Nelson–Aalen Estimator

The Nelson–Aalen estimator is a nonparametric estimator which may be used to estimate the cumulative hazard rate function from censored survival data (*see* **Survival Distributions and Their Characteristics**). Since no distributional assumptions are needed, one important use of the estimator is to check graphically the fit of parametric models, and this is the reason why it was originally introduced by Nelson [10, 11]. Independently of Nelson, Altshuler [2] derived the same estimator in the context of **competing risks** animal experiments. Later, by adopting a counting process formulation, Aalen [1] extended its use beyond the survival data and competing risks setups, and studied its small and large sample properties using martingale methods. The estimator is nowadays denoted the Nelson–Aalen estimator, although other names (the Nelson estimator, the Altshuler estimator, the Aalen–Nelson estimator, the empirical cumulative hazard estimator) are sometimes used as well. Below we present a number of situations where the Nelson–Aalen estimator may be applied and exemplify its use in one particular case. Furthermore, we indicate how counting processes provide a framework which allows for a unified treatment of all these diverse situations, and we summarize the most important properties of the Nelson–Aalen estimator. A detailed account is given in [3, Section IV.1].

Survival Data

Consider first the survival data situation, where we want to study the time to death (or some other event) for a homogeneous population with hazard rate function $\alpha(t)$ and cumulative hazard rate function $A(t) = \int_0^t \alpha(s)\,\mathrm{d}s$. Assume that we have a sample of n individuals from this population. Our observation of the survival times for these individuals will typically be subject to right censoring, meaning that for some individuals we only know that their true survival times exceed certain censoring times. The censoring is assumed to be independent in the sense that the additional knowledge of censorings before any time t does not alter the risk of failure at t (*see* **Censored Data**). We denote by $t_1 < t_2 < \cdots$ the times when deaths are observed and let d_j be the number of individuals who die at t_j.

The Nelson–Aalen estimator for the cumulative hazard rate function then takes the form

$$\widehat{A}(t) = \sum_{t_j \le t} \frac{d_j}{r_j}, \tag{1}$$

where r_j is the number of individuals at risk (i.e. alive and not censored) just prior to time t_j. Thus the Nelson–Aalen estimator is an increasing right-continuous step function with increments d_j/r_j at the observed failure times. The variance of the Nelson–Aalen estimator may be estimated by

$$\widehat{\sigma}^2(t) = \sum_{t_j \le t} \frac{(r_j - d_j)d_j}{(r_j - 1)r_j^2}. \tag{2}$$

It may be shown (see below) that the Nelson–Aalen estimator (1) as well as the variance estimator (2) are almost unbiased. In large samples the Nelson–Aalen estimator, evaluated at a given time t, is approximately normally distributed, so a standard $100(1-\alpha)\%$ confidence interval for $A(t)$ takes the form

$$\widehat{A}(t) \pm z_{1-\alpha/2}\widehat{\sigma}(t), \tag{3}$$

with $z_{1-\alpha/2}$ the $1-\alpha/2$ fractile of the standard normal distribution. The approximation to the normal distribution is improved by using a log transform giving the confidence interval

$$\widehat{A}(t) \exp\left[\pm z_{1-\alpha/2}\frac{\widehat{\sigma}(t)}{\widehat{A}(t)}\right]. \tag{4}$$

This interval is satisfactory for quite small sample sizes [5].

Right censoring is not the only kind of data incompleteness in survival analysis. Often, e.g. in epidemiological applications, individuals are not followed from time zero (in the relevant time scale, typically age), but only from a later entry time (conditional on survival until this entry time). Thus, in addition to right censoring, the survival data are subject to left truncation. For such data we may still use the Nelson–Aalen estimator (1) and estimate its variance by (2). The number at risk, r_j, now is the number of individuals who have entered the study before time t_j and are still in the study just prior to t_j. For left-truncated data the numbers at risk, r_j, may be low for small values of t_j. This will result in estimates $\widehat{A}(t)$ which have large sampling errors. But because the increments of the Nelson–Aalen estimator are

uncorrelated (see below), the uncertainty induced for small time values has no influence on the increment $\widehat{A}(t) - \widehat{A}(s)$ of the Nelson–Aalen estimator over a later time interval $(s, t]$. An estimator for the variance of this increment is $\widehat{\sigma}^2(t) - \widehat{\sigma}^2(s)$.

Quite often we want to estimate the survival distribution function $S(t) = \exp[-A(t)]$, representing the probability that an individual will be alive at time t. This may be done from right-censored and/or left-truncated survival data by the **Kaplan–Meier estimator**. The relation $A(t) = -\ln S(t)$ suggests that the cumulative hazard rate function alternatively may be estimated as minus the logarithm of the Kaplan–Meier estimator. Even though this estimator numerically will be close to the Nelson–Aalen estimator, the latter is the canonical one from a theoretical point of view. Furthermore, the Nelson–Aalen estimator may be used in a number of different situations (see below) while the alternative estimator applies only to the survival data situation.

An Illustration

To give an illustration of the Nelson–Aalen estimator we use data from a randomized clinical trial for patients with histologically verified liver cirrhosis. Patients were recruited from several hospitals in Copenhagen between 1962 and 1969 and were followed until death, lost to follow-up or until the closing date of the study, October 1, 1974. The time variable of interest is time since entry into the study. Patients are right censored if alive on October 1, 1974, or if lost to follow-up before that date.

We consider only the 138 placebo-treated male patients. Their median age at entry was 57 years, while the lower and upper quartiles were 51 and 66 years, respectively. Of the 138 patients, 88 died during the study. The Nelson–Aalen estimate for these patients is shown in Figure 1 with 95% confidence intervals computed according to (4). Even though the cumulative hazard rate function provides a useful summary measure (e.g. [6, Section 2.3]), it is usually the hazard rate function itself which is the entity of real interest. So when interpreting the estimate in Figure 1, we mainly focus on the "slope" of the curve. The estimate of the cumulative hazard rate function is steeper for the first 9–10 months after randomization than at later times. Therefore we have evidence that the risk of dying for these patients is highest just after randomization. (This may, at least in part, be due to heterogeneity which is not accounted for in our simple analysis.) The hazard rate function is approximately 0.3 per year for the first 9–10 months and slightly below 0.2 per year thereafter when estimated as the average slope of the curve over the relevant time periods. More formal procedures for smoothing the Nelson–Aalen estimate in order to obtain an estimate for the hazard rate function itself are available but will not be considered here (*see* **Smoothing Hazard Rates**). A further discussion and analysis of the cirrhosis data is given in [12]. The data were also used for illustrative purposes in [3].

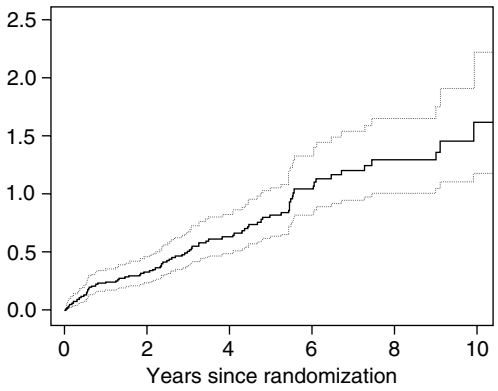

Figure 1 Nelson–Aalen estimate of the cumulative hazard rate function for death for 138 placebo-treated male patients with liver cirrhosis, with 95% log-transformed confidence intervals

Multi-state Models and Recurrent Events

The survival analysis setup considered above may be generalized in two directions. More than one type of event may be considered for each individual under study, and/or the event in question may happen more than once for each individual. Examples of the first type are competing risks with two or more causes of death and the Markov illness–death model with the states "healthy", "diseased", and "dead" (*see* **Counting Process Methods in Survival Analysis**). More generally, we may consider any **Markov process** with a finite number of states which may be used to model the life history of an individual. An example of the second type is an inhomogeneous **Poisson process** with intensity $\alpha(t)$ modeling the occurrence of some recurrent event like episodes of

angina pectoris in patients with coronary heart disease or infections in AIDS patients. For both of these two types of situations we observe the times when events occur for a number of individuals (modeled as iid copies of the relevant process) who need not all be observed over the same interval of time. The Nelson–Aalen estimator may then be applied to estimate cumulative intensities.

To be specific, consider a finite-state Markov process with transition intensities $\alpha_{gh}(t)$ for $g \neq h$. Focusing on fixed g and h in what follows, we drop the subscripts and write just $\alpha(t)$ for the $g \to h$ transition intensity. Furthermore, we denote by $t_1 < t_2 < \cdots$ the times when transitions from g to h are observed. Let d_j be the number of individuals who experience a $g \to h$ transition at t_j, and write r_j for the number of individuals in state g (i.e. at risk for a $g \to h$ transition) just prior to time t_j. Then the cumulative $g \to h$ transition intensity $A(t) = \int_0^t \alpha(s)\,\mathrm{d}s$ may be estimated by (1) and its variance by (2). Similarly, the integrated intensity of an inhomogeneous Poisson process may be estimated with the t_js denoting the times of observed events, and the d_js and r_js being the corresponding numbers of events and numbers at risk, respectively. An illustration of the use of the Nelson–Aalen estimator to estimate integrated Markov transition intensities is given by Keiding & Andersen [9].

Two Other Applications

For the situations considered so far, (1) and (2) apply with r_j the number at risk at t_j for the event in question. The use of the Nelson–Aalen estimator is, however, not restricted to such situations. We mention here two other applications and return to a general discussion below.

Relative Mortality

Our first example considers right-censored and/or left-truncated survival data, but they no longer come from a homogeneous population. Rather, we assume that the hazard rate function of the ith individual may be written as the product $\alpha(t)\mu_i(t)$, where $\alpha(t)$ is a relative mortality common to all individuals and $\mu_i(t)$ is the hazard rate function at time t for a person from an external standard population corresponding to the ith individual (e.g. of the same sex and age as individual i). Typically the $\mu_i(t)$ will be known from published life tables for the general population. In this situation the Nelson–Aalen estimator may be used to estimate the cumulative relative mortality $A(t) = \int_0^t \alpha(s)\,\mathrm{d}s$. All that is required is that r_j in (1) be taken to denote the sum of the external rates $\mu_i(t_j)$ for all individuals at risk just prior to t_j. An illustration of this use of the Nelson-Aalen estimator is provided by Breslow & Day [7, Chapter 5].

An Epidemic Model

A simple model for the spread of an infectious disease in a community is the following (*see* **Epidemic Models, Stochastic**). At the start of the epidemic, i.e. at time $t = 0$, some individuals make contact with individuals from elsewhere and are thereby infected with the disease. There are no further infections from outside the community during the course of the epidemic. Let $S(t)$ and $I(t)$ denote the number of susceptibles and infectives, respectively, just prior to time t. Assuming random mixing, the infection intensity in the community at time t becomes $\alpha(t)S(t)I(t)$, where $\alpha(t)$ is the infection rate per possible contact. We denote by $0 < t_1 < t_2 < \cdots$ the times when individuals are infected and let d_j denote the number infected at t_j. Then the cumulative infection rate, $A(t) = \int_0^t \alpha(s)\,\mathrm{d}s$, may be estimated by the Nelson–Aalen estimator (1) where now $r_j = S(t_j)I(t_j)$; see Becker [4, Section 7.6] for an illustration.

Counting Process Formulation and Small Sample Properties

In general we consider the occurrences of some events of interest (e.g. deaths, occurrences of a disease, infections), and denote by $0 < t_1 < t_2 < \cdots$ the times when an event is observed. We assume that two or more events cannot occur at the same time, so that there are no tied observations. (The handling of ties is discussed briefly below.) Then the process $N(t)$ counting the number of observed events in the time interval $[0, t]$ is a (univariate) counting process. The behavior of $N(t)$ is governed by its intensity process $\lambda(t)$ given heuristically by

$$\lambda(t)\,\mathrm{d}t = \Pr(\text{event occurs in } [t, t + \mathrm{d}t) | \mathcal{F}_{t-}).$$

Here $\mathcal{F}_{t-}$ represents all the information available to the researcher just before time t. The counting process satisfies Aalen's multiplicative intensity model if we may write its intensity process as

$$\lambda(t) = \alpha(t)Y(t), \tag{5}$$

for some unknown function $\alpha(t)$ and some observable process $Y(t)$ whose value at time t is known just prior to t. All the situations considered above give counting processes which fulfill (5). Survival data from a homogeneous population, finite-state Markov processes, and the inhomogeneous Poisson process, all give a $Y(t)$ process which is the number at risk just prior to time t. For the model for relative mortality, $Y(t)$ is the sum of the $\mu_i(t)$ for those at risk just before t, while for the epidemic model, $Y(t) = S(t)I(t)$. The common structure of all these models when formulated as counting processes is the reason why the Nelson–Aalen estimator may be applied to all these diverse problems.

In fact, the counting process formulation provides a framework which makes it simple to study the statistical properties of the Nelson–Aalen estimator. We briefly indicate a few main steps and refer to [3, Section IV.1.1] for a thorough treatment. First, we note that, with $r_j = Y(t_j)$, we may write the Nelson–Aalen estimator (1) as

$$\widehat{A}(t) = \int_0^t \frac{J(s)}{Y(s)}\,\mathrm{d}N(s), \tag{6}$$

where $J(s) = I(Y(s) > 0)$ and 0/0 is interpreted as 0. Then using (5), (6), and the decomposition $N(t) = \int_0^t \lambda(s)\,\mathrm{d}s + M(t)$ of a counting process into a sum of its integrated intensity process and a local square integrable martingale $M(t)$, we obtain

$$\widehat{A}(t) = A^*(t) + M^*(t). \tag{7}$$

Here $A^*(t) = \int_0^t J(s)\alpha(s)\,\mathrm{d}s$ is almost the same as $A(t)$ when there is only a small probability that $Y(s) = 0$ for some $s \le t$, while $M^*(t) = \int_0^t [J(s)/Y(s)]\,\mathrm{d}M(s)$ is a stochastic integral and as such is a local square integrable martingale. Relation (7) is the key to studying the statistical properties of the Nelson–Aalen estimator. Since $M^*(t)$ has expected value zero for any given t, we have $\mathrm{E}\widehat{A}(t) = \mathrm{E}A^*(t)$, so the Nelson–Aalen estimator is almost unbiased. Furthermore, an unbiased estimator for the variance of $M^*(t)$ is its optional variation process $\int_0^t [J(s)/Y(s)^2]\,\mathrm{d}N(s)$. Thus the variance estimator (2) is almost unbiased when there are no ties. Finally, a martingale has uncorrelated increments, and by (7) this is (almost) the case for the Nelson–Aalen estimator as well.

In the presence of ties, i.e. when the number of events d_j at t_j exceeds one, the process $N(t)$ counting occurrences of events in $[0, t]$ may have jumps of size two or larger and is therefore no longer a counting process. Often, however, we may write $N(t) = \sum_{i=1}^n N_i(t)$, where $N_i(t)$ is a counting process registering the events for individual i. If we consider a homogeneous population where the rates of occurrence of the events are the same for all individuals, we may adopt the discrete extension of the model described in [3, pp. 180–181]. For this extended model, the arguments of [8, pp. 94–96], apply, to show that the variance estimator (2) is almost unbiased also in the presence of ties. This justifies the use of the tie-corrected estimator (2) for all situations considered above, except for the model with relative mortality and the epidemic model. Within the framework of the extended model the Nelson–Aalen estimator is a **nonparametric maximum likelihood** estimator; see [3, Section IV.1.5] for details and further discussion.

Weak Convergence and Confidence Bands

By (7) the martingale central limit theorem may be used to prove that, considered as a stochastic process, the Nelson–Aalen estimator (properly normalized) converges weakly to a mean zero Gaussian martingale. In particular, for a fixed t it is asymptotically normally distributed, a fact that was used in connection with the confidence intervals (3) and (4). The weak convergence result also makes it possible to derive confidence bands for A, i.e. limits which contain $A(t)$ for all t in an interval $[\tau_1, \tau_2]$ with a prespecified probability.

One important class of such confidence bands are the equal precision bands. The standard and log-transformed equal precision bands are obtained by replacing $z_{1-\alpha/2}$ in (3) and (4) by $d_{1-\alpha}$, the $1-\alpha$ fractile in the distribution of the supremum of the absolute value of a standardized Brownian bridge (over a certain time interval). This fractile may be found (approximately) by solving (with respect to d)

the nonlinear equation

$$\frac{4\phi(d)}{d} + 2\phi(d)\left(d - \frac{1}{d}\right)\ln\left[\frac{\widehat{\sigma}(\tau_2)}{\widehat{\sigma}(\tau_1)}\right] = \alpha,$$

where $\phi(d)$ is the standard normal density function. The equal precision bands require $\widehat{\sigma}(\tau_1) > 0$, so they cannot be extended all the way down to $t = 0$. Typically, one will also omit the largest values of t. The standard equal precision band has poor small sample properties, so even with sample sizes in the hundreds the use of the log transformed confidence band is recommended [5]. As an illustration we use once more the liver cirrhosis example. Considering the interval from 4 months (1/3 year) to 8 years, we have $\widehat{\sigma}(1/3) = 0.027$ and $\widehat{\sigma}(8) = 0.163$, so that $d_{0.95} = 2.99$. Therefore the 95% log transformed equal precision band for the cumulative hazard rate function between 4 months and 8 years may be obtained from (4) by using the fractile 2.99 instead of the value 1.96 used for the pointwise confidence intervals in Figure 1. A detailed study of the weak convergence of the Nelson–Aalen estimator and the derivation of confidence bands are provided by [3, Section IV.1.2-3]. Here another class of confidence bands, the Hall–Wellner bands, is also discussed.

We finally note that **semi-Markov processes** (or Markov **renewal processes**), where the transition intensities (only) depend on the sojourn times in the states, do not give rise to counting processes which fulfill the multiplicative intensity model (5). Thus the results outlined above do not immediately extend to such models. However, it turns out that enough of the above structure is preserved to be able to define Nelson–Aalen estimators also for such semi-Markov processes and to derive identical asymptotic results for these as for the case of Markov processes; see [3, Section X.1] for a discussion and further references.

References

[1] Aalen, O.O. (1978). Nonparametric inference for a family of counting processes, *Annals of Statistics* **6**, 701–726.

[2] Altshuler, B. (1970). Theory for the measurement of competing risks in animal experiments, *Mathematical Biosciences* **6**, 1–11.

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

[4] Becker, N.G. (1993). *Analysis of Infectious Disease Data*. Chapman & Hall, London.

[5] Bie, O., Borgan, O. & Liestøl, K. (1987). Confidence intervals and confidence bands for the cumulative hazard rate function and their small sample properties, *Scandinavian Journal of Statistics* **14**, 221–233.

[6] Breslow, N.E. & Day, N.E. (1980). *Statistical Methods in Cancer Research*. Vol. **1**: The Analysis of Case-Control Studies, IARC Scientific Publications, Vol. **32**. International Agency for Research on Cancer, Lyon.

[7] Breslow, N.E. & Day, N.E. (1987). *Statistical Methods in Cancer Research*. Vol. **2**: The Design and Analysis of Cohort Studies, IARC Scientific Publications, Vol. **82**. International Agency for Research on Cancer, Lyon.

[8] Fleming, T.R. & Harrington, D.P. (1991). *Counting Processes and Survival Analysis*. Wiley, New York.

[9] Keiding, N. & Andersen, P.K. (1989). Nonparametric estimation of transition intensities and transition probabilities: a case study of a two-state Markov process, *Applied Statistics* **38**, 319–329.

[10] Nelson, W. (1969). Hazard plotting for incomplete failure data, *Journal of Quality Technology* **1**, 27–52.

[11] Nelson, W. (1972). Theory and applications of hazard plotting for censored failure data, *Technometrics* **14**, 945–965.

[12] Schlichting, P., Christensen, E., Andersen, P.K., Fauerholdt, L., Juhl, E., Poulsen, H. & Tygstrup, N., for The Copenhagen Study Group for Liver Diseases (1983). Prognostic factors in cirrhosis identified by Cox's regression model, *Hepatology* **3**, 889–895.

ØRNULF BORGAN