Signed-Rank Tests for Censored Matched Pairs

DOROTA M. DABROWSKA*

I consider the problem of testing bivariate symmetry in matched-pair experiments where the observations are subject to univariate censoring. Thus the observable random variables are given by (Y_1, Y_2) and (δ_1, δ_2), where $Y_j = \min(X_j, C)$ and $\delta_j = I(X_j \le C)$ $(j = 1, 2)$. Here (X_1, X_2) is a random pair of partially observable lifetimes and C is a fixed or random censoring variable. The hypothesis to be tested is that (X_1, X_2) and (X_2, X_1) have the same distribution. Following Woolson and Lachenbruch (1980), I consider censored data generalizations of signed-rank tests such as the sign, signed Wilcoxon, and signed-normal scores tests. I derive the asymptotic distribution of these test statistics under fixed and contiguous alternatives. The efficiencies of the signed-rank tests are considered in a log-linear model and compared with efficiencies of the paired Prentice–Wilcoxon and log-rank tests.

KEY WORDS: Bivariate symmetry; Censored data; Paired-rank tests.

1. INTRODUCTION

I consider the problem of testing whether (X_{1i}, X_{2i}) has the same distribution as (X_{2i}, X_{1i}) $(i = 1, \ldots, n)$, where (X_{1i}, X_{2i}) are iid nonnegative bivariate random vectors representing failure or survival times of paired subjects. Throughout the failure times (X_{1i}, X_{2i}) are subject to univariate right censoring, so the observable random variables are given by (Y_{1i}, Y_{2i}) and $(\delta_{1i}, \delta_{2i})$, where $Y_{ji} = \min(X_{ji}, C_i)$ and $\delta_{ji} = I(X_{ji} \le C_i)$ $(j = 1, 2; i = 1, \ldots, n)$. Here the C_i's are independent random variables representing withdrawal times from the study for reasons unrelated to the failure mechanism. It is assumed that the C's are independent of the X's. The censoring mechanism assumes that for both members of the pair the two time measurements are made on the same time clock. This will occur in the case of matched-pair experiments or twin studies when the subjects undergo the study simultaneously and are censored only if failure does not occur by the end of the study. Batchelor and Hackett (1970), Holt and Prentice (1974), and Woolson and Lachenbruch (1980), for instance, reported data on survival of skin grafts on burn patients, each of whom received two grafts. The donor and the recipient were matched for blood groups and closely or poorly matched for the transplantation antigen system. Censoring occurred at the termination of the study. Another well-known example is the study on remissions in acute leukemia patients (Freireich et al. 1963), where patients were matched according to the remission status. Within each pair, patients were assigned to either placebo or treatment. The possibly censored response variable is given here by the length of the remission period.

For uncensored data, tests for bivariate symmetry can be based on signed-rank statistics; see Doksum (1980), Lehmann (1975), and Woolson and Lachenbruch (1980). In the presence of censoring, define $Z_i = Y_{2i} - Y_{1i}$ and leg ε_i be the sign of Z_i. The censoring mechanism implies that $\varepsilon_i = 0$ and $Z_i = 0$ whenever $\delta_{1i} = \delta_{2i} = 0$, $\varepsilon_i = 1$ whenever $\delta_{1i} = 1$ and $\delta_{2i} = 0$, and $\varepsilon_i = -1$ whenever $\delta_{1i} = 0$ and $\delta_{2i} = 1$. Moreover, if the underlying failure times have a continuous distribution, then for each uncensored pair $\varepsilon_i = 1$ or $\varepsilon_i = -1$ with probability 1. Define sets $B_1 = \{i : \varepsilon_i = 1, \delta_{1i}\delta_{2i} = 1\}$, $B_2 = \{i : \varepsilon_i = -1, \delta_{1i}\delta_{2i} = 1\}$, $B_3 = \{i : \delta_{1i} = 1, \delta_{2i} = 0\}$, and $B_4 = \{i : \delta_{1i} = 0, \delta_{2i} = 1\}$. Finally, for $j = 1, \ldots, 4$ let $N_j(t) = \sum_{i=1}^n N_{ji}(t)$, where $N_{ji}(t) = I[|Z_i| \le t, i \in B_j]$, be processes counting occurrences of uncensored and singly censored pairs $(|Z_i|, \varepsilon_i)$.

To test the hypothesis of bivariate symmetry we consider statistics

$$T = \int K_u d(N_1 - N_2) + \int K_c d(N_3 - N_4), \quad (1.1)$$

where K_u and K_c are some scoring processes. Special cases include (a) the sign test, $K_u = K_c = 1$; (b) the signed Wilcoxon test, $K_u = 1 - \hat{F}_-$ and $K_c = 1 - \hat{F}_-/2$; and (c) the signed-normal scores test, $K_u = \Phi^{-1}(1 - \hat{F}_-/2)$ and $K_c = 2\hat{F}_-^{-1}\phi\{\Phi^{-1}(1 - \hat{F}_-/2)\}$, where ϕ and Φ are the density and the distribution function of the standard normal distribution. Here $\hat{F}_-$ is the left-continuous version of the product integral

$$\hat{F}(t) = \prod_{s \le t} \{1 - \Delta\hat{\Lambda}(s)\} \quad (1.2)$$

with $\hat{\Lambda}(t) = \int_0^t U^{-1} I(U > 0) d(N_1 + N_2)$, where $U(t) = \sum I[|Z_i| \ge t, \varepsilon_i = \pm 1]$. In the absence of censoring, $\hat{\Lambda}(t)$ is the Aalen–Nelson estimator of the cumulative hazard function $|X_{2i} - X_{1i}|$ and $\hat{F}(t)$ is the corresponding empirical survival function. Further interpretation of these statistics is given in Section 2.

In general, assume that $K_u = J_u(1 - \hat{F}_-)$ and $K_c = J_c(1 - \hat{F}_-)$, where the score-generating functions J_u and J_c satisfy the relationship

$$J_u(v) = -\{(1 - v)J_c(v)\}'. \quad (1.3)$$

This choice of the scoring functions is motivated by the

* Dorota M. Dabrowska is Assistant Professor, Department of Biostatistics, School of Public Health, University of California, Los Angeles, CA 90024-1772. She thanks the associate editor and referees for their comments, especially a referee who pointed out mistakes in the original version of the article. This research was partially supported by the University of California Presidential Fellowship and National Science Foundation Grant DMS-8901603.

Journal of the American Statistical Association
June 1990, Vol. 85, No. 410, Theory and Methods

censored-data signed-rank statistics considered by Woolson and Lachenbruch (1980), who discussed these tests in the case of log-linear models. More precisely, let $\log X_{1i} = \theta + \eta_{1i} + \varepsilon_i$ and $\log X_{2i} = \eta_{2i} + \varepsilon_i$, where $\{\eta_{1i}\}_{i=1}^n$ and $\{\eta_{2i}\}_{i=1}^n$ are mutually independent samples from a distribution with density ϕ and $\{\varepsilon_i\}_{i=1}^n$ is a sample independent of η_{1i}'s and η_{2i}'s. Woolson and Lachenbruch's signed-rank test for testing $\theta = 0$ against $\theta > 0$ is based on statistic T with score function $J_u(v) = -\phi'(z)/\phi(z)$ and $J_c(v) = 2\phi(z)/(1 - v)$, where $z = \Phi^{-1}(1/2 + v/2)$ and Φ is the distribution function corresponding to ϕ. These scores arise as the scores of the locally most powerful test derived from the marginal likelihood of signed ranks that is appropriate in the uncensored experiment given the observed pattern of deaths and withdrawals. The sign, signed Wilcoxon, and signed-normal scores tests correspond to double-exponential, logistic, and normal-densities ϕ, respectively.

In this article, I consider the asymptotic distribution of the signed-rank statistics (1.1). In Section 2 I derive their asymptotic distributions under the null hypothesis of bivariate symmetry and under contiguous alternatives. The signed-rank tests are in general inefficient within the class of tests based on the absolute differences Z_i, their signs, and censoring indicators. A test that is efficient within this class of tests assigns to uncensored and singly censored observations scores that depend on the joint distribution of the failure times and the censoring distribution. Moreover, depending on the form of the joint distribution of the failure times, such a test may assign a nonzero score to doubly censored observations. Thus the inefficiency of the signed-rank tests can be attributed to the inappropriate choice of scores assigned to uncensored and singly censored observations and omission of doubly censored observation. This is studied in more detail in Section 3, where I consider a log-linear model and examine the asymptotic relative efficiency of the signed-rank tests with respect to efficient parametric tests based on $(|Z_i|, \varepsilon_i, \delta_{1i}, \delta_{2i})$.

Further, I briefly discuss paired-rank tests such as the Prentice–Wilcoxon and log-rank tests (Albers 1988; Dabrowska 1989; O'Brien and Fleming 1987). These tests have the same form as commonly used rank tests for two-sample comparisons except that the variances of the test statistics are modified so as to take into account the intrapair dependence. For both uncensored and censored data the performance of these tests relative to signed-rank tests strongly hinges on the joint distribution of the underlying failure times. Using a log-linear model, it is shown that these tests may perform both better and worse than signed-rank tests. Moreover, they may have a relatively low efficiency with respect to optimal parametric tests when censoring is heavy and doubly censored observations have a nontrivial contribution to the log-likelihood expansion corresponding to $(|Z_i|, \varepsilon_i, \delta_{1i}, \delta_{2i})$.

Finally, in Section 4 I apply the signed-rank and paired-rank tests to the leukemia remission data of Freireich et al. (1963).

2. ASYMPTOTIC DISTRIBUTION OF THE TEST STATISTICS

2.1 Preliminaries

Throughout we assume that the joint distribution of (X_{1i}, X_{2i}) has density $\psi(x, y)$ and the censoring times C_i have distribution function G and survival function $\overline{G} = 1 - G$. Introduce subsurvival functions $\overline{F}_j(t) = \Pr(|Z_i| \geq t, i \in B_j)$, and let $f_j(t) = -\overline{F}_j(t)'$ be the corresponding improper densities. These are explicitly given by

$$f_1(t) = \int \overline{G}(u)\psi(u - t, u)\, du$$

$$f_2(t) = \int \overline{G}(u)\psi(u, u - t)\, du$$

$$f_3(t) = \int \left\{\int_u^\infty \psi(u - t, y)\, dy\right\} dG(u)$$

$$f_4(t) = \int \left\{\int_u^\infty \psi(x, u - t)\, dx\right\} dG(u). \tag{2.1}$$

Note that $\Sigma\, \overline{F}_j(0) = 1 - p_0 \leq 1$, where

$$p_0 = \Pr(\delta_{1i} = 0, \delta_{2i} = 0)$$
$$= \int \left\{\int_u^\infty \int_u^\infty \psi(x, y)\, dx\, dy\right\} dG(u) \tag{2.2}$$

is the probability of a doubly censored observation.

The functions $f_j(t)$ and $\overline{F}_j(t)$ are related to the intensities of the counting process $N = [\{N_{ji}(t) : j = 1, \ldots, 4; i = 1, \ldots, n\} : 0 < t < \infty]$. Each of the component processes has jumps of size 1, and no two processes jump at the same time. The behavior of the process N is determined by its intensity $\alpha(t) = [\{\alpha_{ji}(t) : j = 1, \ldots, 4; i = 1, \ldots, n\} : 0 < t < \infty]$, where $\alpha_{ji}(t)\, dt = \Pr\{dN_{ji}(t) = 1 \mid \mathfrak{F}_{t-}\}\; j = 1, 2$. Here $dN_{ji}(t)$ denotes the increment of N_{ji} over the interval $[t, t + dt]$, whereas $\{\mathfrak{F}_t\}$ is the self-exciting filtration generated by the null sets and processes $N_{ji}(t)$ $(j = 1, \ldots, 4; i = 1, \ldots, n)$. Thus $\alpha_{ji}(t)\, dt$ is the conditional probability that N_{ji} jumps in an infinitesimal interval of length dt around time t given the history $\mathfrak{F}_{t-}$. It can be easily verified that in our case $\alpha_{ji}(t) = \lambda_j(t)I[|Z_i| \geq t, i \in B_j]$, where $\lambda_j(t) = f_j(t)/\overline{F}_j(t)$.

I can also provide an interpretation of $\hat{\Lambda}(t)$ and $\hat{F}(t)$ in terms of counting processes. For this purpose consider the process $N_0(t) = [\{N_{1i}(t) + N_{2i}(t) : i = 1, \ldots, n\} : 0 < t < \infty]$ counting occurrences of uncensored $|Z_i|$'s. An easy calculation shows that its intensity $\alpha_0(t) = [\{\alpha_{0i}(t); i = 1, \ldots, n\} : 0 < t < \infty]$ is given by $\alpha_{0i}(t) = I[|Z_i| \geq t, \varepsilon_i = \pm 1]\lambda(t)$, where $\lambda = (f_1 + f_2)/\overline{H}$ and $\overline{H}(t) = \sum_{j=1}^4 \overline{F}_j(t) = \Pr(|Z_i| \geq t, \varepsilon_i = \pm 1)$. We have explicitly

$$\overline{H}(t) = \int \overline{G}(u) \left\{\int_u^\infty [\psi(u - t, y) + \psi(y, u - t)]\, dy\right\} du.$$

The process $\hat{\Lambda}(t)$ can be thought of now as the Aalen–

Nelson estimate of the integrated hazard function $\Lambda(t) = \int_0^t \lambda(s)\, ds$, whereas $\hat{F}(t)$ is the Kaplan–Meier (1958) estimate of the product integral

$$\overline{F}(t) = \prod_{s \le t} [1 - \Lambda(ds)] \tag{2.3}$$

associated with $\Lambda(t)$. In the absence of censoring $\overline{F}(t)$ reduces to the survival function of the differences $|X_{2i} - X_{1i}|$ and $\hat{F}(t)$ is the corresponding empirical survival function.

2.2 Asymptotic Normality of the Test Statistics

Consider first the null hypothesis of bivariate symmetry, and assume that the joint distribution of the failure times (X_{1i}, X_{2i}) has a density $\psi(x, y)$ such that $\psi(x, y) = \psi(y, x)$ for all x and y. Clearly, in this case the densities (2.1) satisfy $f_1 = f_2$ and $f_3 = f_4$.

Proposition 2.1. Suppose that the score-generating functions J_u and J_c are continuous and $|J_u(v)| \le a(1 - v)^{-1/2+\delta}$ and $|J_c(v)| \le a(1 - v)^{-1/2+\delta}$ for some constants $a > 0$ and $\delta > 0$. (a) Under the hypothesis of bivariate symmetry, $n^{-1/2}T$ converges weakly to a mean-zero normal distribution with variance

$$\sigma_T^2 = 2\left\{\int J_u^2(1 - \overline{F})f_1\, ds + \int J_c^2(1 - \overline{F})f_3\, ds\right\}. \tag{2.4}$$

(b) A consistent estimate of the asymptotic null variance is given by

$$\hat{\sigma}_T^2 = \sum_{i=1}^{2} \int K_u^2\, dN_i + \sum_{i=3}^{4} \int K_c^2\, dN_i.$$

The proof is in the Appendix. If the score-generating functions are chosen as $J_u(v) = -\phi'(z)/\phi(z)$ and $J_c(v) = 2\phi(z)/(1 - v)$, where $z = \Phi^{-1}(1/2 + v/2)$, ϕ is a symmetric density, and Φ is the corresponding distribution function, then the growth rate conditions assumed in Proposition 2.1 hold for most ϕ's arising in practice. In particular, they are satisfied by logistic, double-exponential ϕ and normal, so under the null hypothesis, the signed Wilcoxon, sign, and signed-normal scores test statistics are asymptotically mean-zero normal with asymptotic variances given by

$$\sigma_W^2 = 2\left\{\int (1 - \overline{F})^2 f_1\, ds + \int (1 - \overline{F}/2)^2 f_3\, ds\right\}$$

$$\sigma_S^2 = 2\left\{\int f_1\, ds + \int f_3\, ds\right\}$$
$$= \Pr(\varepsilon_i = 1) + \Pr(\varepsilon_i = -1)$$

$$\sigma_N^2 = 2\left\{\int w_1^2(\overline{F})f_1\, ds + \int w_2^2(\overline{F})f_3\, ds\right\}. \tag{2.5}$$

Here $w_1(s) = \Phi^{-1}(1 - s/2)$, $w_2(s) = 2s^{-1}\phi\{\Phi^{-1}(1 - s/2)\}$, and ϕ and Φ denote the density and the distribution function of the standard normal distribution. The form of these asymptotic variances was given by Woolson and Lachenbruch (1980).

To derive efficacies of tests based on statistics $n^{-1/2}T$ consider now contiguous alternatives of the form $\psi_n(x, y) = \psi(x, y)\{1 + n^{-1/2}\gamma_n(x, y)\}$, where γ_n is a sequence such that $\gamma_n(x, y) \to \gamma(x, y)$ for almost all (x, y), γ_n and γ are asymmetric functions, and

$$\int \gamma_n(x, y)\psi(x, y)\, dx\, dy = \int \gamma(x, y)\psi(x, y)\, dx\, dy = 0.$$

The last condition ensures that ψ_n is a density. In the case of parametric models, if $\psi_{\theta_0}(x, y)$ is a symmetric density and the alternatives are $\psi_{\theta_n}(x, y)$ with $\theta_n = \theta_0 + cn^{-1/2}$, the function γ reduces to c times the derivative of log $\psi_\theta(x, y)$ at $\theta = \theta_0$.

Let P and P_n denote the joint distributions of $(|Z_i|, \varepsilon_i, \delta_{1i}, \delta_{2i})$ under the null hypothesis and under the alternative, respectively. Then

$$\log dP_n/dP = \sum_{j=1}^{4} \int \log(f_{jn}/f_j)\, dN_j + \log(p_{0_n}/p_0) \sum_{i=1}^{n} \eta_i,$$

where $\eta_i = (1 - \delta_{1i})(1 - \delta_{2i})$. Here p_{0n} and the densities f_{jn} are defined by (2.2) and (2.1), respectively, with ψ replaced by ψ_n.

I shall derive the asymptotic joint distribution of log dP_n/dP and $n^{1/2}T$ under the null hypothesis and use Le Cam's third lemma to obtain efficacies of the test statistics. Set

$$A_0 = p_0^{-1} \int \left\{\int_u^\infty \int_u^\infty \gamma(x, y)\psi(x, y)\, dx\, dy\right\} dG\,(u),$$

$$A_1(t) = f_1(t)^{-1} \int \overline{G}(u)\gamma(u - t, u)\psi(u - t, u)\, du,$$

and

$$A_3(t) = f_3(t)^{-1} \int \left\{\int_u^\infty \gamma(u - t, y) \times \psi(u - t, y)\, dy\right\} dG\,(u),$$

and let A_2 and A_4 be defined similarly except that $\gamma(u - t, u)$ and $\gamma(u - t, y)$ are replaced by $\gamma(u, u - t)$ and $\gamma(y, u - t)$, respectively. In the case of parametric families $\psi_{\theta_n}(x, y)$, if $\theta_n = \theta_0 + cn^{-1/2}$ and θ_0 corresponds to the hypothesis of symmetry, then A_0 and A_j $(j = 1, \ldots, 4)$ correspond to the usual scores and are given by c times the derivatives of log $p_{0\theta}$ and log $f_{j\theta}$ at $\theta = \theta_0$. We need the following condition.

Condition A. As $n \to \infty$, $A_{0n} = 2n^{1/2}[(p_{0n}/p_0)^{1/2} - 1] \to A_0$ and for $j = 1, \ldots, 4$, $\int \{A_{jn} - A_j\}^2 f_j\, ds \to 0$, where $A_{jn} = 2n^{1/2}[(f_{jn}/f_j)^{1/2} - 1]$.

Proposition 2.2. Suppose that the assumptions of Proposition 2.1 and Condition A are satisfied. Then under the null hypothesis, (log dP_n/dP, $n^{-1/2}T$) converge weakly to a normal distribution with mean $(-1/2\sigma_0^2, 0)$ and covariance matrix

$$\begin{pmatrix} \sigma_0^2 & c_T \\ c_T & \sigma_T^2 \end{pmatrix},$$

where σ_T^2 is given by (2.4) and

$$\sigma_0^2 = \int (A_1^2 + A_2^2) f_1 \, ds + \int (A_3^2 + A_4^2) f_3 \, ds + A_0^2 p_0$$

$$c_T = \int J_u(1 - \overline{F})(A_1 - A_2) f_1 \, ds + \int J_c(1 - \overline{F})(A_3 - A_4) f_3 \, ds. \quad (2.6)$$

As a consequence of Le Cam's first lemma (Hájek and Šidák 1967) I conclude that the family of distributions P_n is contiguous to P. Moreover, under the alternatives $\psi_n(x, y)$, the statistic $n^{-1/2}T$ converges weakly to a normal distribution with mean c_T and variance σ_T^2. It follows that the efficacy of the test statistic T is given by c_T^2/σ_T^2. In particular, the efficacies of the sign, signed Wilcoxon, and signed-normal scores tests are given by

$$e_S = \left\{ \int (A_1 - A_2) f_1 \, ds + \int (A_3 - A_4) f_3 \, ds \right\}^2 \Big/ \sigma_S^2,$$

$$e_W = \left\{ \int (A_1 - A_2)(1 - \overline{F}) f_1 \, ds + \int (A_3 - A_4)(1 - \overline{F}/2) f_3 \, ds \right\}^2 \Big/ \sigma_W^2,$$

and

$$e_N = \left\{ \int (A_1 - A_2) w_1(\overline{F}) f_1 \, ds + \int (A_3 - A_4) w_2(\overline{F}) f_3 \, ds \right\}^2 \Big/ \sigma_N^2,$$

where w_1, w_2, σ_W^2, and σ_N^2 are as in (2.5).

From Proposition 2.2 it follows immediately that the signed-rank tests are in general inefficient within the class of tests based on $(|Z_i|, \varepsilon_i, \delta_{1i}, \delta_{2i})$. This is in contrast with tests for two-sample comparisons under the equal-censorship model (Gill 1980; Harrington and Fleming 1982). If the density of the paired survival times (X_1, X_2) belongs to a parametric family $\psi_\theta(x, y)$ with $\theta = \theta_0$ corresponding to symmetry, and if the censoring distribution is known (e.g., in the case of fixed censoring), tests for symmetry can be based on the likelihood ratio statistic. The efficiency of the resulting test is σ_0^2. If the censoring distribution is unknown, tests have to be constructed adaptively using methods appropriate for semiparametric models.

3. DISCUSSION AND SOME COMPARISONS

3.1 Signed-Rank Versus Parametric Tests

I consider now the asymptotic relative efficiency (ARE) of the signed-rank tests with respect to the asymptotically optimal parametric test within the class of tests based on $(|Z_i|, \varepsilon_i, \delta_{1i}, \delta_{2i})$.

If the data are uncensored, then $A_3 = A_4 = 0$ and the functions f_1 and f_2 reduce to the improper densities corresponding to $\Pr(|Z_i| \geq t, \varepsilon_i = 1)$ and $\Pr(|Z_i| \geq t, \varepsilon_i = -1)$, respectively, where $Z_i = X_{2i} - X_{1i}$. If $A_1 = -A_2$ then the signed-rank test (1.1) associated with A_1 is fully efficient within the class of tests based on the absolute differences Z_i and their signs. This will occur whenever the distribution of Z_i has density $\phi(x - \theta)$, where $\phi(x)$ is symmetric about origin. The efficient signed-rank test is then given by (1.1) with $J_u(v) = A_1[\Phi^{-1}(1/2 + v/2)]$, where Φ is the distribution function corresponding to ϕ and $A_1(z) = -\phi'(z)/\phi(z)$. If $A_1 \neq -A_2$, an efficient signed-rank test can be easily constructed by assigning scores $A_1\{\Phi^{-1}[1 - \hat{F}(|Z_i|)/2]\}$ and $A_2\{\Phi^{-1}[1 - \hat{F}(|Z_i|)/2]\}$ to positive and negative Z_i's, respectively. Here $\hat{F}$ is the empirical distribution function of $|Z_i|$'s given by (1.2). Doksum's (1980) signed log-rank test is an example. Asymptotic efficiency of such tests follows from a minor modification of Proposition 2.2.

Whereas in the uncensored case the distribution of the test statistics depends only on the underlying distribution of the differences $X_{2i} - X_{1i}$, in the presence of censoring the finite sample and asymptotic distributions of these statistics depend both on the joint distribution of the failure times and the censoring distribution. The structure of the asymptotic distribution, however, is completely different from that of the log-likelihood expansion, and the efficiency loss is caused by the inappropriate form of the score functions assigned to uncensored and singly censored observations and omission of doubly censored pairs, if $A_0 \neq 0$. Similar inefficiency problems arise in the case of signed-rank tests with differences $X_{2i} - X_{1i}$ having asymmetric densities.

As an illustration let us consider the log-linear model $\log X_{1i} = \theta_1 + \eta_{1i} + \varepsilon_i$ and $\log X_{2i} = \theta_2 + \eta_{2i} + \varepsilon_i$, where $\{\eta_{1i}\}$ and $\{\eta_{2i}\}$ are mutually independent samples from the same distribution and $\{\varepsilon_i\}$ is a sample independent from $\{\eta_{1i}\}$ and $\{\eta_{2i}\}$. The variable ε_i represents the unknown matching effect common to both pair members. Consider two models. In Model 1 I set $\theta_1 = -\theta_2 = \theta/2$, whereas in Model 2 I let $\theta_1 = \theta$ and $\theta_2 = 0$.

In the uncensored case, signed-rank tests derived from the marginal likelihood of the absolute differences of log-failure times and their signs are the same for both models. In the presence of censoring the behavior of their censored-data analogs is, however, different for the two models. In particular, in Model 1 the score A_0 corresponding to doubly censored observations is 0, whereas in Model 2, $A_0 \neq 0$.

For numerical comparisons, η_{1i} and η_{2i} were chosen to have extreme-value distribution with survival function $\exp\{-e^x\}$ and ε_i was chosen to be uniform on the interval $(-a, a)$. The correlation between $\log X_{1i}$ and $\log X_{2i}$ is equal to $\rho = a^2/(a^2 + \pi)$, so $\rho \to 1$ as $a \to \infty$. For $a = 0$, the log-failure times are independent, and in this case Models 1 and 2 reduce to two-sample and one-sample location models. Finally, it was assumed that the censoring variable has an extreme-value distribution with survival function $\exp\{-\lambda e^x\}$. The scale parameter λ determines the heaviness of censoring. In particular, under the null hypothesis of symmetry ($\theta = 0$) the probability of doubly censored observations is given by $p_0 = (2a)^{-1} \log[(2 + \lambda e^a)/(2 + \lambda e^{-a})]$ for $a \neq 0$ and $p_0 = \lambda/(\lambda + 2)$ for $a = 0$. The choice of uniformly distributed ε_i's is no doubt artificial and was made merely to simplify calculations. For distributions other than uniform, the behavior of the test statistics is similar.

The parameter a was chosen so that the correlation between the underlying failure times is $\rho = 0$, .1, .25, .5, .75, and .9. Further, for each of these ρ values the scale parameter λ of the censoring distribution was selected so that under the null hypothesis the probability of a doubly censored pair is equal to $p_0 = .05$, .1, .25, .5, .75, and .9. The calculation of the asymptotic lower bounds of Proposition 2.2 and of the efficacies of the signed-rank tests was carried out numerically using the Gauss–Kronrod rule (IMSL 1987, subroutines QDAG, QDAGI, and QDNG).

For uncensored data and both models, $\log X_{2i} - \log X_{1i}$ follows a logistic location model and the signed Wilcoxon test is fully efficient within the class of all tests based on the absolute differences of the log-failure times and their signs. Table 1 gives the ARE of the censored-data signed Wilcoxon test with respect to the optimal parametric tests based on $(|Z_i|, \varepsilon_i, \delta_{1i}, \delta_{2i})$. In the case of Model 1, the ARE of the signed Wilcoxon test is close to 1 and not much efficiency is lost by considering this test rather than the optimal parametric test. In the case of Model 2, the ARE is a decreasing function of p_0 for all ρ values and it is an increasing function of ρ for all p_0 values. When censoring is light ($p_0 = .05$) the efficiency loss ranges between 11% for $\rho = 0$ and 4.3% for $\rho = .9$. As p_0 increases to .9 the efficiency loss increases and ranges between 47.1% for $\rho = 0$ and 31.8% for $\rho = .9$.

Table 2 gives the ARE of the sign and signed-normal scores tests with respect to the signed Wilcoxon test. The ARE is the same for both models. The efficiency of both tests increases as censoring gets heavier. In particular, for

Table 1. ARE of the Signed Wilcoxon Test With Respect to the Optimal Parametric Test in Models 1 and 2

		ρ					
p_0	Model	0	.10	.25	.50	.75	.90
.05	1	.980	.981	.983	.986	.991	.996
	2	.890	.893	.899	.911	.933	.957
.10	1	.968	.971	.974	.982	.992	.998
	2	.819	.826	.839	.869	.913	.954
.25	1	.959	.963	.970	.982	.994	.999
	2	.702	.718	.745	.806	.891	.954
.50	1	.975	.977	.979	.983	.991	.997
	2	.617	.632	.660	.728	.840	.937
.75	1	.993	.993	.992	.990	.988	.990
	2	.559	.569	.588	.633	.722	.841
.90	1	1.000	1.000	.999	.997	.994	.990
	2	.529	.530	.539	.561	.608	.682

Table 2. ARE of the Sign (S) and Signed Normal Scores (N) Tests With Respect to the Signed Wilcoxon Test

		ρ					
p_0	Test	0	.10	.25	.50	.75	.90
.05	S	.751	.751	.751	.752	.754	.755
	N	.971	.971	.970	.968	.956	.961
.10	S	.753	.753	.755	.757	.760	.762
	N	.983	.981	.979	.975	.967	.961
.25	S	.767	.769	.771	.776	.779	.779
	N	1.001	.996	.992	.982	.970	.961
.50	S	.814	.814	.813	.810	.802	.787
	N	1.005	1.003	1.001	.994	.978	.963
.75	S	.892	.887	.879	.861	.835	.811
	N	1.003	1.001	1.000	.999	.996	.984
.90	S	.952	.948	.941	.925	.895	.859
	N	1.001	1.000	1.000	1.000	.999	.999

the sign test efficiency gain is approximately between 20.1% for $\rho = 0$ and 10.4% for $\rho = .90$ as p_0 increases from 0 to .90. For the signed normal scores test the gain in efficiency is about 5%. For some choices of p_0 and ρ the signed-normal scores test is slightly more efficient than the signed Wilcoxon test, which can be explained by recalling that the signed Wilcoxon test is a "locally most powerful" signed-rank test only in the uncensored version of the experiment and loses this property when the joint distribution of the failure times and the censoring distribution are taken into account.

3.2 Signed-Rank Versus Paired-Rank Tests

For uncensored data a thorough treatment of the paired-rank tests was given by Snijders (1981) in the context of conditional-rank tests. Essentially, the idea is to pool X_{1i}'s and X_{2i}'s and look for the locally most powerful rank test conditionally on the observed configuration of ranks. Given the observed ranks $\{r_{1i}, r_{2i}\}$ of (X_{1i}, X_{2i}), under the null hypothesis of bivariate symmetry the rank r_{1i} is equally likely to be the rank of X_{1i} and X_{2i}. Unfortunately, the scores of these tests are usually too complicated to evaluate. Evaluation of the scores of the paired-rank tests and derivation of their finite sample and asymptotic properties require considering nonlinear rank statistics. In spite of this problem, tests such as the paired Wilcoxon test (Lam and Longnecker 1983; Snijders 1981) or the paired log-rank test (Doksum 1980) have gained some popularity, since in many situations they can be more efficient than the signed-rank tests and parametric tests derived from the likelihood of absolute differences $X_{2i} - X_{1i}$ and their signs, because the latter procedures do not use information or intrapair dependence.

Censored data analogs of paired-rank tests were developed by O'Brien and Fleming (1987), Albers (1988), and Dabrowska (1989), among others. Here we consider a Prentice-type method of ranking of the observations; that is, the paired data are pooled, uncensored observations are ranked among themselves, and each censored observation is assigned the same rank as the nearest uncensored observation on the left. For suitably chosen score functions

$J(u, d)$, with $u \in (0, 1)$ and $d = 1, 0$, the test statistics reject the hypothesis of symmetry for large values of $n^{1/2}W/\hat{\sigma}$, where

$$W = n^{-1}\left\{\sum_{i=1}^{n} J(\hat{S}(Y_{1i}), \delta_{1i}) - \sum_{i=1}^{n} J(\hat{S}(Y_{2i}), \delta_{2i})\right\} \quad (3.1)$$

and $\hat{\sigma}^2$ is an estimator of the asymptotic variance of W of the form

$$\hat{\sigma}^2 = n^{-1}\left\{\sum_{j=1}^{2}\sum_{i=1}^{n} J^2(\hat{S}(Y_{ji}), \delta_{ji}) - 2\sum_{i=1}^{n} J(\hat{S}(Y_{1i}), \delta_{1i})J(\hat{S}(Y_{2i}), \delta_{2i})\right\}.$$

Here $\hat{S}(t)$ is the Kaplan–Meier estimate from the pooled sample or an estimator asymptotically equivalent to it; for example, $\hat{S}(t) = 1 - \exp\{-\hat{\Lambda}(t)\}$, where $\hat{\Lambda}(t)$ is the Aalen–Nelson estimate based on the pooled sample. The choice $J(u, 1) = 2u - 1$ and $J(u, 0) = u$ yields the Prentice–Wilcoxon text statistic, whereas the choice of $J(u, 1) = -1 - \log(1 - u)$ and $J(u, 0) = -\log(1 - u)$ leads to the paired log-rank test.

Asymptotic distributions of tests based on $n^{1/2}\hat{W}/\hat{\sigma}$ were derived in Dabrowska (1989) for arbitrary fixed and converging alternatives. If S is the common marginal distribution function of X_{1i} and X_{2i} under the null hypothesis, the efficacy of tests based on the statistic (3.1) is given by c_W^2/σ_W^2, where

$$c_W = \int \{J[S(x), 1] - J[S(x), 0]\}\Gamma(x)\overline{G}(x)\, dS(x)$$

and

$$\sigma_W^2 = 2\int \{J[S(x), 1] - J[S(x), 0]\}^2\overline{G}(x)\, dS(x) - 2EJ[S(Y_{1i}), \delta_{1i}]J[S(Y_{2i}), \delta_{2i}],$$

provided that $J(u, 1) = -\{(1 - u)J(u, 0)\}'$. In the notation of Proposition 2.2, the function $\Gamma(x)$ is given by

$$\Gamma(x) = \gamma_0(x) \Big/ \int \psi(x, y)\, dy - \int_x^\infty \gamma_0(t)\, dt \Big/ [1 - S(x)]$$

and

$$\gamma_0(x) = \int [\gamma(x, y) - \gamma(y, x)]\psi(x, y)\, dy.$$

Note that the expectation in the second term of σ_W^2 is equal to the asymptotic covariance between $J[S(Y_{1i}), \delta_{1i}]$ and $J[S(Y_{2i}), \delta_{2i}]$ and accounts for the possible intrapair dependence between the underlying failure times. On the other hand, the mean c_W is the same as in the case of two-sample comparisons under the equal-censorship model and depends only on the marginal distributions of the failure times and the censoring distribution.

For both uncensored and censored data, the performance of paired-rank tests relative to optimal parametric tests and signed-rank tests depends heavily on the structure of the joint distribution of the failure times. Table 3 gives the ARE of the paired log-rank and paired Wilcoxon tests with respect to the signed Wilcoxon test in the log-linear model of Section 3.1.

For uncensored data ($p_0 = 0$) both paired-rank tests lose efficiency when the correlation between the paired failure times increases. This is especially pronounced in the case of the log-rank test, where the efficiency loss is approximately 81.6% when ρ increases from 0 to .90. The efficiency loss for the paired Wilcoxon test is only 21.9% for the same range of ρ values, and as ρ increases this test becomes more efficient than the log-rank test. This pattern is present also in the case of censored data, though the amount of efficiency loss decreases for both tests as censoring gets heavier. Moreover, for heavily censored data the ARE approaches 1, so both paired tests are asymptotically as efficient as the signed Wilcoxon test.

The ARE of the paired-rank tests with respect to the optimal parametric test based on $(|Z_i|, \varepsilon_i, \delta_{1i}, \delta_{2i})$ can be obtained by multiplying entries of Tables 1 and 3. Similar to the signed-rank tests, the efficiency loss can be attributed to the omission of doubly censored observations in the test statistics (3.1). Note that in the presence of univariate censoring, each member of a doubly censored pair is assigned the same rank and score; consequently, such pairs do not contribute to (3.1).

The log-rank test is fully efficient only in the case of Model 1 with $\rho = 0$. For $\rho > 0$ or Model 2 the scores of the fully efficient tests depend on the joint distribution of the survival times and their rank counterpart is a nonlinear rank statistic. Similar to the optimal parametric test based on $(|Z_i|, \varepsilon_i, \delta_{1i}, \delta_{2i})$, the fully efficient test assigns score 0 to doubly censored observations in Model 1. In Model 2, however, this score is nontrivial. Numerical integration shows that the ARE of the paired-rank tests with respect to the fully efficient test is in the range 50%–63.7% for the log-rank test and 37.5%–79.5% for the Wilcoxon test

Table 3. ARE of the Paired Log-Rank (L) and Paired Wilcoxon Test (W) With Respect to the Signed Wilcoxon Test

		ρ					
p_0	Test	0	.10	.25	.50	.75	.90
0	L	1.500	1.375	1.241	1.061	.869	.684
	W	1.125	1.096	1.053	.906	.929	.906
.05	L	1.430	1.324	1.203	1.031	.841	.660
	W	1.107	1.084	1.048	.990	.939	.919
.10	L	1.369	1.281	1.175	1.014	.829	.651
	W	1.091	1.073	1.045	.996	.949	.930
.25	L	1.227	1.182	1.118	1.000	.827	.639
	W	1.054	1.049	1.040	1.015	.971	.928
.50	L	1.085	1.078	1.063	1.014	.881	.668
	W	1.018	1.022	1.028	1.028	.984	.874
.75	L	1.019	1.020	1.021	1.017	.980	.839
	W	1.003	1.006	1.010	1.016	1.006	.913
.90	L	1.002	1.003	1.003	1.005	1.004	.983
	W	1.000	1.000	1.001	1.004	1.006	.992

Control:	1	22	3	2	8	17	2	11	8	12	2	5	4	15	8	23	5	11	4	1	8
Treatment:	10	7	22*	23	22	6	16	34*	32*	25*	11*	20*	19*	6	17*	35*	6	13	9*	6*	10*

Figure 1. Paired Remission Times for Leukemia Patients. Censored observations are indicated by asterisks.

in the case of Model 2, whereas in Model 1 the range is 62.1%–100% and 75%–99.7%, respectively. In both models the Wilcoxon test has higher efficiency under stronger dependence and/or heavier censoring. Finally, the ARE of the optimal parametric test based on $(|Z_i|, \varepsilon_i, \delta_{1i}, \delta_{2i})$ with respect to the fully efficient test is in the range 37.5%–95.4% in the case of Model 2 and 75%–99.8% in the case of Model 1.

4. AN EXAMPLE

To conclude, we consider the leukemia remission data of Freireich et al. (1963). This data set was analyzed many times, primarily as a two-sample problem without reference to the original matched-pair setting. Figure 1 gives remission or withdrawal times in weeks, as given by Lachenbruch, Palta, and Woolson (1982).

The standardized log-rank and Spearman rank correlation statistics for testing independence (Cuzick 1982; Dabrowska 1986) are $-.602$ and $-.680$, respectively. Thus it seems that matching does not introduce association between the paired failure times. Oakes (1982) reported similar results based on his modified Kendall's tau.

The standardized sign, signed Wilcoxon, and normal-scores statistics are approximately -3.273, -3.389, and -3.426, respectively, and, using normal approximation, the corresponding two-sided tests have p values of .001, .0006, and .0006. This is consistent with earlier findings of Lachenbruch et al. (1982) for the sign and signed Wilcoxon tests. The standardized paired log-rank and paired Wilcoxon statistics are equal to -3.252 and -3.071, respectively, so the p values of the associated tests are .0012 and .002. By using any of these tests we can reject the hypothesis of symmetry and equal marginal distributions for the paired remission times.

APPENDIX: PROOFS

To prove Propositions 2.1 and 2.2, consider first statistics $n^{-1/2}T_1$ and $\bar{\sigma}_T^2$, where

$$T_1 = \int J_u(1 - F)d(N_1 - N_2) + \int J_c(1 - F)d(N_3 - N_4)$$

and

$$\bar{\sigma}_T^2 = n^{-1} \sum_{i=1}^{2} \int J_u^2(1 - F)\, dN_i + n^{-1} \sum_{i=3}^{4} \int J_c^2(1 - F)\, dN_i.$$

Under the assumed growth-rate conditions on the score functions J_u and J_c, T_1 is a sum of iid mean-zero random variables with variance σ_T^2 $(\sigma_T^2 < \infty)$, so the asymptotic normality of $n^{-1/2}T_1$ follows from the central limit theorem. Similarly, $\hat{\sigma}_T^2$ is a sum of iid random variables with mean σ_T^2 and its consistency follows from the law of large numbers.

To show the joint asymptotic normality of log dP_n/dP_0 and $n^{-1/2}T_1$, note that by Le Cam's second lemma (Hájek and Šidák 1967) it is enough to consider the asymptotic joint distribution of L_n and $n^{-1/2}T_1$, where

$$L_n = n^{-1/2} \sum_{j=1}^{4} \int A_{jn}\, dN_j + n^{-1/2} A_{0n} \sum_{i=1}^{n} \eta_i.$$

Under the null hypothesis,

$$S = n^{-1/2} \sum_{j=1}^{4} \int A_j\, dN_j + n^{-1/2} A_0 \sum_{i=1}^{n} \eta_i$$

is a sum of iid mean-zero random variables with variance σ_0^2. Furthermore, Condition A and a little algebra yield

$$EL_n = -1/4 \sum_{j=1}^{4} \int A_{jn}^2 f_j\, dt - 1/4 A_{0n}^2 p_0 \to -1/4 \sigma_0^2$$

and

$$\operatorname{var}(L_n - S) \le \sum_{j=1}^{4} \int (A_{jn} - A_j)^2 f_j\, dt + (A_{0n} - A_0)^2 p_0 \to 0.$$

It follows that under the null hypothesis L_n is asymptotically normal with mean $-\sigma_0^2/4$ and variance σ_0^2. Le Cam's second lemma completes the proof of the asymptotic normality of the log-likelihood log dP_n/dP. Proposition 2.2 follows then after application of the Cramer–Wold device to $n^{-1/2}T_1$ and S.

It remains to show that $n^{-1/2}(T - T_1) = 0_P(1)$ and $\hat{\sigma}_T^2 - \bar{\sigma}_T^2 = 0_P(1)$. The proofs are analogous, so I consider the first of these statistics only.

Arguments similar to those of Gill (1980) and Gill and Johansen (1989) show that $\hat{F}(t)$ and $\hat{\Lambda}(t)$ converge in probability of $\bar{F}(t)$ and $\Lambda(t)$, respectively, uniformly in $t \in [0, \tau]$, where $\bar{H}(\tau_-) > 0$. For any such τ,

$$n^{-1/2} \int_0^\tau [J_u(1 - \hat{F}) - J_u(1 - F)]d(N_1 - N_2) \to_P 0.$$

Furthermore,

$$\lim_{\tau \uparrow \infty} \overline{\lim}\, \Pr\left[n^{-1/2} \int_\tau^\infty |J_u(1 - \hat{F}) - J_u(1 - F)|\, dN_j > \varepsilon \right] = 0 \tag{A.1}$$

for any $\varepsilon > 0$ and $j = 1, 2$. We have $\hat{F} \ge U/n$, and by theorem 1.1 of Van Zuijlen (1978), for given $\eta > 0$, $(U/n)^{-1/2+\delta} \le \beta \bar{H}^{-1/2+\delta}$ with probability at least $1 - \eta$ and $\beta = \eta^{-1/2+\delta}$. On the set where this holds,

$$n^{-1/2} \int_\tau^\infty |J_u(1 - \hat{F}) - J_u(1 - F)|\, dN_j \le c n^{-1/2} \int_\tau^\infty \bar{H}^{-1/2+\delta}\, dN_j,$$

where $c = a(1 + \beta)$. (A.1) follows now immediately from Markov's inequality applied to the square of this bound. The terms involving the score $J_c(v)$ can be treated analogously.

[Received June 1988. Revised November 1989.]

REFERENCES

Albers, W. (1988), "Combined Rank Tests for Randomly Censored Paired Data," *Journal of the American Statistical Association*, 83, 1159–1162.

Batchelor, J. R., and Hackett, M. (1970), "HL-A Matching in Treatment of Burned Patients With Skin Allografts," *Lancet*, 2, 581–583.

Cuzick, J. (1982), "Rank Tests for Association With Right Censored Data," *Biometrika*, 69, 351–364.

Dabrowska, D. M. (1986), "Rank Tests for Independence for Bivariate Censored Data," *The Annals of Statistics*, 14, 250–264.

——— (1989), "Rank Tests for Matched Pair Experiments With Censored Data," *Journal of Multivariate Analysis*, 28, 88–114.

Doksum, K. A. (1980), "Rank Tests for the Matched Pair Problem With Life Distributions," *Scandinavian Journal of Statistics*, 7, 67–72.

Freireich, E. J., et al. (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.

Gill, R. D. (1980), *Censoring and Stochastic Integrals* (Tract 124), Amsterdam: Mathematical Centre.

Gill, R. D., and Johansen, S. (1989), "Product-Integrals and Counting Processes," Technical report, Centre for Mathematics and Computer Science, Amsterdam.

Hájek, J., and Šidák, Z. (1967), *The Theory of Rank Tests*, New York: Academic Press.

Harrington, D. P., and Fleming, T. R. (1982), "A Class of Rank Test Procedures for Censored Survival Data," *Biometrika*, 69, 553–566.

Holt, J. D., and Prentice, R. L. (1974), "Survival Analyses in Twin Studies and Matched Pair Experiments," *Biometrika*, 61, 17–30.

Kaplan, E. L., and Meier, P. (1958), "Nonparametric Estimation From Incomplete Observations," *Journal of the American Statistical Association*, 53, 457–481.

Lachenbruch, P. A., Palta, M., and Woolson, R. F. (1982), "Analysis of Matched Pairs Studies With Censored Data," *Communications in Statistics*, 11, 549–569.

Lam, F. C., and Longnecker, M. T. (1983), "A Modified Wilcoxon Rank Sum Test for Paired Data," *Biometrika*, 70, 510–513.

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

Oakes, D. (1982), "A Concordance Test for Independence in the Presence of Censoring," *Biometrics*, 38, 451–455.

O'Brien, P. C., and Fleming, T. R. (1987), "A Paired Prentice–Wilcoxon Test for Censored Paired Data," *Biometrics*, 43, 169–180.

Snijders, T. (1981), "Rank Tests for Bivariate Symmetry," *The Annals of Statistics*, 9, 1087–1095.

Woolson, R. F., and Lachenbruch, P. A. (1980), "Rank Tests for Censored Matched Pairs," *Biometrika*, 67, 597–606.

Van Zuijlen, M. C. A. (1978), "Properties of the Empirical Distribution Function for Independent Nonidentically Distributed Random Variables," *The Annals of Probability*, 6, 250–266.