Estimation of the time between repeated events under different sampling patterns

Keiding, Niels (presenting author)
University of Copenhagen, Department of Biostatistics
Øster Farimagsgade 5
P.O.Box 2099
DK-1014 Copenhagen K, Denmark
E-mail: nike@sund.ku.dk

Gill, Richard D.
University of Leiden, Mathematical Institute
Postbus 9512
NL-2300 RA Leiden
Netherlands
E-mail: gill@math.leidenuniv.nl

Two classical problems in nonparametric statistical analysis of recurrent event data are considered, both formalised within the framework of a simple, stationary renewal process.

We first consider observation around a fixed time point, i.e. we observe a backward recurrence time R and a forward recurrence time S. It is well known that the nonparametric maximum likelihood estimator is the Cox-Vardi estimator (Cox 1969, Vardi 1985) derived from the length-biased distribution of the gap time $R+S$. However, Winter & Földes (1988) proposed to use a product-limit estimator based on S, with delayed entry given by R. We clarify the relation of that estimator to the standard left truncation problem, see Keiding & Gill (1990).

The second observation scheme considers a stationary renewal process observed in a finite interval where the left endpoint does not necessarily correspond to an event. (For an application to automobile insurance, see Keiding et al. (1998)). The full likelihood function is complicated, and we briefly survey possibilities for restricting attention to various partial likelihoods, in the nonparametric case again allowing the use of simple product-limit estimators.

A more detailed version of this work, also containing references to earlier work, was recently published (Gill & Keiding, 2010).

Observation of a stationary renewal process around a fixed point

Winter & Földes (1988) studied the following estimation problem. Consider n independent renewal processes in equilibrium with underlying distribution function H, which we shall assume absolutely continuous with density h and support $(0,\infty)$ and hence hazard $\beta(t) = h(t)/(1-H(t))$. Corresponding to a fixed time, say 0, the forward and backward recurrence times S_i and R_i are observed; then $Q_i = R_i + S_i$ is a length-biased observation corresponding to the distribution function H. We quote the following distribution results: let χ be the expectation of H, $\chi = \int_0^\infty [1-H(u)]du$, then the joint distribution of (R,S) has density $\chi^{-1}h(r+s)$, the marginal distributions of R and S are equal with *density* $\chi^{-1}[1-H(r)]$, and the marginal distribution of $Q = R+S$ has density $\chi^{-1}qh(q)$, the length-biased density corresponding to h.

Winter and Földes considered the product-limit estimator

$$1-\tilde{H}(t) = \prod_{i:Q_i \le t}\left(1-\frac{1}{Y(Q_i)}\right)$$

where $Y(t) = \sum_{i=1}^{n} I\{R_i < t \le R_i + S_i\}$ is the *number at risk* at time t. This estimator is the same as the Kaplan-Meier estimator for iid survival data $Q_1,\ldots,Q_n$ left-truncated at $R_1,\ldots,R_n$ (Kaplan & Meier 1958, Andersen et al. 1993). Winter & Földes showed that $1-\tilde{H}$ is strongly consistent for the *underlying* survival function $1-H$.

We shall show how the derivation of this estimator follows from a simple Markov process model similar to the one used by Keiding & Gill (1990) to study the random truncation model.

First notice that the conditional distribution of $Q = R + S$ given that $R = r$ has density

$$\frac{\chi^{-1} h(q)}{\chi^{-1}\left[1 - H(r)\right]}, \; r \le q \le \infty$$

that is, intensity (hazard) $h(q)/\left[1 - H(q)\right]$, which is just the hazard $\beta(q)$ corresponding to the underlying distribution H. Now define for each i (the i is suppressed in the notation) a stochastic process U on $[0, \infty]$ with state space $\{0,1,2\}$ by

$$U(t) = \begin{cases} 0, & 0 \le t < R \\ 1, & R \le t < R + S \\ 2, & R + S \le t \; . \end{cases}$$

We have $P\left\{U(t+h) = 2 \middle| U(u), 0 \le u \le t\right\} = o(h)$ for $U(t) = 0$, and for $U(t) = 1$ (that is, $R \le t < R + S$) this is

$$P\left\{R + S \le t + h \middle| R, R + S > t\right\} = \frac{h(t)}{1 - H(t)} h + o(h)$$

by the above result on the hazard of $R + S \middle| R$. That this depends on t but not on R proves that U is a Markov process

$$\boxed{0} \xrightarrow{\alpha} \boxed{1} \xrightarrow{\beta} \boxed{2}$$

with intensities $\alpha(t) = \left[1 - H(t)\right] \Big/ \int_t^\infty \left[1 - H(r)\right] dr$ (the marginal hazard of R, equal to the residual mean lifetime function of the underlying distribution H) and $\beta(t) = h(t)/\left[1 - H(t)\right]$.

The Markov process framework of Keiding & Gill (1990) now indicates that the product limit estimator $1 - \tilde{H}$ is a natural estimator of the survivor function $1 - H$ of the underlying distribution, and consistency and asymptotic normality may be obtained as shown by Keiding & Gill (1990, Sec. 5).

Winter and Földes stated that (R,S) contain no more information than $R+S$ about H. This is easily seen from the likelihood function based on observation of $(R_1,S_1),\ldots,(R_n,S_n)$, which is $\chi^{-n}\prod_{i=1}^{n} h(r_i+s_i)$ from which the NPMLE of H is readily derived as

$$\hat{H}(t)=\sum_{i=1}^{n}\frac{I\{R_i+S_i\le t\}}{R_i+S_i}\Big/\sum_{i=1}^{n}\frac{1}{R_i+S_i},$$

that is the Cox-Vardi estimator in the terminology of Winter and Földes (Cox 1969, Vardi 1985).

It follows that the estimator $1-\tilde{H}$ is *not* NPMLE. The difference between the situation here and that of the random truncation model studied by Keiding & Gill (1990, Sec. 3) is that not only the intensity $\beta(t)$, but also $\alpha(t)$ depends only on the estimand H.

Observation of a stationary renewal process in a finite interval

We consider again a stationary renewal process on the whole line and assume that we observe it in some interval $[t_1,t_2]$ determined independently of the process. Cook & Lawless (2007, Chapter 4) surveyed the general area of analysis of gap times emphasizing that the assumption of independent gap times implied by simple renewal processes is often unrealistic.

We shall here nevertheless work under the assumption of the simplest possible model as indicated above. Because the nonparametric maximum likelihood estimator is computationally involved it may sometimes be useful to calculate less efficient alternatives, and there are indeed such possibilities.

Under the observation scheme indicated above we may have the following four types of elementary observations

1. Times x_i from one renewal to the next, contributing the density $h(x_i)$ to the likelihood.
2. Times from one renewal T to t_2 (right-censored versions of 1.), contributing factors of the form $(1-H(t_2-T))$ to the likelihood.
3. Times from t_1 to the first renewal T (forward recurrence times), contributing factors of the form $(1-H(T-t_1))/\chi$ to the likelihood.
4. Knowledge that no renewal happened in $[t_1,t_2]$, actually a right-censored version of 3., contributing factors of the form $\int_{t_2-t_1}^{\infty}(1-H(u))\,du/\chi$ to the likelihood.

Our interest is in basing the estimation only on complete or right-censored gap times, i.e. observations of type 1 or 2. When this is possible, we have simple product-limit estimators in the one-sample situation, and we may use well-established regression models (such as Cox regression) to account for covariates. Peña et al. (2001) assumed that observation started at a renewal (thereby

defining away observations of type 3 and 4) and gave a comprehensive discussion of exact and asymptotic properties of product-limit estimators with comparisons to alternatives. The crucial point here is that calendar time and time since last renewal both need to be taken into account, so the straightforward martingale approach displayed by Andersen et al. (1993) is not available.

As noted by Aalen & Husebye (1991) in their attractive non-technical discussion of observation patterns, observation does however often start between renewals. (In the example of Keiding et al. (1998), auto insurance claims were considered in a fixed calendar period). As long as observation starts at a stopping time, inference is still valid, so by starting observation at the first renewal in the interval we can essentially refer back to Peña et al. (2001). A more formal argument could be based on the concept of the *Aalen filter*, see Andersen et al. (1993, p.164). The resulting product-limit estimators will not be fully efficient, since the information in the backward recurrence time (types 3 and 4) is ignored. It is important to realize that the validity of this way of reducing the data depends critically on the independence assumptions of the model. Keiding et al. (1998), cf. Keiding (2002) for details, used this fact to base a goodness-of-fit test on a comparison of the full nonparametric maximum likelihood estimator with the product-limit estimator.

We finally mention that under the assumption of stationarity, we might as well study the phenomena in reverse time. Then type 1 is still type 1, but type 2 becomes type 3 and vice versa, while type 4 remain type 4. One may therefore use the product of the partial likelihoods based on types 1 and 2 observations in forward time and type 1 and 2 observations in reverse time, or in other words each full observation counted twice, observations of type 2 and 3 each counted once as right-censored observation in wither forward time or reverse time, and type 4 observations still omitted.

Acknowledgements
This research was partially supported by a grant (RO1CA54706-12) from the National Cancer Institute and by the Danish Natural Sciences Council grant 272-06-0442 “Point process modelling and statistical inference”.

REFERENCES

Aalen, O.O. & Husebye, E. (1991). Statistical analysis of repeated events forming renewal processes. *Statistics in Medicine* **10**, 1227-1240.

Andersen, P.K., Borgan, Ø., Gill, R.D. & Keiding, N. (1993). *Statistical Models Based on Counting Processes*. Springer Verlag, New York.

Cook, R.J. & Lawless, J.F. (2007). *The statistical analysis of recurrent events*. Springer Verlag, New York.

Cox, D.R. (1969). Some sampling problems in technology. In *New Developments in Survey Sampling* (N.L. Johnson and H. Smith, Jr., eds), 506-527. Wiley, New York.

Gill, R.D. & Keiding, N. (2010). Product-limit estimators of the gap time distribution of a renewal process under different sampling patterns. *Lifetime Data Analysis* **16**, 571-579.

Kaplan, E.L. & Meier, P. (1958). Non-parametric estimation from incomplete observations. *Journal of the American Statistical Association* **53**, 457-481.

Keiding, N. (2002). Two nonstandard examples of the classical stratification approach to graphically assessing proportionality of hazards. In: *Goodness-of-fit Tests and Model Validity* (eds. C. Huber-Carol, N. Balakrishnan, M.S. Nikulin and M. Mesbah). Boston, Birkhäuser, 301-308.

Keiding, N., Andersen, C. & Fledelius, P. (1998). The Cox regression model for claims data in non-life insurance. *ASTIN Bulletin* **28**, 95-118.

Keiding, N. & Gill, R.D. (1990). Random truncation models and Markov processes. *The Annals of Statistics* **18**, 582-602.

Peña, E. A., Strawderman, R. L. & Hollander, M. (2001). Nonparametric estimation with recurrent event data. *Journal of the American Statistical Association* **96,** 1299-1315.

Vardi, Y. (1985). Empirical distributions in selection bias models. *The Annals of Statistics* **13**, 178-203.

Winter, B.B. & Földes, A. (1988). A product-limit estimator for use with length-biased data. *The Canadian Journal of Statistics* **16**, 337-355.