Survival Analysis I (CHL5209H)

Olli Saarela

Dalla Lana School of Public Health
University of Toronto

olli.saarela@utoronto.ca

January 24, 2017

Non-parametric estimators for survival and cumulative hazard functions

Motivation

- Previously we have been focusing on parametric survival models, which specify a parametric form for the hazard function, and therefore, for the event time distribution.
- Estimating hazard functions completely non-parametrically is not possible, as this will always require some form of smoothing (why?), but cumulative hazards and survival functions can be estimated non-parametrically.
- Often we want to look into the survival patterns in the data descriptively, before considering any parametric models, or compare survival visually between different groups.
- We can also test for between group differences in survival non-parametrically.

Consecutive follow-up intervals

- Recall the idea of splitting the follow-up period into N short intervals of length h.
- The risk and rate parameters were then connected through $\pi = \lambda h$.
- The probability of surviving through these N intervals was, through the chain rule of conditional probabilities, $(1 - \pi)^N = (1 - \lambda h)^N$.
- For the chain rule to work, we do not actually need to assume that the rate is constant over time; rather we can allow separate rate for each interval to get a generalized version

$$\prod_{j=1}^{N} (1 - \lambda_j h) .$$

Kaplan-Meier estimator

- If in interval j we observed d_j events, and n_j individuals were at risk, contributing $n_j h$ time units of follow-up time, we can estimate the rate λ_j by

$$\hat{\lambda}_j = \frac{d_j}{n_j h}.$$

- Thus, an estimate for the survival probability is given by

$$\prod_{j=1}^{N} \left(1 - \frac{d_j}{n_j}\right).$$

- Since this changes only when events actually occurred, we can equivalently take the product over the ordered event (and censoring) times t_j observed in the data, until a specific time point t, to get the Kaplan-Meier estimator

$$\hat{S}_{\mathrm{KM}}(t) = \prod_{j:t_j \leq t} \left(1 - \frac{d_j}{n_j}\right).$$

Numerical illustration (C&H 1993, p. 36)

Table 4.2. Cumulative survival probabilities from the Kaplan–Meier method. Non-melanoma deaths (*) are counted as losses.

Month	N	D	L	Conditional probability of death	Conditional probability of survival	Cumulative prob. of survival
0	50	2		0.0400	0.9600	0.9600
1	48	1		0.0208	0.9792	0.9400
2	47	2		0.0426	0.9574	0.9000
3	45	1	1*	0.0222	0.9778	0.8800
8	43	1		0.0233	0.9767	0.8595
10	42	1		0.0238	0.9762	0.8391
12	41	1	1*	0.0244	0.9756	0.8186
13	39	1		0.0256	0.9744	0.7976
15	38	1		0.0263	0.9737	0.7766
18	37		1*			
19	36	1		0.0278	0.9722	0.7551
21	35		1			
27	34		2			
30	32		1			
33	31	1	1	0.0323	0.9677	0.7307
34	29	1		0.0345	0.9655	0.7055
38	28		1			
40	27		1			
41	26	1		0.0385	0.9615	0.6784
43	25		1			
44	24		1			
46	23		1			
54	22		1			
55	21	1		0.0476	0.9524	0.6461
56	20	1		0.0500	0.9500	0.6138
57	19		2			
60	17		1*			

K-M curve (C&H 1993, p. 37)

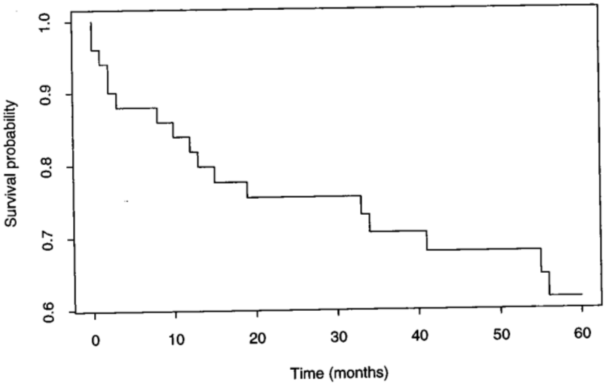

Fig. 4.7. Cumulative survival probability by the Kaplan–Meier method.

The Delta method

- We can obtain pointwise confidence intervals for the survival probabilities if we can obtain a standard error for the KM-estimator.
- For this purpose, note that we can approximate the function g of an estimator $\hat{\theta}$ as

$$g(\hat{\theta}) \approx g(\theta) + g'(\theta)(\hat{\theta} - \theta)$$

  and, if $\hat{\theta}$ unbiased, the expectation of the this as

$$E[g(\hat{\theta})] \approx g(\theta) + g'(\theta)E[\hat{\theta} - \theta] = g(\theta).$$

- Similarly, for the variance we get

$$\begin{aligned} V[g(\hat{\theta})] &\approx E[(g(\hat{\theta}) - g(\theta))^2] \\ &\approx E[(g'(\theta)(\hat{\theta} - \theta))^2] \\ &= (g'(\theta))^2 E[(\hat{\theta} - \theta)^2] = (g'(\theta))^2 V[\hat{\theta}]. \end{aligned}$$

The Delta method (cont.)

- This approach is known as the Delta method; it is useful when we know or can easily calculate the variance of an untransformed statistic, and want the approximate variance of a transformation of this.
- In particular, in the case of the KM-estimator, it turns out to be easier to first calculate the variance of the logarithm of the KM-estimator, and use the Delta method to get the variance of the KM-estimator itself.

Greenwood formula

- The Kaplan-Meier estimator has a variance expression known as the Greenwood formula:

$$\hat{V}\left(\hat{S}_{\mathrm{KM}}(t)\right) = \hat{S}_{\mathrm{KM}}(t)^2 \sum_{j:t_j \leq t} \frac{d_j}{n_j(n_j - d_j)}.$$

- This can be motivated through the following calculation:

$$\begin{aligned}
V\left(\log \hat{S}_{\mathrm{KM}}(t)\right) &\approx \sum_{j:t_j \leq t} V\left(\log\left(1 - \frac{d_j}{n_j}\right)\right) \\
&\approx \sum_{j:t_j \leq t} \frac{1}{(1 - \frac{d_j}{n_j})^2} V\left(\frac{d_j}{n_j}\right) \\
&= \sum_{j:t_j \leq t} \frac{d_j}{n_j(n_j - d_j)}.
\end{aligned}$$

Greenwood formula (2)

- Using the delta method again, we get

$$
\begin{aligned}
V\left(\hat{S}_{\mathrm{KM}}(t)\right) &\approx \hat{S}_{\mathrm{KM}}(t)^2 V\left(\log \hat{S}_{\mathrm{KM}}(t)\right) \\
&\approx \hat{S}_{\mathrm{KM}}(t)^2 \sum_{j:t_j \leq t} \frac{d_j}{n_j(n_j - d_j)}.
\end{aligned}
$$

- This could be used to derive confidence bands for the survival curve through $\hat{S}_{\mathrm{KM}}(t) \pm 1.96\sqrt{V(\hat{S}_{\mathrm{KM}}(t))}$, but has the problem that the interval limits are not bounded between 0 and 1.
- This could be circumvented by using the transformation $\log(-\log \hat{S}_{\mathrm{KM}}(t))$, which can take values in $(-\infty, \infty)$, and the corresponding variance

$$
\hat{V}\left(\log(-\log \hat{S}_{\mathrm{KM}}(t))\right) = \frac{1}{(\log \hat{S}_{\mathrm{KM}}(t))^2} \sum_{j:t_j \leq t} \frac{d_j}{n_j(n_j - d_j)}.
$$

Nelson-Aalen estimator

- Making the time intervals infinitely short, h becomes $\mathrm{d}t$, and thus each $\hat{\lambda}_j h$ becomes equivalent to a corresponding increment in the estimated cumulative hazard, motivating the Nelson-Aalen estimator for the cumulative hazard function $\Lambda(t)$:

$$\hat{\Lambda}_{\mathrm{NA}}(t) = \sum_{j:t_j \leq t} \frac{d_j}{n_j}.$$

- This is less informative than the survival curve, since cumulative hazard is not a probability, but can be used for example for visually checking how constant the hazard rate is over time, since a constant hazard rate corresponds to linear cumulative hazard.

The connection between the two estimators

- The theoretical survival and cumulative hazards have the familiar connection

$$S(t) = \exp\{-\Lambda(t)\} \Leftrightarrow -\log\left(S(t)\right) = \Lambda(t).$$

- The same relationship applies to the Kaplan-Meier and Nelson-Aalen estimators approximately, because

$$\begin{aligned} -\log\left(\hat{S}_{\mathrm{KM}}(t)\right) &= -\sum_{j:t_j \leq t} \log\left(1 - \frac{d_j}{n_j}\right) \\ &\approx -\sum_{j:t_j \leq t} -\frac{d_j}{n_j} \\ &= \hat{\Lambda}_{\mathrm{NA}}(t), \end{aligned}$$

or

$$\hat{S}_{\mathrm{KM}}(t) \approx \exp\left\{-\hat{\Lambda}_{\mathrm{NA}}(t)\right\}.$$

Between-group comparisons of survival

- We can plot two or more KM-curves along with their respective confidence bands in the same figure, and see whether the intervals are overlapping at any given point in time.
- However, the pointwise comparisons do not directly correspond to comparing whether the survival curves as a whole are different between the groups.
- For testing equivalence of two or more survival functions, we can use the non-parametric *log-rank test*.

Log-rank test

- Consider grouping the follow-up data at time t_j as follows.

	Group 0 (reference)	Group 1 (intervention)	Total count
Events	d_{j0}	d_{j1}	d_j
Survivors	$n_{j0} - d_{j0}$	$n_{j1} - d_{j1}$	$n_j - d_j$
At risk	n_{j0}	n_{j1}	n_j

- We can think that the event count in group k at time t_j is distributed under the null as

$$d_{jk} \sim \mathrm{Binomial}(n_{jk}, \lambda_j h).$$

- On the other hand, the event count in group k at time t_j conditional on the total event count at this time is distributed as

$$d_{jk} \mid d_j \sim \mathrm{Hypergeometric}(n_{jk}, n_j, d_j).$$

Hypergeometric distribution

- Under the hypergeometric distribution, the probability that d_{jk} events occurred in group k out of the possible d_j is given by

$$\frac{\binom{n_{jk}}{d_{jk}}\binom{n_j-n_{jk}}{d_j-d_{jk}}}{\binom{n_j}{d_j}} = \frac{\binom{n_{j0}}{d_{j0}}\binom{n_{j1}}{d_{j1}}}{\binom{n_j}{d_j}}.$$

- The corresponding conditional mean and variance from the hypergeometric distribution are given by

$$E[d_{jk} \mid d_j] = \frac{n_{jk} d_j}{n_j} \equiv E_{jk}$$

and

$$V[d_{jk} \mid d_j] = d_j \frac{n_{jk}}{n_j} \left(\frac{n_j - n_{jk}}{n_j} \right) \left(\frac{n_j - d_j}{n_j - 1} \right) \equiv V_j.$$

The test statistic

- The log-rank test statistic aggregates the observed and expected event counts d_{j1} and E_{j1} in the intervention group over the times indexed by j to get

$$\frac{\left(\sum_j d_{j1} - \sum_j E_{j1}\right)^2}{\sum_j V_j},$$

which is approximately χ^2-distributed with one degree of freedom.

- If the null not true, and the groups are different in terms of survival, the test statistic will give large values. (Why?)
- Relabeling the groups does not change the value of the test statistic, so either group can be the reference.
- The test also generalizes for more that two groups being compared.