Statistical Science
2004, Vol. 19, No. 4, 644–651
DOI 10.1214/088342304000000521

Nonparametric Methods in Reliability

Myles Hollander and Edsel A. Peña

Abstract. Probabilistic and statistical models for the occurrence of a recurrent event over time are described. These models have applicability in the reliability, engineering, biomedical and other areas where a series of events occurs for an experimental unit as time progresses. Nonparametric inference methods, in particular, the estimation of a relevant distribution function, are described.

Key words and phrases: Aalen–Nelson estimator, counting process, frailty, Gaussian process, martingale, minimal repair, perfect repair, product-limit estimator.

1. INTRODUCTION

Recurrent events are prevalent in reliability and engineering studies such as monitoring the status of a repairable system; in biomedical and public health settings, for example, keeping track of hospitalization visits of a patient with a chronic disease; in economics such as when there is a drop of 200 points in the Dow Jones index during a trading day; and in sociology such as the commission of a criminal act by a delinquent youth. Therefore, it is imperative to have stochastic models for the occurrence of a recurrent event and to have appropriate statistical methods for making inference about model parameters.

Data that arise from monitoring recurrent events usually come in the form depicted in Figure 1, which is a pictorial representation of the successive migratory motor complex (MMC) periods for 19 subjects in a gastroenterology study. This data set was analyzed by Aalen and Husebye (1991), and one of the questions posed was whether the successive MMC periods have lengths that are independent and identically distributed, the so-called IID renewal assumption. Another problem is to estimate the common distribution of the MMC periods if the IID renewal assumption is valid. For each unit, there is a random monitoring period, and the random number of event occurrences is determined by the interplay between the lengths of the successive periods and the length of the monitoring period. Furthermore, for each unit the last MMC period is always right-censored since its length is only known to exceed the monitoring length minus the (calendar) time of the last observed event occurrence.

In Section 2 we discuss the problem of estimating the common interevent distribution under the IID renewal model when data of the form in Figure 1 are available. Before considering the estimation of the interevent distribution, which was the problem considered by Peña, Strawderman and Hollander (2001), we first point out some subtleties inherent in the data structure. Section 3 deals with a more general model for recurrent event data proposed and discussed by Dorado, Hollander and Sethuraman (1997). This model, introduced in the context of repairable systems, covers the IID renewal model and includes as special cases other models in the literature. Nonparametric estimation of a distribution function for this general model will be discussed. We have chosen to present the interconnected subjects of Sections 2 and 3 because they are important topics that are receiving a lot of attention from researchers who are developing nonparametric methods for reliability and survival analysis, and for which additional results, sparked by those described here, are currently being developed.

Myles Hollander is Robert O. Lawton Distinguished Professor and Chairman, Department of Statistics, Florida State University, Tallahassee, Florida 32306, USA (e-mail: holland@stat.fsu.edu). Edsel A. Peña is Professor, Department of Statistics, University of South Carolina, Columbia, South Carolina 29208, USA (e-mail: pena@stat.sc.edu).

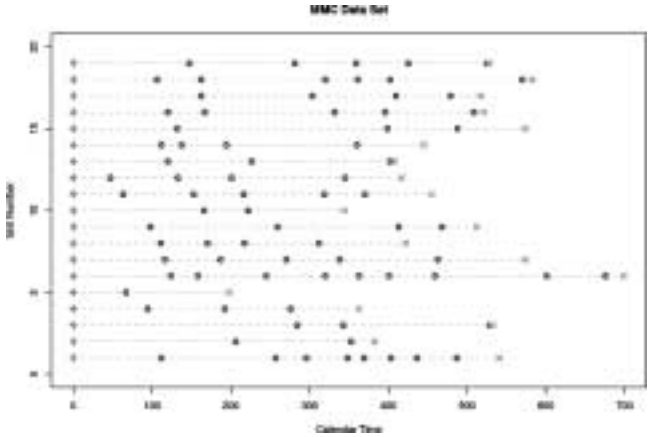

FIG. 1. *Pictorial representation of the MMC data that contains successive MMC periods for* 19 *subjects. The multiplication marks* $(\times)$ *indicate the end of monitoring periods for each of the subjects.*

2. ESTIMATION OF INTEREVENT DISTRIBUTION

We now consider the problem of estimating the interevent distribution when data of the form in Figure 1 are available. To introduce notation for this recurrent event data accrual, we suppose that for unit i out of n units, the recurrent event process is observed over the random period $[0, \tau_i]$, where τ_i, $i = 1, 2, \ldots, n$, are i.i.d. according to distribution G. The successive interevent times, denoted by T_{ij}, $j = 1, 2, \ldots$, are assumed to be i.i.d. from an unknown continuous distribution F. The successive calendar times of event occurrences for the ith unit are denoted by

$$0 \equiv S_{i0} < S_{i1} < S_{i2} < S_{i3} < \cdots \quad \text{with } S_{ij} = \sum_{k=1}^{j} T_{ij}.$$

The number of events that occurred on or before calendar time s for unit i is denoted by $N_i^\dagger(s)$, so

$$\begin{aligned} N_i^\dagger(s) &= \max\bigl\{ j \in \{0, 1, 2, \ldots\} : S_{ij} \le (s \wedge \tau_i) \bigr\} \\ &= \sum_{j=1}^{\infty} I\{S_{ij} \le (s \wedge \tau_i)\}, \end{aligned}$$

where $a \wedge b = \min\{a, b\}$. We denote by $K_i = N_i^\dagger(\infty)$ the total number of observed events over $[0, \tau_i]$ for the ith unit.

We first discuss some subtleties regarding such a data accrual scheme. First, we note that K_i is informative about the distribution F and, in some cases, conditionally on τ_i, it contains all information about F. To see this latter point, consider the case where F is exponential with mean $1/\lambda$ [i.e., $F = \mathrm{EXP}(\lambda)$]. Then K_i has a Poisson distribution with mean $\tau_i \lambda$ and, conditioning on $K_i = k_i$, $(S_{i1}, S_{i2}, \ldots, S_{iK_i})/\tau_i$ has the same distribution as the order statistics of a sample of size k_i from a standard uniform distribution by virtue of a uniformity property of Poisson processes (cf. Resnick, 1992). This implies that K_i is a sufficient statistic for λ, hence it carries all information about λ. Second, observe that there is always a right-censored observation and, even if τ_i is independent of the T_{ij}'s, the censoring variable for T_{iK_i+1} is not independent of the T_{ij}'s since this censoring variable is $C_i = \tau_i - S_{iK_i}$. Thus, the data accrual scheme leads to dependent and informative censoring. These two traits make this recurrent event setting different from the usual random censorship model typically considered in failure-time analysis, and as such demand a more delicate treatment when making inference about F or its associated parameters. In particular, it is erroneous to treat the T_{ij}'s as complete data and the $(\tau_i - S_{iK_i})$'s as censored data in the usual way and to apply estimation methods developed under the random censorship model.

For the nonparametric estimation of the interevent distribution F, we define doubly indexed processes for the ith unit via

$$N_i(s,t) = \sum_{j=1}^{N_i^\dagger(s)} I\{T_{ij} \le t\}, \tag{1}$$

$$\begin{aligned} Y_i(s,t) = &\sum_{j=1}^{N_i^\dagger(s-)} I\{T_{ij} \ge t\} \\ &+ I\bigl\{(s \wedge \tau_i) - S_{iN_i^\dagger(s-)} \ge t\bigr\}. \end{aligned} \tag{2}$$

The process $N_i(s,t)$ represents the number of events that occurred on or before time s whose interevent times are at most t, whereas $Y_i(s,t)$ represents the number of events over $[0, s]$ whose interevent times are at least t plus a count on whether the right-censored last interevent time is also at least t. Observe that both of these processes have a random number of summands, since $N_i^\dagger(s)$ and $N_i^\dagger(s-)$ are both random. In addition, they are dependent on the T_{ij}'s and $(s \wedge \tau_i) - S_{iN_i^\dagger(s-)}$. These dependencies, which arise because of the sum–quota accrual scheme, make this recurrent event setting more difficult and interesting. For instance, as contained in Theorem 1, the expected values of the above processes depend on the renewal function of F, which

is defined according to

$$\rho(t) = \sum_{k=1}^{\infty} F^{*k}(t), \tag{3}$$

where F^{*k} is the kth convolution of F, the distribution of S_{ik}. In the sequel we denote by $\Lambda(t)$ the cumulative hazard function of $F(t)$.

THEOREM 1. *Take* $G_s(t) = G(t)I\{t < s\} + I\{t \geq s\}$ *and* $\bar{G}_s(t) = 1 - G_s(t)$. *Then under the IID renewal model,*

$$\begin{aligned} &E\{N_i(s,t)\} \\ &\quad = \int_0^{\infty} \Big\{ F(w \wedge t) \\ &\qquad\quad + \int_0^w F\big((w-v) \wedge t\big)\rho(dv) \Big\} G_s(dw), \\ &E\{Y_i(s,t)\} = y(s,t) \\ &\qquad \equiv \bar{F}(t)\bar{G}_s(t-) \\ &\qquad\quad \cdot \Big\{ 1 + \frac{1}{\bar{G}_s(t-)} \int_t^{\infty} \rho(w-t) G_s(dw) \Big\}. \end{aligned}$$

In particular, if $F = \mathrm{EXP}(\lambda)$ *and* $G = \mathrm{EXP}(\eta)$, *and we let* $s \to \infty$, *we obtain*

$$E\{N_i(\infty,t)\} = \frac{\lambda}{\eta}\big(1 - \exp\{-(\lambda+\eta)t\}\big)$$

and

$$y(\infty,t) = \left(1 + \frac{\lambda}{\eta}\right) \exp\{-(\lambda+\eta)t\}.$$

PROOF. The proofs of these results use renewal arguments. We present here only the proof for the first result, because that for the second result uses the same ideas. For a given s, let $\tau_i(s) = s \wedge \tau_i$, which has distribution G_s. By first conditioning on $\tau_i(s) = w$, we let $H_t(w) \equiv E\{N_i(\tau_i(s), t) | \tau_i(s) = w\}$. Then we obtain

$$\begin{aligned} &H_t(w) \\ &\quad = \int_0^w E\{N_i(w,t) | T_{i1} = v\} F(dv) \\ &\qquad + \int_w^{\infty} E\{N_i(w,t) | T_{i1} = v\} F(dv) \\ &\quad = \int_0^w \big\{ I\{v \leq t\} \\ &\qquad\quad + E\{N_i(\tau_i(s), t) | \tau_i(s) = w - v\} \big\} F(dv) + 0 \\ &\quad = F(w \wedge t) + \int_0^w H_t(w-v) F(dv), \end{aligned}$$

which forms a renewal equation. Invoking the renewal equation theorem (Resnick, 1992), the solution to this equation is given by

$$H_t(w) = F(w \wedge t) + \int_0^w F\big((w-v) \wedge t\big)\rho(dv).$$

Integrating out with respect to the distribution $G_s(w)$, we therefore obtain the first result. The particular results associated with exponential distributions follow routinely by noting that the renewal function for $\mathrm{EXP}(\lambda)$ is $\rho(v) = \lambda v$. □

Notice that for the special case of $F = \mathrm{EXP}(\lambda)$ and $G = \mathrm{EXP}(\eta)$, since

$$\begin{aligned} \int_0^t y(\infty, w)\, dw &= \left(1 + \frac{\lambda}{\eta}\right) \int_0^t \exp\{-(\lambda+\eta)w\}\, dw \\ &= \frac{1}{\eta}\big(1 - \exp\{-(\lambda+\eta)t\}\big), \end{aligned}$$

we have that $E\{N_i(\infty,t) - \lambda \int_0^t Y_i(\infty, w)\, dw\} = 0$. This is a particular manifestation of the more general result that

$$E\left\{ N_i(s,t) - \int_0^t Y_i(s,w)\Lambda(dw) \right\} = 0,$$

which motivates the Aalen–Nelson estimator $\Lambda(t)$ and, consequently, the product-limit estimator F which is given subsequently.

The aggregated processes based on n units are denoted by

$$N(s,t) = \sum_{i=1}^n N_i(s,t) \quad \text{and} \quad Y(s,t) = \sum_{i=1}^n Y_i(s,t).$$

The resulting product-limit type estimator of $\bar{F} = 1 - F$ based on data that have accrued over the calendar time $[0, s]$ for n units is

$$\hat{\bar{F}}_n(s,t) = \prod_{w=0}^{t} \left[1 - \frac{N(s, dw)}{Y(s,w)} \right]. \tag{4}$$

Utilizing the ideas of Sellke (1988) and Gill (1980), the following asymptotic properties of this product-limit type estimator were established by Peña, Strawderman and Hollander (2001).

THEOREM 2. *If* $t^* \in (0, \infty)$ *is such that* $y(s,t^*) > 0$ *and* $\Lambda(t^*) < \infty$, *then, as* $n \to \infty$:

(i) $\sup_{0 \leq t \leq t^*} |\hat{\bar{F}}_n(s,t) - \bar{F}(t)|$ *converges in probability to zero;*

(ii) *the process* $\{W_n(s,t) = \sqrt{n}[\hat{\bar{F}}_n(s,t) - \bar{F}(t)] : 0 \le t \le t^*\}$ *converges weakly in Skorohod space* $\mathcal{D}[0,t^*]$ *to a zero-mean Gaussian process* $\{W^\infty(s,t): 0 \le t \le t^*\}$ *whose variance function is*

$$\operatorname{Var}\{W^\infty(s,t)\} = \sigma^2(s,t) \equiv \bar{F}(t)^2 \int_0^t \frac{\Lambda(dw)}{y(s,w)}.$$

In particular, if $F = \mathrm{EXP}(\lambda)$ *and* $G = \mathrm{EXP}(\eta)$, *then*

$$\sigma^2(\infty,t) = \frac{\lambda\eta}{(\lambda+\eta)^2}\exp\{-2\lambda t\}\big(\exp\{(\lambda+\eta)t\} - 1\big).$$

A possible nonparametric estimator of the variance of $\hat{\bar{F}}_n(s,t)$ is given by

$$\hat{\sigma}_n^2(s,t) = \hat{\bar{F}}_n(s,t)^2 \int_0^t \frac{N(s,dw)}{Y(s,w)[Y(s,w) - N(s,\Delta w)]}.$$

Together with the weak convergence result, this estimate of the variance could be utilized to form a $100(1-\gamma)\%$ asymptotic confidence interval for $\bar{F}(t)$ given by

$$\big[\hat{\bar{F}}_n(s,t) \pm z_{\gamma/2}\hat{\sigma}_n(s,t)\big],$$

where $z_{\gamma/2}$ is the $100(1-\gamma/2)\%$ percentile of the standard normal distribution. We note in passing that it is possible to develop a Hall and Wellner (1980) simultaneous confidence band in this recurrent event setting.

Notice that the asymptotic results for $\hat{\bar{F}}(s,t)$ are analogous to properties of the product-limit estimator for single-event settings, except that the limiting variance function in this recurrent event setting involves the renewal function ρ of the distribution F. The entry of this renewal function into the limiting variance function is a manifestation of the sum–quota accrual scheme, which forces the number of events for the ith unit observed over $[0,\tau_i]$ to be informative and makes the censoring mechanism of the last event informative and dependent as well.

Peña, Strawderman and Hollander (2001) also considered a model wherein the interevent times for a unit are correlated. This dependence among the interevent times is induced by an unobserved latent or frailty variable. To describe this correlated recurrent event model, it is postulated that there is an unobserved Z_i, with $Z_1, Z_2, \ldots, Z_n$ i.i.d. random variables from a distribution H_Z, which is taken in particular to be a gamma distribution with mean 1 and variance $1/\alpha$, where $\alpha > 0$ is unknown. Note that the gamma distribution for this frailty variable has the same shape and scale parameter, and this is to achieve model identifiability. Given $Z_i = z$, it is assumed that the interevent times $T_{i1}, T_{i2}, \ldots$ are i.i.d. with survivor function

$$\bar{F}(t|z) = P\{T_{ij} > t | Z_i = z\} = [\bar{F}_0(t)]^z. \tag{5}$$

This is equivalent to postulating that the conditional hazard function of T_{ij}, given $Z_i = z$, is $\Lambda(t|z) = z\Lambda_0(t)$, where $\Lambda_0 = -\log \bar{F}_0$ is the hazard function of F_0. As a consequence, the joint survivor function of $(T_{i1}, T_{i2}, \ldots, T_{ik})$ for fixed k is given by

$$\begin{aligned} &P\{T_{i1} > t_1, T_{i2} > t_2, \ldots, T_{ik} > t_k\} \\ &\quad = \int_0^\infty \left[\prod_{j=1}^k \bar{F}_0(t_j)\right]^z \frac{\alpha^\alpha}{\Gamma(\alpha)} z^{\alpha-1}\exp\{-\alpha z\}\,dz \\ &\quad = \left[\frac{\alpha}{\alpha + \sum_{j=1}^k \Lambda_0(t_j)}\right]^\alpha. \end{aligned}$$

From this, by setting $t_j = t$ and $t_l = 0, l \ne j$, we immediately see that the interevent times are dependent and the common marginal survivor function of the interevent times is

$$\bar{F}(t) = P\{T_{ij} > t\} = \left[\frac{\alpha}{\alpha + \Lambda_0(t)}\right]^\alpha. \tag{6}$$

The semiparametric estimation of this marginal survivor function was discussed by Peña, Strawderman and Hollander (2001). Mimicking the ideas of Nielsen, Gill, Andersen and Sørensen (1992), the computation of the estimator relies on the expectation–maximization (EM) algorithm (see Dempster, Laird and Rubin, 1977), where the frailty values are considered as missing values. Given values of $(Z_1, Z_2, \ldots, Z_n)$, say $(\hat{z}_1, \hat{z}_2, \ldots, \hat{z}_n)$, the first part of the M step of the algorithm is to obtain the conditional estimate of Λ_0 given by

$$\hat{\Lambda}_0(s,t|\hat{z}_1,\ldots,\hat{z}_n) = \int_0^t \frac{\sum_{i=1}^n N_i(s,dw)}{\sum_{i=1}^n \hat{z}_i Y_i(s,w)}.$$

The second part of the M step of the algorithm is to maximize a marginal likelihood function for α, given values of $\hat{\Lambda}_0$ and $\hat{z}_i$. To describe this marginal likelihood, define

$$Y_i^\dagger(s) = I\{\tau_i \ge s\} \quad \text{and} \quad R_i(s) = (s \wedge \tau_i) - S_{iN_i^\dagger(s-)}.$$

Note that $R_i(s)$ is the backward recurrence time at s. Then the marginal likelihood for obtaining the estimate

of α is given by

$$L_F(s;\alpha)=\prod_{i=1}^{n}\Bigg\{\frac{\Gamma(\alpha+N_i^{\dagger}(s))}{\Gamma(\alpha)}\cdot\Bigg[\frac{\alpha}{\alpha+\int_0^s Y_i^{\dagger}(v)\,d\Lambda_0[R_i(v)]}\Bigg]^{\alpha+N_i^{\dagger}(s)}\cdot\Bigg(\prod_{v\le s}\Bigg[\frac{Y_i^{\dagger}(v)\,d\Lambda_0[R_i(v)]}{\alpha}\Bigg]^{N_i^{\dagger}(\Delta v)}\Bigg)\Bigg\}.$$

Given $\hat{\Lambda}_0(s,t|\hat{z}_1,\ldots,\hat{z}_n)$, $d\Lambda_0[R_i(v)]$ is replaced by the jump of $\hat{\Lambda}_0(s,\cdot)$ at $t=R_i(v)$. The maximization of this marginal likelihood with respect to α is facilitated by iterative procedures, such as the Newton–Raphson algorithm. On the other hand, the E step of the algorithm proceeds by obtaining the values of the Z_i's, given $\hat{\alpha}$ and $\hat{\Lambda}_0$, according to the formula

$$\hat{z}_i=\frac{\hat{\alpha}+N_i^{\dagger}(s)}{\hat{\alpha}+\int_0^s Y_i^{\dagger}(v)\,d\hat{\Lambda}_0[s,R_i(v)]},\quad i=1,2,\ldots,n.$$

The E and M steps are then iterated alternately until convergence is achieved. Finally, having obtained the estimate $\hat{\Lambda}_0(s,\cdot)$ and $\hat{\alpha}$, the estimate of the marginal survivor function $\bar{F}$ is

$$\hat{\bar{F}}(s,t)=\Bigg[\frac{\hat{\alpha}}{\hat{\alpha}+\hat{\Lambda}_0(s,t)}\Bigg]^{\hat{\alpha}}. \tag{7}$$

A competing estimator that was proposed by Wang and Chang (1999) applies even if the frailty components are not gamma distributed; hence, their estimator is more general. In Peña, Strawderman and Hollander (2001), these two estimators, as well as the estimator that ignored the frailty components, were compared in terms of their bias and mean-squared error functions. It was found that if the gamma frailty model holds, then the semiparametric estimator in (7) outperforms the Wang–Chang estimator. The comparisons, which were done through computer simulation studies, also demonstrated that the estimator that ignored the frailty components has a nonnegligible systematic bias and, hence, is not a viable estimator of the marginal survivor function of the interevent times.

The two estimators of the marginal survivor function discussed above, together with the estimator of Wang and Chang (1999), were illustrated by Peña, Strawderman and Hollander (2001) using the gastroenterology data set presented in Figure 1. The plots of the three survivor function estimates for this data set are provided in Figure 2. As pointed out by Peña, Strawderman and Hollander (2001), the fact that the three curves are quite close to each other indicates that there is no need for the frailty component and that the renewal assumption is viable. The resulting estimate of α obtained from the EM algorithm was $\hat{\alpha}=10.18$, which was judged to indicate a weak association among the interevent times.

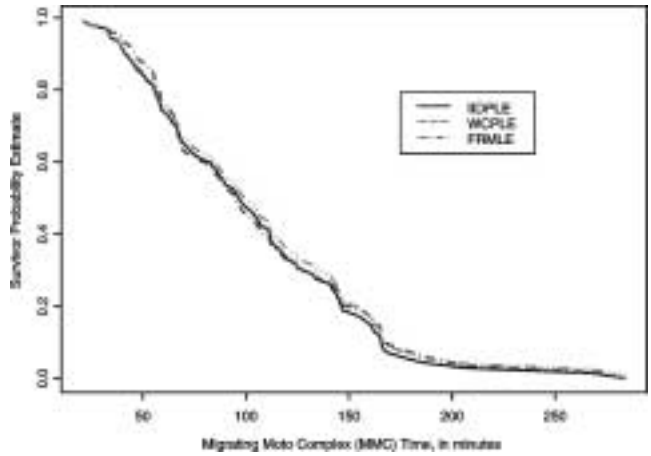

FIG. 2. *Plots of the three survivor function estimates for the MMC data set*: IIDPLE *is the estimate obtained by assuming the no-frailty IID renewal model*; WCPLE *is the estimate of Wang and Chang* (1999); *and* FRMLE *is the gamma frailty-based semiparametric estimate. The maximum likelihood estimate of the frailty parameter* α *under the gamma frailty model is* $\hat{\alpha}=10.17562$ *or, equivalently,* $\hat{\xi}=\hat{\alpha}/(1+\hat{\alpha})=0.9105$.

3. REPAIR MODELS

A more general model which subsumes the IID renewal model considered in the preceding section is that proposed by Dorado, Hollander and Sethuraman (DHS; 1997). Their general repair model also contains many of the models in the literature and introduces new models as well. They considered the family of survival functions $\bar{F}_a^{\theta}(x)=\bar{F}(\theta x+a)/\bar{F}(a)$. The family of distributions $\{F_a^{\theta}\}$ is stochastically ordered in θ. That is, $\theta\le\theta'$ implies $F_a^{\theta}\overset{\text{st}}{\ge}F_a^{\theta'}$ for each a, so that $F_a^{\theta}(t)\le F_a^{\theta'}(t)$ for every t. The DHS general repair model depends on two sequences $\{A_j\}_{j\ge1}$ and $\{\theta_j\}_{j\ge1}$ known as the effective ages and life supplements, respectively. These sequences satisfy

$$A_1=0,\quad \theta_1=1,\quad A_j\ge0,\quad \theta_j\in(0,1],\quad A_j\le A_{j-1}+\theta_{j-1}T_{j-1},\quad j\ge2. \tag{8}$$

The joint distributions of the interfailure times are given as

$$P(T_j\le t|A_1,\ldots,A_j,\theta_1,\ldots,\theta_j,T_1,\ldots,T_{j-1})=F_{A_j}^{\theta_j}(t) \tag{9}$$

for $t > 0$, $j \geq 1$. From (8) and (9) we see that for $j \geq 1$, the effective age of the system after the jth repair is less than the effective age $X_j \stackrel{\text{def}}{=} A_j + \theta_j T_j$ just before the jth failure, and since $\theta_j \leq 1$, X_j is less than the actual age S_j. The use of the term "supplemental life" has the following motivation. If a minimal repair, which restores the unit to an effective age just before it failed, was performed at the time of the first failure, T_2 would have the distribution $F^1_{T_1}$. A longer expected life for T_2 is provided, however, if we use the distribution $F^{\theta_2}_{T_1}$ for some θ_2 satisfying $0 < \theta_2 < 1$. Starting with the distribution $F^{\theta_2}_{T_1}$ for T_2 and applying minimal repair after the second failure, T_3 would have the distribution $F^1_{A_3}$, where $A_3 = T_1 + \theta_2 T_2$. If we seek a longer expected life for T_3, we can use the distribution $F^{\theta_3}_{A_3}$ for some θ_3 satisfying $0 < \theta_3 < 1$. By continuing in this way, we obtain the supplemented life model. Under this model, the system has a larger expected remaining life than it would have under minimal repair.

It is also of interest to consider monotonicity properties of the expected interfailure times. Theorem 3, due to Dorado (1995), is a typical result. We first give the definition of a decreasing mean residual life distribution. The mean residual life (MRL) function corresponding to F is

$$\varepsilon_F(x) = \left\{ \int_x^\infty \bar{F}(y)\, dy \right\} \Big/ \bar{F}(x).$$

DEFINITION 1. A failure distribution F is said to be a decreasing mean residual life (DMRL) distribution if the mean $\varepsilon_F(0)$ is finite and $\varepsilon_F(s) \geq \varepsilon_F(t)$ for all $0 \leq s \leq t$.

THEOREM 3. *Assume in the DHS model of* (9) *that the* $\{\theta_j\}_{j\geq 1}$ *and* $\{A_j\}_{j\geq 1}$ *are increasing sequences and* F *is DMRL. Then* $E(T_j)$ *is decreasing in* j.

PROOF. We have that

$$\begin{aligned} E(T_j) &= \int_0^\infty P(T_j > t)\, dt \\ &= \int_\Omega \int_0^\infty \frac{\bar{F}(\theta_j t + A_j)}{\bar{F}(A_j)}\, dt\, dP \\ &\leq \int_\Omega \int_0^\infty \frac{\bar{F}(\theta_j t + A_{j-1})}{\bar{F}(A_{j-1})}\, dt\, dP \\ &\leq \int_\Omega \int_0^\infty \frac{\bar{F}(\theta_{j-1} t + A_{j-1})}{\bar{F}(A_{j-1})}\, dt\, dP \\ &= E(T_{j-1}). \end{aligned}$$

The first inequality follows from the fact that F is DMRL and that the A_j's are increasing. The second inequality holds since the θ_j's are increasing. □

We now discuss the estimation of the distribution F in this general repair model. In contrast to the situation in the preceding section, we postulate that the repair process is observed until a fixed time T. The effective age X_j prior to the jth failure is $A_j + \theta_j(S_j - S_{j-1})$ if $S_j \leq T$. If $S_{j-1} \leq T < S_j$, we cannot observe X_j and the effective age of the system at time T is $A_j + \theta_j(T - S_{j-1})$, which can be written as $X_j \wedge (A_j + \theta_j(T - S_{j-1}))$, a representation similar to that encountered in censored data. We define the processes

$$N(t) = \sum_j I(X_j \leq t, S_j \leq T),$$

$$Y(t) = \sum_j I\big(A_j < t \leq (X_j \wedge [A_j + \theta_j(T - S_{j-1})])\big).$$

Let $\delta_j = I(S_j \leq T)$ and set $\widetilde{X}_j = X_j \wedge [A_j + \theta_j \cdot (T - S_{j-1})]$. Then the random variables $\{(\widetilde{X}_1, \delta_1), (\widetilde{X}_2, \delta_2), \ldots\}$ can be thought of as observations coming from a censored model. A repair model observed during $[0, T]$ is similar to a survival study where a subject enters the study at A_j (the system at failure time S_{j-1} is repaired to effective age A_j) and either dies during the study at age X_j (a failure occurs) or leaves the study by $A_j + \theta_j(T - S_{j-1})$ (the system that was repaired at time S_{j-1} has not yet, by time T, suffered its next failure). From this viewpoint, $N(t)$ is the number of observed (uncensored) deaths by time t and $Y(t)$ is the number at risk at time t.

Next, we define the process

$$M(t) = N(t) - \int_0^t Y(s)\, d\Lambda(s).$$

Typically, analogous to results in censored data theory (Aalen, 1978; Fleming and Harrington, 1991), it is natural to try to establish that M is a martingale with respect to the history of N. This proved to be difficult, but DHS (1997) were able to show that the M process does have the same mean and covariance structure as if it were a martingale. They proved

$$E(M(t)) = 0, \tag{10}$$

$$\operatorname{cov}\big(M(t), M(t')\big) = \int_0^{t\wedge t'} E(Y)(1 - \Delta\Lambda)\, d\Lambda. \tag{11}$$

These results and techniques of Gill (1980) are sufficient to obtain asymptotic properties of the estimator of F. We sketch the development here and refer the reader to DHS (1997) for details.

We suppose that we observe n independent copies of the processes N and Y on a finite interval $[0, T]$, and let N_n and Y_n denote the sum of the first n copies. We wish to estimate F based on these observations. A natural estimator of the failure rate is N_n/Y_n, the ratio of observed deaths at time t to the number at risk at time t. Thus a natural estimator of the cumulative hazard function is the Aalen–Nelson estimator

$$\hat{\Lambda}_n(t) = \int_0^t \frac{J_n\, dN_n}{Y_n},$$

where $J_n(t) = I(Y_n(t) > 0)$ for $t \in (0, T]$. It is easy to see that F satisfies $F(t) = \int_0^t (1 - F(s-))\, d\Lambda(s)$ and hence we want an estimator $\hat{F}_n$ of F to satisfy $\hat{F}_n(t) = \int_0^t (1 - \hat{F}_n(s-))\, d\hat{\Lambda}_n(s)$. The solution of this Volterra integral equation is

$$\hat{\bar{F}}(t) = \prod_{s \le t} \big(1 - d\hat{\Lambda}_n(s)\big),$$

where $\prod_{s \le t}(1 - d\hat{\Lambda}_n(s))$ denotes the product integral (see Gill and Johansen, 1990; Andersen, Borgan, Gill and Keiding, 1993).

Let $M_n = N_n - \int Y_n\, d\Lambda$. This is the sum of n i.i.d. processes in $D[0, T]$ with mean 0 and covariance function given by (11). Thus $W_n(t) = n^{-1/2} M_n(t)$, $0 \le t \le T$, converges to a Gaussian process if tightness can be established. This was done in Theorem 5.1 of DHS (1997), who showed that

$$\frac{\hat{F}_n(t) - F(t)}{\bar{F}(t)} = \int \frac{\hat{\bar{F}}_n(s-) J_n(s)}{\bar{F}(s)(Y_n(s)/n)}\, dM_n(s). \tag{12}$$

Let

$$C(t) = \int_0^t \frac{dF}{EY(1 - F_-)}.$$

Assume that $F(T) < 1$ and F is an increasing failure rate distribution. From the continuous mapping theorem (see Billingsley, 1968) and a result on the uniform convergence of the integrand in (12), Corollary 5.1 of DHS (1997) shows

$$\sqrt{n}\left(\frac{\hat{F}_n - F}{\bar{F}}\right) \Rightarrow B(C) \quad \text{on } D[0, T],$$

where B denotes the Brownian motion on $[0, \infty)$. They also proved

$$\sqrt{n}\frac{\bar{K}}{\bar{F}}(\hat{F}_n - F) \Rightarrow B^0(K) \quad \text{on } D[0, T],$$

where B^0 denotes a Brownian bridge on $[0, 1]$ and $K = C/(1 + C)$.

Dorado, Hollander and Sethuraman (1997) also derived a simultaneous confidence band for F. For $t \in [0, T]$, let $L_n = I(\hat{F}_n(t) < 1)$ and set

$$\hat{C}_n(t) = \int_0^t J_n L_n\, d\hat{F}_n / [(Y_n/n)(1 - \hat{F}_n)]$$

and

$$\hat{K}_n(t) = \hat{C}_n(t)/\big(1 + \hat{C}_n(t)\big).$$

For t such that $\hat{F}_n(t) = 1$, set $\hat{K}_n(t) = 1$. A nonparametric asymptotic simultaneous confidence band for F with confidence coefficient at least $100(1 - \alpha)\%$ is

$$\big[\hat{F}_n \pm n^{-1/2} \lambda_\alpha \hat{\bar{F}}_n / \hat{K}_n\big], \tag{13}$$

where λ_α is such that $P(\sup_{t \in [0,1]} |B^0(t)| \le \lambda_\alpha) = 1 - \alpha$. Values of λ_α can be obtained from Hall and Wellner (1980) and Koziol and Byar (1975).

Let $X_{(1)}, X_{(2)}, \ldots, X_{(r)}$ be the distinct ordered values of the X's whose corresponding failure times are within $[0, T]$. Also, let δ_j be the number of observations with value $X_{(j)}$. Then for computational purposes we can use the simplified formulas

$$\hat{\bar{F}}_n(t) = \prod_{X_{(j)} \le t} \left(1 - \frac{\delta_j}{Y_n(X_{(j)})}\right)$$

and

$$\hat{C}_n(t) = n \sum_{X_{(j)} \le t} \frac{\hat{F}_n(X_{(j)}) - \hat{F}_n(X_{(j-1)})}{Y_n(X_{(j)})\hat{\bar{F}}_n(X_{(j)})}.$$

In practice, it may be that the data obtained lead to $\hat{F}_n(t_0) = 1$ for some $0 < t_0 < T$. When this happens, the data yield a confidence band only on the interval $[0, \sigma)$, where $\sigma = \inf\{t \in [0, T] : \hat{F}_n(t) = 1\}$.

Finally, note that if in Section 2 we let $\tau_i = T$ for all i and $s = T$, where T is the end of observation period in the current section, and if in addition we set $A_j = 0$ and $\Theta_j = 1$ for all j in the DHS model, then the product limit estimators in these two settings reduce to the estimator considered by Gill (1981).

ACKNOWLEDGMENTS

Myles Hollander's research was supported by NIH Grant 5 R01 DK52329 and NHLBI Grant 7 R01 HL67460. Edsel Peña's research was supported by NSF Grant DMS-01-02870, NIH Grant 2 R01 GM56182 and NIH COBRE Grant RR17698. Both authors thank R. Randles and a reviewer for helpful comments and suggestions.

REFERENCES

AALEN, O. (1978). Nonparametric inference for a family of counting processes. *Ann. Statist.* **6** 701–726.

AALEN, O. and HUSEBYE, E. (1991). Statistical analysis of repeated events forming renewal processes. *Statistics in Medicine* **10** 1227–1240.

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

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

DEMPSTER, A., LAIRD, N. and RUBIN, D. (1977). Maximum likelihood from incomplete data via the EM algorithm (with discussion). *J. Roy. Statist. Soc. Ser. B* **39** 1–38.

DORADO, C. (1995). On a general repair model for repairable systems. Ph.D. dissertation, Florida State Univ.

DORADO, C., HOLLANDER, M. and SETHURAMAN, J. (1997). Nonparametric estimation for a general repair model. *Ann. Statist.* **25** 1140–1160.

FLEMING, T. and HARRINGTON, D. (1991). *Counting Processes and Survival Analysis*. Wiley, New York.

GILL, R. (1980). Nonparametric estimation based on censored observations of a Markov renewal process. *Z. Wahrsch. Verw. Gebiete* **53** 97–116.

GILL, R. (1981). Testing with replacement and the product-limit estimator. *Ann. Statist.* **9** 853–860.

GILL, R. and JOHANSEN, S. (1990). A survey of product-integration with a view toward application in survival analysis. *Ann. Statist.* **18** 1501–1555.

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

KOZIOL, J. and BYAR, D. (1975). Percentage points of the asymptotic distributions of one and two sample K–S statistics for truncated or censored data. *Technometrics* **17** 507–510.

NIELSEN, G., GILL, R., ANDERSEN, P. and SØRENSEN, T. (1992). A counting process approach to maximum likelihood estimation in frailty models. *Scand. J. Statist.* **19** 25–43.

PEÑA, E. A., STRAWDERMAN, R. L. and HOLLANDER, M. (2001). Nonparametric estimation with recurrent event data. *J. Amer. Statist. Assoc.* **96** 1299–1315.

RESNICK, S. (1992). *Adventures in Stochastic Processes*. Birkhäuser, Boston.

SELLKE, T. (1988). Weak convergence of the Aalen estimator for a censored renewal process. In *Statistical Decision Theory and Related Topics IV* (S. Gupta and J. Berger, eds.) **2** 183–194. Springer, New York.

WANG, M.-C. and CHANG, S.-H. (1999). Nonparametric estimation of a recurrent survival function. *J. Amer. Statist. Assoc.* **94** 146–153.