A k-Sample Median Test for Censored Data
Author(s): Ron Brookmeyer and John Crowley
Source: *Journal of the American Statistical Association*, Vol. 77, No. 378 (Jun., 1982), pp. 433-440
Published by: American Statistical Association
Stable URL: http://www.jstor.org/stable/2287264
Accessed: 13/08/2010 16:01

A *k*-Sample Median Test for Censored Data

RON BROOKMEYER and JOHN CROWLEY*

A *k*-sample median test for censored data is proposed. The test is more sensitive to differences in median survival times than to differences in shape of the survival curves. The asymptotic distribution of the test statistic is derived under both the null and the alternative hypotheses and its variance is shown to take a simple form. The median test is compared to several other tests for censored data by computing asymptotic relative efficiencies and performing a Monte Carlo simulation under various conditions. The procedure described is applied to data from a colorectal cancer clinical trial.

KEY WORDS: Censored data; Median test; Survival analysis.

1. INTRODUCTION

In medical follow-up studies with censored data, it has become increasingly common to cite point estimates for the median survival time based on the Kaplan-Meier estimate (1958) of the survival curve. The estimate, however, can be misleading because of its large variability. This is particularly true if the survival curve is flat near the median. Some recent work has addressed this problem by proposing confidence interval procedures for the median. Brookmeyer and Crowley (1982) and Emerson (1982) invert a sign test for censored data to obtain a confidence interval, while Efron (1981) and Reid (1981) suggest bootstrapping techniques.

If one is interested in comparing several treatments by examination of the medians, it is useful to have a formal test statistic. The classical median test (Westenberg 1948) is appropriate with uncensored data; a generalization of the statistic is required for censored data. Prentice (1978) proposed a statistic for testing the equality of k survival distributions with censored data that would be particularly powerful against location shifts in the double exponential distribution; it is in this situation that the median test is the locally most powerful rank test with uncensored data. Prentice's approach consists of evaluating the likelihood of a generalized rank vector followed by differentiation of the log-likelihood to obtain a test statistic. However, the statistic is difficult to calculate in the situation of different censoring distributions for the samples where a permutation variance is inappropriate. Gill (1980) proposed a similar statistic that depends on the pattern of observations only up to the median of the combined sample. However, the statistic does not reduce to the usual median test when there is no censoring.

A more intuitive approach for extending the k-sample median test to censored data is taken here. The test will be particularly useful if one is interested in detecting differences in the median survival times among several treatments. Of course, the appropriateness of any test statistic depends on the type of differences one is most interested in detecting. For example, if one is more concerned with detecting early differences in survival experience, either the Breslow-Gehan or the Peto-Prentice generalized Wilcoxon test (Gehan 1965; Breslow 1970; Peto and Peto 1972; Prentice 1978) is more appropriate; while the Kolmogorov-Smirnov type statistics (Fleming et al. 1980) may be particularly useful in the crossing-hazards situation.

2. A *k*-SAMPLE MEDIAN TEST FOR CENSORED DATA

Suppose n_i independent observations are drawn from the ith population, $1 \le i \le k$. The k-sample median test for uncensored data consists of pooling together the observations from the k samples; determining the median $\hat{M}$ of the pooled sample; then counting the number A_i in the ith sample that exceed $\hat{M}$. The random variables A_i are linear rank statistics and their large-sample behavior is well known (Hájek and Sidák 1967):

$$4 \sum_{i=1}^{k} \frac{(A_i - n_i/2)^2}{n_i} \xrightarrow{D} \chi^2(k-1).$$

In developing a median test for censored data, two questions arise:

1. How should the ambiguous observations be handled; that is, the censored observations that are less than the pooled-sample median?
2. How should the pooled-sample median be defined?

Let X_{ij} be the jth observed survival time from the ith population $1 \le j \le n_i$, $1 \le i \le k$. When observations are subject to arbitrary right censorship, the period of follow-up for the (i, j)th individual is restricted by the censoring time T_{ij}. The observed survival time is $X_{ij} = \min(X_{ij}{}^0, T_{ij})$ where $X_{ij}{}^0$ is the true but often unobserved survival time. One also observes δ_{ij}, which indicates whether X_{ij} is censored or not; thus, if $X_{ij} < X_{ij}{}^0$ the observation is

* Ron Brookmeyer is Assistant Professor, Department of Biostatistics, The Johns Hopkins University, Baltimore, MD 21205. John Crowley is Associate Professor, Department of Biostatistics, University of Washington, and Associate Member, Fred Hutchinson Cancer Research Center, Seattle, WA 98104. This research was partially supported by NIH grant 5-R01 CA 18332-06.

© Journal of the American Statistical Association
June 1982, Volume 77, Number 378
Theory and Methods Section

said to be censored and we set $\delta_{ij} = 0$. On the other hand, if $X_{ij} = X_{ij}{}^0$, it is an observed death and we set $\delta_{ij} = 1$.

The Kaplan-Meier estimate $\hat{S}_i{}^0(t) = 1 - \hat{F}_i{}^0(t)$, which is an estimate of $P(X_{ij}{}^0 > t)$, can be computed for each sample. We define a weighted Kaplan-Meier estimate by

$$\hat{F}_w{}^0(t) = 1 - \hat{S}_w{}^0(t) = \sum_{i=1}^{k} \lambda_i{}^N \hat{F}_i{}^0(t),$$

where $\lambda_i{}^N = n_i/N$ and $N = \sum_{i=1}^k n_i$, so that $\lambda_i{}^N$, is the relative sample size for the ith sample. Then the pooled-sample median is defined as

$$\hat{M} = \inf\{t\colon \hat{F}_w{}^0(t) \geq \tfrac{1}{2}\} = \hat{F}_w{}^{0-1}(\tfrac{1}{2}).$$

The quantity $\hat{F}_w{}^0(t)$ is a natural definition for a weighted Kaplan-Meier estimate, since it is estimating a quantity that involves only the survival functions and not the censoring distributions; thus it represents the survival experience of an "average person" on study. Although other definitions of the weighted Kaplan-Meier estimate could be chosen (for example, the pooled Kaplan-Meier estimate based on the combined sample or a weighted average of the Kaplan-Meier estimates where the weights are based on the number at risk), in general, they estimate quantities that involve not only the survival distributions but the censoring distributions as well. This will be true, for example, when the survival distributions are different for the populations and the censoring distributions are different.

A score is assigned to each of the n_i observations from the ith population. We consider the scoring function

$$\hat{Q}(X_{ij}, \delta_{ij}, \hat{M}) = P(X_{ij}{}^0 > \hat{M} \mid X_{ij}, \delta_{ij}, \hat{S}_i{}^0);$$

in the spirit of Efron (1967), this conditional probability is to be interpreted as if $X_{ij}{}^0$ actually had the distribution $\hat{S}_i{}^0$. Then

$$\hat{Q}(X_{ij}, \delta_{ij}, \hat{M}) = \begin{cases} 1 & X_{ij} > \hat{M} \\ \hat{S}_i{}^0(\hat{M})/\hat{S}_i{}^0(X_{ij}) & X_{ij} \leq \hat{M}, \delta_{ij} = 0 \\ 0 & X_{ij} \leq \hat{M}, \delta_{ij} = 1 \end{cases}$$

The statistic $(1/n_i) \sum_{j=1}^{n_i} \hat{Q}(X_{ij}, \delta_{ij}, \hat{M})$ reduces to $\hat{S}_i{}^0(\hat{M})$ by the property of self-consistency of the Kaplan-Meier estimate (see Efron 1967; Brookmeyer and Crowley 1982). Then the median test for censored data consists of evaluating the Kaplan-Meier estimate for each of the samples at the pooled-sample median.

It may be shown that under the null hypothesis of identical survival functions for the k populations,

$$\tilde{\mathbf{X}}_N = \{\sqrt{N}(\hat{F}_i{}^0(\hat{M}) - \tfrac{1}{2})\} \xrightarrow{D} N_k(0, A \ddagger A'),$$

where $A \ddagger A'$ is a $k \times k$ matrix that can be consistently estimated from the data, say by $A \hat{\ddagger} A'$. Furthermore, it can be proved that

$$\tilde{\mathbf{X}}_N'(A \hat{\ddagger} A')^- \tilde{\mathbf{X}}_N \xrightarrow{D} \chi^2(k - 1),$$

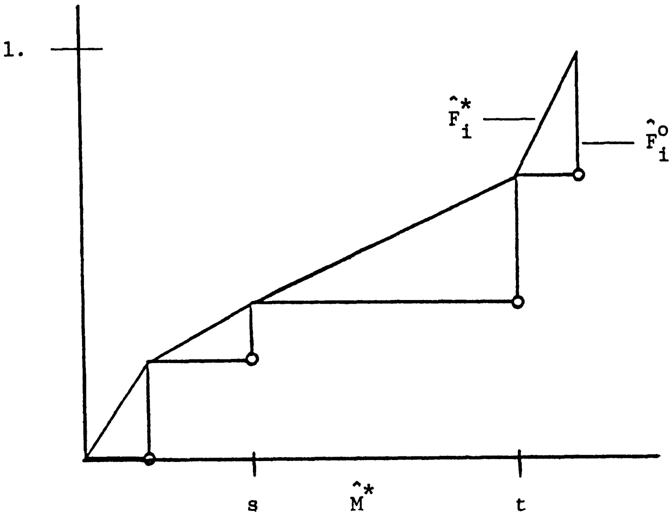

Figure 1. Continuous Version $F_i{}^*$ of the Kaplan-Meier Estimate

where $(A \hat{\ddagger} A')^-$ is a generalized inverse. One would expect such a statistic to be invariant for any choice of the G-inverse; that is, the value of the statistic in finite samples should not depend on the choice of the G-inverse. In general, the statistic will be G-inverse invariant if $\tilde{\mathbf{X}}_N$ is in the column space of $A \hat{\ddagger} A'$ (Graybill 1969). However, this cannot be guaranteed, essentially because the Kaplan-Meier estimate is a step function and thus $\hat{F}_w{}^0(\hat{M})$ need not be precisely $\frac{1}{2}$. One way to avoid the ambiguity caused by choosing a G-inverse is to transform each $\hat{F}_i{}^0(i = 1, \ldots, k)$ into a continuous function $\hat{F}_i{}^*$ by connecting the jump points (see Figure 1). Then let $\mathbf{X}_N = \{\sqrt{N}(\hat{F}_i{}^*(\hat{M}^*) - \frac{1}{2})\}$, where $\hat{M}^*$ is the point satisfying $\sum_{i=1}^k \lambda_i{}^N \hat{F}_i{}^*(\hat{M}^*) = \hat{F}_w{}^*(\hat{M}^*) = \frac{1}{2}$. It is shown in the next section that the statistic $\mathbf{X}_N'(A \hat{\ddagger} A')^- \mathbf{X}_N$, which is G-inverse invariant, is asymptotically $\chi^2(k - 1)$.

3. ASYMPTOTIC DISTRIBUTION OF THE TEST STATISTIC

In this section the joint asymptotic distribution of $(\hat{F}_1{}^*(\hat{M}^*) \cdots \hat{F}_k{}^*(\hat{M}^*))$, suitably normalized, is derived under both the null and alternative hypotheses. It is well known that the Kaplan-Meier estimate converges weakly to a mean-0 Gaussian process (Breslow and Crowley 1974). The problem at hand is that the Kaplan-Meier estimate is being evaluated at the random point $\hat{M}^*$.

The survival function for the ith population is $S_i{}^0(t) = 1 - F_i{}^0(t) = P(X_{ij}{}^0 > t)$; and the censoring times have cumulative distribution function $H_i(t) = P(T_{ij} \leq t)$. Then let $F_w{}^0(t) = \sum_{i=1}^k \lambda_i F_i{}^0(t)$, where it is assumed that $\lambda_i{}^N \to \lambda_i$, $0 < \lambda_i < 1$.

In general, the medians of the k populations are not equal. Thus, $\hat{M}^*$ is not estimating a common median but rather a "pooled median" $c = F_w{}^{0-1}(\frac{1}{2})$. It is shown here that

$$\mathbf{X}_N = \{\sqrt{N}(\hat{F}_i{}^*(\hat{M}^*) - F_i{}^0(c))\}$$

converges in distribution to a multivariate normal distribution. The following set of assumptions is required.

A1: F_i^0, H_i are continuous $i = 1, \ldots, k$.

A2: A model of random censorship: the censoring CDF H_i is independent of the survival CDF F_i^0.

A3: $H_i(c) < 1, i = 1, \ldots, k$.

A4: $(dF_i^0(t)/dt)\mid_c = f_i^0(c) \neq 0$, $(dF_w^0(t)/dt)\mid_c \neq 0$.

The statistic $\sqrt{N}\,(\hat{F}_i^*(\hat{M}^*) - F_i^0(c))$ is decomposed into k parts in Lemma 3.1; the motivation for this decomposition arises from the techniques developed by Pyke and Shorack (1968).

Lemma 3.1.

$$\sqrt{N}\,(\hat{F}_i^*(\hat{M}^*) - F_i^0(c))$$
$$= \frac{(1 - \lambda_i^N \hat{Q}_i)}{(\lambda_i^N)^{1/2}} \sqrt{n_i}\,(\hat{F}_i^*(\hat{M}^*) - F_i^0(\hat{M}^*))$$
$$- \sum_{j \neq i} \sqrt{\lambda_j^N n_j}\, \hat{Q}_i(\hat{F}_j^*(\hat{M}^*) - F_j^0(\hat{M}^*)),$$

where

$$\hat{Q}_i = (F_i^0(\hat{M}^*) - F_i^0(c))/(F_w^0(\hat{M}^*) - \tfrac{1}{2}) \quad i = 1, \ldots, k.$$

Each of the k terms $\sqrt{n_j}\,(\hat{F}_j^*(\hat{M}^*) - F_j^0(\hat{M}^*))$ will converge in distribution to independent normal random variables with mean 0 and variance given by the variance of the standardized Kaplan-Meier process (Breslow and Crowley 1974) at the point c (see Appendix 1). Furthermore, using a Taylor series expansion it can be shown that $\hat{Q}_i$ converges almost surely to

$$Q_i = f_i^0(c) \Big/ \sum_{i=1}^{k} \lambda_i f_i^0(c).$$

Thus, Theorem 3.2 is readily proved with an application of Slutsky's theorem.

Theorem 3.2. Let $\mathbf{X}_N$ be the k-dimensional vector with ith component $\sqrt{N}\,(\hat{F}_i^*(\hat{M}^*) - F_i^0(c))$. Under assumptions A1 through A4, $\mathbf{X}_N \xrightarrow{D} N_k(0, A\,\Sigma\,A')$, where Σ is the $k \times k$ diagonal matrix with diagonal elements V_i given by

$$V_i = [S_i^0(c)]^2 \int_0^c \frac{dF_i^0}{(1 - F_i^0)^2(1 - H_i)}$$

and A is the $k \times k$ matrix

$$A = \begin{pmatrix} \dfrac{1 - \lambda_1 Q_1}{\sqrt{\lambda_1}} & -\sqrt{\lambda_2}\,Q_1 & & -\sqrt{\lambda_k}\,Q_1 \\ -\sqrt{\lambda_1}\,Q_2 & \dfrac{1 - \lambda_2 Q_2}{\sqrt{\lambda_2}} & \cdots & -\sqrt{\lambda_k}\,Q_2 \\ \vdots & & & \\ -\sqrt{\lambda_1}\,Q_k & -\sqrt{\lambda_2}\,Q_k & \cdots & \dfrac{1 - \lambda_k Q_k}{\sqrt{\lambda_k}} \end{pmatrix}.$$

Under the null hypothesis H_0: $F_i^0 = F^0, i = 1, \ldots, k$, $Q_i = 1$ and $F^0(c) = \frac{1}{2}$. Matrix multiplication of $A\,\Sigma\,A'$ gives the asymptotic variance and covariance of the (i, j)th components of $\mathbf{X}_N$:

$$\operatorname{var}(X_i) = \frac{(1 - \lambda_i)^2}{\lambda_i} V_i + \sum_{i \neq j} \lambda_j V_j,$$

$$\operatorname{cov}(X_i, X_j) = \left(\sum_{l=1}^{k} \lambda_l V_l\right) - (V_i + V_j).$$

Consistent estimates $\hat{V}_i$ of the Kaplan-Meier variances V_i are discussed in Section 5 (see expression 5.1, for example). We can then estimate Σ by $\hat{\Sigma}$, say. It is verified in Appendix 2 that the rank of $A\,\Sigma\,A'$ is $k - 1$. Theorem 3.3 follows immediately.

Theorem 3.3. Under assumptions A1 through A4, $\mathbf{X}_N'(A\,\hat{\Sigma}\,A')^-\mathbf{X}_N \xrightarrow{D} \chi^2(k - 1)$.

As discussed in Appendix 2, the statistic $\mathbf{X}_N'(A\,\hat{\Sigma}\,A')^-\mathbf{X}_N$ is G-inverse invariant. It should be noted that in finite samples the rank of $A\,\hat{\Sigma}\,A'$ may be less than $k - 1$. This is because our estimates $\hat{V}_i$ of the diagonal elements of $\hat{\Sigma}$ may be zero; in which case the rank of $A\,\hat{\Sigma}\,A'$ is the number of nonzero diagonal elements of $\hat{\Sigma}$. One should be aware of this when computing the G-inverse, although the statistic should still be compared to a χ^2 distribution with $k - 1$ degrees of freedom.

4. ASYMPTOTIC POWER AND EFFICIENCY CALCULATIONS

The median test statistic for censored data for the two-sample problem takes a simple form. Application of the results of Section 3 shows that

$$\sqrt{N}\,(\hat{F}_1^*(\hat{M}^*) - F_1^0(c)) \xrightarrow{D} N(0, \sigma^2),$$

where

$$\sigma^2 = ((1 - \lambda_1 Q_1)^2/\lambda_1)\,V_1 + (1 - \lambda_1)Q_1^2 V_2.$$

Under the null hypothesis, $\sigma^2 = \sigma_0^2 = ((1 - \lambda_1)/\lambda_1)(\lambda_2 V_1 + \lambda_1 V_2)$; thus the null variance σ_0^2 is a linear combination of the Kaplan-Meier variances that decreases to 0 as λ_1 increases to 1. This suggests a large-sample power approximation that may be useful in sample size determinations: if one is interested in the one-sided hypothesis testing problem H_0: $F_1^0 \equiv F_2^0$ and H_A: $F_1^0 < F_2^0$, then an estimate of the power for two cdf's F_1^0 and F_2^0 is given by $P(Z < K^{**})$, where $Z \sim N(0, 1)$ and

$$K^{**} = \frac{Z_\alpha \sigma_0}{\sigma} - \frac{\sqrt{N}}{\sigma}(F_1^0(c) - \tfrac{1}{2})$$

with σ and σ_0 as defined earlier; in order to compute V_1 and V_2, censoring distributions H_1 and H_2 must be assumed.

Although the median test for uncensored data is the locally most powerful rank test for detecting location differences with double exponentials, it is not as powerful as other rank tests against other alternatives. The asymptotic relative efficiency is a useful way of comparing the median test to its nonparametric competitors. The efficiencies for the median, logrank and Wilcoxon tests have been computed for various alternatives under the assumption of no censorship and equal sample sizes. Table 1 lists the efficiencies relative to the one of the three statistics that has the highest efficacy (see also Tarone and Ware 1977).

The asymptotic relative efficiency for the median test with censored data depends on the censoring distributions in a complicated manner. However, it has been calculated here for the proportional hazards situation: let $1 - F_1^0(t) = S_1^0(t) = S^0(t)$ and $S_2^0(t) = (S^0(t))^\theta$. In addition, we assume that the censoring distributions are also in that proportional hazards family; that is,

$$1 - H_1(t) = (S^0(t))^{\alpha_1}$$

and

$$1 - H_2(t) = (S^0(t))^{\alpha_2}.$$

The asymptotic relative efficiency for two test statistics is the ratio of the square of their efficacies. A calculation shows that

$$(\text{efficacy})^2 = (1 - \lambda_1)\lambda_1(-\ln 2)^2/$$
$$\left(\frac{1 - \lambda_1}{\alpha_1 + 1}(2^{\alpha_1 + 1} - 1) + \frac{\lambda_1}{\alpha_2 + 1}(2^{\alpha_2 + 1} - 1)\right).$$

A special case frequently considered in the literature (Efron 1967; Gehan 1965; Crowley and Thomas 1975) is the exponential case

$$S_1^0(t) = \exp(-\phi t) \qquad S_2^0(t) = \exp(-\theta\phi t)$$
$$1 - H_1(t) = \exp(-t) \quad 1 - H_2(t) = \exp(-\alpha t)$$

In this case the expression for the efficacy with equal sample sizes becomes

$$(\text{efficacy})^2 = (\ln 2)^2/$$
$$2\phi\left[\frac{1}{\phi + 1}(2^{1/\phi + 1} - 1) + \frac{1}{\alpha + \phi}(2^{\alpha/\phi + 1} - 1)\right].$$

Table 1. Asymptotic Efficiencies Relative to the Best of the Three Test Statistics (no censorship)

	Test Statistics		
Alternatives	*Logrank*	*Wilcoxon*	*Median*
Lehmann	1	.75	.48
Logistic	.75	1	.75
Scale-Lognormal	.85	1	.67
Translation-Double Exponential	.48	.75	1
Lehmann:	$S_2(x) = (S_1(x))^\theta$		
Logistic:	$S_2(x) = S(x)\,[S(x) + \epsilon^\theta(1 - S(x)]^{-1}$		
Scale (lognormal):	$S_2(x) = S_1(\theta x)$, $f_1(x) = \exp(-\frac{1}{2}\log^2 x)/\{\sqrt{2\pi}\, x\}$		
Translation (double exponential):	$S_2(x) = S_1(x + \theta)$, $f_1(x) = \frac{1}{2}\exp(-\lvert t\rvert)$		

5. MONTE CARLO RESULTS

The performance of the median test was compared to three other test statistics (logrank, generalized Wilcoxon, and Prentice-median) for the two-sample case by several Monte Carlo experiments. First, the computational forms of the four test statistics in the no-ties situation are reviewed.

1. *Median test.* The version of the median test used in these simulations is

$$T_M = N\left[\frac{\hat{F}_1^*(\hat{M}^*) - \frac{1}{2}}{\sigma_0}\right]^2 \xrightarrow{D} \chi^2(1).$$

An estimate of σ_0^2 is given by

$$\hat{\sigma}_0^2 = \frac{\lambda_2}{\lambda_1}[\lambda_2\hat{V}_1(\hat{M}^*) + \lambda_1\hat{V}_2(\hat{M}^*)].$$

Here, $\hat{V}_i(\hat{M}^*)$ is an estimate of the variance of the standardized Kaplan-Meier process $Z_{n_i}{}^i(\hat{M}^*)$ (see Greenwood 1926),

$$\hat{\text{v}}\text{ar}(Z_{n_i}{}^i(\hat{M}^*) = \hat{V}_i(\hat{M}^*) = [S_i^*(\hat{M}^*)]^2 \times \sum_{\{j|X_{ij} \leq M^*\}} \frac{n_i d_{ij}}{N_i(X_{ij})(N_i(X_{ij}) - d_{ij})}, \tag{5.1}$$

where $N_i(X_{ij})$ is the number from the ith population at risk at X_{ij}, and d_{ij} is the number of observed deaths from the ith population at X_{ij}. In the no-ties situation, $d_{ij} = \delta_{ij}$ (that is, d_{ij} is either 0 or 1). An alternative estimate $\hat{V}_i^*(\hat{M}^*)$ can be obtained that accounts for the smoothing of the Kaplan-Meier estimate and thus the covariance of $\hat{F}_i^0(s)$ and $\hat{F}_i^0(t)$ (where $\hat{M}^*$ falls between the two consecutive observed death times from the ith sample s and t, see Figure 1).

$$\begin{aligned} \hat{V}_i^*(\hat{M}^*) &= \left(\frac{\hat{M}^* - s}{t - s}\right)^2 \hat{\text{v}}\text{ar}(Z_{n_i}{}^i(s)) + \left(\frac{t - \hat{M}^*}{t - s}\right)^2 \hat{\text{v}}\text{ar}(Z_{n_i}{}^i(t)) \\ &\quad + \frac{2(\hat{M}^* - s)(t - \hat{M}^*)}{(t - s)^2}\,\hat{\text{c}}\text{ov}\,(Z_{n_i}{}^i(s), Z_{n_i}{}^i(t)) \\ &= \left[S_i^0(t)\left(\frac{\hat{M}^* - s}{t - s}\right)\right]^2 \sum_{\{j|X_{ij} \leq t\}} \frac{n_i d_{ij}}{N_i(X_{ij})(N_i(X_{ij}) - d_{ij})} \\ &\quad + \left\{\left[S_i^0(s)\left(\frac{t - \hat{M}^*}{t - s}\right)\right]^2 \right. \\ &\quad \left. + \frac{2(\hat{M}^* - s)(t - \hat{M}^*)}{(t - s)^2} S_i^0(s) S_i^0(t)\right\} \\ &\quad \times \sum_{\{j|X_{ij} \leq s\}} \frac{n_i d_{ij}}{N_i(X_{ij})(N_i(X_{ij}) - d_{ij})}. \end{aligned} \tag{5.2}$$

The estimate $\hat{V}_i^*(\hat{M}^*)$ is recommended in small samples, while for moderate samples of size 50 or larger our simulation results indicate that the simpler estimate $\hat{V}_i(\hat{M}^*)$

is adequate.

2. *Logrank test.* (Mantel 1966; Peto and Peto 1972; Cox 1972)

$$T_L = \frac{[\sum_{(i,j)} \delta_{ij}(D_{ij} - N_1(X_{ij})/N_Z(X_{ij}))]^2}{\sum_{(i,j)} \delta_{ij}(N_1(X_{ij})N_2(X_{ij})/[N_Z(X_{ij})]^2} \xrightarrow{D} \chi^2(1),$$

where $D_{ij} = 0$ if the (i, j)th observation is a death from the second sample, and $D_{ij} = 1$ if the (i, j)th observation is a death from the first sample.

$$N_Z(X_{ij}) = N_1(X_{ij}) + N_2(X_{ij}).$$

3. *Generalized Wilcoxon test* (Gehan 1965; Breslow 1970; Tarone and Ware 1977)

$$T_B = \frac{[\sum_{(i,j)} \delta_{ij}N_Z(X_{ij})(D_{ij} - N_1(X_{ij})/N_Z(X_{ij}))]^2}{\sum_{(i,j)} \delta_{ij}N_1(X_{ij})N_2(X_{ij})} \xrightarrow{D} \chi^2(1)$$

4. *Prentice-median (1978) test.* This test statistic is standardized by its permutation variance and is assumed to be asymptotically $\chi^2(1)$.

$$T_P = \frac{\left(\sum_{j=1}^{n_i} C_{1j}\right)^2}{(\sum_{(i,j)} C_{ij}^2) \cdot \frac{n_1 n_2}{N(N-1)}},$$

$$\text{where } C_{ij} = \begin{cases} -1 & \text{if } \hat{G}(X_{ij}) > \frac{1}{2}, \delta_{i,j} = 1 \\ \frac{1 - \hat{G}(X_{ij})}{\hat{G}(X_{ij})} & \text{if } \hat{G}(X_{ij}) > \frac{1}{2}, \delta_{ij} = 0 \\ 1 & \text{if } \hat{G}(X_{ij}) \le \frac{1}{2} \end{cases}$$

$$\text{and } \hat{G}(X_{ij}) = \prod_{\{(m,n)|\delta_{mn}=1,\, X_{mn} \le X_{ij}\}} \left(\frac{N_Z(X_{mn})}{N_Z(X_{mn}) + 1}\right).$$

The estimate $\hat{G}(X_{ij})$ is very similar to the Kaplan-Meier estimate for the pooled sample. A linear transformation of the Prentice scores to $(C_{ij} + 1)/2$ shows that these scores differ from $\hat{Q}(X_{ij}, \delta_{ij}, \hat{M})$ in that both the scores and pooled sample median are based on $\hat{G}$.

An experiment consisted of generating survival and censoring times for two samples of size 50 each; and four test statistics were computed; this was repeated 1,000 times; the fraction of times H_0 was rejected (estimated power) for each test statistic was calculated at several α levels, α = .01, .05, .10, .20, .25.

Experiment 1 assumed that sample 1 was generated from a double exponential (median = 100) while sample 2 was generated from a double exponential (median = θ), θ = 100, 99.8, 99.6, 99.4, 99.2. Censoring for both samples was assumed to be double exponential (100.5). The results at level α = .05 are illustrated graphically in Figure 2. All tests performed at approximately the prescribed α level; in addition the median and Prentice-median test were somewhat more powerful than the logrank test.

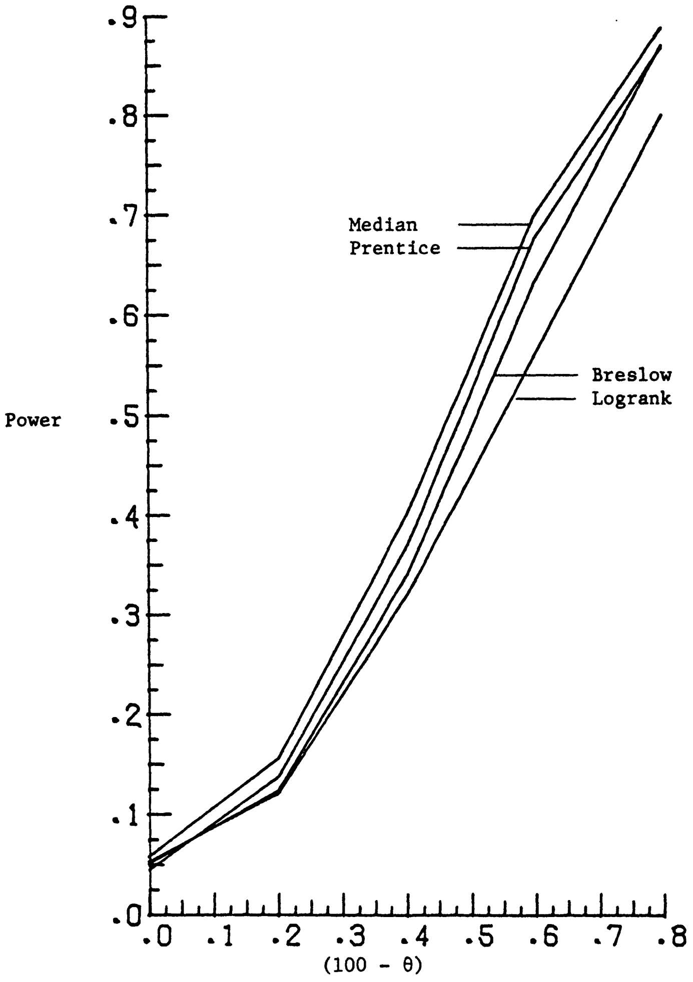

Figure 2. Power Curves for Double Exponentials (α = .05)—Sample 1: n_1 = 50 Double Exp (θ); Sample 2: n_2 = 50 Double Exp (100); Censoring: Double Exp (100.5)

Experiment 2 assumed that sample 1 was generated from an exponential (.01) while sample 2 was generated from an exponential (ϕ), ϕ = .01, .012, .014, .016, .018, .020. The censoring distribution for both samples was assumed to be uniform [0, 250]. The results are illustrated graphically in Figure 3 for α = .05. As expected, the logrank test was more powerful than the other tests for this proportional hazards situation.

Experiment 3 investigated a cross-over survival situation generated from two Weibull distributions with a censoring distribution uniform on [0, 200]. A similar situation was considered by Fleming et al. (1981). A graph of this crossing-hazards situation reveals few early differences between the survival curves with increasing differences later on. The results (see Table 2) show that while the logrank and median test performed well, the generalized Wilcoxon test had very low power. This result should not be surprising in this particular crossing hazards situation since it is well known that this test is

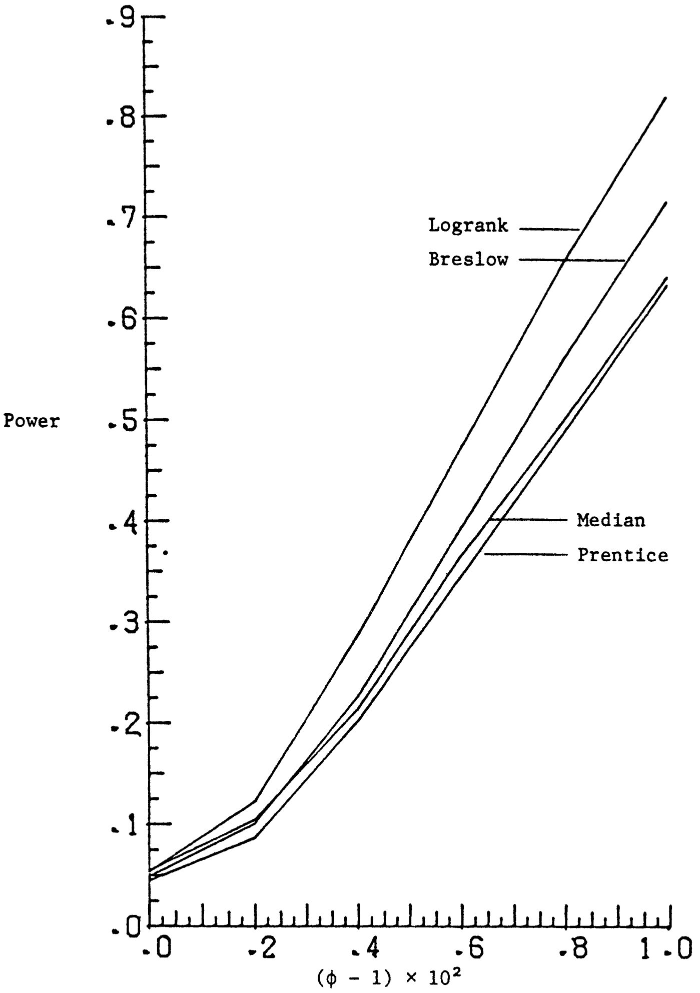

Figure 3. Power Curves for Exponentials ($\alpha = .05$)—Sample 1: $n_1 = 50$ Exp (.01); Sample 2: $n_2 = 50$ Exp (ϕ); Censoring: Uniform (0, 250); Exp (ϕ) = $S^0(t) = \exp(-\phi t)$

not sensitive to later differences in survival. Of course, the generalized Wilcoxon test might be more powerful in other crossing-hazards situations; the superiority of one of these tests depends on the particular cross-over survival situation. In such situations, estimating and graphing of the survival curves become particularly important.

Experiment 4 considered two survival distributions with the same median but very different shapes. Since the median test is primarily sensitive to differences in medians, one would expect the median test to perform poorly here, with a power not much greater than the α-level the test is performed at. Sample 1 was generated from an exponential (.01) while sample 2 was generated from a Weibull with shape parameter 2. The censoring was uniform on [0, 250]. The results (Table 3) suggest that the Prentice-median and median tests are not sensitive to shape differences if the two curves have the same median.

Table 2. Power for Cross-Over Survival Curves[a]

α	*Median*	*Logrank*	*Breslow*	*Prentice*
.01	.542	.775	.099	.406
.05	.712	.912	.229	.618
.10	.794	.954	.322	.719

[a] *X*: Weibull (.0002, 2) 31.4 percent censored; *Y*: Weibull (.05, .5) 63.2 percent censored; (Weibull (ϕ, γ) = $S^o(t) = \exp(-\phi t^\gamma)$.)

Table 3. Power for Survival Curves With Different Shapes and Equal Medians[a]

α	*Median*	*Logrank*	*Breslow*	*Prentice*
.01	.012	.046	.027	.014
.05	.059	.150	.087	.061
.10	.130	.237	.172	.108

[a] *X*: Exp (.01) 36.8 percent censored; *Y*: Weibull (1.443×10^{-4}, 2) 29.5 percent censored.

The behavior of the statistics was investigated in small samples (sample size = 20) to see how well they approximate the asymptotic null distribution. The results are reported in Table 4 for both samples generated from a double exponential (median = 100) with censoring distributions as described in experiment 1; and the results are reported in Table 5 for both samples generated from an exponential (.01) with censoring distributions as described in experiment 2. The tables show that the median test with expression (5.2) more accurately approximates the asymptotic null distribution than with expression (5.1), which rejects H_0 more often than it should. The median test with expression (5.2) and the three other test statistics are performing at approximately the correct α-level.

Table 4. Small Sample Size Results for Double Exponentials ($N_1 = N_2 = 20$)[a]

	α		
	.01	*.05*	*.10*
Median (with eq. (5.1))	.023	.072	.139
Median (with eq. (5.2))	.013	.055	.114
Logrank	.011	.054	.097
Breslow	.017	.055	.112
Prentice	.015	.048	.110

[a] *X*: 37.4 percent censored; *Y*: 38.3 percent censored.

*Table 5. Small Sample Size Results for Exponentials ($N_1 = N_2 = 20$)**

	α		
	.01	*.05*	*.10*
Median (with eq. (5.1))	.019	.073	.124
Median (with eq. (5.2))	.014	.058	.101
Logrank	.009	.050	.103
Breslow	.016	.054	.103
Prentice	.010	.045	.097

[*] In one simulation the pooled median was not reached because of extensive censoring; this simulation was discarded.

[*] *X*: 36.1 percent censored; *Y*: 37.2 percent censored.

Table 6. Sample Sizes and Median Survival Times for Four Treatments of Colorectal Cancer

	#1	#2	#3	#4
Censored	16	8	14	7
Observed Deaths	37	48	44	45
Sample Size	53	56	58	52
Median Survival Time (weeks)	61	41	47	29

6. AN EXAMPLE

The median test procedure is applied to data from a Phase III colorectal cancer clinical trial. Four dosage regimens of 5-Fluorouracil are compared (Ansfield et al. 1977)[1]; Table 6 summarizes the data. Below is a partial list of the Kaplan-Meier estimates at some of the observed death times of the combined sample.

Death Time	$\hat{S}_1^0$	$\hat{S}_2^0$	$\hat{S}_3^0$	$\hat{S}_4^0$
41	.582	.485	.501	.422
43	.559	.465		.401
47			.480	

Calculations give $\hat{M}^* = 41.004$, and

$$\mathbf{X}_N = \{\sqrt{N}(\hat{F}_i^*(\hat{M}^*) - \tfrac{1}{2})\}$$
$$= (-\ 1.21, .22, -\ .015, 1.15).$$

The matrix Σ is the 4 × 4 diagonal matrix with entries (.271, .286, .293, .270). These variances were estimated by expression (5.1), which accounts for the ties in the data. Then $\mathbf{X}_N'(A\ \Sigma\ A')^-\mathbf{X}_N = 2.85$ and comparison with the $\chi^2(3)$ distribution leads us not to reject H_0 ($p > .25$). We note that the logrank and generalized Wilcoxon tests gave statistics of 4.29 and 4.49, respectively, both of which should also be compared with the $\chi^2(3)$ distribution.

Using the techniques of Brookmeyer and Crowley (1982) for computing nonparametric confidence intervals for the median, one obtains the following 95 percent confidence intervals:

Treatment 1	[38, 73)	Treatment 3	[28, 60)
Treatment 2	[31, 51)	Treatment 4	[25, 46)

It is interesting to note that all four confidence intervals overlap.

APPENDIX 1

We shall prove that the random variables $Z_{n_i}{}^i(\hat{M}^*) = \sqrt{n_i}(\hat{F}_i^*(\hat{M}^*) - F_i^0(\hat{M}^*))$ converge in distribution to independent normal random variables.

First, we note that $\hat{M}^* \to c$ a.s. since $\hat{F}_i^*(t) \to F_i^0(t)$ a.s. and thus $\hat{F}_w^*(t) \to F_w^0(t)$ a.s. (Peterson 1977).

Next consider the random functions

$$Z_{n_i}{}^i(\cdot) = \sqrt{n_i}(\hat{F}_i^*(\cdot) - F_i^0(\cdot))$$

in $C[0, T]$, the space of continuous functions on $[0, T]$; here T is such that $T > c$ and $H_i(T) < 1$. We can be assured that for N sufficiently large, $\hat{M}^* < T$ almost surely. Now $Z_{n_i}{}^i(\cdot)$ converges weakly to a mean-0 Gaussian process, say $Z^i(\cdot)$ (see Breslow and Crowley 1974; they work, however, with the nonsmoothed Kaplan-Meier process based on $\hat{F}_i^0(\cdot)$). Furthermore, since $\hat{M}^* \to c$ a.s., we have

$$(Z_{n_i}{}^i, \hat{M}^*) \xrightarrow{w} (Z^i, c),$$

that is, convergence in distribution relative to the product topology. Since the composite function mapping ψ: $C \times R \to R$ defined by $\psi(z, m) = z(m)$ is continuous, we can conclude that $Z_{n_i}{}^i(\hat{M}^*) \xrightarrow{w} Z^i(c)$ (Billingsley 1968). Finally noting that $Z^i(\cdot)$ are independent processes $1 \le i \le k$ by assumption, we have proven the following result: Under assumptions A1 through A4, the k-dimensional vector with ith component $Z_{n_i}{}^i(\hat{M}^*)$ will converge in distribution to $N_k(0, \Sigma)$, where Σ is the $k \times k$ matrix as defined in Theorem 3.2.

APPENDIX 2

We shall prove that the rank of $A\ \Sigma\ A'$ is $k - 1$. Multiplying the ith row of A by $\sqrt{\lambda_i}/Q_i$ gives $A^* = I - pp'$, where I is the $k \times k$ identity matrix and $\mathbf{p}' = (\sqrt{\lambda_i}, \ldots, \sqrt{\lambda_k})$. Since elementary row operations do not affect the rank, rank (A) = rank (A^*). Furthermore, A^* is idempotent since

$$A^* \cdot A^* = (I - \mathbf{pp}')(I - \mathbf{pp}')$$
$$= I - 2\mathbf{p}\,\mathbf{p}' + \mathbf{p}\,\mathbf{p}'\,\mathbf{p}\,\mathbf{p}' = I - \mathbf{pp}'.$$

Hence rank (A^*) = trace $(A^*) = \sum_{i=1}^{k} (1 - \lambda_i) = k - 1$.

Now the null spaces of A and $A\ \Sigma\ A'$ are identical. This follows because all the diagonal elements of Σ are nonzero (guaranteed by our assumptions) and thus Σ is of full rank. Then it follows that rank $(A\ \Sigma\ A')$ = rank $(A) = k - 1$.

The statistic $\mathbf{X}_N'(A\ \Sigma\ A')^-\mathbf{X}_N$ is G-inverse invariant. The argument is straightforward: the orthogonal complement of the column space of $A\ \Sigma\ A'$ is the same as the null space of $A\ \Sigma\ A'$ which contains the null space of A'. Now the vector $\boldsymbol{\lambda}' = (\lambda_1, \lambda_2, \ldots, \lambda_k)$ spans the null space of A'; further $\boldsymbol{\lambda}'\mathbf{X}_N = 0$. Thus, $\mathbf{X}_N$ must lie in the column space of $A\ \Sigma\ A'$.

[*Received December 1980. Revised December 1981.*]

REFERENCES

ANSFIELD, F., KLOTZ, J., and THE CENTRAL ONCOLOGY GROUP (1977), "A Phase III Study Comparing the Clinical Utility of Four Dosage Regimens of 5-Fluorouracil," *Cancer*, 39, 34–40.

BILLINGSLEY, P. (1968), *Convergence of Probability Measures*, New York: John Wiley.

[1] The data analyzed were from an updated version reported in the Central Oncology Group Final Report (COG 7030), Spring 1977.

BRESLOW, N.R. (1970), "A Generalized Kruskal-Wallis Test for Comparing K Samples Subject to Unequal Patterns of Censorship," *Biometrika*, 57, 579–594.

BRESLOW, N.R., and CROWLEY, J. (1974), "A Large Sample Study of the Life Table and Product Limit Estimates Under Random Censorship," *Annals of Statistics*, 2, 437–453.

BROOKMEYER, R., and CROWLEY, J. (1982), "A Confidence Interval for the Median Survival Time," *Biometrics*, to appear.

COX, D.R. (1972), "Regression Models and Life Tables," *Journal of the Royal Statistical Society*, Ser. B, 26, 103–110.

CROWLEY, J., and THOMAS, D.R. (1975), "Large Sample Theory for the Log Rank Test," Technical Report #415, Department of Statistics, University of Wisconsin.

EFRON, B. (1967), "The Two-Sample Problem with Censored Data," *Proceeding of the Fifth Berkeley Symposium on Mathematical Statistics and Probability*, 4, 831–854.

——— (1981), "Censored Data and the Bootstrap," *Journal of the American Statistical Association*, 76, 312–319.

EMERSON, J. (1982), "Nonparametric Confidence Interval for Quantiles in the Presence of Partial Right Censoring," *Biometrics*, to appear.

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.

GEHAN, E.A. (1965), "A Generalized Wilcoxon Test for Comparing Arbitrarily Singly-Censored Samples," *Biometrika*, 52, 203–223.

GILL, R.D. (1980), "Censoring and Stochastic Integrals," Mathematical Centre Tracts 124, Mathematische Centre, Amsterdam.

GRAYBILL, F.A. (1969), *Introduction to Matrices with Applications in Statistics*, Belmont, Calif.: Wadsworth Publishing.

GREENWOOD, M. (1926), "The Natural Duration of Cancer," *Reports on Public Health and Medical Subjects, Her Majesty's Stationery Office*, 33, 1–26.

HÁJEK, J., and SIDÁK, Z. (1967), *Theory of Rank Tests*, New York: Academic Press.

KAPLAN, E.L., and MEIER, P. (1958), "Nonparametric Estimation from Incomplete Observations," *Journal of the American Statistical Association*, 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.

PETERSON, A.V., JR. (1977), "Expressing the Kaplan-Meier Estimator as a Function of Empirical Subsurvival Functions," *Journal of the American Statistical Association*, 72, 854–858.

PETO, R., and PETO, J. (1972), "Asymptotically Efficient Rank Invariant Test Procedures," *Journal of the Royal Statistical Society*, Ser. A, 135, 185–206.

PRENTICE, R.L. (1978), "Linear Rank Tests with Right Censored Data," *Biometrika*, 65, 167–179.

PYKE, R., and SHORACK, G. (1968), "Weak Convergence of a Two-Sample Empirical Process and a New Approach to Chernoff-Savage Theorems," *Annals of Mathematical Statistics*, 39, 755–771.

REID, N. (1981), "Estimating the Median Survival Time," *Biometrika*, 68, 601–608.

TARONE, R.E., and WARE, J. (1977), "On Distribution-Free Tests for Equality of Survival Distributions," *Biometrika*, 64, 156–160.

WESTENBERG, J. (1948), "Significance Test for Median and Interquartile Range," *Koninklijke Nederlandsche Akademie Van Wetenschappen* 51, 252–261.