Statistica Sinica **5**(1995), 279-290

ESTIMATION OF THE MEAN FUNCTION OF POINT PROCESSES BASED ON PANEL COUNT DATA

J. Sun and J. D. Kalbfleisch

University of Waterloo

Abstract. This article discusses estimation of the mean function of point processes when only incomplete data are available. Specifically, we consider situations in which each individual who gives rise to a point process is observed only at discrete time points and no information about the histories of the subject between observation times is available. Data structures of this type occur, for example, in many clinical trials and reliability studies in which it is impractical to keep subjects under observation over the entire study period. The main difficulty in estimating the mean function in such situations is that observation times usually differ between study subjects. In this paper, a simple and consistent estimator of the mean function of point processes is presented. Following two illustrative examples, a small simulation study demonstrates that the presented estimator is satisfactory in the cases considered.

Key words and phrases: Mean function, panel count data, point process, isotonic regression.

1. Introduction

This article considers estimation of the mean function (MF) of point processes when only incomplete data are available. Specifically, we consider situations in which each individual gives rise to a sequence of events over time. Let $N_i(t)$ represent the number of events that occur in $[0,t]$ from the ith subject with $N_i(0)=0$ and let $\Lambda(t)=E[N_i(t)]$ be the mean function of the process N_i, $i=1,\ldots,n$. Suppose that for each subject, observations are taken only at discrete time points and no information about the status between observation times is available. Thus the data for the ith individual consist only of the numbers of the events that have occurred prior to each observation time and the timing of events is not known. We refer to this kind of data as panel count data (see Kalbfleisch and Lawless (1981, 1985) and Groeneboom and Wellner (1992)). Our main focus will be on the estimation of the MF when the data on individuals are panel count data.

Panel count data arise if it is impractical to keep subjects under observation over the entire study period. For example, such data arise in clinical trials, especially of chronic diseases, such as studies of epileptic seizures, bladder cancer and gallstones, in which patients visit the clinic center for periodic evaluation. A reliability study of repairable systems in engineering provides another area in which panel count data are encountered. In this case, it is common to have collections of the same type of equipment or systems that generate point events (failures). Here, again, if the data are obtained from periodic inspection of the items under test and comprise counts of failures up to the inspection times, panel count data are obtained. For example, Gaver and O'Muircheartaigh (1987) discuss three sets of data arising from such reliability studies. Other similar situations occur, for example, in animal tumorgenicity experiments and in sociology. In sociology, panel count data are often referred to as event-count data.

The main difficulty in estimating the MF with panel count data is that observation times usually differ between subjects. Suppose, for example, that there are two individuals in a study and one observation is taken on each of the two subjects at time t_i; thus $N_i(t_i) = n_i$ is the number of events occurring by t_i from the ith individual, $i = 1, 2$. Suppose that the aim is to estimate the MF, $\Lambda(t)$, of the underlying process. It is obvious that only values of Λ at t_1 and t_2 are estimable. If $t_1 = t_2$, $\Lambda(t_1)$ can be simply estimated by $(n_1 + n_2)/2$. However, the estimation is not straightforward if $t_1 \neq t_2$.

An important special case of panel count data is current status data (e.g., see Diamond, McDonald, and Shah (1986) and Sun and Kalbfleisch (1993)), in which only one observation is taken for each subject in the study. For example, Sun and Kalbfleisch (1993) discuss a set of current status data on counts of multiple tumors in a tumorgenicity experiment. Table 1 presents another set of current status data from a reliability study of nuclear power generation systems. With current status data, several authors have supposed a Poisson process as the underlying process for which the estimation of the MF is equivalent to the simultaneous estimation of several Poisson means. For example, Gaver and O'Muircheartaigh (1987) propose an empirical Bayes procedure which assumes that all systems in the study arise from an independent sampling of a superpopulation.

Section 2 presents a simple and consistent estimator of a MF which is based on isotonic regression (Barlow, Bartholomew, Bremner, and Brunk (1972)). The idea is to pool all observations together and estimate the MF by taking into account its monotonic property. A similar estimator in the context of interval-censored data has been discussed by Groeneboom and Wellner (1992). In Section 3, two examples illustrate the approach. Following the examples, a small simulation study is presented which suggests that the proposed method is quite satisfactory. Section 4 concludes with some discussion.

As most authors (e.g., Thall and Lachin (1988)), we assume that the sequence of observation times for each individual is distributed independently of the underlying process. This includes, for example, situations in which the observation times are fixed in advance. If the times at which individuals are observed depend upon the underlying point process, the nature of this dependence would need to be modeled and the methods suggested here would, of course, be inappropriate.

2. Estimation of the MF

Suppose that there are n independent individuals in a study and that each gives rise to a point process. Let N_i and Λ be as defined in the previous section, and suppose that N_i is observed only at discrete time points. Let $0 < t_{i,1} < \cdots < t_{i,m_i}$ denote the observation times for the ith individual and $n_{i,j}$ denote the observation on the ith individual at time $t_{i,j}$, i.e., $n_{i,j} = N_i(t_{i,j})$, $j = 1, \ldots, m_i$, $i = 1, \ldots, n$. In the following, estimation of the $\Lambda(t)$ will be restricted to values of t corresponding to the distinct observation times.

We begin by examing some very simple situations.

2.1. Some preliminary remarks

Suppose first that $m_i = m$ and $t_{i,j} = t_j$, $i = 1, \ldots, n$, $j = 1, \ldots, m$, so that all items are observed at the same times $t_1, \ldots, t_m$. Since $\Lambda(t) = E\{N_i(t)\}$, $i = 1, \ldots n$, a natural estimator of $\Lambda(t_j)$ is

$$\hat{\Lambda}(t_j) = \sum_{i=1}^{n} n_{i,j} / n = \bar{n}_j \, , \; j = 1, \ldots, m \, . \tag{2.1}$$

A simple generalization of the above estimator leads to the Nelson-Aalen estimator (c.f. Andersen and Borgan (1985)). Suppose that the situation is as above except that the ith individual is observed only at times t_j, $j = 1, \ldots, m_i$, with $m_i \leq m$. Let $Y_{i,j} = 1$ if $j \leq m_i$ and $Y_{i,j} = 0$ otherwise. The Nelson-Aalen estimator of $\Lambda(t_j)$ can be written as

$$\hat{\Lambda}(t_j) = \sum_{l=1}^{j} \{ \sum_{i=1}^{n} Y_{i,l}(n_{i,l} - n_{i,l-1}) / \sum_{i=1}^{n} Y_{i,l} \} \, , \tag{2.2}$$

where $n_{i0} = 0$, $i = 1, \ldots, n$. If $m_i = m$, $i = 1, \ldots, m$, the above estimator reduces to (2.1).

Unfortunately, extension of the Nelson-Aalen estimator to more general data structures is not at all straightforward. In the general case discussed in the next section, we consider an alternative approach.

2.2. A general case

Let $s_1, \ldots, s_m$ denote the ordered distinct observation times in the set $\{ t_{i,j} \, ; \, j = 1, \ldots, m_i, \; i = 1, \ldots, n \}$. Let l_j and $\bar{n}_j$ denote the number and mean values respectively of observations made at s_j and let $\Lambda_j = \Lambda(s_j)$, $j = 1, \ldots, m$. Note that because of the order restriction $\Lambda_1 \leq \cdots \leq \Lambda_m$, we cannot simply estimate $\Lambda(s_j)$ by $\bar{n}_j$. The order restriction does not cause any trouble in the above simple case giving rise to (2.1) since, in this case, $\bar{n}_1 \leq \cdots \leq \bar{n}_m$. More generally, the Nelson-Aalen estimator in (2.2) incorporates the order restriction by estimating the increments $\Lambda(s_j) - \Lambda(s_{j-1})$.

We define estimates $\hat{\Lambda}_1, \ldots, \hat{\Lambda}_m$ as those Λ_j's that minimize the sum

$$\sum_{j=1}^{m} l_j \, [\bar{n}_j \, - \, \Lambda_j]^2 \tag{2.3}$$

subject to the order restriction $\Lambda_1 \leq \cdots \leq \Lambda_m$. This is the isotonic regression based on $\{\bar{n}_1, \ldots, \bar{n}_m\}$ with weights $\{l_1, \ldots, l_m\}$. Obviously if $\bar{n}_1 \leq \cdots \leq \bar{n}_m$, $\hat{\Lambda}_j = \bar{n}_j$, $j = 1, \ldots, m$.

Using the max-min formula for isotonic regression (see Barlow, Bartholomew, Bremner, and Brunk (1972)), we find that

$$\hat{\Lambda}_j = \max_{r \leq j} \, \min_{s \geq j} \Big(\sum_{v=r}^{s} l_v \, \bar{n}_v\Big) \Big/ \Big(\sum_{v=r}^{s} l_v\Big) = \min_{s \geq j} \, \max_{r \leq j} \Big(\sum_{v=r}^{s} l_v \, \bar{n}_v\Big) \Big/ \Big(\sum_{v=r}^{s} l_v\Big) \, . \tag{2.4}$$

The $\hat{\Lambda}_j$'s can also be written as the form

$$\hat{\Lambda}_j \; = \sum_{i \in S_r} l_i \, \bar{n}_i \, / \sum_{i \in S_r} l_i \; , \; j \in S_r \; , \; j = 1, \ldots, m \, . \tag{2.5}$$

The S_r's are called blocks and constitute the increasing and adjacent partition of $\{1, \ldots, m\}$ determined by the pool-adjacent-violators, the up-and-down, or other algorithms (see Barlow, Bartholomew, Bremner, and Brunk (1972) and Robertson, Wright, and Dykstra (1988)). In the following, the estimator given by (2.4) or (2.5) is referred to as the isotonic regression (IR) estimator.

Remark A. For current status data, and under a Poisson assumption for the N_i's, the proposed estimator (2.4) is the maximum likelihood estimator of Λ. This follows since, in this case, the likelihood function is

$$\sum_{j=1}^{m} [\, \bar{n}_j \, \log \Lambda_j \, - \, \Lambda_j \,] \, l_j$$

and maximization of the above likelihood function is equivalent to minimization of (2.3) subject to the order restriction $\Lambda_1 \leq \cdots \leq \Lambda_m$. In more general panel count data, under a Poisson assumption, the IR estimator does not reduce to the maximum likelihood estimate. In general, the m.l.e. is very complicated in this case (see Section 4 for more remarks).

Remark B. The IR estimator of the MF is constructed by taking into account the order among observations from different individuals which is required by the property of the MF Λ. Another way to estimate the MF is to estimate the rate function and then use the integral of the rate estimator as an estimator of the MF. For example, Thall and Lachin (1988) describe a method for estimation of the rate function. In their method, the rate function for the ith individual is assumed constant between observation times and estimated as $\hat{\lambda}_i(t) = (n_{i,j} - n_{i,j-1})/(t_{i,j} - t_{i,j-1})$ for $t \in (t_{i,j-1}, t_{i,j}]$, $i = 1, \ldots, n$. At time t, the overall (average) rate function for all individuals on study is estimated as an average of estimated rate functions over individuals on study at t.

Remark C. An estimate of the variance of $\hat{\Lambda}(s_j) = \hat{\Lambda}_j$ $(j = 1, \ldots, m)$ can be heuristically obtained as follows. Let S_r be the block containing j. Note that $\hat{\Lambda}(s_j)$ is the sample mean of the observations obtained at a neighbourhood of s_j and placed in the same block by the algorithm. If, indeed, these observations were independent with the same mean, a simple estimate of the variance of $\hat{\Lambda}_j$ could be obtained from the sample variance,

$$\sum_{i,l} [n_{i,l} - \hat{\Lambda}(s_j)]^2 / b_r^2 \,, \tag{2.6}$$

where the summation is over $U_r = \{(i,l)\,;\, t_{i,l} = s_u,\, u \in S_r\}$ and $b_r = |U_r|$. It should be noted that the observations contributing to (2.6) are in fact not i.i.d. Some simulations (see Section 3) have indicated that it gives a reasonable estimate of precision, however.

The strong consistency of the estimator (2.4) or (2.5) is discussed in the Appendix. Under certain conditions, the proposed estimator converges almost surely to Λ as the number of individuals observed approaches infinity. The main conditions are as follows: first, at each given observation time, the number of the observations taken in any neighbourhood of it goes to infinity as n does. The necessity of this can be seen as follows. Suppose that there is an observation time, t_0 say, which is isolated from all other observation times. Then it is clear that $\hat{\Lambda}$ will not converge to Λ at t_0 (assuming that Λ is strictly increasing in a neighbourhood of t_0). The second condition required for the consistency of the proposed estimator is that no process is observed with a high frequency around a single time point. In other words, the number of observations from one process around a single time point is finite when n goes to infinity. In the case of current status data, uniform consistency of the proposed estimator can be obtained and is also discussed in the Appendix.

In essence, the method proposed here treats each of the $\sum_{i=1}^{n} m_i$ observations as independent observations on current status. This will be a satisfactory approach provided observations on each of the processes are fairly well spaced.

Clearly, for example, if one process is observed very often in the neighbourhood of a time, t_0 say, this approach would give more emphases to this process in estimating $\Lambda(t_0)$. The second condition above is a reflection of this potential difficulty and suggests that the approach described here is best used when the spacings of observations are similar in the observed samples.

3. Examples and Simulation

The estimation of the MF of point processes proposed in the previous section is illustrated in two examples. The first example is also discussed in Gaver and O'Muircheartaigh (1987) and concerns the reliability of nuclear plants. The second example (Thall and Lachin (1988)) concerns a study of the incidence of nausea of patients with floating gallstones. The results of a small simulation study of the estimator (2.4) and its variance estimator (2.6) are also presented.

3.1. Example 1

Table 1 presents the observation times (one per plant) and the corresponding observed numbers of losses of feedwater flow for 30 nuclear plants. The IR estimator of the average number of losses of feedwater flow for a nuclear plant based on these data is displayed in Figure 1. For comparison, the total number of losses of feedwater flow divided by the number of systems at each distinct observation time (the $\bar{n}_j$'s) is also shown. The IR estimator is obtained by pooling the $\bar{n}_j$'s according to the order restriction. These results are similar to those given in Gaver and O'Muircheartaigh (1987), who model the numbers of losses of feedwater flow as homogeneous Poisson processes with i.i.d. intensities.

Table 1. Panel count data about losses of feedwater flow

Observation times t_i (in years) and observed numbers n_i											
System	t_i	n_i	System	t_i	n_i	System	t_i	n_i	System	t_i	n_i
1	15	4	9	4	13	17	2	11	25	1	1
2	12	40	10	3	4	18	2	1	26	3	10
3	8	0	11	4	27	19	2	0	27	2	5
4	8	10	12	4	14	20	1	3	28	4	16
5	6	14	13	4	10	21	1	5	29	3	14
6	5	31	14	2	7	22	1	6	30	11	58
7	5	2	15	3	4	23	5	35			
8	4	4	16	3	3	24	3	12			

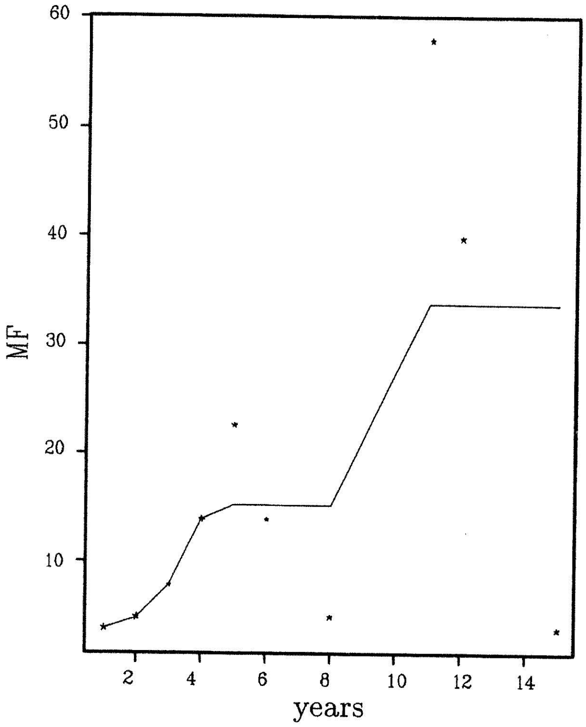

Figure 1. IR estimator of losses of feedwater flow

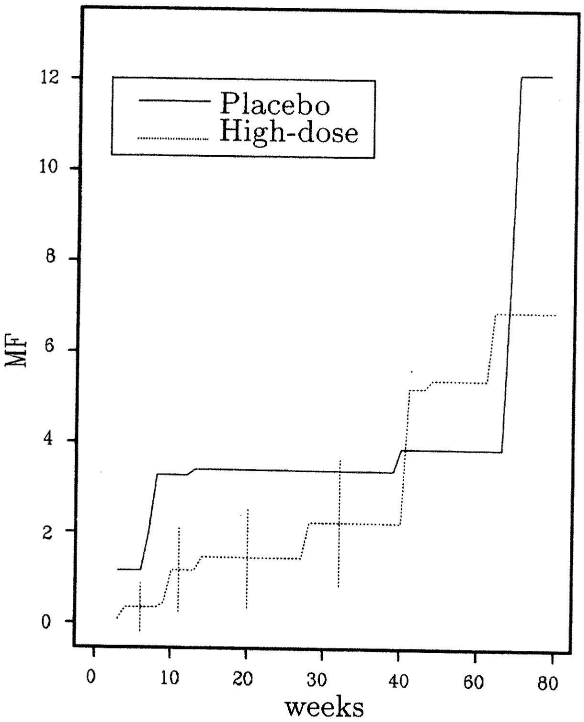

Figure 2. IR estimators of the MF for nausea

3.2. Example 2

Consider the data presented in Table 1 of Thall and Lachin (1988), which give the successive visit-times in study weeks and the associated counts of episodes of nausea for 113 patients with floating gallstones. These data comprise the first year follow up of two groups, placebo (48) and high-dose chenodiol (65), from the National Cooperative Gallstone Study. The whole study consists of 916 patients who were randomized to the placebo, low dose, or high dose group and treated for up to two years.

During the study patients were scheduled to return for clinic visits at 1, 2, 3, 6, 9, and 12 months, and, on each occasion, asked to report the total numbers of each type of symptom that had occurred since the last visit. It can be seen from their table, however, that actual visitation times differ from patient to patient. For example, the first visit times ranged from 3 to 9 weeks and some patients did not complete the full schedule.

Figure 2 shows the estimates of the MF of occurrence of nausea for the patients in the placebo and high-dose groups. The estimated MF for the placebo group is higher than that for the high-dose group over the first 40 weeks. Most of this difference is due to an early difference in the event rate over the first 10 weeks. After 40 weeks, the high-dose group has a slightly higher estimated MF than the placebo group for a period of time. At the end of the year, however, the estimated MF of the placebo group jumps sharply over that of the high-dose group. These seem to differ in a substantial way from the trend exhibited in Figure 2 of Thall and Lachin (1988). That figure shows estimated event rates of the two groups based on the first year of the study and suggests that the event rate of the high-dose group is never lower than that of the placebo group after week 19.

Also shown in Figure 2 are 95% confidence intervals at several time points for the MF of occurrence of nausea for the high-dose group based on the variance formula (2.6). Pointwise comparisons indicate significant differences between the groups at early times (before 30 weeks), but not for later times.

3.3. A simulation study

To investigate the adequacy of the estimator (2.4) and the variance estimator (2.6), a simulation study was performed. For each of $n = 50$ individuals, four observations from a Poisson process with the MF $\Lambda(t) = 10\,t$ were generated. The observation times were generated from the uniform distribution U[0,1]. The process was replicated 100 times with the 200 observation times held fixed. Figure 3 shows the average of the MF estimates (dot line in the middle) and the average of the 95% confidence bands (dot lines) based on the estimators (2.4) and (2.6) respectively. The pointwise sample variances of the MF estimates were also calculated and the corresponding 95% intervals (broken lines) are also displayed in Figure 3. For comparison, the pointwise mean estimate (solid line) of the MF using the parametric assumption $\Lambda(t) = \lambda\,t$ is also shown in Figure 3, where λ is an unknown parameter.

It can be seen from Figure 3 that the two MF estimators, the IR and parametric mean estimators, are very close except at the last several observation times. The two variance estimates of the MF are almost identical, which suggests that the variance estimate given by (2.6) is quite acceptable. Similar results are also shown by other simulation data, in which one, two or eight observations are observed for each individual.

4. Discussion

A simple and consistent estimator of the MF of a recurrent event based on panel count data is presented. The proposed estimator does not assume any explicit model for the underlying process. In the case when an explicit model for the underlying process is assumed, for example, under a Poisson assumption, an alternative to the IR estimator of a MF is the maximum likelihood estimate which can be obtained fairly generally using the EM algorithm, which although being a straightforward method gives rise to rather complicated equations.

As usual, the actual observation times are assumed to be non-informative about the MF in the above discussion. That is, the time at which a subject visits the clinical center is independent of his disease status in the case of clinical trials. It should be noted that this is sometimes not the case. For example, an early visit by a subject may be caused by a worsening of the disease (e.g., Gruger, Kay, and Schumacher (1991)). In this case, methods which take this into account need to be used. The approximate variance formula given in (2.6) will be most satisfactory in large data sets where the b_r's are not small. One case in which this formula is not valid is that the b_r's are equal to one and in this situation, alternative variance estimates are needed.

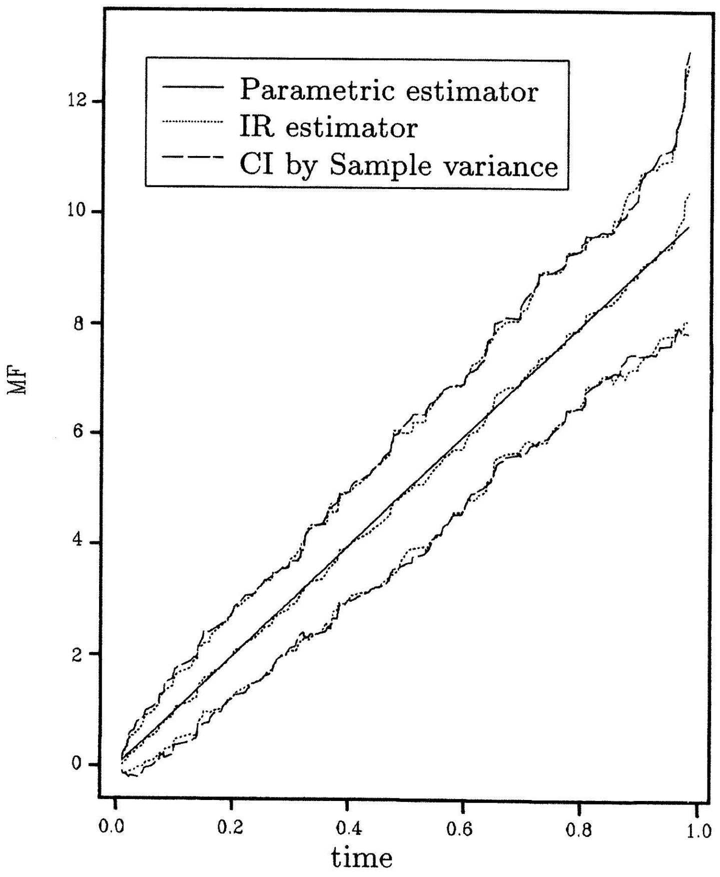

Figure 3. Estimators of the MF from simulation data

It would be useful to derive the asymptotic distribution of the IR estimator. The asymptotic distribution of this type of estimator at a point has been discussed by Brunk (1970) and other authors in the context of isotonic regression. Brunk assumed that for each fixed n, one observation is made at each of the observation points $\{1/n, 2/n, \ldots, n/n\}$ and all observations are independent. Under these and some other assumptions, he derived the asymptotic distribution with norming constants of order $n^{1/3}$ and the asymptotic p.d.f. expressed in term of partial derivatives of a particular solution to the heat equation. Groeneboom and Wellner (1992) also studied asymptotic distributions for estimators of the isotonic-type in the context of interval-censored data.

Acknowledgement

The authors are grateful to Professor Jerry Lawless, the Chair Editor and a referee for helpful comments.

Appendix

Consistency of the estimator (2.2). Suppose that $\Lambda(t)$ is a continuous function. Let $T_n = \{t_{i,j},\ j = 1, \ldots, m_i,\ i = 1, \ldots, n\}$ and $k_n = |T_n|$, the number of observations for the given n. Let $\hat{\Lambda}_n(t_l)$ be defined by (2.4), $l = 1, \ldots, k_n$. The following conditions are sufficient for the strong consistency property of $\hat{\Lambda}_n$.

Condition I. For each $t_l \in \bigcup_{n=1}^{\infty} T_n$ and each interval J containing t_l, $\limsup_{n\to\infty} k_n / I_n(J) < \infty$, where $I_n(J)$ denotes the number of times in T_n which lie in the interval J.

Condition II. For each $t_l \in \bigcup_{n=1}^{\infty} T_n$, there exists an interval J_0 containing t_l such that $\limsup_{n\to\infty} U_{in}(J_0) < \infty$ for all i, where $U_{in}(J_0)$ denotes the number of observations with observation times in $T_n \bigcap J_0$ from the ith process.

Theorem 1. *If Conditions* I *and* II *hold and* $\mathrm{Var}[N(t)] < \infty$, *then for each* $t_l \in \bigcup_{n=1}^{\infty} T_n$, *as* $n \to \infty$,

$$\Pr\{\, \hat{\Lambda}_n(t_l) - \Lambda(t_l) \to 0\} = 1\ .$$

The proof of Theorem 1 uses Theorem 6.1 of Brunk (1958) and follows the same line of the proof as Theorem 4.1 of Brunk (1970).

For current status data, uniform consistency of $\hat{\Lambda}_n$ can be obtained. Suppose that all observation time points belong to the interval $[0,1]$ and $\hat{\Lambda}_n(t)$ is defined as above at the observation times $t_{i,j}$'s. Suppose that $\hat{\Lambda}_n(t)$ is extended and defined on the interval $[0,1]$ subject to the monotonic restriction. Let $I_n(J)$ be defined as in Condition I.

Theorem 2. If for each subinterval J of $(0,1)$, $\limsup_{n\to\infty} n / I_n(J) < \infty$ and for $t \in (0,1)$, $\mathrm{Var}[N(t)] < \infty$, then

$$\Pr\{\lim_{n\to\infty} \sup_{a\le t\le b} |\hat{\Lambda}_n(t) - \Lambda(t)| = 0\} = 1\ ,$$

where $0 < a < b < 1$.

Theorem 2 follows directly from Theorem 4.1 of Brunk (1970).

References

Andersen, P. K. and Borgan, O. (1985). Counting process models for life history data: A review (with discussion). Scand. J. Statist. **12**, 97-158.

Barlow, R. E., Bartholomew, D. J., Bremner, J. M. and Brunk, H. D. (1972). Statistical Inference Under Order Restrictions. John Wiley, New York.

Brunk, H. D. (1958). On the estimation of parameters restricted by inequalities. Ann. Math. Statist. **29**, 437-454.

Brunk, H. D. (1970). Estimation of isotonic regression (with discussion by Ronald Pyke). In Nonparametric Techniques in Statistical Inference, 177-197. Cambridge University Press.

Diamond, I. D., McDonald, J. W. and Shah, I. H. (1986). Proportional hazards models for current status data: Application to the study of differentials in age at weaning in Pakistan. Demography **23**, 607-620.

Gaver, D. P. and O'Muircheartaigh, I. G. (1987). Robust empirical Bayes analysis of event rates. Technometrics **29**, 1-15.

Groeneboom, P. and Wellner, J. (1992). Information Bounds and Nonparametric Maximum Likelihood Estimation. Birkhauser, Basel.

Grüger, J., Kay, R. and Schumacher, M. (1991). The validity of inferences based on incomplete observations in disease state models. Biometrics **47**, 595-605.

Kalbfleisch, J. D. and Lawless, J. F. (1981). Statistical inference for observational plans arising in the study of life history processes. In Symposium on Statistical Inference and Applications In honour of George Barnard's 65th Birthday. University of Waterloo, August 5-8, 1981.

Kalbfleisch, J. D. and Lawless, J. F. (1985). The analysis of panel data under a Markov assumption. J. Amer. Statist. Assoc. **80**, 863-871.

Robertson, T., Wright, F. T. and Dykstra, R. L. (1988). Order restricted statistical inference. John Wiley, New York.

Sun, J. and Kalbfleisch, J. D. (1993). The analysis of current status data on point processes. J. Amer. Statist. Assoc. **88**, 1449-1454.

Thall, P. F. and Lachin, J. M. (1988). Analysis of recurrent events: Nonparametric methods for random-interval count data. J. Amer. Statist. Assoc. **83**, 339-347.

Department of Statistics and Actuarial Science, University of Waterloo, Waterloo, Ontario N2I 3G1, Canada.

(Received March 1993; accepted May 1994)