the Newton methods and the manner of application of the method of false positions. This latter method was used to obtain the M.L.E., $\hat{\theta}$, of θ for a very large number of random samples, from the distribution (1), of sizes $n = 3, 5, 7, 9, 11, 13, 15$ and 19. These results also yielded estimates of the variance of $\hat{\theta}$ for these values of n, together with the standard errors of the estimates. A weighted regression analysis was carried out of var $(\hat{\theta})$ on n and smoothed estimates obtained for var $(\hat{\theta})$. Hence estimates were obtained of the small sample efficiency of the M.L.E. for $n < 20$.

It might be conjectured that, whilst multiple relative maxima of L do arise, the use of an iterative process which starts in the neighbourhood of some consistent estimator, e.g. the median in the Cauchy case, would inevitably lead to the M.L.E. Empirical results show, however, that this is not so. The results presented do, however, show that the proportion of occasions on which the absolute maximum is not the relative maximum closest to the median is very small. But unfortunately there is no guarantee that the method of false positions, starting in the neighbourhood of the median, will necessarily yield the relative maximum closest to the median. If we could use a method which ensured this, the very small probability of not obtaining the M.L.E. in any particular situation might obviate the necessity of a complete scan of the likelihood function, but in the absence of such a surety this seems unavoidable.

From the results of the simulation, details have been extracted of the distribution of the number of relative maxima of L for different n, and these are also presented later.

3. Numerical methods of locating relative maxima

Five possible methods have been described above for locating a relative maximum of L. Some of these are special cases of others, but they are considered separately since they often appear in their own right in different applications. All but the method of false positions prove unsatisfactory on the grounds that they are liable not to converge. The latter method avoids this difficulty and is easily extended to scan the complete likelihood function and locate all relative maxima. The methods are described in more detail in this section and diagrams presented to illustrate their operation, and advantages and disadvantages.

All the methods except false positions are based on a Taylor expansion of the function whose roots we wish to estimate. Thus for maximum likelihood

$$0 = \frac{\partial L}{\partial \theta}\bigg|_{\hat{\theta}} \doteqdot \frac{\partial L}{\partial \theta}\bigg|_{\theta_0} + (\hat{\theta} - \theta_0)\frac{\partial^2 L}{\partial \theta^2}\bigg|_{\theta_0}, \tag{5}$$

where $\hat{\theta}$ is the M.L.E. and θ_0 some neighbouring value of θ. Rearranging (5) we have

$$\hat{\theta} = \theta_0 - \frac{\partial L}{\partial \theta}\bigg|_{\theta_0} \Big/ \frac{\partial^2 L}{\partial \theta^2}\bigg|_{\theta_0} \tag{6}$$

which may be used iteratively to obtain $\hat{\theta}$. The specific form of the iteration procedure depends on whether we re-evaluate the second derivative $\partial^2 L/\partial\theta^2$ at each stage, or choose, for computational convenience, to use a related function or a constant in place of the second derivative. These different approaches are now considered in more detail.

3·1. *The Newton–Raphson method.* Here, the second derivative is evaluated at each stage of the iteration procedure, yielding a second-order process. For the current problem this operates as follows. Starting with a consistent estimator of θ, say the median $\hat{\theta}_0$, we obtain successive iterates by

$$\hat{\theta}_{i+1} = \hat{\theta}_i - \frac{\partial L}{\partial \theta}\bigg|_{\hat{\theta}_i} \Big/ \frac{\partial^2 L}{\partial \theta^2}\bigg|_{\hat{\theta}_i}. \tag{7}$$

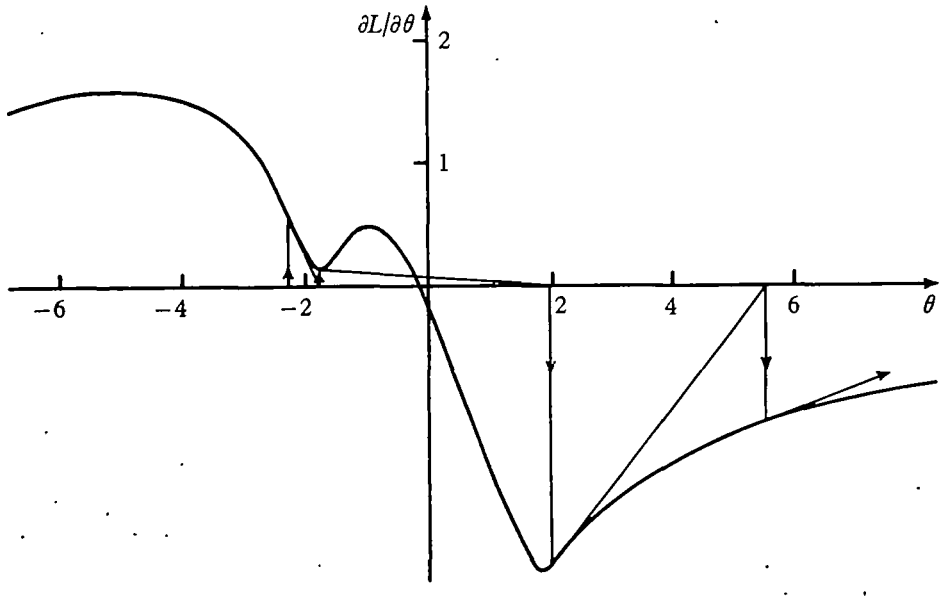

Fig. 1

This has two disadvantages. Any zero of $\partial L/\partial\theta$ which is located may correspond either to a relative maximum or a relative minimum of L. Furthermore, any point of inflexion of L may cause the iterate to go off to infinity or to cycle in the neighbourhood of the point of inflexion. In either case it will not converge to a zero of $\partial L/\partial\theta$. Thus we do not even require multiple roots of the likelihood equation to produce degenerate behaviour with this method; a stationary value of $\partial L/\partial\theta$ may produce this effect. This is best illustrated by studying the geometric representation of the process for the likelihood function of a particular random sample from the Cauchy distribution (see Fig. 1). Here we have $\partial L/\partial\theta$ based on a random sample of 5 observations, and the iterative process starts at the sample median (in this case $-2{\cdot}3$). The iterative process follows the slope of $\partial L/\partial\theta$ at each stage and is seen to go off to $+\infty$ after reaching the region of the point of inflexion of L at $-1{\cdot}9$. We can also see that it might continue to cycle in the region of such a point of inflexion.

3·2. *Fixed-derivative Newton approach.* Rather than re-evaluate $\partial^2 L/\partial\theta^2$ at each stage we might simply choose a sequence of constants a_i and replace the second derivative at the $(i+1)$st stage of the iteration by $-n/a_i$, where n is the sample size, yielding

$$\hat{\theta}_{i+1} = \hat{\theta}_i + \frac{a_i}{n}\frac{\partial L}{\partial\theta}\bigg|_{\hat{\theta}_i}. \tag{8}$$

Hildebrand (1956, pp. 443–50) discusses conditions which must be satisfied by the a_i, and Kale (1961) considers the resulting asymptotic probabilistic properties of the process.

In practice, however, for M.L.E., it is usual to choose a_i to be constant, a common choice being $(-n)$ divided by the value of the second derivative, $\partial^2 L/\partial\theta^2$, at the starting value of the process. Alternatively, $a_i = k$, where k is some arbitrary positive constant, is also often used.

These latter methods might be termed fixed-derivative Newton methods. One advantage for maximum-likelihood studies is that any zero of $\partial L/\partial\theta$ which is located will automatically correspond to a relative *maximum* of L. Unfortunately such methods often fail to converge due to the steepness of $\partial L/\partial\theta$ in the neighbourhood of the relative maximum of L, but continue to cycle around the relative maximum. This will occur if, for a relative maximum at θ_0,

$$-k\frac{\partial L}{\partial\theta}\bigg|_{\theta_0} > 2n. \tag{9}$$

Again this is illustrated for a random sample of 5 observations from the Cauchy distribution, the process starting at the sample median (see Fig. 2). Here, at any stage of the process, we obtain the next iterate by moving with fixed slope $-n/k$ from the current value of $\partial L/\partial\theta$ to the θ-axis.

Obviously for any given example we could avoid this trouble by choosing k in relation to the steepness of $\partial L/\partial\theta$ in the neighbourhood of any zero we are trying to locate. But this is entirely impracticable since it would require a complete knowledge of the form of $\partial L/\partial\theta$ in the example before we start the iteration procedure, which would be prohibitively laborious. We could avoid the disadvantage of the Newton–Raphson technique in the same way, but again this is not reasonable in practice. What we really require is a process which cannot degenerate and which can be applied in a systematic way, without detailed knowledge of the form of $\partial L/\partial\theta$ for the sample under study, to yield all the relative maxima of L. The method of false positions is seen later to satisfy these requirements.

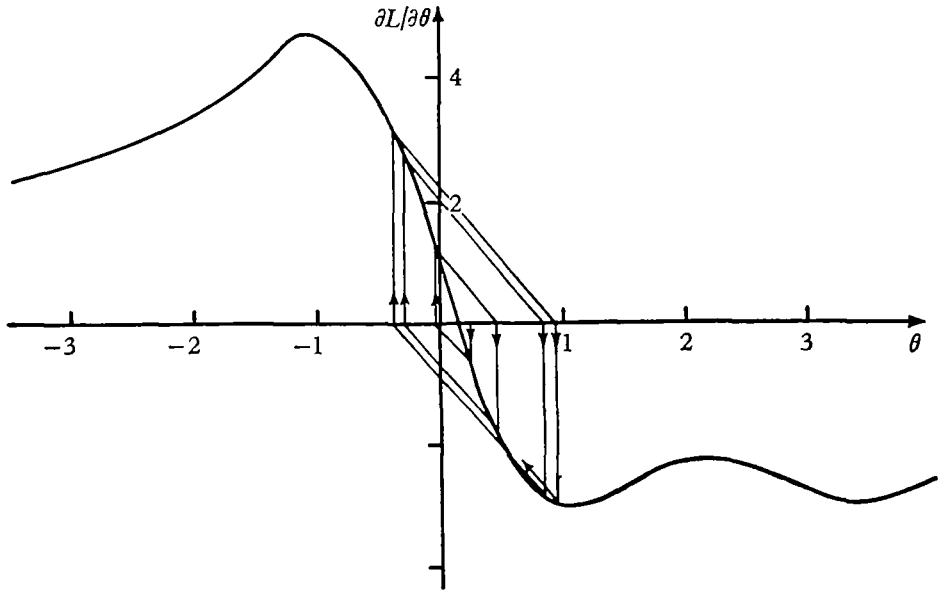

Fig. 2

3·3. '*Scoring for parameters.*' This is a method commonly used for M.L.E., and was introduced for this purpose by Fisher (1925). Here the second derivative is replaced by its expected value and the iterative process becomes,

$$\hat{\theta}_{i+1} = \hat{\theta}_i + \frac{\partial L}{\partial \theta}\Big|_{\hat{\theta}_i} \Big/ \{nI(\hat{\theta}_i)\}, \tag{10}$$

where

$$I(\theta) = E\left(-\frac{\partial^2 L}{\partial \theta^2}\right). \tag{11}$$

No immediate general comments seem possible on the practical utility of this approach for different distributions, but it is again obvious that it cannot, in the spirit of the previous paragraph, be recommended for general application. This is seen by considering the location parameter of the Cauchy distribution where $I(\theta)$ is independent of θ, having the value $\frac{1}{2}$. Thus the method reduces, in this case (and in other cases also, whenever $I(\theta)$ is constant) to the previous method, with its attendant disadvantages. This is the approach used by Kendall & Stuart (1961) in their discussion of the M.L.E. of the location parameter of the Cauchy distribution.

3·4. *A further standard method.* Another method commonly used (see e.g. Whittaker & Robinson, 1944, p. 81) proceeds as follows. If we wish to locate a root of $g(x) = 0$, we rewrite it as $g_1(x) = g_2(x)$, which can be done in many ways, and apply the iterative procedure

$$x_{i+1} = g_1^{-1}(g_2(x_i)), \tag{12}$$

starting with x_0 in the neighbourhood of the root we are seeking. For graphical applications any appropriate functions $g_1(.)$ and $g_2(.)$ may be chosen. But for reasonable accuracy we shall wish to carry out numerical calculations rather than a graphical interpolation, and it

is then necessary to choose a simple form for $g_1(.)$ to facilitate the inversion stage of the process. Usually we would choose $g_1(x) \equiv x$. Thus for M.L.E. estimation we should choose

$$\left.\begin{aligned} g_1(\theta) &= \theta, \\ g_2(\theta) &= \frac{\partial L}{\partial \theta} + \theta, \end{aligned}\right\} \tag{13}$$

and the iterative process becomes

$$\hat{\theta}_{i+1} = \hat{\theta}_i + \left.\frac{\partial L}{\partial \theta}\right|_{\hat{\theta}_i}. \tag{14}$$

For example, this is the one-parameter analogue of the method used by Cohen (1957) to evaluate the M.L.E.'s of the mean and variance of truncated and censored normal distributions. But (14) is again an example of the fixed-derivative Newton method, and will be unacceptable as a general method for the reasons described above. The method is particularly prone to degeneracy due to the relatively large value of the coefficient of $\partial L/\partial\theta$.

More general choices of $g_1(.)$ and $g_2(.)$ complicate the calculations and will, in any case, not guarantee that we do not get similar degenerate behaviour.

3·5. *The method of false positions.* The disadvantages of the previous methods can be avoided by using the method of false positions. With this method we can ensure the location of a relative maximum. We start by choosing two values of θ, a_0 and b_0 say, such that

$$a_0 < b_0, \quad \left.\frac{\partial L}{\partial \theta}\right|_{a_0} > 0, \quad \left.\frac{\partial L}{\partial \theta}\right|_{b_0} < 0. \tag{15}$$

Writing $\partial L/\partial\theta|_x$ as $f(x)$ and putting

$$x_i = \frac{a_{i-1} f(b_{i-1}) - b_{i-1} f(a_{i-1})}{f(b_{i-1}) - f(a_{i-1})},$$

successive a_i, b_i are then obtained as

$$\left.\begin{aligned} a_i &= x_i \\ b_i &= b_{i-1} \end{aligned}\right\} \quad (f(x_i) > 0), \qquad \text{or} \qquad \left.\begin{aligned} a_i &= a_{i-1} \\ b_i &= x_i \end{aligned}\right\} \quad (f(x_i) < 0). \tag{16}$$

With this procedure a_i and b_i continue to enclose and converge on a relative maximum of L, and when $b_i - a_i$ is sufficiently small the process terminates and yields this maximum as $\frac{1}{2}(a_i + b_i)$. It is not immediately obvious, however, how a_0 and b_0 should be chosen in this situation. One possibility is to consider the values

$$\left.\begin{aligned} x_j &= m - j\epsilon \\ y_j &= m + j\epsilon \end{aligned}\right\} \quad (j = 1, 2, \ldots), \tag{17}$$

where m is some consistent estimator, say (in the Cauchy case) the sample median and ϵ a small positive quantity, choosing as a_0 and b_0 the first pair x_j, y_j to satisfy (15). We shall certainly locate a relative maximum in this way, but there is no guarantee that it will be the one closest to m. Geometrically the process operates as shown in Fig. 3, where it is applied to a further random sample of 5 observations from (1). (For illustrative purposes the quantity ϵ has been assigned the unnaturally large value of 4 in this figure. Had ϵ been chosen as 0·25 say, the process would very rapidly have located the relative maximum just above m.)

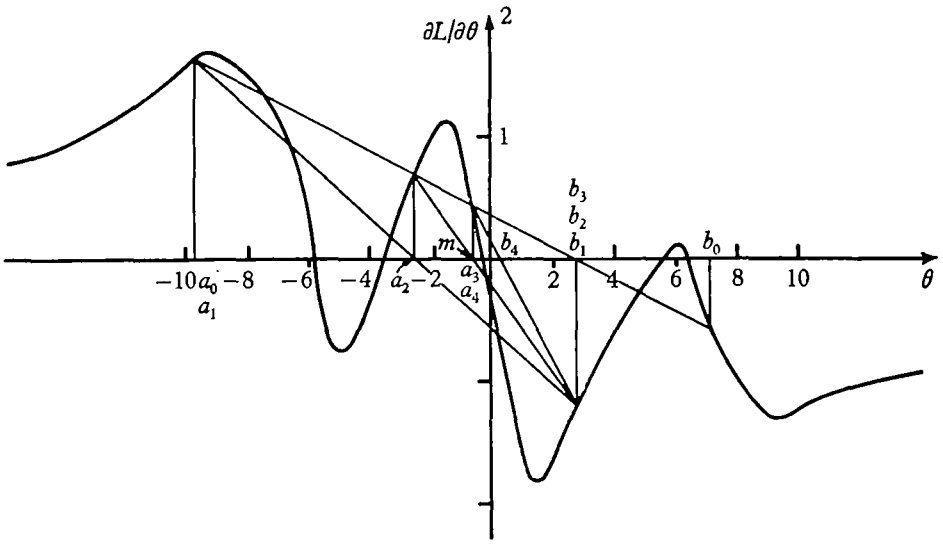

Fig. 3

It is a simple matter to extend this method to locate all the relative maxima in a given situation and, by observation of the values of L at these relative maxima, to determine the M.L.E. $\hat{\theta}$. Obviously in the current context there can be no relative maxima occurring at a value less than the minimum observation $x_{(1)}$ or in excess of the maximum observation $x_{(n)}$. Consequently we have only to consider starting values a_0 and b_0 in this range. Starting with $a_0 = x_{(1)}$, $b_0 = x_{(1)} + \epsilon$ we use the method described above to locate a relative maximum between a_0 and b_0, *if* a_0 *and* b_0 *satisfy* (8). We then define new a_0 and b_0 by the relation

$$\left.\begin{aligned} a_0^{(i)} &= b_0^{(i-1)}, \\ b_0^{(i)} &= b_0^{(i-1)} + \epsilon, \end{aligned}\right\} \tag{18}$$

and carry out this procedure again, continuing in this way until $b_0^{(i)} \geqslant x_{(n)}$. That is, we consider an interval of width ϵ, with a_0 and b_0 at its extremes, and move this interval in steps of ϵ over the range of values of the observations. In this way we locate the various relative maxima of L for our sample, storing information about their positions and associated values of L. Inevitably we must run the risk that our interval of width ϵ might enclose more than one relative maximum, of which only one will be located. But this is not a problem from a practical point of view. We can choose ϵ small enough to ensure that there is negligible probability of this happening. Furthermore, should we not completely avoid this difficulty, the rare occurrence of multiple roots in any small interval will cause no appreciable bias to the estimation of the M.L.E. Any two relative maxima less than a distance ϵ apart (where ϵ is small) must have associated values of L quite close in value. Thus if we adjudge one of them the M.L.E. it is an unimportant formal distinction that the other might have a trivially higher associated value of L; also, even if we regard this formal distinction as important we cannot introduce an error in excess of ϵ. In the same way, we may feel confident that we have not missed a relative maximum with L substantially higher than any of those we have located.

Detailed empirical study of this method for the Cauchy distribution revealed that the choice of $\epsilon = 0{\cdot}25$ was reasonable. No situations were encountered where the above pro-

cedure failed to locate any relative maxima with this choice of ϵ, let alone where the M.L.E. of θ was wrongly determined. In the process of this study, and the above-mentioned examination of the alternative iterative methods, a very large number of random samples (*c.* 30,000) of different sample sizes from 3 to 19 were simulated on a digital computer. As well as yielding the qualitative information on the relative merits of the different numerical methods of obtaining the M.L.E. of θ reported above, this material also produced extensive quantitative information specifically concerned with the Cauchy distribution. In particular it was possible to study the bias, variance and small-sample efficiency of the M.L.E. of the location of this distribution. This is discussed in § 4 of this paper.

3·6. *Discussion.* The remarks above on the utility of the different numerical methods suggest that none of them is practicable except the method of false positions. It should be realized that this is a somewhat extreme viewpoint to adopt, since the rejection of all methods but false positions is based on the fact that such methods may potentially fail (and have been observed to do so). This is to ignore completely the relative frequency of failure of the different methods. It would appear that the incidence of such failures is very low, and that in many practical situations the sample drawn will not occasion any of the difficulties discussed above. In such cases any of the methods may of course be used. To obtain single estimates or a small number of estimates it may well be worthwhile to use one of the Newton methods with its greater ease of application, and to accept the risk of degeneracy. But where multiple roots may occur we can never be sure that we have located all the roots in any example by these methods without a detailed study of the complete likelihood function, and it is unlikely that any real economy of effort over the method of false positions will be obtained.

For more extensive calculations where we require estimates of the M.L.E. for a large number of samples, e.g. in simulation, there is little alternative but to use false positions. Inevitably such work will be mechanized (perhaps conducted on a computer) and it is essential that the calculations can proceed unmonitored without causing any failure. We have no such guarantee of the behaviour of the Newton methods.

4. QUANTITATIVE FEATURES OF THE M.L.E. OF THE LOCATION OF THE CAUCHY DISTRIBUTION

The method of false positions described in § 3·5 was used to determine $\hat{\theta}$ for a large number, k_n, of random samples of n observations from the distribution (1) with $\theta = 0$, for $n = 3(2)\,15, 19$. Random observations, x_i, from the Cauchy distribution were obtained by a direct transformation of random uniform deviates, y_i, generated by the multiplicative congruential method, using

$$x_i = \tan(\pi y_i - \tfrac{1}{2}\pi). \tag{19}$$

All the calculations were carried out on the English Electric K.D.F. 9 computer in the University of Birmingham. The results were used to evaluate, for each n, estimates of the mean and variance of $\hat{\theta}$ and the standard error of the variance of $\hat{\theta}$, the number of occasions, p_n, on which $\hat{\theta}$ was not the relative maximum closest to the sample median, and the relative frequency distribution of the number of relative maxima. These various quantities are presented in Table 1.

Some comments on the contents of this table are needed, and some preliminary conclusions may be drawn.

Table 1

n	k_n	$\hat{E}(\hat{\theta})$	$\hat{\text{var}}(\hat{\theta})$	S.E. $[\hat{\text{var}}(\hat{\theta})]$	p_n	Relative frequency distribution of no. of relative maxima: 1	2	3	4	5	6	7
3	18,000	−0·0010	5·1693	$8{\cdot}58 \times 10^{-1}$	0	0·646	0·262	0·092	—	—	—	—
5	3,000	·0007	0·9414	$6{\cdot}28 \times 10^{-2}$	50	·652	·268	·069	0·011	0·001	—	—
7	3,000	− ·0027	·4940	$3{\cdot}03 \times 10^{-2}$	23	·670	·261	·058	·009	·001	—	—
9	2,250	·0084	·3187	$1{\cdot}75 \times 10^{-2}$	10	·673	·269	·052	·005	—	—	—
11	1,000	·0148	·2475	$1{\cdot}88 \times 10^{-2}$	2	·706	·245	·036	·012	·001	—	—
13	931	− ·0275	·1876	$1{\cdot}33 \times 10^{-2}$	1	·698	·255	·039	·009	—	—	—
15	1,000	− ·0123	·1575	$8{\cdot}90 \times 10^{-3}$	0	·696	·262	·039	·002	·001	—	—
19	784	·0053	·1178	$8{\cdot}42 \times 10^{-3}$	0	·707	·245	·043	·005	—	—	—

(1) The value of k_n, for each n, was chosen with regard to the relative cost of simulation for the given value of n. The fact that k_{13} and k_{19} have apparently arbitrary values is due to the difficulty of assessing computing time with the result that the computer terminated the calculations when the prescribed upper time limit was reached even if the required number of simulations had not been carried out.

(2) The extreme disparity of the accuracy of the estimates of var $(\hat{\theta})$ (as reflected by their estimated standard errors) was not entirely fortuitous. The main interest in this work was in estimating the small sample efficiency of the M.L.E. of θ. The estimated standard error of this latter quantity shows a much smaller variation from one value of n to another, as is shown in the third column of Table 2.

Table 2

n	$\hat{e}_n$	Estimated S.E.	Asymptotic e_n	Smoothed $\hat{e}_n$	S.E.	Sample median f_n
3	0·129	0·021	0·545	0·129	0·0070	—
5	·425	·028	·667	·421	·0072	0·328
7	·578	·035	·737	·586	·0057	·467
9	·697	·038	·783	·688	·0050	·544
11	·735	·056	·815	·756	·0056	·593
13	·820	·058	·839	·806	·0066	·626
15	·847	·048	·857	·843	·0077	·651
19	·894	·064	·884	·895	·0096	·670

(3) There would seem to be no systematic bias in $\hat{\theta}$. It is easily shown theoretically that $\hat{\theta}$ is unbiased for the Cauchy distribution.

(4) The variance of $\hat{\theta}$ increases very rapidly for $n < 10$. We should expect this, since for $n = 1$, $\hat{\theta} = x$, the observed value, which has infinite variance.

(5) The incidence of values of $\hat{\theta}$ not corresponding to the relative maximum nearest the sample median is negligible (the maximum relative frequency being less than 2 % when $n = 5$).

(6) Multiple relative maxima seem to occur on about 30 % of occasions for all values of n. The distribution of the number of relative maxima appears to settle down to a common form for n in excess of about 10, with a single maximum on about 70 % of occasions and two maxima on about 25 % of occasions. (A simple χ^2 test of homogeneity of distributional form shows no significant difference between the observed distributions for $n = 11, 13, 15$ and 19.)

The exact distribution of the number of relative maxima provides an interesting source of further work.

4·1. *Small sample efficiency of* $\hat{\theta}$. If we accept the criterion of efficiency proposed by Cramér (1946) the small sample efficiency of $\hat{\theta}$ may be defined as

$$e_n = \frac{2}{n \operatorname{var}(\hat{\theta})}. \tag{20}$$

Attempts to obtain e_n explicitly have not proved successful due to the intractability of $\operatorname{var}(\hat{\theta})$. Shenton & Bowman (1963), extending some work of Haldane & Smith (1956), have given an expression for the asymptotic form of $\operatorname{var}(\hat{\theta})$ for a general single parameter distribution up to the term in n^{-3}. Thus we have

$$\operatorname{var}(\hat{\theta}) = \frac{1}{nI} + \frac{1}{n^2}\sum_1^4 b_s I^{-s} + \frac{1}{n^3}\sum_1^7 c_s I^{-s}, \tag{21}$$

where

$$I = E\left\{\frac{-\partial^2 \log f}{\partial \theta^2}\right\}, \tag{22}$$

and the coefficients b_s and c_s are rather complicated linear combinations of terms each of which may involve the product of up to four quantities of the form

$$E\left\{\left(\frac{1}{f}\frac{\partial f}{\partial \theta}\right)^i \left(\frac{1}{f}\frac{\partial^2 f}{\partial \theta^2}\right)^j \left(\frac{1}{f}\frac{\partial^3 f}{\partial \theta^3}\right)^k \cdots\right\},$$

in which f is the probability density function of the random variable under consideration.

For the Cauchy distribution $I = \frac{1}{2}$ and we immediately obtain the leading two terms of (21) as

$$\operatorname{var}(\hat{\theta}) \sim \frac{2}{n} + \frac{5}{n^2} \cdots, \tag{23}$$

or

$$e_n \sim \left\{1 + \frac{5}{2n} \cdots\right\}^{-1}. \tag{24}$$

Table 2 presents the simulation estimates, $\hat{e}_n$, of the efficiency together with their estimated standard errors, and also the values of e_n given by the asymptotic form (24). It is obvious that (24) is quite inadequate for values of n less than about 12. Some rather tedious calculations yield the coefficient $-42{\cdot}87$ for the term in n^{-3}. The resulting estimate of e_n is considerably worse than that given by (24), and in fact meaningless if $n < 9$. Consequently it was decided to estimate e_n by smoothing the simulation estimates $\hat{e}_n$. For this purpose a weighted regression analysis of $\hat{e}_n$ on n was carried out, with weights inversely proportional to $\sqrt{n}$ in an attempt to make use of the decreasing accuracy of the $\hat{e}_n$. A model was assumed of the form

$$\left.\begin{aligned} E(\hat{e}_n|n) &= \alpha + \frac{\beta}{n} + \frac{\gamma}{n^2}, \\ \operatorname{var}(\hat{e}_n|n) &= n\sigma^2, \end{aligned}\right\} \tag{25}$$

and the parameters α, β and γ estimated from the data by 'least squares'. This analysis yielded the following estimates of α, β and γ,

$$\hat{\alpha} = 1{\cdot}102, \quad \hat{\beta} = -4{\cdot}133, \quad \hat{\gamma} = 3{\cdot}646,$$

resulting in estimates of e_n (and corresponding standard errors) as given in columns 5 and 6 of Table 2. The final column of Table 2 gives, for the purposes of comparison, the small sample efficiency, f_n, of the sample median. Fig. 4 presents a graphical representation of the e_n and their asymptotic and least-square estimates.

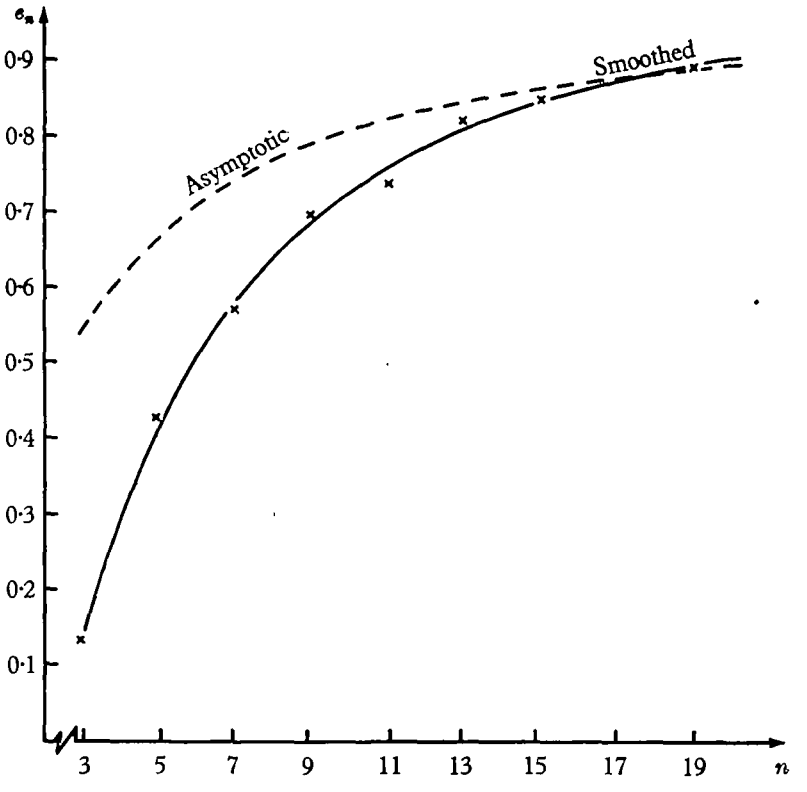

Fig. 4

4·2. *Conclusions.* The extensive calculations needed to ensure the correct evaluation of the M.L.E. of the location parameter of a Cauchy distribution, and the poor efficiency of the M.L.E. for small samples (indicated in Table 2) render the utility of the maximum-likelihood method rather doubtful in this context. Table 2 also shows that the sample median is no more efficient, but it is interesting to observe that the relative efficiency of the sample median compared with the M.L.E. remains very close to its asymptotic value for all sample sizes, i.e. about 80 %.

This raises the question of whether the extreme increase in labour required to obtain the M.L.E. is justified for such a small increase in efficiency, particularly since the fuller use of order statistics in the estimation procedure will inevitably reduce the advantage of the M.L.E. even more for little increase in effort. The author is currently preparing tables of the quantities required to obtain the best linear unbiased estimator from the order statistics, together with a discussion of their small sample efficiency.

Rather than obtaining comparable efficiency for less effort, we might consider the possibility of obtaining more efficient estimates for comparable effort. In the face of multiple relative maxima it is possible that an estimator based on a smoothed version of the likelihood

function may be more efficient than the M.L.E. (cf. Daniels (1960), who suggests the use of such estimators). In the absence of any theoretical work in this direction for small samples it is impossible to judge the value of such an approach. However, a simple simulation study suggests that this might be a useful method. Sample estimates have been obtained of the variance of an estimator of θ based on maximization of the quantity

$$\frac{1}{2\xi}\int_{\theta-\xi}^{\theta+\xi} L(\theta'; \mathbf{x})\, d\theta', \tag{26}$$

rather than L itself. Exactly the same procedure was used as for the standard maximum-likelihood estimation (described above) except that L was replaced by (26) and we sought zeros of

$$\frac{1}{2\xi}\{L(\theta+\xi; \mathbf{x}) - L(\theta-\xi; \mathbf{x})\} \tag{27}$$

rather than $\partial L/\partial\theta$. The same set of 500 random samples of 5 observations were used for 7 different values of ξ, viz. $\xi = 0, 0{\cdot}5, 1{\cdot}0, 1{\cdot}5, 2{\cdot}0, 3{\cdot}0$ and $4{\cdot}0$. The same set of 500 samples was used in order to get a representative idea of the effect of this procedure unclouded by inevitable sampling variations which would have arisen between distinct sets of samples and which may have concealed any true effect. The following results were obtained:

ξ	0	0·5	1·0	1·5	2·0	3·0	4·0
$\hat{e}_n$	0·416	0·426	0·450	0·483	0·494	0·448	0·373

suggesting that we might improve on the M.L.E. by using such a method and obtain in this case an increase in efficiency of the order of 10 % for an optimal value of ξ in the region of 2 (see Fig. 5).

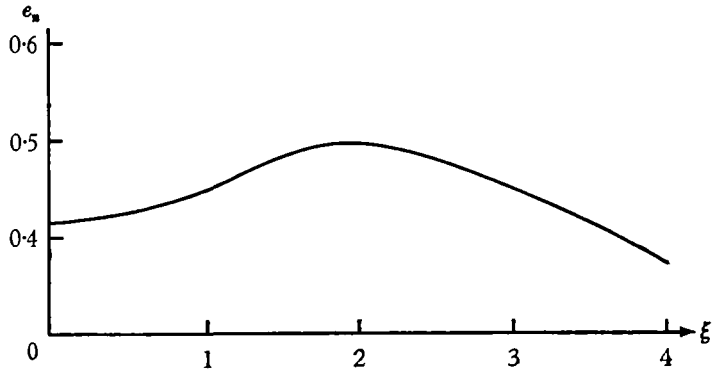

Fig. 5

Before such a procedure could be adopted in practice, there is obviously a great deal of theoretical work needed to describe its form and behaviour. Needless to say we are not restricted to the particular form of smoothing exhibited in (26). The introduction of some function of ξ into the integrand may well produce even more efficient estimates.

REFERENCES

COHEN, A. C. (1957). On the solution of estimating equations for truncated and censored samples from normal populations. *Biometrika*, **44**, 225–36.

CRAMÉR, H. (1946). *Mathematical Methods of Statistics*. Princeton University Press.

DANIELS, H. E. (1960). The asymptotic efficiency of a maximum likelihood estimator. *Proceedings of the Fourth Berkeley Symposium on Mathematical Statistics and Probability*, Berkeley, **1**, 151–63.

FISHER, R. A. (1925). Theory of statistical estimation. *Proc. Camb. Phil. Soc.* **22**, 700–25.
FISHER, R. A. (1950). *Statistical Methods for Research Workers*. Edinburgh: Oliver and Boyd.
HALDANE, J. B. S. & SMITH, S. M. (1956). The sampling distribution of a maximum likelihood estimate. *Biometrika*, **43**, 96–103.
HAMMING, R. W. (1962). *Numerical Methods for Scientists and Engineers*. New York: McGraw-Hill.
HILDEBRAND, F. B. (1956). *Introduction to Numerical Analysis*. New York: McGraw-Hill.
KALE, B. K. (1961). On the solution of the likelihood equation by iteration processes. *Biometrika*, **48**, 452–6.
KALE, B. K. (1962). On the solution of the likelihood equation by iteration processes—the multiparametric case. *Biometrika*, **49**, 479–86.
KENDALL, M. G. & STUART, A. (1961). *The Advanced Theory of Statistics*, **2**. London: Charles Griffin and Co.
NORTON, H. W. (1956). One likelihood adjustment may be inadequate. *Biometrics*, **12**, 79–81.
SHENTON, L. R. & BOWMAN, K. (1963). Higher moments of a maximum-likelihood estimate. *J. R. Statist. Soc.* B, **25**, 305–17.
WHITTAKER, E. T. & ROBINSON, G. (1944). *The Calculus of Observations*. London: Blackie and Son.