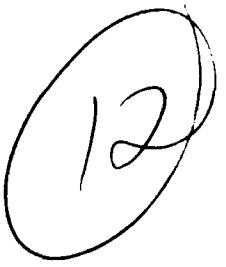
(12)

DEPARTMENT OF OPERATIONS RESEARCH

COLLEGE OF ENGINEERING
CORNELL UNIVERSITY
ITHACA, NEW YORK 14853

SCHOOL OF OPERATIONS RESEARCH
AND INDUSTRIAL ENGINEERING
COLLEGE OF ENGINEERING
CORNELL UNIVERSITY
ITHACA, NEW YORK



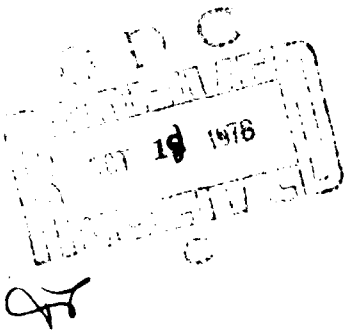
TECHNICAL REPORT NO. 305

July 1976

THE EMPIRICAL DISTRIBUTION FUNCTION WITH ARBITRARILY GROUPED, CENSORED, AND TRUNCATED DATA

by

Bruce W. Turnbull



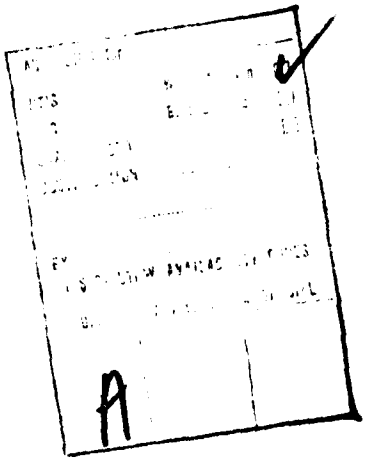
Prepared under contracts

DAHC04-73-C-0008, U.S. Army Research Office - Durham,

N00014-75-C-0586, Office of Naval Research.

THE FINDINGS IN THIS REPORT ARE NOT TO BE CONSTRUED AS AN OFFICIAL DEPARTMENT OF THE ARMY POSITION, UNLESS SO DESIGNATED BY OTHER AUTHORIZED DOCUMENTS.

Table of Contents



Abstract

This paper is concerned with the nonparametric estimation of a distribution function F, when the data are incomplete due to grouping, censoring and/or truncation. Subsets $B_1, B_2, \ldots, B_N$ of the real line are given and there are N independent observations $X_1, X_2, \ldots, X_N$, where X_i is drawn from the truncated distribution $F(x; B_i) = P(X \leq x \mid X \in B_i)$, $x \in B_i$. However X_i may not be observed exactly and is known only to lie in the set $A_i \subseteq B_i$. The situation occurs frequently in survivorship, reliability, and recidivism analysis. Using the idea of self-consistency, a simple algorithm is constructed and shown to converge monotonically to yield a maximum likelihood estimate of F. The procedure compares favourably with the more cumbersome Newton-Raphson method. A test is proposed for comparing two distributions when data on one or both is incomplete and some other applications of the empirical distribution function are indicated.

Keywords: EMPIRICAL DISTRIBUTION FUNCTION; CENSORING; INTERVAL CENSORING, TRUNCATION, GROUPING; SURVIVAL CURVE; MAXIMUM LIKELIHOOD; SELF-CONSISTENCY; NEWTON-RAPHSON; MULTINOMIAL DISTRIBUTION; TWO SAMPLE TEST; LOGRANK TEST; LEHMANN ALTERNATIVES

AMS 1970 subject classifications. Primary 62G05; Secondary 65K05.

I. INTRODUCTION

In this paper we will be mainly concerned with the nonparametric estimation of the distribution F of a real valued random variable X, when the sample data are incomplete due to restricted observation brought about by grouping, censoring and/or truncation. More precisely the situation is as follows. Subsets $B_1, B_2, \ldots, B_N$ of the real line are given and there are N independent observations $X_1 = x_1, \ldots, X_N = x_N$, where X_i $(1 \leq i \leq N)$ is drawn from the truncated distribution $F(x;B_i) = P(X \leq x | X \in B_i)$, $x \in B_i$. Thus X_i is truncated by B_i or, in other words, the experimenter would not have been aware of the existence of that observation had X_i not belonged to B_i. Moreover X_i $(1 \leq i \leq N)$ may not be observed exactly and is known only to lie in the set A_i where $A_i \subseteq B_i$. Thus X_i is censored into the set A_i. Grouped data can be naturally considered as censored, where each observation is censored into one of a fixed collection of disjoint sets. The observed data are then the N pairs $(A_1,B_1), (A_2,B_2), \ldots, (A_N,B_N)$.

The truncating sets $\{B_i\}$ can either be viewed as fixed or as random. We can now think of a partition of the set B_i and A_i is that member of the partition into which X_i falls. Again the partition can be viewed either as fixed or as having arisen from some random mechanism independent of X_i. In many cases, the partition of B_i will be unknown (except for the fact that A_i belongs to it); these assumptions will make knowledge of the partition irrelevant to the estimation of F. The case of grouped data can be considered as one in which the partitions are known and are the same for each i $(1 \leq i \leq N)$.

If $B_i = (-\infty,\infty)$ then X_i is not truncated, and if A_i consists of a single point then X_i is uncensored, i.e. is exact. We say that X_i is interval censored if A_i is of the form $[L_i,R_i]$ and X_i is right (left)

censored if $R_i = +\infty$ $(L_i = -\infty)$. Of course if $L_i = R_i$, then X_i is exact. Interval, right and left truncation are defined similarly. A sample is said to be singly censored if all the data is either exact or right censored; and doubly censored if the data is all either exact, right censored or left censored. This is now standard terminology.

Examples of right censoring are very common e.g. in medical follow-up and industrial life-testing situations. Interval censoring occurs naturally when the $\{X_i\}$ represent response times. Let us suppose that periodic inspections are made at times $t_1 < t_2 < \ldots < t_m$ in order to see whether a certain event has yet happened. If it has already occurred by the first inspection, the observation is left censored in $(-\infty, t_1]$. If the response is first observed to have occurred at the k'th inspection $(2 \leq k \leq m)$ then the observation is censored into $(t_{k-1}, t_k]$, while if the event has still not happened by the last inspection, the observation is right censored in (t_m, ∞). Examples of interval censoring are, for instance, described in Harris, Meier and Tukey (1950), Cohen (1957), Hartley (1958), Gehan (1965), Peto (1973) and Turnbull (1974). Also the bioassay problem discussed by Ayer et al (1955) can be considered an extreme case of double censoring when there is no exact data. The same situation arises in the estimation of gap acceptance distributions in traffic studies (see Miller, A.J. (1974).).

In most practical situations, each set A_i will be an interval or a point. However the problem is not made much more complex if we allow the $\{A_i\}$ to be unions of intervals and points. This could arise if, in grouped data, non-adjacent groups had been pooled; for example, readings off the scale of the measuring instrument whether too high or too low might have been pooled. This more general type of censoring pattern has also been considered by Mantel (1967).

Truncation can occur if the population from which X_i is drawn has been

subject to some screening procedure in which all items with x-values outside B_i have been removed. This situation can arise in consumer product testing, for example. If several data sets have been pooled to produce the sample then the $\{B_i\}$ will not necessarily all be identical. Another example of truncation occurs when the instrument which is measuring X needs a certain minimum level before it will respond at all.

Concerning survivorship analysis, Mantel (1966) mentions left truncation in the context of merging clinical trials. Here a group of survivors at a certain point in time is to be incorporated into ongoing study data when the original size of the group of which these are a remnant is unknown. The reentry problem, also suggested by Mantel, is an example where there can be a more general truncation pattern. This situation occurs when a person can be lost to follow-up, by leaving a health insurance programme for instance, but then he rejoins at a later date. If he had died in the intervening interval we would not have been aware of it. Here B_i is of the form $(-\infty,b_1] \cup [b_2,\infty)$, but one could envisage a more general situation where a person could enter or leave the programme several times. Of course, with some effort, we may uncover information about an individual who might otherwise be lost. However, not only will this be expensive but it could also introduce a bias if the success of the search is influenced by whether or not death has occurred. Thus an unbiased incomplete (truncated) sample may be preferable to complete but biased data. (Another difficulty is to ensure that the person has not rejoined because his health has deteriorated. This would violate our assumptions. We might refer to this situation as "prognostic truncation".)

The problem of the estimation of F when some parametric form for F is assumed has been treated extensively in the literature. Early work has been summarized by Buckland (1964, Ch. 2). For example, the case when F is normal has been considered by Cohen (1957), while recently Selvin (1974) has examined

the Poisson case. Blight (1970) has developed a general method for obtaining the maximum likelihood estimates of the parameters for any distribution in a multiparameter exponential family. (See also Hartley and Hocking (1971) and Sundberg (1974).) Most authors have assumed that the sets $\{B_i\}$ are intervals, the same for each i, and that the observations are either exact or censored into one member of a fixed set comprising several disjoint intervals.

We shall be concerned with deriving the maximum likelihood estimate (MLE) of F when no parametric assumptions about its form are assumed. Of course, if all the data are exact with no truncation, this estimate is given by the empirical (sample) c.d.f. When the data is subject only to right censoring, which is common in survivorship and life-testing situations, Kaplan and Meier (1958) have shown that the MLE of F is given by the product limit (PL) method. This can be adapted to accommodate left truncation as well by treating such data as "negative losses" (see p. 463). Trivially, by reversing the scale, the PL method can be applied to data subject only to left censoring and right truncation. It can also be used in problems with no truncation and very special patterns of double censoring (Turnbull (1974, p. 170)) and of interval censoring (Peto (1973, p. 87)). Explicit estimates are also available for certain particular interval truncation patterns with no censoring (see Section 5).

For obtaining estimates in more general situations, explicit solutions of the likelihood equations are not available and iterative methods must be used. For interval censored data, Peto (1973) employed direct but rather cumbersome Newton-Raphson search methods to maximise the likelihood. Turnbull (1974) used the idea of self-consistency (cf. Efron (1967)) to obtain a simple algorithm in the doubly censored case.

The purpose of this paper is threefold. Firstly, a simple iterative procedure is proposed for finding the MLE of F for the general case of arbitrarily censored and truncated data. This method can be considered the nonparametric analogue of that of Blight (1970). Secondly a new method of proof of the equivalance of self-consistency and maximum likelihood is presented. This method utilizes the relation between the values of the likelihood and successive approximations of the estimates, giving at the same time some insight as to why the theorem should be true. The proof differs from the rather inelegant and lengthier arguments used previously for the easier special cases of single censoring (Efron (1967, Thm 7.1)) and of double censoring (Turnbull (1974)). Finally, the algorithm is shown to converge, and in a monotone fashion, a fact conjectured by Turnbull (1974) on the basis of empirical evidence.

In Section 2, the likelihood function is examined, and the problem reduced to a simpler one of estimating the parameters of a multinomial distribution with censoring and truncation. In Section 3, the self-consistency algorithm is described and, in the following section, is shown to converge to yield the MLE of F. In Section 5, properties of the algorithm are discussed and comparisons made with the Newton-Raphson method. A two sample test when one or both samples may be subject to censoring and truncation is proposed in Section 6. Finally, some further problems such as large sample properties of the estimates and the handling of concomitant variables are discussed.

2. REDUCTION OF THE PROBLEM

We first show that the maximum likelihood estimate, $\hat{F}$, of F increases in only a finite number of disjoint intervals (or points). We shall use the same notation as Peto (1973) who obtained a similar result for interval censoring with no truncation.

Let us assume that each A_i $(1 \leq i \leq N)$ can be expressed as the finite union of disjoint closed intervals, with the convention that an isolated point $\{x\}$ is a closed interval $[x,x]$ and that a semi-infinite interval is semi-closed only. Thus we can write

$$A_i = \bigcup_{j=1}^{k_i} [L_{ij},R_{ij}] \qquad (i=1,2,\ldots,N) ,$$

where $-\infty \leq L_{i1} \leq R_{i1} < L_{i2} \leq \ldots < L_{iK_i} \leq R_{ik_i} \leq \infty$ and $R_{i1} > -\infty$, $L_{ik_i} < \infty$. From a practical point of view, this restriction on the form of A_i, is unimportant. We now construct a set of disjoint intervals whose left and right end points lie in the set $\{L_{ij};\ 1 \leq j \leq k_i,\ 1 \leq i \leq N\}$ and $\{R_{ij},\ 1 \leq j \leq k_i,\ 1 \leq i \leq N\}$ respectively, and which contain no other members of $\{L_{ij}\}$ or $\{R_{ij}\}$ except at their end points. We write these intervals

$$[q_1,p_1],\ [q_2,p_2],\ldots,[q_m,p_m] ,$$

where $q_1 \leq p_1 < q_2 \leq \ldots < q_m \leq p_m$. Also define

$$C = \bigcup_{j=1}^{m} [q_j,p_j] . \qquad (2.1)$$

For example in the case of single censoring, we have $k_i = 1$ for all

i, and $L_{il} = R_{il}$ if x_i is exact while $R_{il} = +\infty$ if x_i is right censored. Let u $(1 \leq u \leq N)$ be defined by $L_{ul} = \max_i L_{il} < \infty$. If $L_{ul} = R_{ul}$ (i.e. the largest observation is uncensored), then m is the number of exact observations and $q_j = p_j$ is the value of the j'th largest exact observation. If $R_{ul} = +\infty$ (i.e. the largest observation is censored), then $m-1$ is the number of exact observations, the last interval $[q_m, p_m]$ is $[L_{ul}, \infty)$, and $q_j = p_j$ $(1 \leq j \leq m-1)$ are the values of the exact observations.

Under the assumptions of Section 1, the likelihood is proportional to

$$L^*(F) = \prod_{i=1}^{N} [P_F(A_i)/P_F(B_i)]$$

$$= \prod_{i=1}^{N} \{ \sum_{j=1}^{k_i} [F(R_{ij}+) - F(L_{ij}-)] \}/P_F(B_i). \quad (2.2)$$

We will assume that $P_F(\cup B_i) = 1$, which occurs for instance if at least one observation is not truncated. The search for that function F that maximises (2.2) is facilitated by the following lemmas.

Lemma 1.

Any c.d.f. which increases outside the set C cannot be a maximum likelihood estimate of F, except in the trivial case when $A_i \cap C = B_i \cap C$ for all i.

Proof. Recall that $A_i \subseteq B_i$ and C is defined by (2.1). Suppose that c.d.f. G assigns non-zero probability p to the set $A_i - C$ for some i. Then the likelihood can be strictly increased by "transferring" probability p from $A_i - C$ to $A_i \cap C$. Similarly if G assigns positive probability to a set $B_i - A_i - C$ or to $\bigcap_{i=1}^{N} B_i^c$, the likelihood can be improved by

"transferring" probability from these sets to C. This improvement is again strict except in the trivial case mentioned.

Remark. When $A_i \cap C = B_i \cap C$ for all i, the maximum likelihood is unity and is achieved by any distribution which assigns zero probability to $\bigcup_{i=1}^{N} B_i - C$. The situation represents one of severe censorship and truncation; we will exclude such cases from our further discussion.

Lemma 2.

For fixed values of $F(p_j+)$, $F(q_j-)$ $(1 \le j \le m)$, the likelihood is independent of the behaviour of F within each interval $[q_j,p_j]$.

The proof is obvious.

Now, for $1 \le j \le m$, define

$$s_j = F(p_j+) - F(q_j-). \tag{2.3}$$

Then the vectors $\underset{\sim}{s} = (s_1,\ldots,s_m)$, where $\sum s_j = 1$ and $s_j \ge 0$, define equivalence classes on the space of distribution functions F which are flat outside C. We will say that two such functions are equivalent if they have the same $\underset{\sim}{s}$-vectors, as defined by (2.3). All functions in the same equivalence class will have the same likelihood by Lemma 2, and Lemma 1 shows that we can restrict our search for an MLE to these classes. Therefore the MLE will, at best, be unique only up to equivalence defined in this way.

For example, for right censored data, the Kaplan-Meier PL estimate is undefined at the exact observation points and in an interval $[L,\infty)$, say, if the largest observation is at L and is censored. Of course one can obviate the ambiguity when $p_j = q_j$ by requiring F to be right continuous.

The foregoing discussion shows the problem of maximising (2.2) reduces to one of maximising

$$L^*(s_1,\ldots,s_m) = \prod_{i=1}^{N} \Big(\sum_{j=1}^{m} \alpha_{ij} s_j \Big/ \sum_{j=1}^{m} \beta_{ij} s_j \Big) , \qquad (2.4)$$

subject to $\sum s_j = 1$, $s_j \geq 0$ $(1 \leq j \leq m)$, where

$$\alpha_{ij} = \begin{cases} 1 & \text{if } [q_j,p_j] \subseteq A_i, \\ 0 & \text{otherwise} , \end{cases}$$

$$\beta_{ij} = \begin{cases} 1 & \text{if } [q_j,p_j] \subseteq B_i, \\ 0 & \text{otherwise} . \end{cases}$$

We remark that we would be able to write down (2.4) immediately as the likelihood if there were a discrete scale for X (i.e. X could only take on values $t_1,t_2,\ldots,t_m$, say). Then we would define $s_j = P(X = t_j)$. This was the situation in the double censoring problem considered by Turnbull (1974), in which it was required to estimate the probabilities that a certain response time fell in the first month, the second month, etc.

Now since $A_i \subseteq B_i$ for all i, we have that $\alpha_{ij} = 1$ implies $\beta_{ij} = 1$. Let $\hat{\underset{\sim}{s}} = (\hat{s}_1,\ldots,\hat{s}_m)$ denote a value of $\underset{\sim}{s}$ for which L^* attains its maximum in the region $R = \{\underset{\sim}{s} | \sum s_j = 1, s_j \geq 0 \ (1 \leq j \leq m)\}$. We assume that neither of the following two trivial situations hold:

(A) There exist j,k with $1 \leq j,k \leq m$ and $j \neq k$ such that $\alpha_{ij} = \alpha_{ik}$ for all i $(1 \leq i \leq N)$.

(B) There exists a subset D such that for each i, $1 \leq i \leq N$, either $B_i \cap C \subseteq D$ or $B_i \cap C \subseteq D^C$.

If (A) occurs, L^* depends on s_j and s_k only through their sum. In case (B) only the ratio $s_j/(\sum_{k\in D} s_k)$ is estimable for $j \in D$ and hence $\hat{s}_j$ is defined only up to a multiplicative constant. (Condition (B) modifies a result of Asano (1965, Thm. 5) concerning necessary and sufficient conditions for the estimability of multinomial probabilities with truncated data.)

If either (A) or (B) occurs, $\hat{s}$ is not unique and the maximum likelihood estimate $\hat{F}$ will be determined only as far as belonging to a certain union of equivalence classes.

3. THE SELF-CONSISTENCY ALGORITHM

In this section, we describe an algorithm for obtaining the MLE of $\underset{\sim}{s}$ based on the equivalence between the property of maximum likelihood and that of self-consistency. This latter property will be defined precisely below; it is an extension of the idea first used by Efron (1967) for right censored data and later by Turnbull (1974) for doubly censored data. The algorithm is related to one proposed by Hocking and Oxspring (1971) for multinomial data subject to censoring without truncation.

For $1 \leq i \leq N$, $1 \leq j \leq m$, let

$$I_{ij} = \begin{cases} 1 & \text{if } x_i \in [q_j, p_j] \\ 0 & \text{otherwise} \end{cases}$$

Because of the censoring the value of I_{ij} may not be known, however its expectation is given by

$$E_{\underset{\sim}{s}}[I_{ij}] = \alpha_{ij} s_j / \sum_{k=1}^{m} \alpha_{ik} s_k \qquad (3.1)$$

$$= \mu_{ij}(\underset{\sim}{s}) \text{ , say.}$$

Thus $\mu_{ij}(\underset{\sim}{s})$ represents the probability that the i'th observation lies in $[q_j, p_j]$ when F belongs to the equivalence class defined by $\underset{\sim}{s} = (s_1, \ldots, s_m)$. Also, because of the truncation, each observation $X_i = x_i$, can be considered a remnant of a group, the size of which is unknown and all (except the one observed) with x-values in B_i^c. (They can be thought of as X_i's "ghosts".) Let J_{ij} be the number in the group

corresponding to the i'th observation which have values in $[q_j, p_j]$. Of course J_{ij} is unknown but its expectation, under $\underset{\sim}{s}$, is given by

$$E_{\underset{\sim}{s}}(J_{ij}) = (1 - \beta_{ij})s_j / \sum_{k=1}^{m} \beta_{ik} s_k \qquad (3.2)$$

$$= \nu_{ij}(\underset{\sim}{s}) \ , \quad \text{say.}$$

If we treated (3.1), (3.2) as observed rather than expected frequencies, the proportion of cbservations in interval $[q_j, p_j]$ is

$$\sum_{i=1}^{N} [\mu_{ij}(\underset{\sim}{s}) + \nu_{ij}(\underset{\sim}{s})]/M(\underset{\sim}{s}) = \pi_j(\underset{\sim}{s}) \ , \qquad (3.3)$$

say, where

$$M(\underset{\sim}{s}) = \sum_{i=1}^{N} \sum_{j=1}^{m} [\mu_{ij}(\underset{\sim}{s}) + \nu_{ij}(\underset{\sim}{s})] \ .$$

Note that $M(s) \geq N$ with equality if there is no truncation for then $\nu_{ij} = 0$ for all i,j. We say that the vector of probabilities $\underset{\sim}{s}$ is <u>self-consistent</u> if

$$s_j = \pi_j(s_1, \ldots, s_m) \quad (1 \leq j \leq m) \ . \qquad (3.4)$$

A <u>self-consistent estimate</u> (s.c.e) of $\underset{\sim}{s}$ is defined to be any solution of the simultaneous equations (3.4). The form of (3.4) immediately suggests an iterative procedure for finding the solution.

A. Obtain initial estimates s_j^0 $(1 \leq j \leq m)$. This can be any set of

positive numbers summing to unity, e.g. $s_j = 1/m$ for all j.

B. Evaluate $\mu_{ij}(\underset{\sim}{s}^0)$ and $\nu_{ij}(\underset{\sim}{s}^0)$ for $1 \leq i \leq N$ and $1 \leq j \leq m$, and hence $M(\underset{\sim}{s}^0)$ and $\pi_j(\underset{\sim}{s}^0)$.

C. Obtain improved estimates s_j^1 by setting

$$s_j^1 = \pi_j(\underset{\sim}{s}^0) \quad \text{for} \quad 1 \leq j \leq m.$$

D. Return to Step B with $\underset{\sim}{s}^1$ replacing $\underset{\sim}{s}^0$, etc.

E. Stop when the required accuracy has been achieved.

(E.g. the rule may be to stop when $\max_{1\leq j\leq m}|s_j^k - s_j^{k-1}| < 0.0001$, say. Alternatively a stopping rule may be based on the difference between successive values of the likelihood.)

The procedure is easy to programme on a computer, requiring only simple operations. If any component of $\underset{\sim}{s}^k$ is small then it is possible for $M(\underset{\sim}{s}^k)$ to become very large. However, rounding errors can be avoided if the sequence of operations for computing the $\{\pi_j\}$ is chosen with care. Of course, the difficulty does not arise if there is no truncation for then $M(\underset{\sim}{s}^k)$ is always equal to N.

Another way to write $\pi_j(\underset{\sim}{s})$, which is useful if relatively few of the $\{A_i\}$ and $\{B_i\}$ are distinct, is

$$\pi_j(\underset{\sim}{s}) = \left[\sum_A \xi_A I_A(j) \frac{s_k}{\sum_{k\in A} s_k} + \sum_B \mu_B (1 - I_B(j)) \frac{s_k}{\sum_{k\in B} s_k} \right] \Big/ M(\underset{\sim}{s}), \tag{3.5}$$

where ξ_A is the number of observations censored into the set A, η_B is the number truncated by the set B, and $I_A(j)$ equals 1 if $[q_j, p_j] \subseteq A$ and is zero otherwise. Thus $\sum_A \xi_A = \sum_B \eta_B = N$ and using this relation $M(\underset{\sim}{s})$, which is the sum over j of the quantities in square brackets in (3.5), can be written more simply as

$$\sum_{B} [\eta_B (\sum_{k \in B} s_k)^{-1}].$$

Therefore the computations in Step 3 of the algorithm involve summing only over the number of distinct $\{A_i\}$ and distinct $\{B_i\}$ which may be considerably less than N in the case when X is discrete, for example.

In the next section we show that this algorithm converges and that self-consistent estimates also maximise the likelihood. In Section 5, the algorithm is discussed further and compared with the general Newton-Raphson method which has been suggested by several authors in connection with various special cases.

4. THE EQUIVALENCE AND CONVERGENCE THEOREM

We now examine the equivalence of the s.c.e. and the m.l.e. From (2.4), we see that the log-likelihood is given by

$$L(\underset{\sim}{s}) = \sum_{i=1}^{N} [\log(\sum_{j=1}^{m} (\alpha_{ij}s_j) - \log(\sum_{j=1}^{m} \beta_{ij}s_j)] \tag{4.1}$$

Consider the effect of increasing a particular component, s_j say, by a small positive amount ϵ and then dividing all the $\{s_k\}$, including $s_j + \epsilon$, by $1 + \epsilon$ in order to keep the sum equal to unity. We let $d_j(\underset{\sim}{s})$ denote the value of the derivative of L with respect to ϵ at $\epsilon = 0$. Therefore

$$d_j(\underset{\sim}{s}) = \frac{\partial}{\partial \epsilon} L \left(\frac{s_1}{1+\epsilon}, \ldots, \frac{s_j + \epsilon}{1+\epsilon}, \ldots, \frac{s_m}{1+\epsilon}\right)\Big|_{\epsilon=0}$$

$$= \frac{\partial L}{\partial s_j} - \sum_{k=1}^{m} s_k \frac{\partial L}{\partial s_k} \tag{4.2}$$

$$= \sum_{i=1}^{N} \left[\frac{\alpha_{ij}}{\sum_{k=1}^{m} \alpha_{ik}s_k} - \frac{\beta_{ij}}{\sum_{k=1}^{m} \beta_{ik}s_k} \right] \tag{4.3}$$

for $1 \leq j \leq m$. From (3.3), we have

$$\pi_j(\underset{\sim}{s}) = \frac{s_j}{M(\underset{\sim}{s})} \cdot \sum_{i=1}^{N} \left[\frac{\alpha_{ij}}{\sum_{k=1}^{m} \alpha_{ik}s_k} + \frac{1 - \beta_{ij}}{\sum_{k=1}^{m} \beta_{ik}s_k} \right]$$

$$= \frac{s_j}{M(\underset{\sim}{s})} \cdot \left[d_j(\underset{\sim}{s}) + \sum_{i=1}^{N} \left(\sum_{k=1}^{m} \beta_{ik}s_k \right)^{-1} \right], \tag{4.4}$$

where we have substituted for $d_j(\underset{\sim}{s})$ by (4.3). However

$$M(\underset{\sim}{s}) = \sum_{i=1}^{N} \sum_{j=1}^{m} \left[\frac{\alpha_{ij} s_j}{\sum_{k=1}^{m} \alpha_{ik} s_k} + \frac{(1 - \beta_{ij}) s_j}{\sum_{k=1}^{m} \beta_{ik} s_k} \right]$$

$$= \sum_{i=1}^{N} \left(\sum_{k=1}^{m} \beta_{ik} s_k \right)^{-1} .$$

Substituting in (4.4), we obtain

$$\pi_j(\underset{\sim}{s}) = \left(1 + \frac{d_j(\underset{\sim}{s})}{M(\underset{\sim}{s})}\right) s_j \qquad (1 \leq j \leq m). \tag{4.5}$$

Now a necessary and sufficient condition for $\underset{\sim}{s}$ to be an MLE is that for each i

$$\underline{\text{either}}\ d_j(\underset{\sim}{s}) = 0 \quad \underline{\text{or}} \quad d_j(\underset{\sim}{s}) \leq 0 \ \text{ with } \ s_j = 0. \tag{4.6}$$

Thus from (4.5) and (4.6), we see immediately that the MLE $\hat{\underset{\sim}{s}}$ satisfies $\pi_j(\hat{\underset{\sim}{s}}) = \hat{s}_j$ for all j, and hence is self-consistent.

Concerning convergence of the algorithm, we let $\underset{\sim}{s}$ and $\underset{\sim}{s}'$ be successive approximations where, by (4.5), $s_j' = [1 + (d_j(\underset{\sim}{s})/M(\underset{\sim}{s}))]s_j$ for $1 \leq j \leq m$. Now by a Taylor series expansion we have

$$L(\underset{\sim}{s}') - L(\underset{\sim}{s}) = \sum_{j=1}^{m} (s_j' - s_j) \frac{\partial L}{\partial s_j} + O(\| s' - s \|^2)$$

$$\approx \frac{1}{M(s)} \sum_{j=1}^{m} s_j d_j(\underset{\sim}{s}) \frac{\partial L}{\partial s_j}$$

$$= \frac{1}{M(\underset{\sim}{s})} \left[\sum_{j=1}^{m} s_j \left(\frac{\partial L}{\partial s_j}\right)^2 - \left(\sum_{j=1}^{m} s_j \frac{\partial L}{\partial s_j} \right)^2 \right]$$

$$= \frac{1}{M(\underset{\sim}{s})} \sum_{j=1}^{m} s_j d_j^2(\underset{\sim}{s}) \geq 0, \qquad (4.7)$$

where we have used (4.2) and have neglected terms of second and higher order. Thus $L(\underset{\sim}{s}') \geq L(\underset{\sim}{s})$ with equality only if, for each j, either $s_j = 0$ or $d_j(\underset{\sim}{s}) = 0$. Thus the algorithm converges monotonically, at least for $\underset{\sim}{s}^0$ close enough to $\hat{\underset{\sim}{s}}$, so that higher order terms can indeed be neglected. Suppose that the limiting value is $\tilde{\underset{\sim}{s}}$. Then $\tilde{\underset{\sim}{s}}$ satisfies (3.4). Hence if all $\tilde{s}_j > 0$, it follows by (4.5) that $d_j(\tilde{\underset{\sim}{s}}) = 0$ for all j and $\tilde{\underset{\sim}{s}}$ is the MLE $\hat{\underset{\sim}{s}}$ of $\underset{\sim}{s}$. Suppose then that $\tilde{s}_j = 0$ for some j and that $d_j(\underset{\sim}{s}) > 0$ in some neighbourhood of $\tilde{\underset{\sim}{s}}$. From the assumption that $s_j^0 > 0$ for all j it follows that $s_j^k > 0$ and $M(\underset{\sim}{s}^k) < \infty$ for $k = 0,1,2,\ldots$. We are assuming that $\underset{\sim}{s}^k$ eventually lies in this neighbourhood where $d_j(\underset{\sim}{s}) > 0$. However (4.5) implies that s_j^k cannot decrease any further towards $\tilde{s}_j = 0$, which is a contradiction. Thus if $\tilde{s}_j = 0$ for some j, it follows that $d_j(\tilde{\underset{\sim}{s}}) \leq 0$. (*In fact the limit of* $d_j(\underset{\sim}{s})$ as $\underset{\sim}{s} \to \tilde{\underset{\sim}{s}}$ may not exist if some $\tilde{s}_j = 0$, in which case we interpret the previous statement, and (4.6), as meaning that $d_j(\underset{\sim}{s}) \geq 0$ throughout a neighbourhood of $\underset{\sim}{s}$.) A similar argument shows that $d_\ell(\tilde{\underset{\sim}{s}}) = 0$ for any ℓ such that $\tilde{s}_\ell > 0$. Hence $\tilde{\underset{\sim}{s}}$ satisfies the condition (4.6) for maximising L and this completes the proof of the equivalence of the s.c.e. and m.l.e.

Note that for given initial vector $\underset{\sim}{s}^0$, the limit $\tilde{\underset{\sim}{s}}$ is unique even if $\hat{\underset{\sim}{s}}$ is not. A maximum likelihood estimate $\hat{F}$ of F is given by

$$\hat{F}(x) = \begin{cases} 0 & \text{if } x < q_1 \\ \tilde{s}_1 + \tilde{s}_2 + \ldots + \tilde{s}_j & \text{if } p_j < x < q_{j+1} \quad (1 \leq j \leq m-1) \\ 1 & \text{if } x > p_m, \end{cases}$$

and is undefined for $x \in [q_j, p_j]$ for $1 \leq j \leq m$. Therefore, when plotted,

$\hat{F}$ consists of a series of $m + 1$ horizontal lines of increasing heights with gaps in between, where the way in which increases occur is arbitrary. The variances of the non-zero $\{\tilde{s}_j\}$ are given by the inverse of the matrix of second derivatives of L with respect to the elements of $(s_1, s_2, \ldots, s_{m-1})$ corresponding to the non-zero elements of $\tilde{\underset{\sim}{s}}$. Thus estimates of the variance of $\hat{F}(x)$ can be calculated for $x \notin C$, from which approximate standard errors can be obtained for the height of each horizontal line.

5. DISCUSSION

Asano (1965) considered the problem of estimating the parameters of a multinomial distribution with truncation (but no censoring). For the "nested" case when $B_1 \supseteq B_2 \supseteq \ldots \supseteq B_N$ or the "chained" case when $B_i \cap B_j \neq \phi$ for $j = i-1, i, i+1$ and the intersection is empty otherwise (with reordering of the X_i if necessary), Asano gave explicit expressions for $\hat{\underset{\sim}{s}}$. Thus these two special cases can be added to those mentioned in Section 1 as being ones where formulae for the MLE can be written down explicitly and an iterative procedure is not needed. For the general case, Asano suggested using constrained Newton-Raphson methods. A similar search method was also proposed by Peto (1973) for the special case of interval censoring only.

However the Newton-Raphson (NR) procedure involves updating a vector of first derivatives and the inverse of a large matrix of second derivatives of L at each stage of the iteration. This can be difficult even for moderate values of m. Furthermore the step size in the NR iterations must be checked to ensure that the boundary of the region R is not violated. Also an improvement at each stage is not guaranteed since the step size may be too large and the maximum "overshot". To avoid this, the likelihood has to be calculated and if it has decreased the exercise must be repeated with a smaller step size, and so on. In contrast the self-consistency algorithm is completely automatic, simple to implement and is intuitively appealing.

In fairness, it should be pointed out that in exceptional cases, the convergence of the self-consistency algorithm can be rather slow. This happens if both $s_j = 0$ and $d_j(\tilde{\underset{\sim}{s}}) = 0$ for some j, i.e. the likelihood has and <u>unconstrained</u> maximum at $s_j = 0$. Why this is so can be seen by Equation (4.7). For example, suppose $m = 3$, $N = 4$ and $L^* = s_1(s_2 + s_3)s_3(s_1 + s_2)$.

This represents the case of no truncation and three intervals with one X in the first interval, one not in the first, one in the third and one not in the third. Starting with $s_j^0 = 1/3$ $(j = 1,2,3)$, we have $s_1^k = s_3^k = (1 - s_2^k)/2$ and $s_2^k = (3 + k)^{-1}$. Hence the convergence towards the MLE $\hat{s}_1 = \hat{s}_3 = 1/2$, $\hat{s}_2 = 0$ is quite slow. However such cases are exceptional and usually the convergence is rapid.

6. APPLICATION TO HYPOTHESIS TESTING

An important application of the MLE $\hat{F}$ is to the two sample problem where it is desired to test the equality of two distributions, and observations on one or both are subject to arbitrary censoring and truncation. (Extension to the K sample problem is immediate.) Let us suppose that $X_1,\ldots,X_{N_1}$ is a sample from Group 1 and the remaining $N_2 = N - N_1$ observations are from Group 2. It is desired to test the null hypothesis H_0 that all N observations have the same underlying F (unspecified). The alternative H_1 is that Group 1 observations have an underlying $F = F_{\theta_0}$ while $F = F_\theta \neq F_{\theta_0}$ for Group 2 observations. We consider Lehmann alternatives, i.e. $F_\theta(x) \equiv G_\theta(F_{\theta_0}(x))$ where G_θ is a specified c.d.f. on $[0,1]$ with $G_{\theta_0}(y) \equiv y$, while F_{θ_0} is unspecified. Peto and Peto (1972) have derived asymptotically efficient rank invariant tests for interval censored data, and their procedure can be naturally extended to the situation with arbitrary censoring and truncation as follows.

The likelihood L_i under H_0 with $F(x) \equiv G_\theta(\hat{F}(x))$ of the i'th observation represented by the pair (A_i,B_i) is

$$L_i = \sum_{j=1}^{m} g(j,\theta,\hat{\underset{\sim}{s}})\alpha_{ij} \Big/ \sum_{j=1}^{m} g(j,\theta,\hat{\underset{\sim}{s}})\beta_{ij}$$

where $g(j,\theta,\underset{\sim}{s}) = G_\theta(s_1 + \ldots + s_j) - G_\theta(s_1 + \ldots + s_{j-1})$ and $\{\alpha_{ij}\}$, $\{\beta_{ij}\}$ are defined as before. An efficient score for the i'th observation is given by $U_i = \partial \log L_i/\partial\theta\big|_{\theta=\theta_0}$, i.e.

$$U_i = \frac{\sum_{j=1}^{m} f(j,\hat{\underset{\sim}{s}})\alpha_{ij}}{\sum_{j=1}^{m} \hat{s}_j\alpha_{ij}} - \frac{\sum_{j=1}^{m} f(j,\hat{\underset{\sim}{s}})\beta_{ij}}{\sum_{j=1}^{m} \hat{s}_j\beta_{ij}} \quad ,$$

where $g(j,\hat{\underset{\sim}{s}}) = \partial g/\partial\theta\big|_{\theta=\theta_0}$ and we have used the fact that $g(j,\theta_0,\underset{\sim}{s}) = s_j$. If we assume that, under H_0, the censoring and truncation mechanism is random and independent of group membership, a test of any given size can be constructed using the permutational distribution of $\sum_{\text{group 1}} U_i$.

The test statistic may not be unique if $\hat{\underset{\sim}{s}}$ is not unique. If this situation occurs, for small samples the test statistic can be evaluated for "extreme" values of the possible $\hat{\underset{\sim}{s}}$ and if the decision concerning acceptance or rejection of H_0 is the same there is no difficulty. In large samples non-uniqueness of $\hat{\underset{\sim}{s}}$ is less likely to occur and if it does the difference between significance levels for the different values of $\hat{\underset{\sim}{s}}$ will be small.

The above discussion assumes that the same random censoring mechanism is operating in each group. Tests that do not make this assumption have been proposed by Efron (1967) for right censored data and Mantel (1967) for arbitrarily censored data, (with no truncation). Efron uses the MLE's of the distributions of the two groups calculated separately to derive an estimate of the probability that an X-value from group 1 is greater than one from group 2 and this is used as a test statistic. In theory the test could be easily extended to arbitrary censoring, however it is difficult to compute the sampling distributions involved. Mantel (1967, Section 7) describes a test for arbitrarily censored data which is a generalisation of that of Gehan (1965) and does not use the estimated c.d.f. A disadvantage of this test is that it requires knowledge, not always available, of the entire pattern of restriction for each observation even if it is exact. Also much of the information in the data is unused and thus the test will be rather inefficient.

7. CONCLUSION

The definition of self-consistent estimates does not directly involve the likelihood function and so their exact coincidence with the MLE's is an aesthetic and perhaps unexpected result. The property was first proved for singly censored data by Efron (1967) and for doubly censored data by Turnbull (1974). However their methods were lengthier and involved converting the likelihood equations into the defining equations for self-consistency rather than the examination of successive values of the estimates given by the algorithm.

An alternative nonparametric approach is to estimate the hazard rate associated with F rather than F itself. The work of several authors is summarised by Barlow (1968, Section 3). Usually the hazard rate is assumed to be a step function, constant within each interval, the set of intervals being fixed. For instance, Harris, Meier and Tukey (1950) treat an interval censoring situation and use a similar "prorating" idea as the basis for an iterative scheme for obtaining approximate MLE's of the hazard rates in the various intervals.

Consistency and other large sample properties of the maximum likelihood estimate $\hat{F}$ will depend on the censoring and truncation mechanism. Consider the case when the range of X is finite, $\{t_1,t_2,\ldots,t_m\}$ say, and when the mechanisms are random as described in Section 1. If we suppose that the sets B and the partitions with non-zero probability are such that conditions (A), (B) as stated in Section 2 do not occur for N sufficiently large, then $q_j = p_j = t_j$ $(1 \leq j \leq m)$ and $\hat{\underset{\sim}{s}}$ is unique - again for N sufficiently large. Then consistency and asymptotic normality of the MLE's of the non-zero s_j follow from the standard theory, regarding the pairs (A_i,B_i) as i.i.d. random variables involving a finite number of parameters including the $\{s_j\}$. Large sample properties of $\hat{F}$ when m does not remain bounded

as $N \to \infty$ is an interesting open question. (Results are known for the case of single censoring - see Breslow and Crowley (1974).)

For the situation when there is concomitant information available for each observation, there appears to be no natural extension of the powerful methods that Cox (1972) has proposed for singly censored data. However in a recent paper on regression with censored data, R.G. Miller (1974) uses $\hat{F}$ as a basis for inference and thus it appears that his methods can be extended to the case of arbitrary censoring and truncation.

It is interesting to note that Sackrowitz and Strawderman (1974) have shown that, for a wide class of reasonable loss functions, the MLE $\hat{F}$ is inadmissible for the case when the range of X is finite and there is extreme double censoring (no exact observations). Thus in general the MLE will be inadmissible. However, unless a prior measure can be assigned to the space of possible c.d.f.'s F, there is no apparent substitute to be preferred to the MLE. Indeed self-consistency provides a justification for using maximum likelihood even in relatively small samples.

8. ACKNOWLEDGEMENTS

The author is grateful to Richard Peto and Nathan Mantel for some useful conversations.

9. REFERENCES

[1] Asano, C. (1965). On estimating multinomial probabilities by pooling incomplete samples. Ann. Inst. Statist. Math., 17, 1-13.

[2] Ayer, M. Brunk, H.D. Ewing, G.M., Reid, W.T. and Silverman, E. (1955) An empirical distribution function for sampling with incomplete information. Ann. Math. Statist., 26, 641-647.

[3] Barlow, R.E. (1968) Some recent developments in reliability theory. Selected Statistical Papers, Mathematical Centre, Amsterdam, 2, 49-66.

[4] Blight, B.J.N. (1970). Estimation from a censored sample for the exponential family. Biometrika, 57, 389-395.

[5] Breslow, N. and Crowley, J. (1974). A large sample study of the life table and product limit estimates under random censorship. Ann. Statist., 2, 437-453.

[6] Buckland, W.R. (1964). Statistical Assessment of the Life Characteristic. London: Griffin.

[7] Cohen, A.C. (1957). On the solution of estimating equations for truncated and censored samples from normal populations. Biometrika, 44, 225-236.

[8] Cox, D.R. (1972). Regression models and life tables. J.R. Statist. Soc. B, 34, 187-220.

[9] Efron, B. (1967). The two sample problem with censored data. In Proc. 5th Berkeley Symp. on Math. Statist. Prob., pp. 831-853. Berkeley: University of California Press.

[10] Gehan, E.A. (1965). A generalized two-sample Wilcoxon test for doubly censored data. Biometrika, 52, 650-653.

[11] Harris, T.E., Meier, P. and Tukey, J.W. (1950). Timing of the distribution of events between observations. Human Biology, 22, 249-270.

[12] Hartley, H.O. (1958). Maximum likelihood estimation from incomplete data. Biometrics, 14, 174-194.

[13] Hartley, H.O. and Hocking, R.R. (1971). The analysis of incomplete data. Biometrics, 27, 783-823.

[14] Hocking, R.R. and Oxspring, H.H. (1971). Maximum likelihood estimation with incomplete multinomial data. J. Am. Statist. Assoc., 66, 65-70.

[15] Kaplan, E.L. and Meier, P. (1958). Nonparametric estimation from incomplete observations. J. Am. Statist. Assoc., 53, 457-481.

[16] Mantel, N. (1966). Evaluation of survival data and two new rank order statistics arising in its consideration. Cancer Chemotherapy Reports, 50, 163-170.

[17] Mantel, N. (1967). Ranking procedures for arbitrarily restricted observations. Biometrics, 23, 65-78.

[18] Miller, A.J. (1974). A note on the analysis of gap-acceptance in traffic. Appl. Statist., 23, 66-73.

[19] Miller, R.G. (1974). Least squares regression with censored data. Technical Report, Stanford Univ.

[20] Peto, R. (1973). Experimental survival curves for interval-censored data. Appl. Statist., 22, 86-91.

[21] Peto, R. and Peto, J. (1972). Asymptotically efficient rank invariant test procedures. J.R. Statist. Soc. A, 135, 185-206.

[22] Sackrowitz, H. and Strawderman, W. (1974). On the admissibility of the M.L.E. for ordered binomial parameters. Ann. Statist. 2, 822-828.

[23] Selvin, S. (1974). Maximum likelihood estimation in the truncated or censored Poisson distribution. J. Am. Statist. Assoc., 69, 234-237.

[24] Sundberg, R. (1974). Maximum likelihood theory for incomplete data from an exponential family. Scand. J. Statist., 1, 49-58.

[25] Turnbull, B.W. (1974). Nonparametric estimation of a survivorship function with doubly censored data. J. Am. Statist. Assoc., 69, 169-173.

Unclassified

SECURITY CLASSIFICATION OF THIS PAGE (When Data Entered)

REPORT DOCUMENTATION PAGE

READ INSTRUCTIONS BEFORE COMPLETING FORM

1. REPORT NUMBER: #305
2. GOVT ACCESSION NO.
3. RECIPIENT'S CATALOG NUMBER

4. TITLE (and Subtitle): THE EMPIRICAL DISTRIBUTION FUNCTION WITH ARBITRARILY GROUPED, CENSORED, AND TRUNCATED DATA.

5. TYPE OF REPORT & PERIOD COVERED: Technical Report

6. PERFORMING ORG. REPORT NUMBER

7. AUTHOR(s): Bruce W. Turnbull

8. CONTRACT OR GRANT NUMBER(s): DAHC04-73-C-0008. N00014-75-C-0586

9. PERFORMING ORGANIZATION NAME AND ADDRESS
School of Operations Research & Industrial Engineering, College of Engineering
Cornell University, Ithaca, NY 14853

10. PROGRAM ELEMENT, PROJECT, TASK AREA & WORK UNIT NUMBERS

11. CONTROLLING OFFICE NAME AND ADDRESS
Sponsoring Military Activity
U.S. Army Research Office
Durham, N.C. 27706

12. REPORT DATE: July 1976

13. NUMBER OF PAGES: 27 (33p.)

II. CONTROLLING OFFICE NAME AND ADDRESS
Sponsoring Military Activity
Statistics and Probability Program
Office of Naval Research
Arlington, Virginia 22217

15. SECURITY CLASS (of this report): Unclassified

15a. DECLASSIFICATION/DOWNGRADING SCHEDULE

16. DISTRIBUTION STATEMENT (of this Report)

Approved for public release, distribution unlimited. (14) TR 376 (1)

17. DISTRIBUTION STATEMENT (of the abstract entered in Block 20, if different from Report)

18. SUPPLEMENTARY NOTES

19. KEY WORDS (Continue on reverse side if necessary and identify by block number)

Empirical distribution function; censoring; interval censoring, truncation, grouping; survival curve; maximum likelihood; self-consistency; Newton-Raphson; multinomial distribution; two sample test; logrank test; Lehmann alternatives

20. ABSTRACT (Continue on reverse side if necessary and identify by block number)

This paper is concerned with the nonparametric estimation of a distribution function F, when the data are incomplete due to grouping, censoring and/or truncation. Subsets $B_1, B_2, \ldots, B_N$ of the real line are given and there are N independent observations $X_1, X_2, \ldots, X_N$, where X_i is drawn from the truncated distribution $F(x;B_i) = P(X<x \mid X \in B_i)$, $x \in B_i$. However X_i may not

DD FORM 1 JAN 73 1473 EDITION OF 1 NOV 65 IS OBSOLETE
S/N 0102-014-6601

403138

Unclassified
SECURITY CLASSIFICATION OF THIS PAGE (When Data Entered)

be observed exactly and is known only to lie in the set $A_i \subseteq B_i$. The situation occurs frequently in survivorship, reliability, and recidivism analysis. Using the idea of self-consistency, a simple algorithm is constructed and shown to converge monotonically to yield a maximum likelihood estimate of F. The procedure compares favourably with the more cumbersome Newton-Raphson method. A test is proposed for comparing two distributions when data on one or both is incomplete and some other applications of the empirical distribution function are indicated.