Statistical Science
2004, Vol. 19, No. 4, 615–623
DOI 10.1214/088342304000000530

Nonparametric Survival Analysis

Michael G. Akritas

Abstract. Some classes of nonparametric procedures with randomly right-censored data are presented. They include procedures for analysis of variance and analysis of covariance designs with independent and dependent ordinal (continuous and discrete) data.

Key words and phrases: Censored data, product-limit estimator, k-sample problem, analysis of variance, analysis of covariance, dependent data.

1. INTRODUCTION

Methods for the analysis of data on an event observed over time and the study of factors associated with the occurrence rates of this event fall under the heading *survival analysis*. For example, in a study of survival rates for cancer patients, the event of interest may be death. Time to event data is different because it is often incomplete. Incompleteness due to the fact that the time for which a subject was under observation is less than the time to the event of interest is called (*right*) *censoring*. The focus here is censoring and, in particular, the type of censoring called *random*. For example, patients in a survival study may be lost to followup (e.g., due to transfer to a nonparticipating institution) or are admitted after the study began (staggered entries). With censored data it is not obvious how to estimate such standard quantities as the mean and variance. Thus different methods need to be developed. The different approaches can be classified as parametric, semiparametric, distribution-free and fully nonparametric (NP).

The term "survival analysis" derives from the historical development of the field. John Graunt's 1662 book *Natural and Political Observations upon the Bill of Mortality*, which classified registered deaths by age, period, gender and cause of death, suggested for the first time that death be regarded as an event which deserves systematic study. Some years later, Edmund Halley devised the first life table, very similar to those still in use today in demographic and actuarial studies, and Greenwood (1926) provided a variance formula for the life table estimator. Broadening the term survival analysis to include data on any event observed over time, not just death or failure, came with the use of such methods in clinical trials and the social sciences, where events such as disease progression or metastasis or first employment after formal education are also of interest.

The seminal paper by Kaplan and Meier (1958) marked a big breakthrough in survival analysis, especially from the NP point of view. It allowed the use of descriptive statistics and fueled the development of all existing NP approaches with censored data. In this article we give a unified presentation of the fully NP approach, concentrating on the analysis of factorial designs. Starting with the well studied one-, two- and k-sample problems, we present generalizations to multifactor designs as well as designs with continuous covariates. Dependent data arising from repeated measures designs are also discussed. Thus, NP approaches to regression problems, including regression curve estimation, which is based entirely on smoothing techniques with no inferential component, and the area of lack-of-fit testing are not included. A fairly complete review (also for truncated data), which includes semiparametric methods, can be found in Akritas and LaValley (1997).

In what follows we use the following notation. For any right-continuous function A, denote the left-hand limit of A at s by $A_-(s) = A(s-)$, denote the "jump" at s by $\Delta A(s) = A(s) - A_-(s)$ and denote the continuous component as $A_c(t) = A(t) - \sum_{s \le t} \Delta A(s)$. If A is nondecreasing, the inverse of A is defined to be $A^{-1}(t) = \inf\{s; A(s) \ge t\}$.

Michael G. Akritas is Professor, Department of Statistics, Pennsylvania State University, University Park, Pennsylvania 16802, USA (e-mail: mga@stat.psu.edu). He is also affiliated with National Technical University of Athens, Athens, Greece.

2. THE ONE-SAMPLE PROBLEM

2.1 Survival and Cumulative Hazard Functions

The *survival function* S and the *cumulative hazard function* Λ that correspond to a distribution function F are defined, respectively, by $S = 1 - F$ and

$$\Lambda(t) = \int_0^t \big(1 - F_-(s)\big)^{-1} dF(s). \tag{2.1}$$

Knowledge of Λ also implies knowledge of F. In fact, it is easy to verify that in the discrete case, $S(t) = \prod_{s \le t}(1 - \Delta\Lambda(s))$, and in the continuous case, $S(t) = \exp(-\Lambda(t))$. In the general case, the two formulas can be combined as

$$S(t) = \prod_{s \le t} \big(1 - \Delta\Lambda(s)\big) \exp(-\Lambda_c(t)). \tag{2.2}$$

The combined formula (2.2) is a consequence of $F(t) = \int_0^t (1 - F_-(s))\, d\Lambda(s)$, and another result stated and proved as Proposition A.4.1 in Gill (1980).

2.2 The Random Censoring Model

A censored data set consists of n independent realizations of the random vector (Z, Δ); thus to subject i there corresponds the random vector (Z_i, Δ_i) for $i = 1, \ldots, n$. The variable Z_i denotes the time for which subject i is under observation. If at the end of the observation period we have occurrence of the event of interest, Z_i is called *uncensored*; otherwise it is called *censored*. The variable Δ_i takes the value 1 if the observation on subject i is uncensored and takes the value 0 if the observation is censored. The random censoring model uses a variable C_i, called the *censoring variable*, for the maximum time subject i can be observed, and assumes that C_i is independent of the *time to the event variable* Y_i. Clearly,

$$Z_i = Y_i \wedge C_i, \quad \Delta_i = I(Z_i = Y_i), \tag{2.3}$$

where $a \wedge b = \min(a, b)$ for any two real numbers a, b and $I(A)$ denotes the indicator function of the event A. The distribution function $F(t) = P(Y \le t)$ of Y is of primary interest, while that of C, $G(t) = P(C \le t)$, is considered to be an unknown nuisance parameter.

Let $H(t) = P(Z_i \le t)$ and $H_1(t) = P(Z_i \le t, \Delta_i = 1)$. By the independence of Y and C,

$$\begin{aligned} 1 - H(t) &= \big(1 - F(t)\big)\big(1 - G(t)\big), \\ H_1(t) &= \int_0^t \big(1 - G_-(s)\big)\, dF(s). \end{aligned} \tag{2.4}$$

2.3 The Kaplan–Meier Estimator

The *Kaplan–Meier* estimator (Kaplan and Meier, 1958) was originally derived as an NP maximum likelihood estimator of F and as a limit of the actuarial estimator as the time axis is partitioned into fine intervals. Because of the latter method of derivation, it is also known as the *product-limit* (PL) estimator. The derivation presented here exploits the connection between F and H, H_1, which are directly estimable from the data. Multiplying and dividing the integrand in (2.1) by $1 - G_-$ gives

$$\begin{aligned} \Lambda(t) &= \int_0^t \frac{1 - G_-}{(1 - F_-)(1 - G_-)}\, dF \\ &= \int_0^t \frac{1}{1 - H_-}\, dH_1, \end{aligned} \tag{2.5}$$

where the second equality follows from the two relationships in (2.4). Since F can be obtained from Λ through (2.2), relationship (2.5) effectively solves the system of equations (2.4) for F.

REMARK 2.1. Note that the first equality in (2.5) is valid only if $1 - G_-(t) > 0$. This imposes a natural limit on the range of t values for which $\Lambda(t)$ can be estimated.

Let $\widehat{H}_n(t) = n^{-1} \sum_{i=1}^n I(-\infty < Z_i \le t)$ and $\widehat{H}_{1n}(t) = n^{-1} \sum_{i=1}^n I(-\infty < Z_i \le t, \Delta_i = 1)$ be the empirical estimators of H and H_1. Then, on the basis of (2.5), Λ can be estimated by

$$\widehat{\Lambda}_n(t) = \int_0^t \frac{1}{1 - \widehat{H}_{n-}}\, d\widehat{H}_{1n}. \tag{2.6}$$

The estimator of the cumulative hazard in (2.6) is known as the *Nelson–Aalen estimator*. The Nelson–Aalen estimator and (2.2) yield the Kaplan–Meier or PL estimator

$$\begin{aligned} \widehat{S}_n(t) &= \prod_{s \le t} \big(1 - \Delta\widehat{\Lambda}_n(s)\big) \\ &= \prod_{i\,:\,Z_{(i)} \le t} \left(\frac{n - i}{n - i + 1}\right)^{\Delta_{(i)}}, \end{aligned} \tag{2.7}$$

where $\Delta_{(i)}$ is the Δ that corresponds to the ith ordered observation $Z_{(i)}$. In case of ties, the formula is valid for any integer ranking of the tied observations. We set $\widehat{F}_n = 1 - \widehat{S}_n$. With uncensored data, $\widehat{F}_n$ reduces to the empirical distribution function. Note that the definition of $\widehat{S}_n$ allows it to be strictly positive and constant to the right of the last observed failure, which is the version of the PL estimator suggested by Gill (1980). (Kaplan

and Meier left the estimator undefined for $t > Z_{(n)}$ if $\Delta_{(n)} = 0$.) It can be shown that the PL estimator is biased upward; in particular (see Andersen, Borgan, Gill and Keiding, 1993, page 259),

$$0 \le E\widehat{S}_n(t) - S(t) \le F(t)\big(1 - H(t)\big)^n.$$

In spite of this, $\widehat{S}_n(t)$ is consistent and asymptotically normal, uniformly in $t \in [0, Z_{(n)}]$ (Gill, 1983). Its asymptotic variance is estimated by the Greenwood formula

$$\widehat{\sigma}^2_{\mathrm{GR}} = \widehat{S}^2_n(t) \sum_{Z_{(i)} \le t} \frac{\Delta_{(i)}}{(n-i)(n-i+1)}. \tag{2.8}$$

The standard asymptotic $100(1-\alpha)\%$ confidence interval for $S(t)$ is

$$\widehat{S}_n(t) \pm z_{\alpha/2}\widehat{\sigma}_{\mathrm{GR}}, \tag{2.9}$$

where $z_{\alpha/2}$ is the upper $\alpha/2$ quantile of the standard normal distribution. This interval is not completely satisfactory because it can include values that fall outside the interval $[0, 1]$. This can be remedied by applying the asymptotic normal distribution to a transformation of $S(t)$ (Thomas and Grunkemeier, 1975). Possible transformations include $g(x) = \log(-\log x)$, $g(x) = \arcsin\sqrt{x}$ and $g(x) = \log(x/(1-x))$. For example, the first of these transformations gives an asymptotic $100(1-\alpha)\%$ confidence interval for $S(t)$ of

$$\widehat{S}_n(t) \exp\big\{\pm z_{\alpha/2}\widehat{\sigma}_{\mathrm{GR}}/[\widehat{S}_n(t)\log\widehat{S}_n(t)]\big\}, \tag{2.10}$$

which takes values in $[0, 1]$. Borgan and Liestøl (1990) indicated that such confidence intervals are quite satisfactory for sample sizes as low as 25, even with 50% censoring.

With censored data, sample quantiles are more commonly used as descriptive statistics than sample moments. This is because under right censoring there is often incomplete information on the right tail of the distribution. The pth quantile F is $\xi_p = F^{-1}(p)$, which is estimated by $\widehat{\xi}_p = \widehat{F}_n^{-1}(p)$. The asymptotic variance of $\widehat{\xi}_p$ is estimated by

$$\frac{(1-p)^2}{\widehat{f}^{\,2}(\widehat{\xi}(p))} \sum_{X_{(i)} \le t} \frac{\Delta_{(i)}}{(n-i)(n-i+1)}, \tag{2.11}$$

where $\widehat{f}$ is an estimator of the density function. Confidence intervals for quantiles that do not require estimation of the density function are obtained by inverting the sign test (Brookmeyer and Crowley, 1982). Adapting this idea to a transformed version of the PL estimator, such a $100(1-\alpha)\%$ confidence interval consists of all values ξ_p^0 which satisfy

$$\frac{|g(\widehat{S}(\xi_p^0) - g(1-p)|}{|g'(\widehat{S}(\xi_p^0))|\widehat{\sigma}_{\mathrm{GR}}(\xi_p^0)} \le z_{\alpha/2},$$

where g is a transformation, such as those mentioned above. Using $g(x) = x$ gives the confidence intervals of Brookmeyer and Crowley (1982). This interval can be read directly from the lower and upper pointwise confidence limits for the survival distribution, just as $\widehat{\xi}_p$ can be read from the Kaplan–Meier curve. This graphical approach to confidence intervals for quantiles was described by Lawless (1982).

COMMENT. Hall and Wellner (1980) constructed confidence bands for the survival function. A bootstrap version of these bands was given by Akritas (1986). Confidence bands provide one of several methods for model validation. For other approaches, see Akritas (1988), Hjort (1990), Hollander and Peña (1992), Kim (1993) and Li and Doss (1993).

3. TWO- AND k-SAMPLE PROBLEMS

Here we briefly describe the most commonly used statistics for testing equality of k treatments based on independent randomly censored samples. These statistics can be obtained by the heuristic arguments of Mantel (1966), the weighted log-rank statistics of Tarone and Ware (1977) and Gill's (1980) general class $\mathcal{K}$ of tests. In particular, Mantel (1966) considered the data as a series of $k \times 2$ tables at each of the distinct failure times, applied the Mantel–Haenszel test for contingency tables and combined the tables as if they were independent. The resulting log-rank test is based on the sum of the vectors of observed minus expected frequencies for each of the $k \times 2$ tables, $LR = \sum_{\ell=1}^{L}(D_{1\ell} - E(D_{1\ell}), \ldots, D_{k\ell} - E(D_{k\ell}))$, where L is the total number of distinct failure times, $D_{i\ell}$ is the number of failures from sample i at the ℓth failure time and $E(D_{i\ell})$ is its expected value under the null hypothesis, conditionally on the risk sets from each of the samples and on the total number of failures. Evaluation of $E(D_{i\ell})$ and some algebra yield the form

$$LR = \sum_{i=1}^{k}\Big(\int_0^\infty K_{1i}(s)\,d\big(\widehat{\Lambda}_1(s) - \widehat{\Lambda}_i(s)\big), \ldots, \int_0^\infty K_{ki}(s)\,d\big(\widehat{\Lambda}_k(s) - \widehat{\Lambda}_i(s)\big)\Big) \tag{3.1}$$

with $K_{ij}(s) = (Y_1.(s)Y_i.(s))/Y..(s)$, where $Y_i.(s)$ is the number at risk from sample i at time $s-$, $Y..(s)$ is

the total number at risk at time $s-$ and $\widehat{\Lambda}_i$ is the Nelson–Aalen estimator of the cumulative hazard function from sample i. Gill's class $\mathcal{K}$ of statistics is (3.1) with $K_{ij}(s)$ any bounded, nonnegative and predictable function with the property that $Y_{i\cdot}(s)Y_{j\cdot}(s) = 0$ implies $K_{ij}(s) = 0$. The commonly used weights are of the form

$$K_{ij}(s) = W(s)\frac{Y_{i\cdot}(s)Y_{j\cdot}(s)}{Y_{\cdot\cdot}(s)}. \tag{3.2}$$

Statistics based on (3.1) but with $K_{ij}(s)$ in (3.2) are called *weighted log-rank* statistics, since $W(s) = 1$ gives the log-rank statistic. In particular, the Gehan weights are $W(s) = Y_{\cdot\cdot}(s)$, the Peto–Prentice weights are $W(s) = \widehat{S}(s-)$ with $\widehat{S}$ the PL estimator from the combined sample and the Fleming–Harrington weights are $W(s) = \widehat{S}(s-)^\rho$, $\rho \geq 0$.

COMMENTS. 1. Scheffé-type multiple comparisons were described by Akritas (1992).

2. For a two-sample test procedure, closer in spirit to the NP procedures for analysis of variance (ANOVA; though only for continuous data), see Pepe and Fleming (1991).

3. Gehan's test (even with its Tarone–Ware variance) is sensitive to different censoring patterns in the two samples. Thus, one of the other tests is preferred with small and moderate sample sizes (see Lawless, 1982, page 423, or Andersen et al., 1993, page 350).

4. Proc Lifetest in SAS performs the log-rank and Gehan tests. Program 1L in BMDP gives the log-rank, Gehan, Tarone–Ware and the Peto–Prentice tests. S-Plus implements the family of tests from Fleming and Harrington (1981) in the function surv.diff and gives the log-rank test as the default.

4. NONPARAMETRIC METHODS: ANALYSIS OF VARIANCE AND ANALYSIS OF COVARIANCE

4.1 Motivation

The aforementioned elegant approach, based on ideas of Mantel (1966), Tarone and Ware (1977) and Gill (1980) for deriving test procedures for two and k samples, does not extend to factorial designs. Consider, for example, an $a \times b$ factorial design, so there are a total of $k = ab$ populations determined by all factor-level combinations. As before, let $D_{(rc),\ell}$ denote the number of failures from group (r, c) at the ℓth failure time from the combined samples. To continue with Mantel's heuristic argument, we need to obtain $E(D_{(rc),\ell})$. While this is possible under the one-way hypothesis (i.e., all $k = ab$ populations are identical), it is impossible to do it in a completely NP way under the specialized hypotheses that are of interest in two-way designs (i.e., no main effects and no interaction). Thus, the analysis of factorial designs is commonly carried out under the assumption of proportional hazards (Cox, 1972). Though widely used, proportional hazards methods perform inference on parameters which lose their interpretation when the proportional hazards assumption is violated. In this section we describe test procedures for ANOVA and analysis of covariance (ANCOVA) designs, developed by Akritas and Brunner (1997) and by Du, Akritas and Van Keilegom (2003), respectively, which use fully nonparametric procedures. Codes that implement these procedures are available from the aforementioned authors.

4.2 Nonparametric Models for ANOVA and ANCOVA

We present the NP models for two-way ANOVA and one-way ANCOVA as the simplest designs where all features of NP modeling can be appreciated. To include all ordinal (continuous and discrete) data in the formulation, all distribution functions (so also conditional ones) are taken as the average of their left- and right-continuous versions.

For the two-way ANOVA design, the time to the event of interest is denoted by Y_{ijk}, $i = 1, \ldots, I$, $j = 1, \ldots, J$, $k = 1, \ldots, n_{ij}$. The observations follow the random censoring model of Section 2.2; thus, (Z_{ijk}, Δ_{ijk}), where $Z_{ijk} = Y_{ijk} \wedge C_{ijk}$, with C_{ijk} being the censoring variable in cell (i, j), and $\Delta_{ijk} = I(Y_{ijk} = Z_{ijk})$.

For the one-way ANCOVA design, the covariate and time to the event of interest are denoted by (X_{ij}, Y_{ij}), $i = 1, \ldots, I$, $j = 1, \ldots, n_i$. Thus, i enumerates the factor levels and j denotes the observations within each factor level. The observations are $(X_{ij}, Z_{ij}, \Delta_{ij})$, $j = 1, \ldots, n_i$, where $Z_{ij} = \min(Y_{ij}, C_{ij})$ and $\Delta_{ij} = I(Y_{ij} = Z_{ij})$, where the censoring variable C_{ij} is conditionally independent of Y_{ij} given X_{ij}.

The NP model for the two-way ANOVA design specifies only that

$$Y_{ijk} \sim F_{ij} \tag{4.1}$$

for some distribution function F_{ij} (Akritas and Arnold, 1994).

The NP model for the one-way ANCOVA design specifies only

$$Y_{ij}|X_{ij} = x \sim F_{ix}, \tag{4.2}$$

that is, that conditionally on $X_{ij} = x$, Y_{ij} has a distribution function that depends on i and x (Akritas, Arnold and Du, 2000). Note that models (4.1) and (4.2) do not specify how the response distribution changes when the levels or covariate value changes. Thus they are completely nonparametric (also nonlinear and nonadditive).

For the two-way ANOVA design set $\overline{F}_{i\cdot}(y) = J^{-1} \cdot \sum_j F_{ij}(y)$ and $\overline{F}_{\cdot j}(y) = I^{-1} \sum_i F_{ij}(y)$. For the one-way ANCOVA design, choose a distribution function $G(x)$ and let

$$\overline{F}_{i\cdot}^{G}(y) = \int F_{ix}(y)\, dG(x),$$
$$\overline{F}_{\cdot x}(y) = \frac{1}{I} \sum_{i=1}^{I} F_{ix}(y). \tag{4.3}$$

If X_{ij} are a random sample, G can be taken as their overall distribution function. Thus, if the covariate has the same distribution in all groups, $\overline{F}_{i\cdot}^{G}(y)$ is the marginal distribution function of Y_{ij}. The hypotheses of interest in model (4.1) or (4.2) follow:

1. The $\overline{F}_{i\cdot}(y)$ do not depend on i or $\overline{F}_{i\cdot}^{G}$ do not depend on i (no main effect).
2. The $\overline{F}_{\cdot j}(y)$ do not depend on j or $\overline{F}_{\cdot x}(y)$ do not depend on x (no main effect).
3. The $F_{ij}(y) = \overline{F}_{i\cdot}(y) + K_j(y)$ or $F_{ix}(y) = \overline{F}_{i\cdot}^{G}(y) + K_x(y)$ (no interaction).
4. The term $F_{ij}(y)$ is independent of i or $F_{ix}(y)$ is independent of i (no simple effect).
5. The term $F_{ij}(y)$ is independent of j or $F_{ix}(y)$ is independent of x (no simple effect).

For the ANCOVA setting, the first hypothesis is sensible even when the model is not additive (i.e., unequal slopes in the classical case), while the third and fourth hypotheses correspond to those for parallelism and equality of regression curves. An important advantage of the nonparametric hypotheses and test procedures is that they are unchanged by monotone transformations in the response. In the classical model, such transformations are often necessary to linearize the expectation and/or equalize the variances.

The above hypotheses can also be described in terms of corresponding NP effects being zero. These NP effects are defined from decompositions of F_{ij} and F_{ix}. The decomposition of F_{ij} (Akritas and Arnold, 1994) is

$$F_{ij}(y) = M(y) + A_i(y) + B_j(y) + C_{ij}(y), \tag{4.4}$$

where $M = \overline{F}_{\cdot\cdot}$, $A_i = \overline{F}_{i\cdot} - M$, $B_j = \overline{F}_{\cdot j} - M$ and $C_{ij} = F_{ij} - \overline{F}_{i\cdot} - \overline{F}_{\cdot j} + M$. The quantities A_i, B_j and C_{ij} are, respectively, the NP main row, main column and interaction effects. The decomposition of F_{ix} (Akritas, Arnold and Du, 2000) is

$$F_{ix}(y) = M^G(y) + A_i^G(y) + B_x^G(y) + C_{ix}^G(y), \tag{4.5}$$

where $M^G(y) = I^{-1} \sum_{i=1}^{I} \int F_{ix}(y)\, dG(x)$, $A_i^G(y) = \overline{F}_{i\cdot}^{G}(y) - M(y)$, $B_x^G(y) = \overline{F}_{\cdot x}(y) - M^G(y)$ and $C_{ix}^G(y) = F_{ix}(y) - M^G(y) - A_i^G(y) - B_x^G(y)$ are, respectively, the NP main factor, main covariate, and interaction effects.

It can be shown that the NP hypotheses imply, but are not implied by, their parametric counterparts. For example, the NP hypothesis of no interaction is equivalent to the statement that the mean of any monotone transformation of the response can be decomposed in an additive fashion. This strong form of additivity captures the *substantive* meaning of no interaction between factors as a scientist might think of it.

4.3 Test Procedures for ANOVA

Because the NP test procedures for ANOVA are not generalizations of the common test procedures for k samples, some comments on the latter are in order. The statistics (3.1) and (3.2) differ from the typical rank statistics with uncensored data in two critical ways. First they are written in terms of contrasts in the estimated cumulative hazard functions instead of the empirical distribution functions $\widehat{F}_i$. Second the integrands used in the comparisons $\widehat{\Lambda}_1 - \widehat{\Lambda}_i$ are different. Nevertheless, when specialized to uncensored data, the relationship $Y_{i\cdot}(t) = \int_{[t,\infty)} dN_{i\cdot}(s)$, where $N_{i\cdot}(s) = n_i \widehat{F}_i(s)$, implies that a weighted (by sample size) contrast of (3.1), with the weights in (3.2), gives comparisons between the distributions functions that use the same integrand in each comparison (see Andersen et al., 1993, Section 3.3). This property, however, does not generally hold with censored data. Since we seek weight functions that result in meaningful expressions when the components of (3.1) are reexpressed as integrals in Kaplan–Meier estimators, most of the commonly used weights for k samples are not included in our formulation.

A general theory of testing hypotheses in two-way and higher ANOVA designs is possible from the fact that all hypotheses can be expressed as $\mathbf{CF} = \mathbf{0}$, where $\mathbf{C}$ is a contrast matrix and $\mathbf{F}$ is the column vector of cell distribution functions. For a method to generate an appropriate contrast matrix for each hypothesis, see Akritas and Brunner (1997).

The Kaplan–Meier estimator of the distribution and survival functions from cell (i, j) are denoted by $\widehat{F}_{ij}$ and $\widehat{S}_{ij}$, respectively. As noted in Section 2.3, the distribution function of the time to the event variable cannot be estimated beyond the largest observation in the randomly censored sample. This implies that the comparison of distributions cannot be extended beyond the minimum of these maximum values from each sample. Let $\mathbf{F} = (F_{11}, \ldots, F_{1b}, \ldots, F_{ab})'$ and $\widehat{\mathbf{F}} = (\widehat{F}_{11}, \ldots, \widehat{F}_{1b}, \ldots, \widehat{F}_{ab})'$ denote the vector of distribution functions and the vector of Kaplan–Meier estimators from each factor-level combination. In view of the preceding discussion, the test statistic for $H_0 : \mathbf{CF} = \mathbf{0}$ is based on

$$\mathbf{C} \int_0^T \widehat{S}_H \, d\widehat{\mathbf{F}}, \tag{4.6}$$

where $T = T_{11} \wedge \cdots \wedge T_{ab}$, with T_{ij} being the largest observation from cell (i, j), and $\widehat{H} = 1 - \widehat{S}_H$ is the empirical distribution function of the censored and uncensored observation from all cells. Of course, (4.6) tests only $H_0^* : \mathbf{CF}(t) = \mathbf{0}, t \le \tau, \tau$ being the limit of T. Let $N_{rc\cdot}(t) = \sum_{k=1}^{n_{rc}} I(X_{rck} \le t, \Delta_{rck} = 1)$, $Y_{rc\cdot}(t) = \sum_{k=1}^{n_{rc}} I(X_{rck} \ge t)$ and $\widehat{h}_{rc}(s) = \widehat{S}_{rc}^2(s-)[\widehat{S}_H(s) - \widehat{S}_{rc}(s)^{-1} \int_{(s,T]} \widehat{S}_H \, d\widehat{F}_{rc}]^2 I(s \le T) n_{rc} Y_{rc\cdot}(s)^{-1}$, and define

$$\widehat{\sigma}_{rc}^2(t) = \int_0^t \widehat{h}_{rc}(s) \left(1 - \frac{\Delta N_{rc\cdot}(s) - 1}{Y_{rc\cdot}(s) - 1}\right) \frac{dN_{rc\cdot}(s)}{Y_{rc\cdot}(s)}.$$

Let $\widehat{\mathbf{V}}$ be the $ab \times ab$ diagonal matrix with diagonal elements $(n/n_{rc})\widehat{\sigma}_{rc}^2(T)$ and let $\mathbf{C}$ be any $\nu \times ab$ full row rank contrast matrix ($\nu < ab$). Then, under H_0^* given above,

$$N\left(\mathbf{C}\int_0^T \widehat{S}_H \, d\widehat{\mathbf{F}}\right)' (\mathbf{C}\widehat{\mathbf{V}}\mathbf{C}')^{-1} \left(\mathbf{C}\int_0^T \widehat{S}_H \, d\widehat{\mathbf{F}}\right) \xrightarrow{\mathcal{L}} \chi_\nu^2$$

in distribution,

where χ_ν^2 denotes the central chi-squared distribution with ν degrees of freedom.

4.4 Test Procedures for ANCOVA

For ANCOVA, we only discuss testing the first type of hypothesis given below (4.3). This includes also testing for covariate adjusted main effects and interactions between factors in higher way ANCOVA designs with one covariate.

To describe the test statistics, set $\overline{\mathbf{F}}_{\cdot}^{G} = (\overline{F}_{1\cdot}^{G}, \ldots, \overline{F}_{I\cdot}^{G})'$ and let $\widehat{\overline{\mathbf{F}}}_{\cdot}^{G} = (\widehat{\overline{F}}_{1\cdot}^{G}, \ldots, \widehat{\overline{F}}_{I\cdot}^{G})'$ be the NP estimator of it described in Remark 4.1. A general theory of testing hypotheses of the aforementioned type in one-way, two-way and higher ANCOVA designs is possible from the observation that any of these hypotheses can be expressed as $H_0 : \mathbf{C}\overline{\mathbf{F}}_{\cdot}^{G} = \mathbf{0}$ for some full-rank contrast matrix $\mathbf{C}$. For reasons explained in Section 4.3, the comparison of distributions must terminate at an appropriate point T (see Du, Akritas and Van Keilegom, 2003). Thus, the test statistic for such a hypothesis is based on

$$\widehat{T}_{\mathbf{C}}^{G} = \mathbf{C} \int_0^T \widehat{\overline{\mathbf{F}}}_{\cdot}^{G} \, d\widehat{H}, \tag{4.7}$$

where $\widehat{H}$ is the empirical distribution function of all Y_{ij}. Of course, (4.7) tests only $H_0^* : \mathbf{C}\overline{\mathbf{F}}_{\cdot}^{G}(t) = \mathbf{0}$, $t \le \tau$, τ being the limit of T.

Let $\widehat{\mathbf{V}}^G$ denote the estimate of the asymptotic covariance matrix of $\int_0^T \widehat{\overline{\mathbf{F}}}_{\cdot}^{G} \, d\widehat{H}$ given in Du, Akritas and Van Keilegom (2003). Then in the aforementioned paper it was shown that under suitable smoothness assumptions and under H_0^* given above,

$$N(\widehat{T}_{\mathbf{C}}^{G})'(\mathbf{C}\widehat{\mathbf{V}}^{G}\mathbf{C}')^{-1}\widehat{T}_{\mathbf{C}}^{G} \to \chi_\nu^2 \quad \text{in distribution,}$$

where ν denotes the rank of $\mathbf{C}$.

REMARK 4.1. An NP estimator of $\overline{F}_{i\cdot}^{G}$ is $\widehat{\overline{F}}_{i\cdot}^{G}(y) = \int \widehat{F}_{ix}(y) \, d\widehat{G}(x)$, where $\widehat{G}$ is the empirical distribution function of all X_{ij} and $\widehat{F}_{ix}(y)$ is Beran's (1981) NP kernel estimator of $F_{ix}(y)$ (cf. Du, Akritas and Van Keilegom, 2003).

5. NONPARAMETRIC METHODS: REPEATED MEASURES

With dependent data, the most common NP procedures pertain to matched pairs and the multivariate two-sample problem. See Woolson and O'Gorman (1992) and Wei and Lachin (1984). For the reasons given in Section 4.1, extension of such methods to general classes of repeated measures designs requires the use of the fully NP models and hypotheses.

The models we describe are called *marginal NP* repeated measures models because the covariance structure of the repeated measurements is left unspecified. In this formulation, factors whose levels are crossed with the subjects are called *column* factors; those factors with subjects nested within their levels are called *row* factors. In the diabetic retinopathy study (see below), type of diabetes is the row factor, while treatment is the column factor. However, gender could have been an additional row factor, while medication such as eye drops could have been an additional column factor. The terms $MM(x; y)$ denote a design with x row factors

and y column factors. We present the formulation for an $MM(1;1)$ design with r and c levels of the row and column factors, respectively. By imposing structure on the subscripts i and j, the above model formulation includes any $MM(x;y)$ design.

The independent random vectors $\mathbf{Y}_{ik} = (Y_{i1k}, \ldots, Y_{ick})'$ represent the c observations of the kth subject nested under the ith level of the row factor. Let also $\mathbf{C}_{ik} = (C_{i1k}, \ldots, C_{ick})'$ be independent random vectors that represent the censoring variables. Vector $\mathbf{Y}_{ik}$ is assumed to be independent of $\mathbf{C}_{ik}$. The observed quantities are

$$\begin{aligned} &\mathbf{Z}_{ik} = (Z_{i1k}, \ldots, Z_{ick})' \quad \text{and} \\ &\mathbf{\Delta}_{ik} = (\Delta_{i1k}, \ldots, \Delta_{ick})', \end{aligned} \tag{5.1}$$

$i = 1, \ldots, r$ and $k = 1, \ldots, n_i$, where $Z_{ijk} = \min(Y_{ijk}, C_{ijk})$ and $\Delta_{ijk} = I(Z_{ijk} = Y_{ijk})$ for $j = 1, \ldots, c$. The marginal NP mixed model specifies only that

$$Y_{ijk} \sim F_{ij} \quad \text{and} \quad C_{ijk} \sim G_{ij}, \qquad i = 1, \ldots, r,\ j = 1, \ldots, c,$$

for some distributions functions F_{ij} and G_{ij}, which are not assumed continuous. The NP hypotheses are defined through a decomposition of the F_{ij} as in Section 4.2. Thus, all hypotheses of interest can be expressed as $\mathbf{CF} = \mathbf{0}$, where $\mathbf{C}$ is a contrast matrix and $\mathbf{F} = (F_{11}, \ldots, F_{1c}, \ldots, F_{r1}, \ldots, F_{rc})'$. Let $\widehat{\mathbf{F}} = (\widehat{F}_{11}, \ldots, \widehat{F}_{1c}, \ldots, \widehat{F}_{r1}, \ldots, \widehat{F}_{rc})'$, where $\widehat{F}_{ij}$ is the Kaplan–Meier estimator from the data in cell (i, j). The statistic for testing $\mathbf{CF} = \mathbf{0}$ is still based on (4.6), but due to dependence, its asymptotic covariance matrix is different from that given in Section 4.3. For details, see O'Gorman and Akritas (2001).

EXAMPLE 1 (Diabetic retinopathy study). This study considers the effectiveness of laser photocoagulation in delaying the onset of blindness in patients with diabetic retinopathy. One eye of each patient was randomly chosen for treatment, while the other received no treatment. This is an $MM(1,1)$ design where the column factor (factor B) is treatment and the row factor (factor A) is type of diabetes (juvenile or adult onset). The response variable is the time until visual acuity in an eye is less than $5/200$. There are 114 juvenile and 83 adult onset patients, and 61% of the observations are censored.

Figure 1 presents the estimated NP main effects for type of diabetes and treatment, and the estimated NP

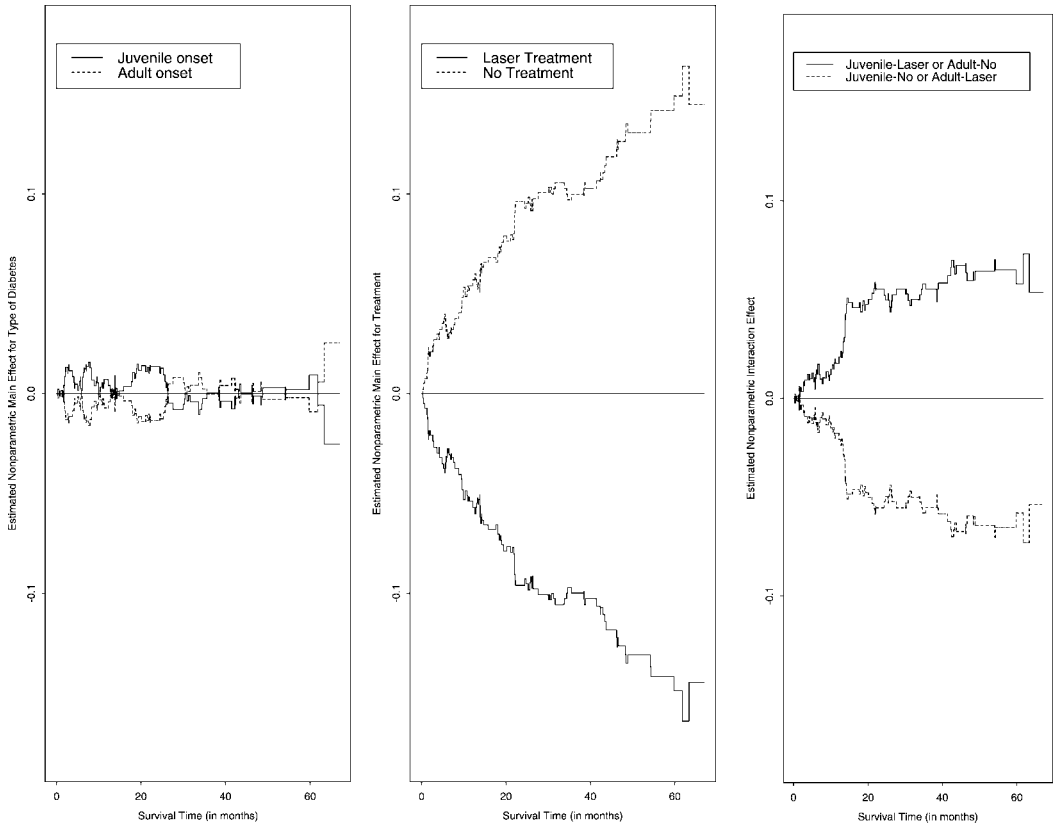

FIG. 1. *Plots of estimated NP effects for Example* 1.

interaction effect. The extensive crossing and the absolute magnitude of the estimated NP type of diabetes effects in the left plot of Figure 1 indicate that the NP type of diabetes main effect is not significant. The middle plot of Figure 1 indicates significant treatment effect. In particular, $\overline{\widehat{F}}_{\cdot 1}(t) > \overline{\widehat{F}}_{\cdot 2}(t)$ means that the average estimated probability that the response variable is less than t is greater for the no treatment group. Finally, the right plot also indicates significant interaction effects.

The test procedures give p values 0.961, less than 0.001 and 0.003 for type of diabetes, treatment and interaction, respectively, which agree with the plots of the NP effects.

It is interesting to compare the above analysis with the NP multivariate two-sample approach of Wei and Lachin (1984), who tested equality of the two multivariate distributions $F_1(t_1, t_2)$ and $F_1(t_1, t_2)$ that correspond to juvenile and adult groups, respectively. Note that this null hypothesis implies our NP hypothesis of no simple effect for type of diabetes (no main effect and no interaction with treatment). Lin (1994) pointed out that the Wei and Lachin (1984) statistic can be calculated with SAS PROC PHREG and a macro from the SAS web page. This gives a p value of 0.0175. Note, however, that the multivariate two-sample approach cannot be used to test for no treatment effect.

ACKNOWLEDGMENT

This work was supported in part by NSF Grant SES-03-18200.

REFERENCES

AKRITAS, M. G. (1986). Bootstrapping the Kaplan–Meier estimator. *J. Amer. Statist. Assoc.* **81** 1032–1038.

AKRITAS, M. G. (1988). Pearson-type goodness-of-fit tests: The univariate case. *J. Amer. Statist. Assoc.* **83** 222–230.

AKRITAS, M. G. (1992). Rank transform statistics with censored data. *Statist. Probab. Lett.* **13** 209–221.

AKRITAS, M. G. and ARNOLD, S. F. (1994). Fully nonparametric hypotheses for factorial designs. I. Multivariate repeated measures designs. *J. Amer. Statist. Assoc.* **89** 336–343.

AKRITAS, M. G., ARNOLD, S. F. and DU, Y. (2000). Nonparametric models and methods for nonlinear analysis of covariance. *Biometrika* **87** 507–526.

AKRITAS, M. G. and BRUNNER, E. (1997). Nonparametric methods for factorial designs with censored data. *J. Amer. Statist. Assoc.* **92** 568–576.

AKRITAS, M. G. and LAVALLEY, M. P. (1997). Statistical analysis with incomplete data: A selective review. In *Robust Inference* (G. S. Maddala and C. R. Rao, eds.). *Handbook of Statistics* **15** 551–632. North-Holland, Amsterdam.

ANDERSEN, P. K., BORGAN, Ø., GILL, R. and KEIDING, N. (1993). *Statistical Models Based on Counting Processes*. Springer, New York.

BERAN, R. (1981). Nonparametric regression with randomly censored survival data. Technical report, Dept. Statistics, Univ. California, Berkeley.

BORGAN, Ø. and LIESTØL, K. (1990). A note on confidence intervals and bands for the survival function based on transformations. *Scand. J. Statist.* **17** 35–41.

BROOKMEYER, R. and CROWLEY, J. J. (1982). A confidence interval for the median survival time. *Biometrics* **38** 29–41.

COX, D. R. (1972). Regression models and life-tables (with discussion). *J. Roy. Statist. Soc. Ser. B* **34** 187–220.

DU, Y., AKRITAS, M. G. and VAN KEILEGOM, I. (2003). Nonparametric analysis of covariance for censored data. *Biometrika* **90** 269–287.

FLEMING, T. R. and HARRINGTON, D. P. (1981). A class of hypothesis tests for one- and two-sample censored survival data. *Comm. Statist. Theory Methods* **10** 763–794.

GILL, R. D. (1980). *Censoring and Stochastic Integrals*. Mathematisch Centrum, Amsterdam.

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

GREENWOOD, M. (1926). The natural duration of cancer. *Reports on Public Health and Medical Subjects* **33** 1–26. Her Majesty's Stationery Office, London.

HALL, W. J. and WELLNER, J. A. (1980). Confidence bands for a survival curve from censored data. *Biometrika* **67** 133–143.

HJORT, N. L. (1990). Goodness-of-fit tests in models for life history data based on cumulative hazard rates. *Ann. Statist.* **18** 1221–1258.

HOLLANDER, M. and PEÑA, E. A. (1992). A chi-squared goodness-of-fit test for randomly censored data. *J. Amer. Statist. Assoc.* **87** 458–463.

KAPLAN, E. L. and MEIER, P. (1958). Nonparametric estimation from incomplete observations. *J. Amer. Statist. Assoc.* **53** 457–481.

KIM, J. H. (1993). Chi-square goodness-of-fit tests for randomly censored data. *Ann. Statist.* **21** 1621–1639.

LAWLESS, J. F. (1982). *Statistical Models and Methods for Lifetime Data*. Wiley, New York.

LI, G. and DOSS, H. (1993). Generalized Pearson–Fisher chi-square goodness-of-fit tests, with applications to models with life history data. *Ann. Statist.* **21** 772–797.

LIN, D. Y. (1994). Cox regression analysis of multivariate failure time data: The marginal approach. *Statistics in Medicine* **13** 2233–2247.

MANTEL, N. (1966). Evaluation of survival data and two new rank order statistics arising in its consideration. *Cancer Chemotherapy Reports Part 1* **50** 163–170.

O'GORMAN, J. T. and AKRITAS, M. G. (2001). Nonparametric models and methods for designs with dependent censored data. *Biometrics* **57** 88–95.

PEPE, M. S. and FLEMING, T. R. (1991). Weighted Kaplan–Meier statistics: Large sample and optimality considerations. *J. Roy. Statist. Soc. Ser. B* **53** 341–352.

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

THOMAS, D. R. and GRUNKEMEIER, G. L. (1975). Confidence interval estimation of survival probabilities for censored data. *J. Amer. Statist. Assoc.* **70** 865–871.

WEI, L. J. and LACHIN, J. M. (1984). Two-sample asymptotically distribution-free tests for incomplete multivariate observations. *J. Amer. Statist. Assoc.* **79** 653–661.

WOOLSON, R. F. and O'GORMAN, T. W. (1992). A comparison of several tests for censored paired data. *Statistics in Medicine* **11** 193–208.