Biostatistics (MATH11230)
Nonparametric estimators of the survival function

Vanda Inácio

University of Edinburgh

Semester 1, 2022/2023

Nonparametric procedures

General context

- An initial step when analysing survival or even times is to provide numerical or graphical summaries of the event times for subjects in a particular group.
- Such summaries may be of interest in their own right or as a preliminary step before a more detailed analysis of the event times is conducted.
- Event times are convenient summarised through estimates of the survival or hazard function.

Nonparametric procedures

Estimating the survival function: noncensored observations

- In the case of noncensoring, an obvious estimator of the survival function is the empirical estimator, given by

$$\widehat{S}(t) = \frac{\text{number of individuals with event times} > \text{t}}{\text{number of individuals in the dataset}} = \frac{\#\{j : t_j > t\}}{n},$$

  where $t_1, \ldots, t_n$ are the event times and n is the number of individuals in the dataset.

- Note that $\widehat{S}(t) = 1$ for values of t below the smallest event time and $\widehat{S}(t) = 0$ for values of t above the largest event time.

- Equivalently, $\widehat{S}(t) = 1 - \widehat{F}(t)$, where $\widehat{F}(t)$ is the empirical cumulative distribution function, that is,

$$\widehat{F}(t) = \frac{\#\{j : t_j \leq t\}}{n}.$$

Nonparametric procedures

Estimating the survival function: noncensored observations

↪ Let us consider the following event times (say, in months):

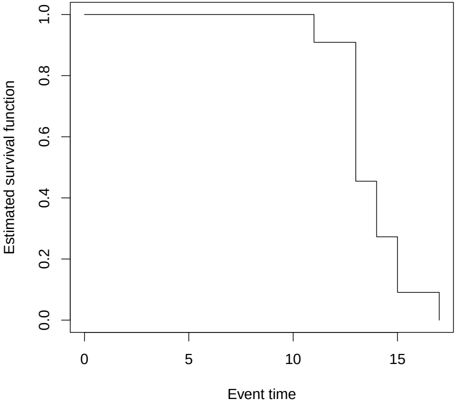
$$11 \quad 13 \quad 13 \quad 13 \quad 13 \quad 13 \quad 14 \quad 14 \quad 15 \quad 15 \quad 17$$

↪ We have that:

$$\widehat{S}(11) = \frac{\#\{j : t_j > 11\}}{11} = \frac{10}{11},$$
$$\widehat{S}(13) = \frac{\#\{j : t_j > 13\}}{11} = \frac{5}{11},$$
$$\widehat{S}(14) = \frac{\#\{j : t_j > 14\}}{11} = \frac{3}{11},$$
$$\widehat{S}(15) = \frac{\#\{j : t_j > 15\}}{11} = \frac{1}{11},$$
$$\widehat{S}(17) = \frac{\#\{j : t_j > 17\}}{11} = \frac{0}{11}.$$

Nonparametric procedures

Estimating the survival function: noncensored observations



Nonparametric procedures

Estimating the survival function: noncensored observations

- ↪ The empirical estimator of the survival function cannot be used when there are censored observations as it actually disregards the information provided by censored observations.
- ↪ Nonparametric estimators of the survival function that take into account the partial information available from the censored observations have been proposed.
- ↪ The two most commonly used nonparametric estimators for right-censored data are the Kaplan–Meier estimator of the survival function and the Nelson–Aalen estimator of the cumulative hazard function.
- ↪ Both estimators allow to make inferences about the distribution of the true event times based on the available information (observed event times and censoring status).

Nonparametric procedures

Estimating the survival function: Kaplan–Meier estimator

- Let us start by introducing some notation.
- Let $0 = t_0 < t_1 < t_2 < \ldots < t_J < t_{J+1} = \infty$ denote the unique uncensored event times, with $d_1, \ldots, d_J$ the corresponding number of events at time j, $j = 1, \ldots, J$.
- Further, let $n_1, \ldots, n_J$ be the size of the risk set at each event time, i.e., n_j is the number of individuals still event free just before t_j.

Nonparametric procedures

Estimating the survival function: Kaplan–Meier estimator

- By the law of total probability, we have that

$$\Pr(T > t_j) = \Pr(T > t_j \mid T > t_{j-1})\Pr(T > t_{j-1}) + \Pr(T > t_j \mid T \leq t_{j-1})\Pr(T \leq t_{j-1}).$$

- The fact that $t_{j-1} < t_j$ implies that

$$\Pr(T > t_j \mid T \leq t_{j-1}) = 0,$$

as it is impossible for an individual to survive past t_j if he or she did not survive an earlier time t_{j-1}.

- Therefore,

$$S(t_j) = \Pr(T > t_j) = \Pr(T > t_j \mid T > t_{j-1})\Pr(T > t_{j-1}).$$

Nonparametric procedures

Estimating the survival function: Kaplan–Meier estimator

- But, by definition of survival function, we have that

$$\Pr(T > t_{j-1}) = S(t_{j-1}),$$

and thus

$$S(t_j) = \Pr(T > t_j) = \Pr(T > t_j \mid T > t_{j-1})S(t_{j-1}).$$

- It also holds that

$$S(t_{j-1}) = \Pr(T > t_{j-1} \mid T > t_{j-2})S(t_{j-2}),$$

and that

$$S(t_{j-2}) = \Pr(T > t_{j-2} \mid T > t_{j-3})S(t_{j-3}),$$

and that ....

- This implies that

$$S(t_j) = \Pr(T > t_j \mid T > t_{j-1}) \times \Pr(T > t_{j-1} \mid T > t_{j-2}) \times \ldots \times \Pr(T > t_2 \mid T > t_1)S(t_1). \tag{1}$$

Nonparametric procedures

Estimating the survival function: Kaplan–Meier estimator

- We must now simply plug-in estimates of each of the terms on the right-hand side of Equation (1).

- We have that

$$\begin{aligned}\widehat{\Pr}(T > t_j \mid T > t_{j-1}) &= 1 - \widehat{\Pr}(T \leq t_j \mid T > t_{j-1}) \\ &= 1 - \frac{\#\text{ number of events in } (t_{j-1}, t_j]}{\#\text{ number of individuals at risk at time } t_j} \\ &= 1 - \frac{d_j}{n_j} \\ &= \frac{n_j - d_j}{n_j}.\end{aligned}$$

- This leads to the Kaplan–Meier estimator of the survival curve

$$\widehat{S}^{\text{KM}}(t) = \prod_{j: t_j \leq t} \frac{n_j - d_j}{n_j}, \quad \text{for } t_j \leq t < t_{j+1},$$

with $\widehat{S}^{\text{KM}}(t) = 1$ for $t < t_1$.

Nonparametric procedures

Estimating the survival function: Kaplan–Meier estimator

- This estimator was originally proposed by Kaplan and Meier in 1958, hence the name Kaplan–Meier estimator.
- This estimator is also often referred to as the product limit estimator.
- This approach is undeniably the most used one to estimate and summarise survival curves.
- This method is so widespread that the original article is the most highly cited article in the history of statistics.

Nonparametric procedures

Estimating the survival function: Kaplan–Meier estimator

$\hookrightarrow$ To demonstrate the computation of $\widehat{S}^{\text{KM}}(t)$ we consider the following hypothetical dataset

Patient	1	2	3	4	5	6	7
Event time (in months)	1	3	3	6	8	9	10
Censoring status	1	1	1	0	0	1	0

$\hookrightarrow$ Here a censoring status equal to 1 means that the corresponding event time is not censored and 0 that it is censored.

$\hookrightarrow$ Let us construct the Kaplan–Meier estimate of the survival curve:

t_j	d_j	n_j	$\frac{n_j-d_j}{n_j}$	$\prod \frac{n_j-d_j}{n_j}$
1	1	7	$\frac{7-1}{7}=\frac{6}{7}$	$\frac{6}{7}$
3	2	6	$\frac{6-2}{6}=\frac{4}{6}$	$\frac{6}{7}\times\frac{4}{6}=\frac{4}{7}$
9	1	2	$\frac{2-1}{2}=\frac{1}{2}$	$\frac{1}{2}\times\frac{4}{7}=\frac{2}{7}$

Nonparametric procedures

Estimating the survival function: Kaplan–Meier estimator

- We thus have that

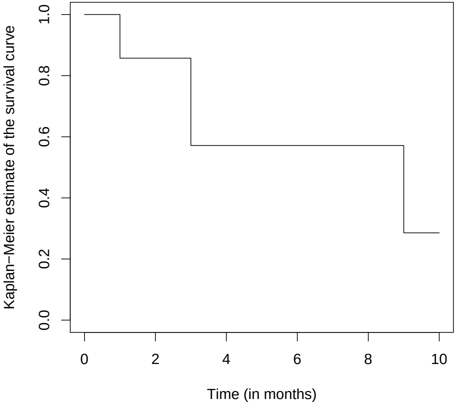
$$\widehat{S}^{\text{KM}}(t) = \begin{cases} 1, & t < 1 \\ \frac{6}{7}, & 1 \leq t < 3 \\ \frac{4}{7}, & 3 \leq t < 9 \\ \frac{2}{7} & 9 \leq t < 10 \end{cases}$$

- Note that the estimate of $\widehat{S}^{\text{KM}}(t)$ is undefined for $t > 10$ because the largest observation is a censored event time and $\widehat{S}^{\text{KM}}(t)$ cannot be estimated consistently beyond this time.

- On the other hand, if the largest event time is an uncensored observation, then $n_J = d_J$, and so $\widehat{S}^{\text{KM}}(t)$ is zero for $t \geq t_J$.

Nonparametric procedures

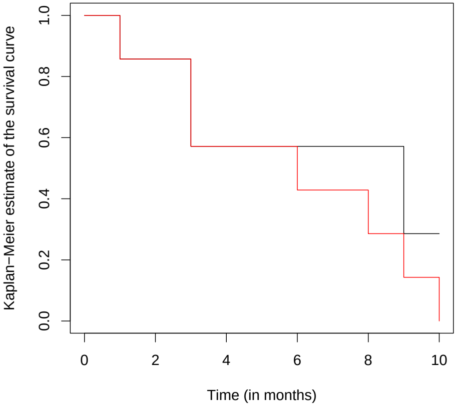
Estimating the survival function: Kaplan–Meier estimator



Nonparametric procedures

Estimating the survival function: Kaplan–Meier estimator

$\hookrightarrow$ Below in red is the empirical estimate of the survival function pretending that the censored event times were the true/observed ones



Nonparametric procedures

Estimating the survival function: Kaplan–Meier estimator

- As we could notice, the Kaplan–Meier estimate of the survival function is a step function, in which the estimated survival probabilities are constant between adjacent event times and decrease at each event time.

- If there are no censored observations in the dataset, the Kaplan–Meier estimator reduces to the empirical estimator of the survival function that we have seen at the beginning of the lecture.

Nonparametric procedures

Estimating the survival function: Kaplan–Meier estimator

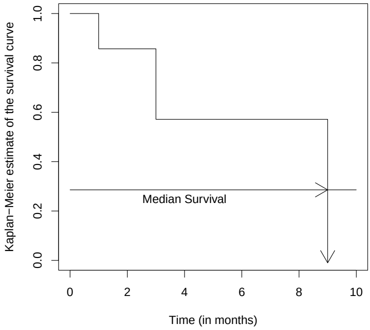
- $\hookrightarrow$ A key summary statistic of the survival function is the **median survival time**.
- $\hookrightarrow$ The median survival time is defined as the smallest time t such that $S(t) \leq 1/2$.
- $\hookrightarrow$ This can be estimated from the Kaplan–Meier plot by finding where the curve intersects the horizontal line $\widehat{S}^{\text{KM}}(t) = 1/2$.

Nonparametric procedures

Estimating the survival function: Kaplan–Meier estimator

- For the toy example in slides 12/13, the median survival time is 9, as this is the smallest time for which the survival curve is (equal) or below 0.5.

- We can get this easily in R as part of the output of the function that fits the Kaplan-Meier estimator (see more in the Supplementary Materials).



Nonparametric procedures

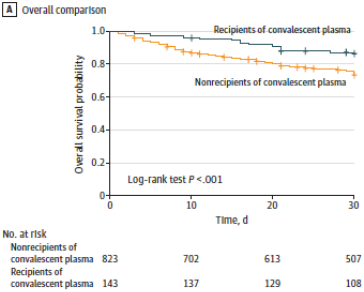
Kaplan–Meier estimator in recent scientific publications

Research

JAMA Oncology | Original Investigation

Association of Convalescent Plasma Therapy With Survival in Patients With Hematologic Cancers and COVID-19

Michael A. Thompson, MD, PhD; Jeffrey P. Henderson, MD, PhD; Pankil K. Shah, MD, MSPH; Samuel M. Rubinstein, MD; Michael J. Joyner, MD; Toni K. Choueiri, MD; Daniel B. Flora, MD, PharmD; Elizabeth A. Griffiths, MD; Anthony P. Gulati, MD; Clara Hwang, MD; Vadim S. Koshkin, MD; Esperanza B. Papadopoulos, MD; Elizabeth V. Robilotti, MD, MPH; Christopher T. Su, MD, MPH; Elizabeth M. Wulff-Burchfield, MD; Zhuoer Xie, MD, MS; Peter Paul Yu, MD; Sanjay Mishra, MS, PhD; Jonathon W. Senefeld, PhD; Dimpy P. Shah, MD, PhD; Jeremy L. Warner, MD, MS; for the COVID-19 and Cancer Consortium

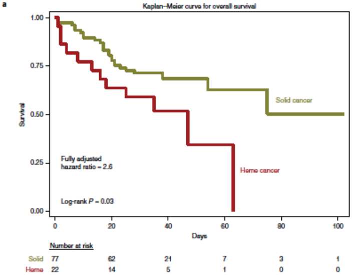


Nonparametric procedures

Kaplan–Meier estimator in recent scientific publications

ARTICLES
https://doi.org/10.1038/s41591-021-01386-7
nature medicine
Check for updates

CD8^{+} T cells contribute to survival in patients with COVID-19 and hematologic cancer



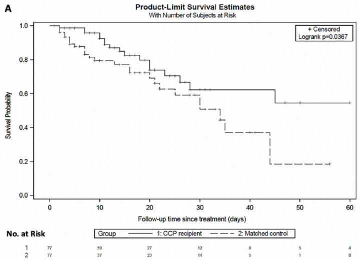
Nonparametric procedures

Kaplan–Meier estimator in recent scientific publications

RESEARCH ARTICLE

Early but not late convalescent plasma is associated with better survival in moderate-to-severe COVID-19

Neima Briggs[1‡], Michael V. Gormally[1‡], Fangyong Li[2], Sabrina L. Browning[3], Miriam M. Treggiari[4], Alyssa Morrison[5], Maudry Laurent-Rolle[6], Yanhong Deng[2], Jeanne E. Hendrickson[7,8], Christopher A. Tormey[7], Mahalia S. Desruisseaux[6]*



Nonparametric procedures

Estimating the survival function: Kaplan–Meier estimator

- The variance of the Kaplan–Meier estimator can be approximated by the so-called Greenwood formula (see, for example, Collett, 2014, chapter 2)

$$\widehat{\text{var}}(\widehat{S}^{\text{KM}}(t)) = \left[\widehat{S}^{\text{KM}}(t)\right]^2 \sum_{j;t_j \leq t} \frac{d_j}{n_j(n_j - d_j)}.$$

- For large samples, the following result holds

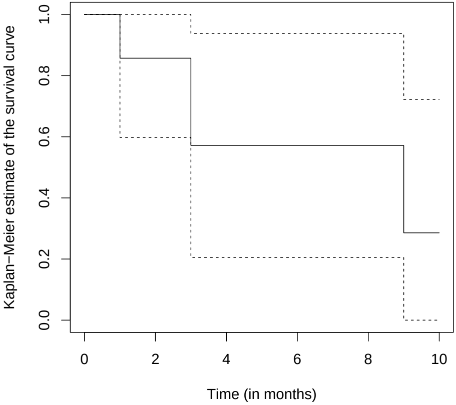
$$\frac{\widehat{S}^{\text{KM}}(t) - S(t)}{\sqrt{\widehat{\text{var}}(\widehat{S}^{\text{KM}}(t))}} \sim \text{N}(0, 1).$$

- This result can be used to derive a confidence interval for $S(t)$

$$\left(\widehat{S}^{\text{KM}}(t) - z_{\alpha/2}\sqrt{\widehat{\text{var}}(\widehat{S}^{\text{KM}}(t))}, \widehat{S}^{\text{KM}}(t) + z_{\alpha/2}\sqrt{\widehat{\text{var}}(\widehat{S}^{\text{KM}}(t))}\right).$$

Nonparametric procedures

Estimating the survival function: Kaplan–Meier estimator



- This CI is not accurate (may produce limits beyond the range of zero or one) when $\widehat{S}^{\mathrm{KM}}(t)$ is close to 0 or 1, so often CIs are first calculated for a transformation, for example, $\log(-\log S(t))$.

- For more details about this, I refer the interested reader to Collett, 2014, Chapter 2.

- The package `survival` implements both approaches. See more in the Supplementary Materials file.

Nonparametric procedures

Estimating the survival function: Nelson–Aalen estimator

- An alternative estimator of the survival function is based on the so called Nelson–Aalen estimator of the cumulative hazard function, proposed independently by Nelson and Aalen in the 70s.

- This estimator is given by
$$\widehat{H}(t) = \sum_{j:t_j \leq t} \frac{d_j}{n_j}.$$

- We can view this as equivalent to estimating the hazard function at each distinct event time t_j as the ratio of the number of event times at that time to the number of individuals still at risk at that time.

- The estimated cumulative hazard up to time t is just the sum of the estimated hazards at all event times up to t.

Nonparametric procedures

Estimating the survival function: Nelson–Aalen estimator

$\hookrightarrow$ From this estimator, one can obtain the Nelson–Aalen estimate of the survival function

$$\begin{aligned}
\widehat{S}^{\text{NA}}(t) &= \exp\{-\widehat{H}(t)\} \\
&= \exp\left\{-\sum_{j:t_j\leq t} \frac{d_j}{n_j}\right\} \\
&= \prod_{j:t_j\leq t} \exp\left\{-\frac{d_j}{n_j}\right\}.
\end{aligned}$$

Nonparametric procedures

Estimating the survival function: Nelson–Aalen estimator

- Interestingly, the Kaplan–Meier estimator of the survival function can actually be regarded as a first-order Taylor expansion approximation, around zero, of the Nelson–Aalen estimator.
- Recall that based on the first order Taylor expansion around zero, we can write

$$f(x) \approx f(0) + (x-0)f'(0).$$

- Letting $f(x) = e^{-x}$, we have that $e^{-x} \approx 1 - x$.
- Thus,

$$\widehat{S}^{\text{NA}}(t) \approx \prod_{j:t_j \leq t} \left(1 - \frac{d_j}{n_j}\right) = \widehat{S}^{\text{KM}}(t).$$

Nonparametric procedures

Estimating the survival function: Nelson–Aalen estimator

↪ For the hypothetical dataset in slide 12, we can also compute the Nelson–Aalen estimate of the survival curve.