MATH 60604A
Statistical modelling
§ 7c - Kaplan–Meier estimator

Léo Belzile

HEC Montréal
Department of Decision Sciences

Notation

We consider a continuous random variable T and an associated sample of size n.

- Suppose that there are D distinct event times
- Let $0 \leq t_1 < t_2 < \cdots < t_D$ denote these ordered D failure times.
- Let r_j denote the number of individuals who are **at risk** at time t_j.
  - That is, these individuals have not had experienced the event (nor been censored) before time t_j.
  - Thus, r_j is the number of known survivors just before time t_j who are "at risk" of experiencing the event at time t_j.
- Let $d_j \in \{0, \ldots, r_j\}$ denote the number of failures at time t_j (there are d_j deaths at time t_j).

Derivation of Kaplan–Meier estimator

The probability of dying in the time window $(t_j, t_{j+1}]$ given survival until t_j is

$$h_j = \mathsf{P}\left(t_j < T \leq t_{j+1} \mid T > t_j\right) = \frac{S(t_j) - S(t_{j+1})}{S(t_j)}.$$

This recursion yields

$$S(t) = \prod_{j : t_j < t} (1 - h_j).$$

The Kaplan–Meier estimator is is non-parametric:

- it does not assume any underlying probability distribution for the variable T_i
- rather, the conditional probabilities $\{h_j\}_{j=1}^{D}$ are treated as parameters of the model.

Likelihood for the discrete observations

- Each failure at time t_j contributes h_j to the likelihood
  - the probability of failure at t_j given survival in the previous time interval.
- The likelihood contribution of survivors at time t_j is $1 - h_j$.
- We may write the log likelihood as

$$\ell(\boldsymbol{h}) = \sum_{j=1}^{D} \{d_j \ln(h_j) + (r_j - d_j) \ln(1 - h_j)\},$$

the sum of contributions of binomial variables at time t_j.

Optimizing the survival probabilities

- Differentiating $\ell(\boldsymbol{h})$ with respect to h_j, we find $\widehat{h}_j = d_j/r_j$.
- The Kaplan–Meier estimator of the survival function is

$$\widehat{S}(t) = \prod_{t_j < t} \left(1 - \frac{d_j}{r_j}\right)$$

- Intuition: d_j/r_j is the sample proportion of death at time t_j relative to the total population still alive at time t_j.

Example

The `breastcancer` data from Sedmak *et al.* (1989) contain informations on patients with breast cancer, including the following variables:

- `time`: time until death, or end of study (in months)
- `death`: indicator variable for death either `0` for right-censored times or `1` for death
- `im`: response to immunohistochemical examination, either negative (`0`) or positive (`1`)

Descriptive statistics for `breastcancer`

Analysis Variable : time				
N	**Mean**	**Std Dev**	**Minimum**	**Maximum**
45	98.33	51.84	19.00	189.00

death	Frequency	Percent
0	21	46.67
1	24	53.33

im	Frequency	Percent
0	36	80.00
1	9	20.00

In practice, Kaplan–Meier estimator requires significant number observations to be a reliable approximation of the true survivor curve ($n \gg 1000$).
Keep in mind censored observations contribute less information than observed failure times.

Estimation of the survival function

SAS code to fit the Kaplan–Meier estimator

```
proc lifetest data=statmod.breastcancer method=km plots=(s(cl));
time time*death(0);
run;
```

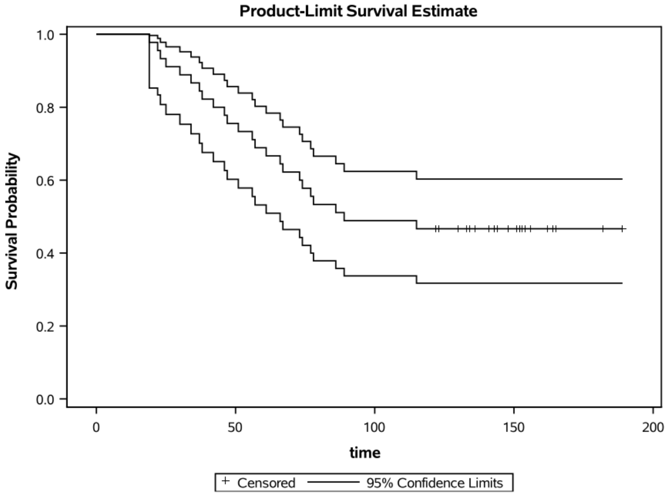
The `time` argument indicates both the response T_i (`time`) and the right-censoring indicator δ_i, with the reference in parenthesis for the right-censored observations (`death=0`)

Estimated survival function

Product-Limit Survival Estimates						
time		Survival	Failure	Survival Standard Error	Number Failed	Number Left
0.000		1.0000	0	0	0	45
19.000		0.9778	0.0222	0.0220	1	44
22.000		0.9556	0.0444	0.0307	2	43
23.000		0.9333	0.0667	0.0372	3	42
25.000		0.9111	0.0889	0.0424	4	41
⋮						
165.000	*	.	.	.	24	2
182.000	*	.	.	.	24	1
189.000	*	.	.	.	24	0

Note: The marked survival times are censored observations.

Plot of the survival function



The survival curve is not consistent: $\widehat{S}(t)$ doesn't decrease to zero because the largest observed time is right-censored.

Breastfeeding duration

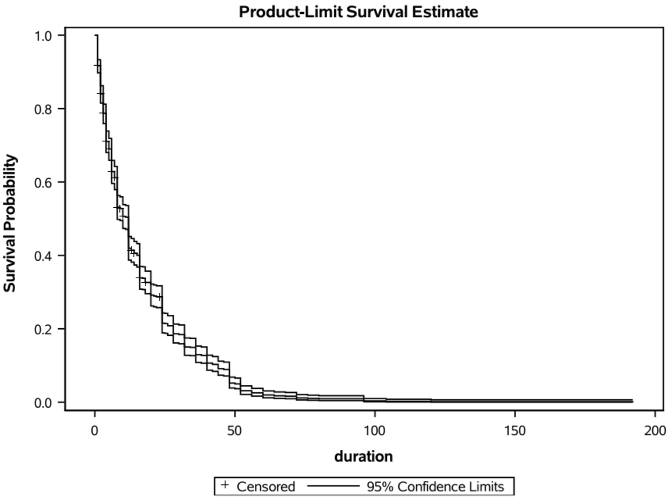
The `breastfeeding` data from the National Longitudinal Survey of Youthcontains information on the time until which mothers stop breastfeeding from birth. We focus on the following explanatories:

- `duration`: duration of breast feeding (in weeks)
- `delta`: indicator for completed breastfeeding
  - yes (1)
  - right-censored (0)

Summary of the Number of Censored and Uncensored Values

Total	Failed	Censored	Percent Censored
927	892	35	3.78

Survival curve for `breastfeeding` data



$\widehat{S}(t)$ reaches zero because the largest survival time is observed, not censored.

Median survival time

The median survival time is the time t_M such that $S(t_m) = 0.5$.

- That is, the median time t_M is such that 50% of people have survived until time t_M.

We can easily find this estimated median time by seeing where the horizontal line $\widehat{S}(t) = 0.5$ intersects the survival curve.

Quartile Estimates

			95% Confidence Interval	
Percent	**Point Estimate**	**Transform**	**[Lower**	**Upper)**
75	.	LOGLOG	.	.
50	89.000	LOGLOG	66.000	.
25	51.000	LOGLOG	34.000	67.000

Mean survival time

For a continuous positive random variable, $T > 0$, it can be shown that

$$\mathsf{E}\left(T\right) = \int_0^\infty S(t)\mathrm{d}t$$

We can estimate the expected survival time $\mathsf{E}\left(T\right)$ simply by calculating the area under the survivor curve $\widehat{S}(t)$.

- For example, the mean survival time for the breastfeeding data is 16.89 weeks with standard error 0.614 weeks.
- If the largest recorded survival time is censored, the estimated survival curve $\widehat{S}(t)$ will plateau and never reaches 0. The area under the curve is infinite!
- In this case, we can estimate instead the restricted mean survival time: $\mathsf{E}\left(\min\{T, \tau\}\right)$ for a chosen τ. It amounts to calculating the average as if the curve dropped to 0 at time τ (`rmst` option in SAS).