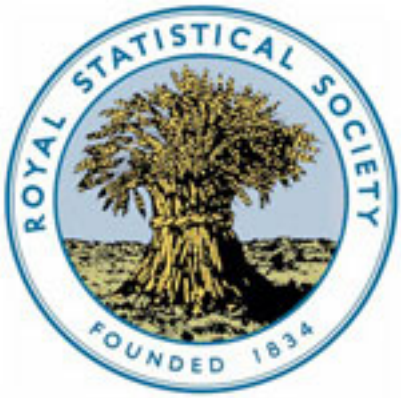

Experimental Survival Curves for Interval-censored Data

By RICHARD PETO

R.P.M. Radcliffe Infirmary, Oxford University

SUMMARY

A method is given for calculating from interval-censored data an estimate of the c.d.f. which is analogous to the estimate derivable from right-censored data by the life-table technique. A Fortran implementation has been constructed by the author.

Keywords: EXPERIMENTAL SURVIVAL CURVE; LIFE TABLE; PRODUCT-LIMIT; RANK TEST; ASYMPTOTIC EFFICIENCY; CENSORING; INTERVAL-CENSORING; RIGHT-CENSORING; LOGRANK; NEWTON–RAPHSON; MAXIMUM-LIKELIHOOD; PROGRAMMED SEARCH; EVENT RATE; SURVIVORSHIP FUNCTION

1. DEFINITIONS

(i) LET z be a real-valued random variable with c.d.f. $F(z)$. z is said to be *censored* into a non-zero interval I if the only information we have about z is that z lies in I. If I is the interval $[T, \infty]$, z is said to be *right-censored.* If z is not censored then z is *exact.*

Example. Two examinations at particular times to see whether a certain event has yet occurred will produce a *censored observation* of that event: whether the observation is left-censored, right-censored or interval-censored depends on whether the event happened before the first examination, after the second examination or between the two examinations.

(ii) Let $G(z) = 1 - F(z)$. $G(z)$ is the *survival curve* for z.

(iii) Let $z_1 \ldots z_N$ be independent random variables with the common survival curve $G(z)$. Suppose we have data consisting of an observation (exact or censored) of each z_i. Let the interval into which z_i is censored be written $[L_i, R_i]$ (where $L_i = R_i$ if z_i was observed exactly). Our data are thus $\mathbf{L} = L_1, \ldots, L_N$ and $\mathbf{R} = R_1, \ldots, R_N$. Under the survival curve G, the likelihood for the ith observation is

$$\{G(L_i - 0) - G(R_i + 0)\}$$

and the likelihood for all the data is a product of N such terms. This overall likelihood depends not only on the data, $\mathbf{L}$ and $\mathbf{R}$, but also on G. If we assumed a different (monotonic) function G^* instead of G, the value of the overall likelihood might be changed. Since each separate likelihood must lie in $[0, 1]$ so must the overall product of them, and so if G^* is a monotonic function which decreases from 1 to 0 the overall likelihood for the data must lie in $[0, 1]$. Different forms for G^* will give different values in $[0, 1]$ for the overall likelihood, and the *experimental survival curve* (e.s.c.) for the data is defined as the monotonic function $H(z)$ which maximizes the overall likelihood. With censored data, the experimental survival curve may not be unique since if a censored observation lies in an interval I then the likelihood for that

observation depends only on the difference between the experimental survival curve values at the end-points of that interval and not at all on the detailed behaviour within the interval.

Example. For completely uncensored (exact) data, the e.s.c. is undefined at the actual observation values, and between them $H(z)$ can be shown to equal the proportion of the observations that exceed z. H is discontinuous, so the likelihood for each observation is finite rather than infinitesimal. Fig. 1 gives an example of a more complex e.s.c.

2. Introduction

Kaplan and Meier (1958) have shown that for data subject only to right censoring the ordinary (product-limit) life-table is identical with the experimental survival curve. A special case of the life-table is for completely uncensored data, when every observation is known exactly: the experimental survival curve $H(z)$ is then given by the proportion of the observations that exceed z. Kaplan and Meier's method can easily be reversed to get the experimental survival curve by a life-table technique for left-censored data and it can be adapted slightly to get the experimental survival curve in certain special cases of interval-censored data.† However, it cannot be adapted to deal with the completely general situation where each observation may be exact or censored into its own particular interval, and some other technique is then required. No algebraic solution with the simplicity of Kaplan and Meier's life-table technique has emerged in the general situation, and the most obvious technique is to write down the total log likelihood as a function of the end-points of each censoring interval (treating an exact observation of x as an observation censored into the closed interval $[x, x]$, this equals the sum for all subjects of

$$\log \{G(\text{left end-point} - 0) - G(\text{right end-point} + 0)\}),$$

and to find the values that maximize this total log likelihood by a programmed search. This has proved practicable with quite large data sets: the work involved eventually increases as the square of the number of data points involved.

3. Example

Annual surveys on 196 girls recorded whether or not, at the time of the survey, sexual maturity had developed. Development was complete in some girls before the first survey, some girls were lost before the last survey and before development was complete, and some girls had not completed development at the last survey. An estimate was required of the proportion who were not yet mature as a function of age without assuming a normal distribution for the time of development. (Data provided by Dr L. A. Malcolm, Regional Health Office, New Guinea, through courtesy of Professor J. M. Tanner, Institute of Child Health, London.) The experimental survival curve was calculated, and is shown in Fig. 1.

4. Theory Underlying the Programmed Search

We have N independent observations, $z_1, \ldots, z_N$. Our knowledge about z_i is restricted to the fact that $L_i \leqslant z_i \leqslant R_i$. If z_i is exact, $L_i = R_i$, and if z_i is right- or left-censored, $R_i = \infty$ or $L_i = -\infty$. From the sets $\{L_i\}$ and $\{R_i\}$ we can derive all the

† For example, if an ordered set of disjoint intervals exists such that each censoring interval contains one interval I in this set, nothing of any of the intervals before I and either all or nothing of the intervals after I.

distinct closed intervals whose left and right end-points lie in the sets $\{L_i\}$ and $\{R_i\}$ respectively and which contain no members of $\{L_i\}$ or $\{R_i\}$ other than at their left and right end-points respectively. Let these intervals be written in order as

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

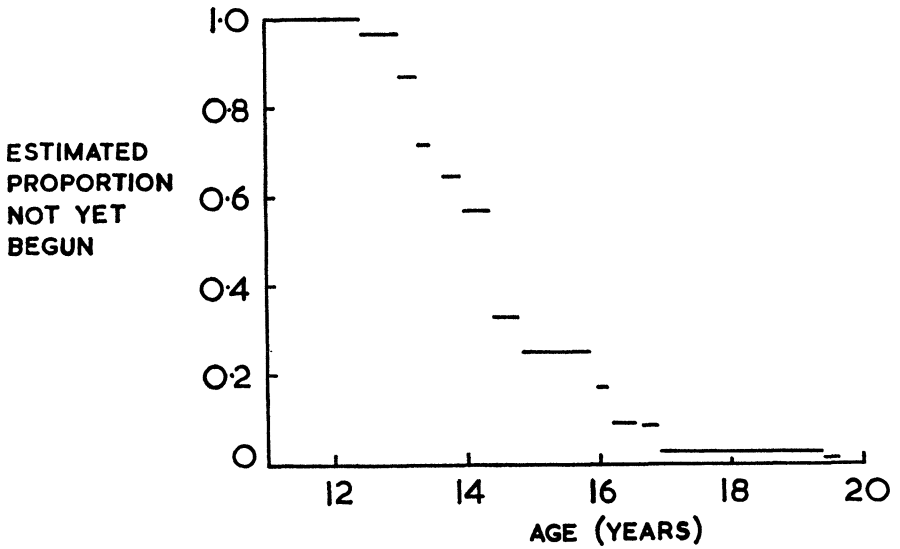

FIG. 1. Onset of puberty: New Guinea females. E.S.C. from interval-censored observations of 196 subjects.

It is sufficient, in our search for the experimental survival curve (e.s.c.) to consider only survival curves which are horizontal everywhere except in these intervals and which decrease in some or all of these intervals. (This is proved in the Appendix.) Moreover, the total likelihood is a function only of the survival curve decreases in these intervals and is independent of how the decreases actually occur, so the e.s.c. is undefined in each $[p_i, q_i]$ and is well defined and flat between these intervals. If we write the sizes of the decreases of a survival curve in these intervals as $s_1, s_2, \ldots, s_m$ (where $s_m = 1 - \sum_{i=1}^{m-1} s_i$), the total likelihood is a function of $s_1, \ldots, s_{m-1}$ and, within the indeterminacy implicit in the e.s.c. being undefined in the intervals in which it decreases, the problem of finding the e.s.c. has been reduced to the finite-dimensional problem of maximizing a function of $s_1 \ldots s_{m-1}$ subject to $s_i \geqslant 0$ and $1 - \sum_{i=1}^{m-1} s_i \geqslant 0$. Except on those of the boundaries of this region on which the likelihood function is zero, the total log likelihood as a function of $s_1, \ldots, s_{m-1}$ is strictly convex, so the values of $s_1, \ldots, s_{m-1}$ that maximize it are unique and can be found by any efficient search algorithm that can be made to respect the m boundaries. Because of the easy availability of all the first and second derivatives, it is possible to use a suitably constrained Newton–Raphson search to locate the absolute maximum of the log likelihood function. (The search technique devised would, of course, be applicable in any other problem involving maximization of any other convex function of direction in the closed positive quadrant of m-space.)

5. Plotting the E.S.C.

The e.s.c. is defined only between the intervals in which it decreases, and so an e.s.c. plot consists of a decreasing sequence of horizontal lines. There may be small blank regions between the end of one line and the beginning of the next (e.g. around 13·5 in Fig. 1) in which the e.s.c. is undefined. This plotted e.s.c. is the analogue for interval-censored data of the life-table for right-censored data, and has analogous advantages and disadvantages when compared with the c.d.f. of a fitted model from a parametric family of distributions such as the normal (Swan, 1969) or exponential. The inverse of the negative matrix of second derivatives of the log-likelihood function with respect to the non-zero elements of $s_1, \ldots, s_{m-1}$ estimates their variance/covariance matrix. From this, the variance of the sum of any particular subset of the non-zero elements of $s_1, \ldots, s_{m-1}$ can be estimated. The position of each line in an e.s.c. plot is unity minus such a sum, so standard errors can be estimated for the position of any line in an e.s.c. plot. These standard errors should, however, only be used descriptively and not in a statistical test, since asymptotically efficient rank invariant methods exist (Peto and Peto, 1972) for detecting differences between groups of independent interval-censored observations.

6. Reducing the Dimension of the Search

If the data are very extensive, the dimension of the search for the e.s.c. may be excessive and it can then be reduced by coarsening the data somewhat by grouping. Search dimensions larger than a hundred or so will very rarely be needed.

Example of dimension reduction

If the development of radiologically visible metastases occurs within a year or so of first diagnosis of a particular cancer and X-ray photographs from several thousand patients are available at various times after diagnosis, then the dimension of the search for the e.s.c. of the distribution of the time to the first visible metastasis will be reduced from a few thousand to around fifty if the times at which the X-ray photographs were taken are recorded in weeks rather than hours from first diagnosis. The rounding can be made finer in regions of particular interest (e.g. immediately after diagnosis) if necessary.

Strictly, left end-points of censoring intervals should be rounded downwards and right end-points upwards. However, if this is being done cross-overs must be avoided; thus, if left end-points are being rounded down to ·000, right end-points should be rounded up to ·999 rather than to ·000.

7. Practical Details of a Suitably Constrained Newton-Raphson Search Algorithm

7.1. *General*

If $m = 1$ then the solution is $s_1 = 1$. Otherwise, we have a function of m (at least two) non-negative step sizes $s_1, \ldots, s_m$ to maximize. Each likelihood is a sum of certain of the s_i, and so the total likelihood may be zero only if one or more of the s_i are zero, and maybe not even then. Write $(s_1, \ldots, s_m) = \mathbf{x}$ and $(s_1, \ldots, s_{m-1}) = \mathbf{y}$. Since $\sum s_i = 1$, the value of $\mathbf{y}$ implies the value of $\mathbf{x}$ and the total log likelihood to be maximized can either be regarded as a function $f(\mathbf{x})$ of $\mathbf{x}$ or as a function $g(\mathbf{y})$ of $\mathbf{y}$. Because $g(\mathbf{y})$ is convex, the latter is more convenient. The possible values of $\mathbf{y}$ are restricted by the conditions $s_i \geqslant 0$ to that part of the positive quadrant $y_i \geqslant 0$ in which

$\sum \mathbf{y}_i \leqslant 1$, and the possible values of $\mathbf{y}$ are still further restricted (by the total likelihood having to be positive rather than zero) to avoid any of the boundaries of this region where all of the values of s_i contributing to a particular likelihood vanish. (This incidentally implies $y_1 > 0$ and $\sum y_i < 1$.)

7.2. *Newton-Raphson*

Consider movement away from $\mathbf{y}$ with certain components of $\mathbf{y}$ constrained to be unaltered in this movement. If the first and second derivatives of $g(\mathbf{y})$ in the unconstrained dimensions of $\mathbf{y}$ are the vector $\mathbf{D}$ and the matrix $-\mathbf{H}$, then the Newton–Raphson search for the suitably constrained value of $\mathbf{y}$ which maximizes $g(\mathbf{y})$ involves going to $\mathbf{y}+\mathbf{D}.\mathbf{H}^{-1}$ (or, if that does not increase g, going a suitable fraction of the way towards $\mathbf{y}+\mathbf{D}.\mathbf{H}^{-1}$), re-evaluating $\mathbf{D}$ and $\mathbf{H}$ at the new point and continuing until the elements of $\mathbf{D}$ are negligible. This is the basis of the present search algorithm. At the point $\mathbf{y}$, the set $B(\mathbf{y}) = \{i \,|\, y_i = 0\}$ of boundaries we are on can be divided into $B^+(\mathbf{y}) = \{i \,|\, y_i = 0 \text{ and } \partial g/\partial y_i > 0\}$ and $B^-(\mathbf{y}) = \{i \,|\, y_i = 0 \text{ and } \partial g/\partial y_i \leqslant 0\}$. Movement off the boundaries $i \in B^-$ initially decreases g, and so we derive the Newton–Raphson step with the elements $y_i \,|\, i \in B^-$ held constant at zero. If any of the changes suggested by this derived step to any of the elements $y_i \,|\, i \in B^+$ are negative, we modify the step by setting these changes to zero. If the step violates any distant boundaries (i.e. boundaries not in $B(\mathbf{y})$) we reduce the step length by a suitable factor so that the nearest of the distant violated boundaries is reached but is not crossed. We now evaluate the function at the new point. If it is worse than at the old point (either because the new point is on a boundary which gives a zero-likelihood function or because we have gone over the optimum and down on the other side), we successively halve the step length until, as must eventually occur, the function value at the new point is better than at the old point. This process is repeated indefinitely until we are at the unique absolute function maximum $\hat{\mathbf{y}}$ (the necessary and sufficient condition for which is that $\partial g/\partial y_i$ is small for all i not in $B^-(\hat{\mathbf{y}})$).

7.3. *Calculation of Derivatives and Step Lengths*

The partial derivatives of g with respect to $\mathbf{y}$ are easily derived from the partial derivatives of f with respect to $\mathbf{x}$ by

$$\partial g/\partial y_i = \partial f/\partial x_i - \partial f/\partial x_m$$

and

$$\partial^2 g/\partial y_i \, \partial y_j = \partial^2 f/\partial x_i \, x_j - \partial^2 f/\partial x_i \, \partial x_m - \partial^2 f/\partial x_m \, \partial x_j + \partial^2 f/\partial x_m^2.$$

The constraints of the changes in $y_i \,|\, i \in B^-$ to be zero can either be effected by reducing the vector $\mathbf{D}$ of first derivatives and the negative matrix $\mathbf{H}$ of second derivatives, inverting the matrix, multiplying them together and re-expanding the product vector with zeros or equivalently by overwriting the original vector and matrix with zeros in the elements to be reduced out (except for the diagonal elements, which become unity), inverting and multiplying. The latter is easier to program.

The calculation of the partial derivatives of f with respect to $\mathbf{x}$ is straightforward. If the e.s.c. steps that contribute to the likelihood l_i for the ith individual are s_j, $F_i \leqslant j \leqslant G_i$, then

$$l_i = \sum_{j=F_i}^{j=G_i} x_j.$$

The first derivative of $\log l_i$ with respect to x_j is thus either $1/l_i$ or zero according to whether $F_i \leqslant j \leqslant G_i$ or not, and $-\partial^2 \log l_i / \partial x_j \partial x_k$ is either $1/l_i^2$ or zero, depending on the values of j and k. The first and second derivatives of the total log-likelihood function $\Sigma \log l_i$ are therefore sums of such terms, and can conveniently be calculated by clearing **D** and **H** and then for each l_i calculating $1/l_i$ and $1/l_i^2$ and accumulating them into the appropriate elements of **D** and **H**.

Appendix

Proof that the e.s.c. is horizontal everywhere except in the intervals $[q_i, p_i]$: Define a set of points $r_1, \ldots, r_{m-1}$ where r_i is some value greater than all the right and less than all the left end-points in $[p_i, q_{i+1}]$. Now if $G(z)$ is a survival curve which is not flat outside the $[q_i, p_i]$ then if $G^*(z)$ is a survival curve which is flat outside these intervals with $G^*(q_1) = 1$, $G^*(p_m) = 0$ and $G^*(p_j) = G^*(q_{j+1}) = G(r_j)$, then the total likelihood under G^* is greater than that under G and G cannot therefore be the e.s.c.

Acknowledgement

I am very grateful to the director, Mr J. Howlett, and the staff of the S.R.C. computer at Chilton for the extensive use of computing facilities, without which this study could not have succeeded.

References

Kaplan, E. L. and Meier, P. (1958). Nonparametric estimation from incomplete observations. *J. Amer. Statist. Ass.*, **53**, 457–481.

Peto, Richard and Peto, Julian (1972). Asymptotically efficient rank invariant test procedures. *J. R. Statist. Soc.* A, **135**, 185–207.

Swan, A. V. (1969). Computing maximum-likelihood estimates for parameters of the normal distribution from grouped and censored data. *Appl. Statist.*, **18**, 65–69.