Nonparametric Estimation of Specific Occurrence/Exposure Rate in Risk and Survival Analysis

GUTTI JOGESH BABU, C. RADHAKRISHNA RAO, and M. BHASKARA RAO*

A cohort of individuals exposed to some risk is followed up to a point of time $M$, and observations on two random variables $(Y, S)$ are recorded for each individual. The variable $Y$ refers to one of the four possible events that can occur for an individual in the period $[0, M]$: (i) dies of a specific disease, say cancer, (ii) dies of a natural cause, (iii) withdraws from the study, and (iv) is alive and still under study at time $M$. The variable $Y$ refers to the time at which an event occurs. Based on such data for $n$ individuals, we consider the problem of estimation of a specific occurrence/exposure rate (SOER), which is a risk ratio defined as the ratio of probability of death due to cancer in the interval $[0, M]$ to the mean lifetime of all individuals up to the time point $M$. The asymptotic distribution of a nonparametric estimator of SOER is shown to be normal, and the asymptotic variance involves unknown parameters. Various ways of bootstrapping are discussed for construction of confidence intervals for SOER and compared. Some numerical illustrations are provided.

KEY WORDS: Berry–Esseen bound; Bootstrap; Competing risks; Kaplan–Meier estimate; Mixed censoring; Strong approximation.

1. INTRODUCTION

The main problem discussed in this article arose in the context of analyzing a particular data set consisting of four types of observations. Records of mortality have been maintained for a certain length of time, $M$, say, on people who enter into a particular hazardous employment. The length of time $M$ is fixed in advance, based on practical and financial considerations. Some of the people have died due to cancer, some others due to other causes, some others have left the employment during the period of observation, and the rest are still alive at the end of the period. Let $\pi$ denote the probability of dying due to cancer for anyone who enters into the hazardous employment and $\pi^*$ the corresponding probability of death due to cancer in the population at large. We assume that the relevant covariates remain the same in these two populations. One of the goals in the analysis of the data described to us is to estimate $\pi$ and $\pi^*$ and compare them. A need also arose to define in a meaningful way the risk of exposure to cancer for any one who enters into the hazardous employment and compare this risk with the one for the general population.

We believe that the solution offered in this article will be helpful in such studies. We formulate this problem somewhat formally as follows.

1.1 Occurrence/Exposure Rate

We consider a population of individuals exposed to some risk and study the problem of assessing the risk rate due to a particular cause of death, say cancer. Let

$$F(t) = \pi F_1(t) + (1 - \pi) F_2(t), \quad 0 \leq t < \infty,$$

be the distribution function (cdf) of the lifetimes of individuals, where $\pi$ is the probability of death due to cancer, $1 - \pi$ that of death due to other causes, $F_1$ is the distribution function of lifetimes of those dying of cancer, and $F_2$ of those dying of natural or other causes. The risk ratio, or the specific occurrence/exposure rate (SOER), for cancer in a given period $[0, M]$, defined by

$$\rho_M = \frac{\pi F_1(M)}{\int_0^M \bar{F}(t) \, dt},$$

(1.2)

where $\bar{F}(t) = 1 - F(t)$ and $\pi_M = F_1(M)$, has been used to assess the risk rate in medical studies. See, for instance, Howe (1983). Roughly speaking, SOER is the rate of deaths due to cancer during an average lifespan. An intuitive interpretation of SOER can be provided as follows. Suppose that it is possible to observe every individual in the population until he or she dies, that is, $M = \infty$. The exposure rate of cancer for any individual in the population can be defined by

$$\rho = \frac{\text{(probability of death due to cancer)/(average life span)}}{t \, dF(t)}.$$

If the period of observation $[0, M]$ is an integral part of the design, which is natural in many studies, it has been pointed out that $\rho$ cannot be estimated nonparametrically. However, $\rho_M$, as defined above, which is a modification of $\rho$, can be estimated nonparametrically.

To make inferences on $\rho_M$, we follow a cohort of $n$ individuals exposed to risk during the period $[0, M]$ and make observations on certain events as described subsequently. In such a study, four possibilities arise with reference to any member of the cohort during the period $[0, M]$: (i) dies of cancer, (ii) dies of other causes, (iii) leaves the study, or (iv) is alive and in the study at $M$, the end of the period.

* Gutti Jogesh Babu is Professor of Statistics and C. Radhakrishna Rao is Eberly Professor of Statistics, Department of Statistics, Pennsylvania State University, University Park, PA 16802. M. Bhaskara Rao is Associate Professor of Statistics, Department of Statistics, North Dakota State University, Fargo, ND 58105. This work was supported by Air Force Office of Scientific Research Grant F33165-58-C-0630. The work of Gutti Jogesh Babu was partially supported by National Security Agency Grant MDA904-90-H-1001 and by National Aeronautics and Space Administration Grant NAGW-1917. The authors thank the associate editor and referees for helpful comments.

© 1992 American Statistical Association
Journal of the American Statistical Association
March 1992, Vol. 87, No. 417, Theory and Methods

84
of study. These possibilities are summarized graphically in Figure 1. The data thus consist of four different types of observations. It is in this framework that we want to extract information on $\rho_M$.

1.2 Formulation of the Problem

Let $T$ be a nonnegative random variable representing the lifetime of an individual and $\tilde{\Delta}$ be a binary random variable taking values

$$\tilde{\Delta} = 1, \quad \text{if an individual dies of cancer},$$
$$\quad = 0, \quad \text{if an individual dies of other causes}.$$

It is seen that $F_1$ in (1.1) is the conditional distribution function of $T$ given $\tilde{\Delta} = 1$ and $F_2$ that of $T$ given $\tilde{\Delta} = 0$. Further, let $C$ be a nonnegative random variable, called the censoring variable, with cdf $G$ independent of $(T, \tilde{\Delta})$. Associated with each individual we have the triplet $(T, \tilde{\Delta}, C)$, which is not observable in its entirety. What can be recorded for each individual is $(Y, \Delta)$, where $Y = \min\{T, C\}$ and

$$\Delta = 1, \quad \text{if } Y = T \text{ and } \tilde{\Delta} = 1,$$
$$\quad = 0, \quad \text{if } Y = T \text{ and } \tilde{\Delta} = 0,$$
$$\quad = -1, \quad \text{if } Y = C,$$
$$\quad = -2, \quad \text{if } Y = M. \quad (1.3)$$

The last category represents the situation in which an individual leaves the study. To avoid any ambiguity in the definition of $(Y, \Delta)$, we assume that $P(T = C) = 0$. The stochastic model just described for $(Y, \Delta)$ is known as the right censoring model (RCM).

Besides the type of right censoring described previously, we have, in practice, another kind of censoring arising out of possible termination of the study at a given time point $M$. In such a case, the observation on each individual is of the form $(Y, \Delta)$, where $Y = \min\{T, C, M\}$ and

$$\Delta = 1, \quad \text{if } Y = T \text{ and } \tilde{\Delta} = 1,$$
$$\quad = 0, \quad \text{if } Y = T \text{ and } \tilde{\Delta} = 0,$$
$$\quad = -1, \quad \text{if } Y = C,$$
$$\quad = -2, \quad \text{if } Y = M. \quad (1.4)$$

The stochastic model (1.4) for $(Y, \Delta)$ is called the mixed censoring model (MCM). To avoid any ambiguity in the definition of $(Y, \Delta)$, we assume that the probability of any two of $T, C, \text{ and } M$ being equal is 0.

In this article we consider the problem of drawing inferences on $\rho_M$, as defined in (1.2), when we have observed data under the MCM. In this connection, a generalized maximum likelihood nonparametric estimator $\hat{\pi}_M$ of $\pi_M$ is given in Appendix A. Based on this estimator, a nonparametric estimator $\hat{\rho}_M$ of $\rho_M$ is proposed, and its asymptotic distribution is obtained (see Corollary 2). Theorem 1 establishes the strong approximation result needed in this connection. The asymptotic variance of $\hat{\rho}_M$ involves unknown parameters. Various ways of bootstrapping are discussed for construction of confidence intervals for $\rho_M$ and compared.

We note that there has been considerable work done under the RSM with a single cause of death, originating with the seminal work of Kaplan and Meier (1958) and followed up by Csörgő and Horváth (1983), Gill (1980), Horváth and Yandell (1987), Johansen (1978), Miller (1981) and others. Some work is also done under the RCM with multiple causes of death (referred to as competing risks). However, some restrictions are placed on the stochastic mechanism giving rise to the distribution (1.1) of lifetimes of individuals. Kalbfleisch and Prentice (1980, p. 163) considered the competing risks between causes of death to be stochastically independent. Langberg, Proshan, and Quinz (1981) assumed a specific dependent structure for lifetimes. The particular framework under which we have formulated the problem is very general and makes no assumptions on the stochastics of competing risks. Further, we treat the problem under the MCM, which usually occurs in practice.

We do not assume any particular models for the cdfs $F_1$, $F_2$, and $G$, and our method of estimation is nonparametric as in the Kaplan-Meier problem.

2. MAIN RESULTS

Let $(Y_1, \Delta_1), (Y_2, \Delta_2), \ldots, (Y_n, \Delta_n)$ be observations on $(Y, \Delta)$ under the MCM as described in (1.4). Based on these data, we consider the inference problem on the SOER.

$$\rho_M = \pi_M / \gamma_M, \quad (2.1)$$

where

$$\pi_M = \pi F_1(M), \quad \gamma_M = \int_{(0,M]} \bar{F}(t) \, dt$$

and

$$F(t) = \pi F_1(t) + (1 - \pi) F_2(t)$$

is as defined in (1.1). For any cdf $B$, let $\bar{B}(t) = 1 - B(t)$ and $B(t^-)$ denote the left limit of $B$ at $t$.

Throughout the main body of this paper, the cdfs $F_1$, $F_2$, and $G$ are assumed to be continuous.

Let $\bar{H} = \bar{F}G$ and assume that $\bar{H}(M) > 0$. From (A.1) and (A.2) of the Appendix, the Kaplan-Meier (KM) estimators [generalized maximum likelihood estimators in the sense of Kiefer and Wolfowitz (1956)] of $\pi_M$ and $\gamma_M$ are given by

$$\hat{\pi}_M = (1/n) \sum_{i=1}^{n} I(\Delta_i = 1) (\bar{G}(Y_i^-))^{-1} \quad (2.2)$$

and
\[ \hat{\gamma}_M = \int_{[0, M]} \hat{F}(t) \, dt, \quad (2.3) \]

where \( \hat{G} \) and \( \hat{F} \) are the usual KM estimators of \( G \) and \( F \).

A natural estimator for \( \rho_M \), based on (2.2) and (2.3), is given by \( \hat{\rho}_M = \hat{\pi}_M / \hat{\gamma}_M \).

To state strong approximation results for \( \hat{F} \) and \( \hat{G} \), we need to define the following for \( 0 \leq t \leq M, \ x \geq 0, \) and \( \delta \) real:

\[ \xi(x, \delta, t) = \hat{F}(t)[g_1(\min\{x, t\}) + (\hat{H}(x))^{-1} I(x \leq t, \delta \geq 0)] \]

and

\[ \eta(x, \delta, t) = \hat{G}(t)[g_2(\min\{x, t\}) + (\hat{H}(x))^{-1} I(x \leq t, \delta < 0)]. \]

where

\[ g_1(x) = \int_0^x (\hat{H}(s))^{-2} \, d\hat{H}(s), \]

\[ g_2(x) = \int_0^x (\hat{H}(s))^{-2} \, d\hat{H}(s), \]

\[ \bar{H}_s(s) = P(Y > s, \Delta \geq 0), \]

and

\[ \bar{H}_s(s) = P(Y > s, \Delta < 0). \]

Note that

\[ \bar{H}_s(s) + \bar{H}_s(s) = \bar{H}(s) I(s < M). \]

Further, \( \xi \) and \( \eta \) are uniformly bounded by \( 2(\hat{H}(M))^{-2} \), for \( 0 \leq t \leq M \). The following lemma on strong approximation is due to Lo and Singh (1986). Let \( a_n = n^{-\delta/4}(\log n)^{\delta/4} \).

**Lemma 1.** We have, with probability 1,

\[ \sup_{0 \leq t \leq M} \left| \hat{F}(t) - \bar{F}(t) + (1/n) \sum_{i=1}^n \xi_i(t) \right| = O(a_n), \quad (2.4) \]

and

\[ \sup_{0 \leq t \leq M} \left| \hat{G}(t) - \bar{G}(t) + (1/n) \sum_{i=1}^n \eta_i(t) \right| = O(a_n), \quad (2.5) \]

where

\[ \xi_i(t) = \xi(Y_i, \Delta_i, t) \]

and

\[ \eta_i(t) = \eta(Y_i, \Delta_i, t). \]

To state our main result on strong approximations, we need the following notation. For \( t \leq M \), let

\[ \mu(t, \delta) = -\int_0^M \xi(t, \delta, s) \, ds \]

and

\[ \theta(t, \delta) = [I(\delta = 1) (\hat{G}(Y))^{-1} + \bar{\eta}(t, \delta)]. \]

where

\[ \bar{\eta}(t, \delta) = E[\eta(t, \delta, Y)(\Delta = 1)(\hat{G}(Y))^{-1}] \]

\[ = E[\{I(\Delta = 1, T \leq M) g_2(\min\{T, t\})\}] \]

\[ + \left( \pi_M - \pi F_1(t) (\hat{H}(t))^{-1} l(\delta < 0) \right] . \]

**Theorem 1.** With probability 1, we have

\[ \hat{\pi}_M = (1/n) \sum_{i=1}^n \theta(Y_i, \Delta_i) + O(a_n) \quad (2.6) \]

and

\[ \hat{\gamma}_M - \gamma_M = (1/n) \sum_{i=1}^n \mu(Y_i, \Delta_i) + O(a_n). \quad (2.7) \]

A proof of this result is given in the Appendix. The following corollaries are immediate consequences of the iid representations given in Theorem 1.

**Corollary 1.** The asymptotic distribution of

\[ n^{1/2}(\hat{\pi}_M - \pi_M, \hat{\gamma}_M - \gamma_M) \]

is \( N(0, \Sigma) \), where \( \Sigma \) is the variance–covariance matrix of the random variables \( \theta(Y, \Delta) \) and \( \mu(Y, \Delta) \).

The formulas for these variances and covariances are now given.

\[ \sigma_{11} = \text{var}(\theta(Y, \Delta)) \]

\[ = \int_0^M a^2(x, M) \, d\bar{H}_x(s) + \int_0^M (\bar{G}(t))^{-1} d(\pi F_1(t) - \pi_M^2), \]

\[ \sigma_{22} = \text{var}(\mu(Y, \Delta)) = -\int_0^M \left[ \int_s^M (\bar{F}(s)/(\bar{H}(s))) \, ds \right]^2 d\bar{H}_x(s), \]

and

\[ -\sigma_{12} = -\text{cov}(\theta(Y, \Delta), \mu(Y, \Delta)) \]

\[ = \int_0^M a(x, M) \left[ \int_s^M (\bar{F}(s)/(\bar{H}(s))) \, ds \right] d\bar{H}_x(s) \]

\[ + \int_0^M \left[ \int_s^M (\bar{F}(t)/(\bar{H}(t))) \, dt \right] d(\pi F_1(s)), \]

where

\[ a(x, M) = \pi F_1(M) - F_1(x)/\bar{H}(x). \]

The expression for \( \sigma_{12} \) follows by observing that for \( 0 \leq t, s \leq M \),

\[ E(\xi(Y, \Delta, t) \eta(Y, \Delta, s)) = 0. \]

**Corollary 2.** The asymptotic distribution of

\[ n^{1/2}(\hat{\pi}_M - \pi_M/\gamma_M) \]

is normal with mean zero and variance

\[ \sigma^2 = \gamma_M^4(\pi_M - \pi_M) \Sigma(\gamma_M - \pi_M)' \quad (2.8) \]

The preceding results can be improved by refining (2.4) and (2.5) by including a U-statistic-type second-order statistic in the representation. The details will be given elsewhere. Using these refinements, one can obtain the following Berry–Essén bound,

\[ |P(n^{1/2}(\hat{\pi}_M - \pi_M/\gamma_M) \leq x) - \Phi_{\pi}(x)| = O(n^{-1/2}), \]

where \( \Phi_{\pi} \) is the cdf of \( N(0, \sigma^2) \).

**Corollary 3.** We have

\[ n^{1/2}(\hat{\pi}_M - \rho_M \hat{\gamma}_M) \rightarrow N(0, \sigma_{11} - 2\rho_M \sigma_{12} + \rho_M^2 \sigma_{22}), \]

as \( n \rightarrow \infty \).

**Remark.** If \( \rho_{1,M} \) and \( \rho_{2,M} \) denote the SOER's in the same time interval \([0, M]\) for two independent populations, the asymptotic distribution of the ratio \( \hat{\rho}_{1,M}/\hat{\rho}_{2,M} \) can easily be obtained following Corollary 2.

The preceding corollaries are of little value in obtaining confidence intervals for \( \pi_M, \gamma_M \), or for the SOER \( \rho_M \) as \( \Sigma \) and \( \sigma^2 \) depend on the unknown distributions \( F \) and \( G \). There are at least two ways of estimating \( \sigma^2 \) and \( \Sigma \). One is by replacing \( \pi F_1, (1 - \pi) F_2, G \), and \( H(x) \) by \( (\pi F_1), (1 - \pi F_2), \)
\( \pi F \), \( G \), and \( H_j(x) = n^{-1} \sum_{i=1}^n I(Y_i \leq x) \). The second is the bootstrap method, which is explained below.

The bootstrap method consists of drawing random samples (with replacement) \((Y_i^*, \Delta^*_i)\), \(i = 1, 2, \ldots, n\), from \((Y_i, \Delta_i)\), \(i = 1, 2, \ldots, n\), each giving pair equal weight and constructing the KM estimators \((\hat{\pi}_M, \gamma_M)\) of \((\pi_M, \gamma_M)\) using \((Y_i^*, \Delta_i^*)\)'s. In practice, by repeating this procedure a large number of times, say \(N\), one obtains estimators \((\hat{\pi}_j\gamma_j\), \(j = 1, 2, \ldots, N\). The sample variance-covariance matrix of \((\hat{\pi}_j\gamma_j\), \(j = 1, 2, \ldots, N\) is a consistent estimator of \(\Sigma\). Using this and the estimators \(\hat{\pi}_M\gamma_M\) of \(\pi_M\) and \(\gamma_M\), we obtain a consistent estimator of \(\sigma^2\) as defined in (2.8).

By arguing as in the proof of Theorem 1 and using equation (2) of Theorem 1 of Lo and Singh (1986), we obtain the following result.

**Theorem 2.** With probability 1
\[
\pi^* = (1/n) \sum_{i=1}^n \theta(Y_i^*, \Delta_i^*) + R_{1,n},
\]
and
\[
\gamma^* = (1/n) \sum_{i=1}^n \mu(Y_i^*, \Delta_i^*) + R_{2,n},
\]
where, for some \(A > 0\),
\[
P^*(\|R_{1,n}\| + \|R_{2,n}\| > Aa_n) \to 0
\]
as \(n \to \infty\), \(P^*\) being the bootstrap measure.

From Corollaries 2 and 3 and Theorem 2, we have the following bootstrap approximation of the sampling distribution.

**Corollary 4.** With probability 1,
\[
P^*(n^{1/2}(\pi^*/\gamma^* - \hat{\pi}_M/\gamma_M) \leq x) - P(n^{1/2}(\hat{\pi}_M/\gamma_M - \pi_M/\gamma_M) \leq x) \to 0
\]
and
\[
P^*(n^{1/2}(\pi^* - \hat{\pi}_M) \leq x) - P(n^{1/2}(\pi^* - \pi_M) \leq x) \to 0
\]
uniformly in \(x\) as \(n \to \infty\).

Now, we present various methods of constructing confidence intervals for \(\rho_M\).

**Method 1.** By Corollary 3, the asymptotic distribution of
\[
n^{1/2}(\hat{\sigma}_M - \rho_M\gamma_M)(\sigma_{11} - 2\rho_M\sigma_{12} + \rho_M^2\sigma_{22})^{-1/2}
\]
is \(N(0, 1)\). In the formulas given for \(\sigma_{11}, \sigma_{12}, \) and \(\sigma_{22}\) after Corollary 1, replace \(F, G, \pi F\), and \(1 - \pi F\) by their corresponding KM estimators to obtain consistent estimators of these \(\sigma_j\)'s. Since the asymptotic distribution of
\[
n^{1/2}(\hat{\sigma}_M - \rho_M\gamma_M)(\sigma_{11} - 2\rho_M\sigma_{12} + \rho_M^2\sigma_{22})^{-1/2}
\]
is \(N(0, 1)\), one can obtain a confidence interval for \(\rho_M\) in the usual way.

**Method 2.** Replacing \(\hat{\sigma}_j\)'s in (2.13) by the corresponding bootstrap estimates, one can obtain confidence intervals for \(\rho_M\) in the usual way.

**Method 3.** By taking a large number \(N\) of bootstrap samples from the given data, one can obtain estimates \(\pi_j^\ast\) and \(\gamma_j^\ast\) of \(\pi_M\) and \(\gamma_M\) and construct the empirical histogram of \(\pi_j^\ast - \hat{\rho}_M\gamma_j^\ast\)'s. From this one can obtain \(x_L\) and \(x_U\), the \((\alpha/2)\)th and \((1 - \alpha/2)\)th quantiles, respectively. Now an approximate \((1 - \alpha)100\%\) confidence interval is given by
\[
x_L \leq \hat{\pi}_M - \hat{\rho}_M\hat{\gamma}_M \leq x_U.
\]

**Method 4.** Instead of looking at the histogram of \(\pi_j^\ast - \hat{\rho}_M\gamma_j^\ast\)'s, one could look at the histogram of
\[
t' = n^{1/2}(\pi_j^\ast - \hat{\rho}_M\gamma_j^\ast)(\sigma_{11} - 2\hat{\rho}_M\sigma_{12} + \hat{\rho}_M^2\sigma_{22})^{-1/2}
\]
and obtain \(t'\) and \(t'_U\), the \((\alpha/2)\)th and \((1 - \alpha/2)\)th quantiles. Then the confidence interval for \(\rho_M\) is obtained from
\[
t'_L \leq n^{1/2}(\hat{\pi}_j - \hat{\rho}_M\hat{\gamma}_j)(\sigma_{11} - 2\hat{\rho}_M\sigma_{12} + \hat{\rho}_M^2\sigma_{22})^{-1/2}
\]
whence \(t' \leq t'_L\).

Now we take up the problem of estimating the survival function associated with \(F_i\) over the interval \([0, M]\). As pointed out in the Appendix, the survival function \(S_i(t) = F_i(t) = 1 - F_i(t)\) of cancer deaths is not estimable. However, the modified survival function over the interval \([0, M]\) defined by
\[
\hat{S}_i(t) = (F_i(M) - F_i(t))/F_i(M) = (\pi_M - \pi F_i(t))/\pi_M
\]
is estimable.

From Lemma 1, (A.2), (B.3), and from the proof of Theorem 1, we have uniformly in \(0 \leq t \leq M\), with probability 1,
\[
(\hat{S}_i(t)) = 1 - \left[ \left(\frac{1}{n} \sum_{i=1}^n \theta(Y_i, \Delta_i, t)I(Y_i \leq t) + O(a_n) \right) \times \left(\frac{1}{n} \sum_{i=1}^n \theta(Y_i, \Delta_i) + O(a_n) \right) \right]^{-1}
\]
where
\[
\theta(s, \delta, t) = \eta(\delta = 0, s \leq t)(\hat{G}(Y_i))^{-1}
\]
\[
+ E[\theta(s, \delta, Y_i)(\hat{G}(Y_i))^{-2}].
\]
To show that
\[
w_i(t) = n^{-1/2} \sum_{i=1}^n [\theta(Y_i, \Delta_i, t)I(Y_i \leq t) - E(\theta(Y, \Delta, t)I(Y_i \leq t))]
\]
converges weakly to a Gaussian process on \([0, M]\) with covariance function
\[
r(t, s) = \text{cov}[\theta(Y, \Delta, t)I(Y \leq t), \theta(Y, \Delta, s)I(Y \leq s)].
\]
we show that \([w_n]_a\) is tight in \([0, M]\). Since \(\theta\) is uniformly bounded, using Hölder's inequality it can be shown that, for \(0 \leq t_1 \leq t \leq t_2 \leq M\),
\[
E[w_n(t) - w_n(t_1)^2] w_n(t) - w_n(t_2)^2 \leq (A(t) - A(t_1))(A(t_2) - A(t))
\]
for some continuous nondecreasing function \(A\). This implies tightness of \([w_n]_a\) in \([0, M]\). See Billingsley (1968,
p. 128). Let \( w \) be the limiting distribution of \( w_n \). Then the process
\[
g = n^{1/2}(S - \hat{S})
\]
converges weakly in \( D[0, M] \) to
\[
w^\circ = w/F_1(M) - \hat{S}w(M).
\]
By problem 3 on page 136 of Billingsley (1968), \( w^\circ \) is a gaussian process with continuous paths. Similar results hold for \( S(t) \), the modified survival function of deaths due to other causes. The limiting process for \( \hat{G}(t) \) follows from (2.5).

3. NUMERICAL ILLUSTRATION

Suppose we have a sample of \( n = 7 \) individuals, of whom two died of cancer, three of other causes, and the remaining two left the study. Suppose the following are the summary statistics in the notation of Appendix A:

\[
\begin{align*}
u_i & : 1.7 \ 2.3 \ 4.4 \ 4.5 \ 4.9 \ 6.0 \ 6.1  \\
d_i & : 1 \ 0 \ 0 \ 1 \ 0 \ 0 \ 0  \\
s_i & : 0 \ 1 \ 0 \ 0 \ 0 \ 1 \ 1  \\
c_i & : 0 \ 0 \ 1 \ 0 \ 1 \ 0 \ 0  \\
n & : 7 \ 6 \ 5 \ 4 \ 3 \ 2 \ 1.
\end{align*}
\]

Since two out of seven died of cancer, a crude estimate of \( \pi \) would be \( 2/7 \), which needs to be adjusted to take into account the fact that two individuals left the study. Since \( c_i = 0 \), that is, at the last observable time point no one left the study, we have indeed a generalized maximum likelihood estimate of \( \pi \) given by
\[
\hat{\pi} = 9/28 > 2/7.
\]
For a revised estimate of \( \pi \), it is important how the times at which cancer deaths occurred are interspersed with the times of individuals who left the study. In the following example, in which the times of cancer deaths precede censoring times, \( \hat{\pi} = 2/7 \) remains the same as the crude death rate:

\[
\begin{align*}
u_i & : 1.7 \ 2.3 \ 4.4 \ 4.5 \ 4.9 \ 6.0 \ 6.1  \\
d_i & : 1 \ 1 \ 0 \ 0 \ 0 \ 0 \ 0  \\
s_i & : 0 \ 0 \ 1 \ 0 \ 1 \ 0 \ 1  \\
c_i & : 0 \ 0 \ 0 \ 1 \ 1 \ 0 \ 0  \\
n & : 7 \ 6 \ 5 \ 4 \ 3 \ 2 \ 1.
\end{align*}
\]

APPENDIX

A.1 A Generalized Maximum Likelihood Nonparametric Estimator \( \hat{\pi}_M \) of \( \pi_M \)

Let \( \gamma_i \) be a realization of \( \gamma_i \) and \( \delta_i \) a realization of \( \delta_i \), for \( i = 1, 2, \ldots, n \). Let
\[
u_i < u_2 < \cdots < u_k
\]
be distinct values among \( y_1, y_2, \ldots, y_n \) and \( M \). Let
\[
d_i = \text{number dying due to cancer at time } u_i,
\]
\[
s_i = \text{number dying due to other causes at time } u_i,
\]
\[
c_i = \text{number leaving the study at time } u_i,
\]
and
\[
n_i = \sum_{j=1}^{u_i} (d_j + s_j + c_j)
\]
= number of subjects in the study active throughout the time interval \([0, u_i] \),

for \( i = 1, 2, \ldots, k \). We want \( c_i \) to be defined in a special way. The number \( c_i \) should include those \( c_i \) who survived the entire project in addition to those who left at time \( u_i \). If everyone in the sample can be observed until everyone either dies or leaves the study \( (M = \infty) \), of course, \( c_i = 0 \). Note that \( n_i = n \) and that \( u_i > 0 \).

Following Johansen (1978) and using a crucial idea due to Kiefer and Wolfowitz (1956), it can be shown that the generalized maximum likelihood estimators of \( F \) and \( G \) are the usual \( KM \) estimators.

If \( c_i = 0 \), then \( \pi \) can be estimated by
\[
\hat{\pi} = \sum_{i=1}^{k} \frac{d_i}{n_i} \prod_{j=1}^{i-1} (1 - (d_j + s_j)/n_j).
\]

What this means is that, if every member of the sample is either dead at \( u_i \) or earlier or leaves the study before \( u_i \), then we can estimate \( \pi \). If \( c_i > 0 \), there is more than one solution to the likelihood equations. Hence \( \pi, F_1 \), and \( F_2 \) are not estimable. Note that \( P(c_i > 0) > 0 \) if \( F(t) > 0 \) and \( G(t) > 0 \), for all \( t > 0 \).

In the case the project is terminated at \( M \), we can estimate \( \pi_M \). Recall \( \pi_M = \pi F_1(M) \). The generalized maximum likelihood estimator of \( \pi_M \) can be shown to be
\[
\hat{\pi}_M = \sum_{i=1}^{n} \prod_{j=1}^{i-1} (1 - (d_j + s_j)/n_j).
\]

If \( F \) and \( G \) are continuous, then, with probability \( 1, k < n \) and \( d_i + s_i + c_i = 1 \), for all \( j = 1, 2, \ldots, k - 1 \). It can be verified that
\[
\hat{\pi}_M = (1/n) \sum_{i=1}^{n} d_i \prod_{j=1}^{i-1} (1 - c_j/n_j)^{-1}
= (1/n) \sum_{i=1}^{n} d_i \frac{[\hat{G}(u_i - 1)]^{-1}}{[\hat{G}(u_i - 1)]^{-1}},
\]

where \( \hat{G} \) is the \( K-M \) estimator of \( G \). Thus with probability one, the generalized maximum likelihood estimator of \( \pi_M \) is given by
\[
\hat{\pi}_M = (1/n) \sum_{i=1}^{n} I(\Delta_i = 1)[\hat{G}(Y_i - 1)]^{-1}.
\]

Similarly, the generalized maximum likelihood estimator of \( \pi F_1(t) \) for \( 0 \leq t \leq M \) is given by
\[
\hat{\pi}_F(t) = (1/n) \sum_{i=1}^{n} I(\Delta_i = 1, Y_i \leq t)[\hat{G}(Y_i - 1)]^{-1}
\]

A.2 Proof of Theorem 1

We give a proof of Theorem 1 here. The following lemmas are helpful in this connection.

Lemma B.1. Let
\[
\overline{\eta}(t) = (1/n) \sum_{i=1}^{n} \eta(t).
\]

Then we have
\[
E \left( n^{-1} \sum_{i=1}^{n} \left( \overline{\eta}(Y_i) \right)^{k} \right) = O(n^{-k}),
\]
and
\[
\sum_{i=1}^{n} P \left( n^{-1} \sum_{i=1}^{n} \left( \overline{\eta}(Y_i) \right)^{k} > \alpha \right) < \infty.
\]

Proof. Note that \( \overline{\eta}(Y_i) \), for \( i = 1, 2, \ldots, n \), are identically distributed. Then the left side of (A.3) is less than or equal to
\[
E \left( \sum_{i=1}^{n} \overline{\eta(\gamma_i)} \right)^4 = E(\overline{\eta(\gamma_i)})^4 \\
\leq O(n^{-\delta}) + E \left( \sum_{i=1}^{n} \eta(\gamma_i) \right)^4 \\
= O(n^{-\delta}) + E \left( \sum_{i=1}^{n} \eta(\gamma_i) \right)^4 + O(n^{-\delta}) \\
= O(n^{-\delta}).
\]
The last equality follows since, given \(Y_1, \eta(\gamma_i)\), for \(j = 2, 3, \ldots, n\), are iid bounded random variables with \(E(\eta(\gamma_i)|Y_1) = 0\). Now (A.4) follows from (A.3).

**Lemma B.2.** Let \(Z_1, Z_2, \ldots, Z_n\) be iid random vectors. Let \(h\) be a bounded measurable function on the domain of \((Z_1, Z_2)\) such that \(E(h(Z_1, Z_2)|Z_1) = 0\), for \(i = 1, 2\). Then
\[
E \left( \sum_{i=1}^{n} W_j \right)^4 = O(1),
\]
where
\[
W_j = \frac{1}{n} \sum_{i=1}^{n} h(Z_i, Z_j).
\]

**Proof.** As the \(W_j\)’s are identically distributed,
\[
E \left( \sum_{i=1}^{n} W_j \right)^4 = nE(W_i^4) + 6n(n-1)(E(W_i^2W_j^2) + E(W_iW_j^3)) \\
+ 12n(n-1)(n-2)E(W_iW_j^2) \\
+ 24n(n-1)(n-2)(n-3)E(W_iW_jW_kW_u).
\]
Let \(h(t) = h(Z_i, t)\). Since, for distinct \(i, j, u, v > 1\),
\[
E(h(Z_i)h(Z_j)h(Z_i)h(Z_j)) = 0,
\]
it follows that \(EW_i^4 = O(n^{-1})\). Since, for \(i \neq 1, j \neq u, v \neq 2\), and \(j \neq 2\),
\[
E(h_i(Z_i)h_j(Z_j)h_i(Z_j)h_j(Z_j)) = 0,
\]
and, for \(\{i, u\} \neq \{j, v\}, i \neq u, j \neq v, 1 \neq \{i, u\}, \) and \(2 \neq \{j, v\}\),
\[
E(h_i(Z_i)h_j(Z_j)h_j(Z_j)h_i(Z_i)) = 0,
\]
It follows that \(E(W_i^2W_j^2) = O(n^{-2})\). These, along with similar estimates, yield that
\[
E \left( \sum_{i=1}^{n} W_j \right)^4 = O(1).
\]

**Proof of Theorem 1.** The estimate (2.7) is an immediate consequence of (2.3) and Lemma 1. Note that, with probability 1,
\[
(1/n)\#\{j: Y_j = M\} = (1/n)\#\{j: \min(T_j, C_j) > M\} \rightarrow H(M).
\]
As a consequence, we have, uniformly in \(i\), with probability 1, that
\[
I(\Delta_i = 1)\overline{G(Y_i - \overline{G(Y_i)})} \\
\leq I(\Delta_i = 1)(\#\{j: Y_j \geq Y_i\})^{-1} \\
\leq (\#\{j: Y_j = M\})^{-1} = O(n^{-1}). \quad (A.5)
\]
From (A.5), (2.2), and (2.5), we have, with probability 1,
\[
\hat{\sigma}_M = (1/n) \sum_{i=1}^{n} I(\Delta_i = 1)(\overline{G(Y_i)})^{-1} \left[ 1 + (\overline{\eta(\gamma_i)}/\overline{\eta(\gamma_i)}) \right] \\
+ O(\sigma_M + O(\overline{\eta(\gamma_i)^2})) \\
= (1/n) \sum_{i=1}^{n} I(\Delta_i = 1)(\overline{G(Y_i)})^{-1} \left[ 1 + (1/n) \sum_{i=1}^{n} J_i \overline{\eta(\gamma_i)} \right] \\
+ O(\sigma_M + O(\overline{\eta(\gamma_i)^2})). \quad (A.6)
\]
where
\[
J_i = I(\Delta_i = 1)\overline{G(Y_i)}^{-2}.
\]
Recall that
\[
\eta(t, \delta) = E(\eta(t, \delta, Y_i, J_i).
\]
By applying Lemma B.2 to
\[
h(Y_i, \Delta_i, (Y_i, \Delta_i)) = \eta(Y_i, \Delta_i, Y_i, J_i - \eta(Y_i, \Delta_i),
\]
we get, with probability 1,
\[
(1/n) \sum_{i=1}^{n} J_i \overline{\eta(\gamma_i)} = (1/n) \sum_{i=1}^{n} \eta(\gamma_i, \Delta_i) + O(\sigma_M)
\]
The result now follows from (A.4) and (A.6).

[Received February 1990. Revised March 1991.]

**REFERENCES**


