BIOMETRICS 41, 643–652
September 1985

A Two-Sample Test Sensitive to Crossing Hazards in Uncensored and Singly Censored Data

Donald M. Stablein

The EMMES Corporation, 11325 Seven Locks Road, Suite 214,
Potomac, Maryland 20854, U.S.A.

and

I. A. Koutrouvelis

University of Patras, Department of Engineering Science, Rion-Patras, Greece

SUMMARY

Savage score statistics are employed to develop a test for comparing survival distributions with right-hand singly censored data. The procedure is motivated by the interest in developing a powerful method for determining differences when true survival distributions cross. Examination of small-sample characteristics under the null hypothesis indicate that asymptotic critical values yield a slightly conservative test. Power of the test compares favorably with other criteria, including the modified Smirnov procedure, particularly if there is a single crossing of the survival curves.

1. Introduction

The comparison of time-to-response distributions, particularly with censored data, has been receiving considerable attention in the statistical literature. The log rank test, first proposed by Mantel (1966), has optimal power among all unbiased rank-invariant tests for detecting small constant differences in the relative hazard function. Generalizations of the Wilcoxon test (Gehan, 1965; Peto and Peto, 1972) offer frequent competitors to the log rank procedure. More recently, Harrington and Fleming (1982) have discussed a class of linear rank statistics of which both the log rank and Peto generalization of the Wilcoxon test are members.

There is frequently expressed concern by practitioners that alternative distributions may not be from a set of location shift alternatives, and thus, the efficiency properties for the above procedures are not always applicable. Specifically, under many crossing hazard function alternatives, the power of the tests may be small. Fleming et al. (1980) noted this and developed a Kolmogorov–Smirnov-type procedure which performs considerably better in some crossing hazard situations with only a modest loss of power under a proportional hazards alternative. It is well known that the omnibus nature of Kolmogorov–Smirnov-type tests, although affording wide applicability, often offers reduced power versus common desirable alternatives.

It is the intention of this paper to develop in the succeeding section a two-sample testing procedure applicable to singly censored data, sensitive to simple crossing hazards alternatives. Section 3 will examine the size of the procedure in finite samples, and Section 4 will examine, in several situations, the power of this procedure versus that of the log rank and generalized Smirnov tests. Section 5 will give an illustration of the use of the procedure.

Key word: Log rank test.

2. Development of the Test

Following the notation of Koziol and Petkau (1978), consider $X_1, \ldots, X_m$, and $X_{m+1}, \ldots, X_N$ to be independent random samples of sizes m and $n = N - m$ from continuous populations π_1 and π_2, respectively. To test that the respective population distribution functions F_1 and F_2 are equal, one can employ the Savage statistic

$$S_N = \sum_{i=1}^{N} d_i[b_N(i) - 1],$$

where $b_N(i) = \sum_{j=1}^{i} (N - j + 1)^{-1}$ and $d_1, \ldots, d_N$ is a set of indicator variables, such that d_i equals 1 when the ith element of the ordered survival times $\{X_{(1)}, \ldots, X_{(N)}\}$ is from π_1. Under the null hypothesis H_0: $F_1 = F_2$,

$$\mathrm{var}(S_N \mid H_0) = \frac{mn}{N-1}\left(1 - \frac{b_N(N)}{N}\right).$$

If the data are singly censored at observation r and assuming $r/N \to p$ as $N \to \infty$ $(0 < p \leqslant 1)$, an optimal modification of the Savage statistic is

$$S_r^* = \sum_{i=1}^{r} d_i[b_N(i) - 1] + \sum_{i=r+1}^{N} d_i \sum_{l=r+1}^{N} [b_N(l) - 1]/(N - r),$$

where the d_i for $i > r$ are arbitrary ones and zeroes with the number of $d_i = 1$ equalling the number of censored individuals from π_1 (Gastwirth, 1965; Johnson and Mehrotra, 1972). The latter term, which represents the average remaining score to be assigned, can be simplified (Kalbfleisch and Prentice, 1980, p. 146), yielding

$$S_r^* = \sum_{i=1}^{r} d_i[b_N(i) - 1] + \sum_{i=r+1}^{N} d_i b_N(r).$$

Koziol and Petkau (1978) have derived and used the asymptotic distributions of the statistics

$$S_{N,r}^{(1)} = \max_{0 \leqslant k \leqslant r} \{S_k^*/[\mathrm{var}(S_N \mid H_0)]^{1/2}\}, \quad S_{N,r}^{(2)} = \max_{0 \leqslant k \leqslant r} \{\mid S_k^* \mid /[\mathrm{var}(S_N \mid H_0)]^{1/2}\}$$

to develop one-tailed and two-tailed sequential tests for H_0.

It is our purpose to develop, for singly censored data, a test at a single point in time which is sensitive to differences in distribution attributable to crossing hazard functions. Consider that if the hazard function for π_2 exceeds the hazard function for π_1 at early time points, and the reverse occurs at later time points, then $S_k^*/[\mathrm{var}(S_N \mid H_0)]^{1/2}$ will initially trend to larger and subsequently to smaller values. Asymptotically, one can test H_0 versus a one-sided alternative with the statistic $A_{N,r} = \max_{0 \leqslant k \leqslant r} A_{N,r}^{(k)}$, where

$$A_{N,r}^{(k)} = [S_k^* - (S_r^* - S_k^*)]/[\mathrm{var}(S_N \mid H_0)]^{1/2} = (2S_k^* - S_r^*)/[\mathrm{var}(S_N \mid H_0)]^{1/2}. \quad (1)$$

Motivation for the statistic is as follows. It is well known that for a constant relative hazard function, the summed contrasts of observed minus expected deaths, as defined by the log rank test, provide for a locally optimal test. Likewise, if a constant valued relative hazard favors one treatment prior to t^* and the other treatment after this point, one may reverse signs in the contrasts after t^* (Schoenfeld, 1981). For any i, S_i^* is related to the summed contrast for $t < t_i$ so that the statistic $A_{N,r}$ can be considered as the same conceptual contrast, with reversed signs for early and late events, extended to deal with the situation of t^* unknown. At times, one can, a priori, envision the potential for crossing hazard functions. For example, surgical decompression of biliary obstructed patients likely has initially high hazard rates associated with the operation. Alternative management techniques avoid this early risk, but as they may not be as efficacious, the latter risks of hepatic failure would be greater. Likewise, one could speculate that organ transplant studies could have similar crossing hazard patterns.

Let Φ be the standard normal distribution function. The following lemma provides the asymptotic distribution of $A_{N,r}$.

Lemma. If $\min(m, n) \to \infty$ and $r/N \to p$ $(0 < p \leqslant 1)$, then under H_0 and for every $x > 0$,

$$\lim_{N\to\infty} \Pr(A_{N,r} \geqslant x) = 2\{1 - \Phi(x/\sqrt{p}) + x(2\pi p)^{-1/2}\exp[-x^2/(2p)]\}. \tag{2}$$

The proof is given in the Appendix. It should be noted that the expression in the right-hand side of (2) equals $\Pr(\chi^2_{(3)} > x^2/p)$. An asymptotic critical value for a one-sided size α test based on $A_{N,r}$ is therefore $(p\chi^2_{3,\alpha})^{1/2}$, where $\chi^2_{3,\alpha}$ is the $100(1 - \alpha)$ percentile of the chi-square distribution with 3 degrees of freedom.

When crossing survival distributions are present, one is unlikely to want to perform a one-tailed test, as one must correctly specify beforehand where the statistically weighted advantage is likely to occur. Thus, to test H_0: $F_1 = F_2$ versus H_A: $F_1 \neq F_2$, one would use the statistic $B_{N,r} = \max_{0\leqslant k\leqslant r}|A_{N,r}(k)|$, where $A_{N,r}(k)$ is given by (1). The asymptotic distribution of $B_{N,r}$ is given by the following theorem.

Theorem. If $\min(m, n) \to \infty$ and $r/N \to p$ $(0 < p \leqslant 1)$, then under H_0 and for every $x > 0$,

$$\lim_{N\to\infty} \Pr(B_{N,r} \geqslant x) = 2 - \frac{1}{2}\,\Phi\!\left(\frac{3x}{\sqrt{p}}\right) - \frac{3}{2}\,\Phi\!\left(\frac{x}{\sqrt{p}}\right) + \frac{4x}{(2\pi p)^{1/2}}\exp\!\left(\frac{-x^2}{2p}\right) - C_p(x), \tag{3}$$

where

$$C_p(x) = \frac{2x}{(2\pi p)^{1/2}} \sum_{j=1}^{\infty} (-1)^{j+1}\exp\!\left[\frac{-(2j + 1)^2x^2}{2p}\right]. \tag{4}$$

The proof is given in the Appendix. For a two-tailed test of size α, an approximate asymptotic critical value, x, can be found from (3). For large x, $\Phi(3x/\sqrt{p}) \approx 1$ and the quantity $C_p(x)$ given by (4) is negligible. Thus, x $(x > 0)$ is the solution of the equation

$$1.5[1 - \Phi(x/\sqrt{p})] + 4x(2\pi p)^{-1/2}\exp[-x^2/(2p)] = \alpha \tag{5}$$

and an approximate test of size α would reject H_0 if $B_{N,r} \geqslant x$.

The above procedure can be applied to test the equality of any distributions. As it applies to singly censored data, it is particularly applicable to certain survival studies. Frequently in animal experiments all subjects begin the study at the same time and an identical follow-up period applies to all, thus yielding singly censored data at the termination of the study. Human experiments are less likely to be singly censored since study entry is usually staggered. However, when duration of follow-up is long relative to the entry period, especially when "cures" occur, failure times from clinical studies may be singly censored. Likewise, a fixed-length follow-up period might apply to all study subjects because the acute repair process is of maximal interest. For example, in a study of a surgical technique, all patients might be followed for 6 months, since later "failures" at the wound site seldom occur. Thus, singly censored data arise in several ways in biologic experiments.

3. Finite-Sample Null Distribution Results

Application of the statistic $B_{N,r}$ requires only the evaluation of $\text{var}(S_n \mid H_0)$ and the partial sum statistics S_k^* for $0 \leqslant k < r$. From small-sample null distribution results of the progressive application of the Savage test (Joe, Koziol, and Petkau, 1981), one might anticipate that the proposed test is conservative in finite samples. Simulations were performed using the unit exponential to examine the small-sample characteristics of the two-tailed test statistic under the null hypothesis. For 2000 random samples of given $m = n$, the test statistic was calculated and censoring was applied corresponding to $p = 1$, .9, .75, and .5. Figure 1 displays the results of the simulated $\alpha = .10$, .05, and .01 levels of the statistic $B_{N,r}/\sqrt{p}$.

In sample sizes frequently encountered in practice, the statistic is consistently conservative. As can be seen, the conservativeness is present even up to sample sizes of 1000. Thus, the full asymptotic nature of the test is approached quite slowly. For a sample size of 80 patients, the nominal .05 asymptotic level test has an observed size in the simulations ranging between 2.60% and 3.95%. Thus the bias, though definite, tends to be small.

As in Joe et al. (1981), equations of the form $X_{\alpha,N} = A_1 - A_2\exp(-A_3N^{1/2})$ were fit for each α to the simulated critical values in order to smooth the results and to simplify interpolations for other sample sizes. In these equations A_1 represents the asymptotic critical value x determined by solving equation (5). Censored critical values were scaled by $1/\sqrt{p}$, and a single curve fit for all p. The resulting curves are plotted in Figure 1 while the parameter values are listed in Table 1.

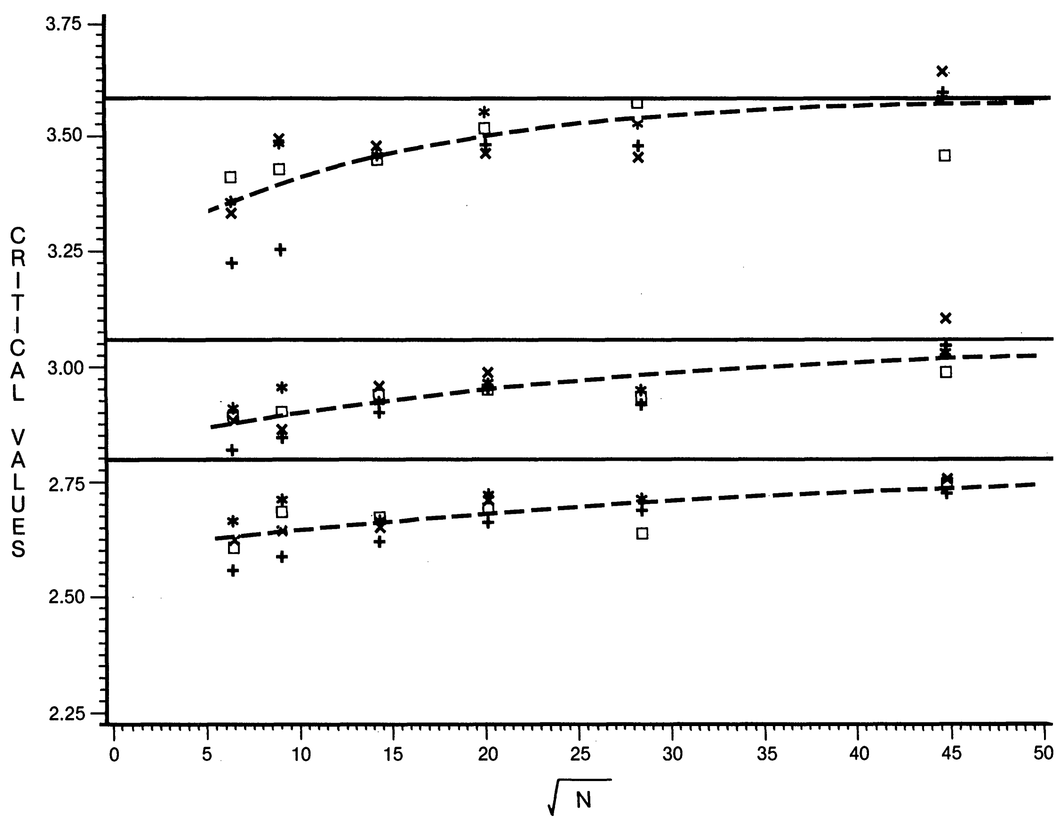

Figure 1. Small-sample critical values.
——— Approximate asymptotic critical values
----- Fitted critical value functions
Data points correspond to:
+ $p = 1.0$ × $p = .90$
∗ $p = .75$ □ $p = .50$

Table 1
Parameter estimates for critical value smoothing equations
$X_{\alpha,N} = A_1 - A_2\exp(-A_3N^{1/2})$

α	A_1	A_2	A_3
.10	2.7681	.18064	.03549
.05	3.0366	.21890	.04597
.01	3.5699	.37411	.08217

4. Simulated Power Results

Power for the statistic $B_{N,r}$, the modified Smirnov procedure, the Koziol–Petkau statistic $S_{N,r}^{(2)}$, and the log rank test have been examined in several situations with both crossing hazard and crossing survival distributions. The Wilcoxon test has not been included as either the Smirnov or log rank procedure had greater power in the situations examined by Fleming et al. (1980). Although it was not specifically designed for this purpose, the Koziol–Petkau procedure has been included both because it is a simpler procedure than $B_{N,r}$ and because it may not be vulnerable to the decreases in power that affect the log rank test when crossing hazards are present. Figure 2 displays the survival curves for each of the studies. Note that the situations considered include each of the sets of paired survival curves considered by Fleming et al. (1980). Power results from 1000 replications for sample sizes of 25 and 50 uncensored observations per treatment are detailed in Table 2. The observed power for the .05 and .01 level tests are provided.

In the first situation examined, that of proportional hazards, the proposed statistic performs poorly and, as is expected, the log rank statistic performs best. In each of the remaining situations both the log rank and the Koziol–Petkau tests have less power than both of the other procedures. Thus, for all other cases, only the proposed and modified Smirnov tests need to be considered. In situations B and C, which are both comparisons of

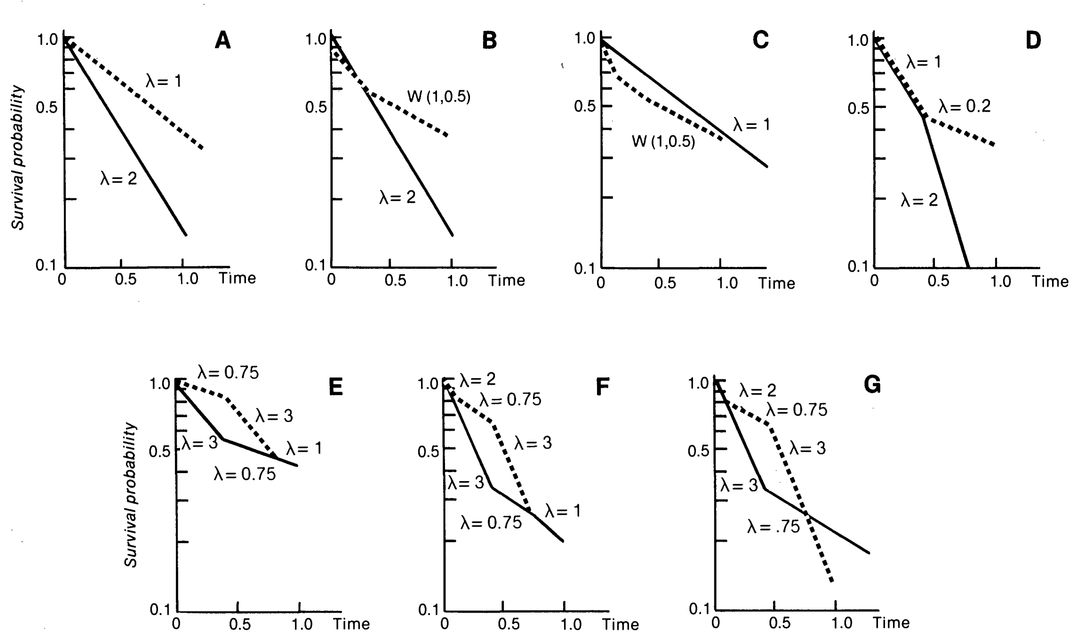

Figure 2. Survival distribution plots for power simulations.

Table 2
Power simulations in uncensored data

	$m=n$	$B_{N,r}$ $\alpha=.05$	$B_{N,r}$ $\alpha=.01$	Modified Smirnov $\alpha=.05$	Modified Smirnov $\alpha=.01$	Koziol–Petkau $\alpha=.05$	Koziol–Petkau $\alpha=.01$	Log rank $\alpha=.05$	Log rank $\alpha=.01$
A $\lambda_1=2, \lambda_2=1$	25	.319	.082	.555	.341	.588	.302	.658	.405
	50	.762	.539	.831	.637	.908	.724	.928	.805
B W(2, 1),[a] W(1, .5)	25	.737	.438	.572	.378	.339	.121	.469	.229
	50	.987	.930	.899	.725	.683	.412	.811	.587
C W(1, 1), W(1, .5)	25	.561	.302	.279	.098	.086	.012	.087	.015
	50	.908	.763	.554	.248	.209	.034	.148	.050
D $\lambda_1=1, \lambda_2=1 : t\in(0,.8)$	25	.571	.244	.802	.652	.429	.156	.596	.305
$\lambda_1=2, \lambda_2=.2 : t\in(.8,\infty)$	50	.954	.802	.985	.937	.782	.506	.883	.699
E $\lambda_1=3, \lambda_2=.75: t\in(0,.2)$	25	.340	.175	.377	.171	.155	.029	.082	.025
$\lambda_1=.75, \lambda_2=3 : t\in(.2,.4)$	50	.705	.485	.749	.474	.374	.092	.098	.033
$\lambda_1=1, \lambda_2=1 : t\in(.4,\infty)$									
F $\lambda_1=2, \lambda_2=2 : t\in(0,.1)$	25	.385	.172	.409	.239	.242	.068	.115	.044
$\lambda_1=3, \lambda_2=.75: t\in(.1,.4)$	50	.753	.541	.744	.604	.510	.218	.167	.070
$\lambda_1=.75, \lambda_2=3 : t\in(.4,.7)$									
$\lambda_1=1, \lambda_2=1 : t\in(.7,\infty)$									
G $\lambda_1=2, \lambda_2=2 : t\in(0,.1)$	25	.730	.487	.530	.305	.266	.068	.065	.013
$\lambda_1=3, \lambda_2=.75: t\in(.1,.4)$	50	.986	.940	.868	.685	.508	.230	.053	.012
$\lambda_1=.75, \lambda_2=3 : t\in(.4,\infty)$									

[a] W(λ, α) is a Weibull distribution with survival function $1 - F(t) = \exp[-(\lambda t)^\alpha]$.

a constant hazard versus a Weibull distribution such that there is a late survival advantage for the Weibull curve, the proposed test has much better power than the modified Smirnov test. Alternatively, power comparisons for the late difference described in situation D, where no crossing of the hazard functions occurs, favor the Smirnov procedure. In situations E and F, where there is no crossing of the survival distributions, there is little difference between the Smirnov and the proposed statistics. If situation F is modified to permit the curves to cross in the last quartile one obtains situation G. Here the proposed test statistic achieves considerably increased power relative to the other tests.

Note that smoothed simulated critical values are being used for $B_{N,r}$ and the Koziol–Petkau $S^{(2)}_{N,r}$. Increases in power for the new procedure range from only 1% to 7% as a result of using the smoothed rather than asymptotic critical values for the case $m = n = 50$, $\alpha = .05$.

In summary, for each of the above simulations with crossing survival functions (situations B, C, and G), the new test has considerably more power than the Smirnov test. For cases with crossing hazards without crossing survival functions (situations E and F), the two approaches have for practical purposes similar power characteristics.

Effects of right-hand single censoring on the procedures are presented for three tests and the single situation E. The modified Smirnov procedure is included although it may be used for singly censored data only providing that the survival times are statistically independent of the single censoring mechanism. One thousand replications were performed for five censoring levels with sample sizes of $m = n = 25$. The resulting power of the .05 and .01 level tests are detailed in Table 3. Examination of the results for the log rank test indicates the penalty that is paid for long-term follow-up when one has crossing hazards of this type. After observing 30% of the failure times, $B_{N,r}$ has less power than the Smirnov test, which has power slightly less than the log rank. Observation of 50%–90% of the

Table 3
Simulated power for various censoring levels
$m = n = 25$
For survival distributions situation E

	$B_{N,r}$		Modified Smirnov		Log rank	
p	.05	.01	.05	.01	.05	.01
.3	.465	.230	.542	.249	.570	.331
.5	.680	.451	.481	.234	.169	.070
.7	.567	.338	.361	.151	.085	.023
.9	.404	.195	.367	.152	.075	.018
1.0	.340	.175	.377	.171	.082	.025

failures reverses these relationships. The modified Smirnov is relatively insensitive to follow-up past the point where the curves return together, for it is dependent only on the maximum difference between the curves. The power of the new test is diminished with follow-up past this point due to the increase in the critical value with no potential, on average, for an increase in the size of the test statistic. The relative value of these three testing procedures in singly censored data will depend on the relative hazard function and the amount of censoring permitted.

5. Illustration

The Gastrointestinal Tumor Study Group (1982) reported on the results of a trial comparing chemotherapy versus combined chemotherapy and radiation therapy in the treatment of locally unresectable gastric cancer. Survival times in days, for the 45 patients on each treatment are as follows:

Chemotherapy:

1,	63,	105,	129,	182,	216,	250,	262,	301,
301,	342,	354,	356,	358,	380,	383,	383,	388,
394,	408,	460,	489,	499,	523,	524,	535,	562,
569,	675,	676,	748,	778,	786,	797,	955,	968,
1000,	1245,	1271,	1420,	1551,	1694,	2363,	2754^+,	2950^+

Chemotherapy plus Radiotherapy:

17,	42,	44,	48	60,	72,	74,	95,	103,
108,	122,	144,	167,	170,	183,	185,	193,	195,
197,	208,	234,	235,	254,	307,	315,	401,	445,
464,	484,	528,	542,	567,	577,	580,	795,	855,
1366,	1577,	2060,	2412^+,	2486^+,	2796^+,	2802^+,	2934^+,	2988^+

Figure 3 is a plot of the estimated survival distributions. A test for treatment equality with the log rank statistic provides for $\chi^2 = .23$, $P = .64$. The largest value of the function $|A_{90,82}(k)|$, with $A_{N,r}(k)$ given by (1), is observed for $k = 35$ (day 315) and equals $|2(-9.80) - (-2.10)|/4.633 = 3.78$. Thus, the observed value of the statistic, $B_{90,82} = 3.78$, is significant at the .01 level where the approximate critical value is $3.57(82/90)^{1/2} = 3.41$.

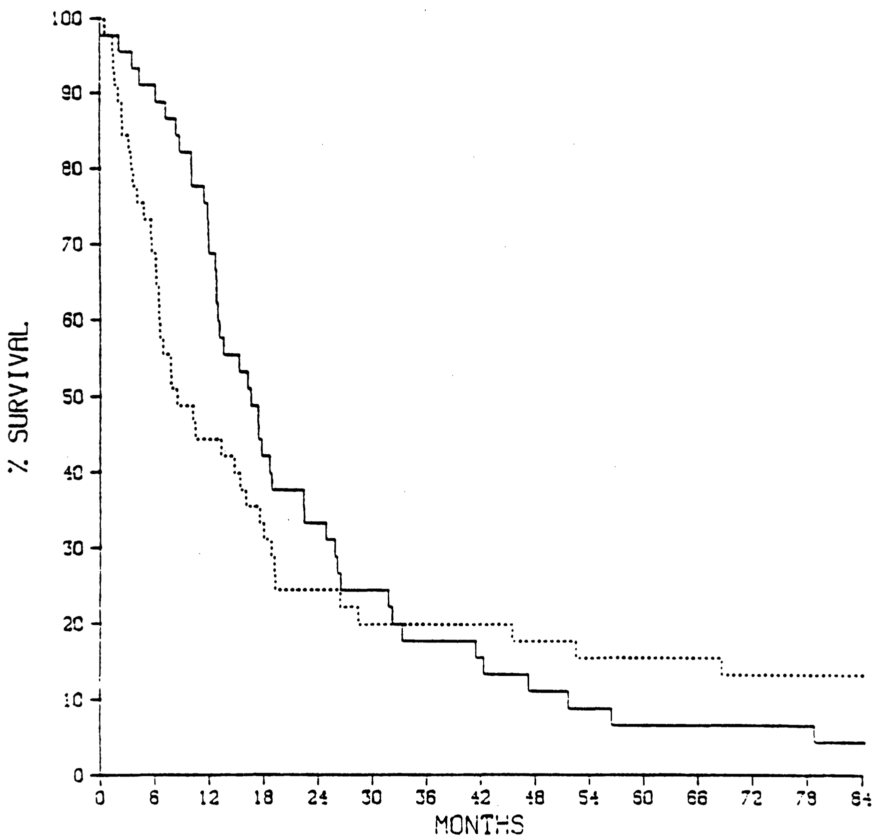

Figure 3. Product limit estimate of survival distributions.
——— Chemotherapy
...... Chemotherapy and radiation

6. Discussion

The developed procedure is intended to supplement rather than supplant current two-sample testing procedures. In a sense, we consider this procedure to add a framework for testing for equality of survival distributions. One can regard the log rank procedure as testing for main effects, i.e., a constant relative hazard function, since that is the situation for which it is optimal. The statistic discussed in this paper often provides for substantially increased power if there is an interaction of time and treatment. Investigation of simultaneous application of both testing procedures is warranted. It is possible to extend this procedure to develop tests sensitive to multiple crossings of the survival functions. Note, though, that as the number of crossings increases, the practical import attached to the distributional differences frequently decreases. Finally, application of this procedure to clinical situations would be increased by extension to situations of random censorship.

References

Chatterjee, S. K. and Sen, P. K. (1973). Nonparametric testing under progressive censorship. *Calcutta Statistical Association Bulletin* **22,** 13–50.

Fleming, T. R., O'Fallon, J. R., O'Brien, P. C., and Harrington, D. P. (1980). Modified Kolmogorov–Smirnov test procedures with application to arbitrarily right-censored data. *Biometrics* **36,** 607–625.

Gastrointestinal Tumor Study Group: Schein, P. D., Stablein, D. M., Bruckner, H. W., Douglass, H. O., Mayer, R., et al. (1982). A comparison of combination chemotherapy and combined modality therapy for locally advanced gastric carcinoma. *Cancer* **49,** 1771–1777.

Gastwirth, J. L. (1965). Asymptotically most powerful rank tests for the two-sample problem with censored data. *Annals of Mathematical Statistics* **36,** 1243–1247.

Gehan, E. A. (1965). A generalized Wilcoxon test for comparing arbitrarily single censored samples. *Biometrika* **52,** 203–223.

Harrington, D. P. and Fleming, T. R. (1982). A class of rank test procedures for censored survival data. *Biometrika* **69,** 553–566.

Joe, H., Koziol, J. A., Petkau, A. J. (1981). Comparison of procedures for testing the equality of survival distributions. *Biometrics* **37,** 327–340.

Johnson, R. A. and Mehrotra, K. G. (1972). Locally most powerful rank tests for the two-sample problem with censored data. *Annals of Mathematical Statistics* **43,** 828–831.

Kalbfleisch, J. D. and Prentice, R. L. (1980). *The Statistical Analysis of Failure Time Data.* New York: Wiley.

Karlin, S. and Taylor, H. (1975). *A First Course in Stochastic Processes,* 2nd edition. New York: Academic Press.

Koziol, J. A. and Petkau, A. J. (1978). Sequential testing of the equality of two survival distributions using the modified Savage statistic. *Biometrika* **65,** 615–623.

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

Peto, R. and Peto, J. (1972). Asymptotically efficient rank-invariant test procedures. *Journal of the Royal Statistical Society, Series A* **135,** 85–207.

Schoenfeld, D. (1981). The asymptotic properties of nonparametric tests for comparing survival distributions. *Biometrika* **68,** 316–319.

Received May 1984; *revised July* 1985.

APPENDIX

Proof of Lemma in Equation (2) *and Theorem in Equation* (3)

Define $W(t)$ to be a standard Wiener process on [0, 1] and let

$$W_{1,p} = \sup_{0\leqslant t\leqslant p} W(t),$$

$$W_{2,p} = \inf_{0\leqslant t\leqslant p} W(t),$$

$$Q_{1,p} = 2W_{1,p} - W(p),$$

$$Q_{2,p} = 2W_{2,p} - W(p),$$

$$Q_p = \sup_{0\leqslant t\leqslant p} |2W(t) - W(p)|.$$

Under the assumptions of the lemma or theorem and from the results of Chatterjee and Sen (1973), it follows that

$$\lim_{N\to\infty} \Pr(A_{N,r} \geqslant x) = \Pr(Q_{1,p} \geqslant x)$$

and

$$\lim_{N\to\infty} \Pr(B_{N,r} \geqslant x) = \Pr(Q_p \geqslant x).$$

Now for $x > 0$,

$$\Pr(Q_{1,p} \geqslant x) = \int_{-\infty}^{\infty} \Pr\left[W_{1,p} \geqslant \frac{x+w}{2}\,\middle|\, W(p) = w\right] f_{W(p)}(w)\, dw,$$

where

$$f_{W(p)}(w) = (2\pi p)^{-1/2}\exp[-w^2/(2p)].$$

If $|w| > x$, the conditional probability factor under the integral sign is equal to 1 since $Q_{1,p} \geq |W(p)|$ holds with probability 1. Therefore,

$$\Pr(Q_{1,p} \geqslant x) = 2\left[1 - \Phi\left(\frac{x}{\sqrt{p}}\right)\right] + \int_{-x}^{x} \Pr\left[W_{1,p} \geqslant \frac{x+w}{2} \;\middle|\; W(p) = w\right] f_{W(p)}(w)\,dw.$$

The latter term can be shown to equal

$$\frac{2x}{(2\pi p)^{1/2}} \exp\left(-\frac{x^2}{2p}\right)$$

since

$$\Pr\left[W_{1,p} \geqslant \frac{x+w}{2} \;\middle|\; W(p) = w\right] = \Pr\left[W_{1,1} \geqslant \frac{x+w}{2\sqrt{p}} \;\middle|\; W(1) = \frac{w}{\sqrt{p}}\right]$$
$$= \exp[-(x^2 - w^2)/(2p)], \quad |w| < x,$$

where the second equality follows from the well-known property (cf. Karlin and Taylor, 1975, p. 388)

$$\Pr[W_{1,1} \geqslant a \mid W(1) = b] = \exp[-2a(a-b)], \quad a > 0, \quad b < a. \tag{A.1}$$

Thus the lemma follows.

We can write

$$\Pr(Q_p \geqslant x) = \Pr[\max(Q_{1,p}, -Q_{2,p}) \geqslant x]$$
$$= 2\Pr(Q_{1,p} \geqslant x) - \Pr(Q_{1,p} \geqslant x,\ Q_{2,p} \leqslant -x).$$

Now, $\Pr(Q_{1,p} \geqslant x)$ is given by the right-hand side of (2) and

$$\Pr(Q_{1,p} \geqslant x,\ Q_{2,p} \leqslant -x)$$
$$= 2\left[1 - \Phi\left(\frac{x}{\sqrt{p}}\right)\right] + \int_{-x}^{x} \Pr\left[W_{1,p} \geqslant \frac{x+w}{2},\ W_{2,p} \leqslant \frac{-x+w}{2} \;\middle|\; W(p) = w\right] f_{W(p)}(w)\,dw. \tag{A.2}$$

To complete the proof of the theorem it is sufficient to show

$$I_p(x) = \frac{1}{2}\left[\Phi\left(\frac{3x}{\sqrt{p}}\right) - \Phi\left(\frac{x}{\sqrt{p}}\right)\right] + C_p(x), \tag{A.3}$$

where $I_p(x)$ denotes the integral in (A.2) and $C_p(x)$ is given by (4). Application of the reflection principle gives

$$\Pr\left[W_{1,p} \geqslant \frac{x+w}{2},\ W_{2,p} \leqslant \frac{-x+w}{2} \;\middle|\; W(p) = w\right]$$
$$= \sum_{j=1}^{\infty} (-1)^{j+1} \Pr\left[W_{1,p} \geqslant (j+1)\left(\frac{x+w}{2}\right) + j\left(\frac{x-w}{2}\right) \;\middle|\; W(p) = w\right]$$
$$+ \sum_{j=1}^{\infty} (-1)^{j+1} \Pr\left[W_{1,p} \geqslant (j+1)\left(\frac{x-w}{2}\right) + j\left(\frac{x+w}{2}\right) \;\middle|\; W(p) = w\right],$$

where $|w| < x$. Substituting in $I_p(x)$ and using property (A.1), we obtain after straightforward algebra and integration

$$I_p(x) = \frac{1}{2} \sum_{j=1}^{\infty} (-1)^{j+1} \left\{\Phi\left[\frac{(2j+3)x}{\sqrt{p}}\right] - \Phi\left[\frac{(2j-1)x}{\sqrt{p}}\right]\right\} + C_p(x). \tag{A.4}$$

Equation (A.3) then follows by noting that the sum in (A.4) equals $\Phi(3x/\sqrt{p}) - \Phi(x/\sqrt{p})$.