SISCER 2022 Mod 12

Survival Analysis

Lecture 2

Ying Qing Chen, Ph.D.

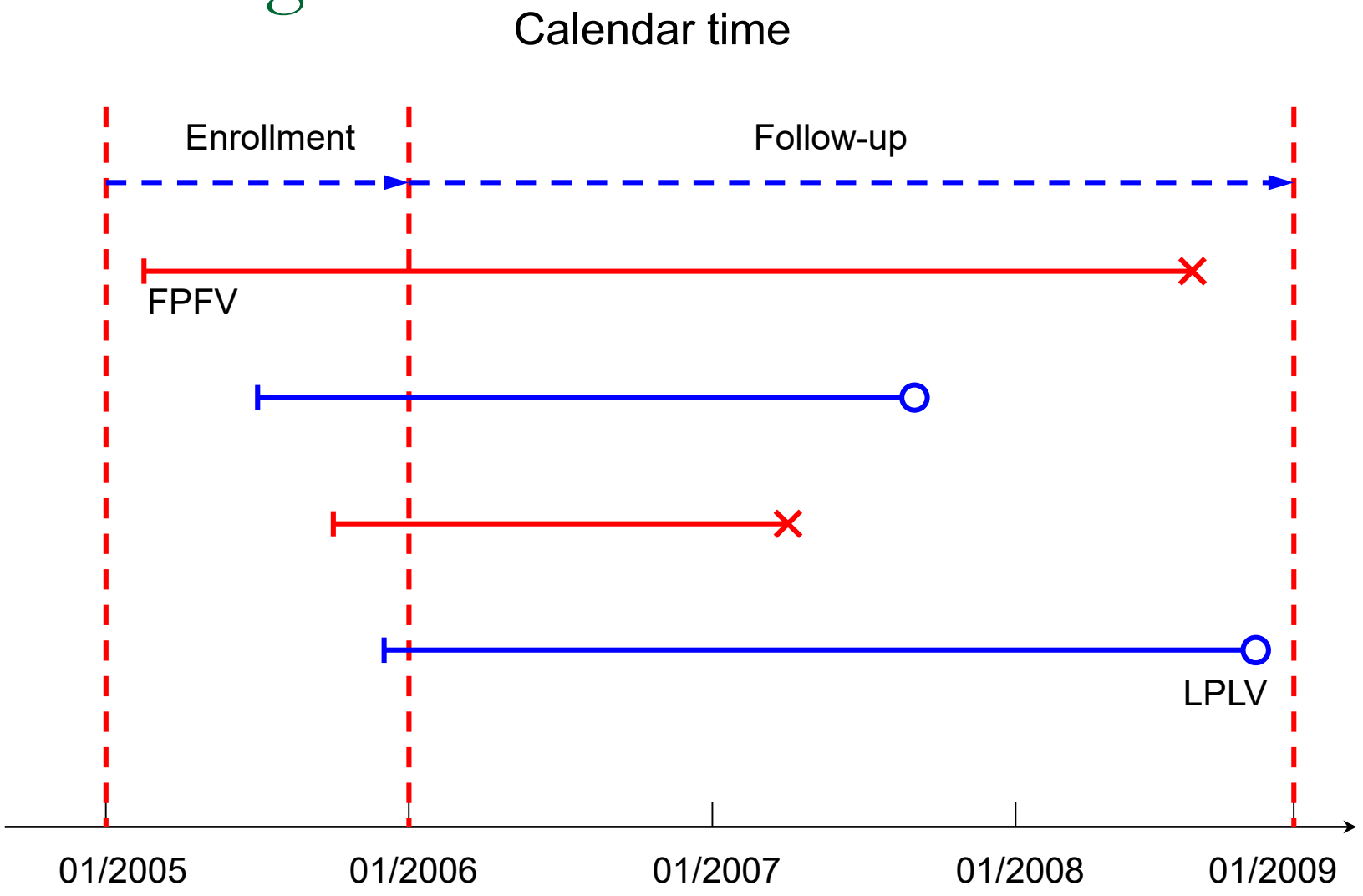
Department of Medicine
Stanford University

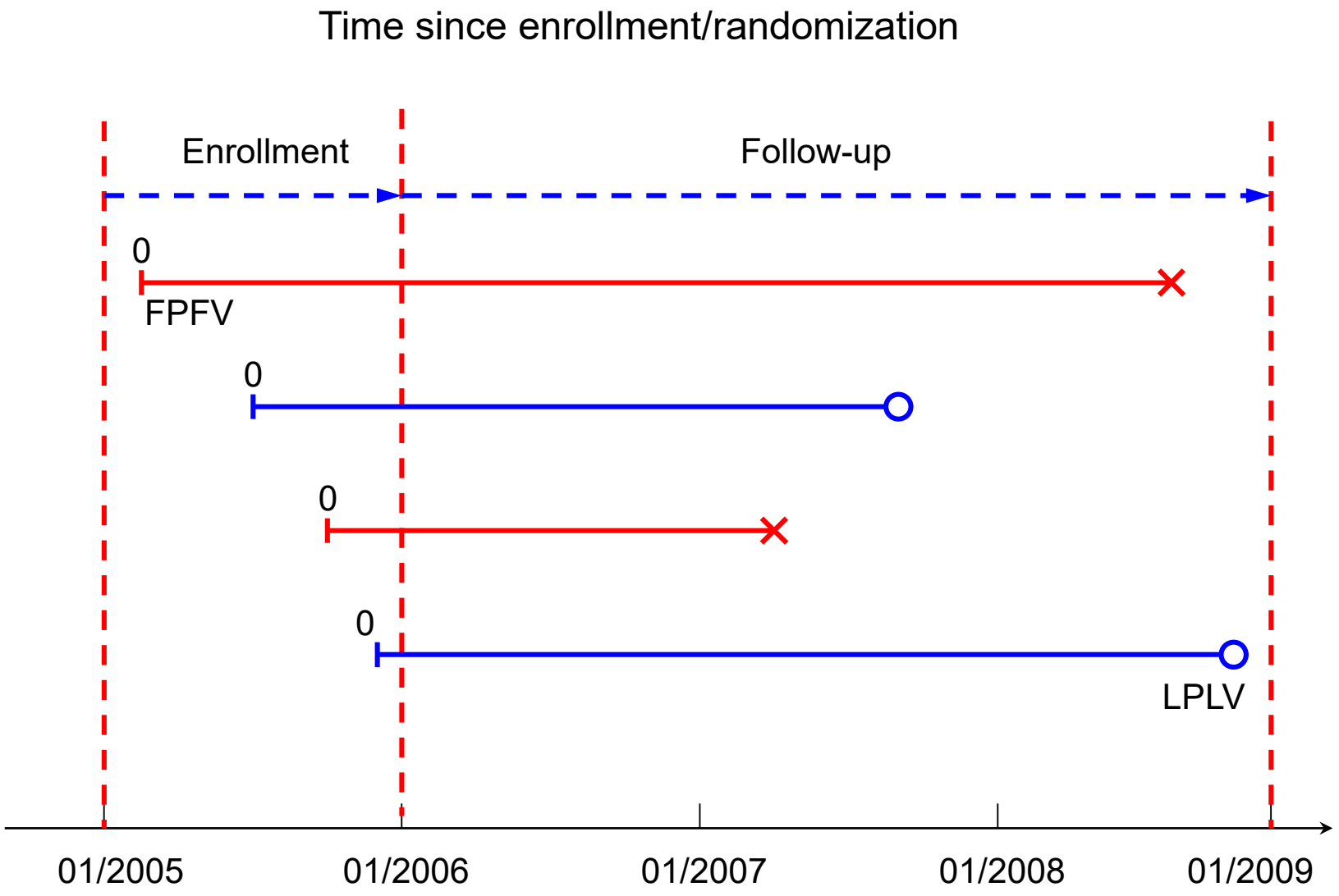
Department of Biostatistics
University of Washington

In Lecture 1

- What we discussed
  - Time-to-event and censoring
  - Parametric methods for analysis of censored time-to-event outcomes
- SAS Users
  - Allison, P (2010) *Survival Analysis Using SAS* : *A Practical Guide, 2nd Ed.* Cary, NC: SAS Institute.
  - SAS support web sites: http://support.sas.com
    - PROC LIFEREG
    - PROC LIFETEST
    - PROC PHREG

Censoring


Calendar time
Enrollment
Follow-up
FPFV
LPLV
01/2005
01/2006
01/2007
01/2008
01/2009


Time since enrollment/randomization
Enrollment
Follow-up
0
FPFV
0
0
0
LPLV
01/2005
01/2006
01/2007
01/2008
01/2009

Some theory: parameter estimation via MLE

Parameter estimation: maximum likelihood estimation (MLE)

- likelihood function:

$$\mathscr{L}(\theta) = \prod_{i=1}^{n} f(x_i \mid Z_i; \theta)^{\delta_i} S(x_i \mid Z_i; \theta)^{1-\delta_i}$$
$$= \prod_{i=1}^{n} \lambda(x_i \mid Z_i; \theta)^{\delta_i} S(x_i \mid Z_i; \theta)$$

This means we are taking "arguments," to maximize it for an estimate

- parameter estimation: $\widehat{\theta} = \arg\max_\theta \mathscr{L}(\theta) = \arg\max_\theta l(\theta)$

$$l(\theta) = \log \mathscr{L}(\theta)$$
$$= \sum_{i=1}^{n} [\delta_i \log \lambda(x_i \mid Z_i; \theta) + \log S(x_i \mid Z_i; \theta)]$$
$$= \sum_{i=1}^{n} [\delta_i \log \lambda(x_i \mid Z_i; \theta) - \Lambda(x_i \mid Z_i; \theta)]$$
$$= \sum_{i=1}^{n} \left[\delta_i \log \lambda(x_i \mid Z_i; \theta) - \int_0^{x_i} \lambda(u \mid Z_i; \theta) du\right]$$

It is equivalent to maximizing the log of likelihood function

Solve $l'(\theta) = 0$ for the MLE $\hat{\theta}$

Statistical theory of MLE

- Consistency:

$$\hat{\theta} \to \theta_0$$

- Normality:

$$\frac{\hat{\theta} - \theta_0}{\sqrt{\text{var}(\hat{\theta})}} \sim \text{Normal}(0, 1)$$

where $\text{var}(\hat{\theta})$ can be estimated by observed Fisher information

$$-\left\{ \frac{d^2}{d\theta^2} l(\hat{\theta}) \right\}^{-1}$$

Parametric methods

- Advantages when parametric models are correctly specified
  - Usually requires less number of parameters
  - Usual MLE can be used to obtained parameter estimates
  - Statistical properties can be easily established
  - Computational routines can be easily adapted
- Major challenges
  - Heavily relies on parametric assumptions
  - Incorrect model specification can lead to biased estimates and wrong inferences
  - Less robust to model misspecification

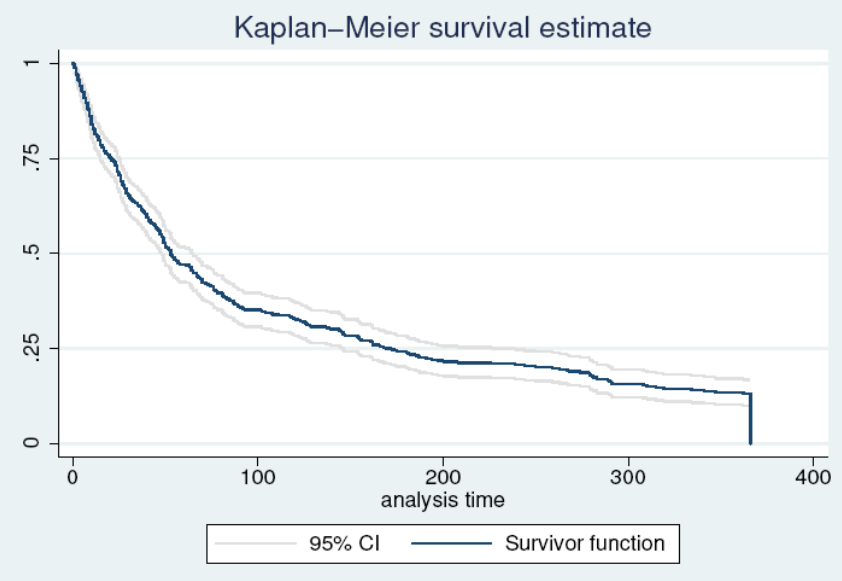
Nonparametric methods

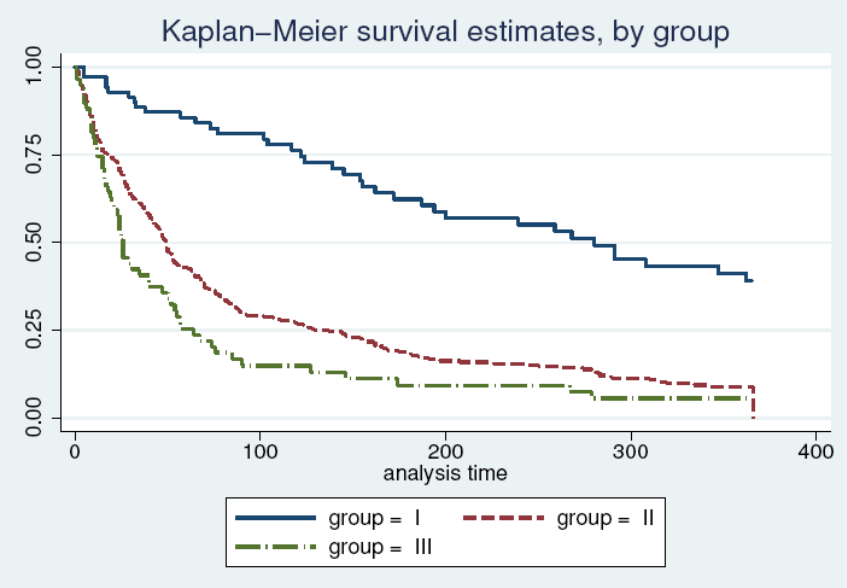
- Features
  - Less distributional assumptions
  - More robust to model misspecification
  - Appealing to data description and model assessment
- Examples
  - Summary statistics: sample mean, sample variance, median, percentiles
  - Distributions: histograms, empirical cumulative distribution function (ECDF)
  - Rank-based test statistics
  - Functional data analysis (FDA)

Kaplan-Meier curves/estimates

- Kaplan-Meier curves
  - Nonparametric estimate
  - Survival function for censored time-to-event outcomes
  - Not rely on any parametric assumptions

Some typical Kaplan-Meier curves





- Features
  - Always starts at S(0)=1
  - Monotonic decreasing (non-increasing)
  - Step functions
  - May not go down to zero all the way when time progresses
  - Shows time-varying profile of absolute risk

Calculate Kaplan-Meier curves

- Easiest way
  - All you need to do is to get your data ready and use any statistical software that you will learn in the labs
  - But it may not help you understand how and why we would like to estimate survival functions the Kaplan-Meier's way
  - In particular for those interested in statistical methods development, it doesn't help with you what assumptions involved or how to apply the theory underlying the Kaplan-Meier estimates to other similar settings

Empirical estimates of survival function

- What do we want to estimate?
  - Population parameter: survival function of an event time

$$\begin{aligned} S(t) &= \mathrm{P}(T > t) \\ &= \text{Population fraction surviving beyond } t \end{aligned}$$

  - Interpretation
    - Percentage of the population not experiencing the disease outcomes (those still at risk at time t)
    - Absolute risk

- What if there is no censoring?
  - Observed data

$$t_1, t_2, \ldots, t_n$$

  - ECDF

$$\hat{S}(t) = \frac{\#t_i > t}{n} = \frac{1}{n} \sum_{i=1}^{n} I(t_i > t)$$

Variance of ECDF

Define

$$
\begin{aligned}
B(t) &= \sum_{i=1}^{n} I(T_i > t) = a \text{ Binomial variable} \\
B(t) &\sim \text{Binomial}(n, p = S(t)) \\
\mathrm{E}[\hat{S}(t)] &= \frac{1}{n} \cdot np = p = S(t) \\
\mathrm{Var}[\hat{S}(t)] &= \frac{1}{n^2}\mathrm{Var}(B(t)) = \frac{1}{n^2} npq \\
&= \frac{S(t)(1 - S(t))}{n}
\end{aligned}
$$

When n is large,

$$
\hat{S}(t) \overset{\text{approx}}{\sim} \text{Normal}\left(S(t), \frac{S(t)(1 - S(t))}{n}\right) .
$$

95% Confidence interval

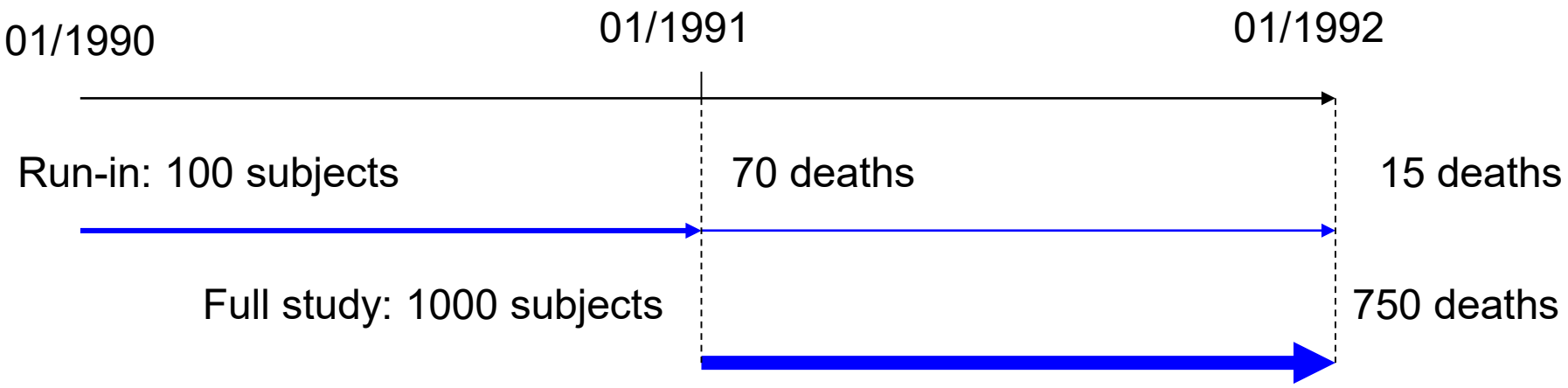
A 95% confidence interval for $S(t)$ is

$$\left(\hat{S}(t) - 1.96\sqrt{\frac{\hat{S}(t)(1-\hat{S}(t))}{n}},\ \hat{S}(t) + 1.96\sqrt{\frac{\hat{S}(t)(1-\hat{S}(t))}{n}} \right).$$

- If n is small ($n < 20$), it is more appropriate to find confidence intervals using the binomial distribution tables (see Mood, Graybill and Boes, Chapter 8).
- If n is large ($n \geq 30$), use the normal approximation to derive confidence intervals.
- The normal approximation works better when $0 << S(t) << 1$ (that is, $S(t)$ is not close to 0 or 1). When $S(t)$ is close to 0 or 1, the Poisson approximation technique is better.

Kaplan-Meier estimates

Example. A prospective study recruited 100 patients in January, 1990 and recruited 1000 patients in January, 1991. The study ended in January, 1992. Survival time T = time from treatment (enrollment) to death. Suppose 70 patients died in year 1 and 15 patients died in year 2 from the first cohort (recruited in 90), and 750 patients died in year 1 from the second cohort.



How do we estimate 2-year survival?

Approach 1 Reduced sample estimate

Only use information from individuals who had been followed for at least two years. That is, use only group 1 data to derive

$$\hat{S}(2) = \frac{100 - 70 - 15}{100} = \frac{15}{100} = 0.15$$

This estimate is statistically appropriate but inefficient. It is appropriate in the sense that $\hat{S}(2)$ is very close to $S(2)$ when n_1 is large. It is inefficient because only part of the data is used. Here

$$\hat{\text{var}}(\hat{S}(2)) = \frac{\hat{S}(2)(1 - \hat{S}(2))}{100}.$$

Inappropriate approaches

— Assume 250 individuals from group 2 died in year 2,

$$\hat{S}(2) = \frac{15}{1100} = 0.014$$

— Assume 250 individuals from group 2 remained alive in year 2

$$\hat{S}(2) = \frac{15 + 250}{1100} = 0.241$$

— Exclude 250 patients from the analyzed data (Watch out! A common mistake!)

$$\hat{S}(2) = \frac{15}{1100 - 250} = 0.018.$$

The Kaplan-Meier approach

Approach 3 (A simple case of the Kaplan-Meier estimate).
Decompose the survival function into conditional probabilities.

$$\begin{aligned} S(2) = \mathrm{P}(T > 2) &= \frac{Pr(T \geq 2)}{Pr(T \geq 1)} \cdot \frac{Pr(T \geq 3)}{Pr(T \geq 2)} \\ &= Pr(T \geq 2 | T \geq 1) \cdot Pr(T \geq 3 | T \geq 2) \end{aligned}$$

$$\hat{Pr}(T \geq 2 | T \geq 1) = \frac{30 + 250}{1100} = \frac{280}{1100}$$

$$\hat{Pr}(T \geq 3 | T \geq 2) = \frac{15}{30}$$

Thus

$$\hat{S}(2) = \frac{280}{1100} \cdot \frac{15}{30} = 0.127.$$

This estimator is more efficient than the reduced sample estimate.

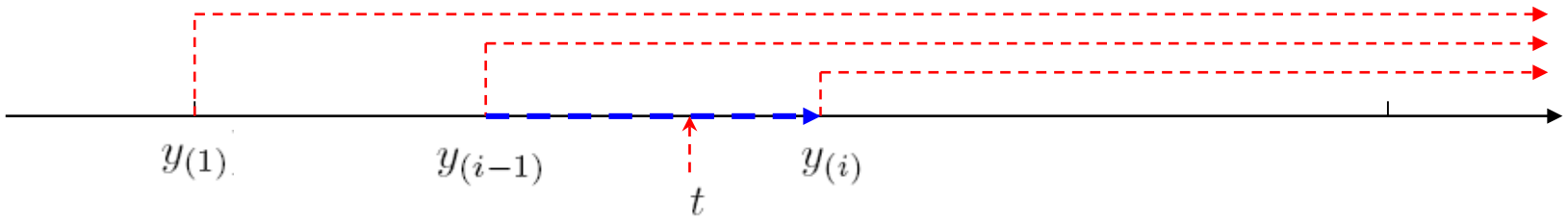
General Kaplan-Meier algorithm

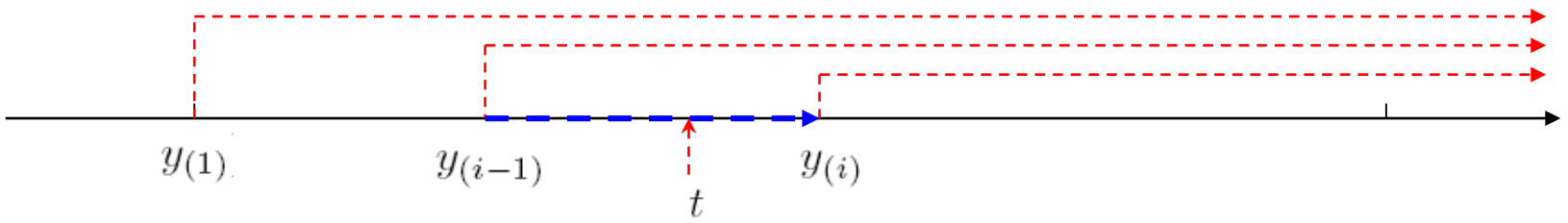
Suppose $y_{(i-1)} \leq t < y_{(i)}$. A principle of nonparametric estimation of S is to assign positive probability to and only to uncensored failure times. Therefore, we try to estimate

$$S(t) \approx \frac{Pr(T \geq y_{(2)})}{Pr(T \geq y_{(1)})} \cdot \frac{Pr(T \geq y_{(3)})}{Pr(T \geq y_{(2)})} \cdots \frac{Pr(T \geq y_{(i)})}{Pr(T \geq y_{(i-1)})}.$$

How to estimate $S(t)$? Define

$$\begin{aligned} R_{(j)} &= \{y_k \ : \ y_k \geq y_{(j)}\} \\ d_{(j)} &= \#\text{ of failures at } y_{(j)} \\ N_{(j)} &= \#\text{ of individuals at risk at } y_{(j)} = \#R_{(j)} \end{aligned}$$





Now estimate $\frac{Pr(T \geq y_{(j+1)})}{Pr(T \geq y_{(j)})}$ by $\frac{N_{(j)} - d_{(j)}}{N_{(j)}}$, $j = 1, 2, \ldots, i-1$. The Kaplan-Meier estimate is thus

$$\hat{S}(t) = \left(1 - \frac{d_{(1)}}{N_{(1)}}\right)\left(1 - \frac{d_{(2)}}{N_{(2)}}\right)\ldots\left(1 - \frac{d_{(i-1)}}{N_{(i-1)}}\right)$$

$$= \boxed{\prod_{y_{(j)} \leq t}\left(1 - \frac{d_{(j)}}{N_{(j)}}\right)}$$

Example

Example Using the previous example $3\ 2^+\ 0\ 1\ 5^+\ 3\ 5$

$$\begin{aligned} N_{(1)} &= 7,\ N_{(2)} = 6,\ N_{(3)} = 4,\ N_{(4)} = 2 \\ d_{(1)} &= 1,\ d_{(2)} = 1,\ d_{(3)} = 2,\ d_{(4)} = 1. \quad //// \end{aligned}$$

$3,\ 2^+,\ 0, 1,\ 5^+,\ 3,\ 5$

uncensored times	0	1	3	5
$d_{(i)}$	1	1	2	1
$N_{(i)}$	7	6	4	2

Ranking the failure times

S(0+)

$$\begin{aligned} \hat{S}(0) &= \left(1 - \frac{1}{7}\right) = \frac{6}{7} = 0.86 \\ \hat{S}(1) &= \frac{6}{7}\left(1 - \frac{1}{6}\right) = \frac{5}{7} = 0.71 \\ \hat{S}(3) &= \frac{5}{7} \cdot \left(1 - \frac{2}{4}\right) = \frac{5}{14} = 0.36 \\ \hat{S}(5) &= \frac{5}{14}\left(1 - \frac{1}{2}\right) = \frac{5}{28} = 0.18 \end{aligned}$$

Remarks

Remark In general, if the largest observed time is uncensored, the Kaplan-Meier estimate will reach the value 0 as $t \geq$ the largest observed time. if the largest observed time is censored, the Kaplan-Meier estimate will not go down to 0 and is unreliable for $t >$ largest y_i. In this case, we say that $\hat{S}(t)$ is undetermined for $t >$ the largest uncensored time.

Variance calculation

Greenwood's formula

The next question is how to identify the variance of the Kaplan-Meier estimate. The idea is sketched for grouped data. First group the data using the uncensored times

$y_{(1)} < y_{(2)} < \ldots < y_{(k)}$.

For each risk set $R_{(j)} = \{y_i \ : \ y_i \geq y_{(j)}\}$, counting the number of failures is a binomial experiment. Thus

$d_{(j)} \sim$ Binomial $(N_{(j)}, \lambda_{(j)})$, where $\lambda_{(j)}$ is the hazard at $y_{(j)}$. Let $q_{(j)} = 1 - \lambda_{(j)}$. For $y_{(i-1)} \leq t < y_{(i)}$,

$$\begin{aligned} \text{var}(\log \ \hat{S}(t)) &= \text{var}(\log\{\hat{q}_{(1)}\hat{q}_{(2)}, \ldots, \hat{q}_{(i-1)}\}) \\ &= \text{var}(\log \hat{q}_{(1)} + \log \hat{q}_{(2)} + \ldots + \log \hat{q}_{(i-1)}) \\ &= \sum_{j=1}^{i-1} \text{var}(\log \hat{q}_{(j)}) \end{aligned}$$

The variances are additive because the risk sets at $y_{(1)}$, $y_{(2)}$, ..., $y_{(k)}$ are nested ($R_{(1)} \supset R_{(2)} \supset \ldots$). Thus, by statistical theory, we can treat $\log \hat{q}_{(1)}, \log \hat{q}_{(2)} \ldots$ as uncorrelated terms.

Use the delta method, for a transformation ϕ of an estimate $\hat{\theta}$, we have

$$\mathrm{var}(\phi(\hat{\theta})) \approx [\phi'(\theta)]^2 \mathrm{var}(\hat{\theta}).$$

Thus

$$\begin{aligned} \mathrm{var}(\log \hat{q}_{(j)}) &\approx \left[\frac{1}{q_{(j)}}\right]^2 \mathrm{var}(\hat{q}_{(j)}) = \frac{1}{q_{(j)}^2} \cdot \frac{\lambda_{(j)} q_{(j)}}{N_{(j)}} = \frac{\lambda_{(j)}}{q_{(j)} N_{(j)}}, \\ \mathrm{var}(\log \hat{S}(t)) &= \sum_{j=1}^{i-1} \mathrm{var}(\log \hat{q}_{(j)}) \approx \sum_{y_{(j)} \le t} \left(\frac{\lambda_{(j)}}{q_{(j)} N_{(j)}} \right) \end{aligned}$$

Use the delta method again,

$$\begin{aligned} \sigma(t)^2 = \mathrm{var}(\hat{S}(t)) &= \mathrm{var}(\underset{\phi}{\exp}(\underset{\hat{\theta}}{\log \hat{S}(t)})) \\ &\approx [S(t)]^2 \cdot \mathrm{var}(\log \hat{S}(t)) \end{aligned}$$

Plug in $\hat{\lambda}_{(j)} = d_{(j)}/N_{(j)}$ and $\hat{q}_{(j)} = \frac{N_{(j)} - d_{(j)}}{N_{(j)}}$. The Greenwood's formula, for estimating the variance of the Kaplan-Meier estimate, is

$$\boxed{\hat{\mathrm{var}}(\hat{S}(t)) \approx [\hat{S}(t)]^2 \sum_{y_{(j)} \le t} \frac{d_{(j)}}{N_{(j)}(N_{(j)} - d_{(j)})}}$$

- The Greenwood standard errors can generate a confidence interval that extends below 0.0 or above 1.0.
- Alternative: provide a CI for a transformation of $S(t)$ and then "back-transform."
- Kalbfleisch & Prentice (1980) suggest

$$\begin{aligned} \mathrm{SE}(\log\{-\log[\widehat{S}(t)]\}) &= \sqrt{\sum_{i\,:\,t_{(i)}\le t} \frac{d_i}{N_i \cdot S_i}} \Big/ \sum_{i\,:\,t_{(i)}\le t} \log(S_i/N_i) \\ &= \mathrm{SE}_{KP} \end{aligned}$$

- Confidence interval: for $\log\{-\log[S(t)]\}$

$$\log\{-\log[\widehat{S}(t)]\} \pm Z_{1-\alpha/2} \cdot \mathsf{SE}_{KP}$$

- Confidence interval: for $S(t)$

$$\text{CI} \quad : \quad \exp[-\exp(\text{upper above})],\ \exp[-\exp(\text{lower above})]$$

$$: \quad \left([\widehat{S}(t)]^{\exp(+Z_{1-\alpha/2}\cdot \mathsf{SE}_{KP})},\ [\widehat{S}(t)]^{\exp(-Z_{1-\alpha/2}\cdot \mathsf{SE}_{KP})} \right)$$

- This interval is implemented in STATA.
- This interval lies within [0,1].

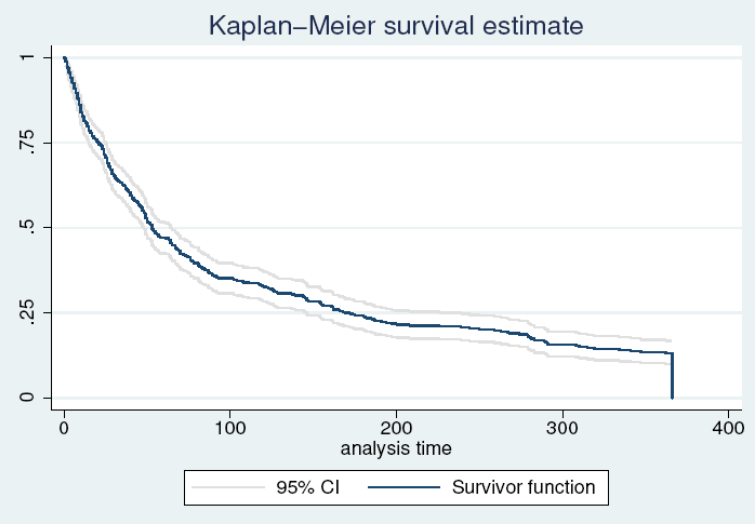
STATA codes/outputs

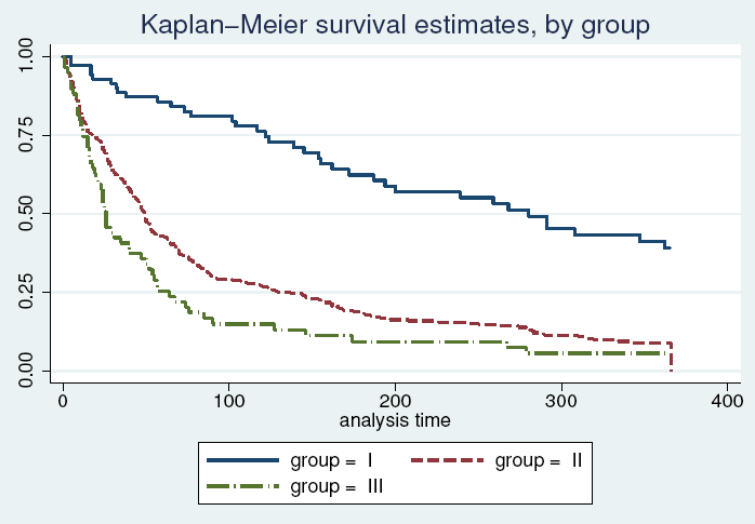
```
*** set the outcome
stset rectime, failure(censor)

*** create output table
sts list

*** create graph(s)
sts graph, gwood
graph export c:\courses\survival\km\HerpesKMplot.eps, as(eps)

sts graph, by(group)
graph export c:\courses\survival\km\HerpesKMgroup.eps, as(eps)
```





The "scary" part

- Underlying theory
  - Nonparametric MLE
  - Likelihood function

$$\mathcal{L} \propto \prod_{i=1}^{n} \left[f(y_i)^{\delta_i} S(y_i)^{1-\delta_i} \right]$$

$$\mathcal{L} \propto \prod_{i=1}^{n} \left[f(y_i)^{\delta_i} S(y_i)^{1-\delta_i} \right] = \prod_{i=1}^{n} \left\{ \frac{f(y_i)}{S(y_i)} \right\}^{\delta_i} \{S(y_i)\}$$

Nonparametric MLE

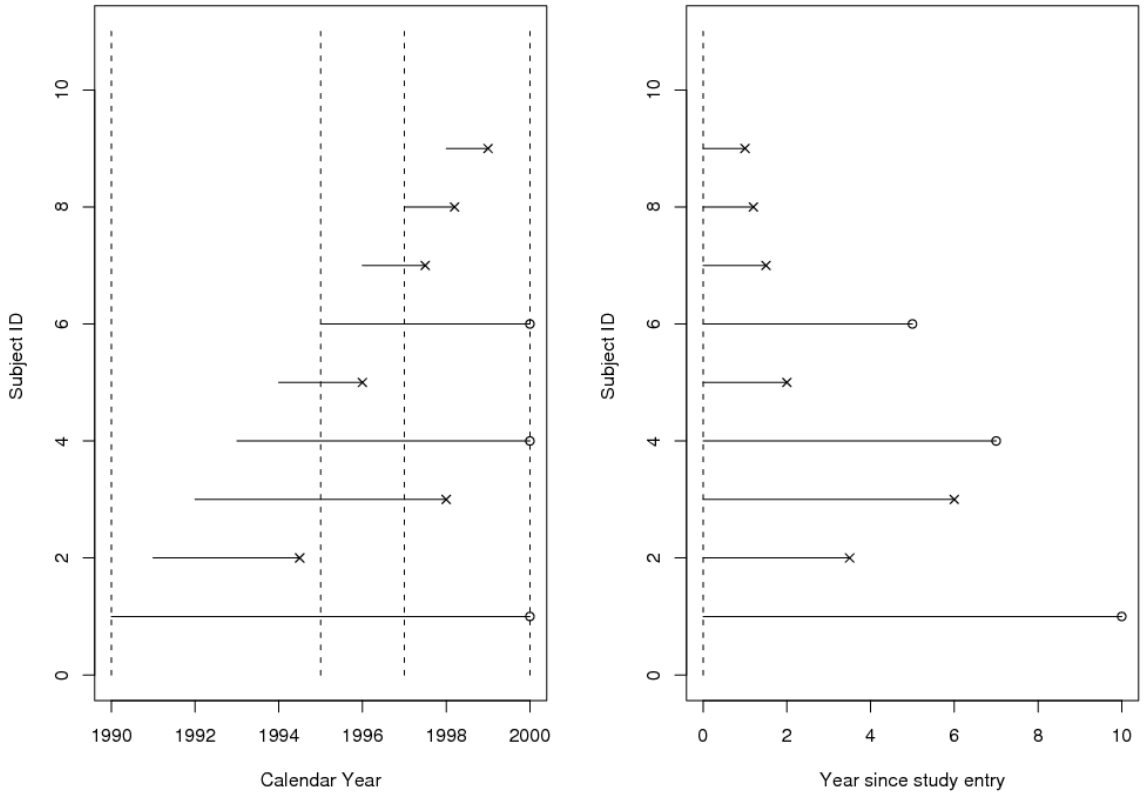
$$\mathcal{L} \propto \prod_{i=1}^{n} \left[f(y_i)^{\delta_i} S(y_i)^{1-\delta_i} \right] = \prod_{i=1}^{n} \left\{ \frac{f(y_i)}{S(y_i)} \right\}^{\delta_i} \{S(y_i)\}$$

- How do we maximize it
  - Based on our data
  - Can we maximized if the underlying survival functions to be continuous at observed failure times?
    - Likelihood function would be always zero is underlying survival functions are continuous at observed failure times!
    - So likelihood function can be only positive when underlying survival functions are step functions with positive jumps at observed failure times

Modern survival analysis based on counting processes

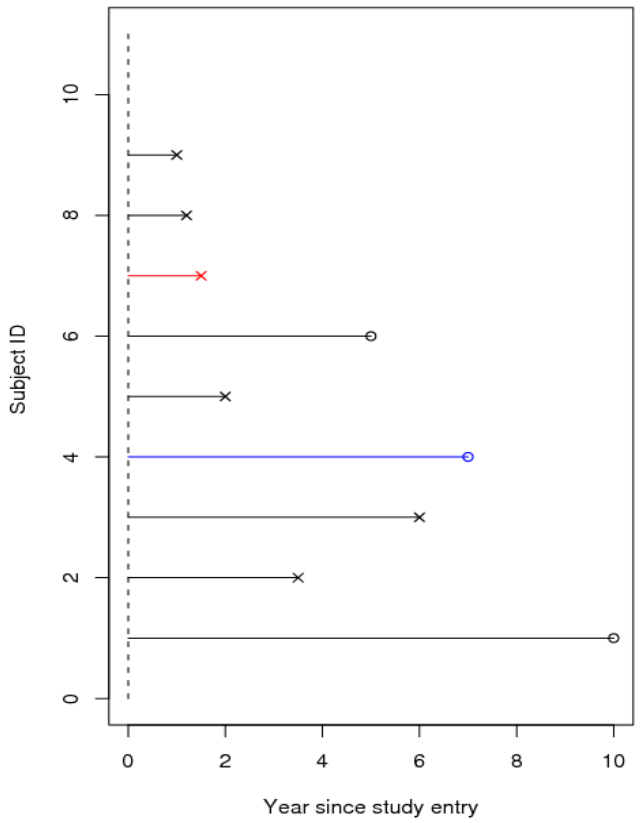
- Data revisited

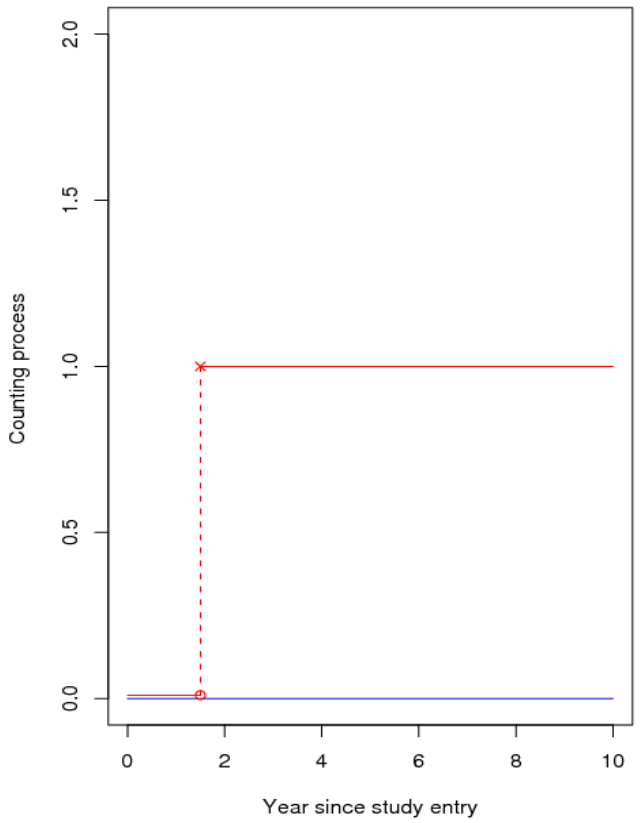
$$(X_i, \Delta_i, Z_i)$$



Counting processes

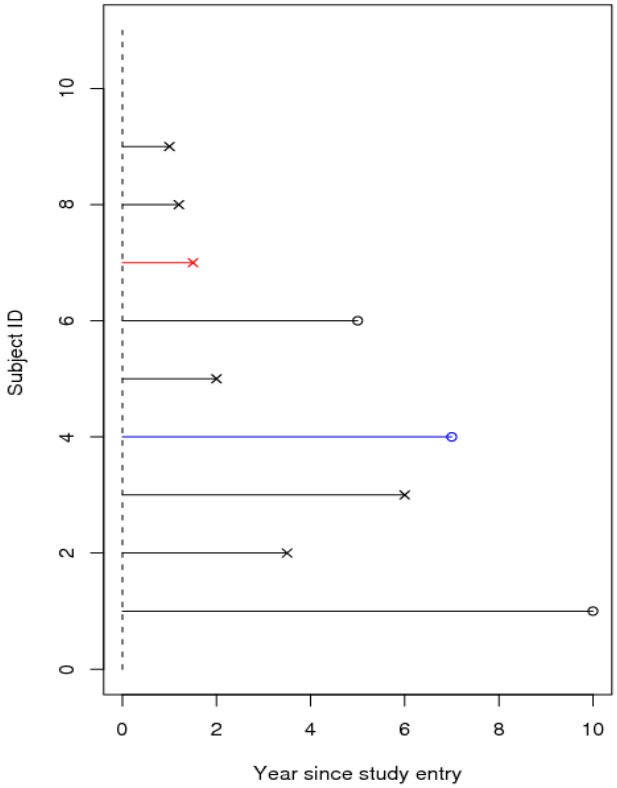
Counting Processes: $N_i(t) = I(X_i \leq t, \Delta_i = 1)$

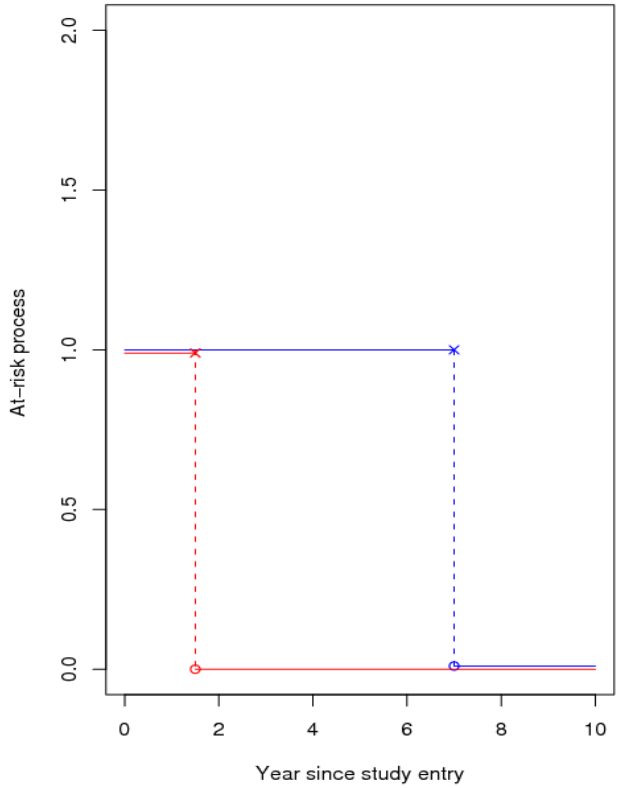




At-risk process

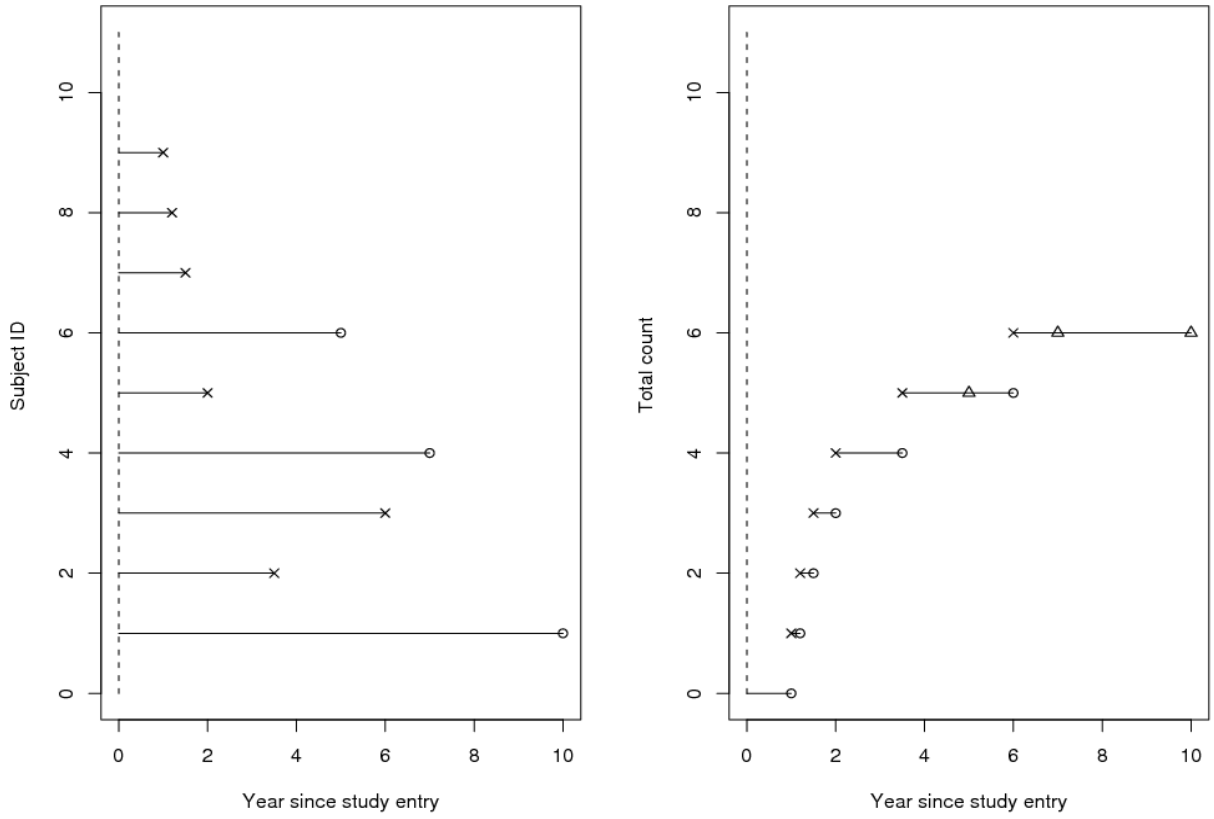
At-risk process: $Y(t) = I(X \geq t)$





Total counting processes

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



A taste of modern survival analysis

Assume T is continuous $\Rightarrow d_j = 1$ mostly

1. $d_j = N(t_j) - N(t_j-) = dN(t_j)$
2. $n_j = Y(t_j) > 0$
3. $\sum_{t_j \le t} \frac{d_j}{n_j} = \int_{u \le t} I(Y(u) > 0) dN(u)/Y(u)$

- Example: calculate $E[dN(t) \mid \mathscr{F}_{t-}]$
  1. $dN(t) = 0$ or 1
  2. $E[dN(t) \mid \mathscr{F}_{t-}] = \Pr\{dN(t) = 1 \mid \mathscr{F}_{t-}\}$
  3. if $Y(t) = 0 \Longrightarrow Y(t-) = 0$, then $\Pr\{dN(t) = 1 \mid \mathscr{F}_{t-}\} = 0$
  4. if $Y(t) = 1 \Longrightarrow Y(t-) = 1$, then

     $$\Pr\{dN(t) = 1 \mid \mathscr{F}_{t-}\} = \Pr\{t \le X \le t + dt, \Delta = 1 \mid X \ge t\}$$

     $X = \min(T, C) \ge t \Leftrightarrow T \ge t, C \ge t$, if T and C are independent, then

     $$\Pr\{dN(t) = 1 \mid \mathscr{F}_{t-}\} = \Pr\{t \le T \le t + dt \mid T \ge t\} = \lambda(t)dt$$

  5. $E[dN(t) \mid \mathscr{F}_{t-}] = Y(t)\lambda(t)dt$
  6. $M(t) = N(t) - \int_0^t Y(u)\lambda(u)du$ is a martingale

Odd Aalen: Heart of the French school of probability theory

$$\mathcal{L} \propto \prod_{i=1}^{n} \left[f(y_i)^{\delta_i} S(y_i)^{1-\delta_i} \right] = \prod_{i=1}^{n} \left\{ \frac{f(y_i)}{S(y_i)} \right\}^{\delta_i} \{S(y_i)\}$$

$$= \left\{ \prod_{(i)} \lambda_{(i)}^{d_{(i)}} \right\} \left\{ \prod_{i=1}^{n} \prod_{y_{(j)} < y_i} (1 - \lambda_{(j)}) \right\} = \prod_{(i)} \lambda_{(i)}^{d_{(i)}} (1 - \lambda_{(i)})^{N_{(i)} - d_{(i}$$

Thus, the unique mle of $\lambda_{(i)}$ is $d_{(i)}/N_{(i)}$ and the Kaplan-Meier estimate is the unique mle.

- References
  - Kaplan & Meier (1958, *JASA*)
  - Kalbfleisch & Prentice (2002)
  - Fleming & Harrington (1991)

Example

Example. (Lee, p29) Forty-two patients with acute leukemia were randomized into a treatment group and a placebo group to assess the treatment effect to maintain remission. T : remission time.

- 6-MP (6-mercaptopurine) group, $n_1 = 21$

  $6, 6, 6, 7, 10, 13, 16, 22, 23, 6^+, 9^+, 10^+, 11^+, 17^+,$
  $19^+, 20^+, 25^+, 32^+, 32^+, 34^+, 35^+$ (months)

- Placebo group, $n_2 = 21$

  $1, 1, 2, 2, 3, 4, 4, 5, 5, 8, 8, 8, 8, 11, 11, 12, 12, 15,$
  $17, 22, 23$ (months)

- Placebo group, $n_2 = 21$

  $1, 1, 2, 2, 3, 4, 4, 5, 5, 8, 8, 8, 8, 11, 11, 12, 12, 15,$

  $17, 22, 23$ (months)

$$\widehat{\text{var}}(S(4)) = \frac{(0.67)(0.33)}{21}$$

$$\widehat{SD}(\hat{S}(4)) = \sqrt{\frac{(0.67)(0.33)}{21}} = 0.103$$

the placebo group

$$\hat{S}(0) = \frac{21}{21} = 1 \qquad \hat{S}(3) = \frac{16}{21}$$

$$\hat{S}(1) = \frac{19}{21} \qquad \hat{S}(4) = \frac{14}{21} = 0.67$$

$$\hat{S}(2) = \frac{17}{21} \qquad \vdots$$

A 95% confidence interval at $t = 4$ is

$$(0.67 - 1.96 \times 0.103, \quad 0.67 + 1.96 \times 0.103) = (0.47, \ 0.87)$$

- 6-MP (6-mercaptopurine) group, $n_1 = 21$

  $6, 6, 6, 7, 10, 13, 16, 22, 23, 6^+, 9^+, 10^+, 11^+, 17^+,$

  $19^+, 20^+, 25^+, 32^+, 32^+, 34^+, 35^+$ (months)

$$\begin{aligned} \widehat{\text{var}}(\hat{S}(10)) &= (0.753)^2 \left(\frac{3}{21 \times 18} + \frac{1}{17 \times 16} + \frac{1}{15 \times 14} \right) \\ &= 0.0093 \end{aligned}$$

the K-M estimate

$$\begin{aligned} \hat{S}(5) &= 1 \\ \hat{S}(6) &= \left(1 - \frac{3}{21}\right) \\ \hat{S}(7) &= \left(1 - \frac{3}{21}\right)\left(1 - \frac{1}{17}\right) \\ \hat{S}(\dot{1}\dot{0}) &= \left(1 - \frac{3}{21}\right)\left(1 - \frac{1}{17}\right)\left(1 - \frac{1}{15}\right) = 0.753 \end{aligned}$$

A 95% confidence interval for $S(10)$ is

$$(0.753 - 1.96\sqrt{0.0093}\ ,\ 0.753 + 1.96\sqrt{0.0093}) = (0.564\ ,\ 0.942)$$

Remark 1 The K-M estimate is a nonparametric method which can be applied to either discrete or continuous data. For a rigorous development of statistical theory, see Kalbfleisch and Prentice (1980).

Remark 2 The accuracy of the K-M estimate and Greenwood's formula relies on large sample size of uncensored data. Make sure that you have at least, say, 20 or 30 uncensored failure times in your data set before using the methods.

Remark 3 Greenwood's formula is more appropriate when $0 << S(t) << 1$. Using Greenwood's formula, the confidence interval limits could be above 1 or below 0. In these cases, we usually replace these limit points by 1 or 0. For example, a 95% confidence interval could be $(0.845, 1.130)$, we will use $(0.845, 1)$ instead.

Cumulative hazard functions

$$\lambda(t) = \frac{f(t)}{S(t)} \dashrightarrow \lambda(t) = \frac{\mathrm{P}(T=t)}{\mathrm{P}(T \geq t)} = \frac{f(t)}{S(t^-)}$$

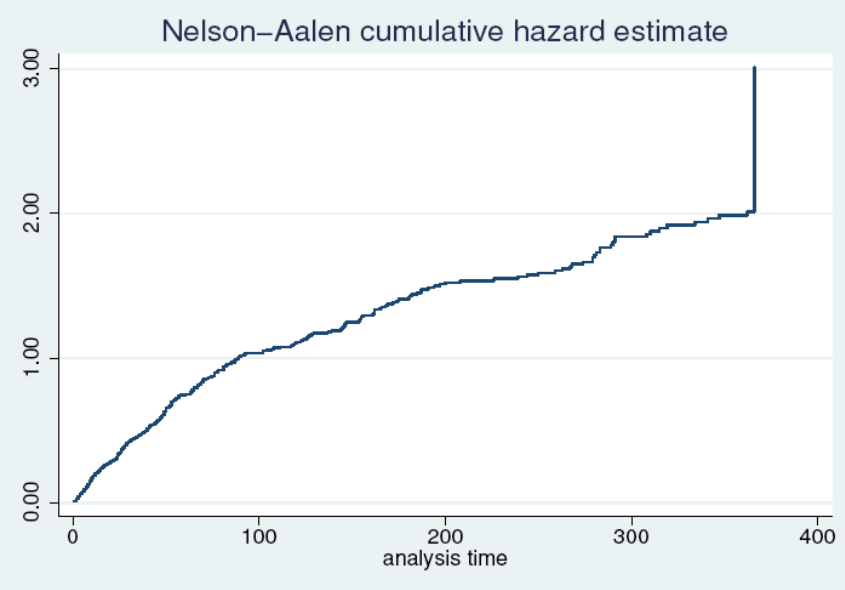
$$f(t) = F'(t)$$

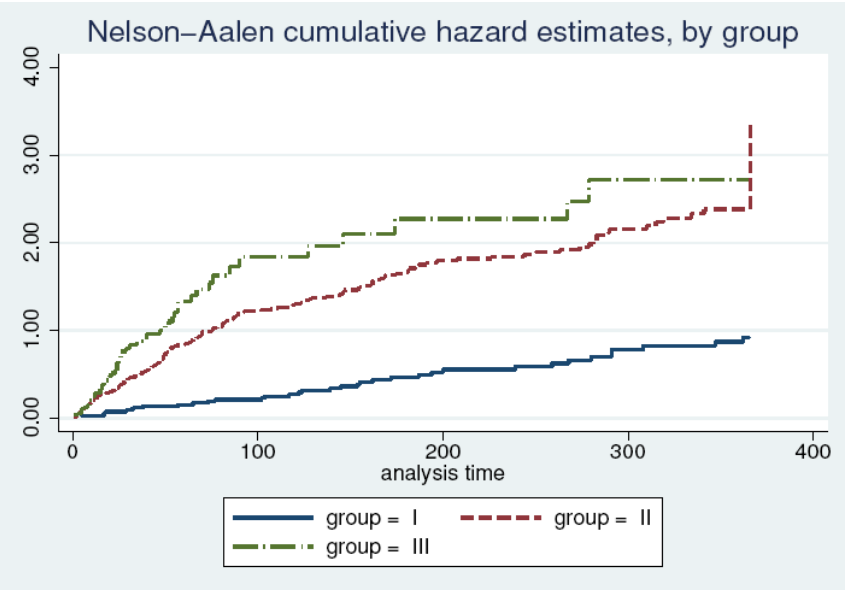
$$S(t) = e^{-\Lambda(t)} = e^{-\int_0^t \lambda(u)du}$$

$$F(t) = 1 - S(t) \dashleftarrow\dashrightarrow \Lambda(t) = \int_0^t \lambda(u)du = \int_0^t \frac{f(u)}{S(u)}du$$

$$= 1 - e^{-\Lambda(t)} \approx 1 - \{1 - \Lambda(t)\} = \Lambda(t)$$

Example of cumulative hazard functions (cumulative incidences)





- Recall: $\Lambda(t) = \int_{s=0}^{t} \lambda(s)ds$
- **Q**: Why estimate $\Lambda(t)$?
  - ▷ Stepping stone to $\lambda(t)$ estimation
  - ▷ Diagnostics (e.g. linear $\Rightarrow$ exponential distn.)
- Note that d_i/N_i can be thought of as estimating

$$P[T \in (t_{(i-1)}, t_{(i)}] \mid T \geq t_{(i-1)}] \approx (t_{(i)} - t_{(i-1)}) \cdot \lambda(t_{(i)})$$

- Given this approximation we have

$$\begin{aligned} \Lambda(t) &= \int_{s=0}^{t} \lambda(s)ds \\ &\approx \sum_{t_{(i)} \leq t} \lambda(t_{(i)}) \cdot (t_{(i)} - t_{(i-1)}) \approx \sum_{t_{(i)} \leq t} d_i/N_i \end{aligned}$$

- The estimator $\widehat{\Lambda}(t) = \sum_{t_{(i)} \leq t} d_i/N_i$ is known as the Nelson-Aalen Cumulative Hazard Estimator.
- Standard error:

$$\text{var}\widehat{\Lambda}(t) = \sum_{t_{(i)} \leq t} \frac{d_i}{N_i S_i}$$

- As $N \to \infty$ the number of jumps gets bigger, and the jump sizes, d_i/N_i, get smaller – so the estimate becomes closer to a continuous function.
- STATA: `sts graph, na`

Kaplan-Meier and Nelson-Aalen

- Kaplan-Meier estimator

$$\widehat{S}(t) = \prod_{t_j \le t} \left(1 - \frac{d_j}{n_j}\right) \approx \prod_{t_j \le t} e^{-\frac{d_j}{n_j}} = e^{-\sum_{t_j \le t} \frac{d_j}{n_j}}$$

- Assume T is continuous $\Rightarrow d_j = 1$ mostly
  1. $d_j = N(t_j) - N(t_j-) = dN(t_j)$
  2. $n_j = Y(t_j) > 0$
  3. $\sum_{t_j \le t} \frac{d_j}{n_j} = \int_{u \le t} I(Y(u) > 0) dN(u)/Y(u)$

- Nelson-Aalen estimator

$$\widehat{\Lambda}(t) = -\log \widehat{S}(t) = \int_0^t \frac{I(Y(u) > 0) dN(u)}{Y(u)}$$

Kernel estimate of hazard functions

- Bad Idea: Let the estimate of $\lambda(t)$ just be

$$\widehat{\lambda}(t) = \frac{\widehat{\Lambda}(t+\frac{1}{2}) - \widehat{\Lambda}(t-\frac{1}{2})}{(t+\frac{1}{2}) - (t-\frac{1}{2})}$$

  - When time is measured in days (minutes) this estimate will just be d_i/N_i at $t_{(i)}$ and then zero everywhere else.

- Better Idea: Don't just use d_i/N_i or 0.0, but rather look in the "neighborhood" of t and average values that are close by.

- Choose a weight function to assign weight to values:
  kernel function $K(x)$ where $\int K(x)dx = 1$
  - $K(x) = 1(x - \frac{1}{2} \leq x \leq x + \frac{1}{2})$
  - $K(x) = \exp(-\frac{1}{2}x^2)/\sqrt{2\pi}$
  - $K(x) = 1(x - 1 \leq x \leq x + 1) \cdot (1 - x^2) \cdot 0.75$

- Choose a bandwidth to determine the (effective) size of the neighborhood over which you average: bandwidth b.
- Estimate

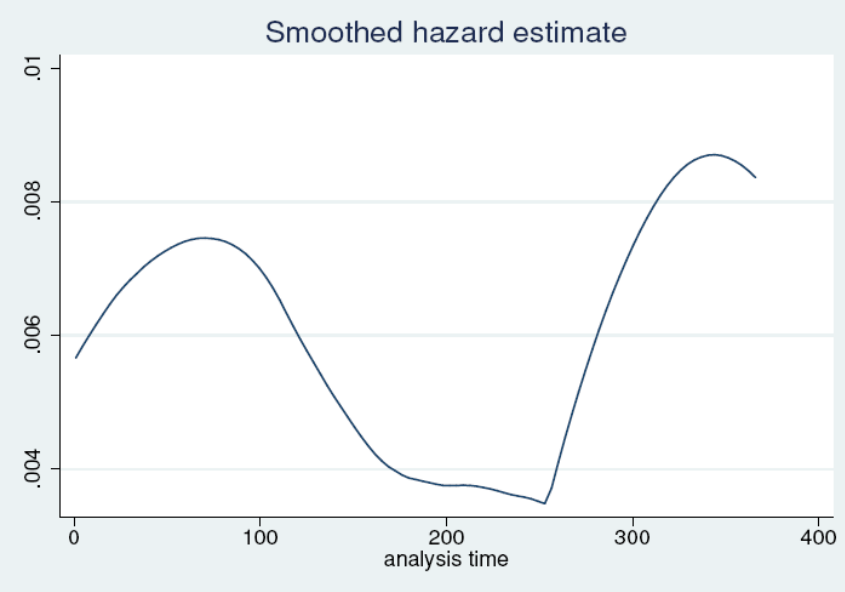
$$\widehat{\lambda}(t) = \sum_i \frac{1}{b} \cdot K\left(\frac{t - t_{(i)}}{b}\right) \cdot \frac{d_i}{N_i}$$

- Standard Error:

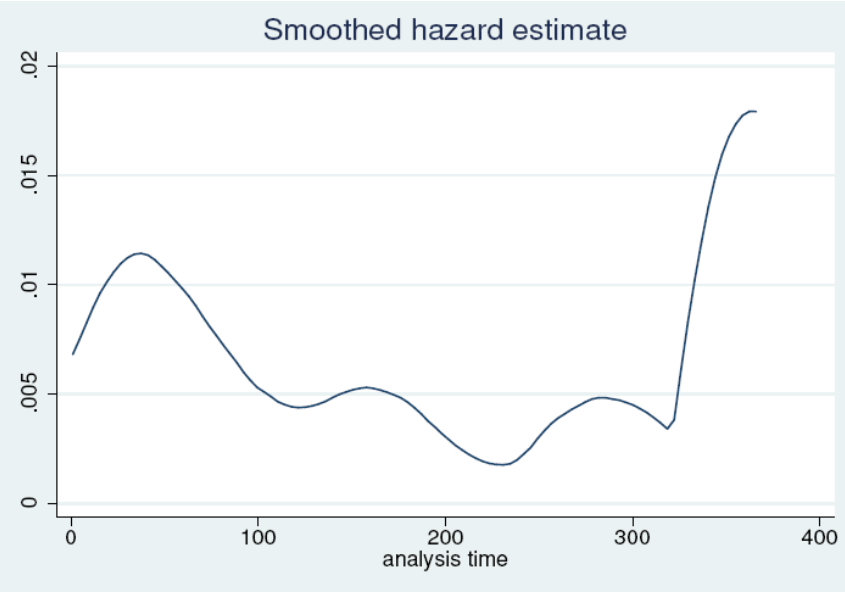
$$\mathsf{SE}[\widehat{\lambda}(t)] = \frac{1}{b}\left\{\sum_i \left[K\left(\frac{t - t_{(i)}}{b}\right)\right]^2 \cdot \frac{d_i}{N_i S_i}\right\}^{1/2}$$

- Large value of $b \Rightarrow$ smooth estimate.
- Small value of $b \Rightarrow$ fluctuating estimate.
- STATA: `sts graph, hazard` with `width` option.

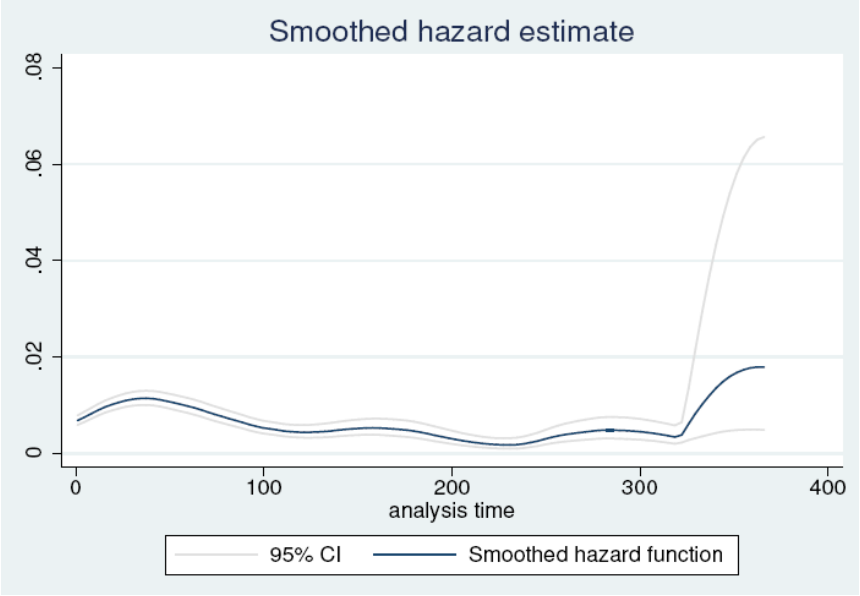
Hazard Estimate: `width(50)`



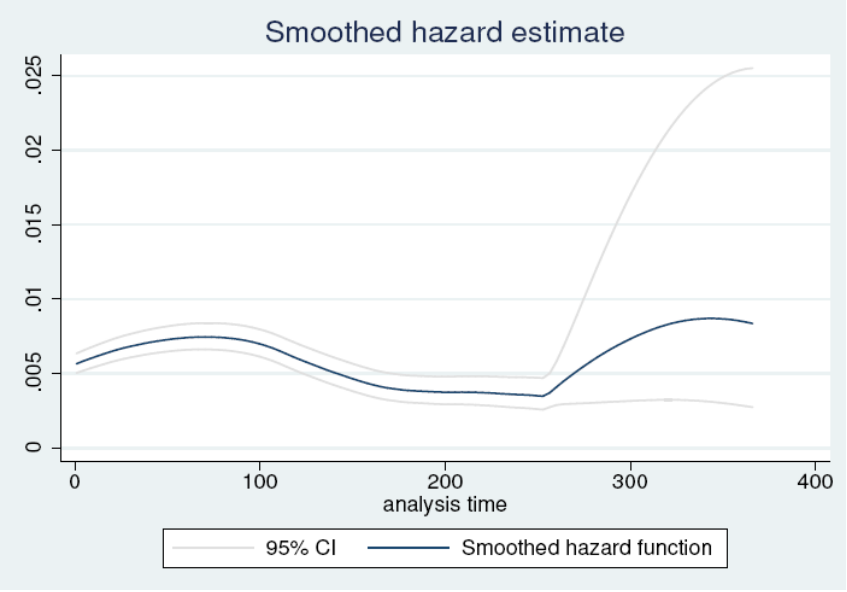
Hazard Estimate: `width(20)`



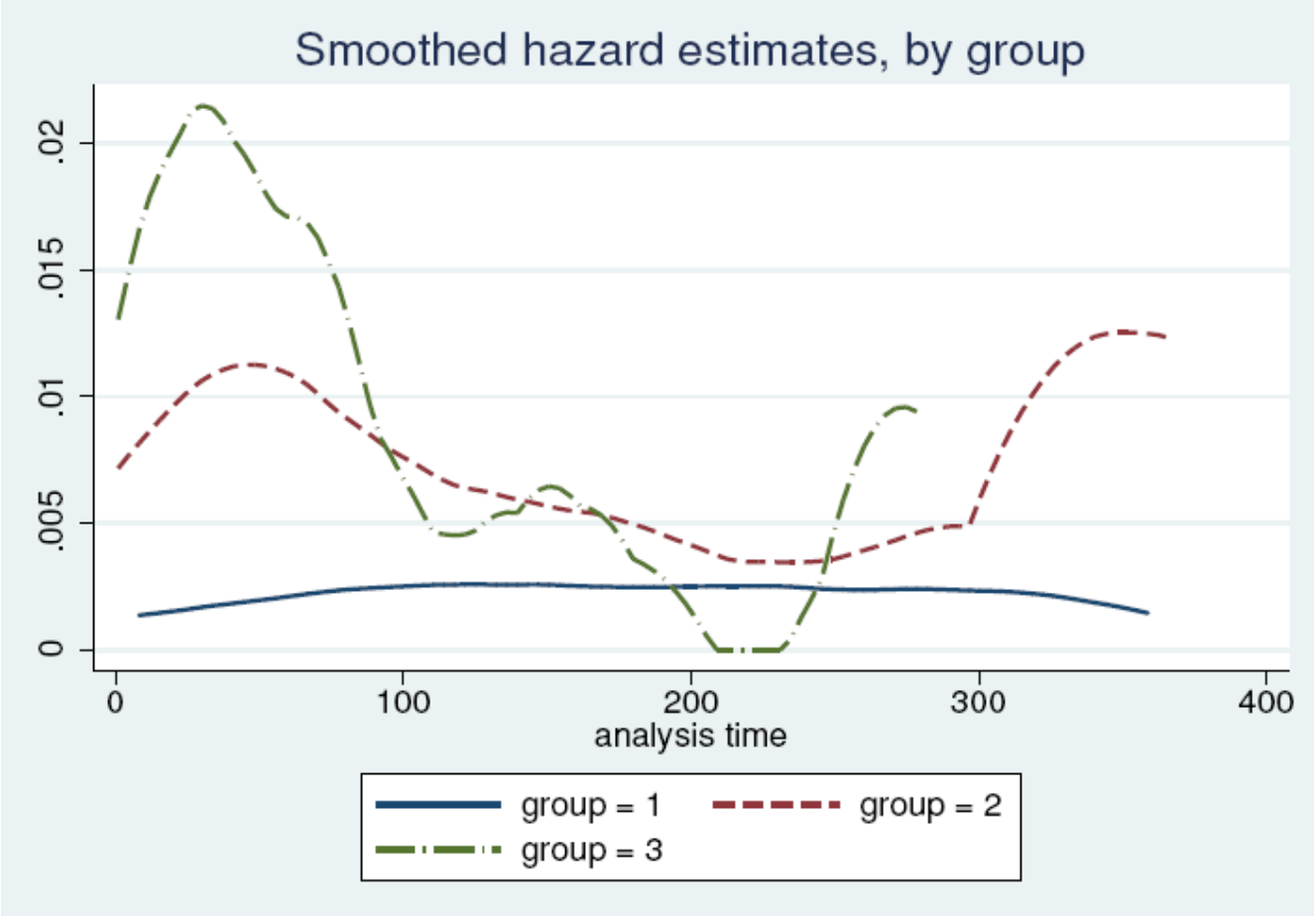
Hazard Estimate: `width(20) cihazard`



Hazard Estimate: `width(50) cihazard`



Hazard Estimates by Group: `default`



Summary

- We have illustrated (2) main statistical approaches
  - Parametric model
    - Assume a model form (e.g. Weibull)
    - Estimate parameter(s) using Maximum Likelihood
    - Use model to estimate mean, median, $S(t)$, $\lambda(t)$
  - Non-parametric methods
    - No specified model form
    - Kaplan-Meier Estimator of $S(t)$ (and percentiles such as median)
    - Nelson-Aalen Estimator of $\Lambda(t)$
    - Kernel Estimator of $\lambda(t)$

- Parametric methods will generate estimates for the mean even when a large fraction of observations are censored – thoughts?
- Non-parametric methods produce estimates of $S(t)$ with increasing variance with increasing time.
- All methods have assumed that censoring is independent of the survival time (i.e. no selection bias).
- Summaries can be produced for groups (subsets) of subjects.
- **Q**: How can we make inference regarding group differences?