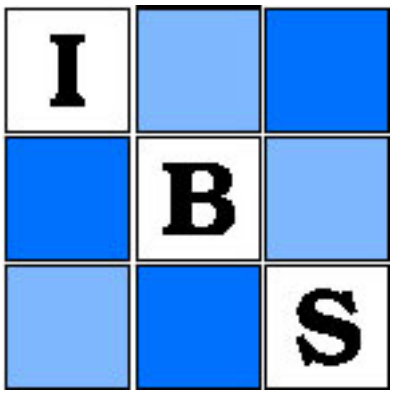

BIOMETRICS 40, 587–600
September 1984

A Comparison of Reflected Versus Test-Based Confidence Intervals for the Median Survival Time, Based on Censored Data

Eric V. Slud

Department of Mathematics, University of Maryland, College Park, Maryland 20742, U.S.A.

David P. Byar and Sylvan B. Green

Clinical and Diagnostic Trials Section, National Cancer Institute, Landow Building, Bethesda, Maryland 20205, U.S.A.

SUMMARY

The small-sample performance of some recently proposed nonparametric methods of constructing confidence intervals for the median survival time, based on randomly right-censored data, is compared with that of two new methods. Most of these methods are equivalent for large samples. All proposed intervals are either 'test-based' or 'reflected' intervals, in the sense defined in the paper. Coverage probabilities for the interval estimates were obtained by exact calculation for uncensored data, and by simulation for three life distributions and four censoring patterns. In the range of situations studied, 'test-based' methods often have less than nominal coverage, while the coverage of the new 'reflected' confidence intervals is closer to nominal (although somewhat conservative), and these intervals are easy to compute.

1. Introduction

Estimates of median survival times are frequently presented in medical reports to characterize the survival experience of groups of patients. For censored data this measure is certainly preferable to estimated mean survival as it is easy to estimate and has a clear interpretation; however, biostatisticians (for example Peto *et al.*, 1977) have pointed out that estimates of median survival can be seriously misleading because of their relatively large variation, particularly if the survival curve does not change rapidly near the median. Despite this warning, published reports containing confidence intervals for median survival are unusual, probably because methods for computing such intervals have not been generally available.

The recent papers of Brookmeyer and Crowley (1982a), Efron (1981), Emerson (1982), Reid (1981) and Simon and Lee (1982) have provided a number of competing nonparametric interval estimators for the median survival time from randomly right-censored data. Each paper illustrates the properties of its interval estimate with a small-sample (typically 25–75 observations) simulation study. However, there is no guidance in these papers on how to choose among the newly available methods of constructing confidence intervals for the median survival time in a small clinical study. In addition, new proposals for confidence bands for the Kaplan–Meier estimated survival curve (see Kaplan and Meier, 1958) lead

Key words: Median survival time; Independent right-censoring; Confidence interval; Test-based interval; Reflected interval; Kaplan–Meier survival curve.

naturally to new confidence intervals for quantiles of the survival distribution, as suggested in the recent work of Anderson, Bernstein and Pike (1982).

For censored data, Brookmeyer and Crowley (1982a), Emerson (1982), and Simon and Lee (1982) have all presented methods that involve the inversion of a generalized sign test for the location of the median survival time. This paper distinguishes between the method of Emerson (1982)—whose confidence interval exemplifies what we call a 'reflected' interval—and the 'test-based' 'intervals of Brookmeyer and Crowley (1982a) and Simon and Lee (1982). After drawing the general distinction in §2 between test-based and reflected intervals, we introduce two additional reflected intervals: the first mentioned by Efron (1981, §6, with acknowledgement to K. Bailey), and the second a new proposal of our own involving a transformation of the time scale. The 'bootstrapped confidence interval' proposed by Efron requires extensive Monte Carlo simulation in its definition and, as discussed in §2, turns out to be closely related to the Brookmeyer–Crowley test-based interval. The 'conditional bootstrap' interval of Reid (1981) is readily computed without simulations and also has a natural interpretation as a test-based interval.

The purpose of the present work is first to compare the nonparametric confidence intervals for median survival under the headings of 'reflected' and 'test-based' intervals; second, to advance the new reflected intervals (simple reflected and transformed reflected) in light of their success on small censored samples; third, to show the asymptotic large-sample equivalence of these reflected intervals to the interval of Brookmeyer and Crowley (1982a); and finally, to present exact calculations illustrating performance on uncensored data, simulation results for small-sample censored data, and recommendations as to which methods to use for small samples of patients. In §2 we give notations and definitions for all proposed methods. Large-sample considerations are briefly dealt with in §3, while §4 contains a discussion of the special modifications and the occasional 'fix-ups' all methods may need when samples are small or heavily censored. Section 5 describes analytically the properties of confidence intervals when there is no censoring. Our simulation experiments are summarized in §6 and are used to compare the intervals for various life and censoring distributions. Section 7 contains a worked example. Discussion and recommendations appear in §8.

2. Notation and Definitions

We assume throughout that the observed survival data (T_i, Δ_i), $i = 1, \ldots, N$, are generated from independent pairs of independent death and censoring times (X_i, Y_i), where $T_i \equiv \min(X_i, Y_i)$ and where $\Delta_i = 1$ if $X_i \leq Y_i$ and $\Delta_i = 0$ if $X_i > Y_i$. We write $S(t) \equiv \text{pr}(X > t)$ and $S_Y(t) \equiv \text{pr}(Y > t)$, and assume that $S(\cdot)$ is continuous with density $f(\cdot)$ and cumulative hazard $\Lambda(\cdot)$. Then there will be no tied death times, X_i and X_j, or simultaneous death and censoring times, X_i and Y_j, and the median, μ, of X_i is uniquely defined as $S^{-1}(\frac{1}{2})$. For such data, the survival curve, S, has the approximately unbiased (Efron, 1967) nonparametric estimator due to Kaplan and Meier (1958)

$$\hat{S}(t) \equiv \prod_{j:T_j \leq t} (1 - \Delta_j/r_j),$$

where $r_j \equiv$ number of T_i which are greater than or equal to T_j, $i = 1, \ldots, N$. For large N, the asymptotic variance of $\hat{S}(t)$ is known to be

$$\phi_N(t) \equiv -N^{-1}S(t)^2 \int_0^t \{S(x)^2 S_Y(x)\}^{-1}\, dS(x)$$

(Breslow and Crowley, 1974) which is consistently estimated (see Brookmeyer and

Crowley, 1982a) by the Greenwood formula

$$\hat{\phi}_G(t) = \hat{S}(t)^2 \sum_{j:T_j \leqslant t} \Delta_j r_j^{-1}(r_j - 1)^{-1}.$$

Since $S(\mu) = \frac{1}{2}$, we define $\tilde{\phi}_G(t) = \frac{1}{4} \sum_{j:T_j \leqslant t} \Delta_j r_j^{-1}(r_j - 1)^{-1}$ to estimate variance near the median. Peto *et al.* (1977, p. 37, Statistical Note 6) proposed a simplified variance estimator in which $\hat{S}(t)$ is treated as an estimate of a binomial parameter based on an effective sample size, $r_i/\hat{S}(t)$, where $i \equiv \max(j{:}\Delta_j = 1,\ T_j \leqslant t)$. Specifically, Peto *et al.* estimated $\operatorname{var}\{\hat{S}(t)\}$ by $\hat{\phi}_P(t) = \hat{S}(t)^2\{1 - \hat{S}(t)\}/r_i$. Depending on the censoring distribution $S_Y(\cdot)$, $\hat{\phi}_P(t)$ can be seriously biased, as we show in §3. Simon and Lee (1982) estimated $\phi_N(t)$ by $\tilde{\phi}_P(t) \equiv \hat{S}(t)/\{4(r_i - 1)\}$, which is always larger than $\hat{\phi}_P(t)$.

For future reference, we define an estimate of the cumulative hazard (Nelson, 1972):

$$\hat{\Lambda}(t) \equiv \sum_{j:T_j \leqslant t} \Delta_j r_j^{-1}.$$

Another possible estimator for $\Lambda(t)$ is $-\ln \hat{S}(t) \equiv \hat{H}(t)$; however, for constructing a confidence interval on a transformed time scale, $\hat{H}$ seems less effective than $\hat{\Lambda}$, as we point out in connection with Table 2. The variance of $\hat{\Lambda}(t)$ can be estimated (see §3) by $\hat{\phi}_G(t)/\hat{S}(t)^2$, which equals 4 $\tilde{\phi}_G(t)$. Whenever we refer to inverses of the right-continuous step functions $\hat{S}$ and $\hat{\Lambda}$, we mean the inverses defined as $\hat{S}^{-1}(x) \equiv \inf\{t{:}\hat{S}(t) \leqslant x\}$ and $\hat{\Lambda}^{-1}(x) \equiv \inf\{t{:}\hat{\Lambda}(t) \geqslant x\}$. We adopt the estimator $\hat{\mu} = \hat{S}^{-1}(\frac{1}{2})$ for the median μ; this estimator is used by all the authors we cite except Simon and Lee (1982), who estimated μ by inverting a linearly interpolated $\hat{S}$.

Our notation for upper binomial tail probabilities is

$$\bar{B}(k, N, p) \equiv \sum_{j=k}^{N} \binom{N}{j} p^j (1 - p)^{N-j}.$$

Then for noninteger y, we define an interpolated tail probability

$$\tilde{B}(y, N, p) \equiv ([y] + 1 - y)\bar{B}([y], N, p) + (y - [y])\bar{B}([y] + 1, N, p),$$

where $[y]$ denotes the largest integer less than or equal to y.

We now distinguish the general notions of reflected and test-based confidence intervals for the median survival time μ. Suppose that for each t, $\{\hat{S}_1(t), \hat{S}_2(t)\}$ is a level-α confidence interval for $S(t)$. We describe as 'test-based' a confidence interval of the form $[t{:}\ \frac{1}{2} \in \{\hat{S}_1(t), \hat{S}_2(t)\}]$, while the corresponding 'reflected' interval is $[t{:}\ \hat{S}(t) \in \{\hat{S}_1(\hat{\mu}), \hat{S}_2(\hat{\mu})\}]$. The difference between these two approaches to the construction of confidence intervals is best seen graphically (Fig. 1) when survival-curve estimators are drawn as continuous curves for clarity. The lower endpoint, R_1, of the reflected interval is obtained by horizontally projecting the upper confidence bound $\hat{S}_2(\hat{\mu})$, until it meets $\hat{S}(\cdot)$; the t corresponding to the meeting point is taken as R_1, that is $R_1 = \hat{S}^{-1}\{\hat{S}_2(\hat{\mu})\}$. An analogous procedure using $\hat{S}_1(\hat{\mu})$ defines R_2. For this reason the reflected interval requires calculation of $\hat{S}_1(t)$ and $\hat{S}_2(t)$ only at $t = \hat{\mu}$. By contrast, the test-based interval requires knowledge of $\hat{S}_1(\cdot)$ and $\hat{S}_2(\cdot)$ at many time points. The lower limit, T_1, is the time t at which $\hat{S}_1(t) = \frac{1}{2}$, and T_2 that at which $\hat{S}_2(t) = \frac{1}{2}$.

Let $\hat{S}_1(t) \equiv \hat{S}(t) - z_{\alpha/2}\{\hat{\phi}(t)\}^{1/2}$ and $\hat{S}_2(t) \equiv \hat{S}(t) + z_{\alpha/2}\{\hat{\phi}(t)\}^{1/2}$, where $\hat{\phi}(t)$ is an estimator of the asymptotic variance $\phi_N(t)$. When $\hat{\phi}(t)$ is taken to be $\hat{\phi}_G(t)$ or $\tilde{\phi}_P(t)$, the test-based interval for μ is, respectively, the confidence interval defined by Brookmeyer and Crowley (1982a) or that defined by Simon and Lee (1982). We refer to the corresponding reflected interval with $\hat{\phi}(t) \equiv \tilde{\phi}_G(t)$ and $\{\hat{S}_1(\hat{\mu}), \hat{S}_2(\hat{\mu})\} \equiv [\frac{1}{2} \pm z_{\alpha/2}\{\tilde{\phi}_G(\hat{\mu})\}^{1/2}]$ as the 'simple reflected

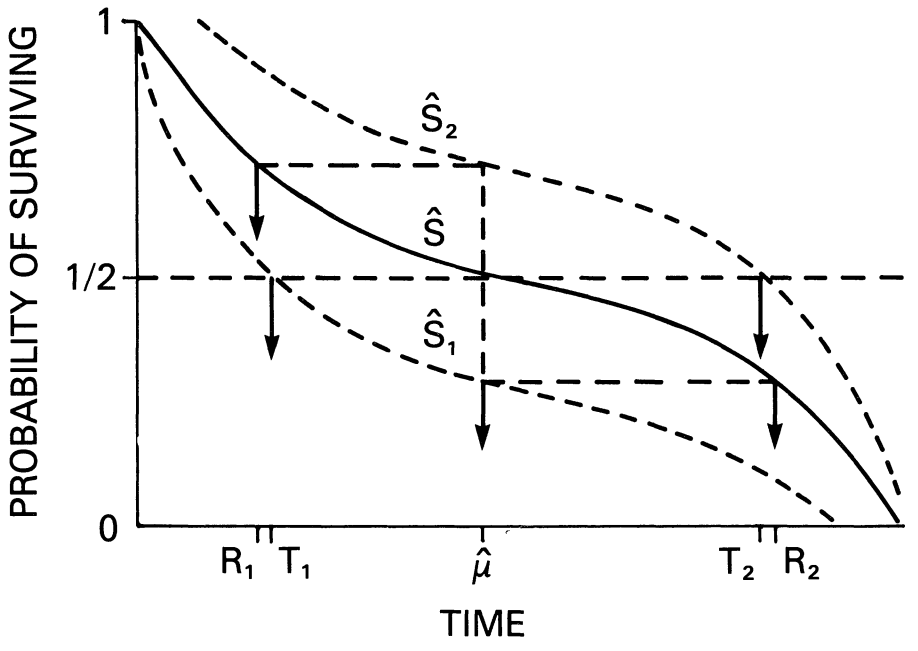

Figure 1. Schematic representation of test-based and reflected intervals. T_i are endpoints of test-based interval; R_i are endpoints of reflected interval.

interval' due to Bailey and Efron (see Efron 1981, §6), although they defined only the reflected interval with respect to the bootstrap variance estimator $\phi^*(t)$.

Next consider $\{\hat{S}_1(t), \hat{S}_2(t)\}$, given as

$$[p\colon 0 \leqslant p \leqslant 1, \quad \tilde{B}\{Np, N, \hat{S}(t)\} \geqslant \tfrac{1}{2}\alpha, \tilde{B}\{N(1-p), N, 1-\hat{S}(t)\} \geqslant \tfrac{1}{2}\alpha].$$

At $t = \hat{\mu}$ and with $\hat{S}(\hat{\mu})$ replaced by $\frac{1}{2}$, the reflected interval for μ in this setting is (except for endpoints) the interval proposed by Emerson (1982). Similarly, if

$$\{\hat{S}_1(t), \hat{S}_2(t)\} \equiv [p\colon 0 \leqslant p \leqslant 1, \tfrac{1}{2}\alpha \leqslant \bar{B}\{N(1-p), N, 1-\hat{S}(t)\} \leqslant 1-\tfrac{1}{2}\alpha],$$

the test-based interval is the unsmoothed 'conditional bootstrap' interval for μ, due to Reid (1981). With regard to the unconditional bootstrap interval, Efron (1981, §3) argued that the Greenwood estimate $\hat{\phi}_G(t)$ agrees closely with the bootstrap variance estimate $\phi^*(t)$ for $\hat{S}(t)$, and moreover [see Efron, 1981, Equation (6.1)] that the bootstrap interval for μ is equal to the set of t for which the bootstrapped confidence interval for $S(t)$ contains $\frac{1}{2}$. In other words, Efron's bootstrap interval for μ is test-based [with bootstrap confidence bands for $S(t)$] and, if N is large enough so that $\hat{S}(\mu)$ is approximately normally distributed, should agree closely with the test-based interval due to Brookmeyer and Crowley (1982a), defined above.

Other confidence intervals $\{\hat{S}_1(t), \hat{S}_2(t)\}$ for $S(t)$ will correspond to new test-based and reflected confidence intervals for μ. In fact, Anderson *et al.* (1982) have recently proposed test-based intervals for survival quantiles based on (some simple transformations of) the interval for $S(\cdot)$ due to Rothman (1978). Although these authors have shown that their intervals for μ will have proper coverage for large censored samples, they have not published results concerning small-sample performance.

Reflected and test-based intervals can be defined as well in terms of the cumulative hazard function $\Lambda(t)$ as in terms of $S(t)$. We propose a new 'transformed reflected' interval for μ defined as follows: let $\{\hat{\Lambda}_1(t), \hat{\Lambda}_2(t)\} \equiv [\hat{\Lambda}(t) \pm 2z_{\alpha/2}\{\tilde{\phi}_G(t)\}^{1/2}]$ be an approximate level-α confidence interval for $\Lambda(t)$; then the new reflected interval is $[t\colon \hat{\Lambda}(t) \in \{\hat{\Lambda}_1(\hat{\mu}), \hat{\Lambda}_2(\hat{\mu})\}]$.

To close this section, we list the six competing interval-estimators for μ:

Reflected intervals

$I_1 = [t\colon \{\hat{S}(t) - \frac{1}{2}\}^2 \leqslant \chi^2_{1,\alpha}\,\tilde{\phi}_G(\hat{\mu})]$ (simple reflected; Efron, 1981);

$I_2 = (t\colon \tilde{B}\{N\hat{S}(t), N, \frac{1}{2}\} \geqslant \frac{1}{2}\alpha,\ \tilde{B}[N\{1 - \hat{S}(t)\}, N, \frac{1}{2}] \geqslant \frac{1}{2}\alpha)$ (Emerson, 1982);

$I_3 = [t\colon \{\hat{\Lambda}(t) - \hat{\Lambda}(\hat{\mu})\}^2 \leqslant 4\chi^2_{1,\alpha}\,\tilde{\phi}_G(\hat{\mu})]$ (transformed reflected);

Test-based intervals

$I_4 = [t\colon \{\hat{S}(t) - \frac{1}{2}\}^2 \leqslant \chi^2_{1,\alpha}\hat{\phi}_G(t)]$ (Brookmeyer and Crowley, 1982a);

$I_5 = [t\colon \{\hat{S}(t) - \frac{1}{2}\}^2 \leqslant \chi^2_{1,\alpha}\tilde{\phi}_p(t)]$ (Simon and Lee, 1982);

$I_6 = [t\colon \frac{1}{2}\alpha \leqslant \bar{B}\{\frac{1}{2}(N+1), N, 1 - \hat{S}(t)\} \leqslant 1 - \frac{1}{2}\alpha]$ (unsmoothed; Reid, 1981).

Our definitions differ from those of previous authors only with regard to inclusion of endpoints. Reid (1981) also defined an interval based on a *smoothed* conditional bootstrap distribution, which we denote by I_6' and describe in §4.

3. Large-Sample Considerations

In subsequent sections, we compare the performance of the various methods of forming confidence intervals, and concentrate on small-sample properties. In the present section, we demonstrate that intervals I_1, I_3 and I_4 are asymptotically (in large samples) equivalent with proper (i.e. nominal) coverage, that I_5 is asymptotically conservative (coverage greater than nominal) when there is censoring, and that I_2 and I_6 are asymptotically equivalent and strictly anticonservative when there is censoring. All methods are equivalent in the absence of censoring. Here 'asymptotic equivalence of I_k and I_j' means that as $N \to \infty$, the events $\mu \in I_k$ and $\mu \in I_j$ differ by an event of asymptotically negligible probability.

For large N, Sander, (in Technical Report No. 5, Biostatics, Stanford University, 1975) showed that $\hat{\mu}$ is asymptotically normal with mean μ and variance $\phi_N(\mu)/f^2(\mu)$ if $f(\mu) \neq 0$, and that $f(\cdot)$ is continuous in a neighborhood of μ. Since $N^{1/2}\{\hat{S}(\hat{\mu}) - \frac{1}{2}\} \to 0$ in probability as $N \to \infty$, and since the results of Breslow and Crowley (1974) imply that $\hat{S}(\mu)$ is asymptotically normal with mean $\frac{1}{2}$ and variance $\phi_N(\mu)$ and that $N^{1/2}\{\hat{S}(\hat{\mu}) - S(\hat{\mu}) - \hat{S}(\mu) + S(\mu)\} \to 0$ in probability, we find (as did Brookmeyer and Crowley, 1982b) by the delta method that intervals I_1 and I_4 are both asymptotically equivalent to the interval

$$I_0 \equiv [\hat{\mu} \pm \{z_{\alpha/2}\hat{\phi}_G^{1/2}(\hat{\mu})\}/f(\hat{\mu})].$$

Similarly, $N^{1/2}\{\hat{\Lambda}(\hat{\mu}) - \ln 2\} \to 0$, $N^{1/2}\{\hat{\Lambda}(\hat{\mu}) - \Lambda(\hat{\mu}) - \hat{\Lambda}(\mu) + \Lambda(\mu)\} \to 0$, and $\hat{\Lambda}(\mu)$ is asymptotically normal with mean $\ln 2$ and variance $4\phi_N(\mu)$, so that by the delta method, I_3 is also asymptotically equivalent to I_0.

We have remarked that I_5, the test-based interval due to Simon and Lee (1982) differs from I_4, the interval due to Brookmeyer and Crowley (1982a), only in that the simple estimator $\tilde{\phi}_p(\cdot)$ is used in place of $\hat{\phi}_G(\cdot)$. In fact, $\hat{\phi}_p(t)$ and $\tilde{\phi}_p(t)$ strictly overestimate the variance of $\hat{S}(t)$ in large samples with censored data, as can be seen from the almost-sure

limiting relations

$$\begin{aligned}\lim_{N\to\infty} N\hat{\phi}_G(t) &= -S(t)^2 \int_0^t \{S(x)^2 S_Y(x)\}^{-1}\, d\{S(x)\} \\ &\leqslant -S(t)^2\{S_Y(t)\}^{-1} \int_0^t \{S(x)\}^{-2}\, dS(x) \\ &= [S(t)\{1 - S(t)\}]/S_Y(t) \\ &= \lim_{N\to\infty} N\hat{\phi}_p(t).\end{aligned}$$

In particular, when $S(\cdot) \equiv S_Y(\cdot)$ (which implies 50% censoring), it is easy to check that $\lim \hat{\phi}_p(t)/\hat{\phi}_G(t) = 2\{1 + S(t)\}^{-1}$, which is $\frac{4}{3}$ for $t = \mu$. In all cases in which there is no censoring, $\hat{\phi}_p(t)/\hat{\phi}_G(t) \to 1$ as $N \to \infty$. Because of these considerations, we did not include the Simon–Lee interval I_5 in our simulations (§6).

By the normal approximation to binomial tail probabilities, and recalling that $\hat{S}(t)$ will be approximately $\frac{1}{2}$ for t in any of the intervals I_1 to I_6 when N is large, one can easily check that I_2 and I_6 (or the 'smoothed' I_6' defined below) are asymptotically equivalent to

$$\{t\colon |\hat{S}(t) - \tfrac{1}{2}| \leqslant \tfrac{1}{2} z_{\alpha/2} N^{-1/2}\}$$

which is equivalent to

$$[\hat{\mu} - z_{\alpha/2}/\{2f(\hat{\mu})N^{1/2}\},\quad \hat{\mu} + z_{\alpha/2}/\{2f(\hat{\mu})N^{1/2}\}]$$

and does not take the level of censoring into account. In fact, this last interval differs from I_0 only in that the Greenwood variance estimator is replaced by $(4N)^{-1}$, a strictly anticonservative estimator unless $S_Y \equiv 1$ on $[0, \mu]$, as can be seen from $-\int_0^\mu \{S^2(x)S_Y(x)\}^{-1}\, dS(x) > -\int_0^\mu S^{-2}(x)\, dS(x) = 1$.

4. Special Modifications in Small Samples

Special difficulties in defining intervals may arise when the number of observed deaths is very small. For example, when $S(t) = S_Y(t) \equiv \exp(-t)$ and $N \leqslant 25$, there is a probability of several percent that $\hat{S}(\hat{\mu}) = 0$. To avoid degeneracy, when $r_j = 1$, we have replaced $(r_j - 1)$ by r_j in the final term of the summation for $\tilde{\phi}_G(\hat{\mu})$ which we have used in the simple reflected and transformed reflected intervals. Another difficulty peculiar to small samples is that the inverse values $\hat{S}^{-1}$ and $\hat{\Lambda}^{-1}$, which define the endpoints for reflected intervals, may not exist. At the lower endpoint, we define the uncomputable inverse value to be the smallest observed death time (min T_i for which $\Delta_i = 1$); this fix-up was chosen for comparability to test-based intervals, where the lowest possible endpoint is the smallest death time (unless survival curve estimators are interpolated). If we want to avoid semi-infinite intervals, there is no natural value to substitute for uncomputable upper endpoints. If we want the upper endpoint of an interval never to be larger than the largest observed death time, then we note that

$$\text{pr}\{\mu > \max(T_i\colon\ 1 \leqslant i \leqslant N,\quad \Delta_i = 1)\} = \left\{\text{pr}(\Delta = 0) + \int_0^\mu S_Y(x) f(x)\, dx\right\}^N.$$

This probability is $(\frac{1}{2})^N$ when there is no censoring but is $(\frac{7}{8})^N$ when $S(\cdot) \equiv S_Y(\cdot)$, and in the latter case, $N \leqslant 22$ implies that errors of coverage of this sort have probability of at least .05. Another possible choice (which we adopt for both reflected and test-based intervals) is to define an uncomputable upper endpoint to be the largest observed time T_i. Then $\text{pr}\{\mu > \max(T_i\colon\ 1 \leqslant i \leqslant N)\} = \{1 - \frac{1}{2}S_Y(\mu)\}^N$, which is $(\frac{3}{4})^N$ when $S(\cdot) \equiv S_Y(\cdot)$. This

probability [which is actually less than .003 if $S(\cdot) \equiv S_Y(\cdot)$] is small enough for $N > 20$ and for typical $S_Y(\cdot)$ and $S(\cdot)$ that there would be no practical advantage in choosing a larger fix-up value. This fix-up is equivalent to replacing $\hat{S}$ by the self-consistent estimator due to Efron (1967), i.e. to setting $\hat{S}(t) = 0$ for $t \geq \max\{T_i\}$, and is in accord with the practice of Brookmeyer and Crowley (1982a) and Emerson (1982). In spite of these arguments, there will be many small data sets for which it is preferable to report semi-infinite intervals.

None of the intervals I_1 to I_6 use interpolated (i.e. continuous) survival-curve estimators, although Emerson's interval I_2 does rely on interpolated binomial tail probabilities. It is clear that variant intervals of each type could be produced based on an interpolated variant of the Kaplan–Meier curve. Following Efron (1981), Reid (1981) did propose that her conditional bootstrap interval be based on 'smoothed' and interpolated binomial tail probabilities. At the jth ordered death time Reid first averaged the corresponding tail probability with that of the preceding death time, which is equivalent to connecting the midpoints of the vertical steps; then, using these smoothed values she interpolated linearly to find the time points corresponding to probabilities $\frac{1}{2}\alpha$ and $1 - \frac{1}{2}\alpha$. We use this smoothed interval I_6' in our computations below.

There are alternative median estimators, variance estimators, smoothing techniques and interpolation methods which could be chosen in various combinations in an attempt to improve small-sample performance of the intervals we discuss. While we have studied some of these, our purpose here is to survey and unify existing ideas of confidence-interval construction.

5. Behavior of Interval Estimators with Uncensored Data

It is relatively simple, as shown, for example, by Lehmann (1975, pp. 181–185), to construct nonparametric confidence intervals for the median from uncensored data. In fact, for an uncensored sample of size N the exact probability with which the true median lies in the interval $(T_{(J)}, T_{(K)})$ from the Jth-order statistic to the Kth is $\sum_{m=J}^{K-1} \binom{N}{m} 2^{-N}$. Although our primary concern is with confidence intervals based on *censored* samples, there is some interest in the remark that all the intervals I_1 to I_6 have the form $(T_{(J)}, T_{(K)})$ on uncensored samples, where J and K are nonrandom and depend only on the type of interval, the nominal confidence level, and the sample size N. An endpoint (the left one, say) of the smoothed conditional bootstrap interval I_6' for uncensored data always falls a fixed fraction, γ, of the distance between adjacent order-statistics $T_{(J)}$ and $T_{(J+1)}$ and therefore, for fixed N and α, can be conveniently reported as the fractional order-statistic $T_{(J+\gamma)} \equiv T_{(J)} + \gamma(T_{(J+1)} - T_{(J)})$.

For example, with $N = 21$ and no censoring, $\hat{\mu} = T_{(11)}$, $\tilde{\phi}_G(\hat{\mu}) = \frac{1}{4}\{1/(21 \times 20) + \cdots 1/(11 \times 10)\} = .0131$ and $\hat{\Lambda}(\hat{\mu}) = 1/21 + \cdots + 1/11 = .7164$. Therefore when $\alpha = .05$ we can read off from the lists of values $\hat{S}(T_{(i)}) = 1 - i/21$ and $\hat{\Lambda}(T_{(i)}) = \sum_0^{i-1} (21 - k)^{-1}$ the intervals $I_1 = \hat{S}^{-1}(.5 \pm z_{\alpha/2}.0131^{1/2}) = (T_{(6)}, T_{(16)})$ and $I_3 = \hat{\Lambda}^{-1}(.716 \pm 2z_{\alpha/2}.0131^{1/2}) = (T_{(6)}, T_{(15)})$. Slightly lengthier calculations show (still for $N = 21$ and $\alpha = .05$) that the intervals I_4 and I_5 are $(T_{(7)}, T_{(15)})$, that I_2 (even in censored cases) is precisely $\{\hat{S}^{-1}(.7404), \hat{S}^{-1}(.2596)\}$ which in the uncensored case is $(T_{(6)}, T_{(16)})$, and that $I_6' = \{T_{(6.53)}, T_{(15.47)}\}$. Table 1 summarizes the results of similar calculations on uncensored samples with selected values of N and $\alpha = .05$ and $\alpha = .10$, including exact binomial coverage probabilities (which have been linearly interpolated for I_6').

Table 1 is instructive because it shows how the performance of the different intervals (on uncensored data) varies with the sample size, simply because of the discreteness of the binomial distributions. For example, the interval I_5 of Simon and Lee (1982) performs very well at sample size 25, with exact coverage probabilities .957 and .892 in place of the

Table 1

Endpoint order-statistic numbers and exact coverage probabilities for α-level confidence intervals based on uncensored samples of size N

N	Simple reflected, I_1	Emerson, I_2	Transformed reflected, I_3	Brookmeyer–Crowley, I_4	Simon–Lee, I_5	Reid, smoothed, I'_6
$\alpha = .05$						
21	6, 16	6, 16	6, 15	7, 15	7, 15	6.53, 15.47
	.973	.973	.948	.922	.922	.946
22	7, 16	6, 17	6, 15	7, 16	7, 16	6.49, 15.63
	.948	.983	.925	.948	.948	.942
25	8, 18	8, 18	7, 18	8, 18	8, 18	8.16, 17.84
	.957	.957	.971	.957	.957	.946
40	14, 27	14, 27	13, 26	15, 26	14, 27	13.92, 26.17
	.962	.962	.951	.919	.962	.945
41	15, 27	14, 28	14, 27	15, 27	15, 27	14.78, 27.22
	.940	.972	.956	.940	.940	.948
42	15, 28	15, 28	14, 27	15, 28	15, 28	14.73, 27.35
	.956	.956	.946	.956	.956	.945
60	23, 38	22, 39	22, 37	23, 38	23, 38	22.45, 37.59
	.948	.973	.940	.948	.948	.946
61	23, 39	23, 39	23, 38	24, 38	23, 39	23.37, 38.63
	.960	.960	.944	.929	.960	.948
62	24, 39	23, 40	23, 38	24, 39	24, 39	23.33, 38.71
	.944	.970	.936	.944	.944	.947
$\alpha = .10$						
21	7, 15	7, 15	7, 15	7, 15	7, 15	7.23, 14.77
	.922	.922	.922	.922	.922	.896
25	9, 17	8, 18	8, 17	9, 17	9, 17	8.93, 17.07
	.892	.957	.924	.892	.892	.897
41	16, 26	15, 27	15, 26	16, 26	16, 26	15.74, 26.26
	.883	.940	.912	.883	.883	.898

nominal .95 and .90. However, their method at $N = 21$ has the defect of producing identical intervals (with coverage .922) at both nominal probabilities .95 and .90. Similar comments apply to the Brookmeyer–Crowley test-based intervals I_4. Some general conclusions for uncensored data can be drawn from the extended form of Table 1 (summarized in Table 2) covering all N from 21 through 75. The test-based intervals as a group have, with relatively few exceptions, smaller coverage than the reflected intervals. The intervals I_4 due to Brookmeyer and Crowley and I_5 due to Simon and Lee are typically anticonservative, while the simple reflected interval I_1 and especially Emerson's I_2 are noticeably conservative. Between I_1 and I_2 (respectively, I_4 and I'_6), the larger coverage is usually attained by the interval I_2 (respectively, I'_6) based on binomial rather than normal tail probabilities. This might have been expected because of the known inequalities (see Slud, 1977) bounding binomial tails above their normal approximants. The new transformed reflected interval I_3, which is not markedly conservative or anticonservative at level .05 and is slightly conservative at level .10, was chosen for its approximately nominal behavior over the variants

$$I'_3 \equiv [t: \{\hat{\Lambda}(t) - \ln 2\}^2 \leq 4\, \chi^2_{1,\alpha}\, \tilde{\phi}_G(\hat{\mu})]$$

and

$$I''_3 \equiv [t: \{-\ln \hat{S}(t) - \ln 2\}^2 \leq 4\, \chi^2_{1,\alpha}\, \tilde{\phi}_G(\hat{\mu})].$$

The interval I''_3 behaves somewhat anticonservatively (see Table 2), while I'_3 is systematically

Table 2

Counts of how many of the 55 sample sizes ($N = 21, \ldots, 75$) have exact coverage probabilities on uncensored data* (i) *below .935 for $\alpha = .05$ or below .88 for $\alpha = .10$*; (ii) *above .965 for $\alpha = .05$ or above .92 for $\alpha = .10$*; *and* (iii) *above nominal level*

	Simple reflected,	Emerson,	Trans-formed reflected			Brookmeyer–Crowley,	Simon–Lee,	Reid, smoothed,
	I_1	I_2	I_3	I_3'	I_3''	I_4	I_5	I_6'
$\alpha = .05$								
(i) No. < .935	2	0	3	0	9	18	5	0
(ii) No. > .965	6	27	4	3	1	0	2	0
(iii) No. > .95	33	55	32	39	25	15	26	0
$\alpha = .10$								
(i) No. < .88	6	0	0	1	3	19	9	0
(ii) No. > .92	12	34	6	12	1	1	9	0
(iii) No. > .90	33	55	40	40	26	20	30	0

* Binomial coverage probabilities are interpolated for the interpolated endpoints of I_6' (see discussion at the end of §4).

shifted slightly to the left of I_3 and its average length tends to be the same or slightly larger. Finally, the performance of Interval I_6', due to Reid (1981), is quite impressive on uncensored data, always giving very slightly less than nominal coverage. Comparisons with the unsmoothed interval I_6 (which is badly anticonservative) suggest that interpolation plays an important role (which, however, will become less important as sample size increases).

6. Simulation Results

Our simulation experiments cover three lifetime distributions: (i) constant hazard, exponential with parameter 1, denoted by exp(1); (ii) decreasing hazard, Weibull with scale parameter 1.0 and shape parameter 0.7, denoted by Weib (1, 0.7); and (iii) increasing hazard, Weib (1, 1.5). The censoring distributions simulated are (i) exp (1); (ii) uniform on the interval 0 to 2, denoted unif (0, 2); (iii) exp (.3); and (iv) unif (0, 4.5).

All simulations were run with 7600 replications, so that the standard error due to simulation was approximately .0025 for nominal .95 probabilities, and .0034 for nominal .90 probabilities. Tables 3 and 4 show the relative frequencies with which the intervals contained the true median for the underlying life distribution for sample sizes 21 and 41, respectively.

The Brookmeyer–Crowley interval, I_4, was anticonservative in all cases, strikingly so at $N = 21$ and somewhat less so at $N = 41$. The improvement with larger sample size would be predicted from the asymptotic results in §3. Both the Emerson interval, I_2, and the Reid interval, I_6', tended to become noticeably anticonservative with heavier censoring, and this result was more marked at $N = 41$ than at $N = 21$. Again, this would be expected from the large-sample considerations of §3. However, the results of our simulations of Method I_2 differ from the simulation results of Emerson (1982), who reported conservative coverage at all levels of censoring with values of N of 25, 50 and 100. We requested the simulation program used by Emerson, and we noted that this program arbitrarily widened the confidence interval for censored data in a manner not discussed by that author and for which we could imagine no logical justification.

In the simulations the new reflected intervals, I_1 and I_3, maintained conservative coverage at $\alpha = .05$ for all lifetime and censoring combinations, and both were quite conservative at

Table 3
Sample size 21, *empirical coverage for confidence level* .95 (*and* .90 *in parentheses*)

Lifetime	Censoring distribution	Expected percentage censored	Simple reflected, I_1	Emerson, I_2	Transformed reflected, I_3	Brookmeyer–Crowley, I_4	Reid, smoothed, I'_6
exp (1)	exp (1)	50.0	.972(.946)	.922(.863)	.965(.947)	.913(.858)	.941(.890)
	unif (0, 2)	43.2	.976(.940)	.950(.908)	.968(.943)	.924(.874)	.943(.895)
	exp (.3)	23.1	.964(.922)	.957(.921)	.964(.929)	.923(.876)	.945(.893)
	unif (0, 4.5)	22.0	.968(.929)	.964(.930)	.967(.936)	.929(.881)	.948(.901)
Weib (1, 0.7)	exp (1)	47.6	.975(.944)	.940(.889)	.966(.945)	.916(.862)	.950(.898)
	unif (0, 2)	42.3	.974(.944)	.959(.918)	.966(.945)	.924(.874)	.949(.902)
	exp (.3)	24.6	.969(.933)	.962(.932)	.971(.940)	.930(.883)	.950(.904)
	unif (0, 4.5)	24.9	.968(.927)	.963(.928)	.966(.934)	.924(.884)	.949(.901)
Weib (1, 1.5)	exp (1)	52.7	.973(.939)	.896(.832)	.962(.939)	.901(.847)	.926(.875)
	unif (0, 2)	43.8	.975(.939)	.940(.890)	.970(.942)	.921(.867)	.939(.888)
	exp (.3)	22.5	.964(.922)	.956(.916)	.965(.928)	.926(.869)	.942(.889)
	unif (0, 4.5)	20.0	.965(.925)	.962(.925)	.963(.931)	.928(.879)	.946(.896)

$\alpha = .10$, especially with heavier censoring. The transformed reflected interval, I_3, tended to give closer to nominal coverage at $\alpha = .05$, and the simple reflected interval, I_1, tended to do so at $\alpha = .10$. Both methods performed better at $N = 41$ than at $N = 21$, in consonance with the large-sample asymptotics.

Among the large-sample consistent intervals, I_1, I_3 and I_4, Table 5 shows (for $\alpha = .05$) that the conservative reflected intervals, I_1 and I_3, were considerably longer on average for $N = 21$ than the anticonservative Brookmeyer–Crowley interval, I_4, and somewhat longer for $N = 41$. For all types of censoring, all average interval lengths decreased as the lifetime hazards changed from decreasing to constant to increasing.

7. A Worked Example

As an illustration of the six methods we are comparing, we have calculated confidence intervals for the median survival time in a 6-MP treatment group of leukemia patients (in a well-known data set of Freireich *et al.*, 1963, as presented by Gehan, 1975). Our confidence intervals are shown in Table 6. Since this data set contains tied observations, our formulas for $\hat{S}$, $\hat{\phi}_G$ and $\hat{\Lambda}$ must be modified by replacing Δ_i with d_i, the number of deaths at event-time t_i. The final six columns of our table contain the test-statistics Z_1^2, P_2, Z_3^2, Z_4^2, Z_5^2 and P'_6 (where the Z_i^2 have approximate χ_1^2 distributions, and P_2 and P'_6 are doubled *P*-

Table 4
Sample size 41, *empirical coverage for confidence level* .95 (*and* .90 *in parentheses*)

Lifetime	Censoring distribution	Expected percentage censored	Simple reflected, I_1	Emerson, I_2	Transformed reflected, I_3	Brookmeyer–Crowley I_4	Reid, smoothed, I'_6
exp (1)	exp (1)	50.0	.966(.928)	.910(.858)	.961(.931)	.931(.880)	.918(.857)
	unif (0, 2)	43.2	.964(.923)	.939(.887)	.961(.921)	.933(.880)	.930(.873)
	exp (.3)	23.1	.961(.913)	.957(.913)	.956(.920)	.941(.890)	.945(.889)
	unif (0, 4.5)	22.0	.955(.907)	.954(.910)	.956(.914)	.935(.881)	.940(.884)
Weib (1, 0.7)	exp (1)	47.6	.969(.929)	.932(.876)	.962(.932)	.936(.881)	.931(.875)
	unif (0, 2)	42.3	.963(.921)	.944(.899)	.958(.917)	.931(.884)	.934(.880)
	exp (.3)	24.6	.960(.914)	.957(.917)	.958(.917)	.938(.887)	.945(.892)
	unif (0, 4.5)	24.9	.956(.912)	.955(.917)	.957(.917)	.936(.886)	.942(.894)
Weib (1, 1.5)	exp (1)	52.7	.964(.929)	.893(.832)	.962(.932)	.929(.878)	.906(.842)
	unif (0, 2)	43.8	.958(.918)	.927(.871)	.957(.918)	.932(.876)	.919(.857)
	exp (.3)	22.5	.961(.913)	.953(.906)	.956(.918)	.940(.887)	.939(.882)
	unif (0, 4.5)	20.0	.954(.909)	.949(.906)	.953(.914)	.934(.883)	.936(.883)

Table 5
Empirical average lengths for intervals I_1, I_3 and I_4 for confidence level .95, sample size 21 (and 41 in parentheses)

Lifetime	Censoring distribution			
	exp (1)	unif (0, 2)	exp (.3)	unif (0, 4.5)
I_1 (Simple reflected)				
Weib (1, 0.7)	1.586(1.226)	1.269(0.962)	1.539(0.905)	1.471(0.884)
exp (1)	1.350(0.975)	1.124(0.803)	1.131(0.713)	1.094(0.697)
Weib (1, 1.5)	1.088(0.732)	0.929(0.612)	0.811(0.532)	.0792(0.520)
I_3 (Transformed reflected)				
Weib (1, 0.7)	1.577(1.219)	1.261(0.952)	1.461(0.884)	1.400(0.863)
exp (1)	1.372(0.983)	1.142(0.805)	1.101(0.705)	1.064(0.692)
Weib (1, 1.5)	1.136(0.744)	0.967(0.619)	0.813(0.533)	0.791(0.520)
I_4 (Brookmeyer–Crowley)				
Weib (1, 0.7)	1.388(1.209)	1.116(0.934)	1.319(0.849)	1.265(0.826)
exp (1)	1.114(0.914)	0.960(0.764)	0.956(0.663)	0.930(0.648)
Weib (1, 1.5)	0.825(0.649)	0.752(0.564)	0.676(0.491)	0.663(0.479)

values from binomial tails); the nonsignificant values of these test statistics form the confidence intervals I_1 to I_6', respectively. In other words, denoting by $C_j(t)$ the jth column entry at $t_i = t$ in Table 6, we have constructed Columns 9 through 13 with the formulas

$$C_9(t) = 4\{C_4(t) - \tfrac{1}{2}\}^2/C_6(\hat{\mu}),$$

$$C_{10}(t) = 2\tilde{B}[N \max\{C_4(t),\ 1 - C_4(t)\},\ N,\ \tfrac{1}{2}],$$

and

$$C_{11}(t) = \{C_5(t) - C_5(\hat{\mu})\}^2/C_6(\hat{\mu}),$$

$$C_{12}(t) = \{C_4(t) - \tfrac{1}{2}\}^2/C_8(t),$$

$$C_{13}(t) = \{C_4(t) - \tfrac{1}{2}\}^2/C_7(t),$$

where, for this data set, one can easily see from Column 4 that $\hat{\mu}$ is 23 weeks. Column 14 is calculated by the method given in §4. From the six final columns of Table 6, we read off

Table 6
Construction of Intervals I_1, I_2, I_3, I_4, I_5 and I_6' for the 6-MP leukemia-treatment-group data from Freireich et al. (1963)

(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)	(10)	(11)	(12)	(13)	(14)
t_i	r_i	d_i	$\hat{S}(t_i)$	$\hat{\Lambda}(t_i)$	$\sum_{j:T_j \le t_i} \frac{d_j}{r_j(r_j - d_j)}$	$\tilde{\phi}_P(t_i)$	$\hat{\phi}_G(t_i)$	Z_1^2	P_2	Z_3^2	Z_4^2	Z_5^2	P_6'
6	21	3	.857	.143	.008	.011	.006	5.66	.001	4.12	21.9	12.0	.000
7	17	1	.807	.202	.012	.012	.008	4.17	.008	3.36	12.4	7.81	.001
9	16	0	.807	.202	.012	.013	.008	4.17	.008	3.36	12.4	7.46	.001
10	15	1	.753	.268	.016	.013	.009	2.84	.036	2.60	6.89	5.01	.007
11	13	0	.753	.268	.016	.014	.009	2.84	.036	2.60	6.89	4.42	.007
13	12	1	.690	.352	.024	.015	.011	1.60	.134	1.78	3.17	2.46	.039
16	11	1	.627	.442	.033	.015	.013	0.72	.349	1.06	1.25	1.11	.148
17	10	0	.627	.442	.033	.016	.013	0.72	.349	1.06	1.25	1.04	.148
19	9	0	.627	.442	.033	.017	.013	0.72	.349	1.06	1.25	0.93	.148
20	8	0	.627	.442	.033	.020	.013	0.72	.349	1.06	1.25	0.83	.148
22	7	1	.538	.585	.057	.020	.016	0.06	.901	0.31	0.09	0.07	.477
23	6	1	.448	.752	.090	.019	.018	0.12	.802	0.	0.15	0.14	.952
25	5	0	.448	.752	.090	.022	.018	0.12	.802	0.	0.15	0.12	.952
32	4	0	.448	.752	.090	.028	.018	0.12	.802	0.	0.15	0.10	.952
34	2	0	.448	.752	.090	.056	.018	0.12	.802	0.	0.15	0.05	.952
35	1	0	.448	.752	.090	.112	.018	0.12	.802	0.	0.15	0.02	.315

the estimated 95% confidence intervals (in weeks) $I_1 = (10, 35)$, $I_2 = I_4 = I_5 = (13, 35)$, $I_3 = (7, 35)$ and $I_6' = (13.30, 35)$, where the left endpoint of I_6' is interpolated from Column 14 as 13 + (16–13)(.05–.039)/(.148–.039). The upper endpoint, 35, is chosen as the largest t_i, since none of the Z_j attain significance for large values of t_i. Ordinarily, the upper endpoint of the estimated interval is the smallest t_i for which $Z_j^2(t)$ is significant at all $t \geq t_i$. It is worth remarking that in this example, instead of using time intervals $(k - .5, k + .5]$ in months, we have rounded follow-up times to k weeks, and we have adopted the usual convention of treating all censorship in this time interval as occurring after all deaths 'at $t_i = k$'. The survival and censoring distributions have thereby been discretized, and the inclusion of endpoints in the estimated confidence intervals is open to interpretation.

It is instructive to compare the median estimate and confidence interval for the 6-MP group in our example with the estimate and interval for the placebo group (Gehan, 1975). In the placebo group, $\hat{\mu} = 8$ weeks and $I_1 = (3, 12)$, $I_2 = I_3 = (4, 12)$, $I_4 = I_5 = (4, 11)$ and $I_6' = (3.87, 11.77)$. Even though the sample sizes for the two groups are small ($N = 21$ for both), the fact that these confidence intervals barely overlap (with I_2, I_4, I_5 and I_6' they do not overlap) suggests that the large observed difference in the median estimates is not due to chance but to a true treatment effect. In fact the Cox–Mantel test comparing the two treatments yields a normal deviate at 4.10 ($P < .001$). Despite the nice separation of confidence intervals in this example, the imprecision of the median estimate for the heavily censored 6-MP group is apparent when we consider that the confidence interval I_1 for this group covers 71% of the time axis, ranging from 0 to the largest observed time, here the censored observation at 35 weeks. The comparable figure for the placebo group is 35%.

8. Discussion and Recommendations

While all intervals in this paper can be derived from considerations of hypothesis testing, the estimator of asymptotic null variance at the single point $\hat{\mu}$ characterizes what we have called the reflected intervals. The geometric interpretation given in Fig. 1 sets reflected intervals clearly apart from their test-based counterparts.

Although in practice one could encounter a broader spectrum of death and censoring distributions than were studied here, our calculations and simulations support the following conclusions and recommendations:

(i) the intervals I_2 (Emerson, 1982), I_5 (Simon and Lee, 1982) and I_6' (Reid, 1981) all have asymptotically incorrect coverage in the presence of censoring;

(ii) the test-based confidence interval I_4 (Brookmeyer and Crowley, 1982a), although large-sample consistent, gives short and anticonservative interval estimates for the median survival time in small samples of censored survival data when the lifetime is not too far from exponential;

(iii) the reflected confidence intervals I_1 and I_3 have asymptotically correct coverage, and in our small-sample simulations of censored survival data, these intervals maintain (at least) nominal coverage for all sample sizes, confidence levels and lifetime-censoring combinations studied, but sometimes gave markedly above-nominal coverage;

(iv) Intervals I_1 and I_3 are both easy to compute, and have similar small-sample behavior. Since the simple reflected interval, I_1, uses the Kaplan–Meier curve and Greenwood variance directly, it is the most likely choice for reporting confidence intervals for the median lifetime when data are censored.

Our preference for the reflected intervals I_1 and I_3 conforms with the common perspective,

which we share, that confidence intervals should achieve at least nominal coverage. There is however a clear trade-off between coverage and interval length, and in some situations one might prefer the anticonservative Brookmeyer–Crowley interval I_4 for its shorter length.

Median estimates based on limited data are highly variable. For this reason readers of medical journals may seriously misinterpret median estimates quoted without estimated confidence limits, for example, by believing that two medians differ significantly when they do not, or by believing that because the ratio of two medians is 2, say, patients are dying twice as fast on one treatment as on another. The results presented here suggest that the simple reflected interval (or perhaps the transformed reflected interval) should be calculated to provide confidence limits for the median survival time in the presence of censoring.

Acknowledgements

The authors are grateful to Bob Banks of the IMS Corporation for programming simulation experiments, to the referees for numerous suggestions, and to Julie Paolella for her expert typing.

Résumé

La comportement, sur de petits échantillons, de quelques tests non-paramétriques récemment proposés pour construire des intervalles de confiance du temps moyen de survie sur des échantillons tronqués à droite au hasard est comparé à celui de deux nouvelles méthodes. La plupart de ces méthodes sont équivalentes sur de grands échantillons. Tous les intervalles proposés sont soit 'basés sur un test', soit 'réfléchis' dans un sens défini dans l'article. Les probabilités associées aux estimateurs des intervalles sont obtenues exactement pour des données non tronquées, et par simulation pour trois distributions de durée de vie et quatre modes de troncature. Pour les situations envisagées, les méthodes 'basées sur un test' ont une probabilité associée plus petite que celle annoncée, alors que les nouveaux intervalles 'réfléchis' ont une probabilité associée plus proche de la probabilité annoncée (quoique légèrement conservatrice) et sont facilement calculables.

References

Anderson, J. R., Bernstein, L. and Pike, M. C. (1982). Approximate confidence intervals for probabilities of survival and quantiles in life-table analysis. *Biometrics* **38,** 407–416.

Breslow, N. and Crowley, J. (1974). A large sample study of the life table and product limit estimates under censorship. *Annals of Statistics* **2,** 437–453.

Brookmeyer, R. and Crowley, J. (1982a). A confidence interval for the median survival time. *Biometrics* **38,** 29–41.

Brookmeyer, R. and Crowley, J. (1982b). A K-sample median test for censored data. *Journal of the American Statistical Association* **77,** 433–440.

Efron, B. (1967). The two-sample problem with censored data. In *Proceedings of the Fifth Berkeley Symposium on Mathematical Statistics and Probability, Vol.* IV, L. LeCam and J. Neyman (eds), 831–853. Berkeley: University of California Press.

Efron, B. (1981). Censored data and the bootstrap. *Journal of the American Statistical Association* **76,** 312–319.

Emerson, J. (1982). Nonparametric confidence intervals for the median in the presence of right censoring. *Biometrics* **38,** 17–27.

Freireich, E. J., Gehan, E., Frei, E., Schroeder, L. R., Wolman, R. A., Burgert, O. E., Mills, S. D., Pinkel, D. Selawry, O. S., Moon, J. H., Gendel, B. R., Spurr, C. L., Storrs, R., Haurani, F., Hoogstraten, B. and Lee, S. (1963). The effect of 6-mercaptopurine on the duration of steroid induced remissions in acute leukemia: A model for evaluation of other potentially useful therapy. *Blood* **21,** 699–716.

Gehan, E. (1975). Statistical methods for survival time studies. In *Cancer Therapy: Prognostic Factors and Criteria of Response*, M. J. Staquet (ed.), 7–35. New York: Raven Press.

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

Lehmann, E. (1975). *Nonparametrics: Statistical Methods Based on Ranks.* San Francisco: Holden-Day.

Nelson, W. (1972). Theory and applications of hazard plotting for censored failure data. *Technometrics* **14,** 945–966.

Peto, R., Pike, M. C., Armitage, P., Breslow, N. E., Cox, D. R., Howard, S. V., Mantel, N., McPherson, K., Peto, J. and Smith, P. G. (1977). Design and analysis of randomized clinical trials requiring prolonged observation of each patient II. *British Journal of Cancer* **35,** 1–39.

Reid, N. (1981). Estimating the median survival time. *Biometrika* **68,** 601–608.

Rothman, K. (1978). Estimation of confidence limits for the cumulative probability of survival in life table analysis. *Journal of Chronic Diseases* **31,** 557–560.

Simon, R. and Lee, Y. J. (1982). Nonparametric confidence limits for survival probabilities and median survival time. *Cancer Treatment Reports* **66,** 37–42.

Slud, E. (1977). Distribution inequalities for the binomial law. *Annals of Probability* **5,** 404–412.

Received January 1982; *revised February and June* 1983