NON-PARAMETRIC ESTIMATION OF A SURVIVORSHIP FUNCTION WITH DOUBLY CENSORED DATA

BY

BRUCE W. TURNBULL

TECHNICAL REPORT NO. 32
JUNE 29, 1972

PREPARED UNDER THE AUSPICES
OF
PUBLIC HEALTH SERVICE GRANT USPHS-5TI GM 25-14

DEPARTMENT OF STATISTICS
STANFORD UNIVERSITY
STANFORD, CALIFORNIA

LELAND STANFORD JUNIOR UNIVERSITY
ORGANIZED 1891

NON-PARAMETRIC ESTIMATION OF A SURVIVORSHIP FUNCTION WITH DOUBLY CENSORED DATA

by

BRUCE W. TURNBULL

TECHNICAL REPORT NO. 32
June 29, 1972

PREPARED UNDER THE AUSPICES

OF

PUBLIC HEALTH SERVICE GRANT USPHS-5TI GM 25-14

DEPARTMENT OF STATISTICS
STANFORD UNIVERSITY
STANFORD, CALIFORNIA

Abstract

Non-parametric estimation of a survivorship function with doubly censored data

by

Bruce W. Turnbull

In medical follow-up, life-testing and other situations, the problem is to estimate the survivorship function $P(t) = Prob(T > t)$, where T is the time taken for some event of interest to occur. An iterative procedure is proposed for obtaining non-parametric estimates of $P(t)$ when some of the data are censored on the left and some are censored on the right. The procedure is based on the product-limit method of Kaplan and Meier (1958) for singly censored data and it also uses the idea of self-consistency due to Efron (1965). Under certain assumptions concerning the grouping of the observations, the estimates are shown to have the maximum likelihood property. Using this fact, expressions for their asymptotic variances and covariances are derived. Some special cases are considered where explicit solutions can be found. Also, an iterative method is given for obtaining the self-consistent estimates in the more general case of arbitrary censoring.

KEY WORDS: survivorship function; survival curve; medical follow-up; self-consistency; life-testing; censoring; maximum likelihood; multinomial distribution.

1. Introduction and summary.

A common problem in statistical analysis is the determination of the distribution of the time, T, taken for an event of interest to occur. For instance, in medical follow-up studies, the event of interest is the relapse or death of a patient; and in life-testing it is the time to failure of an item that is under investigation. In this paper all such events will be termed "deaths", and thus the problem is to estimate the lifetime distribution, that is $F(t) = \text{Prob}(T \leq t)$ for $t \geq 0$.

In a sample of N observations $T_1, T_2, \ldots, T_N$, where each lifetime T_i is observed precisely, the natural estimate is the sample distribution function $\hat{F}(t)$, which is the proportion of observations with values less than or equal to the argument t. In this paper we shall consider the situation where not all the T_i are observed exactly but some are censored on the right and some are censored on the left. For each item i $(1 \leq i \leq N)$, we assume that there are limits of observation L_i and U_i (with $L_i \leq U_i$), which are either fixed constants or random variables independent of the $\{T_i\}$. Thus (L_i, U_i) is a "window" of observation and the recorded information is:

$$X_i = \max[\min(T_i, U_i), L_i].$$

Also, for each item i, it is known whether $X_i = L_i$ (i.e.

$T_i \leq L_i$ and the item is a "late entry"), or $X_i = U_i$ (i.e. $T_i > U_i$ and the item is a "loss"), or $X = T_i$ (i.e. $L_i < T_i \leq U_i$). We can denote a loss at time t by the symbol "$>t$", a late entry by "$\leq t$", and a precise observation by simply "t".

Losses can occur for several reasons -- for instance, in medical follow-up, contact may have been lost with the patient after a certain time, he may have died from an unrelated cause or a relapse may not have occurred before termination of the study. In the last case, the $\{U_i\}$ vary according to when the patient entered the study. Late entries occur when an item is inspected and found to be dead with no information on when the event had actually occurred prior to this time.

An example where both losses and late entries occurred together was in a recent study of African infant precocity by Leiderman et al. (1972). His purpose was to establish norms for infant development for a community in Kenya in order to make comparisons with known standards in the United States and the United Kingdom. The sample consisted of 65 children born between 1 July and 31 December 1969. Starting in January 1970, each child was tested monthly to see if he had learned to accomplish certain standard tasks (see Bayley 1969). Here T would represent the time from birth to first learn to perform a particular task. Late entries occurred when it was found that, at the very first test, some children could already perform the task; whereas losses occurred when some infants were still unsuccessful by the end of the study.

The more common case is when there are no late entries (all $L_i = 0$) and this has been treated extensively in the literature. Often some parametric form for F is assumed such as an exponential, lognormal or Weibull. The method of maximum likelihood in such a situation was first used by Boag (1949) and most recently by Herman and Patell (1971) and Moeschberger and David (1971). (In fact, these two papers looked at the more general multi-risk model where there are several competing causes of death.) Non-parametric estimates can be obtained by the actuarial method (see, for instance, Berkson and Gage 1950), or by the product-limit (PL) or reduced-sample (RS) methods described in Kaplan and Meier (1958). Non-parametric two sample tests for comparing two such lifetime distributions have been proposed by Halperin (1960), Gilbert (1962), Gehan (1965a), Efron (1965), Mantel (1966) and Myers (1967). Two sample tests with doubly censored data have been treated by Gehan (1965b) and Mantel (1967).

In this paper we discuss the estimation problem when there is both left and right censoring. We assume that the data is grouped and that lifetimes are recorded as belonging to one of the m intervals $(0,t_1],(t_1,t_2],\ldots,(t_{m-1},t_m]$. The assumptions about how this grouping is done are stated in Section 2. The remainder of this paper gives methods of constructing maximum likelihood estimates for $P(t_1),P(t_2),\ldots,P(t_m)$, where $P(t) = 1-F(t)$ is the survivorship function. In Section 3, certain special cases are considered, where the data are in such a form that explicit expressions for the estimates are available. In Section 4, an iterative procedure for the general case is proposed and, in Section 5, shown to yield maximum likelihood estimates. This procedure is based on the method of Kaplan and Meier (1958). The variances and covariances of the estimates are derived in Section 6. In Section 7, some suggestions are made for modifying the estimates when assumptions other than those of Section 2 are appropriate. Lastly, an iterative method is proposed to find estimates in the more general model, first treated in detail by Harris, Meier and Tukey (1950), in which for each item it is known only that death had not occurred at one known time and that it had occurred before another known time.

2. Some notation and assumptions.

We divide up the time scale into intervals $(t_0,t_1],(t_1,t_2]$, $.....(t_{m-1},t_m],[t_m,\infty)$, where $0 = t_0 < t_1 < t_2 <....< t_m$. Here the $\{t_i\}$ may be chosen arbitrarily but would usually represent the ages of the items at possible inspection times. For example, in the Leiderman study mentioned in Section 1, these times would correspond to 1 month, 2 months, ... etc. Of course, for the cohort of all babies born in July 1969, the possible ages at inspection would start at 6 months since the tests were not started until January 1970. Thus different cohorts would have different possible ages of inspection.

Let δ_i be the number of items observed to have died in age period $(t_{i-1},t_i]$, μ_i be the number of late entries at age t_i, and let λ_i be the number of losses at t_i $(1 \leq i \leq m)$. The situation is illustrated by Table I.

Table I

Age	t_1	t_2	....	t_m
Deaths	δ_1	δ_2		δ_m
Losses (>)	λ_1	λ_2		λ_m
Late entries (≤)	μ_1	μ_2		μ_m

We have made the assumption that the late entries μ_i all occur at the end of age period $(t_{i-1},t_i]$ and that the losses λ_i all occur at the beginning of $(t_i, t_{i+1}]$. This will be valid if the inspection procedure is as follows: Examine a cohort of items all of age t, observe the number δ of deaths since the last examination, record the number μ of late entries who are known only to have died at or before t; finally lose contact with a number λ of the survivors. When all cohorts are considered in this way, the data can be displayed as in Table I. This procedure applies in the Leiderman study and is often the case in medical follow-up. It is also consistent with the conventions followed in Section 1.4 of Kaplan and Meier (1958). They point out, however, that there is a major exception which occurs when "the losses are random but cannot affect items that have already died." In this case it is reasonable to proceed as if half the losses precede and half follow the deaths. This "half-rule" is treated in, for instance, Berkson and Gage (1950) and Harris, Meier and Tukey (1950), and will be discussed further in Section 7.

It can be seen that the assumptions of grouping and of the timing of late entries and losses essentially reduces the problem to one of estimating the parameters $P(t_1), \ldots, P(t_m)$ of a multinomial distribution.

3. Some special cases.

In this section we consider three special cases for which explicit expressions can be derived for the maximum likelihood estimates $\{\hat{P}_i\}$ of the $\{P(t_i)\}$, $(1 \leq i \leq m)$.

3.1. No late entries.

Suppose that $\mu_i = 0$ for all i. This was the case considered by Kaplan and Meier (1958). They showed the maximum likelihood estimates for the $\{P(t_i)\}$ are given by:

$$\begin{aligned} \hat{P}_1 &= q_1 \\ \hat{P}_j &= q_j \hat{P}_{j-1} \qquad (j = 2,3,\ldots,m); \end{aligned} \tag{3.1}$$

where $q_j = (n_j - \delta_j)/n_j$, and $n_j = \Sigma_{i=j}^{m} (\lambda_i + \delta_i)$. Thus q_j is the proportion of those "at risk" at age t_j who survive past t_j and it estimates $\text{Prob}(T > t_j | T > t_{j-1})$. The $\{\hat{P}_j\}$ were called the "product-limit" (PL) estimates by Kaplan and Meier. Under the assumptions about the timing of the losses made in Section 2, these will coincide with the actuarial estimates.

Efron (1965, Corollary 7.1) showed that these estimates have an interesting property which he called "self-consistency". Consider the λ_i items lost at age t_i. Of these, the expected number that will die in period $(t_{j-1}, t_j]$ for $j > i$ is $\beta_{ij}\lambda_i$ where $\beta_{ij} = \text{Prob}[t_{j-1} < T \leq t_j | T > t_i] = [P(t_{j-1}) - P(t_j)]/P(t_i)$. Thus the total expected number of deaths or "adjusted" deaths for the period $(t_{j-1},t_j]$ is:

$$\delta_j' = \delta_j + \Sigma_{i=1}^{j-1} \beta_{ij}\lambda_i \qquad (1 \leq j \leq m).$$

If we knew the $\{\delta_j'\}$, then we would estimate $P(t_j)$ by the simple binomial estimate:

$$\hat{P}_j' = \frac{1}{N} \Sigma_{j=i+1}^{m} \delta_j' \tag{3.2}$$

where $N = \Sigma_{j=1}^{m} \delta_j' = \Sigma_{j=1}^{m}(\delta_i+\lambda_i)$ is the total sample size. Of course, unless all $\lambda_i = 0$, we do not know the $\{\delta_j'\}$ because we do not know the $\{\beta_{ij}\}$. Now suppose that, in the expression for β_{ij}, we replace $P(t_j)$ by its estimate $\hat{P}_j$ as given by (3.1) (and similarly for $P(t_i)$), and then we form the $\{\delta_j'\}$. It then happens that the Formula (3.2) yields the same estimates; i.e. $\hat{P}_j' = \hat{P}_j$, for all j. For obvious reasons, Efron called such estimators "self-consistent". In Section 4, we shall extend this idea to enable us to solve the problem when late entries are also present.

3.2. No losses.

Suppose that all $\lambda_i = 0$, except perhaps λ_m. In this case it is easier to work with the distribution function $F(t) = \text{Prob}[T \leq t]$.

Define $r_j = \text{Prob}[T \leq t_j | T \leq t_{j+1}]$ and $F(t_0) = 0$. Then

$F(t_j) = F(t_m) \prod_{i=j}^{m-1} r_i$, and the likelihood function is given by:

$$L = [1-F(t_m)]^{\lambda_m} \cdot \prod_{j=1}^{m} [F(t_j)-F(t_{j-1})]^{\delta_j} \cdot [F(t_j)]^{\mu_j}$$

$$= [1-F(t_m)]^{\lambda_m} \cdot [F(t_m)]^{n_m} \cdot \prod_{j=1}^{m-1} r_j^{n_j}(1-r_j)^{\delta_{j+1}}$$

where $n_j = \Sigma_{i=1}^{j}(\delta_i+\mu_i)$ is the number of items known to have died at or before t_j. Each factor is maximized separately by the binomial estimates $\hat{r}_j = n_j/(n_j+\delta_{j+1})$, $(1 \le j \le m-1)$ and $\hat{F}(t_m) = n_m/(n_m+\lambda_m)$. Hence the maximum likelihood estimates are:

$$\hat{F}(t_m) = n_m/(n_m+\lambda_m),$$

$$\hat{F}(t_j) = n_j\hat{F}(t_{j+1})/(n_j+\delta_{j+1}), \qquad j = m-1, m-2, \ldots, 1.$$

The proof is the same as that of Kaplan and Meier (1958, Section 5), which was applicable in Section 3.1, but with the time scale reserved.

3.3. The case where losses and late entries do not overlap.

Suppose that $\mu_i = 0$ for $i > I$ and $\lambda_i = 0$ for $I < J$ where $I \le J$. This situation is represented by Table II.

Table II

Ages	1	....	I	I+1	....	J-1	J	....	m
Deaths	δ_1		δ_I	δ_{I+1}		δ_{J-1}	δ_J		δ_m
Losses	0		0	0		0	λ_J		λ_m
Late entries	μ_1		μ_I	0		0	0		0

In fact, the data for many of the tests conducted in the Leiderman study, as described in Section 1, fell into this category. In this case, the maximum likelihood estimates $\{\hat{P}_i\}$ for the $\{P(t_i)\}$ are given by:

$$\hat{P}_i = 1 - \frac{1}{N}\Sigma_{j=1}^{i}(\delta_j+\mu_j), \qquad i = I,I+1,\ldots,J;$$

$$\hat{P}_i = \left[1 - \frac{\delta_i}{\Sigma_{j=i}^{m}(\lambda_j+\delta_j)}\right] \cdot \hat{P}_{i-1}, \qquad i = J+1,\ldots,m;$$

$$\hat{P}_i = 1 - (1-\hat{P}_{i+1})\left[1 - \frac{\delta_{i+1}}{\Sigma_{j=1}^{i}(\delta_j+\mu_j)+\delta_{i+1}}\right], \qquad i = I-1,\ldots,2,1.$$

This follows directly by combining the corresponding proofs in the previous cases: no censoring, right censoring only, and left censoring only. (I am grateful to Dr. Helena C. Kraemer who suggested this special case and its treatment to me.)

4. An iterative method for solving the general case.

In Section 3.1 we noted that, if all $\mu_i = 0$, then explicit maximum likelihood estimates can be obtained by the method of Kaplan and Meier and are given by (3.1). We also saw the sense in which these estimates are self-consistent. For the general case when late entries are also present, we will demand a similar property of self-consistency from our estimates $\{\hat{P}_i\}$. Consider the μ_i late entries at age i. Then the estimated mean number of these that die in period $(t_{j-1},t_j]$ for $j \leq i$ is $\mu_i\alpha_{ij}$ where $\alpha_{ij} = (\hat{P}_{j-1}-\hat{P}_j)/(1-\hat{P}_i)$ is an estimate of $\text{Prob}(t_{j-1} < T \leq t_j | T \leq t_i)$. If we now take all μ_i to be zero and replace each δ_j by an "adjusted" number of deaths, $\delta_j' = \delta_j + \Sigma_{i=j}^m \mu_i\alpha_{ij}$, then we are in the special case considered in Section 3.1. Suppose we obtain the PL estimates $\{\hat{P}_i'\}$ using this "adjusted" data set, then we shall say that the estimates $\{\hat{P}_j\}$ are self-consistent if $\hat{P}_j' = \hat{P}_j$ $(i \leq j \leq m)$.

Therefore the problem is to find numbers $1 \geq \hat{P}_1 \geq \hat{P}_2 \geq \dots \geq \hat{P}_m \geq 0$, which satisfy the following implicit equations:

$$\begin{aligned} \hat{P}_1 &= q_1 \\ \hat{P}_j &= q_j\hat{P}_{j-1} \qquad (j = 2,3,\dots,m) \end{aligned} \tag{4.1}$$

where $q_j = (n_j'-\delta_j')/n_j'$, $n_j' = \Sigma_{i=j}^m(\lambda_i+\delta_i')$, $\delta_j' = \delta_j + \Sigma_{i=j}^m \mu_i\alpha_{ij}$ and $\alpha_{ij} = (\hat{P}_{j-1}-\hat{P}_j)/(1-\hat{P}_i)$ for $j \leq i$.

An iterative method for solving these equations immediately suggests itself.

A. Obtain starting values $\{P_i^0, 1 \leq i \leq m\}$. For instance these could the PL estimates (3.1) with all μ_i taken to be zero. (See also the note at the end of this section.)

B. Form $\alpha_{ij}^0 = (P_{j-1}^0 - P_j^0)/(1-P_i^0)$ all $j \leq i$.

and set $\delta_j' = \delta_j + \Sigma_{i=j}^m \mu_i \alpha_{ij}$ $1 \leq j \leq m$.

C. Obtain improved estimates by taking all $\mu_i = 0$, replacing δ_i by δ_i' $(1 \leq i \leq m)$ and forming the PL estimates on this "adjusted" data set, i.e.

$$P_1^1 = 1 - \delta_1'/n_1' ,$$

and

$$P_j^1 = q_j P_{j-1}^1 , \qquad (j = 2,3,\ldots,m);$$

where

$$q_j = (n_j' - \delta_j')/n_j'$$

and

$$n_j' = \Sigma_{i=j}^m (\lambda_i + \delta_i').$$

D. Return to Step B with the $\{P_j^0\}$ replaced by $\{P_j^1\}$ etc.

E. Stop when the required accuracy has been achieved. (E.g. the rule may be to stop when $\max_{1 \leq i \leq m} |P_i^\ell - P_i^{\ell-1}| < 0.001$, say.)

The procedure is simple to program on a computer and converges fairly rapidly. As a small numerical example, consider the data in Table III.

Table III

Age	t_1	t_2	t_3	t_4
Deaths	12	6	2	3
Losses	3	2	0	3
Late entries	2	4	2	5

The initial values are:

$$\underset{\sim}{\delta} = 12.0, \quad 6.0, \quad 2.0, \quad 3.0$$

$$\underset{\sim}{P}^0 = 0.613, \quad 0.383, \quad 0.287, \quad 0.144 .$$

The first iteration yields:

$$\underset{\sim}{\delta}' = 19.9, \quad 9.5, \quad 2.8, \quad 3.8$$

$$\underset{\sim}{P}^1 = 0.549, \quad 0.303, \quad 0.214, \quad 0.094 .$$

After three iterations, the values settle down on:

$$\underset{\sim}{\delta}' = 20.3, \quad 9.3, \quad 2.7, \quad 3.6$$

$$\hat{\underset{\sim}{P}} = 0.538, \quad 0.295, \quad 0.210, \quad 0.095 .$$

Note. For the starting values $\{P_i^0\}$, any decreasing sequence of m numbers between 0 and 1 will suffice. If the number of late entries is small compared to the number of losses, then the starting values suggested in Step A should be used. However, if the number of losses is smaller than the number of late entries, better starting values can be obtained by assuming $\lambda_i = 0$ for $1 \leq i \leq m-1$ and using the method of Section 3.2. Alternatively, if the $\{\lambda_i\}$ are small for small i and the $\{\mu_i\}$ small for large i, then the method of Section 3.3 can be applied to obtain starting values which are closer to the final solution.

5. The maximum likelihood derivation.

We now state and prove the fundamental theorem of this paper.

Theorem.

If $\delta_i > 0$ $(1 \leq i \leq m)$, then the solution $\underset{\sim}{\hat{P}} = (\hat{P}_1, \hat{P}_2, \ldots, \hat{P}_m)$ of the Equations (4.1), obtained by the iterative procedure of Section 4, is the unique maximum likelihood estimate of $(P(t_1), P(t_2), \ldots, P(t_m))$.

Proof.

Under the assumptions made in Section 2, the likelihood function is proportional to:

$$\prod_{j=1}^{m} (P_{j-1} - P_j)^{\delta_j} P_j^{\lambda_j} (1-P_j)^{\mu_j} ,$$

where $P_j = P(t_j)$ $(1 \leq j \leq m)$ and $P_0 = 1$. The log-likelihood L is given by:

$$L = \Sigma_{j=1}^{m} [\delta_j \log(P_{j-1} - P_j) + \lambda_j \log P_j + \mu_j \log(1-P_j)].$$

The maximum likelihood estimates will be those values of $\{P_j\}$ which maximize L subject to the condition $1 \geq P_1 \geq \cdots \geq P_m \geq 0$.

First note that if $\lambda_m = 0$, then L is maximized by taking $\hat{P}_m = 0$, and the problem can be treated as one with $m-1$ periods and λ_{m-1} replaced by $\lambda_{m-1} + \delta_m$. In this case μ_m contributes no information to the estimation of the $\{P_i\}$ and this agrees with intuition. Thus without loss of generality we may assume $\lambda_m > 0$.

Differentiating L with respect to $P_1, P_2, \ldots, P_m$ and setting the derivatives equal to zero, we obtain:

$$\frac{\partial L}{\partial P_j} = -\frac{\delta_j}{P_{j-1}-P_j} + \frac{\delta_{j+1}}{P_j-P_{j+1}} + \frac{\lambda_j}{P_j} - \frac{\mu_j}{1-P_j} = 0, \qquad (j = 1,2,\ldots,m-1);$$

$$\frac{\partial L}{\partial P_m} = -\frac{\delta_m}{P_{m-1}-P_m} + \frac{\lambda_m}{P_m} - \frac{\mu_m}{1-P_m} = 0 . \tag{5.1}$$

In Lemma A1 of the Appendix, we show that any solution $\{\hat{P}_i\}$ of (4.1) also satisfies the likelihood equations (5.1). Also the estimates obtained in Section 4 clearly satisfy the condition $1 > \hat{P}_1 > \ldots\ldots > \hat{P}_m > 0$, since we have assumed $\lambda_m > 0$ and $\delta_i > 0$ $(1 \leq i \leq m)$. Thus the $\{\hat{P}_i\}$ give stationary values of the likelihood function. To show that this is a unique maximum we examine the matrix $\underset{\sim}{D}$ of second derivatives.

Let $D_{ij} = \frac{\partial^2 L}{\partial P_i \partial P_j}$, then

$$D_{ii} = -\frac{\delta_i}{(P_{i-1}-P_i)^2} - \frac{\delta_{i+1}}{(P_i-P_{i+1})^2} - \frac{\lambda_i}{P_i^2} - \frac{\mu_i}{(1-P_i)^2} \qquad (i = 1,2,\ldots,m-1)$$

$$D_{mm} = -\frac{\delta_m}{(P_{m-1}-P_m)^2} - \frac{\lambda_m}{P_m^2} - \frac{\mu_m}{(1-P_m)^2} \tag{5.2}$$

$$D_{i+1,i} = D_{i,i+1} = \frac{\delta_{i+1}}{(P_i-P_{i+1})^2} \qquad i = 1,2,\ldots,m-1$$

$$D_{ij} = 0, \qquad \text{for } |i-j| \geq 2 .$$

In Lemma A2 of the Appendix, we show that all the leading principal minors of $\underset{\sim}{D}$ are negative and thus $\underset{\sim}{D}$ is negative definite. Hence all solutions of (5.1) yield maxima. But L is a continuous function for $1 > P_1 > \ldots\ldots > P_m > 0$ and so if L has two maxima there must be a minimum between them. There are no minima and so L has a unique maximum. This completes the proof. □

6. Variances and covariances of the estimates.

The Fisher information matrix $\underset{\sim}{J}$ is the matrix $-\underset{\sim}{D}$, where $\underset{\sim}{D}$ is given by (5.2). $\underset{\sim}{J}$ is a symmetric Jacobi matrix, i.e. it is of the form:

$$\begin{pmatrix} c_1 & d_1 & 0 & 0 & \cdots\cdots & 0 & 0 \\ d_1 & c_2 & d_2 & 0 & \cdots\cdots & 0 & 0 \\ 0 & d_2 & c_3 & d_3 & \cdots\cdots & 0 & 0 \\ 0 & 0 & d_3 & c_4 & \cdots\cdots & 0 & 0 \\ \vdots & \vdots & \vdots & \vdots & & \vdots & \vdots \\ 0 & 0 & 0 & 0 & \cdots\cdots & c_{m-1} & d_{m-1} \\ 0 & 0 & 0 & 0 & \cdots\cdots & d_{m-1} & c_m \end{pmatrix}$$

where $c_i = -D_{ii}$ $(1 \leq i \leq m)$ and $d_i = -D_{i,i+1}$ $(1 \leq i \leq m-1)$. The inverse $\underset{\sim}{V}$ of $\underset{\sim}{J}$ is a Green's matrix (see for example Karlin (1968, Chapter 3.3)) and is given by:

$$V_{ij} = a_{\min(i,j)} \cdot b_{\max(i,j)},$$

where

$$a_i = \frac{(-1)^i}{\det(J)} J\begin{pmatrix} 1,2,\ldots,i-1 \\ 1,2,\ldots,i-1 \end{pmatrix} d_i \ldots d_{m-1} \qquad (2 \leq i \leq m-1)$$

$$b_j = (-1)^j J\begin{pmatrix} j+1,\ldots,m \\ j+1,\ldots,m \end{pmatrix} \cdot \frac{1}{d_j d_{j+1} \ldots d_{m-1}} \qquad (1 \leq j \leq m-1)$$

$$a_1 = - d_1 d_2 \ldots d_{m-1} / \det(J)$$

$$a_m = (-1)^m \, J\begin{pmatrix} 1,2,\ldots m-1 \\ 1,2,\ldots m-1 \end{pmatrix} \Big/ \det(J)$$

$$b_m = (-1)^m ,$$

and $J\begin{pmatrix} i_1 i_2 \ldots i_r \\ j_1 j_2 \ldots j_r \end{pmatrix}$ represents the determinant of the matrix formed from $\underset{\sim}{J}$ by removing all rows and columns except rows $i_1, i_2, \ldots, i_r$ and columns $j_1, j_2, \ldots, j_r$.

Denote $\underset{\sim}{V}(\hat{\underset{\sim}{P}})$ as the value of the matrix $\underset{\sim}{V}$ when the $\{P_i\}$ are replaced by their maximum likelihood estimates $\{\hat{P}_i\}$. Then $\underset{\sim}{V}(\hat{\underset{\sim}{P}})$ is an asymptotically unbiased estimate of the variance-covariance matrix of the $\{\hat{P}_i\}$. Thus confidence sets and tests of hypotheses concerning the $\{P_i\}$ can now be constructed. For the numerical example given in Section 4, Table III, the matrix $\underset{\sim}{V}(\hat{\underset{\sim}{P}})$ is (only the upper triangular part is shown):

$$\begin{pmatrix} 7.59 & 3.42 & 2.28 & 0.91 \\ & 5.98 & 3.98 & 1.60 \\ & & 5.05 & 2.02 \\ & & & 2.58 \end{pmatrix} \times 10^{-3}$$

7. The assumptions concerning the timing of the losses and late entries.

For the results up to now we have assumed that the μ_j late entries all occurred at the end of period $(t_{j-1},t_j]$, whereas the λ_{j-1} losses occurred at the beginning of this interval. In Section 2 we asserted that these assumptions were valid in many situations. However, if the λ_{j-1} losses occurred at the end of the interval (i.e. after all the δ_j deaths), then Step C in the algorithm of Section 4 should be modified by taking $q_j = (n_j' + \lambda_{j-1} - \delta_j')/(n_j' + \lambda_{j-1})$. Alternatively, if we assume that the losses occur randomly throughout the interval, we may take either the "joint risk" estimate, namely

$$q_j = \left(\frac{n_j' - \delta_j'}{n_j' + \lambda_{j-1}}\right)^{\delta_j'/(\delta_j'+\lambda_{j-1})}$$

or the "half-rule" estimate of:

$$q_j = \frac{n_j' - \delta_j' + \lambda_{j-1}/2}{n_j' + \lambda_{j-1}/2} .$$

For a full discussion of these estimates see Kaplan and Meier (1958, Section 4.1).

If the late entries are not brought to our attention only at times of inspection $\{t_i\}$, but can occur randomly throughout the intervals, then the estimates of Section 4 will no longer have the property of maximum likelihood. In this case some extra parametric assumption is needed. For instance, Harris, Meier and Tukey (1950), who look at the timing of events between inspections, assume that the instantaneous death rate is constant within each interval.

We claim, however, that if the width of the intervals can be chosen to be small, the estimates obtained by the iterative procedure of Section 4 will be approximately maximum likelihood, whatever the timing of the events.

8. A more general problem: arbitrary censoring.

A harder problem, considered by Harris, Meier and Tukey (1950) is one where the age of death is never known exactly, but is known only to fall in some interval, perhaps semi-infinite, where this interval differs from item to item. This idea was also discussed in Mantel (1967).

For each pair of times (t_i, t_j) $(i = 0,1,\ldots,m; j = i+1,\ldots,m+1)$, define μ_{ij} to be the number of deaths known to have occurred between t_i and t_j with $t_0 = 0$ and $t_{m+1} = \infty$. Thus, for example, $\mu_{i,m+1}$ represents the number of losses at age t_i and μ_{0i} is the number of late entries at t_i.

Essentially this was the model considered by Harris, Meier and Tukey. They derived maximum likelihood estimates assuming a constant instantaneous risk of death within each interval.

Here we propose a generalization of the procedure of Section 4 to derive estimates for the $P(t_i)$ $(1 \leq i \leq m)$. The iterative procedure is again based on the idea of "self-consistency" and is as follows:

A. Obtain starting values P_i^0, $(1 \leq i \leq m)$. For instance we can take $P_i^0 = n_i/n_1$ where $n_i = \Sigma_{j=i+1}^{m+1} \mu_{j-1,j}$ $(1 \leq i \leq m)$ i.e. we first consider only the deaths which we can observe precisely. Also take $P_0^0 = 1$, $P_{m+1}^0 = 0$.

B. Obtain estimates of the conditional probabilities, $\text{Prob}[t_{i-1} < T \leq t_i | t_a < T \leq t_b]$ by:

$$\alpha_{abi} = \frac{P^0_{i-1} - P^0_i}{P^0_a - P^0_b} \qquad (0 \le a \le m,\ a+1 \le b \le m+1,\ a+1 \le i \le b.)$$

C. Obtain the "adjusted" deaths in period $(t_{i-1},t_i]$ by:

$$\delta^1_i = \Sigma^{i-1}_{a=0} \Sigma^{m+1}_{b=i} \mu_{ab} \alpha_{abi} \qquad (1 \le i \le m+1)$$

D. Improved estimates will now be given by

$$P^1_i = \Sigma^{m+1}_{j=i+1} \delta^1_j/N \qquad (1 \le i \le m) \qquad (8.1)$$

where N is the total sample size. Also take $P^1_0 = 1$, $P^1_{m+1} = 0$.

E. Now return to Step B with $\{P^1_i\}$ replacing $\{P^0_i\}$ etc.

The $\{P^k_i\}$ will converge to values $\{\hat{P}_i\}$ which will have the self-consistency property. That is if each observation known only to be in the interval $(t_a,t_b]$, with $b > a$, is "reapportioned" to the intervals $(t_{i-1},t_i]$, $i = a+1,\ldots,b$ according to the conditional probabilities α_{abi}, then the $\{\hat{P}_i\}$, are the same as the binomial estimates using the adjusted deaths and given by (8.1).

Again assuming that the censoring times are independent of T, the time until death, the likelihood function is proportional to:

$$\prod_{a=0}^{m} \prod_{b=a+1}^{m+1} (P_a - P_b)^{\mu_{ab}} .$$

Setting the derivatives of the log-likelihood equal to zero, we obtain:

$$- \Sigma_{a=0}^{i-1} \frac{\mu_{ai}}{P_a - P_i} + \Sigma_{b=i+1}^{m+1} \frac{\mu_{ib}}{P_i - P_b} = 0 \qquad (8.2)$$

$$(i = 1,2,\ldots,m)$$

We conjecture that the iterative procedure described in this section leads to estimates $\{\hat{P}_i\}$ which satisfy the likelihood equations (8.2), and are in fact maximum likelihood.

9. Directions of future research.

In a later paper it is hoped to apply the self-consistent estimates derived in Section 4 to the two sample problem where two survivorship curves are to be compared and observations on one or both are doubly censored. This would be an extension of the method of Efron (1965, Section 8). It would then be of interest to compare this test with that of Gehan (1965b).

10. Acknowledgement.

I am grateful to Dr. Helena C. Kraemer who first showed me this problem.

REFERENCES

Bayley, N. (1969). Bayley scales for infant development. Psychological Corp., New York.

Berkson, J. and Gage, R.P. (1950). "Calculation of survival rates for cancer," Proceedings of the Staff Meetings of the Mayo Clinic 25, 270-286.

Boag, J.W. (1949). "Maximum likelihood estimates of the proportion of patients cured by cancer therapy," Journal of the Royal Statistical Society, Ser. B., 11, 15-53.

Efron, B. (1965). "The two sample problem with censored data," Proceedings of Fifth Berkeley Symposium on Mathematical Statistics and Probability, 4, 831-853, University of California Press.

Gehan, E.A. (1965a). "A generalized Wilcoxon test for comparing arbitrarily singly censored samples," Biometrika, 52, 203-223.

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

Gilbert, J.P. (1962). "Random censorship," unpublished Ph.D. thesis, University of Chicago.

Halperin, M. (1960). "Extension of the Wilcoxon-Mann-Whitney test to samples censored at the same fixed point," Journal of the American Statistical Association, 55, 125-138.

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

Herman, R.J. and Patell, R.K.N. (1971). "Maximum likelihood estimation for multi-risk model," Technometrics, 13, 385-396.

Kaplan, E.L. and Meier, P. (1958). "Nonparametric estimation from incomplete observations," Journal of the American Statistical Association, 53, 457-481.

Karlin, S. (1968). Total positivity I, Stanford University Press.

Leiderman, P.H., Babu, B., Kagia, J., Kraemer, H.C. and Leiderman, G.F. (1972). "African infant precocity: some social influences during the first year," to appear.

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

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

Moeschberger, M.L. and David, H.A. (1971). "Life tests under competing causes of failure and the theory of competing risks," Biometrics, 27, 909-933.

Myers, M.H. (1966). "A computing procedure for a significance test of the difference between two survival curves," Methodological Note No. 18, Biometry Branch, National Cancer Institute.

Appendix.

Here we prove two lemmas needed in the proof of the theorem of Section 5. The notation used is the same as that in Sections 4 and 5.

Lemma A1. Let $\underset{\sim}{P} = (P_1, P_2, \ldots, P_m)$ be the self-consistent estimates defined by (4.1). (We shall omit the caret signs.) Then $\underset{\sim}{P}$ satisfies the likelihood equations (5.1).

Proof. First note that if $\lambda_m, \delta_1, \ldots, \delta_m$ are positive then $1 > P_1 > \ldots > P_m > 0$. It is required to prove $\frac{\partial L}{\partial P_i} = 0$ for $i = 1,2,\ldots,m$. We do this first for $i = m$ and then proceed by induction to show it is true for $i = m-1, m-2, \ldots, 1$.

Now $P_m = P_{m-1} \cdot \frac{\lambda_m}{\lambda_m + \delta_m'}$, where $\delta_m' = \delta_m + \frac{P_{m-1} - P_m}{1-P_m} \cdot \mu_m$.

Therefore, substituting for δ_m', we obtain:

$$\lambda_m P_{m-1} = (\delta_m + \lambda_m) P_m + (P_{m-1} - P_m) \mu_m P_m / (1-P_m),$$

or

$$\frac{\delta_m}{P_{m-1} - P_m} - \frac{\lambda_m}{P_m} + \frac{\mu_m}{1-P_m} = 0.$$

Hence $\frac{\partial L}{\partial P_m} = 0$. For fixed $i < m$, assume $\frac{\partial L}{\partial P_j} = 0$ for $j = m, m-1, \ldots, i+1$.

Now $P_i = P_{i-1} \cdot \frac{n_i' - \delta_i'}{n_i'} = P_{i-1} \cdot \frac{n_{i+1}' + \lambda_i}{n_{i+1}' + \lambda_i + \delta_i'}$.

Substituting for $\delta_i' = \delta_i + (P_{i-1}-P_i)\Sigma_{j=i}^{m} \frac{\mu_j}{1-P_j}$ and rearranging terms, we obtain:

$$\frac{n'_{i+1}+\lambda_i}{P_i} - \frac{\delta_i}{P_{i-1}-P_i} - \Sigma_{j=i}^{m} \frac{\mu_j}{1-P_j} = 0 \tag{A1}$$

By the induction hypothesis, $\Sigma_{j=i+1}^{m} \frac{\partial L}{\partial P_j} = 0$, or

$$- \frac{\delta_{i+1}}{P_i-P_{i+1}} + \Sigma_{j=i+1}^{m} \frac{\lambda_j}{P_j} - \Sigma_{j=i+1}^{m} \frac{\mu_j}{1-P_j} = 0$$

Substituting in (A1), we obtain:

$$- \frac{\delta_i}{P_{i-1}-P_i} + \frac{\delta_{i+1}}{P_i-P_{i+1}} + \frac{\lambda_i}{P_i} - \frac{\mu_i}{1-P_i} + \frac{n'_{i+1}}{P_i} - \Sigma_{j=i+1}^{m} \frac{\lambda_j}{P_j} = 0 \tag{A2}$$

We now claim that, for $0 \le i \le m-1$:

$$\frac{n'_{i+1}}{P_i} - \Sigma_{j=i+1}^{m} \frac{\lambda_j}{P_j} = 0 \tag{A3}$$

Since $P_m = P_{m-1} \cdot \lambda_m/(\lambda_m+\delta'_m)$, we have $n'_m/P_{m-1} = \lambda_m/P_m$ and (A3) is true for $i = m-1$. Assume (A3) is true for $i = \ell$. Then $\Sigma_{j=\ell}^{m} \frac{\lambda_j}{P_j} = \frac{n'_{\ell+1}+\lambda_\ell}{P_\ell} = \frac{n'_\ell-\delta'_\ell}{P_\ell} = \frac{n'_\ell}{P_{\ell-1}}$. Thus (A3) is true for $i = \ell-1$ and by induction true for all $0 \le i \le m-1$.

Combining (A2) and (A3) we obtain:

$$\frac{\partial L}{\partial P_i} = -\frac{\delta_i}{P_{i-1}-P_i} + \frac{\delta_{i+1}}{P_i-P_{i+1}} + \frac{\lambda_i}{P_i} - \frac{\mu_i}{1-P_i} = 0.$$

The proof of Lemma A1 now follows by induction. □

Lemma 2. The matrix $\underset{\sim}{D}$ given by (5.2) is negative definite.

Proof. We will show that the matrix $\underset{\sim}{J} = -\underset{\sim}{D}$ is positive definite. For $j = 1,2,\ldots,m$, let J_j denote the value of the j'th leading principal minor. It suffices to show that $J_1, J_2, \ldots, J_m$ are all positive.

For $1 \leq i \leq m$, let $x_i = \delta_i/(P_{i-1}-P_i)^2$ and $y_i = (\lambda_i/P_i^2) + (\mu_i/(1-P_i)^2)$. Note that $y_i \geq 0$ and $x_i > 0$ since all $\delta_i > 0$ under the hypotheses of the theorem. Then by (5.2),

$$J = \begin{pmatrix} z_1 & -x_2 & 0 & \cdots & 0 \\ -x_2 & z_2 & -x_3 & \cdots & 0 \\ 0 & -x_3 & z_3 & \cdots & 0 \\ \vdots & \vdots & \vdots & & \vdots \\ 0 & 0 & 0 & \cdots & z_m \end{pmatrix}$$

where $z_i = x_{i+1} + x_i + y_i$ for $1 \leq i \leq m$, and $x_{m+1} = 0$.

Thus $J_1 = x_2+x_1+y_1 > 0$

$$J_2 = (x_3+x_2+y_2)(x_2+x_1+y_1) - x_2^2 > 0$$

and

$$J_i = (x_{i+1}+x_i+y_i)J_{i-1} - x_i^2 J_{i-2} \qquad (A4)$$

If we set $J_0 = 1$ and $J_{-1} = 0$, then (A4) holds for $i = 1,2,\ldots,m$. We proceed by induction. Assume $J_1, J_2, \ldots, J_{i-1}$ are all positive. Then using (A4) we have:

$$\begin{aligned} J_i &> x_i J_{i-1} - x_i^2 J_{i-2} \\ &= x_i(z_{i-1}J_{i-2} - x_{i-1}^2 J_{i-3} - x_i J_{i-2}) \\ &> x_i x_{i-1}(J_{i-2} - x_{i-1}J_{i-3}), \end{aligned}$$

since $J_{i-2} > 0$, $x_{i-1} > 0$.

Iterating, we have:

$$\begin{aligned} J_i &> x_i x_{i-1} \cdots x_1 (J_0 - x_0 J_{-1}) \\ &= x_i x_{i-1} \cdots x_1 > 0. \end{aligned}$$

The proof of Lemma A2 now follows by induction. □