PhD course in “Advanced survival analysis”

Introduction. The Nelson-Aalen and Kaplan-Meier estimators

Examples and general approach follow the textbook: “Statistical Models Based on Counting Processes” by Andersen, Borgan, Gill and Keiding, Springer-Verlag, 1993 but the same material is to a large extent covered by Martinussen and Scheike, Ch. 2-4.

Per Kragh Andersen: 4 October 2021

Empirical phenomena modeled and analysed:

- **Discrete events** occurring in
- **continuous time**
- Main example - survival data:
  - Ex. I.3.1: Malignant melanoma
  - Ex. I.3.2: Insulin-dependent diabetes
- but also - multi-state models:
  - Ex. I.3.9: Malignant melanoma
  - Ex. I.3.11: Diabetes, nephropathy
  - Ex. I.3.12: Liver cirrhosis, prothrombin

Malignant melanoma

- 205 patients with malignant melanoma (skin cancer) operated (1962-77) at Odense University Hospital, DK
- Followed from time of operation to death or 31 Dec. 1977
  - 57 died from the disease
  - 14 died from other causes
  - 134 were alive 31 Dec. 1977
- Purpose of study: assess the effect of risk factors on survival:
  - sex, age at operation
  - tumour thickness, ulceration
  - cell types, ...
- Difficulty (inherent in survival data):
  - **right-censoring**

Fig. II.1.1

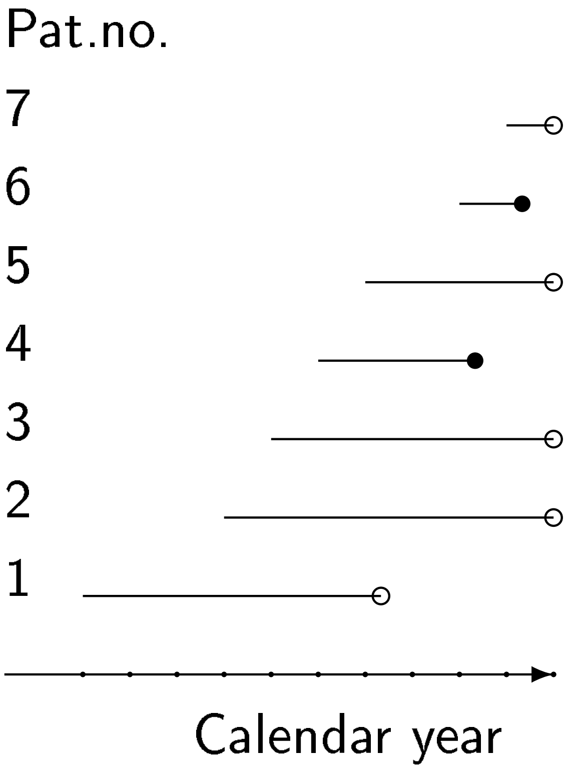
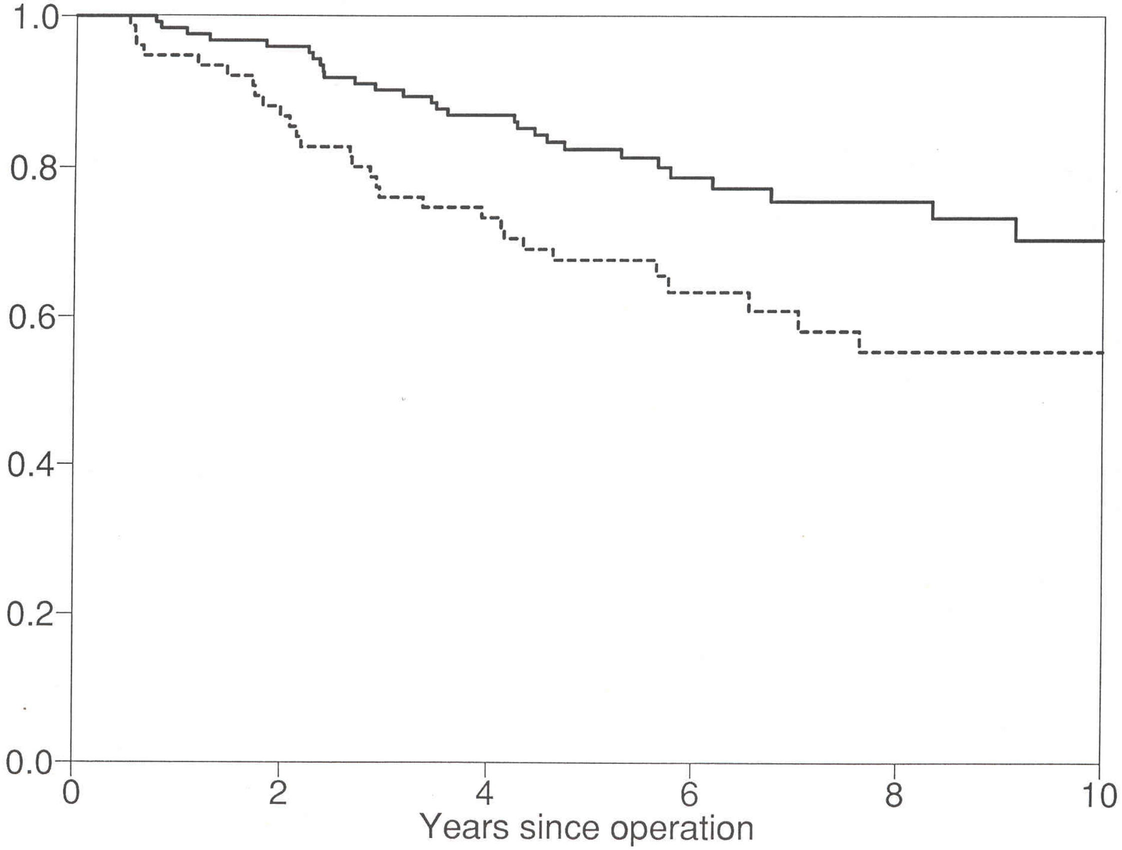


FIG.I.3.2



Females: solid, males: dashed (We cheat a bit and consider only deaths from the disease. Competing risks is properly treated later in the course)

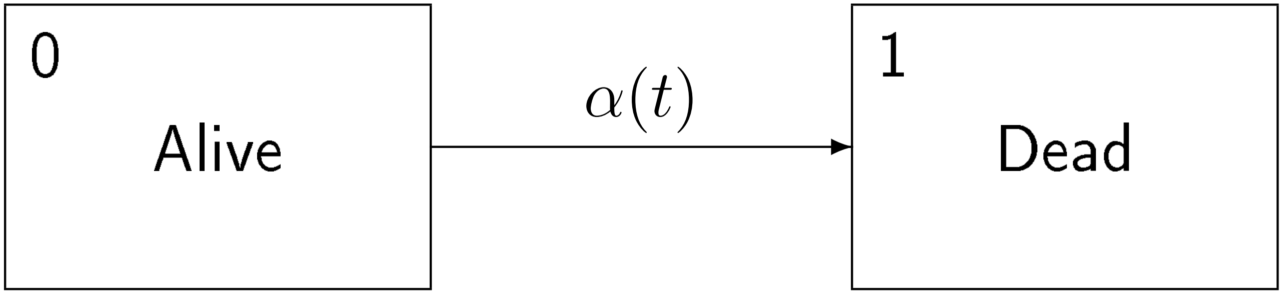
Insulin-dependent diabetes

- 1 July 1973: 1499 persons in Fyn county, DK suffered from IDDM. Ascertained via prescriptions from a 5 month period in 1973
- The diabetics are followed from 1 July 1973 to:
  - death (491)
  - emigration (2) or 31 Dec. 1981 (1006)
- Purpose of study: evaluate age- and sex- specific mortality among diabetics (adjusting for disease duration, standard mortality) – *prevalent cohort study*
- Difficulties:
  - right-censoring (as in malignant melanoma study)
  - delayed entry (left-truncation) - individuals followed from age (or disease duration) at 1 July 1973 and not from birth (or disease onset)

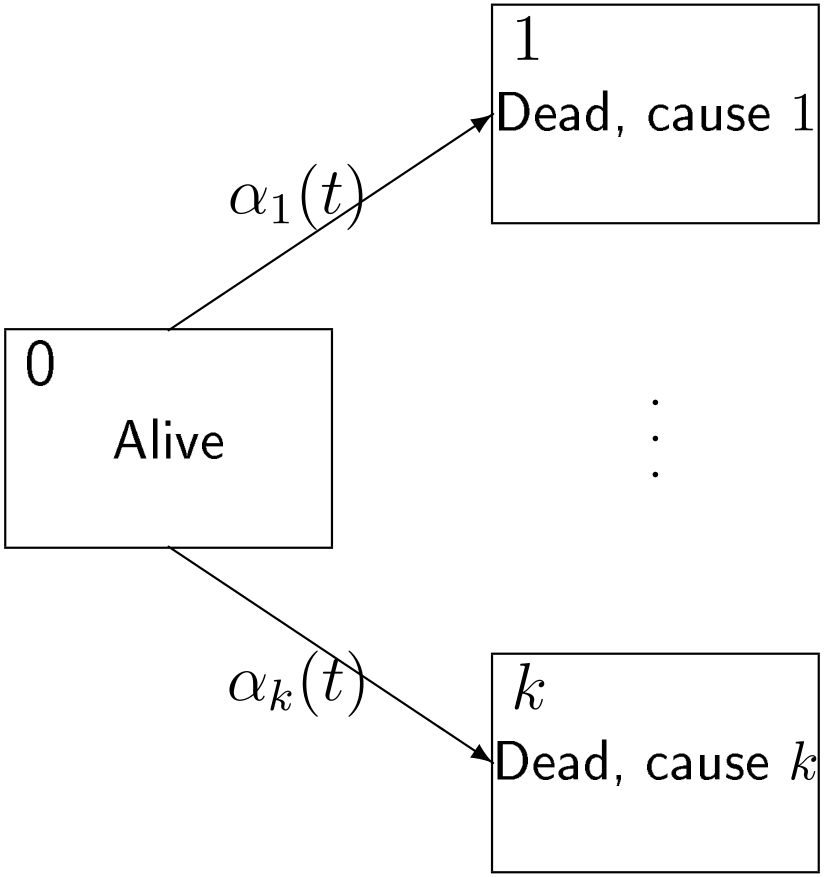
Insulin-dependent diabetes - Table I.3.1

Age 1 July 1973	No. alive 1 July 1973	Deaths before Dec. 1981	Proportion of deaths
		Males	
0-29	207	9	.043
30-39	115	22	.191
40-49	124	25	.202
50-59	122	43	.352
60-69	126	78	.619
70+	89	77	.865
		Females	
0-29	146	6	.041
30-39	89	10	.112
40-49	83	10	.120
50-59	107	29	.271
60-69	143	69	.483
70+	148	113	.769

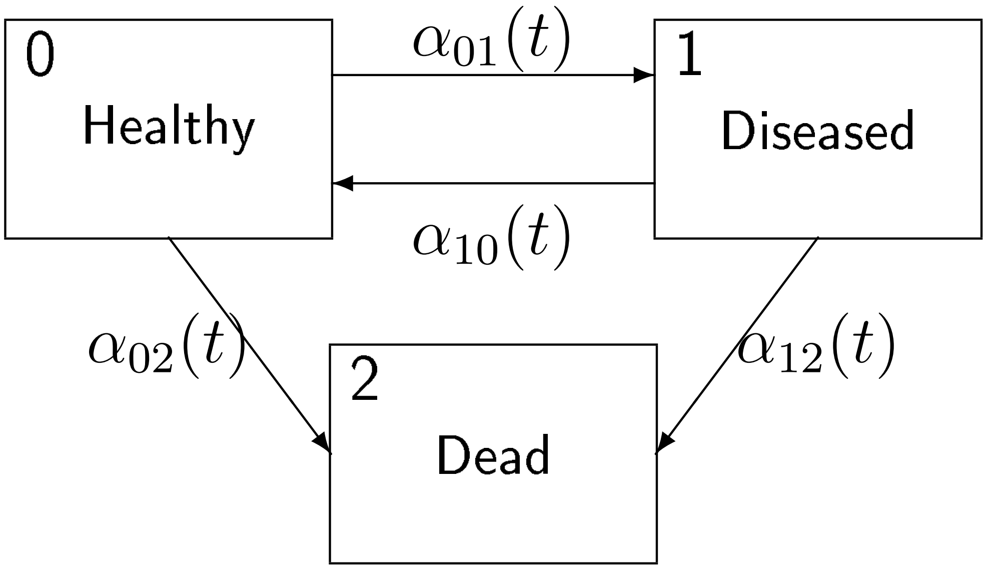
Two-state model for survival data: Fig. I.3.1



Competing risks: Fig. I.3.5



Illness-death model (recurrent disease): Fig. I.3.6



Summary

- Data: transitions between states in stochastic process
- Incomplete observation: right-censoring, left-truncation (delayed entry)
- Other kinds of incomplete observation may occur but are not further discussed: left- (interval-) censoring, right-truncation
- Main example: survival data
- Often several possible time variables

Survival data

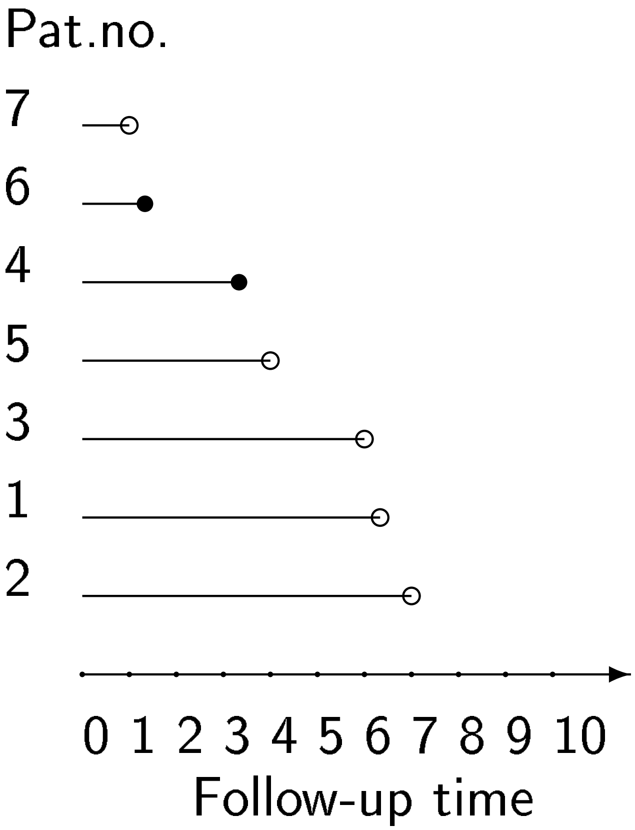
Why is there a need for special methods? $(*)$

- **Categorical data:**, e.g. comparison of 1 year survival when certain cancer patients are treated with drug A or drug B
- compare x_A/n_A with x_B/n_B (#deaths/#patients)
- problems: why 1 year? what about patients followed for <1 year?
- **Quantitative data:** use methods based on means, SD's or ranks
- problem: what about patients still alive by time of analysis?

Answer to $(*)$: because of the inevitable incomplete observation:

censoring

Fig. II.1.2



Meaningful computations?

Fraction of observation times exceeding 5 years: 3/7.

Average of all observation times:

$$(0.9 + 1.2 + 3.0 + 3.5 + 6.1 + 6.3 + 7.2)/7 = 28.2/7 = 4.0.$$

Average of completely observed observation times:

$$(1.2 + 3.0)/2 = 2.1.$$

Survival data

Why consider survival data as a

stochastic process?

Possible answers:

- It gives a framework where generalizations to other kinds of event history data are possible
- It gives a framework which makes some powerful mathematical results available
- It provides you with the right way of thinking of survival data

One uncensored survival time: Stochastic processes

X survival time with *survival function*:

$$S(t) = P(X > t) = \exp(-\int_0^t \alpha(u)du),$$

$\alpha(t)$ *hazard function*

$$\alpha(t) = -\frac{\mathrm{d}}{\mathrm{d}t}\log S(t) = -\frac{\mathrm{d}}{\mathrm{d}t}S(t)\frac{1}{S(t)},$$

i.e.,

$$\alpha(t)dt \approx P(X < t + dt \mid X > t).$$

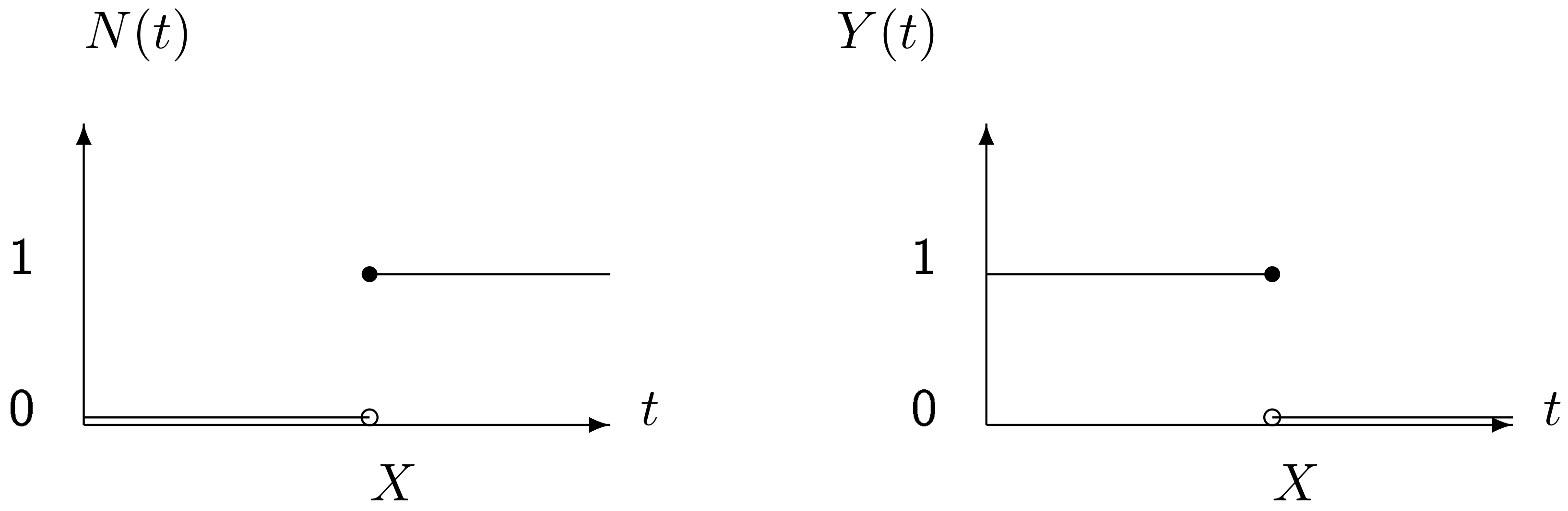
Let

$$N(t) = I(X \leq t),$$

counting process, counting +1 at X.

Right-continuous step function with a step of size 1 at X.

The processes $N(t)$ and $Y(t)$



Intensity process, compensator

Imagine that an individual is followed from time 0 and onwards:

At any time t, ask: what is the conditional probability that $N(\cdot)$ jumps "now" given "the past":

$$P(N(t+dt) - N(t-) = 1 \mid \mathcal{F}_{t-})$$

where $(\mathcal{F}_t)$: "history, filtration" - family of $\sigma-$algebras.

This is:

$$\approx \alpha(t)dt = \alpha(t)Y(t)dt, \text{ if } X \geq t$$

$$= 0 = \alpha(t)Y(t)dt, \text{ if } X < t$$

where $Y(t) = I(X \geq t)$ is the indicator of the individual being "still alive" just before time t (see figure on previous page).

Intensity process, compensator

Let

$$dN(t) = N(t + dt) - N(t-).$$

Then $dN(t) = 1$ or 0 and

$$E(dN(t) \mid \mathcal{F}_{t-}) = P(dN(t) = 1 \mid \mathcal{F}_{t-}) = \alpha(t)Y(t)dt = \lambda(t)dt.$$

Here, $\lambda(t)$ is the *intensity process* for $N(\cdot)$ (wrt. P and $(\mathcal{F}_t)$)

(NB: a stochastic process).

Let

$$\Lambda(t) = \int_0^t \lambda(u)du,$$

integrated intensity process or *compensator* for $N(\cdot)$.

Martingales

Look at: $M(t) = N(t) - \Lambda(t)$ or

$$dM(t) = dN(t) - d\Lambda(t) = dN(t) - \lambda(t)dt.$$

Then

$$E(dM(t) \mid \mathcal{F}_{t-}) = E(dN(t) - \lambda(t)dt \mid \mathcal{F}_{t-})$$

$$= \lambda(t)dt - E(\alpha(t)Y(t)dt \mid \mathcal{F}_{t-})$$

$$= \alpha(t)Y(t)dt - \alpha(t)Y(t)dt = 0,$$

since $Y(t)$ is *adapted to* $(\mathcal{F}_{t-})$, i.e. $M(\cdot)$ is a *martingale* (wrt. P and $(\mathcal{F}_t)$):

$$E(M(t) \mid \mathcal{F}_s) = M(s), s < t.$$

Martingales

This is because

$$E(M(t) \mid \mathcal{F}_s) - M(s)$$

$$= E(M(t) - M(s) \mid \mathcal{F}_s)$$

$$= E(\int_{(s,t]} dM(u) \mid \mathcal{F}_s)$$

$$= \int_{(s,t]} E(dM(u) \mid \mathcal{F}_s)$$

$$= \int_{(s,t]} E(E(dM(u) \mid \mathcal{F}_{u-}) \mid \mathcal{F}_s)$$

$$= 0.$$

The Doob-Meyer decomposition

The equation

$$N(t) = \Lambda(t) + M(t)$$

is the *Doob-Meyer* decomposition of (the sub-martingale) $N(t)$ into its *compensator* $\Lambda(t)$ and the martingale $M(t)$.

The decomposition is unique when *predictability* is assumed for $\Lambda(t)$, i.e. $\Lambda(t)$ is left-continuous and adapted to $(\mathcal{F}_{t-})$, in other words the value of $\Lambda(t)$ is "fixed given $(\mathcal{F}_{t-})$" or "known just before time t".

Right censoring

Some times we only observe X if $X \leq U$, a (potential) *right censoring time*.

Let the observed data be: $D = I(X \leq U), \quad \widetilde{X} = X \wedge U$.

Let $N(t) = I(\widetilde{X} \leq t, D = 1)$: *counting process* with *intensity process* given by:

$$\lambda(t)dt = P(dN(t) = 1 \mid \mathcal{F}_{t-}).$$

This is:

$$= 0, \text{ if } \widetilde{X} < t$$

$$= \alpha(t)dt, \text{ if } \widetilde{X} \geq t$$

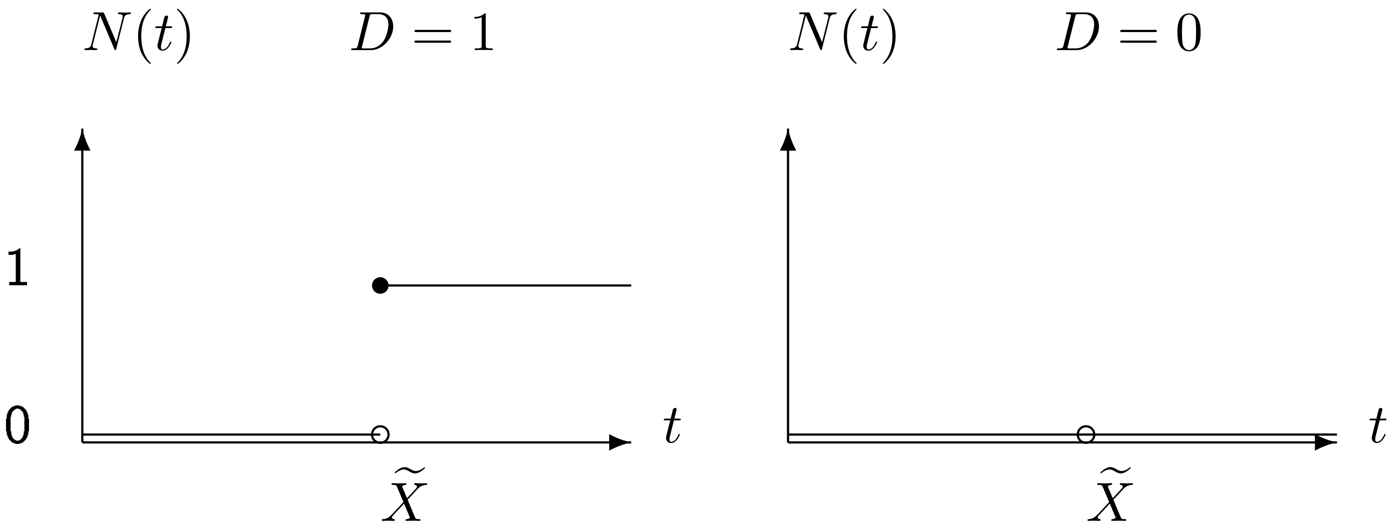
and censoring is "independent" - a *crucial* assumption, Sect.III.2.2.

Independent censoring

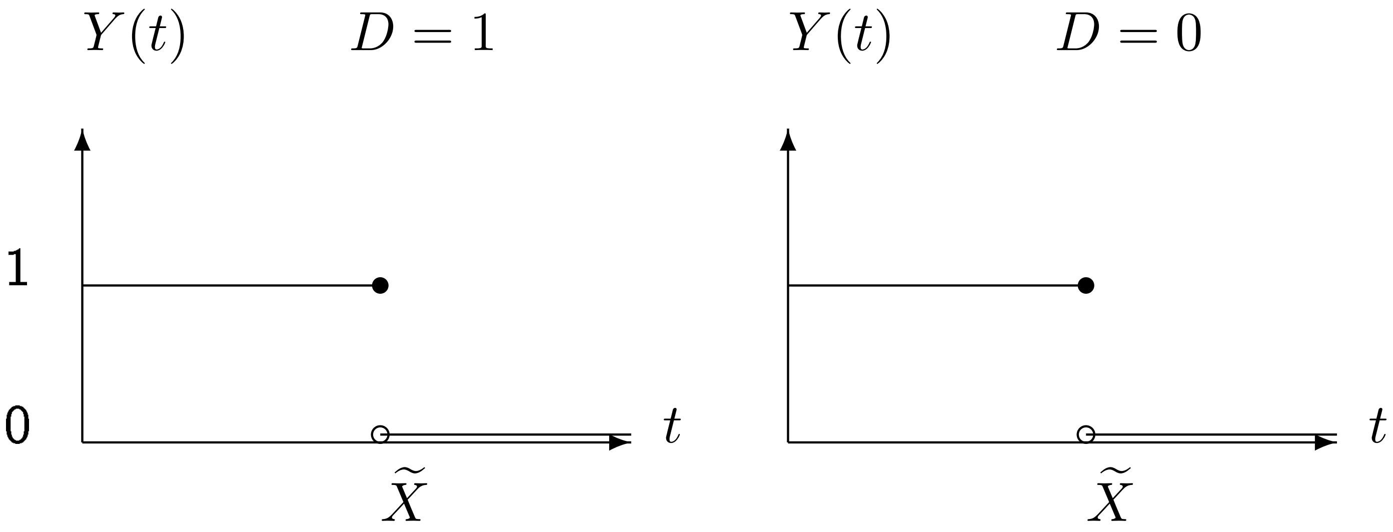
The assumption of independent censoring (by some denoted non-informative censoring). means that individuals censored at any given time t should not be a biased sample of those who are at risk at time t. Stated in other words: the hazard $\alpha(t)$ gives the event rate at time t, i.e. the failure rate given that the subject is still alive $(X \geq t)$. Independent censoring then means that the extra information that the subject is not only alive, but also uncensored at time t does not change the failure rate.

Typically, independent censoring cannot be tested from the available data - it is a matter of discussion. Censoring caused by being alive at the end of study can usually safely be taken to be 'independent'. However, one should be more suspicious to other kinds of loss to follow-up before end of study. It is strongly advisable always to keep track of subjects who are lost to follow-up and to note the reasons for loss to follow-up (e.g., drop-out of follow-up schedule or emigration).

The counting process $N(t)$ for $D = 0$ or 1



The process $Y(t)$ for $D = 0$ or 1



Right censoring

That is:

$$\lambda(t) = \alpha(t) Y(t), \text{where} \;\; Y(t) = I(\widetilde{X} \geq t).$$

Same structure as without censoring but:

- $Y(\cdot)$ is different
- $(\mathcal{F}_t)$ is different

$M(t) = N(t) - \Lambda(t)$ is still a martingale.

The Nelson-Aalen estimator

A sample of n right-censored survival times:

$$(\widetilde{X}_1, D_1), \ldots, (\widetilde{X}_n, D_n).$$

Model: the *un*censored life times $X_1, \ldots, X_n$ are *i.i.d.* with hazard $\alpha(\cdot)$. Censoring is assumed to be *independent*.

Define:

$$N_i(t) = I(\widetilde{X}_i \leq t, D_i = 1), \quad Y_i(t) = I(\widetilde{X}_i \geq t).$$

Then $\mathbf{N} = (N_1, \ldots, N_n)$ is a *multivariate counting process* (no two components jump simultaneously),

and $N_. = \sum_{i=1}^{n} N_i$ is a *univariate counting process* counting the number of *observed* deaths.

The Nelson-Aalen estimator

Intensity process for $N_{\cdot}$ is:

$$\lambda_{\cdot}(t) = \sum_{i=1}^{n} \lambda_i(t) = \sum_{i=1}^{n} \alpha(t) Y_i(t) = \alpha(t) Y_{\cdot}(t)$$

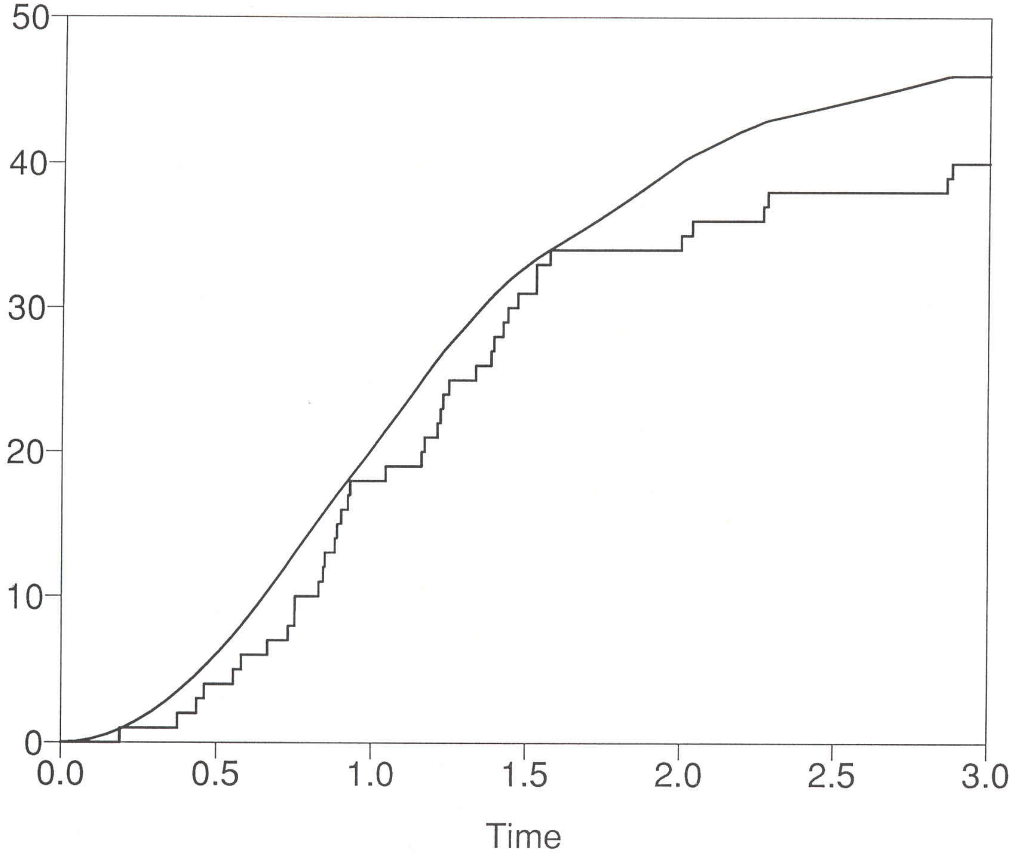
where $Y_{\cdot}(t) = \sum_i Y_i(t)$ is the number of individuals *observed to be at risk* just before time t.

This is an example of *Aalen's multiplicative intensity model*: $\lambda(t) = \alpha(t) Y(t)$ where

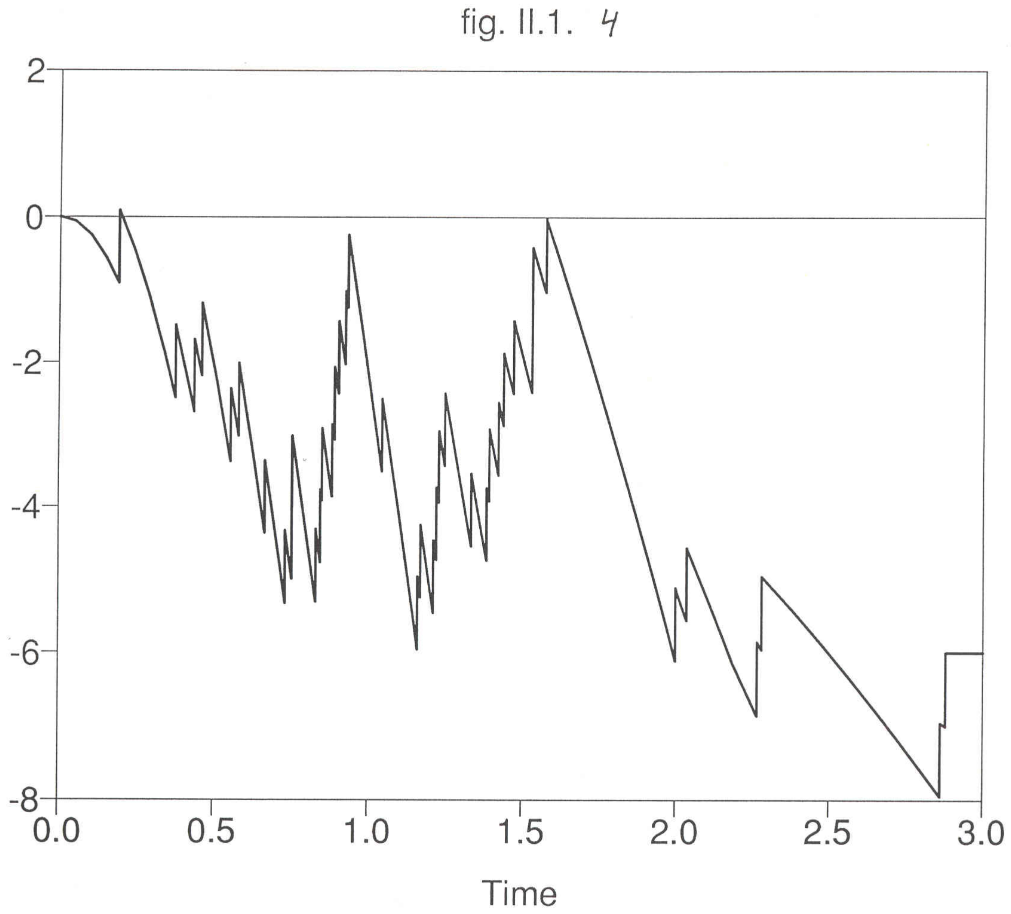
- $\alpha(\cdot)$ is deterministic
- $Y(\cdot)$ is predictable

$M_{\cdot}(t) = N_{\cdot}(t) - \Lambda_{\cdot}(t)$ is a martingale, see simulation illustration with $\alpha(t) = t$ and $n = 50$ (+ censoring).

fig. II.1. 3



Counting process and compensator



Counting process minus compensator (martingale)

The Nelson-Aalen estimator

Why are we happy because $M_.$ is a martingale?

- it provides us with some natural estimating equations for $\alpha(\cdot)$:

  $N_.(t) = \int_0^t \alpha(u) Y_.(u) du + M_.(t)$

  $dN_.(t) = \alpha(t) Y_.(t) dt + dM_.(t)$

  Now $dM_.(t)$ is "noise" so the second equation gives:

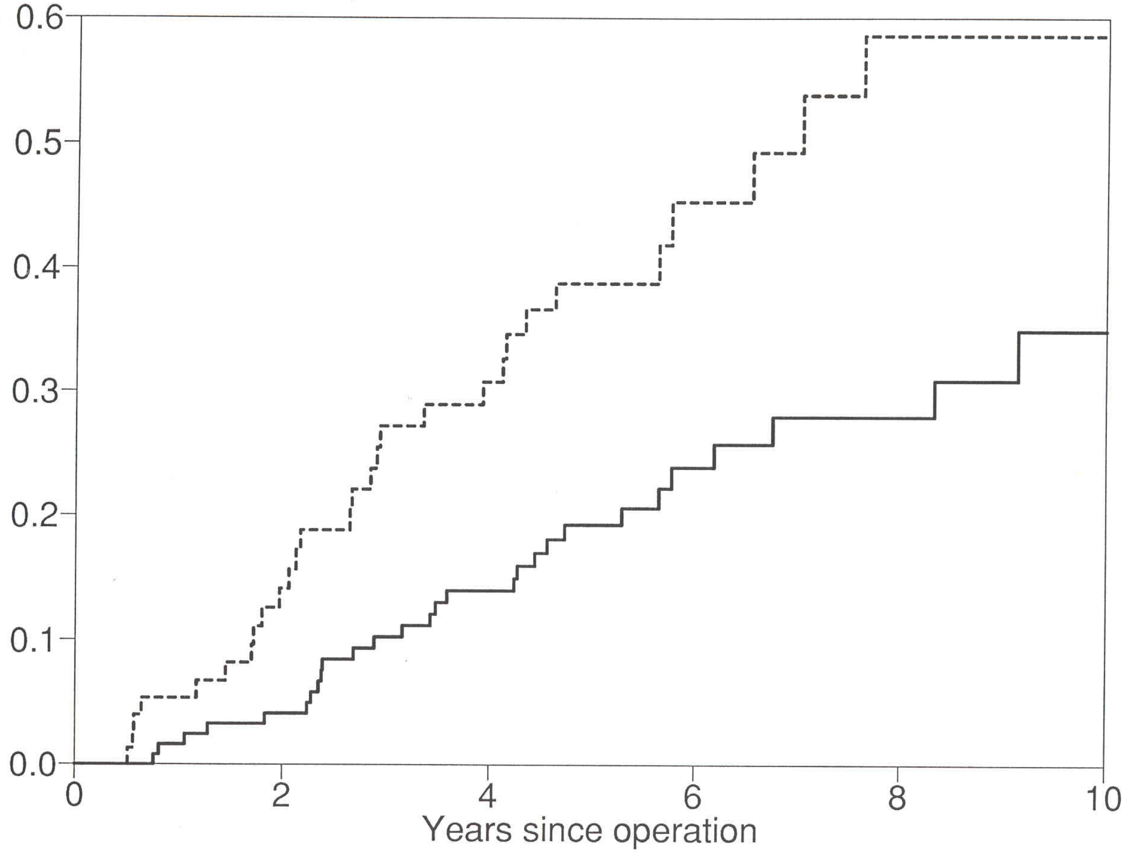
  $\widehat{\alpha}(t)dt = \frac{dN_.(t)}{Y_.(t)}$.

  Estimate $A(t) = \int_0^t \alpha(u) du$ by the *Nelson-Aalen estimator*

  $$\widehat{A}(t) = \int_0^t \frac{dN_.(u)}{Y_.(u)}.$$

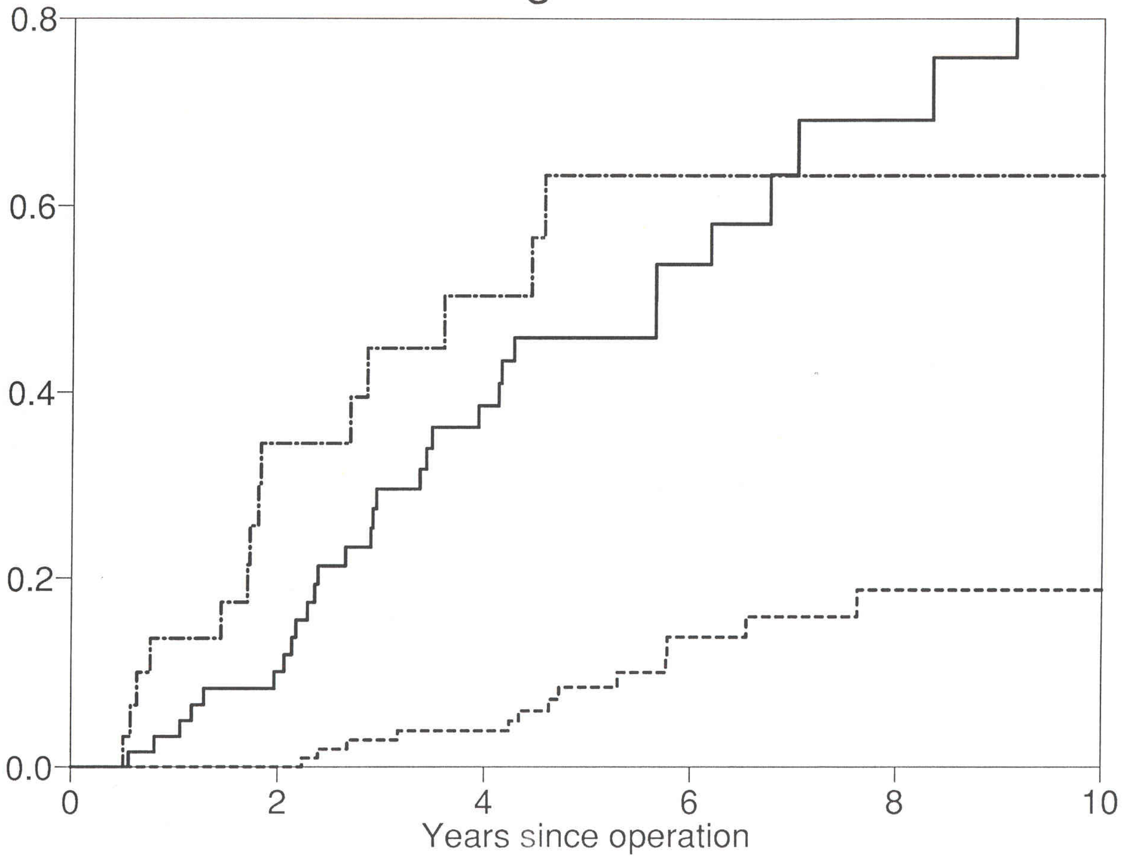
- it provides us with a way of deriving (e.g., large sample) properties of the estimator (more below)

FIG.IV.1.1



Malignant melanoma: females (solid) and males (dashed)

Fig. V.3.1



Malignant melanoma: tumor thickness 0-2, 2-5, 5+ *mm*

Fyn diabetics (Ex. I.3.2)

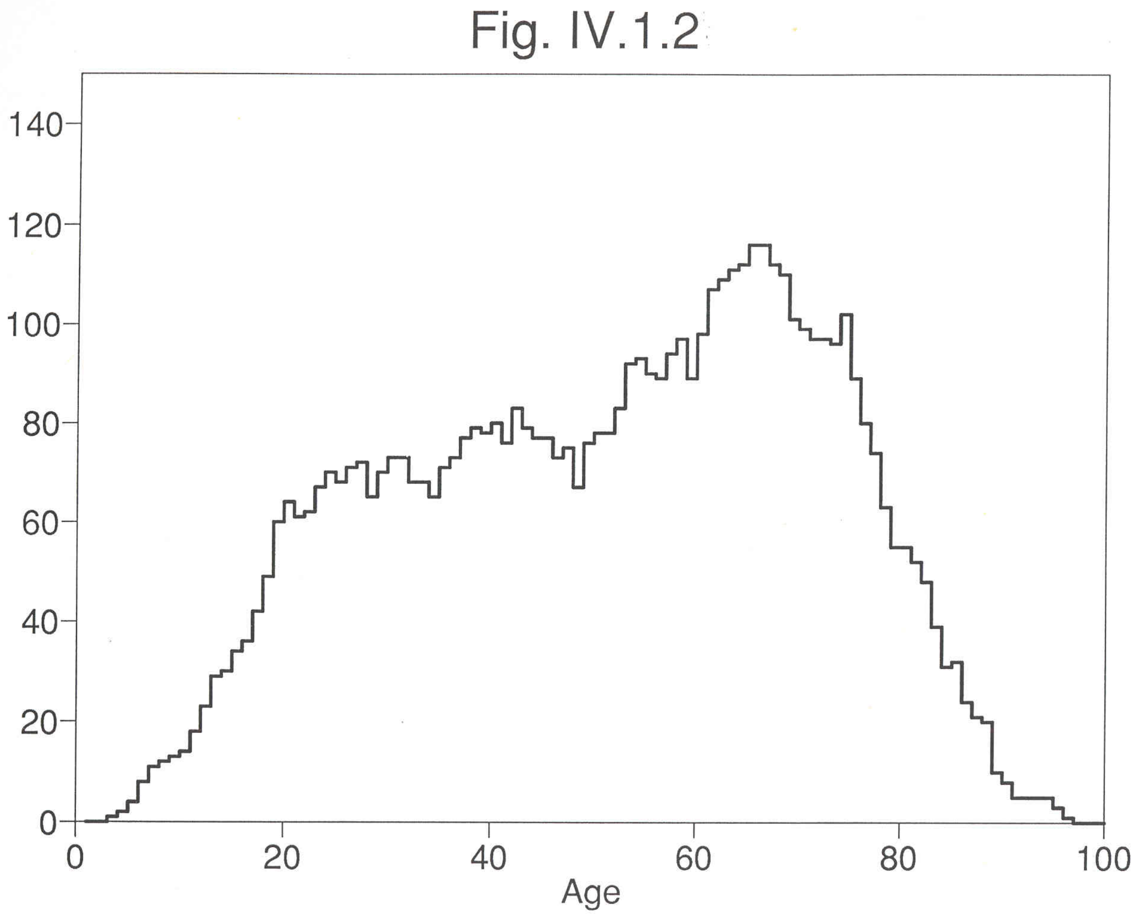
In this example there is *left-truncation*. Let $\widetilde{X}_i$ be age at death/censoring, V_i age at 1 July 1973. Then we define the counting processes $N_i(t)$ and the indicators $Y_i(t)$ (where $t =$age):

$$N_i(t) = I(V_i < \widetilde{X}_i \leq t, D_i = 1), \quad Y_i(t) = I(V_i < t \leq \widetilde{X}_i)$$

and (under an assumption of ‘independent truncation’) we still have Aalen’s multiplicative intensity model:

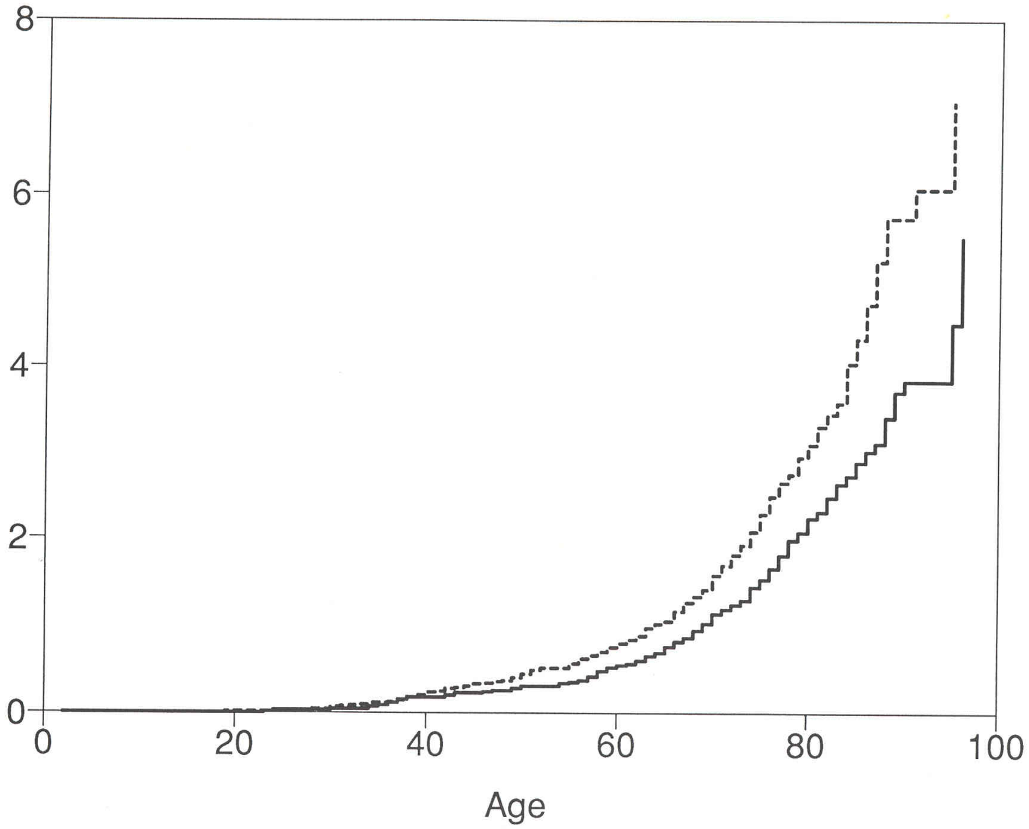
$$N.(t) = \sum_i N_i(t) = \sum_i (\int_0^t \alpha(u) Y_i(u) du + M_i(t))$$

$$= \int_0^t \alpha(u) Y.(u) du + M.(t).$$



Fyn diabetics: number of females at risk by age

Fig.IV.1. 3



Fyn diabetics: females (solid) and males (dashed)

Properties of the Nelson-Aalen estimator

$$dN_{\cdot}(t) = \alpha(t)Y_{\cdot}(t)dt + dM_{\cdot}(t)$$

$$\frac{dN_{\cdot}(t)}{Y_{\cdot}(t)} = I(Y_{\cdot}(t) > 0)\alpha(t)dt + \frac{I(Y_{\cdot}(t) > 0)}{Y_{\cdot}(t)} dM_{\cdot}(t)$$

Here, $H(t) = \frac{I(Y_{\cdot}(t)>0)}{Y_{\cdot}(t)}$ is *predictable* and, therefore,

$$E(H(t)dM_{\cdot}(t) \mid \mathcal{F}_{t-}) = H(t)E(dM_{\cdot}(t) \mid \mathcal{F}_{t-}) = 0$$

showing that the *stochastic integral*

$$M'(t) = \int_0^t H(u)dM_{\cdot}(u)$$

is a martingale.

Properties of the Nelson-Aalen estimator

Thus:

$$\widehat{A}(t) - \int_0^t I(Y_\cdot(u) > 0)\alpha(u)du = \int_0^t \frac{I(Y_\cdot(u) > 0)}{Y_\cdot(u)} dM_\cdot(u)$$

is a martingale and we have *approximate unbiasedness*:

$$E\widehat{A}(t) = E\int_0^t I(Y_\cdot(u) > 0)\alpha(u)du$$

$$= EA^*(t) = \int_0^t P(Y_\cdot(u) > 0)\alpha(u)du$$

$$\leq A(t).$$

Estimating the variance

In general: $\mathrm{var}(M(t)) = E(M(t))^2$ since the mean is zero.

$$E(M(t))^2 = E\int_0^t d(M^2(u)) = \int_0^t Ed(M^2(u))$$

$$= \int_0^t E(E(d(M^2(u)) \mid \mathcal{F}_{u-})) =$$

(since $M(u) = M(u-) + dM(u)$)

$$\int_0^t E(E((dM(u))^2 \mid \mathcal{F}_{u-}) + E(2M(u-)dM(u) \mid \mathcal{F}_{u-}))$$

$$= E\int_0^t \mathrm{var}(dM(u) \mid \mathcal{F}_{u-}).$$

Predictable variation process

We need to be able to calculate $\mathrm{var}(dM(u) \mid \mathcal{F}_{u-})$. This is

$$= \mathrm{var}(dN(u) - \lambda(u)du \mid \mathcal{F}_{u-}) = \mathrm{var}(dN(u) \mid \mathcal{F}_{u-})$$

$$= \lambda(u)du(1 - \lambda(u)du) \doteq \lambda(u)du.$$

The process with increments $\mathrm{var}(dM(u) \mid \mathcal{F}_{u-})$ is denoted $\langle M \rangle(u)$: *predictable variation process* for M (it is the compensator of M^2).

Compare with a Poisson process $N(t)$ for which $E(N(t)) = \mathrm{var}(N(t))$.

Here, $E(dN(t) \mid \mathcal{F}_{t-}) = \mathrm{var}(dN(t) \mid \mathcal{F}_{t-})$ - the counting process behaves locally as a Poisson process.

Stochastic integrals

For a stochastic integral:

$$
\begin{aligned}
M'(t) &= \int_0^t H(u)dM(u): \\
d\langle M'\rangle(t) &= \mathrm{var}(dM'(t) \mid \mathcal{F}_{t-}) \\
&= \mathrm{var}(H(t)dM(t) \mid \mathcal{F}_{t-}) \\
&= H^2(t)\mathrm{var}(dM(t) \mid \mathcal{F}_{t-}) \\
&= H^2(t)d\langle M\rangle(t),
\end{aligned}
$$

i.e.,

$$\langle M'\rangle(t) = \int_0^t H^2(u)d\langle M\rangle(u).$$

Nelson-Aalen estimator

For the Nelson-Aalen estimator:

$$(\widehat{A} - A^*)(t) = \int_0^t \frac{I(Y_.(u) > 0)}{Y_.(u)} dM_.(u),$$

i.e.,

$$\langle \widehat{A} - A^* \rangle(t) = \int_0^t \frac{I(Y_.(u) > 0)}{(Y_.(u))^2} d\langle M_. \rangle(u)$$

Since

$$d\langle M_. \rangle(u) = \lambda_.(u)du = \alpha(u)Y_.(u)du$$

we can estimate $\mathrm{var}(\widehat{A}(t))$ by

$$\widehat{\sigma}^2(t) = \int_0^t \frac{dN_.(u)}{(Y_.(u))^2}.$$

Large sample properties: consistency

Use *Lenglart's inequality*:

$$P\left(\sup_{[0,\tau]} \mid M \mid > \eta\right) \leq \frac{\delta}{\eta^2} + P(\langle M\rangle(\tau) > \delta)$$

on the martingale $M' = \widehat{A} - A^*$ to get

$$P\left(\sup_{[0,t]} \mid \widehat{A}(s) - A^*(s) \mid > \eta\right) \leq \frac{\delta}{\eta^2} + P(\int_0^t \frac{I(Y_.(s) > 0)}{Y_.(s)} \alpha(s) ds > \delta).$$

We need $\inf_{[0,t]} Y_.(s) \rightarrow_P \infty$ as $n \rightarrow \infty$.

Large sample properties

What does $\sqrt{n}(\widehat{A} - A^*)(t)$ look like for large n?

When n is large, $\frac{1}{n}Y_.(t) \simeq y(t)$ (law of large numbers), i.e.,

(1) "Lindeberg" - the jumps of $\sqrt{n}(\widehat{A} - A^*)$:
$d(\sqrt{n}(\widehat{A} - A^*)) = \sqrt{n}d\widehat{A} = \sqrt{n}\frac{dN_.}{Y_.} \simeq \frac{1}{\sqrt{n}}\frac{1}{y}$ "get small" and

(2) the conditional variances:
$\mathrm{var}(d(\sqrt{n}(\widehat{A} - A^*)(t)) \mid \mathcal{F}_{t-}) = \frac{n}{(Y_.(t))^2}\alpha(t)Y_.(t)dt \simeq \frac{\alpha(t)}{y(t)}dt$
"become deterministic".

(1) and **(2)** are sufficient conditions for $\sqrt{n}(\widehat{A} - A^*)$ to look asymptotically like a Gaussian process with variance function $\int \frac{\alpha}{y}$.

Martingale central limit theorem

(A continuous process with independent increments is Gaussian.)

Confidence interval for integrated hazard

Simple “linear”:

$$\widehat{A}(t) \pm c_{\alpha/2}\widehat{\sigma}(t) \quad (c_\alpha \text{ Normal quantile}).$$

Confidence interval for $g(A(t))$ is according to the $\delta-$method:

$$g(\widehat{A}(t)) \pm c_{\alpha/2} \mid g'(\widehat{A}(t)) \mid \widehat{\sigma}(t)$$

“transform back” by g^{-1}, e.g. $g = \log$.

Such transformations may improve the small sample behaviour of the confidence limits considerably.

Simultaneous confidence *bands* may also be derived (use transformation of limiting process to a *Brownian bridge*).

Survival distributions, cumulative hazards, product-integrals

(ABGK, p.57, sect. II.6)

- Uncensored survival time: X
- Survival function: $S(t) = P(X > t)$
- Hazard function: $\alpha(t) = \lim_{\Delta t \to 0} \frac{1}{\Delta t} P(X < t + \Delta t \mid X > t)$
- Cumulative hazard function: $A(t) = \int_0^t \alpha(u) \mathrm{d}u$

Note that the Nelson-Aalen estimator may be used beyond the simple survival data situation, e.g. competing risks and other multi-state models.

This is in contrast to the Kaplan-Meier estimator (to be defined now).

Relations between survival function and hazard

- $\alpha(t) = A'(t) = -\frac{S'(t)}{S(t)} = -\frac{\mathrm{d}}{\mathrm{d}t} \log S(t)$
- $S(t) = \exp\left(-\int_0^t \alpha(u)\mathrm{d}u\right) = \exp(-A(t))$

Here it is assumed that the distribution is *absolutely continuous*.

For a general distribution we have

$$A(t) = -\int_0^t \frac{\mathrm{d}S(u)}{S(u-)}.$$

Product-integrals

The product-integral is needed to describe the relation between a ‘cumulative hazard’ and the survival function for general distributions.

Partition the interval $[0, t]$ into small intervals: $[s_{i-1}, s_i), i = 1, 2...$

Then

$$S(t) = \lim_{\max|s_i - s_{i-1}| \to 0} \prod (1 - (A(s_i) - A(s_{i-1})))$$

and define this to be the *product-integral*

$$= \prod_{0 \leq s \leq t} (1 - \mathrm{d}A(s)).$$

Product-integrals

For the continuous case we have

$$\mathop{\pi}_{0\leq s\leq t}(1-\mathrm{d}A(s))=\exp(-A(t)).$$

For the discrete case we have

$$\mathop{\pi}_{0\leq s\leq t}(1-\mathrm{d}A(s))=\prod_{0\leq s\leq t}(1-\Delta A(s))$$

where $\Delta A(s)=P(X=s\mid X\geq s)$ is the increment of the cumulative hazard at s.

For the general case we have a mixture of the two.

The product-integral as a *mapping* is (compactly) differentiable.

The Kaplan-Meier estimator

ABGK sect. IV.3.1.

Censored survival times: $(\widetilde{X}_i, D_i), i = 1, \ldots, n$.

Model: the *uncensored* survival times are *i.i.d.* with hazard $\alpha(t)$. Censoring is assumed to be independent.

Counting process for i: $N_i(t) = I(\widetilde{X}_i \leq t, D_i = 1)$,

Intensity process: $\lambda_i(t) = \alpha(t) I(\widetilde{X}_i \geq t) = \alpha(t) Y_i(t)$.

Aggregated counting process: $N.(t) = \sum_{i=1}^{n} N_i(t)$

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

with $Y.(t) = \sum_{i=1}^{n} Y_i(t)$, the number at risk just before time t.

The Kaplan-Meier estimator

Recall the Nelson-Aalen estimator:

$$\widehat{A}(t) = \int_0^t \frac{\mathrm{d}N.(u)}{Y.(u)}.$$

Plug this into the product-integral expression for the survival function to get the finite product:

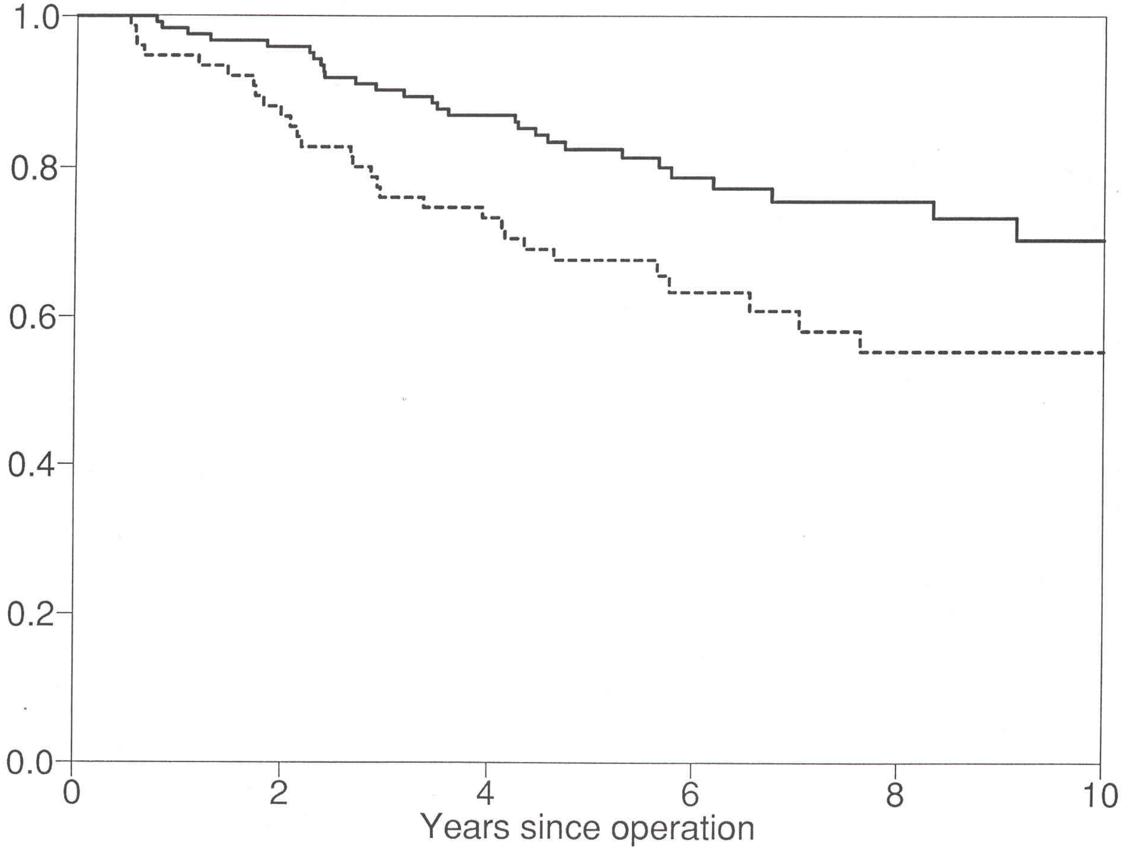
$$\widehat{S}(t) = \underset{0 \leq s \leq t}{\pi} (1 - \mathrm{d}\widehat{A}(s)) = \prod_{0 \leq s \leq t} (1 - \frac{dN.(s)}{Y.(s)}).$$

This is the Kaplan-Meier estimator.

The alternative estimator $\exp(-\widehat{A}(t))$ available in many programs is a strange mixture where continuous-time results are applied to a discrete estimator.

Survival curves for male and female melanoma patients

FIG.I.3.2



Males: dashed, females: solid (we still treat death from other causes as censoring).

Properties of Kaplan-Meier estimator

Recall $A^*(t) = \int_0^t I(Y.(s) > 0)\alpha(s)\mathrm{d}s \approx A(t)$

and introduce its product-integral
$S^*(t) = \prod_{0\leq s\leq t}(1 - \mathrm{d}A^*(s)) \approx S(t)$.

One may show that ("Duhamel's equation"):

$$\frac{\widehat{S}(t)}{S^*(t)} - 1 = \int_0^t \frac{\widehat{S}(s-)}{S^*(s)}\mathrm{d}(\widehat{A} - A^*)(s) \approx -(\widehat{A}(t) - A(t)).$$

Since $\frac{\widehat{S}(s-)}{S^*(s)} \approx 1$ we have for large n that:
$\sqrt{n}(\frac{\widehat{S}(t)}{S(t)} - 1) \approx -\sqrt{n}(\widehat{A}(t) - A(t))$ and thereby, asymptotically

$$\widehat{S}(t) - S(t) \approx -S(t)(\widehat{A}(t) - A(t)).$$

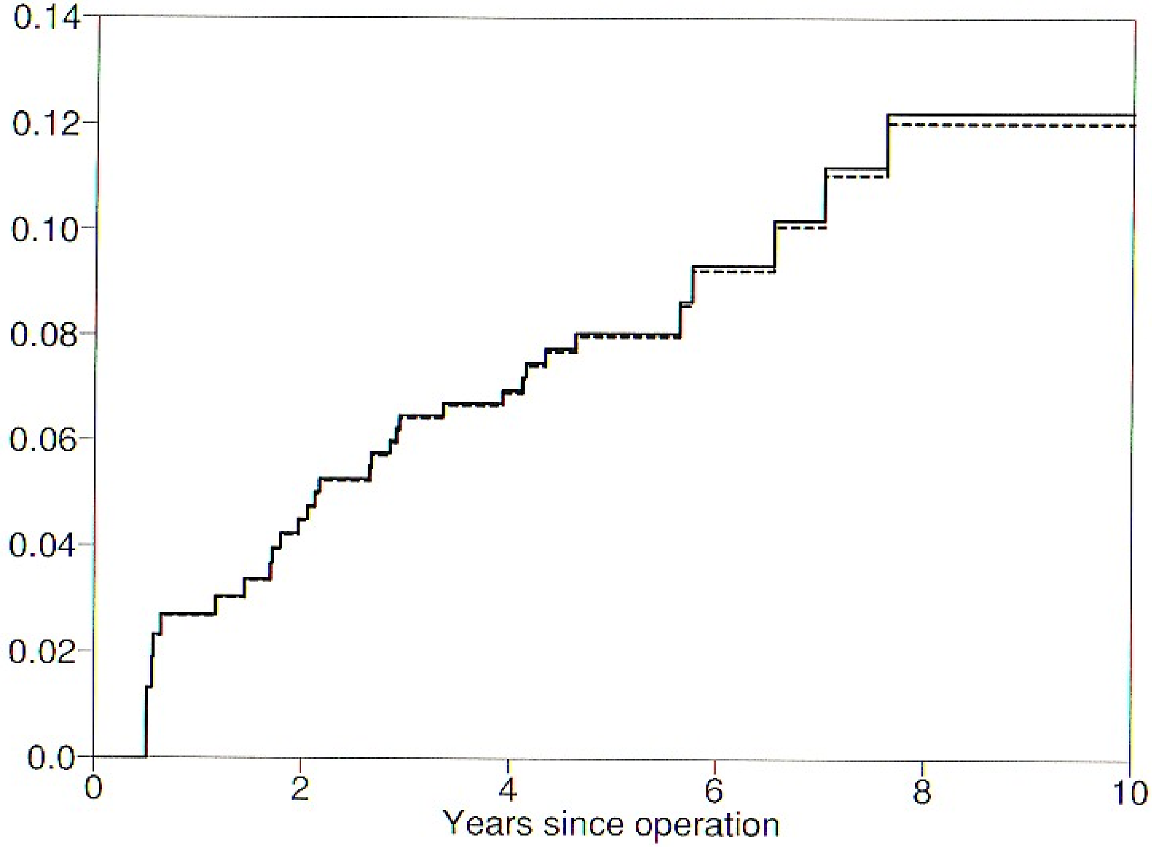
Properties of Kaplan-Meier estimator

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

- $\mathrm{var}(\widehat{S}(t)) \approx (S(t))^2 \mathrm{var}(\widehat{A}(t))$
- $\widehat{\mathrm{var}}(\widehat{S}(t)) = (\widehat{S}(t))^2 \widehat{\sigma}^2(t)$
- with $\widehat{\sigma}^2(t) = \int_0^t \frac{\mathrm{d}N.(s)}{(Y.(s))^2}$
- Alternatively: Greenwood's formula, replace $\widehat{\sigma}^2(t)$ by $\widetilde{\sigma}^2(t) = \int_0^t \frac{\mathrm{d}N.(s)}{Y.(s)(Y.(s)-\Delta N.(s))}$.
- $\widehat{S}(t)$ is asymptotically normally distributed around $S(t)$

Variance estimates, male melanoma patients

FIG.IV.3.1



“Greenwood” (larger) and “Aalen” estimator of standard deviation.

Confidence intervals and bands

Pointwise confidence intervals for $S(t)$:

Linear:

$$\widehat{S}(t) \pm c_{\alpha/2} \widehat{S}(t) \widehat{\sigma}(t).$$

Log-log transformed:

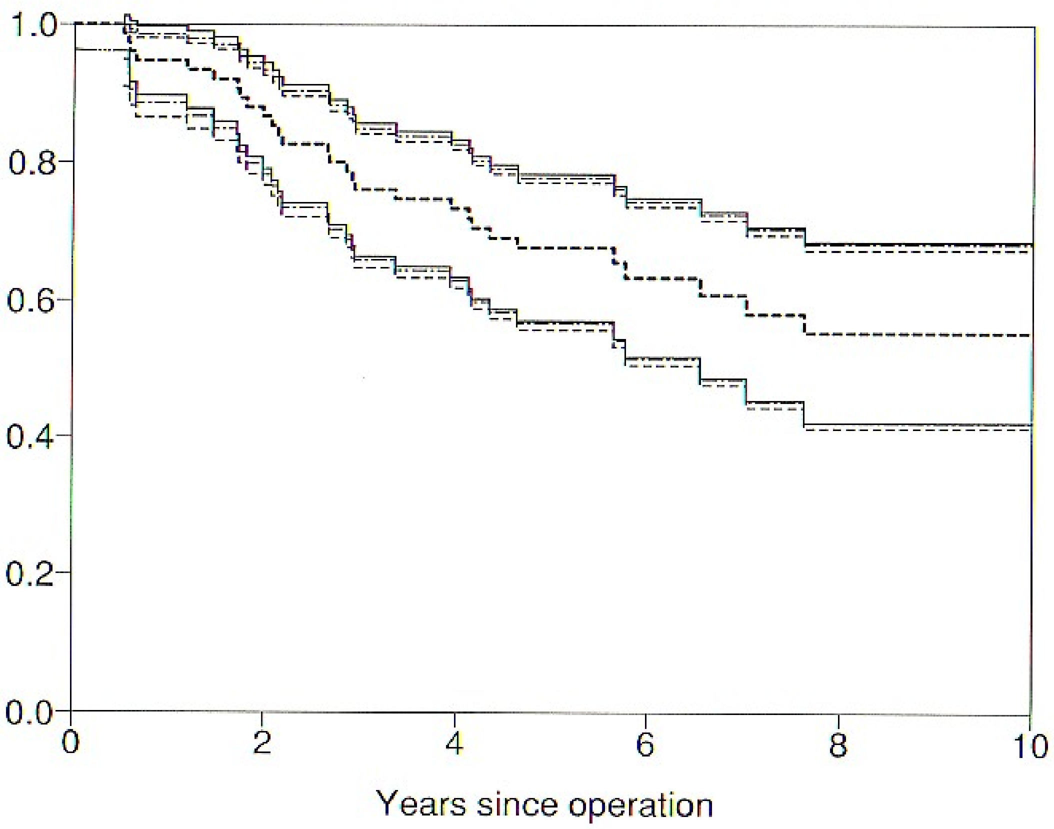
$$\widehat{S}(t)^{\exp(\pm c_{\alpha/2} \widehat{\sigma}(t) / \log \widehat{S}(t))}$$

Log-transformed: R's default, but not optimal choice!

Confidence bands: as for Nelson-Aalen using properties of the Brownian bridge.

Survival curve with c.i., male melanoma patients

Figure IV.3.2



Quantiles

The pth quantile, ξ_p, of the survival distribution is given by:

$$p = P(X \leq \xi_p) = 1 - S(\xi_p)$$

and may be estimated by

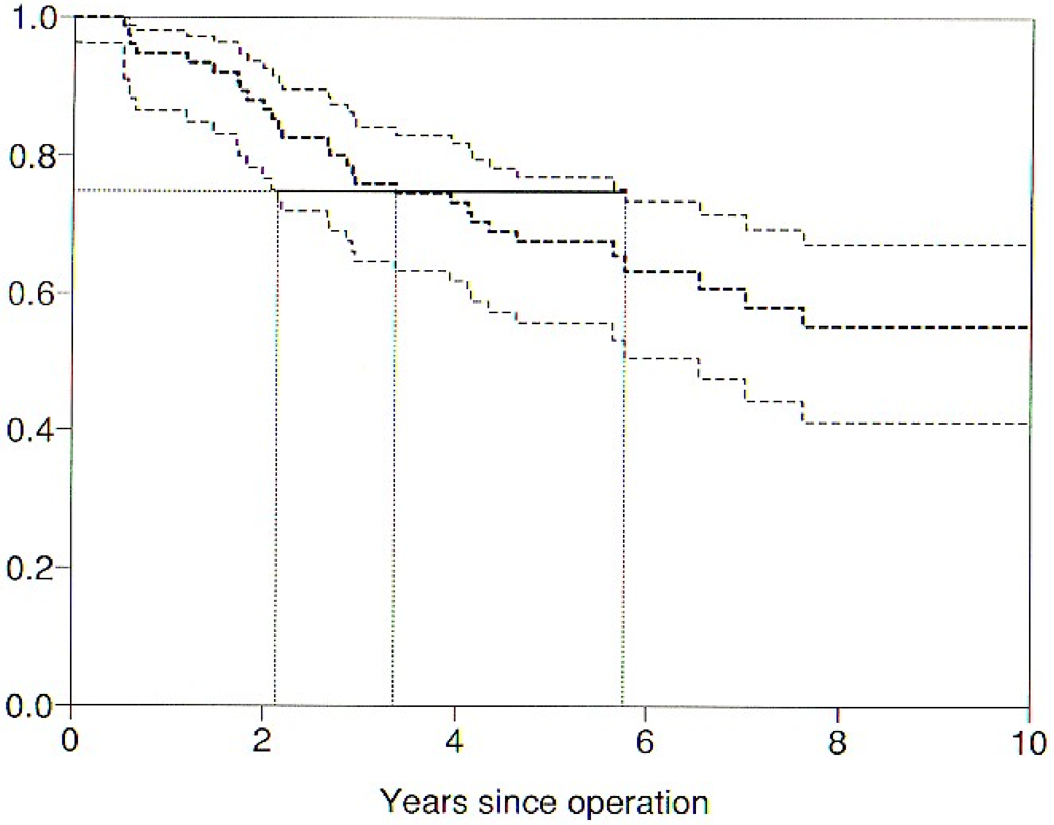
$$\widehat{\xi}_p = \inf\left(t : \widehat{S}(t) \leq 1 - p\right).$$

Both this estimator and confidence limits may be obtained from a Kaplan-Meier plot with pointwise confidence limits.

Example: male melanoma patients, lower quartile, $\widehat{\xi}_{0.25} = 3.36$ with 95% c.i. from 2.13 to 5.76 (years).

Estimation of lower quartile, male melanoma patients

Figure IV.3. 10



Mean survival time

The mean survival time is

$$\mu = E(X) = E\int_0^X 1\mathrm{d}t$$

$$= E\int_0^\infty I(X > t)\mathrm{d}t = \int_0^\infty S(t)\mathrm{d}t$$

and may be estimated by the Kaplan-Meier integral (up to the largest observation time).

Because the tail of the distribution is ill-determined (right-censoring), some times, the *restricted* mean is studied. This is

$$\mu_\tau = E(X \wedge \tau) = \int_0^\tau S(t)\mathrm{d}t$$

and is easily estimated from the Kaplan-Meier estimator.

Left-truncation

- All these results also hold when delayed entry (left-truncation) is present
- However, for small t estimates may get unstable due to few individuals at risk
- This is serious for Kaplan-Meier - a *global* quantity
- The problem is less serious for Nelson-Aalen where the slope estimates the hazard - a *local* quantity.

Left-truncation

One may get a meaningful estimate of the *conditional* survival function given $X > t_0$:

$$S(t \mid t_0) = P(X > t \mid X > t_0) = S(t)/S(t_0)$$

for a suitable t_0 with a reasonable number at risk between t_0 and t.

Estimator:

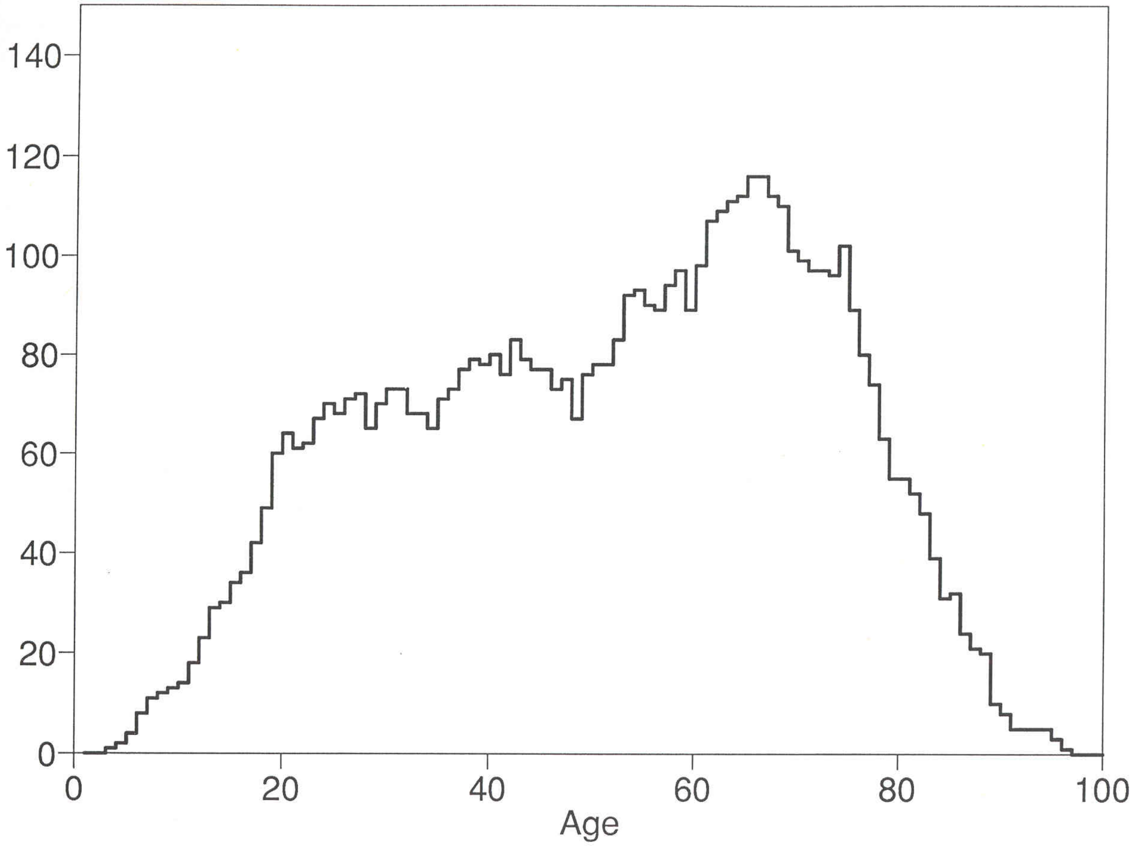
$$\widehat{S}(t \mid t_0) = \widehat{S}(t)/\widehat{S}(t_0) = \prod_{t_0 < s \leq t} \left(1 - \frac{dN.(s)}{Y.(s)}\right).$$

Properties as for ordinary Kaplan-Meier.

Example: female diabetics from Fyn, survival functions truncated at 20, 30, 40 years.

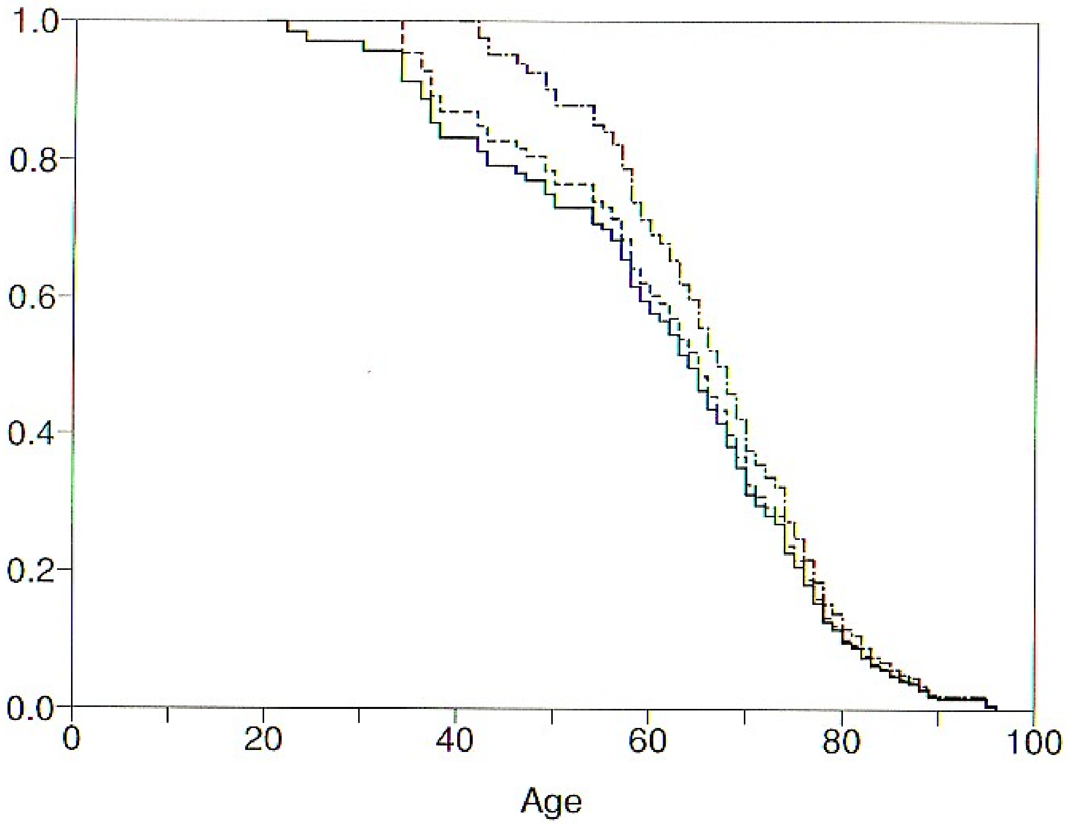
Risk set by age, female diabetics

Fig. IV.1.2



Conditional survival curves, female diabetics

FIG IV.3.7



Survival analysis in R

To use the facilities built into R for doing survival analysis we must first get hold of the proper functions using the library(survival) command.

Then we can create and work with Survival-objects. These consist of "the usual pairs" (time,status) so, assume that t dc represent the survival time and the failure indicator, dc=1 meaning failure.

Then we may write

```
survdata<-Surv(t,dc==1)
```

Note the logical expression dc==1. If delayed entry is present and in represents the entry time the syntax is

```
survdata<-Surv(in,t,dc==1)
```

Survival analysis in R

We shall be working with R -functions which are able to operate on Survival objects:

```
survfit
```

```
survdiff
```

These functions compute Kaplan-Meier/Nelson-Aalen estimators, logrank and similar tests, respectively.

The command: `survfit(survdata~x)`

gives a table of estimated mean and median survival times etc. *in subgroups given by* x. If we want to see the estimated Kaplan-Meier survival curves then we may use the plot function:

```
plot(survfit(survdata~x))
```

and the plot appears in a separate window.

Survival analysis in `R`

The `plot` function allows a great number of options controlling axes, labels, line types etc., see what comes out of the `help` commands:

`?plot` or `?plot.survfit`

For the `plot(survfit(...))` there are special options for controlling confidence limits, and the *Nelson-Aalen* estimator may be obtained using the

`,type=’’fleming-harrington’’), fun=‘‘cumhaz’’)` options.

The `survdiff` function:

```
survdiff(survdata~x)
```

gives the logrank test (or other non-parametric tests - to be discussed later) for comparing the survival distributions (or hazard functions) among subgroups of `x`.

The malignant melanoma study

The data file `melanom.txt` contains data from the study on prognostic factors in malignant melanoma, cf. ABGK Ex. I.3.1. There are 205 records and 9 variables (variable names in first line):

- `dc` death/cens. indicator 1 = death from mal. mel., 2 = alive on 31DEC77, 3 = death from other causes
- `days` time in days from operation
- `level` level of invasion, 0, 1 or 2
- `ici` inflammatory cell infiltration (ICI),0, 1, 2, or 3
- `ecel` presence of epithelioid cells, 1=no, 2=yes
- `ulc` presence of ulceration, 1=yes, 2=no
- `thick` tumour thickness (in 1/100 mm)
- `sex` 1=F, 2=M
- age at operation (years)

The Fyn diabetes data.

The data file `fyn.txt` contains data from the Fyn county diabetes study, cf. ABGK Ex. I.3.2. There are 1499 records and 9 variables (variable names in first line):

- `id` patient id.
- `sex` M=1, F=0
- `fail` indicator of death (1), alive 1JAN82 (0), emigration (2)
- `inage` age in years 1JUL73
- `exage` age in years at exit
- `debage` age in years at disease onset
- `exdat` date (MM/DD/YYYY) of exit
- `bthdat` date (MM/DD/YYYY) of birth
- `debdat` date (MM/DD/YYYY) of disease onset

Exercise: The malignant melanoma study

The data set `melanom.txt` contains the melanoma data (Ex. I.3.1).

1. Read the data into `R`.
2. Plot the Kaplan-Meier and Nelson-Aalen curves for men and women *for all-cause mortality*. Confidence limits may be added to the plot - see documentation for `survfit`
3. Compare the curves using the logrank test and other non-parametric tests.
4. Compute a new variable from the tumour thickness by grouping at the cutpoints *2 mm* and *5 mm*. Repeat 2.-3. for this variable.

Exercise: The Fyn diabetes data.

The file `fyn.txt` contains the data from the study of diabetics in Fyn found in ABGK, Example I.3.2 (p. 14 ff.).

1. Read the data and estimate the cumulative age-specific hazards for men and women.
2. Same question for the disease duration-specific hazards.
3. Compare, for both time variables, the hazards for men and women using the logrank test.