Survival functions, cumulative hazards, and product integrals: the general case

Uncensored survival time T

Survival function: $S(t) = P(T > t)$

For the absolute continuous case, the hazard is given by:

$$\alpha(t) = \lim_{\Delta t \to 0} \frac{1}{\Delta t} P(T < t + \Delta t \mid T \geq t)$$

Cumulative hazard: $A(t) = \int_0^t \alpha(u)du$

We have the relations:

$$\alpha(t) = A'(t) = -\frac{S'(t)}{S(t)} \qquad S(t) = \exp\{-A(t)\}$$

For a general distribution the hazard rate is not defined, but we may define the cumulative hazard rate as (generalizing the first relation above):

$$A(t) = -\int_0^t \frac{dS(u)}{S(u-)}$$

For the discrete case $A(t)$ is a step function with increments

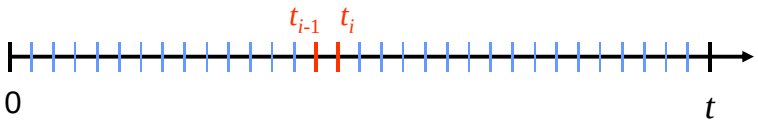
$$\Delta A(u) = -\frac{\Delta S(u)}{S(u-)}$$

$$= P(T = u \mid T \geq u)$$

How can the second relation above be generalized?

Need product-integrals to achieve this generalization

Partition $[0,t]$ into small time intervals:



$$S(t) = \lim_{\max|t_i - t_{i-1}| \to 0} \prod \left(1 - \{A(t_i) - A(t_{i-1})\}\right)$$

$$\stackrel{\text{def}}{=} \prod_{0 \leq u \leq t} \left(1 - dA(u)\right)$$

The limit is a product-integral

For the continuous case we have:

$$\prod_{0 \leq u \leq t} \left(1 - dA(u)\right) = \exp\{-A(t)\}$$

For the discrete case we have:

$$\prod_{0 \leq u \leq t} \left(1 - dA(u)\right) = \prod_{u \leq t} \left(1 - \Delta A(u)\right)$$

where $\Delta A(u) = P(T = u \mid T \geq u)$ is the increment of the cumulative hazard at time u

For the general case we have a mixture of the two

The Kaplan-Meier estimator

For right censored survival data we observe:

$\tilde{T}_i = \min\{\text{survival time } T_i \text{ , censoring time } C_i\}$

$D_i = I\{\tilde{T}_i = T_i\}$

Model: the uncensored survival times T_i are *iid* with hazard rate $\alpha(t)$

Counting and intensity processes:

$$N_i(t) = I\left\{\tilde{T}_i \le t, D_i = 1\right\}$$

$$\lambda_i(t) = I\left\{\tilde{T}_i \ge t\right\}\alpha(t) = Y_i(t)\alpha(t)$$

Aggregated counting process:

$$N(t) = \sum_{i=1}^{n} N_i(t)$$

Intensity process:

$$\lambda(t) = \sum_{i=1}^{n} \lambda_i(t) = Y(t)\alpha(t)$$

with

$$Y(t) = \sum_{i=1}^{n} I\left\{\tilde{T}_i \ge t\right\}$$

the number at risk just before time t

Nelson-Aalen estimator:

$$\hat{A}(t) = \int_0^t \frac{dN(u)}{Y(u)} = \sum_{u \le t} \frac{\Delta N(u)}{Y(u)} = \sum_{T_j \le t} \frac{1}{Y(T_j)}$$

Plug this into the product-integral expression for the survival function (Nelson-Alen is a step-function):

$$\hat{S}(t) = \prod_{0 \le u \le t} \left(1 - d\hat{A}(u)\right) = \prod_{u \le t} \left(1 - \Delta\hat{A}(u)\right)$$

$$= \prod_{u \le t} \left(1 - \frac{\Delta N(u)}{Y(u)}\right) = \prod_{T_j \le t} \left(1 - \frac{1}{Y(T_j)}\right)$$

This is the Kaplan-Meier estimator

Example 3.8: Second births

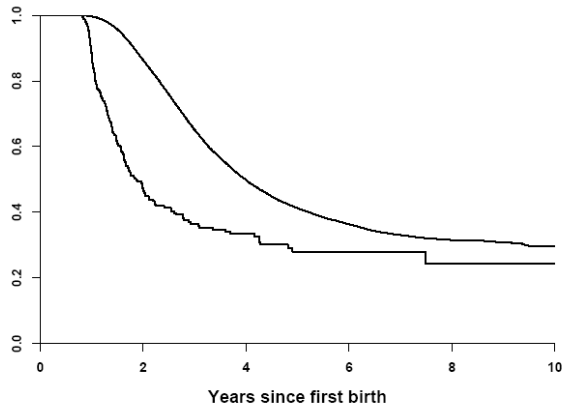


Fig. 3.11 *Kaplan-Meier estimates for the time between first and second birth. Upper curve: first child survived one year; lower curve: first child died within one year.*

Example 3.9: Third births

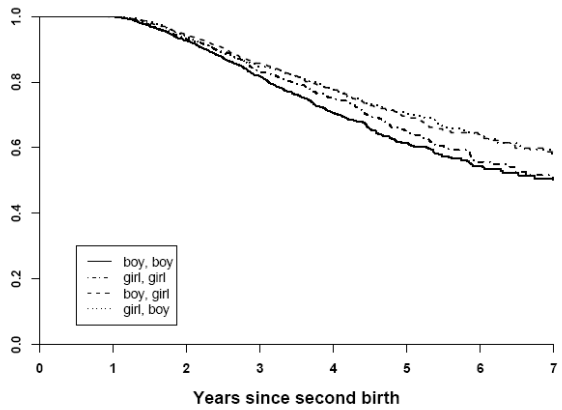


Fig. 3.12 *Kaplan-Meier estimates for the time between the second and third births depending on the gender of the two older children.*

An alternative estimator of the survival function is

$$\tilde{S}(t) = \exp\left\{-\hat{A}(t)\right\}$$

$$= \exp\left\{-\sum_{T_j \le t} \frac{1}{Y(T_j)}\right\}$$

$$= \prod_{T_j \le t} \exp\left\{-\frac{1}{Y(T_j)}\right\}$$

For practical purposes there is little difference between the two estimators

But from a theoretical point of view, the Kaplan-Meier estimator is the natural one (and it may be generalized to Markov models)

Kaplan-Meier estimator: Properties

$$A^*(t) = \int_0^t J(s)\alpha(s)ds \approx A(t)$$

$$S^*(t) = \prod_{0 \le s \le t} \left(1 - dA^*(s)\right) = \exp\left\{-A^*(t)\right\} \approx S(t)$$

May show that (this is Duhamel's equation)

$$\frac{\hat{S}(t)}{S^*(t)} - 1 = -\int_0^t \underbrace{\frac{\hat{S}(s-)}{S^*(s)}}_{\approx 1} d(\hat{A} - A^*)(s) \approx -\left(\hat{A}(t) - A^*(t)\right)$$

Asymptotically: $\frac{\hat{S}(t)}{S(t)} - 1 \approx -\left(\hat{A}(t) - A(t)\right)$

Thus:

$$\hat{S}(t) - S(t) \approx -S(t) \cdot \left(\hat{A}(t) - A(t)\right)$$

The statistical properties for Kaplan-Meier may be derived from those of Nelson-Aalen:

- $\mathrm{Var}\{\hat{S}(t)\} \approx \{S(t)\}^2 \cdot \mathrm{Var}\{\hat{A}(t)\}$
- Variance estimator: $\hat{\tau}^2(t) = \{\hat{S}(t)\}^2 \cdot \hat{\sigma}^2(t)$
  with $\hat{\sigma}^2(t) = \int_0^t \{Y(s)\}^{-2} dN(s)$
- $\hat{S}(t)$ is as.normally distributed around $S(t)$

Usually the variance is estimated by *Greenwood's formula:*

$$\tilde{\tau}^2(t) = [\hat{S}(t)]^2 \cdot \tilde{\sigma}^2(t)$$

$$\text{with} \quad \tilde{\sigma}^2(t) = \int_0^t [Y(s)\{Y(s) - \Delta N(s)\}]^{-1} dN(s)$$

Only minor difference between the two variance estimators

Pointwise 95% confidence limits for $S(t)$

Linear: $\hat{S}(t) \pm 1.96 \cdot \hat{S}(t) \cdot \hat{\sigma}(t)$

Log-log-transformed: $\hat{S}(t)^{\exp\left\{\pm 1.96 \cdot \hat{\sigma}(t)/\log \hat{S}(t)\right\}}$

(cf. Exercise 3.6)

Estimation of mean survival time

The mean survival time is given by (exercise 1.3)

$$\mathrm{E}(T) = \int_0^\infty S(u)du$$

Due to censoring, this may usually not be estimated

We may consider the expected survival in $[0,t]$:

$$\mu_t = \int_0^t S(u)du$$

This may be estimated by

$$\mu_t = \int_0^t \hat{S}(u)du$$

Estimation of median survival time and other fractiles of the survival distribution

The *p*-th fractile ξ_p of the survival distribution is given by (exercise 1.2)

$$F(\xi_p) = p \quad \text{or equivalently} \quad S(\xi_p) = 1 - p$$

It is estimated by

$$\hat{\xi}_p = \inf\left\{t : \hat{S}(t) \le 1 - p\right\}$$

Confidence intervals may be found by "inverting" the confidence intervals for the survival function (exercise 3.8)

Example 3.10: Median time between first and second births

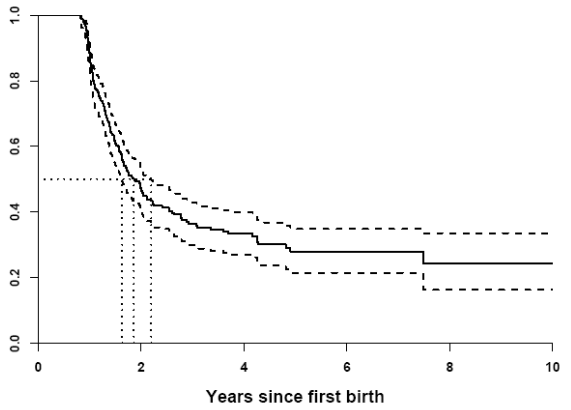


Fig. 3.13 *Kaplan-Meier estimate with 95% log-log-transformed confidence intervals for the time between first and second birth for women who lost the first child within one year after birth. It is indicated at the figure how one may obtain the estimated median time with confidence limits.*