PhD course in “Advanced survival analysis”

Introduction. The Nelson-Aalen and Kaplan-Meier estimators

1 October 2012


Per Kragh Andersen

Empirical phenomena modeled and analysed:

• DISCRETE (POINT) EVENTS occurring in (CONTINUOUS) TIME

• Main example - SURVIVAL DATA:
  – Ex. I.3.1: Malignant melanoma
  – Ex. I.3.2: Insulin-dependent diabetes
  – Ex. I.3.4: Liver cirrhosis

• 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 (approx.)

Pat.no.
7
6
5
4
3
2
1

Calendar year
Females: solid, males: dashed (we cheat a bit and consider only deaths from the disease)

Insulin-dependent diabetes - Table I.3.1

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

A randomized trial in liver cirrhosis: CSL1.

First trial conducted by the Copenhagen Study group for Liver diseases.

- Patient accrual 1962-69: 488 patients
- Randomization to: prednisone (251)/placebo (237)
- Patients followed until death (142+150), drop-out or Sept. 74
- Main purpose: effect of prednisone on survival
- Secondary purposes:
  - effect of prognostic factors
  - treatment interactions
  - effect of treatment on development over time of risk factors

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)
- Difficulties:
  - RIGHT-CENSORING (as in Ex. I.3.1)
  - DELAYED ENTRY (LEFT-TRUNCATION) - individuals followed from age at 1 July 1973
Multi-state models, two-state model for survival data: Fig. I.3.1.

Several types of failure (competing risks): Fig. I.3.5.

An illness-death model (recurrent disease): Fig. I.3.6.
Summary

• Data: transitions between states in stochastic process
• Incomplete observation: right censoring, left truncation (or delayed entry)
• Other kinds of incomplete observation: left (interval) censoring, right truncation (Sec. III.4, HIV positive → AIDS)
• 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.

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

A drawback: the mathematics is somewhat heavy

Fig. II.1.2 (approx.)

<table>
<thead>
<tr>
<th>Pat.no.</th>
</tr>
</thead>
<tbody>
<tr>
<td>7</td>
</tr>
<tr>
<td>6</td>
</tr>
<tr>
<td>5</td>
</tr>
<tr>
<td>4</td>
</tr>
<tr>
<td>3</td>
</tr>
<tr>
<td>2</td>
</tr>
<tr>
<td>1</td>
</tr>
</tbody>
</table>

Follow-up time

0 1 2 3 4 5 6 7 8 9 10
Stochastic processes associated with 1 uncensored survival time

$X$ survival time with survival function:

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

**$\alpha(t)$ hazard function**, i.e.

$$\alpha(t) \Delta t \approx P(X < t + \Delta t \mid X \geq t).$$

Let

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

counting process, counting +1 at $X$.

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

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”:

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

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

This is:

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

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

where $Y(t) = I(X \geq t)$ is the indicator of the individual being “still alive” just before time $t$.

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

Let $dN(t) = N(t + \Delta t) - N(t-)$. Then $dN(t) = 1$ or 0 and $P(dN(t) = 1 \mid \mathcal{F}_{t-}) = \mathbb{E}(dN(t) \mid \mathcal{F}_{t-}) = \alpha(t)Y(t)dt = \lambda(t)dt$.

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

(NB: a stochastic process)

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

integrated intensity process or compensator for $N(\cdot)$.
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 \( (\mathcal{F}_t) \)):
\[
  E(M(t) \mid \mathcal{F}_s) = M(s), s < t.
\]

21

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 \( t \)”.

22

This is because
\[
  E(M(t) \mid \mathcal{F}_s) - M(s) \\
  = E(M(t) - M(s) \mid \mathcal{F}_s) \\
  = E\left( \int_{\{s,t\}} dM(u) \mid \mathcal{F}_s \right) \\
  = \int_{\{s,t\}} E(dM(u) \mid \mathcal{F}_s) \\
  = \int_{\{s,t\}} E(E(dM(u) \mid \mathcal{F}_u) \mid \mathcal{F}_s) \\
  = 0.
\]

23

Right censoring.

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

Let the observed data be: \( D = I(X \leq U), \tilde{X} = X \wedge U \).
Let \( N(t) = I(\tilde{X} \leq t, D = 1) \): counting process with intensity process given by:
\[
  \lambda(t)dt = P(dN(t) = 1 \mid \mathcal{F}_{t-}).
\]
This is:
\[
  = 0, \text{if } \tilde{X} < t \\
  = \alpha(t)dt, \text{if } \tilde{X} \geq t
\]
AND censoring is “independent” - a crucial assumption, Sect.III.2.2.
The counting process $N(t)$ for $D = 0$ or $1$

That is:
$$\lambda(t) = \alpha(t)Y(t), \text{where } Y(t) = I(\tilde{X} \geq t).$$

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

The Nelson-Aalen estimator.

A sample of $n$ right-censored survival times:
$$(\tilde{X}_1, D_1), \ldots, (\tilde{X}_n, D_n).$$

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

Define: $N_i(t) = I(\tilde{X}_i \leq t, D_i = 1), Y_i(t) = I(\tilde{X}_i \geq t)$.

Then $N = (N_1, \ldots, N_n)$ 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.
Intensity process for $N$ is:

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

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

where $Y(t)$ 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

$M(t) = N(t) - \Lambda(t)$ is a martingale, see simulation illustration with $\alpha(t) = t$ and $n = 50$ (+ censoring).

Why are we happy because $M_t$ is a martingale?

- it provides us with some natural estimating equations for $\alpha(\cdot)$
- it provides us with a way of deriving (large sample) properties of the estimator

$$N_t = \int_{0}^{t} \alpha(u)Y(u)du + M_t$$

$$dN_t = \alpha(t)Y(t)dt + dM_t$$

Now $dM_t$ is “noise” so the second equation gives:

$$\delta(\cdot)dt = \frac{dN(t)}{Y(t)}.$$

Estimate $A(t) = \int_{0}^{t} \alpha(u)du$ by the Nelson-Aalen estimator

$$\hat{A}(t) = \int_{0}^{t} \frac{dN(u)}{Y(u)}.$$
Fyn diabetics (Ex. 1.3.2)

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

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

and we still have Aalen’s multiplicative intensity model:

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

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

Fyn diabetics: number of females at risk by age
Properties of the Nelson-Aalen estimator.

\[
dN(t) = \alpha(t)Y(t)dt + dM(t)
\]

\[
\frac{dN(t)}{Y(t)} = I(Y(t) > 0)\alpha(t)dt + \frac{I(Y(t) > 0)Y(t)}{Y(t)}dM(t)
\]

Here, \(H(t) = \frac{I(Y(t) > 0)}{Y(t)}\) is predictable and

\[
E(H(t)dM(t) | \mathcal{F}_{t^-}) = H(t)E(dM(t) | \mathcal{F}_{t^-}) = 0
\]

showing that the stochastic integral

\[
M'(t) = \int_0^t H(u)dM(u)
\]

is a martingale.

Thus:

\[
\hat{A}(t) - \int_0^t I(Y(u) > 0)\alpha(u)du = \int_0^t \frac{I(Y(u) > 0)}{Y(u)}dM(u)
\]

is a martingale and we have approximate unbiasedness:

\[
E\hat{A}(t) = E \int_0^t I(Y(u) > 0)\alpha(u)du
\]

\[
= EA^*(t) = \int_0^t P(Y(u) > 0)\alpha(u)du
\]

\[
\leq A(t).
\]

What about the variance?

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

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

\[
= \int_0^t E(E(d(M^2(u)) | \mathcal{F}_{u^-})) =
\]

(since \(M(u) = M(u^-) + dM(u)\))

\[
\int_0^t E(E((dM(u))^2 | \mathcal{F}_{u^-}) + E(2M(u^-)dM(u) | \mathcal{F}_{u^-}))
\]

\[
= E \int_0^t \text{var}(dM(u) | \mathcal{F}_{u^-}).
\]
We need to be able to calculate \( \text{var}(dM(u) \mid F_{u-}) \). This is
\[
= \text{var}(dN(u) - \lambda(u)du \mid F_{u-}) = \text{var}(dN(u) \mid F_{u-})
= \lambda(u)du(1 - \lambda(u)du) = \lambda(u)du.
\]
The process with increments \( \text{var}(dM(u) \mid F_{u-}) \) is denoted \( \langle M \rangle(u) \): predictable variation process for \( M \) (it is the compensator of \( M^2 \)). Compare with the Poisson process.

For the Nelson-Aalen estimator:
\[
\hat{A}(t) = \int_0^t \frac{I(Y(u) > 0)}{Y(u)} M(u) du,
\]
i.e.,
\[
\langle \hat{A} - \hat{A}^* \rangle(t) = \int_0^t \frac{I(Y(u) > 0)}{(Y(u))^2} d\langle M \rangle(u)
\]
Since
\[
d\langle M \rangle(u) = \lambda(u)du = \alpha(u)Y(u)du
\]
we can estimate \( \text{var}(\hat{A}(t)) \) by
\[
\hat{\sigma}^2(t) = \int_0^t \frac{dN(u)}{(Y(u))^2}.
\]

**Stochastic integrals.**

For a stochastic integral:
\[
M'(t) = \int_0^t H(u)dM(u) ;
\]
\[
d\langle M' \rangle(t) = \text{var}(dM'(t) \mid F_{t-})
= \text{var}(H(t)dM(t) \mid F_{t-})
= H^2(t)\text{var}(dM(t) \mid F_{t-})
= H^2(t)d\langle M \rangle(t),
\]
i.e.,
\[
\langle M' \rangle(t) = \int_0^t H^2(u)d\langle M \rangle(u).
\]

**Large sample properties: consistency**

Use Lenglart’s inequality:
\[
P \left( \sup_{[0,\tau]} | M | > \eta \right) \leq \frac{\delta}{\eta^2} + P(\langle M \rangle(\tau) > \delta)
\]
on the martingale \( M' = \hat{A} - \hat{A}^* \) to get
\[
P \left( \sup_{[0,t]} | \hat{A}(s) - \hat{A}^*(s) | > \eta \right) \leq \frac{\delta}{\eta^2} + P\left( \int_0^t \frac{I(Y(s) > 0)}{Y(s)^2} \alpha(s)ds > \delta \right).
\]
We need \( \inf_{[0,t]} Y(s) \to P \infty \) as \( n \to \infty \).
Large sample properties.

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

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

1. “Lindeberg” - the jumps of $\sqrt{n}(\hat{A} - A^*)$:
   $$d(\sqrt{n}(\hat{A} - A^*)) = \sqrt{n}d\hat{A} = \sqrt{n}\frac{1}{\sqrt{n}} \approx \frac{1}{\sqrt{n}} \text{get smaller}$$

2. the conditional variances:
   $$\text{var}(d(\sqrt{n}(\hat{A} - A^*)(t)) \mid \mathcal{F}_t) = \frac{n}{(Y(t))^2} \alpha(t)Y(t)dt \approx \frac{\alpha(t)}{y(t)} dt$$

   become “more deterministic”.

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

Martingale central limit theorem

Confidence interval for integrated hazard.

Simple “linear”:

$$\hat{A}(t) \pm c_{\alpha/2} \hat{\sigma}(t).$$

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

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

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

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

Simultaneous confidence bands.

Use that $\sqrt{n}(\hat{A} - A^*) \approx W(\cdot)$, Gaussian process on $[0, \tau]$ with mean 0, covariance $\sigma^2$ and transform $W(\cdot)$ into a Brownian bridge $W^0(\cdot)$ on $[0, 1]$:

$$e.g. \ W^0(x) = W(x\tau)/\sqrt{\tau - xW(\tau)/\sqrt{\tau}}$$

when $\sigma^2(t) = t$ and use results of $\text{sup} | W^0(\cdot) |$.

Can be used in connection with transformations as above; again with considerable improvements.
The Nelson-Aalen estimator

Note that the Nelson-Aalen estimator may be used beyond the simple survival data situation (in contrast to the Kaplan-Meier estimator), e.g. competing risks and other multi-state models.

Survival distributions, cumulative hazards and product-integrals.

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

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

Relations between survival function and (cumulative) hazard.

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

Here it is assumed that the distribution is absolutely continuous.

For a general distribution we have

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

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

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

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

$$\pi_{0 \leq s \leq t} (1 - dA(s)).$$

For the continuous case we have

$$\pi_{0 \leq s \leq t} (1 - dA(s)) = \exp(-A(t)).$$

For the discrete case we have

$$\pi_{0 \leq s \leq t} (1 - dA(s)) = \prod_{0 \leq s \leq t} (1 - \Delta A(s))$$

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 $t$.

Recall the Nelson-Aalen estimator (a step function):

$$\widehat{A}(t) = \int_{0}^{t} \frac{dN(u)}{Y(u)}.$$

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

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

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.

Properties of Kaplan-Meier estimator.

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

- $\text{var}(\hat{S}(t)) \approx (S(t))^2 \text{var}(\hat{A}(t))$
- $\text{var}(\hat{S}(t)) = (\hat{S}(t))^2 \tilde{\sigma}^2(t)$
- with $\tilde{\sigma}^2(t) = \int_0^t \frac{\text{d} N(s)}{(Y(s))^2}$
- Alternatively: Greenwood’s formula, replace $\tilde{\sigma}^2(t)$ by $\tilde{\sigma}^2(t) = \int_0^t \frac{\text{d} N(s)}{Y(s) \sqrt{Y(s) - \text{d} N(s)}}$.
- $\hat{S}(t)$ is asymptotically normally distributed around $S(t)$

Variance estimates, male melanoma patients.

"Greenwood" (larger) and "Aalen"
Confidence intervals and bands.

Pointwise confidence intervals for $S(t)$:
Linear:
$$\hat{S}(t) \pm c_{\alpha/2} \hat{S}(t)\hat{\sigma}(t).$$

Log-log transformed:
$$\hat{S}(t)^{\exp(\pm c_{\alpha/2} \hat{\sigma}(t)/\log \hat{S}(t))}.$$

Log-transformed: R’s default, but bad choice!
Confidence bands: as for Nelson-Aalen using properties of the Brownian bridge.

Quantiles

The $p$th quantile, $\xi_p$, of the survival distribution is given by:
$$p = P(X \leq \xi_p) = 1 - S(\xi_p)$$
and may be estimated by
$$\hat{\xi}_p = \inf \left( t : \hat{S}(t) \leq 1 - p \right).$$

Both this estimator and confidence limits may be read off a Kaplan-Meier plot with pointwise confidence limits.
Example: male melanoma patients, lower quartile, $\hat{\xi}_{0.25} = 3.36$ with 95% c.i. from 2.13 to 5.76 (years).
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.

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:

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

Properties as for ordinary Kaplan-Meier.

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

This brief text does not intend to give a thorough introduction to the statistical R package. Rather, it attempts to help you getting started analysing survival data using R.

Invoking R

R is invoked clicking the R button! You then immediately get access to a window which you can leave again typing q().

R works on objects. These may be of a number of different types of which we will only meet a few in the following. Most commands in R are functions in which case the parentheses () should always be remembered.

E.g., a list of available objects is obtained from the ls() command.

R is case sensitive.

Data frames

We will be using data.frame objects which we will create by reading in existing text files as follows:

my.data<-read.table('filename')

If the first line of filename contains the names of the variables then we add the ,header=T option. If the names are t dc in x y then the variables may be referred to as

my.data$t my.data$dc my.data$in my.data$x my.data$y

By attaching the data frame my.data using the attach(my.data) command this may be abbreviated into

t dc in x y

Alternatively, the first variable may be referred to as the first column in the data matrix: my.data[,1]. Similarly, the first row is my.data[1,].

New variables may be computed using assignment statements like:

z<- x - y

A list of the variables in the data frame may be obtained using names(my.data). If we later want another data frame to be our default then my.data should first be detached typing detach(my.data).

Survival analysis

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 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)

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.
survfit

The command: \texttt{survfit(survdata \textasciitilde x)}
gives a table of estimated mean and median survival times etc. \textit{in subgroups given by x}. If we want to see the estimated Kaplan-Meier survival curves then we may use the \texttt{plot} function:
\texttt{plot(survfit(survdata \textasciitilde x))}
and the plot appears in a separate window. The \texttt{plot} function allows a great number of options controlling axes, labels, line types etc., see what comes out of the \texttt{help} commands:
\texttt{?plot} or \texttt{?plot.survfit}

For the \texttt{plot(survfit(...))} there are special options for controlling confidence limits, and the \textit{Nelson-Aalen} estimator may be obtained using the \texttt{,type=''fleming-harrington'', fun=''cumhaz''}) options.

survdiff

The \texttt{survdiff} function:
\texttt{survdiff(survdata \textasciitilde x)}
gives the logrank (or other non-parametric) tests for comparing the survival distributions (or hazard functions) among subgroups of \textit{x}.

The malignant melanoma study

The data file \texttt{melanom.txt} contains data from the study on prognostic factors in malignant melanoma, cf. ABGK Ex. I.3.1. There are 205 records.

The 9 variables in the data set are:

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

The first line in the file contains the names of the variables.
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.

The 9 variables in the data set are:

- **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

The first line in the file contains the names of the variables.

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.