Statistica Sinica 2(1992), 265-283

A CLASS OF NONPARAMETRIC PROCEDURES FOR COMPARING TWO SURVIVAL DISTRIBUTIONS OVER AN INTERVAL BASED ON RANDOMLY RIGHT CENSORED DATA

Shie-Shien Yang

Kansas State University

Abstract: A class of nonparametric tests for comparing two survival distributions $\bar{F}$ and $\bar{G}$ based on randomly right censored data is proposed. The tests developed assume neither the proportionality of the hazard functions of $\bar{F}$ and $\bar{G}$ nor the equality of the censoring distributions. A subclass of the proposed tests compares favorably to the Gehan and logrank tests for testing standard hypotheses. It can be used to test hypotheses concerning F and G over a prespecified interval.

Key words and phrases: Censored data, crossed survival distributions, quantile.

1. Introduction

In clinical studies, some patients may withdraw from or drop out of the study; and some may still be alive at the end of the study. Therefore, the lifetimes of the patients are not all directly observable. Instead, we observe only right censored lifetimes. There are many efficient nonparametric procedures for comparing two survival functions in the presence of right censoring. The two most notable ones are Gehan's (1965) extension of the Wilcoxon test and Mantel's (1966) logrank test which is closely related to Cox's (1972) regression test and Mantel-Haenszel's (1959) test. Other extensions of these tests can be found in Efron (1967), Peto and Peto (1972), Tarone and Wane (1977) and Prentice (1978).

Gehan's test, the logrank test, and other related tests tend to be insensitive to the "cross-hazards alternative". Fleming, O'Fallon, O'Brien and Harrington (1980) proposed a modified Kolmogorov-Smirnov test to handle this type of alternative. Empirical results seem to show that their test is more powerful than Gehan's test and the logrank test under the "cross-hazards alternative". The modified Kolmogorov-Smirnov test is designed to detect a difference in the distributions which is particularly significant for at least one point in time.

In this article, a class of nonparametric tests is proposed. A subclass of the tests can be used to compare two survival distributions over a prespecified interval of time. Testing a hypothesis of this type may be desirable in some situations. For example, life testing experimentation is often terminated over a prespecified time interval, and thus it is natural to test the equality of survival distributions over this interval. Also, in clinical trials involving aggressive treatment, patients, who are given such treatment, generally may experience one of the following two consequences: (1) They may experience an excellent short-term improvement in survival rate but little or no long-term survival advantage over individuals subjected to the standard treatment. (2) They may experience much improved long-term survival even though their survival during the initial stage of the treatment may be worse than that of those given the standard treatment. In both cases, researchers would like to compare the survival distributions over a prespecified interval of time. The rest of this article is organized as follows. The proposed class of tests is described in Section 2. The asymptotic distribution and properties of the tests are derived in Section 3. The results of a Monte Carlo simulation study are presented in Section 4. Application of the proposed tests to real data sets is presented in Section 5. Section 6 contains concluding remarks.

2. The Proposed Class of Test

Suppose that true lifetimes $X_1^0, \ldots, X_m^0$ form a sample drawn from a distribution F and that $Y_1^0, \ldots, Y_n^0$ form an independent sample drawn from a distribution G. Let $U_1, \ldots, U_m$ and $V_1, \ldots, V_n$ be independent samples of censoring variables from the censoring distributions H and I respectively. The censoring variables are assumed to be independent of the true lifetimes. The censoring distributions can be different.

We observe $X_i = X_i^0 \wedge U_i$, $\delta_i = I_{[X_i^0 = X_i]} (i = 1, \ldots, m)$; $Y_j = Y_j^0 \wedge V_j$, $\epsilon_j = I_{[Y_j^0 = Y_j]} (j = 1, \ldots, n)$. Throughout this article the notations $a \wedge b$ and $a \vee b$ denote respectively the smaller and larger of a and b; and $I_{[A]}$ denotes the indicator function of event A. Based on (X_i, δ_i) $(i = 1, \ldots, m)$ and (Y_j, ϵ_j) $(j = 1, \ldots, n)$, we wish to compare the two survival distributions $\bar{F} = 1 - F$ and $\bar{G} = 1 - G$ over a prespecified interval of time.

Let $W(x)$ and $\hat{W}(x)$ be two bounded nondecreasing functions. Assume that $\hat{W}(x)$ is also a function of $(X_1, \delta_1), \ldots, (X_m, \delta_m)$ and $(Y_1, \epsilon_1), \ldots, (Y_n, \epsilon_n)$ and that $\hat{W}(x)$ converges to $W(x)$ with probability one as $m, n \to \infty$. Let $\hat{F}$ and $\hat{G}$ denote respectively the Kaplan-Meier (1958) product limit estimators of F and G. One natural class of test statistics for comparing F and G over the positive real line is of the form

$$S = \int_0^1 [\hat{G} \circ \hat{Q}(p) - \hat{F} \circ \hat{Q}(p)] \, d\hat{W} \circ \hat{Q}(p).$$

Throughout this paper, let $f \circ g(x) = f(g(x))$, $Q(p) = \inf\{x | F(x) \geq p\}$, and $\hat{Q}(p) = \inf\{x | \hat{F}(x) \geq p\}$. Let $0 < p_1 < \cdots < p_k \leq q < 1$ be the points at which $\hat{W} \circ \hat{Q}(p)$ has jumps. Let $\hat{\xi}_i = \hat{Q}(p_i)$ $(i = 1, \ldots, k)$ and define $\hat{\xi}_0 = -\infty$. Then we can write

$$S = \sum_{i=1}^{k} [\hat{G}(\hat{\xi}_i) - \hat{F}(\hat{\xi}_i)][\hat{W}(\hat{\xi}_i) - \hat{W}(\hat{\xi}_{i-1})].$$

In the next section, it is shown that under the null hypothesis $F = G$ on $(0, \infty)$ and some regularity conditions as m, $n \to \infty$ with $n/m \to \lambda$, $0 < \lambda < \infty$, $n^{1/2} S$ converges in distribution to a normal random variable with zero mean and variance

$$\sigma_S^2 = \lambda \int_0^q \int_0^q \Gamma_F(Q(s), Q(t))\, dW \circ Q(s) dW \circ Q(t) + \int_0^q \int_0^q \Gamma_G(Q(s), Q(t))\, dW \circ Q(s) dW \circ Q(t)$$

where for $u \leq v$,

$$\Gamma_F(u, v) = [1 - F(u)][1 - F(v)] \int_0^u \frac{d\tilde{F}(x)}{[1 - FH(x)]^2} \tag{1}$$

$$\Gamma_G(u, v) = [1 - G(u)][1 - G(v)] \int_0^u \frac{d\tilde{G}(y)}{[1 - GI(y)]^2} \tag{2}$$

$$\tilde{F}(x) = P(X_i \leq x, \delta_i = 1) \tag{3}$$

$$FH(x) = 1 - [1 - F(x)][1 - H(x)] \tag{4}$$

$$\tilde{G}(y) = P(Y_i \leq y, \epsilon_i = 1) \tag{5}$$

$$GI(y) = 1 - [1 - G(y)][1 - I(y)]. \tag{6}$$

The expression for σ_S^2 is still true (see Section 3) even if $F \neq G$. In fact σ_S^2 can be simplified when $F = G$.

Let $X'_{(1)} < X'_{(2)} < \cdots < X'_{(m')}$ and $Y'_{(1)} < \cdots < Y'_{(n')}$ be the m' distinct ordered values of the X_i's and the n' distinct ordered values of the Y_i's respectively. Let rx_i and ry_i be the total number of individuals that are alive or dead at $X'_{(i)}$ and $Y'_{(i)}$ respectively; and dx_i and dy_i be the total number of individuals dead at $X'_{(i)}$ and $Y'_{(i)}$ respectively. Let $\hat{D}_i = [\hat{W}(\hat{\xi}_i) - \hat{W}(\hat{\xi}_{i-1})]$, $i = 1, \ldots, k$. Then, an estimate of σ_S^2 is

$$\hat{\sigma}_S^2 = \sum_{r=1}^{k} \sum_{s=1}^{k} [(n/m)\hat{\Gamma}_F(\hat{\xi}_r, \hat{\xi}_s)\hat{D}_r \hat{D}_s + \hat{\Gamma}_G(\hat{\xi}_r, \hat{\xi}_s)\hat{D}_r \hat{D}_s], \tag{7}$$

where for $\hat{\xi}_r \le \hat{\xi}_s$,

$$\hat{\Gamma}_F(\hat{\xi}_r,\hat{\xi}_s) = [1-\hat{F}(\hat{\xi}_r)][1-\hat{F}(\hat{\xi}_s)]\sum_{i=1}^{m'}\left\{\frac{[X'_{(i)} \le \hat{\xi}_r](dx_i/m)}{(rx_i/m)(rx_i - dx_i)/m}\right\} \tag{8}$$

$$\hat{\Gamma}_G(\hat{\xi}_r,\hat{\xi}_s) = [1-\hat{G}(\hat{\xi}_r)][1-\hat{G}(\hat{\xi}_s)]\sum_{i=1}^{n'}\left\{\frac{[Y'_{(i)} \le \hat{\xi}_r](dy_i/n)}{(ry_i/n)(ry_i - dy_i)/n}\right\}. \tag{9}$$

Note that $\hat{\Gamma}_F(\hat{\xi}_r,\hat{\xi}_r)/m$ is the Greenwood (1926) formula for the variance estimate of $\hat{F}(\xi_r)$. Also note that, for example, if $X_{(m)}$ is an uncensored observation, then $\hat{\xi}_k \ge X_{(m)}$ and $\hat{\Gamma}_G(\hat{\xi}_k,\hat{\xi}_k) = \infty$. Hence in practice, if either $X_{(m)}$ or $Y_{(n)}$ is uncensored, we sum over all $\hat{\xi}_i$'s that are less than $X_{(m)} \wedge Y_{(n)}$ in the computation of S and $\hat{\sigma}_S$. Now, $Z_S = n^{1/2}S/\hat{\sigma}_S$, which is approximately standard normal under H_0, can be used to test the hypothesis concerning F and G.

Suppose we wish to compare the two survival distributions $\bar{F} = 1 - F$ and $\bar{G} = 1 - G$ over some prespecified interval (a,b). If a and b are known, then we can choose $\hat{W}(x)$ and $W(x)$ so that $d\hat{W}(x) = dW(x) = 0$ for x outside (a,b). If a and b are unknown, and consistent estimates $\hat{a}$ and $\hat{b}$ of a and b are available, then we can choose $\hat{W}(x)$ and $W(x)$ so that $d\hat{W}(x) = 0$ outside the interval $(\hat{a},\hat{b})$. Let $S(a,b)$ denote the proposed test statistic with $d\hat{W}(x) = 0$ outside either (a,b) or $(\hat{a},\hat{b})$ depending on whether or not a and b are known. $n^{1/2}S(a,b)/\hat{\sigma}_S$, which is also approximately standard normal under H_0, can be used to test hypotheses of the following types in the usual way:

(1) The null hypothesis is $\bar{F}(x) = \bar{G}(x)$ for $a < x < b$ and the alternative is $\bar{F}(x) \ne \bar{G}(x)$ for some $x \in (a,b)$.
(2) The null hypothesis is $\bar{F}(x) \le (\ge)\bar{G}(x)$ for $a < x < b$ and the alternative is $\bar{F}(x) > (<)\bar{G}(x)$ for $a < x < b$.

Two examples, where a and b need to be estimated, are presented below. In the first example, we assume that F is the distribution function of the life times of a population treated by a standard treatment (control population). It may be desirable to compare $\bar{F}$ and $\bar{G}$ over an interval defined by two prespecified percentiles of F, $a = Q(p_1)$ and $b = Q(p_2)$. In this case, a and b can be consistently estimated by $\hat{a} = \hat{Q}(p_1)$ and $\hat{b} = \hat{Q}(p_2)$. In the next example, we assume that $\bar{F}(x) < \bar{G}(x)$ cross exactly at one point a $(0 < a < \infty)$. Suppose that we want to compare $\bar{F}(x)$ and $\bar{G}(x)$ on the interval $(-\infty,a)$, then the estimated interval $(-\infty,\hat{a})$ is the largest interval on which either $\hat{F}(x) < \hat{G}(x)$ or $\hat{F}(x) > \hat{G}(x)$. A similar idea can be used to compare $\bar{F}$ and $\bar{G}$ over (a,∞). Note that $\hat{a}$ is a consistent estimator of a.

Next, we discuss the choice of a weight function $\hat{W}(x)$. In principle, if H, I and, apart from some unknown parameters, the functional forms of F and G are known, it is possible to choose a weight function $W(x)$ which will yield an asymptotically optimal test. Since, in practice, F, G, H and I are rarely known, we do not pursue it here. However, some prior information about F and G can help us choose a weight function. For example, if distribution differences are expected to be evident late in time, we should choose $\hat{W}(x)$ with larger jumps $d\hat{W}(x)$ at later times and smaller jumps $d\hat{W}(x)$ at earlier times. In this article, we study only the weight function $\hat{W}(x) = \hat{G}(\hat{a}^-)I(x < \hat{a}) + \hat{G}(x)I(\hat{a} \leq x < \hat{b}) + \hat{G}(b)I(x \geq \hat{b})$ where $\hat{G}(\hat{a}^-)$ is the left hand limit of $\hat{G}$ at $x = \hat{a}$. Note that if $a = 0$ and $b = \infty$, S is an estimate of $1/2 - P(X^0 > Y^0)$.

With the above choice of weight function, $S(a,b)$ is not symmetric in F and G. Experience from application of $S(a,b)$ to real and simulated data seems to suggest that labeling F and G so that $F < G$ over the interval (a,b) of interest will generally lead to a larger value of $S(a,b)$ and thus a more powerful test. This may be explained by the following facts.

Let $p_1 < p_2 < \cdots < p_k$ be the points at which $\hat{Q}(p)$ has jumps. The $\hat{\xi}_i = \hat{Q}(p_i)$ $(i = 1, \ldots, k)$ are the ordered uncensored observations of the X-sample. If $\hat{G}(x) > \hat{F}(x)$ on some interval (a,b), then, on any interval $(\hat{\xi}_i, \hat{\xi}_{i+1})$ contained in (a,b), $\hat{G}(x)$ is likely to have a large number of jumps in $(\hat{\xi}_i, \hat{\xi}_{i+1})$. Therefore, the jump sizes $d\hat{G}\circ\hat{Q}(p)$ of $\hat{G}\circ\hat{Q}(p)$ over the interval in which $\hat{G}(x) > \hat{F}(x)$ will tend to be larger than those over the interval in which $\hat{G}(x) \leq \hat{F}(x)$. Hence, in the computation of S, $[\hat{G}\circ\hat{Q}(p) - \hat{F}\circ\hat{Q}(p)]$ will tend to receive larger weights over the interval in which $\hat{G}(x) > \hat{F}(x)$ and smaller weights (sometime zero weight) over the interval in which $\hat{G}(x) \leq \hat{F}(x)$. We adopt this labeling convention to construct the proposed test with this choice of weight function.

3. Some Asymptotic Results

Assume, throughout, that F, G, H and I are continuous. Let $\xi_p = Q(p)$ and $\hat{\xi}_p = \hat{Q}(p)$, and T_1 and T_2 be the uppermost support points of 1-FH and 1-GI respectively. Let q be some positive number between 0 and 1 such that $\xi_q < T_1 \wedge T_2$.

The following general conditions will be needed to prove the asymptotic results.

1. F has bounded second derivative on $[0, \xi_q + \delta)$ for some $\delta > 0$.
2. $\inf_{0 \leq p \leq q} F'(\xi_p) > 0$.
3. $F'(x)$ is bounded on $[0, \xi_q]$.

Theorem 1. *Suppose conditions* 1 *to* 3 *hold. Then as* $m, n \to \infty$ *with* $n/m \to \lambda (0 < \lambda < \infty)$, $m^{1/2}[\hat{F}\circ\hat{Q}(p) - F\circ\hat{Q}(p)]$ *and* $n^{1/2}[\hat{G}\circ\hat{Q}(p) - G\circ\hat{Q}(p)]$ *converge* *weakly to mean zero Gaussian processes* $Z_F(p)$ *and* $Z_G(p)$, *respectively, defined on* $0 \le p \le q$. *The covariance structures of* $Z_F(p)$ *and* $Z_G(p)$ *are given for* $0 < s \le t \le q$ *by* $\Gamma_F(Q(s), Q(t))$ *and* $\Gamma_G(Q(s), Q(t))$ *respectively as defined by equations* (1) *and* (2). *Also,* $Z_F(p)$ *and* $Z_G(p)$ *are independent.*

Proof. We shall only prove the result for $\hat{G}(\hat{\xi}_p)$. Exactly the same argument can be used to prove the result for $\hat{F}(\hat{\xi}_p)$. By Corollary 1 (i) of Cheng (1984), as $n \to \infty$

$$\sup_{0 \le p \le q} |\hat{\xi}_p - \xi_p| = O(n^{-1/2}(\log\log n)^{1/2}).$$

The above fact and Theorem 1 of Cheng (1984) imply that, with probability one, as $m, n \to \infty$ with $n/m \to \lambda$,

$$\sup_{0 \le p \le q} |n^{1/2}[\hat{G}(\hat{\xi}_p) - G(\hat{\xi}_p)] - n^{1/2}[\hat{G}(\xi_p) - G(\xi_p)]| = O(n^{-1/4}(\log n)^{3/4}). \quad (10)$$

Note that conditions 1 to 3 are needed to apply Cheng's results. By Theorem 5 of Breslow and Crowley (1974), as $n \to \infty$ for $0 < p < q$, $n^{1/2}[\hat{G}(\xi_p) - G(\xi_p)]$ converges weakly to a Gaussian process $Z_G(p)$ on $(0, q)$. The independence of $Z_F(p)$ and $Z_G(p)$ follows from that of $\hat{F}(\xi_p)$ and $\hat{G}(\xi_p)$. This completes the proof of the theorem.

Using (2.3) of Cheng (1984) and Corollary 6.1 of Burke, Csörgö and Horváth (1981), we have, with probability one,

$$\sup_{0 \le p \le q} |m^{1/2}[p - \hat{F}(\xi_p)] - Z_F(p)| = O(m^{-1/3}(\log m)^{3/2})$$

as $m \to \infty$. Hence by Equation (10), we also have the following Corollary:

Corollary 1. *Suppose the conditions of Theorem* 1 *are satisfied, and* $0 \le p_1 < p_2 \le q$. *Then, with probability one, as* $m, n \to \infty$, *we have*

$$\sup_{p_1 \le p \le p_2} |m^{1/2}[\hat{F}(\hat{\xi}_p) - F(\hat{\xi}_p)] - Z_F(p)| = O(m^{-1/4}(\log m)^{3/2})$$
$$\sup_{p_1 \le p \le p_2} |n^{1/2}[\hat{G}(\hat{\xi}_p) - G(\hat{\xi}_p)] - Z_G(p)| = O(n^{-1/4}(\log n)^{3/2}).$$

Theorem 2. *Let* $p_1 = F(a)$ *and* $p_2 = F(b)$ *where* $0 \le p_1 < p_2 \le q < 1$. *Suppose that* $\hat{a}$ *and* $\hat{b}$ *converge respectively to* a *and* b *with probability one. Suppose that* *assumptions* 1 *to* 3 *are satisfied. Let* $\hat{p}_1 = \hat{F}(\hat{a})$ *and* $\hat{p}_2 = \hat{F}(\hat{b})$. *Define*

$$T(a,b) = \int_{\hat{p}_1}^{\hat{p}_2} [\hat{G}\circ\hat{Q}(p) - \hat{F}\circ\hat{Q}(p)]\, d\hat{W}\circ\hat{Q}(p) - \int_{\hat{p}_1}^{\hat{p}_2} [G\circ\hat{Q}(p) - F\circ\hat{Q}(p)]\, d\hat{W}\circ\hat{Q}(p)$$

$$J_K(\hat{p}_1,\hat{p}_2) = \int_{\hat{p}_1}^{\hat{p}_2} M^{1/2}[\hat{K}\circ\hat{Q}(p) - K\circ\hat{Q}(p)]\, d[\hat{W}\circ\hat{Q}(p) - W\circ Q(p)]$$

where $K = F$ *or* G *and* $M = m$ *if* $K = F$ *and* $M = n$ *if* $K = G$. *If, as* $m, n \to \infty$ *with* $n/m \to \lambda$ $(0 < \lambda < \infty)$, *and* $J_G(\hat{p}_1,\hat{p}_2)$ *and* $J_F(\hat{p}_1,\hat{p}_2)$ *converge to zero in probability, then* $n^{1/2}T(a,b)$ *converges in law to a normal random variable with zero mean and variance,*

$$\sigma_S^2(p_1,p_2) = \lambda \int_{p_1}^{p_2}\int_{p_1}^{p_2} \Gamma_F(Q(s),Q(t))\, dW\circ Q(s)\, dW\circ Q(t) + \int_{p_1}^{p_2}\int_{p_1}^{p_2} \Gamma_G(Q(s),Q(t))\, dW\circ Q(s)\, dW\circ Q(t). \tag{11}$$

Proof. Write

$$\hat{p}_1 - p_1 = \hat{F}(\hat{a}) - F(a) = \hat{F}(\hat{a}) - F(\hat{a}) + F(\hat{a}) - F(a).$$

By the continuity of F and uniform convergence of $\hat{F}$ to F, $\hat{p}_1$ converges to p_1 with probability one. Similarly, $\hat{p}_2$ converges to p_2 with probability one. Write

$$n^{1/2}T(a,b) = \int_{\hat{p}_1}^{\hat{p}_2} n^{1/2}[\hat{G}\circ\hat{Q}(p) - G\circ\hat{Q}(p)]\, dW\circ Q(p) + J_G(\hat{p}_1,\hat{p}_2) - \left(\frac{n}{m}\right)^{1/2} \int_{\hat{p}_1}^{\hat{p}_2} m^{1/2}[\hat{F}\circ\hat{Q}(p) - F\circ\hat{Q}(p)]\, dW\circ Q(p) - \left(\frac{n}{m}\right)^{1/2} J_F(\hat{p}_1,\hat{p}_2). \tag{12}$$

Consider the first integral

$$\int_{\hat{p}_1}^{\hat{p}_2} n^{1/2}[\hat{G}\circ\hat{Q}(p) - G\circ\hat{Q}(p)]\, dW\circ Q(p) = \int_{p_1}^{p_2} n^{1/2}[\hat{G}\circ\hat{Q}(p) - G\circ\hat{Q}(p)]\, dW\circ Q(p) + \int_0^q \{I_{[\hat{p}_1\le p\le\hat{p}_2]} - I_{[p_1\le p\le p_2]}\} n^{1/2}[\hat{G}\circ\hat{Q}(p) - G\circ\hat{Q}(p)]\, dW\circ Q(p). \tag{13}$$

By Theorem 1 and the invariance principle for weak convergence, the first integral of the right hand side of Equation (13) converges in law to a normal random variable V_G with zero mean and variance given by the second integral of (11). Since $\sup_x n^{1/2}|\hat{G}(x) - G(x)|$ converges in law and $\hat{p}_i$ converges to p_i $(i = 1,2)$ with probability one, the second integral of (13) converges to zero in probability as $m, n \to \infty$. Hence, the first integral of (12) converges in distribution to V_G. Likewise, the third integral of (12) converges in law as $m, n \to \infty$ with $n/m \to \lambda$ to a normal random variable V_F with zero mean and variance given by the first integral of (11). Also, V_G and V_F are independent. By assumption, $J_G(\hat{p}_1, \hat{p}_2)$ and $J_F(\hat{p}_1, \hat{p}_2)$ converge to zero in probability as $m, n \to \infty$ with $n/m \to \lambda$. Hence, the result of the theorem follows.

With additional conditions on $\hat{W}(x)$ and $W(x)$, we can show that as $m, n \to \infty$, $J_F(\hat{p}_1, \hat{p}_2)$ and $J_G(\hat{p}_1, \hat{p}_2)$ converge to zero with probability one.

Theorem 3. *Suppose the conditions of Theorem 2 are satisfied. In addition, assume that $W(x)$ is a continuous function and that $\sup_x |\hat{W}(x) - W(x)| \to 0$ with probability one as m and $n \to \infty$. Then $J_F(\hat{p}_1, \hat{p}_2)$ and $J_G(\hat{p}_1, \hat{p}_2)$ converge to zero with probability one as m and $n \to \infty$.*

Proof. We shall only prove the result for $J_G(\hat{p}_1, \hat{p}_2)$. The same argument can be used to prove the result for $J_F(\hat{p}_1, \hat{p}_2)$.

Let

$$\hat{D}(p) = \hat{W} \circ \hat{Q}(p) - W \circ Q(p)$$
$$\hat{Z}_G(p) = n^{1/2}[\hat{G} \circ \hat{Q}(p) - G \circ \hat{Q}(p)].$$

Then write

$$\begin{aligned} J_G(\hat{p}_1, \hat{p}_2) &= \int_{\hat{p}_1}^{\hat{p}_2} \hat{Z}_G(p)\, d\hat{D}(p) \\ &= \int_0^q \{I_{[\hat{p}_1 \le p \le \hat{p}_2]} - I_{[p_1 \le p \le p_2]}\} \hat{Z}_G(p)\, d\hat{D}(p) + \int_{p_1}^{p_2} \hat{Z}_G(p)\, d\hat{D}(p) \\ &= I_1 + I_2. \end{aligned}$$

$$\begin{aligned} |I_1| \le \sup_x |n^{1/2}[\hat{G}(x) - G(x)]| \int_0^q |I_{[\hat{p}_1 \le p \le \hat{p}_2]} - I_{[p_1 \le p \le p_2]}|\, d\hat{W} \circ \hat{Q}(p) \\ + \sup_x |n^{1/2}[\hat{G}(x) - G(x)]| \int_0^q |I_{[\hat{p}_1 \le p \le \hat{p}_2]} - I_{[p_1 \le p \le p_2]}|\, dW \circ Q(p). \end{aligned}$$

Since, with probability one for sufficiently large m and n, $(\hat{p}_1, \hat{p}_2) \cap (p_1, p_2) \neq \emptyset$, it follows that

$$\begin{aligned}&\int_0^q |I_{[\hat{p}_1 \le p \le \hat{p}_2]} - I_{[p_1 \le p \le p_2]}|\, d\hat{W}\circ\hat{Q}(p)\\ \le&[\hat{W}\circ\hat{Q}(\hat{p}_1 \vee p_1) - \hat{W}\circ\hat{Q}(\hat{p}_1 \wedge p_1)] + [\hat{W}\circ\hat{Q}(\hat{p}_2 \vee p_2) - \hat{W}\circ\hat{Q}(\hat{p}_2 \wedge p_2)].\end{aligned}$$

Write

$$\begin{aligned}&[\hat{W}\circ\hat{Q}(\hat{p}_i \vee p_i) - \hat{W}\circ\hat{Q}(\hat{p}_i \wedge p_i)]\\ =&[\hat{W}\circ\hat{Q}(\hat{p}_i \vee p_i) - W\circ\hat{Q}(\hat{p}_i \vee p_i)] - [\hat{W}\circ\hat{Q}(\hat{p}_i \wedge p_i) - W\circ\hat{Q}(\hat{p}_i \wedge p_i)]\\ &\quad + [W\circ\hat{Q}(\hat{p}_i \vee p_i) - W\circ Q(p_i)] - [W\circ\hat{Q}(\hat{p}_i \wedge p_i) - W\circ Q(p_i)].\end{aligned} \tag{15}$$

Since $\hat{p}_i \vee p_i$ and $\hat{p}_i \wedge p_i$ converge to p_i almost surely, $\sup_{0\le p\le q} |\hat{Q}(p) - Q(p)|$ and $\sup_x |\hat{W}(x) - W(x)|$ converge to zero almost surely, and $W(x)$ is continuous, the expression in (15) converges to zero almost surely. It follows that I_1 converges to zero almost surely.

Now write

$$I_2 = \int_{p_1}^{p_2} [\hat{Z}_G(p) - Z_G(p)]\, d\hat{D}(p) + \int_{p_1}^{p_2} Z_G(p) d\hat{D}(p) = I_3 + I_4.$$

By Corollary 1, I_3 converges to zero with probability one as m and $n \to \infty$.

To show that $I_4 \to 0$, consider any sample sequence:

$$(X_1, \delta_1), (X_2, \delta_2), \ldots ; (Y_1, \epsilon_1), (Y_2, \epsilon_2), \ldots$$

for which $\sup_{0\le p\le q} |\hat{D}(p)| \to 0$ as $m, n \to \infty$. Along any such sample sequence and any fixed m and n, I_4 is a Stieltjes integral of a continuous function $Z_G(p)$ on $[p_1, p_2]$.

Let $p_1 < q_1 < q_2 < \ldots < q_k = p_2$ and $q_0 = p_1$. For $i = 1, \ldots, k$, let $m_i = \min_{q_{i-1}\le p\le q_i} Z_G(p)$, $M_i = \max_{q_{i-1}\le p\le q_i} Z_G(p)$. Define

$$\begin{aligned}B_k&=\sum_{i=1}^k M_i[W\circ Q(q_i) - W\circ Q(q_{i-1})], & b_k&=\sum_{i=1}^k m_i[W\circ Q(q_i) - W\circ Q(q_{i-1})],\\ \hat{B}_k&=\sum_{i=1}^k M_i[\hat{W}\circ\hat{Q}(q_i) - \hat{W}\circ\hat{Q}(q_{i-1})], & \hat{b}_k&=\sum_{i=1}^k m_i[\hat{W}\circ\hat{Q}(q_i) - \hat{W}\circ\hat{Q}(q_{i-1})],\\ B&=\int_{p_1}^{p_2} Z_G(p)\, dW\circ Q(p), & \hat{B}&=\int_{p_1}^{p_2} Z_G(p)\, d\hat{W}\circ\hat{Q}(p).\end{aligned}$$

Given any $\epsilon > 0$, $q_1, q_2, \ldots, q_k$ are chosen so that $|B - B_k| < \epsilon$ and $|B - b_k| < \epsilon$.

Then write

$$(\hat{B}_k - B) = (\hat{B}_k - B_k) + (B_k - B) < (\hat{B}_k - B_k) + \epsilon$$

$$|\hat{B}_k - B_k| \le \sum_{i=1}^{k} |M_i|[|\hat{W}\circ\hat{Q}(q_i) - W\circ Q(q_i)| + |\hat{W}\circ\hat{Q}(q_{i-1}) - W\circ Q(q_{i-1})|].$$

Hence, along such a sample sequence, $\lim_{m,n\to\infty}(\hat{B}_k - B_k) = 0$, and $\overline{\lim}_{m,n\to\infty}(\hat{B}_k - B) < \epsilon$. Similarly, we can show that $\underline{\lim}_{m,n\to\infty}(\hat{b}_k - B) > -\epsilon$. Since $(\hat{b}_k - B) \le (\hat{B} - B) \le (\hat{B}_k - B)$, along such a sample sequence, it follows that $-\epsilon < \underline{\lim}_{m,n\to\infty}(\hat{B} - B) \le \overline{\lim}_{m,n\to\infty}(\hat{B} - B) < \epsilon$. Since ϵ can be arbitrarily small, $(\hat{B} - B)$ converges to zero as $m, n \to \infty$ along any such sample sequence. Hence, $I_4 \to 0$ with probability one as $m, n \to \infty$ and consequently $I_2 \to 0$ with probability one. This completes the proof of the theorem.

Remark 1. If $F(x) = G(x)$ for $x \in [0, \infty)$, then $\int_0^q [G\circ\hat{Q}(p) - F\circ\hat{Q}(p)]\, d\hat{W}\circ\hat{Q}(p) = 0$ and $T(0, \xi_q)$ reduces to $S(0, \xi_q) = \int_0^q [\hat{G}\circ\hat{Q}(p) - \hat{F}\circ\hat{Q}(p)]\, d\hat{W}\circ\hat{Q}(p)$ and $n^{1/2}S(0, \xi_q)$ converges in law to a normal random variable with zero mean and variance $\sigma_S^2(0, q)$. If the support of the censoring and survival distribution is $(0, \infty)$, we can choose q arbitrarily close to 1. Hence, in application, we may assume $q = 1$ in this case.

Now suppose that $F(x) = G(x)$ for $x \in [a, b]$. Given any δ, $0 < \delta < (p_2 - p_1)/2$, define $\hat{q}_1 = \hat{p}_1 + \delta$, $\hat{q}_2 = \hat{p}_2 - \delta$, $q_1 = p_1 + \delta$, and $q_2 = p_2 - \delta_2$. Then there exist q_1' and q_2' such that $[q_1, q_2] \subset [q_1', q_2'] \subset [p_1, p_2]$, and, with probability one for sufficiently large m and n,

$$[\hat{Q}(\hat{q}_1), \hat{Q}(\hat{q}_2)] \subset [Q(q_1'), Q(q_2')] \subset [a, b].$$

Since $F(x) = G(x)$ for $x \in [a, b]$, $n^{1/2}\int_{\hat{q}_1}^{\hat{q}_2}[G\circ\hat{Q}(p) - F\circ\hat{Q}(p)]\, d\hat{W}\circ\hat{Q}(p) = 0$ with probability one for sufficiently large m and n. Hence, as $m, n \to \infty$ with $(n/m) \to \lambda$, $n^{1/2}S(\xi_{q_1}, \xi_{q_2})$ converges in law to a normal random variable with zero mean and variance $\sigma_S^2(q_1, q_2)$. Since we can make q_i arbitrarily close to p_i $(i = 1, 2)$, we may use $n^{1/2}S(\xi_{q_1}, \xi_{q_2})$ to test the null hypothesis $F(x) = G(x)$ for $x \in [a, b]$.

Remark 2. If the conditions of Theorem 3 are satisfied, then

$$\hat{\Delta} = \int_{\hat{p}_1}^{\hat{p}_2} [G\circ\hat{Q}(p) - F\circ\hat{Q}(p)]\, d\hat{W}\circ\hat{Q}(p)$$

converges with probability one to

$$\Delta = \int_{p_1}^{p_2} [G \circ Q(p) - F \circ Q(p)] \, dW \circ Q(p)$$

as m and $n \to \infty$. This implies that the test based on $S(a,b)$ is consistent. To see this, write

$$\begin{aligned}
&\hat{\Delta} - \Delta \\
&= \int_{\hat{p}_1}^{\hat{p}_2} \left\{ [G \circ \hat{Q}(p) - F \circ \hat{Q}(p)] - [G \circ Q(p) - F \circ Q(p)] \right\} dW \circ Q(p) \\
&\qquad + \int_{\hat{p}_1}^{\hat{p}_2} [G \circ \hat{Q}(p) - F \circ \hat{Q}(p)] \, d[\hat{W} \circ \hat{Q}(p) - W \circ Q(p)] \\
&\qquad + \int_0^1 \left\{ I_{[\hat{p}_1 \le p \le \hat{p}_2]} - I_{[p_1 \le p \le p_2]} \right\} [G \circ Q(p) - F \circ Q(p)] \, dW \circ Q(p).
\end{aligned} \tag{16}$$

Clearly, the first and third integrals of the right hand side of (16) converge to zero with probability one as m and $n \to \infty$. Using integration by parts, the second integral of the right hand side of (16) can be written as

$$\begin{aligned}
&[G \circ \hat{Q}(\hat{p}_2) - F \circ \hat{Q}(\hat{p}_2)][\hat{W} \circ \hat{Q}(\hat{p}_2) - W \circ Q(\hat{p}_2)] \\
&\qquad - [G \circ \hat{Q}(\hat{p}_1) - F \circ \hat{Q}(\hat{p}_1)][\hat{W} \circ \hat{Q}(\hat{p}_1) - W \circ Q(\hat{p}_1)] \\
&\qquad - \int_{\hat{p}_1}^{\hat{p}_2} [\hat{W} \circ \hat{Q}(p) - W \circ Q(p)] \, d[G \circ \hat{Q}(p) - F \circ \hat{Q}(p)].
\end{aligned}$$

By assumptions, the above expressions converge to zero with probability one.

If $W = G$, then the above integral can be written as $\int_0^q [G \circ Q(p) - F \circ Q(p)] \frac{g \circ Q(p)}{f \circ Q(p)} \, dp$ where $g(x) = G'(x)$ and $f(x) = F'(x)$. Hence, if the difference of G and F is positive and increasing over an interval, then $g \circ Q(p)/f \circ Q(p)$ will be greater than one and increasing over this interval. The reverse will be true if the difference is negative and decreasing over an interval. This may explain why labeling F and G so that $G(x) > F(x)$ for $a \le x \le b$ will generally lead to a more powerful test in detecting the alternative that $G(x) > F(x)$ for $a \le x \le b$; because choosing $W = G$ will give $[G \circ Q(p) - F \circ Q(p)]$ greater weight on $[a,b]$ than choosing $W = F$.

4. A Simulation Study

Monte Carlo simulation was conducted to determine how well the normal distribution approximates the distribution of the test statistics S and $S(a,b)$ and to compare the power of the S test with that of the Gehan and logrank tests. The simulation was done on a Zenith-386 microcomputer in Fortran 77 programming language.

The Gehan's extension of the Mann-Whitney form of the Wilcoxon test is based on the statistic $W = \sum_{i=1}^{m} \sum_{j=1}^{n} U_{ij}$ where

$$U_{ij} = \begin{cases} 1 & \text{if } X_i > Y_j \text{ and } \epsilon_j = 1 \text{ or } X_i = Y_j, \delta_i = 0, \epsilon_j = 1, \\ -1 & \text{if } X_i < Y_j \text{ and } \delta_i = 1 \text{ or } X_i = Y_j, \delta_i = 1, \epsilon_j = 0, \\ 0 & \text{otherwise.} \end{cases}$$

Let $t_{(1)} < \cdots < t_{(k)}$ be the distinct uncensored observations of the combined sample and d_i and d_{1i} be respectively the total number of deaths at $t_{(i)}$ from the two groups and from the first group (X-sample). Let A_i be the proportion of the total risk set at $t_{(i)}$ which belongs to the first group. The logrank test is based on the statistic

$$L = \sum_{i=1}^{k} (d_{1i} - A_i d_i).$$

Clearly from the definitions of these tests, they can perform poorly in detecting distribution differences when F and G cross. For example, if $F(x) > G(x)$ for $x < a$ and $F(x) < G(x)$ for $x > a$, the U_{ij} will likely give a negative contribution to W for X_i and $Y_j < a$ and a positive contribution to W for X_i and $Y_j > a$. Similar thing happens to L when the hazard functions of F and G cross.

The lifetime distributions included in the study were exponential, Weibull, and lognormal. The lognormal random variables were obtained from normal random variables that were generated by the Box and Muller (1958) transformation. The random variables of the other two distributions were generated by utilizing a cumulative distribution transformation to uniform random variables. The censoring distribution was uniform distribution over some interval $[0, T]$.

The rejection rates of the tests based on S and $S(a, b)$ at levels = 0.1, 0.05, 0.01 for exponential and lognormal lifetimes with $F = G$ and various sample sizes are presented in Table 1. The (a, b) were chosen to be $(Q(p_1), Q(p_2))$ with $(p_1, p_2) = (0, 0.4)$, $(0.2, 0.7)$ and $(0.5, 1)$. The S test corresponds to the choice $(p_1, p_2) = (0, 1)$. Also, the censoring rates were fixed at 43% for the exponential distribution and at 55% for the lognormal distribution. The sizes of the test were estimated from 1,000 simulation samples. For $m = n = 12$ and in the intervals $(\hat{Q}(0.2), \hat{Q}(0.7))$ and especially $(\hat{Q}(0.5), \hat{Q}(1))$, there were few cases where there was only one or no uncensored observation. For these cases, the null hypothesis were accepted because there was not enough evidence to reject the null hypothesis.

Table 1. Estimated sizes of $S(a,b)$

		Sample Size		Level of Test		
Distribution	(p_1, p_2)	m	n	0.1	0.05	0.01
Exponential	(0,1.0)	12	12	0.084	0.049	0.014
	(0,1.0)	20	20	0.097	0.045	0.012
	(0,0.4)	20	20	0.090	0.061	0.009
	(0.2,0.7)	20	20	0.070*	0.042	0.014
	(0.5,1.0)	20	20	0.083	0.042	0.008
Lognormal	(0,1.0)	12	12	0.098	0.044	0.016
	(0,1.0)	20	20	0.089	0.049	0.018*
	(0,0.4)	20	20	0.104	0.056	0.016
	(0.2,0.7)	20	20	0.087	0.040	0.005
	(0.5,1.0)	20	20	0.093	0.047	0.008

Apart from the ones marked by "*", the estimated sizes were all within two standard deviations from the nominal level. Hence, Table 1 shows that for the $(\hat{a}, \hat{b})$ considered, the normal distribution approximates the distributions of $S(a,b)$ and S quite adequately.

The power comparison for the S tests, Gehan test, and logrank test under the standard situation where F and G do not cross is presented in Table 2. The levels of these tests are set at $\alpha = 0.05$. The rejection rules of these tests were all based on their asymptotic critical values. The power estimates are based on 500 simulation samples. When sampling from the exponential distribution, the scale parameters were fixed at 1.0 for distribution F and were chosen to be 1.0 (0.2) 2.0 successively for distribution G. When sampling from the Weibull distribution the scale and shape parameters were fixed respectively at 0.01 and 1.0 for distribution F. For distribution G, the shape parameter is 1.0 for scale parameters chosen to be 0.01, and is 0.85 for scale parameter chosen to be 0.0121, 0.0136, 0.0155, 0.0181, and 0.0218. When sampling from the lognormal distribution, the variances of the log-transformed variable were fixed at 1.0 for both distributions F and G. The means of the log-transformed variables were fixed at 0 for distribution G but were chosen to be 0 (0.2) 1.2 successively for distribution F.

Since asymptotic critical values were used, the observed levels of significance of the tests are different – some are lower and some are higher than the asymptotic nominal level $\alpha = 0.05$. Therefore, in order to have a fair comparison of these tests, the sums of α and β, the probability of type II error, are also presented in Table 2.

Table 2. Power estimates of the Gehan, logrank and S tests under exponential, Weibull and lognormal distributions. Asymptotic nominal level $\alpha = 0.05$

	m	n	Power			$\alpha + \beta$		
Distribution	(% censored)		Gehan	Logrank	S	Gehan	Logrank	S
Exponential	20	20	0.038	0.048	0.054			
	(43	43)	0.078	0.084	0.094	0.960	0.964	0.960
			0.084	0.074	0.146	0.954	0.974	0.908
			0.142	0.162	0.198	0.896	0.886	0.856
			0.172	0.186	0.268	0.866	0.862	0.786
			0.300	0.302	0.344	0.738	0.746	0.710
	20	20	0.044	0.032	0.058			
	(43	0)	0.078	0.078	0.084	0.966	0.954	0.974
			0.102	0.110	0.146	0.942	0.922	0.912
			0.182	0.210	0.224	0.862	0.822	0.834
			0.240	0.332	0.318	0.804	0.700	0.740
			0.404	0.470	0.412	0.640	0.562	0.646
	30	15	0.052	0.042	0.052			
	(43	0)	0.084	0.094	0.114	0.968	0.948	0.938
			0.128	0.152	0.196	0.924	0.890	0.856
			0.198	0.258	0.280	0.854	0.784	0.772
			0.330	0.430	0.370	0.722	0.612	0.682
			0.382	0.480	0.474	0.670	0.562	0.578
Weibull	20	20	0.044	0.050	0.052			
	(55	55)	0.082	0.070	0.072	0.962	0.980	0.980
			0.188	0.150	0.136	0.856	0.900	0.916
			0.276	0.234	0.352	0.768	0.816	0.700
			0.532	0.446	0.466	0.512	0.604	0.586
			0.680	0.556	0.644	0.364	0.494	0.460
	20	20	0.042	0.022	0.052			
	(55	0)	0.080	0.032	0.112	0.962	0.990	0.940
			0.172	0.072	0.196	0.870	0.950	0.856
			0.276	0.114	0.398	0.766	0.908	0.654
			0.526	0.260	0.568	0.516	0.762	0.484
			0.648	0.334	0.704	0.394	0.688	0.348
	30	15	0.060	0.036	0.052			
	(55	0)	0.066	0.056	0.118	0.994	0.980	0.934
			0.212	0.134	0.260	0.848	0.902	0.792
			0.356	0.242	0.416	0.704	0.794	0.636
			0.512	0.382	0.554	0.548	0.654	0.498
			0.700	0.528	0.680	0.360	0.508	0.372
Lognormal	20	20	0.038	0.048	0.054			
	(43	43)	0.126	0.092	0.126	0.912	0.956	0.928
			0.106	0.076	0.190	0.932	0.972	0.864
			0.164	0.166	0.246	0.874	0.882	0.808
			0.234	0.216	0.320	0.814	0.832	0.734
			0.422	0.366	0.454	0.616	0.682	0.600
	20	20	0.044	0.032	0.058			
	(43	0)	0.106	0.088	0.124	1.038	0.944	0.934
			0.118	0.084	0.178	0.926	0.948	0.880
			0.184	0.158	0.240	0.860	0.874	0.818
			0.270	0.290	0.362	0.774	0.742	0.696
			0.490	0.502	0.532	0.554	0.530	0.526
	30	15	0.052	0.042	0.046			
	(43	0)	0.124	0.098	0.146	1.028	0.954	0.900
			0.154	0.124	0.196	0.898	0.918	0.850
			0.212	0.220	0.278	0.840	0.822	0.768
			0.364	0.362	0.414	0.688	0.680	0.632
			0.480	0.492	0.540	0.572	0.550	0.506

Table 2 suggests that for exponential lifetimes, there is no clear winner among the three tests. However, for the cases of unequal censoring rate, the performance, as measured by power and $\alpha + \beta$, of the logrank test, is better for the last two scale parameter values. If the lifetime distribution is Weibull, the performance, as measured by power and $\alpha+\beta$, of the S test, is consistently better than that of the Gehan and logrank tests. For lognormal lifetime, the logrank test performs poorly compared to the Gehan and S tests. The performances of the Gehan and S tests are similar but the S test seems to have a slight edge over the Gehan test for the cases of unequal censoring rates. The poor performance of the logrank test for Weibull and lognormal lifetimes can be explained by the nonproportionality of the hazard functions. It is interesting to note that the observed level of the S test is quite stable – it ranges from 0.046 to 0.058. The observed level of the logrank test tends to be unstable – it ranges from 0.022 to 0.048.

In order to compare the tests under "crossing-hazard alternatives", we considered three cases:

A. F and G are piecewise exponential distributions with respective hazard functions:

$$\lambda_F(x) = \begin{cases} .5, & x \in (0,.2) \\ 3, & x \in (.2,.4) \\ 1, & x \in (.4,\infty) \end{cases} \qquad \lambda_G(x) = \begin{cases} 3, & x \in (0,.2) \\ .5, & x \in (.2,.4) \\ 1, & x \in (.4,\infty) \end{cases}.$$

In this case, the differences in F and G are most evident early in time.

B. F and G are piecewise exponential with hazard functions:

$$\lambda_F(x) = \begin{cases} 2, & x \in (0,.1) \\ .5, & x \in (.1,.4) \\ 3, & x \in (.4,.7) \\ 1, & x \in (.7,\infty) \end{cases} \qquad \lambda_G(x) = \begin{cases} 2, & x \in (0,.1) \\ 3, & x \in (.1,.4) \\ .5, & x \in (.4,.7) \\ 1, & x \in (.7,\infty) \end{cases}.$$

In this case, the differences in F and G are most evident in the middle of time.

C. $F = W(.5,.5)$ and $G = W(2,2)$, where $W(\lambda,\alpha)$ denotes a Weibull distribution with distribution function $1-\exp[-(\lambda t)^\alpha]$. In this case, the differences occur late in time. F and G cross at $a = Q(0.33)$ with $G(x) > F(x)$ for $x \in (a,\infty)$.

Here, we compare the powers of the tests for testing the null hypothesis $H_0 : F(x) = G(x)$ against the alternative hypothesis $H_a : F(x) < G(x)$ over some interval (a,b) at asymptotic nominal levels $\alpha = 0.05$ and 0.01. The censoring was taken to be uniformly distributed over $[0,1]$ or $[0,2]$, and the sample sizes $m = n = 20$ and $m = n = 50$ were considered. One thousand pairs of samples were generated for each selected configuration of survival and censoring distributions and sample sizes. The power estimates are presented in Table 3.

In addition to the tests considered in Table 2, a test based on $S(a,b)$ was included in this simulation study. The interval (a,b) is the interval on which $F(x) < G(x)$ for each of the cases considered.

Table 3. Power estimates of the Gehan, logrank, S and $S(a,b)$ tests under "crossing-hazard alternatives"

Distribution			Gehan		Logrank		S		$S(a,b)$	
Survival	Censoring	n	0.01	0.05	0.01	0.05	0.01	0.05	0.01	0.05
A	U(0,1)	20	.171	.422	.074	.206	.503	.597	.571	.727
		50	.512	.785	.138	.351	.743	.844	.944	.969
	U(0,2)	20	.147	.345	.048	.174	.430	.526	.617	.747
		50	.379	.638	.078	.257	.637	.736	.948	.969
B	U(0,1)	20	.122	.306	.099	.258	.394	.483	.327	.450
		50	.322	.594	.247	.514	.686	.790	.656	.763
	U(0,2)	20	.095	.308	.065	.198	.441	.522	.412	.526
		50	.303	.548	.156	.341	.640	.730	.670	.782
C	U(0,1)	20	.014	.058	.098	.252	.120	.142	.071	.109
		50	.016	.047	.256	.534	.262	.355	.387	.457
	U(0,2)	20	.068	.183	.322	.605	.244	.309	.263	.340
		50	.104	.282	.830	.966	.551	.656	.668	.746

From Table 3, it is clear that for cases A and B, the tests based on S and $S(a,b)$ are clearly more powerful than the Gehan and logrank tests. As expected, the logrank test performs poorly in case A when distribution differences occur early in time. For case B, when distribution differences are evident in the middle of time, both the Gehan and logrank tests perform poorly. For case C, when distribution differences occur late in time, the logrank test is clearly superior and Gehan's test performs very poorly. However, the tests based on $S(a,b)$ and S are not too much inferior to the logrank test and perform much better than Gehan's test. The $S(a,b)$-test has higher estimated power than the S-test in cases A and C; but in case B, the S-test is better than the $S(a,b)$-test when the censoring distribution is $U(0,1)$.

Finally, it should be pointed out that among the four tests considered, technically speaking, only the $S(a,b)$-test is specifically designed to test the equality of the survival distributions over an interval (a,b).

5. Examples

The proposed tests were applied to two data sets which were first analyzed by Fleming et al. (1980). The data will not be presented here. Interested readers

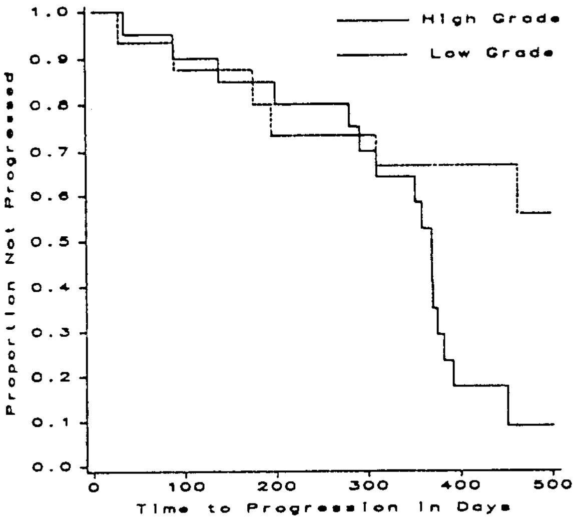

Figure 1. Survival distributions of time to progression of disease for patients with low and high grade ovarian carcinoma

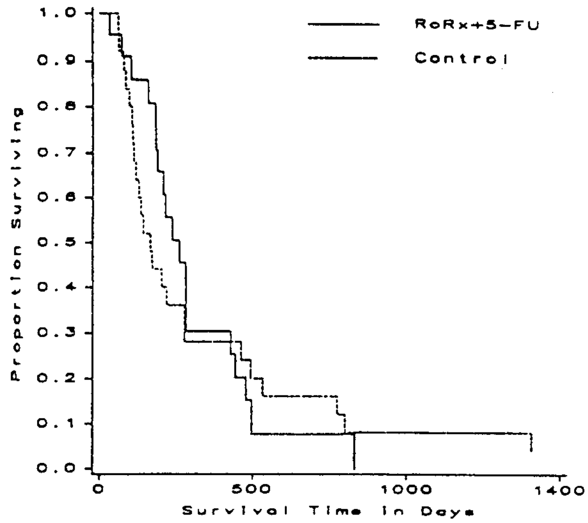

Figure 2. Survival distributions for the control group and the group receiving RoRx+5-FU

are referred to the paper by Fleming et al. (1980).

Example 1. The data of this example were collected in a Mayo Clinic study to determine whether or not grade of disease was associated with time to progression of disease. The first sample consists of times to progression of disease for 15 patients with low-grade ovarian cancer and has nine censored observations. The second sample consists of times to progression of disease for 20 patients with high-grade ovarian cancer and has four censored observations. Here F is the distribution of the low-grade group and G is that of the high-grade group.

Application of various tests to these data for testing the alternative $\bar{F} \neq \bar{G}$ yield the following results. M-K-S is the abbreviation of the modified Kolmogorov-Smirnov test.

Table 4. p-values for Example 1

Test	p-value
Gehan	0.134
Logrank	0.023
M-K-S	0.002
S	0.032

The result of the S test is compatible with that of the logrank and M-K-S tests. $S(Q(0.3), \infty)$ was also applied to compare $\bar{F}$ and $\bar{G}$ over the interval $(Q(0.3), \infty)$. The p-value of this test is 0.04 and thus indicates significance at 0.05 level.

Example 2. The data of this example were collected in another Mayo Clinic study to determine whether a group of patients treated with a combination of radiation treatment (RoRx) and drug (5-FU) would survive significantly longer than a control group. The treated group has 22 survival times with three censored observations. The control group has 25 survival times with no censored observations. Figure 2 shows that the Kaplan-Meier estimators of the survival distributions of the two samples cross at two places. Here F is the distribution of the treated group, and G is that of the control group. $\hat{F}(\hat{\xi}_p) > \hat{G}(\hat{\xi}_p)$ for $p < 0.14$ or $p > 0.7$ and $\hat{F}(\hat{\xi}_p) < \hat{G}(\hat{\xi}_p)$ for $0.14 \leq p \leq 0.7$.

Application of various tests to this data yields the following results.

Table 5. p-values for Example 2

Test	p-value
Gehan	0.127
Logrank	0.418
M-K-S	0.048
S	0.055

For the one-sided alternative $\bar{F} > \bar{G}$, the S test is almost significant at the 0.05 level while the Gehan and logrank tests are not significant even at the 0.1 level. $S(Q(0.14), Q(0.7))$ was also applied to compare F and G over the interval $(Q(0.14), Q(0.7))$. The p-value of this test is 0.036 and thus we may conclude that the treated group has significantly (at 0.05 level) higher survival probability than the control group over the interval $(Q(0.14), Q(0.7))$.

6. Conclusion

The simulation results indicate that the normal distribution approximates the distribution of the proposed tests quite adequately. Also, the proposed tests based on S compares favorably to the Gehan and logrank tests when the lifetime distributions are Weibull and lognormal. However, based on the results in Table 3, apart from case C where distribution differences occur late in time, the tests based on S and $S(a, b)$ are clearly more powerful than the Gehan and logrank tests. In case C, although the logrank test is the best test, the tests based on S and $S(a, b)$ are not too much inferior to the logrank test. The tests based on $S(a, b)$ can be used to test hypothesis concerning F and G over an interval (a, b).

The derivation of the test assumes neither the equality of the censoring distributions nor the proportionality of the hazard functions. Since the proposed test is based on a comparison of the Kaplan-Meier estimators $\hat{F}$ and $\hat{G}$, the conclusion obtained from the proposed tests can be easily reinforced and explained by comparing the graphs of $\hat{F}$ and $\hat{G}$.

Acknowledgements

The major part of this research was done during my visit to the Department of Statistics at Iowa State University in the summer of 1988 and was supported by A.R.O. grant No. 421-21-12. I would like to give my special thanks to Herbert A. David for his interest in and comments on this project, and for inviting me to I.S.U. Thanks are also due to Darlene Wicks, Retha Parker, and Jane Cox for typing various versions of this paper.

References

Box, G. E. P. and Muller, M. E. (1958). A note on the generation of random normal deviates. *Ann. Math. Statist.* **29**, 610-611.

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

Cheng, K. F. (1984). On almost sure representation for quantiles of the product limit estimator with applications. *Sankhyā Ser. A* **46**, 426-443.

Cox, D. R. (1972). Regression models and life tables (with discussion). *J. Roy. Statist. Assoc. Ser. B* **34**, 187-230.

Efron, B. (1967). The two-sample problem with censored data. *Proc. 5th Berkeley Symp. Probab. Statist.* **4**, 831-853.

Fleming, T. R., O'Fallon, J. O., 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.

Gehan, A. (1965). A generalized Wilcoxon test for comparing arbitrarily singly-censored samples. *Biometrika* **52**, 203-233.

Greenwood, M. (1926). The natural duration of cancer. Reports on Public Health and Medical Subjects, No. 33, His Majesty's Stationery Office.

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

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

Mantel, N. and Haenzel, W. (1959). Statistical aspects of the analysis of data from retrospective studies of disease. *J. National Cancer Inst.* **22**, 719-748.

Peto R. and Peto, J. (1972). Asymptotically efficient rank invariant test procedures (with discussion). *J. Roy. Statist. Soc. Ser. A* **135**, 185-207.

Prentice, R. L. (1978). Linear rank tests with right-censored data. *Biometrika* **65**, 167-179.

Tarone, R. E. and Ware, J. (1977). On distribution-free tests for equality of survival distributions. *Biometrika* **64**, 156-160.

Department of Statistics, Kansas State University, Manhattan, Kansas 66506, U.S.A.

(Received May 1989; accepted June 1991)