Maximum Likelihood Estimation from Incomplete Data

H. O. Hartley

Biometrics, Volume 14, Issue 2 (Jun., 1958), 174-194.

Stable URL:
http://links.jstor.org/sici?sici=0006-341X%28195806%2914%3A2%3C174%3AMLEFID%3E2.0.CO%3B2-L

MAXIMUM LIKELIHOOD ESTIMATION FROM INCOMPLETE DATA

H. O. HARTLEY

Iowa State College, Ames, Iowa, U.S.A.

I. Introduction

If a random sample of size N has been drawn from a population involving unknown parameters $\theta_1 \cdots \theta_k$, the latter may be estimated from the sample by the well-known technique of maximum likelihood the properties of which have been frequently examined in the past. The equations resulting from this technique are often simple but sometimes require iterative procedures which are usually supported by special aid-tables.

However, situations frequently arise in which the data are 'incomplete' in the sense that the information contained in the random sample of size N is not completely available for estimation. The two best known examples of such 'incomplete' information have become known under the names of 'truncated' and 'censored' samples but there are numerous other situations. The literature of the past decades is abundant with papers dealing with special cases of such 'incomplete information' for special distributions (see e.g. Finney, D. J. [1949]). Many authors, considering the computational work involved in maximum likelihood estimation prohibitive, have suggested alternative methods. (See e.g. Moore [1952].)

The purpose of the present paper is to simplify and unify the maximum likelihood computations of estimates from 'incomplete data.' It should be stressed that we are using here standard maximum likelihood estimation as introduced by Fisher and as used by many statisticians, and further, that we are *not* concerned here with discussing the small sample efficiency of these methods. We do, however, offer a new approach to the computational evaluation of these estimators from data in incomplete samples which represents a considerable improvement on the methods in current use. We shall in fact show how maximum likelihood estimation from 'incomplete data' can be simply reduced to that from a complete sample. The method is akin

to the 'missing plot technique' in analysis of variance, in fact the latter may be regarded as a special case of our simple iterative method described below. This procedure is applicable to any situation in which a maximum likelihood procedure for the *complete* sample is available. It covers a wide class of problems in which incomplete data arise incorporating as special cases problems of truncation or censoring at either end of the distribution or in the central section or any combination thereof. It is not claimed that in every particular problem of truncation or censorship the present method is simpler than an already existing method specifically developed for that situation and supported by special aid tables, nevertheless this will be so in many cases. The main purpose of the paper is to provide a method of complete generality depending only on tables of the probabilities of the distribution and not on aids or methods which vary with the nature of the incompleteness of the data, and more particularly provide a workable method for the multitude of cases in which no special methods are available. We commence by describing the method for discrete distributions leaving the discussion of continuous distributions to a second paper in which we shall also discuss the fit of a truncated Gamma distribution to rainfall data which problem gave rise to the present study.

2. *Discrete distributions with missing frequencies (truncation)*

We first describe the method in terms of a simple numerical example which we employ to introduce the notation; we then give a general definition of this case and proof of the formulas used.

Example 1.

Snedecor [1956, p. 483] quotes data of Leggatt [1933] on the pollution of seeds of Phleum pratense by the presence of a few noxious weed seeds. Table 1 shows the frequency with which 2, 3, 4, $\cdots$ of these weed seeds were found in 78 quarter-ounce samples. Actually Snedecor also gives the frequencies of quarter ounces with 0 or 1 weed seed but we assume, for purposes of illustration, that these frequencies are not available, that we must estimate the mean number of weed seeds per quarter ounce from the remaining frequencies and that this count x follows a Poisson a distribution $e^{-\theta}\theta^x/x!$ We introduce the following notation:

$$n_i = \text{set of observed frequencies } \textstyle\sum_i n_i = n, \tag{1}$$

$$_cn_j = c\text{-th estimate of } j\text{-th missing frequency } \textstyle\sum_j {}_cn_j = {}_cn', \tag{2}$$

$$f(i, \theta) = \text{probability of } x = i \text{ (here } e^{-\theta}\theta^i/i!), \tag{3}$$

$$f(j, \theta) = \text{probability of } x = j \text{ (here } e^{-\theta}\theta^j/j!), \text{ and} \tag{4}$$

$$f(\theta) = \sum_j f(j, \theta). \tag{5}$$

The iterative procedure described below and illustrated numerically in Table 1 yields the maximum likelihood estimate of the Poisson mean θ (as will be shown below) and consists of the following steps:

(a_0) Estimate initial values of the missing frequencies ${}_0n_j$ (in Table 1 ${}_0n_0 = 4$, ${}_0n_1 = 14$, so that ${}_0n' = 18$).

(b_0) Using the n_i and the ${}_0n_j$ compute an initial estimate ${}_1\theta$ of the parameter from the maximum likelihood equation appropriate to the complete sample. In the example this equation is the mean of the complete distribution so that

$${}_1\theta = (\sum_i in_i + \sum_j j_0n_j)/(n + {}_0n') = (279 + 14)/96 = 3.052. \tag{6}$$

(a_1) Using ${}_1\theta$ compute 'improved' estimates, ${}_1n_j$, of the 'missing frequencies' by proportional allocation based on the Poisson distribution, i.e.

$${}_1n_j = nf(j, {}_1\theta)/(1 - f). \tag{7}$$

In the example (see Table 1) we have

$${}_1n_0 = 4.564, \quad {}_1n_1 = 13.935, \quad \text{so that} \quad {}_1n' = 18.499.$$

The reason for the high decimal accuracy in these preliminary calculations will be apparent in Section 5.

(b_1) Using the n_i and the 'improved' ${}_1n_j$, compute an 'improved' estimate ${}_2\theta$ of θ from the maximum likelihood equation. In the example

$${}_2\theta = (\sum_i in_i + \sum_j j_1n_j)/(n + {}_1n') = (279 + 13.935)/(78 + 18.499) = 3.036.$$

This value is already close to the previous one of ${}_1\theta = 3.052$ but for higher accuracy two more cycles, shown in Table 1, yield ${}_4\theta = 3.026$ which is the maximum likelihood estimate of θ to almost 3 decimal accuracy.

The convergence of the process has been found to be extremely rapid in some 30 examples worked, provided care is taken in the initial choice of the ${}_0n_j$ and provided that the truncation is not 'too drastic,' i.e. provided $n'/(n + n')$ is below $\frac{1}{3}$. Above this value the convergence may still be serviceable if accelerated by such methods as described in

TABLE 1

MAXIMUM LIKELIHOOD FIT TO POISSON DISTRIBUTION WITH MISSING FREQUENCIES

Count x = i j	n_i	${}_0n_j$	${}_1n_j$	${}_2n_j$	${}_3n_j$	$f(j, {}_1\theta)$	$f(j, {}_2\theta)$	$f(j, {}_3\theta)$
0		4	4.564	4.655	4.690	.0473	.0481	.0484
1		14	13.935	14.119	14.205	.1444	.1459	.1466
2	26					From table of Poisson distribution		
3	16							
4	18							
5	9							
6	3							
7	5							
8	0							
9	1							
10								
Totals	$n=78$	${}_0n'=18$	18.499	18.774	18.895	$f=.1917$	.1940	.1950
	$\sum in_i=279$	$\sum j_0n_j=14$	13.935	14.119	14.205			
		${}_0\theta=3.052$	3.036	3.029	3.026			

Section 4. [In the example $n' = 19$, $n = 78$, $n'/(n + n') = 0.2$.] The theoretical background of the convergence will be discussed in a second paper and the variance computation of the resulting maximum likelihood estimates in Section 5. The effect on convergence of choosing poorer starting values is discussed in Section 4.

The above is an example of what is known as a sample from a 'truncated distribution' or more precisely we are speaking of a distribution truncated at the lower tail end. In such a situation we omit from the original distribution (here the Poisson distribution) all frequencies for which $x \leq i^*$ where i^* is the point of truncation (here $i^* = 3$). We then proceed to draw a random sample of fixed size n from the 'truncated distribution' consisting of the (Poisson) frequencies for which $x \geq i^*$ divided by their total. An alternative description of the above process of sampling is to draw random samples of unknown size from the complete distribution and then to consider the subpopulation of samples for which the number of x values with $x \geq i^*$ is a constant sample size n, whilst the remaining values with $x \leq i^*$ are

unknown. The first description of the sampling procedure provides a clear theoretical background for the method of estimation of the parameters of the truncated, and hence of the complete distribution (e.g. by maximum likelihood). The second description is supposed to explain how such 'truncated samples' may arise in practice and has recently been critically discussed by F. N. David and N. L. Johnson [1952], who raise the question whether there may be practical situations where this method of sampling does in fact not apply. Nevertheless it has been almost universally accepted in the past.

The above case of 'truncation' can be described briefly as a sample in which all frequencies for counts $x < i^*$ are 'missing.'

In the present treatment we deal with a more general situation of missing frequencies: We assume that the set of all possible counts $x = 0, 1, 2, \cdots$ is subdivided into two subsets, namely:

(a) the counts $x = i$ for which frequencies n_i were observed

(b) the counts $x = j$ for which frequencies n_j were not observed. (In Table 1, $j = 0, 1$ and $i = 2, 3, 4, \cdots$.)

Again, it is assumed that $\sum_i n_i = n$ is fixed in repeated sampling whilst $\sum_j n_j$ is of course unknown.

We now turn to the problem of estimating the parameters of the distribution which has given rise to the counts x. We give the proof that the procedure described in the above example yields the maximum likelihood solution. We do this for a general distribution depending on k parameters θ_t $(t = 1, 2, \cdots, k)$ which must be estimated.

Let us consider a discrete variate x capable of attaining integral values in the two mutually exclusive and exhaustive ranges denoted by 'i' and 'j' with respective probabilities $f(i, \theta_1 \cdots \theta_k)$ and $f(j, \theta_1 \cdots \theta_k)$. For these we shall use the short notation $f(i, \theta)$ and $f(j, \theta)$ given by (3) and (4) denoting by θ the dependence of f on the multi-parameter vector $\theta_1 \cdots \theta_k$. We have

$$\sum_i f(i, \theta) + \sum_j f(j, \theta) = 1 \qquad (7')$$

where the summations are respectively over the two mutually exclusive ranges i and j of the variate x. Further we write

$$\sum_j f(j, \theta) = f(\theta). \qquad (8)$$

We now draw a random sample from this population and denote by n_i the frequencies observed for the set of integers 'i'. No frequencies are available for the set of integers 'j'. The maximum likelihood equations for the estimation of the θ_t are then given by

$$\sum_i n_i \left\{ \frac{f_t(i, \theta)}{f(i, \theta)} + \frac{f_t(\theta)}{1 - f(\theta)} \right\} = 0 \qquad t = 1, 2, \cdots k \qquad (9)$$

where the subscript t denotes differentiation with respect to θ_t . We now introduce auxiliary variables n_j given by the proportional allocation formulas

$$n_j = nf(j, \theta)/[1 - f(\theta)]. \tag{10}$$

Substituting $\sum n_i = n$, (8) differentiated with regard to θ_t and (10) in the second term of (9) we obtain

$$\sum_i n_i \frac{f_t(i, \theta)}{f(i, \theta)} + \sum_j n_j \frac{f_t(j, \theta)}{f(j, \theta)} = 0 \tag{11}$$

so that the system of equations (10) and (11) is identical with the set of maximum likelihood equations (9). But equations (11) are formally identical with the maximum likelihood equations for a 'complete sample' of size $n + n'$ with observed 'cell frequencies' n_i and n_j and $\sum n_i = n$, $\sum n_j = n'$. Therefore the iterative procedure described above, if it converges, must yield a solution $\hat{\theta}_t = \lim_{c\to\infty} {}_c\theta_t$, $\hat{n}_j = \lim_{c\to\infty} {}_cn_j$ of equations (10) and (11) and hence the vector $\hat{\theta}$ will be a solution of the maximum likelihood equations (9). When it is known that these latter equations (9) can only have a unique solution the iterative process must yield this solution.

It will be appreciated that the present method is particularly suited to situations where the maximum likelihood equations for the *complete* sample are considerably simpler to solve than those for the *incomplete* problem, as is the case for the Poisson. Nevertheless the method will still be of great help when the maximum likelihood equations for the complete sample must be solved by iterative methods. This is shown for the negative binomial in Example 4.

3. Discrete distributions with grouped frequencies (censoring)

The classical situation of 'censoring' arises when a random sample of counts, x, has been drawn from a population and for the counts x which exceed the points of censorship i^* ($x > i^*$) only the total number of counts N' is known whilst the precise values of the counts are unknown (have been censored). The counts x with $x \leq i^*$ are all known. (See e.g. the example at the end of this section where $N = 240$, $N' = 128$ and $i^* = 2$.)

In this situation it is assumed that in repeated sampling the *total* sample size N is fixed whilst the number of counts N' in the censored section is an observed variable, i.e. is known but not fixed from sample to sample. Again, in this section, we consider a more general situation of 'censoring' which we have termed 'grouped frequencies': we assume that certain of the sample frequencies have been 'pooled' or 'grouped'

so that only the group totals of these frequencies are known. This is illustrated in Example 2 below. It will be clear that 'censored samples' are covered as the special case when all 'groups' consist of single frequency groups except for one (fairly large) group in which all frequencies for the extreme x-values are pooled. The method of estimation is described in terms of the following example for which the computations are shown in Table 2 below.

Example 2.

Snedecor [1956, p. 478] gives for 106 litters of 8 pigs the frequencies* of litters with 0 or 1 or 2 males ($N_1 = 14$), with 3 or 4 or 5 males ($N_2 = 73$) and with 6 or 7 or 8 males ($N_3 = 19$). Assuming that the number of males per litter, x, follows a binomial distribution, to estimate the sex ratio of such a distribution of pigs, we require the following notation: The variate x is capable of attaining integral values which are arranged in G groups, $g = 1, 2, \cdots G$. The variate values in each group are numbered $j = 1, 2, \cdots$, and

$$f(j, g; \theta) = \Pr\{x = j\text{-th integer in the } g\text{-th group}\} \tag{12}$$

$$F(g; \theta) = \sum_j f(j, g; \theta) = \Pr\{x \text{ in } g\text{-th group}\} \tag{13}$$

$$N_g = \text{total observed frequency for } g\text{-th group} \tag{14}$$

$${}_cn_{jg} = c\text{-th estimate of } j\text{-th frequency in } g\text{-th group.} \tag{15}$$

The procedure of obtaining the maximum likelihood solution is now as follows:

(a_0) From an inspection of the group frequencies N_g estimate their partitions into individual frequencies ${}_0n_{jg}$. In the example we split $N_1 = 14$ into ${}_0n_{11} = 1$, ${}_0n_{12} = 4$, ${}_0n_{13} = 9$ and likewise $N_2 = 73$ into 23, 26, 24 and $N_3 = 19$ into 12, 5, 2. This split is guided by the knowledge that the distribution is binomial, but very rough guesses are quite serviceable here.

(b_0) Using the ${}_0n_{jg}$ compute an estimate of the θ_t from the maximum likelihood equations applicable to a complete sample. In the example

$${}_1\theta = \frac{1}{nN}\sum_{jg} x_0 n_{jg} = \frac{1}{8 \times 106}(22 + 293 + 123) = .516.$$

*Actually Snedecor gives the complete set of frequencies which are here grouped as above for purposes of illustration.

TABLE 2

ESTIMATION OF BINOMIAL PROPORTION FROM GROUPED DATA

x	g	j	${}_0n_{1j}$	${}_0n_{2j}$	${}_0n_{3j}$	${}_1n_{1j}$	${}_1n_{2j}$	${}_1n_{3j}$	$f(jg; {}_1\theta)$
0	1	1	1			.34			.0030
1		2	4			2.89			.0257
2		3	9			10.77			.0958
3	2	1		23			21.04		.2043
4		2		26			28.04		.2723
5		3		24			23.92		.2323
6	3	1			12			14.13	.1238
7		2			5			4.30	.0377
8		3			2			.57	.0050
	N_g		14	73	19	14	73	19	1.0000
	$\sum_j x\, n_{jg}$		22	293	123	24.43	294.88	119.44	From table of binomial distribution
			${}_1\theta = 0.516$			${}_2\theta = .51742$			

(a_1) Using the initial estimate ${}_1\theta$ compute improved values of the individual frequencies from ${}_1n_{jg} = N_g\, f(j, g; {}_1\theta)/F(g; {}_1\theta)$. In Example 2 the binomial frequencies $f(x; \theta) = \binom{8}{x} {}_1\theta^x (1 - {}_1\theta)^{8-x}$ can either be obtained from tables or by direct computation, so that we obtain ${}_1n_{11} = .34$, ${}_1n_{12} = 2.89$, ${}_1n_{13} = 10.77$ and the remaining ${}_1n_{gj}$ as shown in Table 2.

(b_1) Using the ${}_1n_{gj}$ compute an improved estimate ${}_2\theta$ from the maximum likelihood equations for complete samples. In Example 2 ${}_2\theta = (24.43 + 294.88 + 119.44)/(8 \times 106) = 0.5174$.

A further cycle not shown in the table yields the maximum likelihood solution ${}_3\theta$ equal to 3 decimal accuracy. The proof that the above procedure yields the maximum likelihood solution to the problem is almost identical with that in Section 2. The maximum likelihood equations for the grouped distribution are

$$\sum_g N_g F_t(g, \theta)/F(g, \theta) = 0 \tag{16}$$

where the subscript t denotes differentiation with regard to θ_t. Splitting

up the N_g by the proportional allocation

$$n_{ig} = N_g f(j, g; \theta)/F(g, \theta) \tag{17}$$

and noting that (13) implies

$$F_t(g, \theta) = \sum_i f_t(j, g; \theta) \tag{18}$$

we may rewrite (16) as

$$\sum_g \sum_i n_{ig} f_t(j, g; \theta)/f(j, g; \theta) = 0. \tag{19}$$

But equation (19) is formally identical with the maximum likelihood equations for a 'complete sample' of $N = \sum_{ig} n_{ig}$ observed frequencies. Thus the iteration process $a_0 b_0 a_1 b_1 \cdots$, if it converges, will yield a solution of (17) and (19) and hence of the maximum likelihood equations (16).

We conclude this section by giving an example of a fairly drastically censored distribution to illustrate that the convergence of the procedure may still be reasonable when more than half of the distribution is 'censored' (i.e. pooled in one group).

Example 3.

Bliss [1953] quoting data of Jones, Mollison, and Quenouille [1948] on microscopic counts of soil bacteria gives (Table 4, p. 188) the following distribution of number of colonies per field:

x = colonies per field	0	1	2	3	4	5	6+	Total
n_x = frequency	11	37	64	55	37	24	12	240.

In this example only the extreme tail of the distribution $x \geq 6$ is 'pooled' for the χ^2-test, but the Poisson distribution was apparently fitted to the complete data. We repeat here the fit of the Poisson distribution drastically censored at $x = 3$, i.e. carry out the estimation of the Poisson parameter for the distribution

x = colonies per field	0	1	2	3+	Total
N_g = group frequency	11	37	64	128	240.

The work is set out in Table 3. There are 4 'groups,' the first 3($g = 1, 2, 3$) consisting of the single x values of $x = 0$, $x = 1$, $x = 2$ respectively and the fourth ($g = 4$) of the tail $x \geq 3$. The initial estimates ${}_0n_{4i}$ were originally taken as 50, 40, 20, 10, 5, 1, 0; seeing that this total is still short by 2 of the required total of $N_4 = 128$, the largest frequency was increased to 52. This turned out to be a rather lucky 'guess' as

TABLE 3

MAXIMUM LIKELIHOOD FIT TO POISSON WITH CENSORED TAIL-SUM

x	g	j	N_g	${}_0n_{4j}$	${}_1n_{4j}$	${}_2n_{4j}$	$f(j,g;\ {}_1\theta)$	$f(j,g;\ {}_2\theta)$	
0	1	1	11						
1	2	1	37						
2	3	1	64						
3		1		52	52.65	52.38	.2230	.2233	
4		2		40	37.54	37.51	.1590	.1599	
5		3		20	21.39	21.49	.0906	.0916	From
6	4	4		10	10.18	10.27	.0431	.0438	table
7		5		5	4.16	4.20	.0176	.0179	of
8		6		1	1.46	1.50	.0062	.0064	Poisson
9		7		0	.47	.49	.0020	.0021	distribution
10		8			.12	.14	.0005	.0006	
					.02	.02	.0001	.0001	
Totals			112	128	127.99	128.00	.5421	.5457	
$\sum x\,n$			165	519	522.59	523.68			
${}_c\theta$				2.850	2.8650	2.8695			

the convergence is established after 2 steps. In Section 5 we shall estimate the variance of our estimate as .0155 [see equation (30)]. Finally, attention is drawn here to Bliss [1948] dealing with this special case of a Poisson distribution with a single censored tail frequency.

It is well known that the convergence of an iteration can be accelerated by the judicious choice of starting values and in many situations special short-cut methods are advocated for their computation. Nevertheless, even if these precautions are taken, convergence may be slow in certain problems. We shall therefore describe here a process of accelerating the convergence of the present iteration process known to computers under the name of 'approach to the geometric limit.' This method is usually found to be effective when the convergence of the iteration process is approximately like that of a geometric series. We shall illustrate the use of this procedure for two purposes:

(i) The conversion of poor starting values to much improved values.

(ii) The direct computation of the maximum likelihood estimates in a poorly converging process.

(i) Suppose that in Example 1 we had made a rather poor choice of starting values as ${}_0n_0 = 15$, ${}_0n_1 = 30$, shown in Table 4 below. Carrying out the cycles $(a_0)(b_0)$, $(a_1)(b_1)$ and (a_2) as described

TABLE 4

MAXIMUM LIKELIHOOD FIT TO POISSON DISTRIBUTION WITH MISSING FREQUENCIES (EXAMPLE 1, TABLE 1, REPEATED WITH POOR STARTING VALUES)

j	n_i	${}_0n_j$	${}_1n_j$	${}_2n_j$	${}_3n_j$	${}_4n_j$	${}_5n_j$	$f(j, {}_1\theta)$	$f(j, {}_2\theta)$	$f(j, {}_3\theta)$	$f(j, {}_4\theta)$	$f(j, {}_5\theta)$
0		15	8.86	6.48	4.44	4.60	4.667	.0812	.0633	.0462	.0476	.0482
1		30	22.22	17.88	13.66	14.01	14.115	.2037	.1747	.1412	.1450	.1462
		See Table 1										
	$n = 78$	${}_cn' = 45$	31.08	24.36	18.10	18.61	18.822	.2849	.2380	.1883	.1926	.1944
	$\sum in_i = 279$	${}_c\theta = 2.512$	2.761	2.900	3.045	3.033	3.028					
$\delta\theta$			.249	.139		.012	.005					
θ^*				3.075			3.024					

in Section 1 but carrying fewer figures we reach the following three consecutive values ${}_1\theta = 2.512$, ${}_2\theta = 2.761$, ${}_3\theta = 2.900$ with first differences $\delta_{1(1/2)} = 2.761 - 2.512 = .249$, $\delta_{2(1/2)} = 2.900 - 2.761 = .139$. The ratio, q, of the second to the first difference is given by

$$q = \delta_{2(1/2)}/\delta_{1(1/2)} = .139/.249 = .558 \tag{20}$$

and, if we assume that a continuation of the process would generate a geometric series with this value of q, we would expect to reach as a limit

$$\theta^* = {}_3\theta + \delta_{2(1/2)}q/(1 - q) = 2.900 + (.139)(.558)/.442 = 3.075. \tag{21}$$

Since the geometric progression of the differences δ cannot be relied upon with certainty it will be necessary to start another cycle of the iteration with a starting value of $\theta^* = 3.075$ and to carry this until either a limit is reached or until after $2\frac{1}{2}$ cycles the final value can be computed by a second geometric-series projection. In Table 4 the value of ${}_5\theta = 3.028$ is reached and a geometric projection from here yields 3.024. This example illustrates that even with poor starting values two sets of $2\frac{1}{2}$ cycles may yield the limit.

(ii) If good starting values were chosen but the convergence of the end figures is tedious (as in Table 3) one would proceed to the geometric limit after $2\frac{1}{2}$ cycles of the iteration process would have have yielded ${}_1\theta$, ${}_2\theta$ and ${}_3\theta$. In Table 1 we would obtain immediately $\delta_{1(1/2)} = -.016$, $\delta_{2(1/2)} = -.007$; $q = .44$ and $\theta^* = 3.029 - .007(.44)/.56 = 3.023$. The method should be used with caution for purpose (i) and where convenient short-cut methods are available for the computation of initial values, these will be usually preferable. The method is, however, rather effective for (ii) particularly for improving the end figures of the maximum likelihood estimator. (See the 3 computations in Table 5, Section 6.)

5. Computation of variance estimates for the maximum likelihood estimators

Variance and covariance estimates of maximum likelihood estimators ${}_t\hat{\theta}$ are usually computed from the expectations of the second derivatives $E[L_{t\tau}(\theta)]$ of the likelihood function $L(\theta)$ by substituting the estimates ${}_t\hat{\theta}$. This computation is often facilitated by special aid tables of the elements of the information matrix and such tables have in fact been prepared for certain isolated cases of maximum likelihood estimation from incomplete data. Since situations of 'incomplete data' are very varied it

does not appear to be practical to provide such aid tables for all such situations. It is therefore preferable to use in these situations a numerical method of estimation based on finite difference calculus and closely related to one sometimes used by R. A. Fisher [see e.g. 1953]. This appears to be particularly suited to situations in which the maximum likelihood equations themselves must be solved by iterative methods requiring repeated evaluation of the first derivatives L_t of the empirical likelihood function (called 'scores' by R. A. Fisher). This method can be adapted so that it can be used with the present iterative method described in the above sections and it will be shown that these iterative computations provide, with little extra work, estimates of the variances and covariances.

We first deal with problems depending on a single parameter θ and in the next section with the case of two parameters. Starting with the case of missing frequencies of Section 2 we note that the first derivative(s) of the likelihood function $L_t(\theta)$ with regard to θ_t are given by the left-hand sides of (9) and hence, using auxiliary quantities defined by (10), by the left-hand sides of (11). When the maximum likelihood estimates $\theta = \hat{\theta}$ are substituted for θ we have of course $L_t(\hat{\theta}) = 0$, but when a preliminary estimate (say ${}_1\theta$) is substituted we shall obtain a value $L_t({}_1\theta)$ which will be $\neq 0$ in all situations in which the maximum likelihood equation has a unique solution. We can therefore compute an estimate of $L_{\theta\theta}(\hat{\theta})$ from the first order divided difference

$$\hat{L}_{\theta\theta} = L_\theta({}_c\theta)/({}_c\theta - \hat{\theta}) \tag{22}$$

a quantity called 'rate of change of score' by Fisher. Higher order divided differences may, of course, be used for checking and adjusting this computation but, since it is an estimate only, the resulting differences are usually small compared with the error involved in the estimate.

As a check it will usually be wise to compute $\hat{L}_{\theta\theta}$ from (22) twice using two of the trial values ${}_c\theta$ and compare the results. Finally, we have an estimate of variance in the form

$$\widehat{\mathrm{Var}}\ (\hat{\theta}) = -1/\hat{L}_{\theta\theta}\ . \tag{23}$$

It should be noted that this estimate of variance introduced by Fisher differs from one used on other occasions, viz.

$$\widehat{\widehat{\mathrm{Var}}}\ (\hat{\theta}) = -1/L_{\theta\theta}(\hat{\theta}). \tag{23'}$$

In the latter formula (23') the maximum likelihood estimate $\hat{\theta}$ is substituted in a mathematical formula for $1/L_{\theta\theta}$, whilst in the former formula (23) $\hat{\theta}$ and a neighboring value ${}_c\theta$ are substituted in the formula for $L_{\hat{\theta}}$. The second differential $L_{\theta\theta}$ is then estimated by the finite difference formula (22).

The writer is not aware of any theoretical small sample work on the comparison of the relative efficiency of these estimators, either for complete or incomplete samples.

To illustrate the use of (22) we apply it to our examples. For the Poisson distribution of Example 1 (see Table 1) we have for the 'score'

$$\begin{aligned} L_\theta({}_1\theta) &= (n + {}_1n) - (\sum_i in_i + \sum j_1 n_j)/{}_1\theta \\ &= 96.499 - (279 + 13.935)/3.052 = .518 \end{aligned} \tag{24}$$

and hence

$$-\hat{L}_{\theta\theta} = (.518)/.026 = 20$$

or

$$\widehat{\text{Var}}\ \hat{\theta} = .050.$$

A check value computed from ${}_1\theta = 2.512$ in Table 4 yields $\widehat{\text{Var}}\ \hat{\theta} = .047$. For an estimate $\hat{\theta}$ for a complete sample of 96 counts we would have a variance estimate of $\hat{\theta}/96 = .031$. Turning now to the case of pooled frequencies in Section 3 it is easy to show by an argument almost identical to the one given above that equations (22) and (23) still hold for this problem.

In Example 2 we have for the binomial distribution fitted to grouped data

$$L_\theta(\theta) = \frac{1}{1-\theta}(\sum_{ji} n_{ji}/\theta) - nN \tag{25}$$

and hence we obtain with ${}_1\theta = .51600$ and $\hat{\theta} = .51742$

$$L_\theta({}_1\theta) = \frac{1}{.484}\left(\frac{438.77}{0.516}\right) - 8 \times 106 = 4.82 \tag{26}$$

and hence from (22)

$$\hat{L}_{\theta\theta}(\theta) = 4.82/(-.00142) = -339(0) \tag{27}$$

so that the estimated variance of $\hat{\theta}$ is

$$\widehat{\text{Var}}\ (\hat{\theta}) = -1/\hat{L}_{\theta\theta}(\theta) = .000295. \tag{28}$$

For a complete binomial sample of $N \times n = 8 \times 106$ pigs with sex ratio $\theta = 0.5176$ the variance formula is

$$(.5176)(.4824)/8 \times 106 = .000294. \tag{29}$$

Although both the above figures are estimates of variances only, they indicate that there is no great loss of information through grouping in

this example. In the heavily censored Poisson distribution of Example 3 we find for the trial value, ${}_1\theta = 2.850$.

$$\begin{aligned} L_\theta({}_1\theta) &= n - (\sum x_1 n_{\theta i} + \sum x N_\theta)/{}_1\theta \\ &= 240 - (522.59 + 165)/2.850 = -1.260, \\ \hat{L}_{\theta\theta} &= -1.260/.0195 = -64.6, \\ \widehat{\mathrm{Var}}\ (\hat{\theta}) &= -1/\hat{L}_{\theta\theta} = .0155. \end{aligned} \tag{30}$$

The estimated variance for a complete sample of size 240 from a Poisson with estimated mean $\hat{\theta} = 2.87$ is $2.87/240 = .0119$.

*6. An example of two parameter estimation with truncation**

As a final example we apply our method to a distribution depending on two parameters obtaining their estimates together with estimates of their variances and covariances. The example is the negative binomial distribution with missing 0-class.

Example 4.

Sampford [1955] quoting data of a chromosome breakage study by E. C. Ford gives the distribution of 32 cells with regard to the number of breaks per cell. Since the susceptible cells which do not show any breaks are not distinguishable from the cells not susceptible to breaks, the frequency of the zero-class is not available. We therefore have the truncated distribution shown in the third column of Table 5, and as given by Sampford. We now use these data to illustrate the fit of the negative binomial distribution

$$f(x; k, p) = \frac{(k + x - 1)!}{(k - 1)!x!} \frac{p^x}{(1+p)^{k+x}} \quad (x = 1, 2, \cdots) \tag{31}$$

with unknown parameters k and p. Bliss and Fisher [1953] have described a most convenient method of estimating k and p from a complete sample and this will form the basis of our procedure. For a complete sample of size $n + n_0 = N$ the first derivatives of the likelihood function (called scores by Fisher) are given by

$$K(k, p) = \frac{\partial L}{\partial k} = \sum_{x=0}^{\infty} \frac{A_x}{k + x} - (n + n_0) \ln (1 + p) \tag{32}$$

and

$$P(k, p) = \frac{\partial L}{\partial p} = \frac{n + n_0}{p(1 + p)} (\bar{x} - kp) \tag{33}$$

*This section requires a knowledge of elementary finite difference calculus.

where

$$A_x = \sum_{i=x+1}^{\infty} n_i \tag{34}$$

is the cumulative sum of the observed frequencies shown in the fourth column of Table 5 and $\bar{x}$ is the mean of the complete distribution of

TABLE 5

MAXIMUM LIKELIHOOD FIT TO NEGATIVE BINOMIAL WITH ZERO CLASS MISSING

Columns (1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	
x	i	j	n_i	A_x	${}_c n_0$	${}_c \delta n_0$	${}_c p = \bar{x}/k$	$f(o) = (1+p)^{-k}$	
0		0		32	40		4.583	.5637	
						1.344			
1	1		11	21	41.344		4.499	.5665	
						.474			
2	2		6	15	41.818				
					— — — — — — — — — — —				$k = \frac{1}{3}$
3	3		4	11		.258			
					— — — — — — — — — — —				
4	4		5	6	42.076		4.4549	.56808	
					42.088				
5	5		0	6					
					20		4.231	.4373	
7	7		0	5		4.869			
					24.869		3.869	.4532	
8	8		2	3		1.653			
					26.522		3.759		
9	9		1	2	— — — — — — — — — — —				$k = \frac{1}{2}$
						.850			
10	10		0	2	— — — — — — — — — — —				
					27.372		3.7055	.46100	
11	11		1	1	27.369				
12	12		0	1					
13	13		1	0	11		2.558	.2811	
						1.512			
					12.512		2.471	.2881	
$\sum in_i = 110$						.438			$k = 1$
					12.950				
					— — — — — — — — — — —				
						.179			
					— — — — — — — — — — —				
					13.129		2.4375	.29091	
					13.128				

$n + n_0$ frequencies. For the truncated problem we must find the roots k, p and n_0 of the equations $\partial L/\partial k = 0$, $\partial L/\partial p = 0$ and

$$n_0 = nf(0; k, p)/[1 - f(0; k, p)]. \tag{35}$$

We proceed to solve these equations as follows: three computationally convenient 'pivotal values' $k = 1$, $k = \frac{1}{2}$, $k = \frac{1}{3}$ are chosen and for each of these values only the two equations $\partial L/\partial p = 0$ and (35) are solved by our iterative procedure as shown in Table 5. We describe the method for $k = \frac{1}{3}$, note that the equation $\partial L/\partial p = 0$ is equivalent to

$$p = \bar{x}/k, \tag{36}$$

and choose a trial value ${}_0n_0 = 40$ as shown in column (5) to compute

$${}_1p = \bar{x}/k = \sum in_i/(n + {}_0n_0)k = 110/(32 + 40)\tfrac{1}{3} = 4.583 \tag{37}$$

shown in column (7). Next we compute the zero-class frequency for the negative binomial from the equation

$$f(0; k, {}_1p) = 1/(1 + {}_1p)^k = 1/\sqrt[3]{5.583} = .5637 \tag{38}$$

by using a table of cube roots. This is shown in column (8). Finally we obtain an improved value for n_0 as

$${}_1n_0 = nf(0; k, {}_1p)/[1 - f(0; k, {}_1p)] = 41.344 \tag{39}$$

which brings us back to column (5). After $2\frac{1}{2}$ cycles the geometric projection (see Section 4) of ${}_0n_0 = 40$, ${}_1n_0 = 41.344$, ${}_2n_0 = 41.818$ is computed from

$${}_3n_0 = {}_2n_0 + \delta_{1(1/2)}q/(1 - q) = 42.076 \tag{40}$$

with $q = \delta_{1(1/2)}/\delta_{1/2} = .474/1.344$ obtained from the first differences of the ${}_cn_0$ shown in column (6). The projected value ${}_3n_0 = 42.076$ is then confirmed by one final cycle of the computations. This process is repeated for $k = \frac{1}{2}$ and $k = 1$ and yields the three maximum likelihood estimates $\hat{p}(k)$ of p for the three pivotal values of $k = \frac{1}{3}, \frac{1}{2}, 1$, viz.

$$\hat{p}(\tfrac{1}{3}) = 4.4549, \qquad \hat{p}(\tfrac{1}{2}) = 3.7055, \qquad \hat{p}(1) = 2.4375. \tag{41}$$

These are set out in Table 6 together with their differences along with k and the auxiliary, equidistant tabular argument $u = 1/k$. Also shown in Table 6 are the three scores $\partial L/\partial k = K[k, \hat{p}(k)]$ computed from (32). We are now ready to find the final maximum likelihood solutions k and p. To do this we must find the value of k for which $\partial L/\partial k = 0$. We find by inverse interpolation in the table of K that $\hat{u} = 1/\hat{k} = 2.0287$ and hence

$$\hat{k} = .4929 \tag{42}$$

TABLE 6

MAXIMUM LIKELIHOOD FIT TO NEGATIVE BINOMIAL (CONTINUED)
FINAL SOLUTION AND ESTIMATION OF VARIANCES AND COVARIANCES

k	$u = 1/k$	$\hat{p}(k)$	$\delta' p$	$\delta'' p$	$\sum_x A_x/(k+x)$	$(n+\hat{n})\cdot\ln(1+p)$	$K = \partial L/\partial k$	$\delta' K$	$\delta'' K$
1	1	2.4375			54.6647	55.7213	−1.0566		
	1.5		1.2680					1.0327	
½	2	3.7055		−.5196	91.9239	91.9466	−.0239		.3044
	2.5		.7494					.7287	
⅓	3	4.4549			126.3966	125.6918	+.7048		

and by direct interpolation in the table of $\hat{p}$

$$\hat{p} = 3.730. \tag{43}$$

These solutions agree with Sampford's (p. 68) obtained by a method requiring a special aid table. Estimates of the variances and covariances of $\hat{k}$ and $\hat{p}$ can now be obtained immediately from Table 6. Since the variances and covariances are computed from the elements of the inverse of the information matrix

$$-\begin{bmatrix} \frac{\partial^2 L}{\partial k^2} & \frac{\partial^2 L}{\partial k\, \partial p} \\ \frac{\partial^2 L}{\partial k\, \partial p} & \frac{\partial^2 L}{\partial p^2} \end{bmatrix} \tag{44}$$

they can be directly estimated as the partial derivatives of k and p with regard to the scores $K = \partial L/\partial k$ and $P = \partial L/\partial p$. We have

$$\widehat{\text{var}}\ \hat{k} = -\frac{\partial k}{\partial K}, \tag{45}$$

$$\widehat{\text{var}}\ \hat{p} = -\frac{\partial p}{\partial P}, \tag{46}$$

$$\widehat{\text{cov}}\ (\hat{k}, \hat{p}) = -\frac{\partial k}{\partial P} = -\frac{\partial p}{\partial K}. \tag{47}$$

A method analogous to equation (45) has, in fact, been used by Bliss and Fisher for the variance estimation from the complete sample. The partial derivatives $\partial k/\partial K$ and $\partial p/\partial K$ can now be estimated immediately from Table 6. Since the $\hat{p}(k)$ satisfy the equation $P = \partial L/\partial p = 0$ the table of the score K as a function of k represents a relation for which P is held constant at $P = 0$. Hence the partial $\partial k/\partial K$ can be directly computed from the differences of this table as follows:

$$\left.\begin{aligned} \frac{\partial K}{\partial u}\ (\text{at}\ \ u = 1.5) &= 1.0327 \\ \frac{\partial K}{\partial u}\ (\text{at}\ \ u = 2.5) &= .7287 \end{aligned}\right\} \text{read off from the differences } \delta' K \text{ in Table 6.}$$
$$\left.\frac{\partial K}{\partial u}\ (\text{at}\ \ u = 2.0287) = .872\right\} \text{by linear interpolation.} \tag{48}$$

Hence

$$\frac{\partial k}{\partial K}\ (\text{at}\ \ u = 2.0287) = -(\hat{k})^2 \Big/ \frac{\partial K}{\partial u} = (.4929)^2/.872 = -.278. \tag{49}$$

Likewise

$$\frac{\partial p}{\partial K}(\text{at } u = 2.0287) = \frac{\partial p}{\partial u}(\text{at } u = 2.0287) \Big/ \frac{\partial K}{\partial u} \tag{50}$$

$$= .994/.872 = 1.14.$$

Finally we obtain the partial $\partial p/\partial P$ by using the section for $k = \frac{1}{2}$ in Table 5. The difference between the trial value ${}_2p = 3.869$ and the final value $\hat{p}(\frac{1}{2}) = 3.7055$ can be approximately represented by the first order Taylor expansion

$${}_1p - p(\tfrac{1}{2}) = \frac{\partial p}{\partial P}\Delta P + \frac{\partial p}{\partial K}\Delta K. \tag{51}$$

Here

$$\Delta P = \frac{\partial L}{\partial p}(k = \tfrac{1}{2}, p = {}_2p) - 0 \tag{52}$$

so that, using (33), we have

$$\Delta P = \frac{32 + 26.522}{(3.869)(4.869)}(3.759 - 3.869) = -.171. \tag{53}$$

Likewise we find

$$\Delta K = \frac{\partial L}{\partial k}(k = \tfrac{1}{2}, p = {}_2p) - \frac{\partial L}{\partial k}[k = \tfrac{1}{2}, p = \hat{p}(\tfrac{1}{2})] \tag{54}$$

or, using (32), we write

$$\begin{aligned}\Delta K &= -(32 + 26.522)\ln 4.869 + (32 + 27.369)\ln 4.7055 \\ &= -92.6333 + 91.9466 \\ &= -.687.\end{aligned} \tag{55}$$

Therefore equation (51) can be written as

$$.1635 = -\frac{\partial p}{\partial P}(.171) = \frac{\partial p}{\partial K}(.687). \tag{56}$$

Equation (56) is a relation between $\partial p/\partial P$ and $\partial p/\partial K$ at $P = 0$, $K = -.0239$. Using the same relation* at $P = 0$, $K = 0$ we obtain $\partial p/\partial P$ by substituting the value of $\partial p/\partial K = 1.14$ from (50). We therefore obtain

$$\frac{\partial p}{\partial P} = -5.52 \tag{57}$$

so that the variance-covariance estimates are given by

$$\widehat{\text{var}}\ \hat{k} = .278, \quad \widehat{\text{cov}}\ \hat{k}\hat{p} = -1.14, \quad \widehat{\text{var}}\ \hat{p} = 5.52. \tag{58}$$

*In the present example $\hat{k} = .4929$ is very close to the pivotal value of $k = \frac{1}{2}$; in general this relation would have to be obtained by interpolation between two corresponding relations set up at the two pivotal k values neighboring $\hat{k}$.

Sampford, using the rather complex formulas for the expectation of the information matrix [44, p. 65] and inverting this matrix reaches values [p. 68] equivalent to

$$\widehat{\text{var}}\ \hat{k} = .276, \quad \widehat{\text{cov}}\ \hat{k}\hat{p} = -1.06, \quad \widehat{\text{var}}\ \hat{p} = 4.95. \qquad (59)$$

In comparing (58) and (59) it should be remembered that these are values resulting from different methods of estimation.

Two final remarks are appropriate:

The pivotal values of k should be simple integers or simple fractions so as to facilitate the evaluation of $(1 + p)^k$. They should also be so chosen that they cover the maximum likelihood root $\hat{k}$. To guide this choice some may consider it necessary first to compute a trial k by a short-cut method. This may, however, not be necessary as the iterative process in Table 5 is fast. One would therefore normally compute $\hat{p}(k)$ for $k = 1$ and then judge from the sign of $\partial L/\partial K$ (required in Table 6) whether one should use further pivotal values of $k > 1$ or $k < 1$.

The method described in the above example of the negative binomial is directly applicable to other two parameter situations. One would choose the simpler of the two maximum likelihood equations and solve it for the more convenient parameter ($\partial L/\partial p = 0$ is solved for p in the above example). This would be done for (say 3) selected pivotal values of the other parameter (k in the above example). Such a procedure would automatically produce a table of the second score ($\partial L/\partial K$) for such values for which the first score ($\partial L/\partial p$) is zero. When this procedure is used the simple method of variance-covariance estimation is applicable.

REFERENCES

Bliss, C. I. and Fisher, R. A. [1953]. Fitting the negative binomial distribution to biological data; note on the efficient fitting of the negative binomial. *Biometrics 9*, 176–196; 197–9.

Bliss, C. I. [1948]. Estimation of the mean and its error from incomplete Poisson distributions. *Bulletin 513*, Connecticut Agricultural Station, New Haven, Conn.

David, F. N. and Johnson, N. L. [1952]. The truncated Poisson. *Biometrics 8*, 275.

Finney, D. J. [1949]. The truncated binomial distribution. *Annals of Eugenics 14*, 319.

Jones, P. C. T., Mollison, J. E., and Quenouille, M. H. [1948]. A technique for the quantitative estimation of soil microorganisms. *J. Gen. Microbiology 2*, 54–69.

Moore, P. G. [1952]. The estimation of the Poisson parameter from a truncated distribution. *Biometrika 39*, 247–251.

Sampford, M. R. [1955]. The truncated negative binomial distribution. *Biometrika 42*, 58–69.

Snedecor, G. W. [1956]. *Statistical Methods*. 5th ed., The Iowa State College Press, Ames, Iowa.