J. R. Statist. Soc. B (1991)
53, *No.* 2, *pp.* 341–352

Weighted Kaplan–Meier Statistics: Large Sample and Optimality Considerations

By MARGARET SULLIVAN PEPE† and THOMAS R. FLEMING

Fred Hutchinson Cancer Research Center and University of Washington, Seattle, USA

[Received November 1988. Final revision December 1989]

SUMMARY

We propose a cumulative weighted difference in the Kaplan–Meier estimates as a test statistic for equality of distributions in the two-sample censored data survival analysis problem. For stability of such a statistic, the absolute value of the possibly random weight function must be bounded above by a multiple of $(C^-)^{1/2+\delta}$ where $1-C^-$ is the left continuous censoring distribution function and $\delta > 0$. For these weighted Kaplan–Meier (WKM) statistics, asymptotic distribution theory is presented along with expressions for the efficacy under a sequence of local alternatives. A simple censored data generalization of the two-sample difference in means test (z-test) is a member of this class and in large samples is seen to be quite efficient relative to the popular log-rank test under a range of alternatives including the proportional hazards alternative. Optimal weight functions are also calculated. The optimal WKM statistic is as efficient as the optimal weighted log-rank statistic for any particular sequence of local alternatives. Stratified statistics and trend statistics are also presented.

Keywords: CENSORED DATA: PITMAN EFFICIENCY; TWO-SAMPLE TESTS; WEIGHTED LOG-RANK STATISTIC

1. INTRODUCTION

The familiar two-sample t-test is not easily generalized to right-censored survival data. Such data are frequently encountered in randomized clinical trials and for ease of exposition here we shall use the two-arm clinical trial as the particular application of interest. A natural approach to extending the test statistic to censored data is by standardizing an estimator of the difference in means of survival truncated at τ, the study duration. For example, if $\hat{S}_i(t)$ is the Kaplan–Meier estimator of the survival function $S_i(t)$, $i = 1, 2$, then

$$\int_0^\tau \{\hat{S}_1(t) - \hat{S}_2(t)\}\, dt$$

is such an estimator, reducing to the usual difference in means in uncensored data. Unfortunately $\hat{S}_i(t) - S_i(t)$ is known to be unstable for t close to τ if the probability of surviving the length of the study is non-zero and if censoring at τ is continuous in the sense that the probability of censoring by time t converges to unity as $t \to \tau$. By instability here we mean that, standardized by $1/\sqrt{n}$, the statistic is not asymptotically bounded in probability. Since in most clinical trials censoring is largely determined by

†*Address for correspondence*: Division of Public Health Sciences, Fred Hutchinson Cancer Research Center, MP-665, 1124 Columbia Street, Seattle, WA 98104, USA.

 0035–9246/91/53341 $2.00

enrolment into the study, which is usually continuous in nature, the mean difference estimate above is not a useful basis for a two-sample test. For $t < \tau$, $\hat{S}_i(t) - S_i(t)$ is stable, so that an alternative strategy is to use

$$\int_0^{t_0} \{\hat{S}_1(t) - \hat{S}_2(t)\}\, dt$$

for some $t_0 < \tau$ as the numerator of the test statistic. Although this is stable, the closer t_0 is to τ the larger its variance will be. The choice of t_0 is somewhat arbitrary and hence is difficult, requiring some weighting of the possible loss of information in choosing t_0 too small against the possible instability and resultant loss in power by choosing t_0 too close to τ.

Another approach to producing stability is to use a possibly random weight function $\hat{W}(t)$ to downweight the difference $\hat{S}_1(t) - \hat{S}_2(t)$ in the integrand over later time periods. This yields the weighted Kaplan–Meier (WKM) statistics,

$$\text{WKM} = \int_0^{\tau} \hat{W}(t)\{\hat{S}_1(t) - \hat{S}_2(t)\}\, dt,$$

a general class of two-sample location test statistics (Pepe and Fleming, 1989). In the special case where $\hat{W}$ is a simple function of the censoring patterns observed in the data, a censored data generalization of the ubiquitous two-sample z-test for a difference in means is obtained.

In Section 2 we present large sample distribution theory for these statistics and in doing so we develop conditions on $\hat{W}(\cdot)$ which, when satisfied, render the WKM statistics stable and useful. Since WKM statistics are not rank based, this development of large sample distributional results differs from that for generalized linear rank statistics where simple and elegant results from martingale theory can be invoked.

The most widely used two-sample statistic for censored survival data is the log-rank statistic, a generalization of the exponential scores linear rank statistic (Kalbfleisch and Prentice, 1980). The class of generalized linear rank statistics (also termed weighted log-rank statistics) behaves fundamentally differently from the non-rank-based WKM class (see Pepe and Fleming (1989)). In Section 3, we consider the optimal efficient weight function $W(\cdot)$ and compare the asymptotic efficiency of optimal and non-optimal WKM statistics relative to the log-rank statistic in a variety of situations. In particular, the asymptotic relative efficiency of the generalized z-test mentioned previously is explored. K-sample versions and stratified versions of WKM statistics are proposed and discussed briefly in Section 4. We conclude with some thoughts on the usefulness of these test statistics in practice.

2. LARGE SAMPLE THEORY

In this section we describe the development of asymptotic distribution theory for WKM statistics. Our aim is to develop conditions on $W(\cdot)$ to ensure stability of the statistic under the null hypothesis. We shall also develop an expression for the asymptotic efficacy of WKM statistics under a sequence of local alternatives. To keep the exposition as readable and brief as possible we have omitted some technical details, which can be found in O'Sullivan (1986).

2.1. *Notation*

Suppose that n possibly right-censored positive observations are available of which n_i come from sample i, $i = 1, 2$. We assume that the random censorship model holds so that each observation is the minimum of two independent random variables, the survival and censoring times. Let S_i denote the continuous survivor function for the survival times which are independent and identically distributed within each sample. Similarly, let C_i denote the right continuous survivor function for the independent and identically distributed censoring random variables in sample i. We shall use f^- for the left continuous version of the function f. The largest possible observation time in sample i is τ_i, often the study duration but, more generally,

$$\tau_i = \sup[t: \min\{S_i(t), C_i(t)\} > 0].$$

Similarly if $\hat{S}_i$ and $\hat{C}_i$ denote the Kaplan–Meier estimators of the corresponding survivor functions then the largest actual observation is

$$T_i = \sup[t: \min\{\hat{S}_i(t), \hat{C}_i(t)\} > 0].$$

Let $T = \min(T_1, T_2)$ and $\tau = \min(\tau_1, \tau_2) < \infty$.

We shall derive large sample theory not only under the null and under a fixed alternative but also under a sequence of local alternatives. A superscript n will denote the underlying configuration at sample size n and we assume convergence of the configurations in the sense that $\lim_{n\to\infty}\{S_i^n(t)\} = S_i(t)$ and $\lim_{n\to\infty}\{C_i^n(t)\} = C_i(t)$ for all $t \in [0, \tau_i)$. Again, to simplify the presentation we suppose that $\tau_i^n = \tau_i$, for all n, $i = 1, 2$. The weight function $\hat{W}(\cdot)$ estimates a deterministic positive bounded weight function $W(\cdot)$, and for local alternatives we assume that the $W^n(\cdot)$ are converging to some $W(\cdot)$ with

$$\lim_{n\to\infty}\left\{\sup_{t\in(0,\,\tau_i)} |W^n(t) - W(t)|\right\} = 0.$$

Finally we require that, asymptotically, the fraction of the total in sample i is non-negligible, i.e. $\lim_{n\to\infty}(n_i/n) = p_i > 0$, for $i = 1, 2$.

2.2. *Distribution Theory for Weighted Kaplan–Meier Statistics*

Asymptotic theory for functionals of $\hat{S}$ was proven by Gill (1983). In theorem 1 of Appendix A we have extended his results to a convergent sequence of configurations $(S^n(\cdot), C^n(\cdot))$. Substituting

$$h^n(t) = \int_t^\tau W^n(u)\, S^n(u)\, du$$

and

$$h(t) = \int_t^\tau W(u)\, S(u)\, du$$

in theorem 1 we see that

$$\sqrt{n}\int_0^T W^n(t)\{\hat{S}(t) - S^n(t)\}\, dt \xrightarrow{d} N(0, \sigma^2) \tag{2.1}$$

(by evaluating expression (A.3) at $t=\tau$) if

$$\sigma^2 = \int_0^\tau \left[\left\{ \int_t^\tau W(u)\, S(u)\, du \right\}^2 \Big/ S^2(t)\, C^-(t)\, dt \right] < \infty. \tag{2.2}$$

It is clear from equation (2.2) that, to ensure that σ^2 is finite for all choices of the unknown underlying survival and censoring distributions, a sufficient and almost necessary condition to be satisfied by W is that $W(t)^2/C^-(t)$ should be bounded uniformly in t on $[0, \tau)$.

To use a random rather than deterministic weight function, we can replace $W^n(t)$ with $\hat{W}(t)$ in expression (2.1) if

$$\sup_{t\in(0,\,T)} \left| \frac{\hat{W}(t) - W^n(t)}{C^{n-}(t)^{1/2}} \right| \xrightarrow{p} 0. \tag{2.3}$$

This follows from the fact that $\sup_{t\in(0,\,\tau)} |C^{n-}(t)^{1/2}\{\hat{S}(t) - S^n(t)\}\sqrt{n}|$ is bounded in probability, which in turn is also a consequence of theorem 1. If $\hat{W}$ is uniformly consistent for W then both conditions (2.2) and (2.3) are satisfied if the weight function decreases at a slightly faster rate than $C^-(t)^{1/2}$ as $t \to \tau$. To be precise,

$$\sqrt{n} \int_0^T \hat{W}(t)\{\hat{S}(t) - S^n(t)\}\, dt \xrightarrow{d} N(0, \sigma^2)$$

if

$$\sup_{t\in(0,\,\tau)} |\hat{W}(t) - W^n(t)| \xrightarrow{p} 0,$$

$$|W^n(t)| \leqslant \Gamma\, C^{n-}(t)^{1/2+\delta} \qquad t \in (0,\, \tau)$$

and

$$|\hat{W}(t)| \leqslant \Gamma\, \hat{C}^-(t)^{1/2+\delta} \qquad t \in (0,\, T)$$

for some $\Gamma > 0$ and $\delta > 0$. If $\hat{W}(\cdot)$ is non-random, then only the second condition with $\delta = 0$ need be satisfied. The proof of condition (2.3) as a consequence of these three conditions relies on the fact that $\hat{C}(t)/C^n(t)$ is bounded in probability.

Assuming that the conditions are satisfied for both samples, we see that

$$\left(\frac{n_1 n_2}{n}\right)^{1/2} \int_0^T [\hat{W}(t)\{\hat{S}_1(t) - \hat{S}_2(t)\} - \hat{W}(t)\{S_1^n(t) - S_2^n(t)\}]\, dt \xrightarrow{d} N(0,\, \sigma^2_{\mathrm{WKM}}) \tag{2.4}$$

where

$$\sigma^2_{\mathrm{WKM}} = -\sum_{i=1}^{2} (1-p_i) \int_0^\tau \left[\left\{ \int_t^\tau W(u)\, S_i(u)\, du \right\}^2 \Big/ S_i^2(t)\, C_i^-(t) \right] dS_i(t).$$

A consistent estimator of the variance is found by substituting estimators for S_i, C_i^- and W in σ^2_{WKM}. Under the null hypothesis, the Kaplan–Meier estimator for survival using the pooled sample can be used instead of $\hat{S}_i$, $i=1, 2$, to provide a consistent estimator $\hat{\sigma}^2_{\mathrm{WKM}}$.

The null hypothesis H_0: $S_1 = S_2$ can be tested by comparing $(n_1 n_2/n)^{1/2}\mathrm{WKM}/\hat{\sigma}_{\mathrm{WKM}}$

with a standard normal distribution. We see from expression (2.4) that this test is consistent against any alternative H_1: $S_1 \neq S_2$ where

$$\left| \int_0^\tau W(t)\{S_1(t) - S_2(t)\}\, dt \right| > 0,$$

since $\hat{\sigma}^2_{\text{WKM}}$ is bounded. In particular, this test is consistent against any stochastic ordering alternative if $W(\cdot)$ is positive, a result which is not necessarily true of the weighted log-rank tests.

To simplify an already involved presentation we have calculated the WKM statistic at $T = \min(T_1, T_2)$. This is the end point used by the log-rank statistic, which incorporates information over the time interval during which subjects in both samples are at risk. In uncensored data with $\hat{W} = 1$ this WKM statistic corresponds to calculating a difference in means after truncating all data at the smaller of the two largest observations from the two samples. This loss of data seems to be inappropriate in a non-rank-based test statistic.

A more suitable generalized location test statistic takes the form

$$\text{WKM} = \int_0^{T_c} \hat{W}(t)\{\hat{S}_1(t) - \hat{S}_2(t)\}\, dt$$

where $T_c = T$ if T is a censored observation (in which case it may not be possible to compare S_1 and S_2 thereafter) and $T_c = \max(T_1, T_2)$ if T is an uncensored observation. The above large sample theory holds for this WKM statistic with a few more mild conditions required on survival and censoring over $[T, \tau_i)$. Additionally the results hold if the end point τ_i^n varies with n. Positivity of neither $W^n(\cdot)$ nor $W(\cdot)$ is necessary. The technical details are not particularly insightful and may be found in O'Sullivan (1986).

3. EFFICIENCY: WEIGHTED KAPLAN-MEIER VERSUS WEIGHTED LOG-RANK STATISTICS

3.1. *Asymptotic Efficacies*

Under a sequence of local alternatives such that, uniformly on $[0, \tau)$

$$\left(\frac{n_1 n_2}{n}\right)^{1/2} \{S_1^n(t) - S_2^n(t)\} \to D(t) \tag{3.1}$$

for some bounded function D, we see that

$$\frac{(n_1 n_2/n)^{1/2}\text{WKM}}{\hat{\sigma}_{\text{WKM}}} \xrightarrow{d} N\left(\frac{\int_0^\tau W(t)\, D(t)\, dt}{\sigma_{\text{WKM}}}, 1\right).$$

Appealing to the notion of Pitman efficiency, we find that the efficacy $e_{\text{WKM}}(W)$ of the test against the local alternatives

$$H_1^n: S_1^n(\cdot) \neq S_2^n(\cdot)$$

satisfying expression (3.1) with limiting survival function $S(\cdot)$ is given by

$$\left(\int_0^\tau W(t)\, D(t)\, \mathrm{d}t\right)^2 \Bigg/ - \int_0^\tau \frac{\left(\int_t^\tau W(u)\, S(u)\, \mathrm{d}u\right)^2}{S^2(t)} \frac{p_1 C_1^-(t) + p_2 C_2^-(t)}{C_1^-(t)\, C_2^-(t)}\, \mathrm{d}S(t). \tag{3.2}$$

Gill (1980) derived a similar expression for the efficacy of a weighted log-rank test with weight function $K(\cdot)$ under a sequence of alternatives satisfying

$$\left(\frac{n_1 n_2}{n}\right)^{1/2} \{\lambda_1^n(t) - \lambda_2^n(t)\} \to L(t) \tag{3.3}$$

where $\lambda_i(\cdot)$ denotes the hazard function, $i = 1, 2$. Briefly, the numerator of the weighted log-rank statistic LGK(K) can be written as the sum over distinct death times t_j of {(the observed number of deaths in sample 1 at t_j) − (the proportion of the total number alive and uncensored at t_j which are in sample 1)} × $K(t_j)$ (or an estimate of $K(t_j)$), and the efficacy of the log-rank test is given by

$$e_{\mathrm{LGK}}(K) = \frac{\left\{\int_0^\tau K(t)\, S(t) \frac{C_1^-(t)\, C_2^-(t)}{p_1 C_1^-(t) + p_2 C_2^-(t)} L(t)\, \mathrm{d}t\right\}^2}{\int_0^\tau K^2(t)\, S(t)\, \frac{C_1^-(t)\, C_2^-(t)}{p_1 C_1^-(t) + p_2 C_2^-(t)} \lambda(t)\, \mathrm{d}t} \tag{3.4}$$

where $\lambda(\cdot)$ denotes the limiting hazard function − $\mathrm{d}\{\log S(t)\}/\mathrm{d}t$.

3.2. *Generalized Two-sample z-test*

In any particular study the weight function $\hat{W}$ should be chosen on the basis of the kind of alternative hypothesis which *a priori* seems most likely or interesting (Pepe and Fleming, 1989). It should also satisfy the regularity conditions of Section 2. A simple weight function which satisfies these regularity conditions is

$$\hat{W}_c(t) = \frac{\hat{C}_1^-(t)\, \hat{C}_2^-(t)}{(n_1/n)\hat{C}_1^-(t) + (n_2/n)\hat{C}_2^-(t)}.$$

Since $\hat{W}_c(t) = 1$ in uncensored data the test can be regarded as a generalization to censored data of the two-sample z-test (or t-test) for a difference in means. It is nonparametric in the sense that the statistic is asymptotically stable without regard to the underlying survival or censoring distributions. Thus its performance relative to the nonparametric log-rank statistic is of interest.

Small sample properties were studied in Pepe and Fleming (1989). We have calculated the test's asymptotic efficiency relative to that of the log-rank test under the various stochastic ordering alternatives displayed in Fig. 1. These include proportional hazards and crossing hazards configurations with uniform and no censoring for piecewise exponential survival distributions. The log-rank test, which is asymptotically most powerful for all proportional hazards alternatives with equal censoring distributions, is seen to be only marginally better than the generalized z-test when the base-line survival distribution is exponential and censoring is uniform. In fact, in uncensored data for the classical exponential scale family, the z-test is seen to be fully

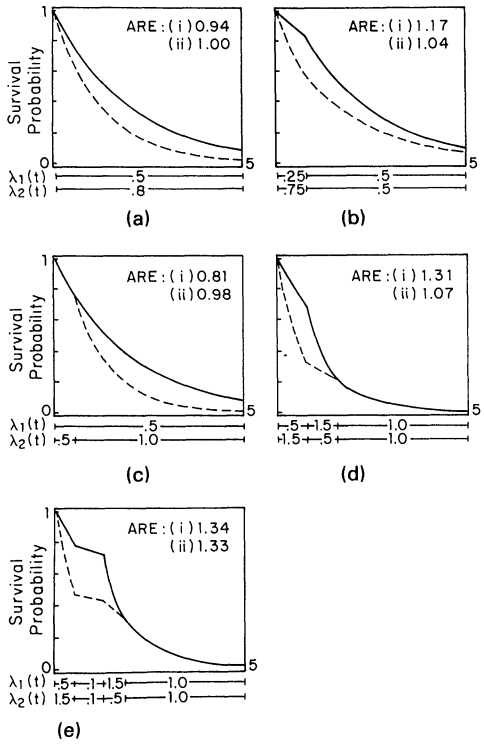

Fig. 1. Asymptotic relative efficiencies (AREs) of WKM(W_c) relative to the log-rank test under (i) uniform (0, 2) censoring and (ii) no censoring (the plots display the survival curves which are piecewise exponential with hazard rates $\lambda_i(t)$ for groups 1 and 2; the bottom axis is time t): (a) exponential scale (proportional hazards); (b) early hazard differences; (c) late hazard differences; (d) crossing hazards; (e) crossing hazards

efficient. Under the crossing hazards alternatives which preserve stochastic ordering considered here the generalized z-test is more efficient than the log-rank test. Indeed, this is not surprising since the log-rank test is known to perform poorly in such cases. Note that the asymptotic relative efficiency of the two statistics will, of course, depend on the base-line survival distribution S, and that in very heavy-tailed data rather poor efficiency of the generalized z-test would be expected. Censored data, however, are rarely heavy tailed.

3.3. *Optimal Weight Functions*

We now turn to the task of finding the optimal statistic within the WKM class for a sequence of local alternatives. Similarly we shall find the optimal weighted log-rank statistic. Although there is no guarantee that the optimality properties extend beyond their respective classes, the asymptotic efficiency of the optimal WKM statistic relative to the optimal weighted log-rank statistic will serve to compare the two classes with regard to efficiency.

The essential step is to maximize the efficacy expressions (3.2) and (3.4) with respect to the weight functions W and K respectively. In both $e_{\mathrm{WKM}}(W)$ and $e_{\mathrm{LGK}}(K)$ the efficacy expression is invariant under a scalar multiplication of the weight

function. Thus a constraint needs to be imposed to optimize the efficacy with respect to the weight function. A natural constraint is to require that the denominator component be unity. By the Cauchy–Schwartz inequality

$$e_{\mathrm{LGK}}(K) = \int_0^\tau K(t)\lambda(t)^{1/2}\, L(t)\frac{1}{\lambda(t)^{1/2}}\, S(t)\frac{C_1^-(t)\, C_2^-(t)}{p_1\, C_1^-(t)+p_2\, C_2^-(t)}\, \mathrm{d}t$$

$$\leqslant \int_0^\tau K^2(t)\, \lambda(t)\, S(t)\frac{C_1^-(t)\, C_2^-(t)}{p_1\, C_1^-(t)+p_2\, C_2^-(t)}\, \mathrm{d}t$$

$$\times \int_0^\tau \frac{L^2(t)}{\lambda(t)}\, S(t)\frac{C_1^-(t)\, C_2^-(t)}{p_1\, C_1^-(t)+p_2\, C_2^-(t)}\, \mathrm{d}t$$

which is independent of $K(\cdot)$ since the first component is unity. The inequality is an equality if and only if $K(t)\lambda(t)^{1/2} = L(t)/\lambda(t)^{1/2}$, so that the optimal weight function is $K_{\mathrm{opt}}(t) = L(t)/\lambda(t)$ and the maximal efficacy is given by

$$e_{\mathrm{LGK}}(K_{\mathrm{opt}}) = \int_0^\tau \frac{L^2(t)}{\lambda(t)}\, S(t)\, \frac{C_1^-(t)\, C_2^-(t)}{p_1\, C_1^-(t)+p_2\, C_2^-(t)}\, \mathrm{d}t. \tag{3.5}$$

This result was derived by Gill (1980) using the Lagrange multiplier method. This method is not directly useful for maximizing $e_{\mathrm{WKM}}(W)$ since the denominator component is a complicated function of $W(\cdot)$. Instead we transform the problem to that of maximizing $e_{\mathrm{WKM}}(W)$ with respect to

$$h(\cdot) = \int_\cdot^\tau W(t)\, S(t)\, \mathrm{d}t,$$

in which case we can rewrite the efficacy as

$$\left\{\int_0^\tau \frac{D(t)}{S(t)}\, \mathrm{d}h(t)\right\}^2 \Big/ \int_0^\tau \frac{h(t)^2}{S(t)}\, \frac{p_1\, C_1^-(t)+p_2\, C_2^-(t)}{C_1^-(t)\, C_2^-(t)}\, \lambda(t)\, \mathrm{d}t. \tag{3.6}$$

On noting that the numerator

$$\left\{\int_0^\tau \frac{D(t)}{S(t)}\, \mathrm{d}h(t)\right\}^2 = \left[\int_0^\tau h(t)\, \frac{\mathrm{d}}{\mathrm{d}t}\left\{\frac{D(t)}{S(t)}\right\}\mathrm{d}t\right]^2$$

we use the same application of the Cauchy–Schwartz inequality to yield the optimal $h(\cdot)$, the corresponding optimal weight function W_{opt} and the maximal efficacy as follows:

$$h_{\mathrm{opt}}(t) = \frac{S(t)}{\lambda(t)}\, \frac{C_1^-(t)\, C_2^-(t)}{p_1\, C_1^-(t)+p_2\, C_2^-(t)}\, \frac{\mathrm{d}}{\mathrm{d}t}\left\{\frac{D(t)}{S(t)}\right\};$$

$$W_{\mathrm{opt}}(t) = \frac{1}{S(t)}\, \frac{\mathrm{d}}{\mathrm{d}t}\, h_{\mathrm{opt}}(t);$$

$$e_{\mathrm{WKM}}(W_{\mathrm{opt}}) = \int_0^\tau \frac{S(t)}{\lambda(t)}\, \frac{C_1^-(t)\, C_2^-(t)}{p_1\, C_1^-(t)+p_2\, C_2^-(t)}\left[\frac{\mathrm{d}}{\mathrm{d}t}\left\{\frac{D(t)}{S(t)}\right\}\right]^2 \mathrm{d}t. \tag{3.7}$$

The required regularity conditions have been omitted to simplify the presentation, though obviously differentiability of the components of h_{opt} is required. In particular, this requires continuity of the censoring distributions.

If

$$L(t) = \frac{\mathrm{d}}{\mathrm{d}t}\left\{\frac{D(t)}{S(t)}\right\}$$

then the maximal efficacies (3.5) and (3.7) are the same. This is the case for the classical types of local alternatives where the nth alternative is parameterized by $\alpha_n = a_0(n/n_1n_2)^{1/2}$ for some constant a_0, with $S_1^n(t) = S(\alpha_n, t)$ and $S_2^n(t) = S(0, t)$. To see this note that

$$D(t) = \lim_{n\to\infty}\left(\frac{n_1n_2}{n}\right)^{1/2}\{S_1^n(t) - S_2^n(t)\} = a_0\left\{\frac{\mathrm{d}S}{\mathrm{d}\alpha}(\alpha, t)\right\}_{\alpha=0};$$

interchanging the order to differentiation yields

$$\frac{\mathrm{d}}{\mathrm{d}t}\left\{\frac{D(t)}{S(t)}\right\} = a_0\frac{\mathrm{d}}{\mathrm{d}\alpha}\left\{\frac{\mathrm{d}}{\mathrm{d}t}\log S(\alpha, t)\right\}_{\alpha=0} = a_0\frac{\mathrm{d}\lambda}{\mathrm{d}\alpha}(\alpha, t) = L(t).$$

Although the maximal efficacies $e_{\text{LGK}}(K_{\text{opt}})$ and $e_{\text{WKM}}(W_{\text{opt}})$ are always equal for the standard parameterization of local alternatives, recall that the classes of statistics are essentially different. The weighted log-rank statistics are generalized rank statistics, whereas the WKM statistics are generalized location test statistics.

4. EXTENSIONS

4.1. *Stratified Statistics*

Possible confounding by factors related to the outcome can be adjusted for through stratification. If the direction of the difference in survival in the two groups is thought to be the same within each of the k_0 strata, then a natural stratified WKM statistic is

$$\text{WKM}_s = \sum_{k=1}^{k_0}\frac{1}{\sqrt{k_0}}\left(\frac{n_1^kn_2^k}{n^k}\right)^{1/2}\int_0^{T_c^k}\hat{W}^k(t)\{\hat{S}_1^k(t) - \hat{S}_2^k(t)\}\,\mathrm{d}t/\hat{\sigma}_{\text{WKM}}^k$$

where the superscript k denotes the stratum. $\hat{W}^k$ may or may not be stratum specific. If the two-sided alternative H_1: $S_1^k(\cdot) \neq S_2^k(\cdot)$ for some $k = 1, \ldots, k_0$ is of interest, the components of the summand may be squared and the statistic compared with a $\chi^2_{k_0}$ distribution instead. Asymptotic theory requires that the conditions of Section 2 hold within each stratum. How small the number of observations can be within each stratum to use the large sample results validly has yet to be explored.

4.2. *k_0-sample Trend Statistics*

Suppose that we wish to test whether or not a discrete covariate V with k_0 levels is associated with survival. A test for trend might be based on a comparison of the standard normal distribution with a standardized version of

$$\mathrm{WKM}_V = \sum_{k=1}^{k_0} \left[n^k(V^k - \bar{V}) \Big/ \left\{ \sum_{k=1}^{k_0} n^k(V^k - \bar{V})^2 \right\}^{1/2} \right] \int_0^{T_c^k} \hat{W}^k(t)\{\hat{S}^k(t) - \hat{S}(t)\}\, \mathrm{d}t.$$

Here, a superscript k indicates the kth level of the covariate, $\bar{V}$ is the average value of the covariate in the sample and $\hat{S}(\cdot)$ is the Kaplan–Meier estimator calculated from the total sample. Rigorous large sample theory can be found in O'Sullivan (1986). In essence, the conditions of Section 2 are required to hold at each level of the covariate. Under the null hypothesis that $S^k(\cdot) = S(\cdot)$, $k = 1, \ldots, k_0$, WKM_V is asymptotically normal with mean zero and variance

$$\sigma^2_V = \sum_{k=1}^{k_0} p^k \frac{\{V^k - E(V)\}^2}{\sum_{k=1}^{k_0} p^k\{V^k - E(V)\}^2} \int_0^{\tau^k} \frac{\left\{ \int_t^{\tau^k} W^k(u)\, S(u)\, \mathrm{d}u \right\}^2}{S^2(t)\, C^{k^-}(t)}\, \mathrm{d}S(t).$$

Large positive values of WKM_V will indicate a positive association between increasing values of the covariate and increasing survival.

Interestingly, in the same way that $\mathrm{WKM}(W_c)$ reduces to the familiar two-sample z-test for a difference in means in uncensored data, if $\hat{W}^k(\cdot) = 1$ in uncensored data, WKM_V reduces to the familiar least squares slope estimator from the linear model

$$\text{survival time} = \alpha + \beta V^k + \text{error}.$$

However, regardless of its interpretation as a slope estimator, WKM_V would seem to be a natural k_0-sample trend test statistic.

5. CONCLUDING REMARKS

The results presented in this paper are sufficiently general to incorporate right censoring but those in Section 3 will be of particular interest for classical uncensored data. In that context, efficacy expressions have been derived for tests based on differences in L-statistics and the optimal such statistic within the family specified by $D(\cdot)$ and $S(\cdot)$ has been obtained. Moreover, the optimal L-statistic within this family is as efficient as the optimal linear rank statistic. As a special case, we have seen that the two-sample z-test for a difference in means is as efficient as the Savage exponential scores test for the exponential scale alternatives.

The class of WKM test statistics adds to the already rich arsenal of two-sample nonparametric tests for censored data. This arsenal includes generalizations of the Kolmogorov–Smirnov test (Fleming *et al.*, 1980; Schumacher, 1984), the Cramér–von Mises test (Schumacher, 1984), the median test (Slud *et al.*, 1984), the Breslow acceleration test (Breslow *et al.*, 1984) and the supremum weighted log-rank tests (Fleming *et al.*, 1987), in addition to the popular generalized linear rank tests (Tarone and Ware, 1977; Harrington and Fleming, 1982; Gill, 1980).

Do we need yet another class of test statistics for this problem? Which of the tests is most appropriate for a particular application depends on the scientific question being addressed, a fact of which many naïve users are not aware. For example, if a particularly large difference in the probability of survival to a specific time point t_0 is of interest then a test based on $\hat{S}_1(t_0) - \hat{S}_2(t_0)$ directly addresses the question of interest. If a difference in the median survival is of interest then the median test is appropriate. If

the concern is with the hazard ratio which is assumed to be constant over time then the log-rank test is most appropriate. Yet, particularly among physicians involved with clinical trials, only the log-rank test is commonly quoted without regard to their scientific question concerning survival.

A WKM statistic with positive weight function directly addresses the stochastic ordering alternative H_1: $S_1(\cdot) \geqslant S_2(\cdot)$, $S_1 \neq S_2$. This is not addressed directly by the other test procedures mentioned here (Pepe and Fleming, 1989). In addition WKM tests are location tests and therefore are not solely based on the ranks of the data. All the other aforementioned procedures (except for the median test) are rank based and hence they inherently do not incorporate information on the size of the difference in survival on the timescale. The magnitude of the difference in survival time will be of particular interest in some applications and is an integral part of the WKM statistic because it is a location test statistic. Finally, the WKM class includes generalizations to censored data of the classical two-sample z-test for a difference in means which will be of interest to some researchers.

ACKNOWLEDGEMENTS

This work was supported by the National Cancer Institute grant CA-32693 and by grant GM-24472 from the National Institutes of Health.

APPENDIX A

The following result is basic to the derivation of asymptotic distribution theory for WKM statistics and was proven by Gill (1983) in the special case of a fixed configuration. The subscript i is dropped here since the result pertains to a single sample.

Theorem 1. Let h^n, $n=1, 2, \ldots$, and h be non-negative, non-increasing continuous functions on $[0, \tau]$ such that

$$\lim_{n\to\infty}\left[\int_0^\tau |\mathrm{d}\{h^n(v)-h(v)\}|\right] = 0$$

and

$$\lim_{n\to\infty}(\sigma_n^2) = \lim_{n\to\infty}\left\{-\int_0^\tau \frac{h^n(v)^2}{S^n(v)^2\, C^{n^-}(v)}\,\mathrm{d}S^n(v)\right\} = -\int_0^\tau \frac{h(v)^2}{S(v)^2\, C^-(v)}\,\mathrm{d}S(v) = \sigma^2 < \infty.$$

If $Z^n(t) = -\sqrt{n}\{\hat{S}(t)-S^n(t)\}/S^n(t)$ and $Z^\infty(t)$ is a Brownian motion with variance function

$$-\int_0^\tau \frac{\mathrm{d}S(v)}{S(v)^2\, C^-(v)}$$

then

$$h^n(t\wedge T)\, Z^n(t\wedge T) \Rightarrow h(t\wedge\tau)\, Z^\infty(t\wedge\tau), \tag{A.1}$$

$$\int_0^{t\wedge T} h^n(v)\,\mathrm{d}Z^n(v) \Rightarrow \int_0^{t\wedge\tau} h(v)\,\mathrm{d}Z^\infty(v) \tag{A.2}$$

and

$$\int_0^{t\wedge T} Z^n(v)\, \mathrm{d}h^n(v) \Rightarrow \int_0^{t\wedge\tau} Z^\infty(v)\, \mathrm{d}h(v), \tag{A.3}$$

where $\Rightarrow$ denotes weak convergence in $D[0, \infty]$ (Billingsley, 1968), $t\wedge T = \min(t, T)$ and by definition (Gill, 1983)

$$\int_0^s h(v)\, \mathrm{d}Z^\infty(v) = h(s)\, Z^\infty(s) - \int_0^s Z^\infty(v)\, \mathrm{d}h(v).$$

Proof. The proof follows that of Gill (1983) quite closely. In essence weak convergence of $Z^n(\cdot)$ to $Z^\infty(\cdot)$ on $[0, \tau)$ follows from the fact that $Z^n(\cdot)$ is a martingale. Therefore weak convergence of functionals of $Z^n(\cdot)$ on $[0, \tau)$ also follows. The finite asymptotic variance σ^2 ensures the existence of the limiting processes at τ and tightness at the end point τ is guaranteed by the conditions on h^n and h. The result then follows by applying theorem 4.2 of Billingsley (1968). □

REFERENCES

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

Breslow, N. E., Edler, L. and Berger, J. (1984) A two-sample censored data rank test for acceleration. *Biometrics*, **40**, 1049–1062.

Fleming, T. R., Harrington, D. P. and O'Sullivan, M. (1987) Supremum versions of the log rank and generalized Wilcoxon statistics. *J. Am. Statist. Ass.*, **82**, 312–320.

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

Gill, R. D. (1980) Censoring and stochastic integrals. *Mathematical Centre Tracts 124*. Amsterdam: Mathematical Centre.

——— (1983) Large sample behavior of the product-limit estimator on the whole line. *Ann. Statist.*, **11**, 49–58.

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

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

O'Sullivan, M. P. (1986) A new class of statistics for the two-sample survival analysis problem. *PhD Thesis*. University of Washington, Seattle.

Pepe, M. S. and Fleming, T. R. (1989) Weighted Kaplan–Meier statistics: a class of distance tests for censored survival data. *Biometrics*, **45**, 497–507.

Schumacher, M. (1984) Two-sample tests of Cramér–von Mises and Kolmogorov–Smirnov type for randomly censored data. *Int. Statist. Rev.*, **52**, 263–281.

Slud, E. V., Byar, D. P. and Green, S. B. (1984) A comparison of reflected versus test based confidence intervals for the median survival time based on censored data. *Biometrics*, **40**, 587–600.

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