STK4080/9080
Survival and event history analysis

Lecture 4:

- Nelson-Aalen estimator
- Multiplicative intensity model
- Survival functions, cumulative hazards and product integrals
- Kaplan-Meier estimator
- Estimating restricted means and fractiles

Nelson-Aalen estimator

We have right censored and/or left truncated survival data for a sample of n individuals from a population with hazard rate $\alpha(t)$

Let $N_i(t)$ count the observed number of occurrences (0 or 1) of the event of interest for individual i

Provided censoring/truncation is independent, the corresponding intensity process takes the form:

$$\lambda_i(t) = \underbrace{Y_i(t)}_{\text{at risk indicator}} \underbrace{\alpha(t)}_{\text{hazard rate}}$$

The aggregated counting process

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

has intensity process

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

where

$$Y(t) = \sum_{i=1}^{n} Y_i(t)$$

is the number at risk "just before" time t

We will estimate the cumulative hazard

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

We have the decomposition:

$$dN(t) = \lambda(t)dt + dM(t)$$
$$= \underbrace{Y(t)\cdot dA(t)}_{\text{signal}} + \underbrace{dM(t)}_{\text{noise}}$$

Estimating equation (when $Y(t) > 0$)

$$dN(t) = Y(t)\cdot d\hat{A}(t)$$

Thus (when $Y(t) > 0$):

$$d\hat{A}(t) = \frac{dN(t)}{Y(t)}$$

The Nelson-Aalen estimator

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

is a sum over the jump times $T_1 < T_2 < \cdots$ of $N(t)$

We will show (later) that the Nelson-Aalen estimator is (almost) unbiased with a variance that may be estimated by

$$\hat{\sigma}^2(t) = \int_0^t \frac{dN(s)}{Y(s)^2} = \sum_{T_j \le t} \frac{1}{Y(T_j)^2}$$

Approximate 95% pointwise confidence limits:

Standard: $\hat{A}(t) \pm 1.96 \cdot \hat{\sigma}(t)$

Log-transformed: $\hat{A}(t) \cdot \exp\{\pm 1.96 \cdot \hat{\sigma}(t) / \hat{A}(t)\}$

(cf. Exercise 3.3)

Example 3.1: Second births

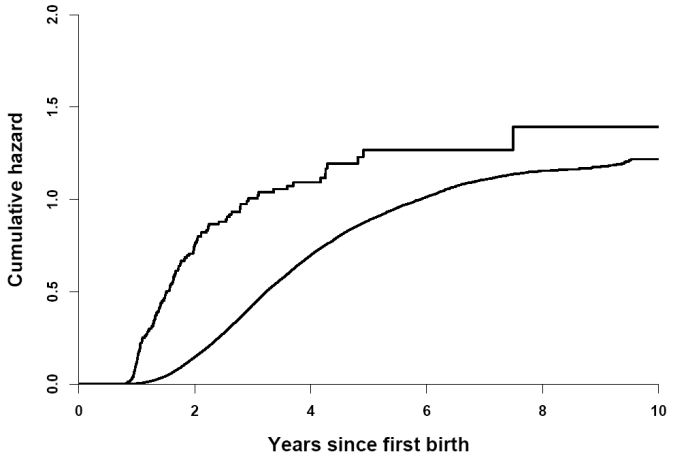


Fig. 3.1 *Nelson-Aalen estimates for the time between first and second births. Lower curve: first child survived one year; upper curve: first child died within one year.*

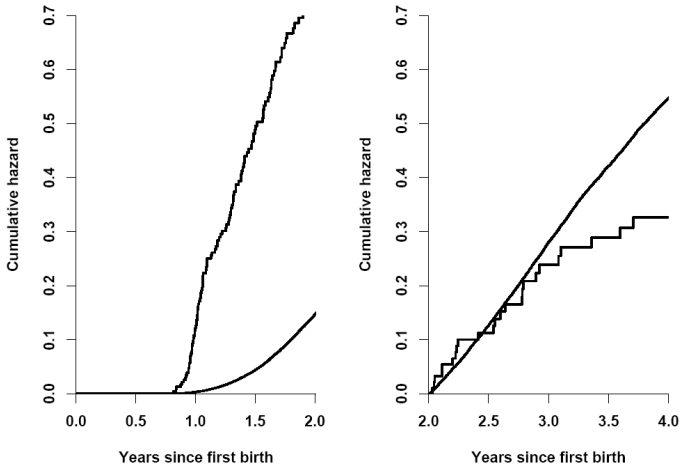


Fig. 3.3 *Nelson-Aalen estimates for the time between first and second births. Left panel: first two years after the first birth; right panel: two to four years after the first birth. Smooth curve: first child survived one year; irregular curve: first child died within one year.*

Example 3.2: Third births

Fig. 3.4 *Nelson-Aalen estimates for the time between the second and third births depending on the gender of the two older children.*

Using R: exercise 3.1

Placebo	1	1	2	2	3	4	4	5	5	8	8
	8	8	11	11	12	12	15	17	22	23	
6-MP	6	6	6	6*	7	9*	10	10*	11*	13	16
	17*	19*	20*	22	23	25*	32*	32*	34*	35*	

```
# Read data:

leukemia=
read.table("http://folk.uio.no/borgan/abg-2008/data/leukemia.txt",header=T)

# Compute Nelson-Aalen estimates and plot them in one figure
# (using the default "efron" method for handling tied failure times):

fit=coxph(Surv(time,status)~strata(treat), data=leukemia)

surv=survfit(fit)

plot(surv, fun="cumhaz", mark.time=F, xlim=c(0,25), ylim=c(0,4),
     xlab="Weeks", ylab="Cumulative hazard", lty=1:2)

legend("topleft", c("Placebo","6-MP"), lty=1:2)
```


The multiplicative intensity model

Consider a counting process $N(t)$ with intensity process of the multiplicative form

$$\lambda(t) = Y(t)\alpha(t)$$

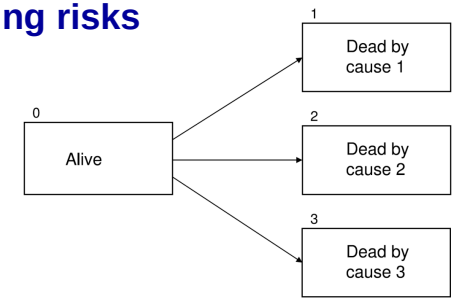
Here $Y(t)$ is *predictable process* that does not depend on unknown parameters and $\alpha(t)$ is a non-negative *parameter function*

The Nelson-Aalen estimator applies to all counting processes fulfilling the multiplicative intensity model

The main example of the multiplicative intensity model is right censored and/or left truncated survival data as considered above

Example 3.3: Competing risks

Consider a competing risks model with *k* causes of death



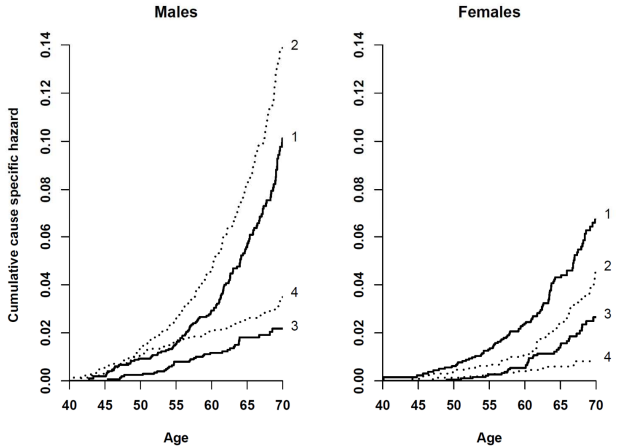
For each cause *h* we define the cause-specific hazard $\alpha_{0h}(t)$ given by

$$\alpha_{0h}(t)dt = P(\text{die from cause } h \text{ in } [t, t+dt) \mid \text{alive at } t-)$$

Based on a sample from a population, we let $N_{0h}(t)$ count the number of observed 0 –> *h* transitions in [0, *t*] and let $Y_0(t)$ be the number at risk (i.e. in state 0) just prior to time *t*

The intensity process of $N_{0h}(t)$ takes the multiplicative form $\lambda_{0h}(t) = \alpha_{0h}(t)Y_0(t)$ so Nelson-Aalen applies

Nelson-Aalen estimates for the cause-specific mortality according to cause of death and sex (data from health screenings in three Norwegian counties):



1) Cancer
2) Cardiovascular disease
3) Other medical
4) Alcohol abuse, violence, accidents

Example 3.4: Relative mortality

Consider censored/truncated survival data, let $N_i(t)$ count the observed number of deaths (0 or 1) for individual i, and assume that its intensity process takes the form:

$$\lambda_i(t) = \underbrace{Y_i(t)}_{\text{at risk indicator}} \underbrace{\alpha(t)}_{\text{relative mortality}} \underbrace{\mu_i(t)}_{\text{population mortality (known)}}$$

The aggregated counting process $N(t) = \sum_{i=1}^{n} N_i(t)$ has intensity process of the multiplicative form

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

with $Y(t) = \sum_{i=1}^{n} Y_i(t)\,\mu_i(t)$, so Nelson-Aalen applies

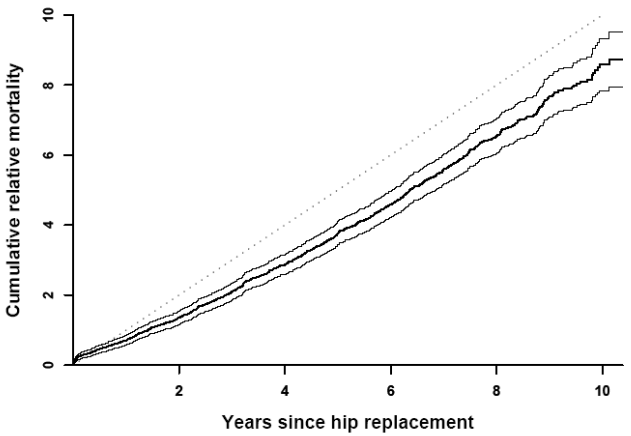


Fig. 3.7 *Nelson-Aalen estimate of the relative cumulative relative mortality with 95% standard confidence intervals for patients who have had a hip replacement in Norway in the period 1987–97. A dotted line with unit slope is included for easy reference.*

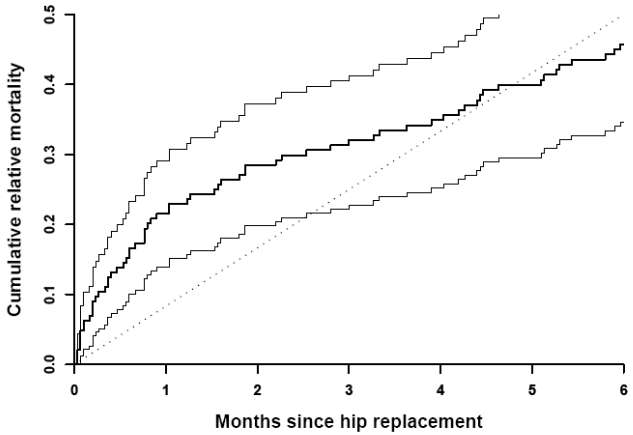


Fig. 3.8 *Nelson-Aalen estimate of the relative cumulative relative mortality with 95% standard confidence intervals for the first six months after the operation for patients who have had a hip replacement in Norway in the period 1987–97.*

Example 3.6: Mating of Drosophila flies

30 female virgin flies and 40 male virgin flies are put in a plastic bowl ("pornoscope") and times on initiatings of matings are recorded.

Two experiments: one experiment with "ebony" flies and one with "oregon" flies

Ebony	143	180	184	303	380	431	455	475	500	514
	521	552	558	606	650	667	683	782	799	849
	901	995	1131	1216	1591	1702	2212			
Oregon	555	742	746	795	934	967	982	1043	1055	1067
	1081	1296	1353	1361	1462	1731	1985	2051	2292	2335
	2514	2570	2970							

$N(t)$ counts the number of matings in $[0, t\,]$

Assuming random mating, the intensity process takes the multiplicative form

$$\lambda(t) = \alpha(t) f(t) m(t)$$

where

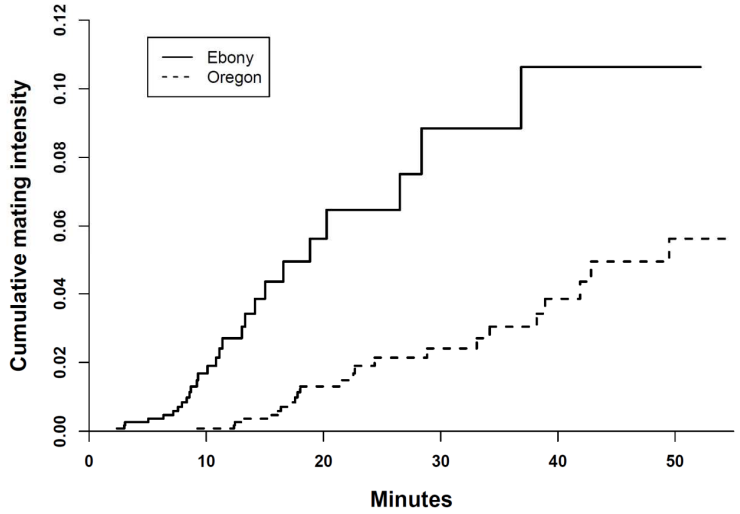
$$f(t) = 30 - N(t-)$$
$$m(t) = 40 - N(t-)$$

are the number of virgin female and male flies just before time *t*

Nelson-Aalen estimator of cumulative mating intensity

$$\hat{A}(t) = \sum_{T_j \le t} \frac{1}{f(T_j)\, m(T_j)}$$

A similar approach may be used to model the spread of an epidemic (if the times of infections are observed)



Statistical properties of Nelson-Aalen estimator

To handle the possibility that *Y*(*t*) = 0, we introduce the indicator $J(t) = I\{Y(t) > 0\}$ and interpret 0/0 as 0

Then we may write

$$\hat{A}(t) = \int_0^t \frac{dN(s)}{Y(s)} = \int_0^t \frac{J(s)}{Y(s)} dN(s)$$

$$= \int_0^t \frac{J(s)}{Y(s)} \{Y(s)\alpha(s)ds + dM(s)\}$$

$$= \underbrace{\int_0^t J(s)\alpha(s)ds}_{= A^*(t)} + \underbrace{\int_0^t \frac{J(s)}{Y(s)} dM(s)}_{= I(t)}$$

Thus we have the decomposition:

$$\hat{A}(t) = \underbrace{A^*(t)}_{\text{systematic part}} + \underbrace{I(t)}_{\text{random part}}$$

The random part is the stochastic integral:

$$I(t) = \int_0^t H(s) dM(s)$$

where $H(t) = J(t)/Y(t)$ is a predictable process (i.e. its value at time *t* is known "just before" *t*)

We may use properties of stochastic integrals to study the Nelson-Aalen estimator

The stochastic integral $I(t)$ is a mean zero martingale

In particular

$$\mathrm{E}\{I(t)\} = 0$$

Thus:

$$\begin{aligned}\mathrm{E}\{\hat{A}(t)\} &= \mathrm{E}\{A^*(t) + I(t)\} \\ &= \mathrm{E}\{A^*(t)\} + 0 \\ &= \int_0^t P(Y(s) > 0)\,\alpha(s)\,ds \\ &\approx A(t)\end{aligned}$$

The Nelson-Aalen estimator is approximately unbiased

For the martingale $I(t)$ we have that

$$I^2(t) - [I](t)$$

is a mean zero martingale

It follows that

$$\mathrm{Var}\{I(t)\} = \mathrm{E}\{I^2(t)\} = \mathrm{E}[I](t)$$

Thus an unbiased estimator for the variance of $I(t)$ and an approximately unbiased estimator for the variance of the Nelson-Aalen estimator is

$$\begin{aligned}\hat{\sigma}^2(t) = [I](t) &= \int_0^t \left(J(s)/Y(s)\right)^2 dN(s) \\ &= \sum_{T_i \le t} \frac{1}{Y(T_i)^2}\end{aligned}$$

Suppose that $N(t)$ is an aggregated process obtained from observation of n subjects (as in all examples except the Drosophila flies)

Assume that

$$Y(s)/n \to y(s) > 0 \text{ as } n \to \infty \quad \text{for all} \quad s \in [0, \tau]$$

By the martingale central limit theorem we then have that (cf. lecture 3, slides 26-27)

$$\sqrt{n}\left(\hat{A}(t) - A(t)\right) \to U(t) \qquad \text{(in distribution)}$$

where $U(t)$ is a mean zero Gaussian martingale with variance function

$$\sigma^2(t) = \int_0^t \frac{\alpha(s)}{y(s)} ds$$

In particular for a fixed value of t the Nelson-Aalen estimator $\hat{A}(t)$ is approximately normally distributed around the true value of $A(t)$ with a variance that may be estimated by $\hat{\sigma}^2(t)$

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 at the bottom of the previous slide):

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

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

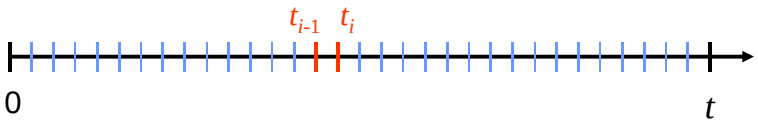
$$\Delta A(u) = -\frac{\Delta S(u)}{S(u-)}$$

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

How can we generalize the second relation at the bottom of the previous slide?

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-Aalen 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

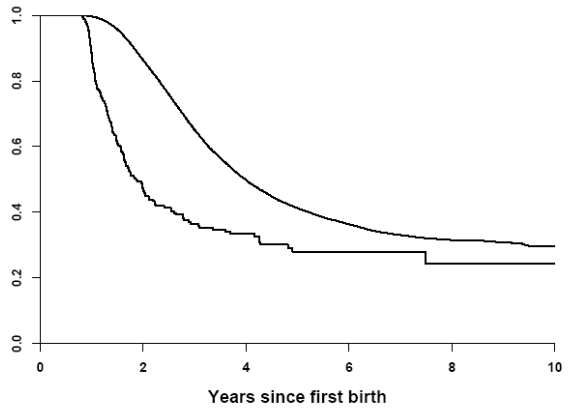


Fig. 3.11 *Kaplan-Meier estimates for the time between first and second birth. Upper curve: first child survived one year; lower curve: first child died within one year.*

Example 3.9: Third births

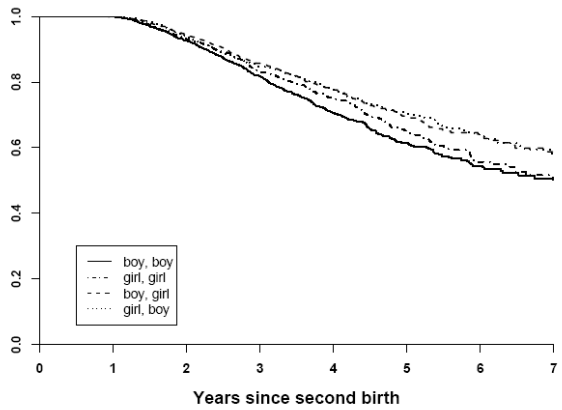


Fig. 3.12 *Kaplan-Meier estimates for the time between the second and third births depending on the gender of the two older children.*

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 $S(t)$

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

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

(cf. Exercise 3.6)

Using R exercises 3.4 and 3.7

Placebo	1	1	2	2	3	4	4	5	5	8	8
	8	8	11	11	12	12	15	17	22	23	
6-MP	6	6	6	6*	7	9*	10	10*	11*	13	16
	17*	19*	20*	22	23	25*	32*	32*	34*	35*	

```
# Read data:

leukemia=
read.table("http://folk.uio.no/borgan/abg-2008/data/leukemia.txt",header=T)

# Compute Kaplan-Meier estimates and plot them in one figure

fit=survfit(Surv(time,status)~treat, data=leukemia, conf.type="none")
plot(fit, mark.time=F, xlim=c(0,25), xlab="Weeks", ylab="Survival", lty=1:2)
legend("topright", c("Placebo","6-MP"), lty=1:2)

#  Kaplan-Meier estimate for placebo group with standard confidence limits

fit.p=survfit(Surv(time,status)~1, data=leukemia, conf.type="plain",subset=(treat==1))
plot(fit.p, mark.time=F, xlim=c(0,25), xlab="Weeks", ylab="Survival",main="Placebo")
```

```
#  Kaplan-Meier estimate for placebo group with log-log transformed
#  confidence limits

fit.p=survfit(Surv(time,status)~1, data=leukemia, conf.type="log-log",subset=(treat==1))
plot(fit.p, mark.time=F, xlim=c(0,25), xlab="Weeks", ylab="Survival",main="Placebo")
```

The default in R is to used a log-transformed confidence interval, but that is not a good idea

To obtain confidence intervals for the survival function, you should use conf.type="plain" or conf.type="log-log"

When the survfit-command is used to obtain confidence intervals for the cumulative hazard (cf slide 9), conf.type="log" will give a standard confidence interval, for the cumulative hazard while conf.type="log-log" will give a log-transformed confidence interval

Estimation of the restricted mean

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

But we may consider the restricted mean, i.e. the expected survival in $[0,t]$:

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

This may be estimated by

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

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

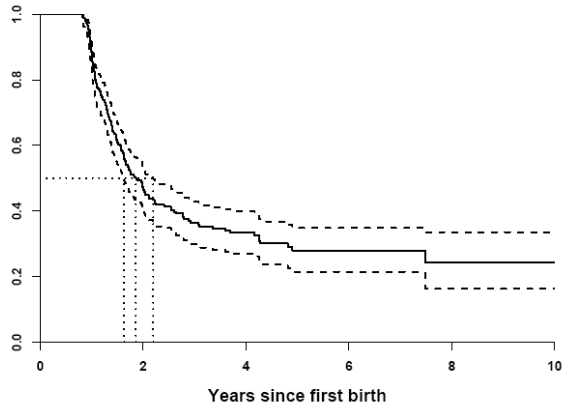


Fig. 3.13 *Kaplan-Meier estimate with 95% log-log-transformed confidence intervals for the time between first and second birth for women who lost the first child within one year after birth. It is indicated at the figure how one may obtain the estimated median time with confidence limits.*

Using R: exercise 3.7 and more

Placebo	1	1	2	2	3	4	4	5	5	8	8
	8	8	11	11	12	12	15	17	22	23	
6-MP	6	6	6	6*	7	9*	10	10*	11*	13	16
	17*	19*	20*	22	23	25*	32*	32*	34*	35*	

```
# Estimate of restricted mean lifetime and median lifetime
#  with standard confidence limits
fit=survfit(Surv(time,status)~treat, data=leukemia, conf.type="plain")
print(fit, rmean=30)

# One may find the median (and other percentiles) with confidence limits from
# the output of the summary command
summary(fit.p)
# Or they may be found directly by the command
quantile(fit,probs=c(0.25,0.50,0.75))
```