Chapter 3

(AST405) Lifetime data analysis

Md Rasel Biswas

Lecture Outline

1. 3. Some Nonparametric and Graphical Procedures
   - 3.1 Introduction
   - 3.2 Non-parametric Estimation of a Survivor Function and Quantiles

Section 1

3. Some Nonparametric and Graphical Procedures

Subsection 1

3.1 Introduction

3.1 Introduction

- Graphs and simple data summaries are important for both description and analysis of data.
- They are closely related to nonparametric estimates of distributional characteristics; many graphs are just plots of some estimate.
- This chapter introduces nonparametric estimation and procedures for portraying univariate lifetime data.
- Tools such as frequency tables and histograms, empirical distribution functions, probability plots, and data density plots are familiar across different branches of statistics.
- For lifetime data, the presence of censoring makes it necessary to modify the standard methods.

3.1 Introduction

- To illustrate, let us consider one of the most elementary procedures in statistics, the formation of a relative-frequency table.
- Suppose we have a complete (i,e., uncensored) sample of n lifetimes from some population.
- Divide the time axis $[0, \infty)$ into $k+1$ intervals $l_j = [a_{j-1}, a_j), j = 1, \quad , k+1$, where $0 = a_0 < a_1 < \quad < a_k < a_{k+1} = \infty$; with a_k being the upper limit on observation.
- Let d_j be the observed number of lifetimes that lie in I_j.
- A frequency table is just a list of the intervals and their associated frequencies, d_j, or relative fequencies, d_j/n.
- A relative-frequency histogram, consisting of rectangles with bases on $[a_{j-1}, a_j\ )$ and areas $d_j/n (j = 1, \ldots, k)$; is often drawn to portray this.

3.1 Introduction

- When data are censored, however, it is generally not possible to form the frequency table, because if a lifetime is censored, we do not know which interval, I_j, it lies in. As a result, we cannot determine the d_j.
- Section 3.6 describes how to deal with frequency tables when data are censored; this is referred to as life table methodology.
- First, however, we develop methods for ungrouped data.
- Section 3.2 discusses nonparametric estimation of distribution, survivor, or cumulative hazard functions under right censoring.
  - This also forms the basis for descriptive and diagnostic plots, which are presented in Section 3.3.
- Sections 3.4 and 3.5 deal with the estimation of hazard functions and with nonparametric estimation from some other types of incomplete data.

Subsection 2

3.2 Non-parametric Estimation of a Survivor Function and Quantiles

3.2 Non-parametric Estimation of a Survivor Function and Quantiles

Recall: Parametric estimation of survivor function

This method assumes a parametric model (e.g., exponential distribution) of the data and we estimate the parameter first, then form the estimator of the survival function. In Parametric approach, we assume that we model the distribution as an exponential distribution with unknown parameter λ. Then we find an estimator of λ, which is $\hat{\lambda}$. Then we estimate the survival function using

$$\widehat{S}(t) = \hat{\lambda} e^{-\widehat{\lambda} t}$$

Non-parametric estimation of a survivor function

- As an example, consider the following sample of n complete observations

$$\{t'_1, \dots, t'_n\}$$

Non-parametric estimation of a survivor function

- *Empirical survivor function* (ESF) for a specific value $t > 0$ is defined as

$$\hat{S}(t) = \widehat{Pr}(T \geqslant t) = \frac{\text{number of observations} \geqslant t}{n} \tag{1}$$

  - $\hat{S}(t)$ is a step function that decreases by $(1/n)$ just after each observed lifetime if all observations are distinct
  - Generally, the ESF drops by (d/n) just past t if d lifetimes equal to t

- For a specific value $t > 0$, ESF can also be defined as

$$\hat{S}(t^+) = \widehat{Pr}(T > t) = \frac{\text{number of observations} > t}{n} \tag{2}$$

Non-parametric estimation of a survivor function

Acute myeloid leukemia (AML)

- AML patients who reached a remission status after the treatment of chemotherapy were randomly assigned to one of the two treatments
  - maintenance chemotherapy
  - no-maintenance chemotherapy (control group)
- Time of interest: Length of remission (in weeks)
  - *maintained*: 13, 161^+, 9, 13^+, 18, 28^+, 31, 23, 34, 45^+, 48
  - *control*: 5, 8, 12, 5, 30, 33, 8, 16^+, 23, 27, 43, 45

Does maintenance chemotherapy prolong the time until relapse?

Non-parametric estimation of a survivor function

- Estimate the survival function for the following sample of 11 complete observations of control group ($n = 11$)

5	5	8	8	12	23	27	30	33	43	45

Non-parametric estimation of a survivor function

- Estimates of survival function for $t = 0, 4, 5, 8, ...$

$$\hat{S}(0^+) = \widehat{Pr}(T > 0) = (11/11) = 1$$

$$\hat{S}(5^+) = \widehat{Pr}(T > 5) = (9/11) = 0.818$$

$$\hat{S}(8^+) = \widehat{Pr}(T > 8) = (7/11) = 0.636$$

$$\hat{S}(12^+) = \widehat{Pr}(T > 12) = (6/11) = 0.545 \text{ and so on}$$

  - Find $\hat{S}(9)$ or $\hat{S}(9^+)$

Non-parametric estimation of a survivor function

Sorted lifetimes

$$5, 5, 8, 8, 12, 23, 27, 30, 33, 43, 45$$

- Estimated survivor function

$$\hat{S}(t_j^+) = \frac{r_j}{n} \tag{3}$$

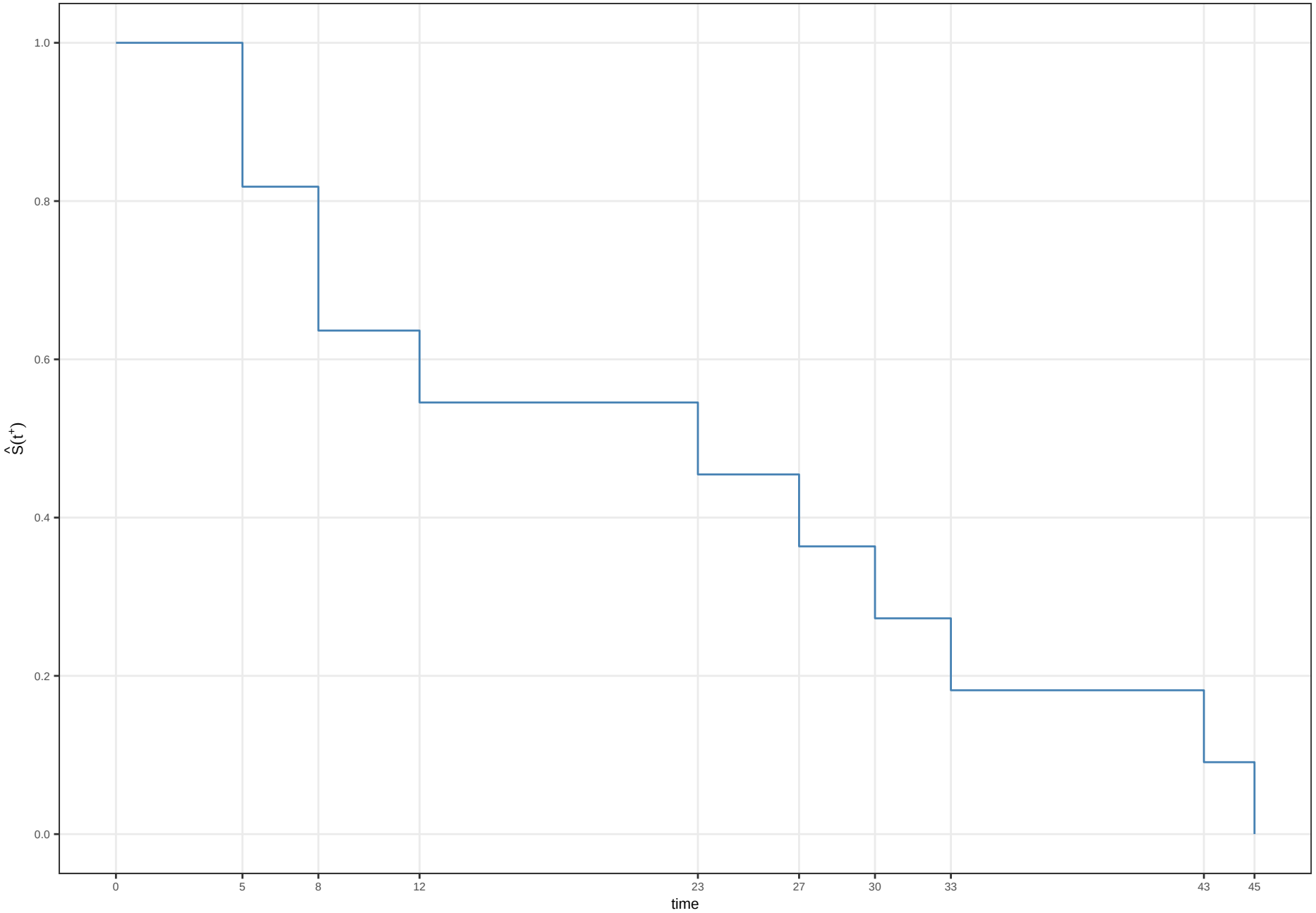
$$r_j = \sum_{i=1}^{n} I(t_i' > t_j) \rightarrow \text{ number of observations } > t_j$$

$$n \rightarrow \text{ total number of observations}$$

Non-parametric estimation of a survivor function

t_j	r_j	$\hat{S}(t_j^+)$
0	11	1.000
5	9	0.818
8	7	0.636
12	6	0.545
23	5	0.455
27	4	0.364
30	3	0.273
33	2	0.182
43	1	0.091
45	0	0.000

Non-parametric estimation of a survivor function



Non-parametric estimation of a survivor function

Exercise

The following are life times of 21 lung cancer patients receiving control treatment (with no censoring):

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

- Draw the ESF
- How would we estimate $\mathrm{S}(10)$, the probability that an individual survives to time 10 or later?

Non-parametric estimation of a survivor function

Let's get back to the AML example:

Sorted lifetimes: $5, 5, 8, 8, 12, 23, 27, 30, 33, 43, 45$

t_j	n_j	d_j	$\hat{S}(t_j^+)$
0	11	0	1.000
5	11	2	0.818
8	9	2	0.636
12	7	1	0.545
23	6	1	0.455
27	5	1	0.364
30	4	1	0.273
33	3	1	0.182
43	2	1	0.091
45	1	1	0.000

$n_j =$ number of subjects alive (or ar risk) just before time t_j

$d_j =$ number of subjects failed at time t_j

Non-parametric estimation of a survivor function

t_j	n_j	d_j	$\hat{p}_j$	$\hat{S}(t_j^+)$
0	11	0	1.000	1.000
5	11	2	0.818	0.818
8	9	2	0.778	0.636
12	7	1	0.857	0.545
23	6	1	0.833	0.455
27	5	1	0.800	0.364
30	4	1	0.750	0.273
33	3	1	0.667	0.182
43	2	1	0.500	0.091
45	1	1	0.000	0.000

$$\begin{aligned} \hat{p}_j &= \widehat{Pr}(T > t_j \,|\, T \geq t_j) \\ &= 1 - \frac{d_j}{n_j} \end{aligned} \tag{4}$$

Non-parametric estimation of a survivor function

Relationship between $\hat{p}_j$ and $\hat{S}(t_j^+)$

t_j	n_j	d_j	$\hat{p}_j$		$\hat{S}(t_j^+)$
0	11	0	1.000	1.000 =	1.000
5	11	2	0.818	1.000*0.818 =	0.818
8	9	2	0.778	1.000*0.818*0.778 =	0.636
12	7	1	0.857	' '	0.545
23	6	1	0.833	' '	0.455
27	5	1	0.800	' '	0.364
30	4	1	0.750	' '	0.273
33	3	1	0.667	' '	0.182
43	2	1	0.500	' '	0.091
45	1	1	0.000	' '	0.000

Non-parametric estimation of a survivor function

- Sorted unique lifetimes

$$5,\ 8,\ 12,\ 23,\ 27,\ 30,\ 33,\ 43,\ 45$$

$$
\begin{aligned}
P(T > 8) &= P(T > 8 \,|\, T \geq 8) P(T \geq 8) && (5)\\
&= P(T > 8 \,|\, T \geq 8) P(T > 5) && (6)\\
&= P(T > 8 \,|\, T \geq 8) P(T > 5 \,|\, T \geq 5) P(T \geq 5) && (7)\\
&= P(T > 8 \,|\, T \geq 8) P(T > 5 \,|\, T \geq 5) P(T \geq 0) && (8)\\
&= 0.778 \times 0.818 \times 1.0 = 0.636 && (9)
\end{aligned}
$$

Non-parametric estimation of a survivor function

- Sorted unique lifetimes

$$5,\ 8,\ 12,\ 23,\ 27,\ 30,\ 33,\ 43,\ 45$$

$$\begin{aligned}
P(T > 10) &= P(T > 10 \,|\, T \geq 10) P(T \geq 10) \\
&= P(T > 10 \,|\, T \geq 10) P(T > 8) \\
&= P(T > 10 \,|\, T \geq 10) P(T > 8 \,|\, T \geq 8) P(T \geq 8) \\
&= P(T > 10 \,|\, T \geq 10) P(T > 8 \,|\, T \geq 8) P(T > 5 \,|\, T \geq 5) P(T \geq \\
&= 1.0 \times 0.778 \times 0.818 \times 1.0 = 0.636
\end{aligned}$$

Non-parametric estimation of a survivor function

t_j	n_j	d_j	$\hat{p}_j$		$\hat{S}(t^+)$	I_j
0	11	0	1.000	1.000 =	1.000	[0, 5)
5	11	2	0.818	1.000*0.818 =	0.818	[5, 8)
8	9	2	0.778	1.000*0.818*0.778 =	0.636	[8, 12)
12	7	1	0.857	''	0.545	[12, 23)
23	6	1	0.833	''	0.455	[23, 27)
27	5	1	0.800	''	0.364	[27, 30)
30	4	1	0.750	''	0.273	[30, 33)
33	3	1	0.667	''	0.182	[33, 43)
43	2	1	0.500	''	0.091	[43, 45)
45	1	1	0.000	''	0.000	[45, Inf)

Non-parametric estimation of a survivor function

Notations:

- Observed times: $t'_1, t'_2, \ldots, t'_n$
- Ordered observed unique time points: $t_1 < t_2 < \cdots < t_k$
- Intervals

$$
\begin{aligned}
I_1 &= [t_1, t_2) \\
I_2 &= [t_2, t_3) \\
I_3 &= [t_3, t_4) \\
\ldots &\qquad \ldots \\
I_k &= [t_k, \infty)
\end{aligned}
$$

- Intervals are constructed so that each of which starts at an observed lifetime and ends just before the next observed lifetime
  - E.g. $I_j = [t_j, t_{j+1})$

Non-parametric estimation of a survivor function

- Sorted unique lifetimes

$$5,\ 8,\ 12,\ 23,\ 27,\ 30,\ 33,\ 43,\ 45$$

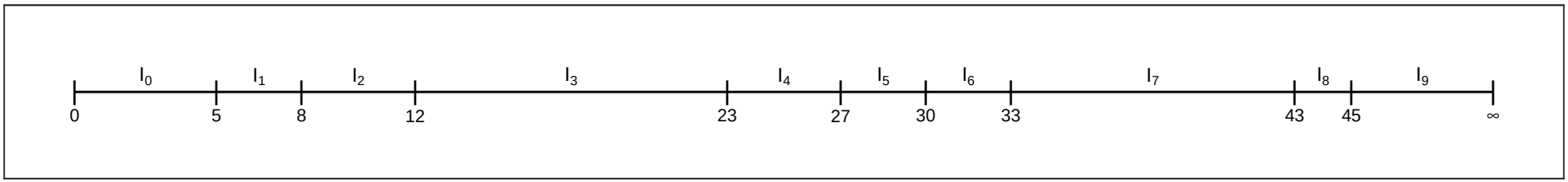


Figure 2

Non-parametric estimation of a survivor function

- Expressing $\hat{S}(t)$ in terms of $\hat{p}$

$$\hat{S}(t^+) = \widehat{Pr}(T > t) = \prod_{t_j \leqslant t} \hat{p}_j \tag{10}$$

$$\hat{S}(t) = \widehat{Pr}(T \geqslant t) = \prod_{t_j < t} \hat{p}_j \tag{11}$$

This method is known as **Kaplan-Meier** or Product-limit estimator of survivor function.

  - We saw that this method is equivalent to the ESF approach:

$$\hat{S}(t^+) = \widehat{Pr}(T > t) = \frac{\text{number of observations} > t}{n} \tag{12}$$

- But the advantage of Kaplan-Meier method is that **it can handle censored observations too**.

Censored sample

If we had censored data, then?

For the control group of AML example, now include the censored observation 16^+.

$$5, 8, 12, 5, 30, 33, 8, 16^+, 23, 27, 43, 45$$

- $\hat{S}(t) = ??$

Censored sample

- Censored sample: $5, 8, 12, 5, 30, 33, 8, 16^+, 23, 27, 43, 45$
- Sorted censored sample

$$5, 5, 8, 8, 12, 16^+, 23, 27, 30, 33, 43, 45$$

t_j	n_j	d_j
5	12	2
8	10	2
12	8	1
16	7	0
23	6	1
27	5	1
30	4	1
33	3	1
43	2	1
45	1	1

Censored sample

$$\hat{p}_j = \widehat{Pr}(T > t_j \,|\, T \geq t_j) = 1 - \frac{d_j}{n_j}$$

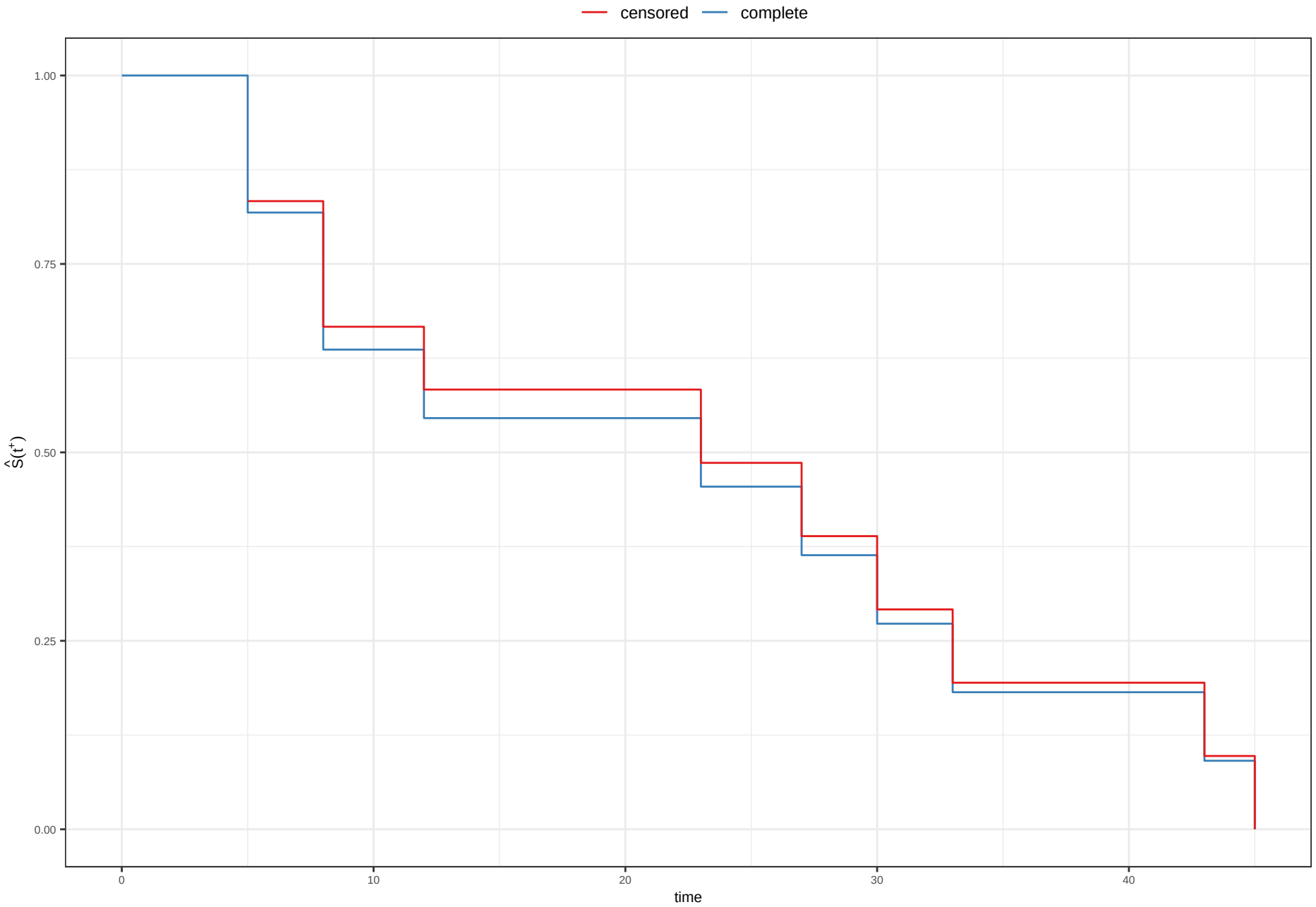
t_j	n_j	d_j	$\hat{p}_j$
5	12	2	0.833
8	10	2	0.800
12	8	1	0.875
16	7	0	1.000
23	6	1	0.833
27	5	1	0.800
30	4	1	0.750
33	3	1	0.667
43	2	1	0.500
45	1	1	0.000

Censored sample

$$\hat{S}(t^+) = \prod_{j:\, t_j \le t} \hat{p}_j$$

t_j	n_j	d_j	$\hat{p}_j$	$\hat{S}(t_j^+)$
5	12	2	0.833	0.833
8	10	2	0.800	0.667
12	8	1	0.875	0.583
16	7	0	1.000	0.583
23	6	1	0.833	0.486
27	5	1	0.800	0.389
30	4	1	0.750	0.292
33	3	1	0.667	0.194
43	2	1	0.500	0.097
45	1	1	0.000	0.000

Censored sample



Subsection 3

Kaplan-Meier estimator

Kaplan-Meier estimator

- Let (t_i', δ_i) be a censored random sample of lifetimes $i = 1, \ldots, n$
- Suppose that there are k $(k \leq n)$ distinct lifetimes at which deaths (event) occurs

$$t_1 < \cdots < t_k$$

Kaplan-Meier estimator

- Define for jth time $j = 1, \ldots, k$
  - $d_j = \sum_i I(t_i' = t_j, \delta_i = 1) = \sum_{i=1}^{n} dN_i(t_j) \rightarrow$ no. of deaths observed at t_j
  - $n_j = \sum_i I(t_i' \geq t_j) = \sum_{i=1}^{n} Y_i(t_j) \rightarrow$ no. of individuals at risk at time t_j, i.e. number of individuals alive an uncensored just prior time t_j

Kaplan-Meier estimator

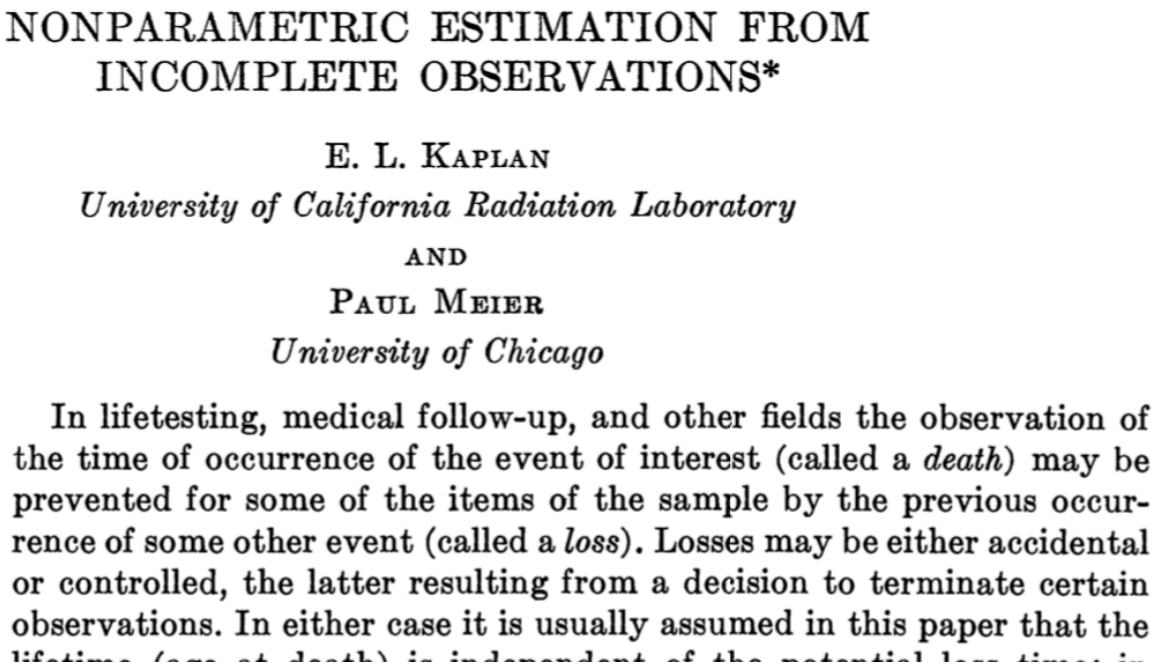
- A non-parametric estimator of survivor function $S(t)$

$$\hat{S}(t) = \prod_{j:\, t_j < t} \hat{p}_j = \prod_{j:\, t_j < t} \left(1 - \frac{d_j}{n_j}\right) = \prod_{j:\, t_j < t} \frac{n_j - d_j}{n_j} \tag{13}$$

- It is known as *Kaplan-Meier* (KM) or *Product-limit* (PL) estimator of survivor function (Kaplan and Meier 1958)
- Similarly

$$\hat{S}(t^+) = \prod_{j:\, t_j \leqslant t} \hat{p}_j \tag{14}$$

Kaplan-Meier estimator



NONPARAMETRIC ESTIMATION FROM INCOMPLETE OBSERVATIONS*

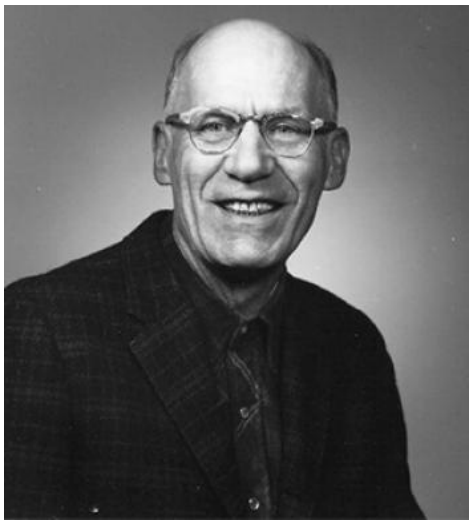
E. L. KAPLAN
University of California Radiation Laboratory
AND
PAUL MEIER
University of Chicago

In lifetesting, medical follow-up, and other fields the observation of the time of occurrence of the event of interest (called a *death*) may be prevented for some of the items of the sample by the previous occurrence of some other event (called a *loss*). Losses may be either accidental or controlled, the latter resulting from a decision to terminate certain observations. In either case it is usually assumed in this paper that the

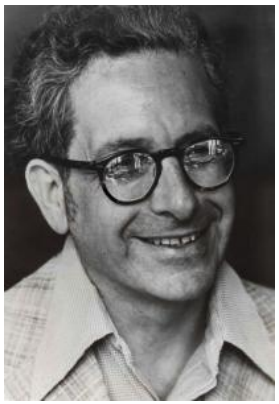
- The paper was published in *the Journal of American Statistical Association* in 1958
- Number of citations 66,345 (Google Scholar, 02 November 2024)

Kaplan-Meier estimator

Edward L Kaplan (1920–2006)



Paul Meier (1924–2011)



Subsection 4

Kaplan-Meier estimator as an MLE

PL estimator as an MLE

- Assume $T_1, \ldots, T_n$ have a discrete distribution with survivor function $S(t)$ and hazard function $h(t)$
- Without loss of generality, assume $t = 0, 1, 2, \ldots$

PL estimator as an MLE

- The general expression of likelihood function (from Eq. 2.2.12)

$$L = \prod_{t=0}^{\infty} \prod_{i=1}^{n} \left[h_i(t)\right]^{dN_i(t)} \left[1 - h_i(t)\right]^{Y_i(t)\left(1-dN_i(t)\right)} \tag{15}$$

  - $t_i \rightarrow$ lifetime of the ith individual
  - $\delta_i = I(t_i$ is a lifetime$)$
  - $Y_i(t) = I(t_i \geq t)$
  - $dN_i(t) = I(t_i = t, \delta_i = 1)$

PL estimator as an MLE

- Since $h_i(t) = h(t) \;\; \forall \; i$

$$L = \prod_{t=0}^{\infty} \prod_{i=1}^{n} \left[h(t)\right]^{dN_i(t)} \left[1 - h(t)\right]^{Y_i(t)\left(1 - dN_i(t)\right)}$$

$$= \prod_{t=0}^{\infty} \left[h(t)\right]^{d_t} \left[1 - h(t)\right]^{n_t - d_t} \tag{16}$$

  - $d_t = \sum_i dN_i(t) \rightarrow$ number of observed lifetimes equal to t, i.e. number of observed deaths at t
  - $n_t = \sum_i Y_i(t) \rightarrow$ number of subjects at risk (alive and uncensored) at time t

PL estimator as an MLE

- The parameters of the lifetime distribution

$$\mathbf{h} = \big(h(0), h(1), h(2), \dots\big)'$$

- The likelihood function

$$L(\mathbf{h}) = \prod_{t=0}^{\infty} \big[h(t)\big]^{d_t} \big[1 - h(t)\big]^{n_t - d_t} \tag{17}$$

- The log-likelihood function

$$\ell(\mathbf{h}) = \sum_{t=0}^{\infty} \Big\{ d_t \log h(t) + (n_t - d_t) \log(1 - h(t)) \Big\} \tag{18}$$

PL estimator as an MLE

- The MLE of $h(0)$

$$\left.\frac{\partial \ell(\mathbf{h})}{\partial h(0)}\right|_{h(0)=\hat{h}(0)} = 0 \tag{19}$$

- The score function evaluated at $\hat{\mathbf{h}}$

$$\frac{d_0}{\hat{h}(0)} = \frac{n_0 - d_0}{1 - \hat{h}(0)}$$

$$\hat{h}(0) = \frac{d_0}{n_0}$$

PL estimator as an MLE

- In general

$$\hat{h}(t) = \frac{d_t}{n_t}, \quad t = 0, 1, 2, \ldots, \tau \tag{20}$$

  - $\tau = \max\{t : n_t > 0\}$

PL estimator as an MLE

- The mle of $S(t)$

$$\hat{S}(t) = \prod_{s=0}^{t-1} \left(1 - \hat{h}(s)\right) = \prod_{s=0}^{t-1} \left(1 - \frac{d_s}{n_s}\right) \tag{21}$$

- If $d_\tau < n_\tau$ (which would happen if the largest observed lifetime is a censored observation) then $\hat{S}(\tau^+) > 0$ and undefined beyond τ^+,
- If $d_\tau = n_\tau$ then $S(\tau^+) = 0$ and $S(t) = 0$ for all $t > \tau$

Standard error of $\widehat{h}(t)$

- The rth diagonal element of the information matrix $\mathbf{I}(h)$

$$I_{rr}(\mathbf{h}) = E\left[\frac{-\partial^2 \ell}{\partial h(r)^2}\right] = E\left[\frac{d_r}{h(r)^2} + \frac{n_r - d_r}{(1-h(r))^2}\right]$$
$$= \frac{n_r}{h(r)\left[(1-h(r))\right]}$$

- Using the assumption

$$d_r \sim \text{Binomial}(n_r, h(r))$$

- Off diagonal elements of $I(\mathbf{h})$ are zero

Standard error of $\hat{h}(t)$

- The asymptotic variance of $\hat{h}(r)$

$$\widehat{\mathsf{Var}}(\hat{h}(r)) = I_{rr}^{-1}(\mathbf{h}) = \frac{\hat{h}(r)(1-\hat{h}(r))}{n_r} \tag{22}$$

Standard error of the PL estimator $\hat{S}(t)$

- Variance of $\log\left(\hat{S}(t)\right)$

$$\begin{aligned}
\widehat{\mathrm{Var}}\left[\log\left(\hat{S}(t)\right)\right] &= \sum_{s=0}^{t-1} \widehat{\mathrm{Var}}\left[\log(1-\hat{h}(s))\right] \\
&= \sum_{s=0}^{t-1} \frac{1}{(1-\hat{h}(s))^2} \widehat{\mathrm{Var}}\left[\hat{h}(s)\right] \\
&= \sum_{s=0}^{t-1} \frac{\hat{h}(s)\left[(1-\hat{h}(s))\right]^{-1}}{n_s}
\end{aligned} \tag{23}$$

  - Using the delta method $\mathrm{Var}(\hat{S})$ is obtained from $\mathrm{Var}(\log \hat{S})$

Standard error of the PL estimator $\hat{S}(t)$

- Using the delta method

$$\widehat{\text{Var}}\big[\log\big(\hat{S}(t)\big)\big] = \frac{1}{\hat{S}(t)^2}\,\widehat{\text{Var}}\big(\hat{S}(t)\big) \tag{24}$$

$$\begin{aligned}
\widehat{\text{Var}}\big(\hat{S}(t)\big) &= \hat{S}(t)^2\,\widehat{\text{Var}}\big[\log\big(\hat{S}(t)\big)\big] \\
&= \hat{S}(t)^2 \sum_{s=0}^{t-1} \frac{\hat{h}(s)\big[\big(1-\hat{h}(s)\big)\big]^{-1}}{n_s} \\
&= \hat{S}(t)^2 \sum_{s=0}^{t-1} \frac{d_s}{n_s(n_s-d_s)} \qquad (25)
\end{aligned}$$

- This formula of variance of PL estimator is known as the *Greenwood's formula*

Standard error of the PL estimator $\hat{S}(t)$

- Censored sample: $5, 8, 12, 5, 30, 33, 8, 16^{+}, 23, 27, 43, 45$

t_j	n_j	d_j	$\hat{S}(t_j^+)$
5	12	2	0.833
8	10	2	0.667
12	8	1	0.583
16	7	0	0.583
23	6	1	0.486
27	5	1	0.389
30	4	1	0.292
33	3	1	0.194
43	2	1	0.097
45	1	1	0.000

Standard error of the PL estimator $\hat{S}(t)$

$$\widehat{\text{Var}}(\hat{S}(t^+)) = [\hat{S}(t^+)]^2 \sum_{j:\, t_j \le t} \frac{d_j}{n_j(n_j - d_j)}$$

t_j	n_j	d_j	$\hat{S}(t_j^+)$	$\frac{d_j}{n_j(n_j-d_j)}$
5	12	2	0.833	0.017
8	10	2	0.667	0.025
12	8	1	0.583	0.018
16	7	0	0.583	0.000
23	6	1	0.486	0.033
27	5	1	0.389	0.050
30	4	1	0.292	0.083
33	3	1	0.194	0.167
43	2	1	0.097	0.500
45	1	1	0.000	Inf

Standard error of the PL estimator $\hat{S}(t)$

$$\widehat{\mathrm{Var}}(\hat{S}(t^+)) = [\hat{S}(t^+)]^2 \sum_{j:\, t_j \le t} \frac{d_j}{n_j(n_j - d_j)}$$

t_j	n_j	d_j	$\hat{S}(t_j^+)$	$\frac{d_j}{n_j(n_j-d_j)}$	$\sum_{j:t_j\le t} \frac{d_j}{n_j(n_j-d_j)}$
5	12	2	0.833	0.017	0.017
8	10	2	0.667	0.025	0.042
12	8	1	0.583	0.018	0.060
16	7	0	0.583	0.000	0.060
23	6	1	0.486	0.033	0.093
27	5	1	0.389	0.050	0.143
30	4	1	0.292	0.083	0.226
33	3	1	0.194	0.167	0.393
43	2	1	0.097	0.500	0.893
45	1	1	0.000	Inf	Inf

Standard error of the PL estimator $\hat{S}(t)$

$$\widehat{\text{Var}}(\hat{S}(t^+)) = [\hat{S}(t^+)]^2 \sum_{j:\, t_j \le t} \frac{d_j}{n_j(n_j - d_j)}$$

t_j	n_j	d_j	$\hat{S}(t_j^+)$	$\frac{d_j}{n_j(n_j-d_j)}$	$\sum_{j:t_j\le t} \frac{d_j}{n_j(n_j-d_j)}$	$\widehat{\text{Var}}(\hat{S}(t^+))$
5	12	2	0.833	0.017	0.017	0.012
8	10	2	0.667	0.025	0.042	0.019
12	8	1	0.583	0.018	0.060	0.020
16	7	0	0.583	0.000	0.060	0.020
23	6	1	0.486	0.033	0.093	0.022
27	5	1	0.389	0.050	0.143	0.022
30	4	1	0.292	0.083	0.226	0.019
33	3	1	0.194	0.167	0.393	0.015
43	2	1	0.097	0.500	0.893	0.008
45	1	1	0.000	Inf	Inf	NaN

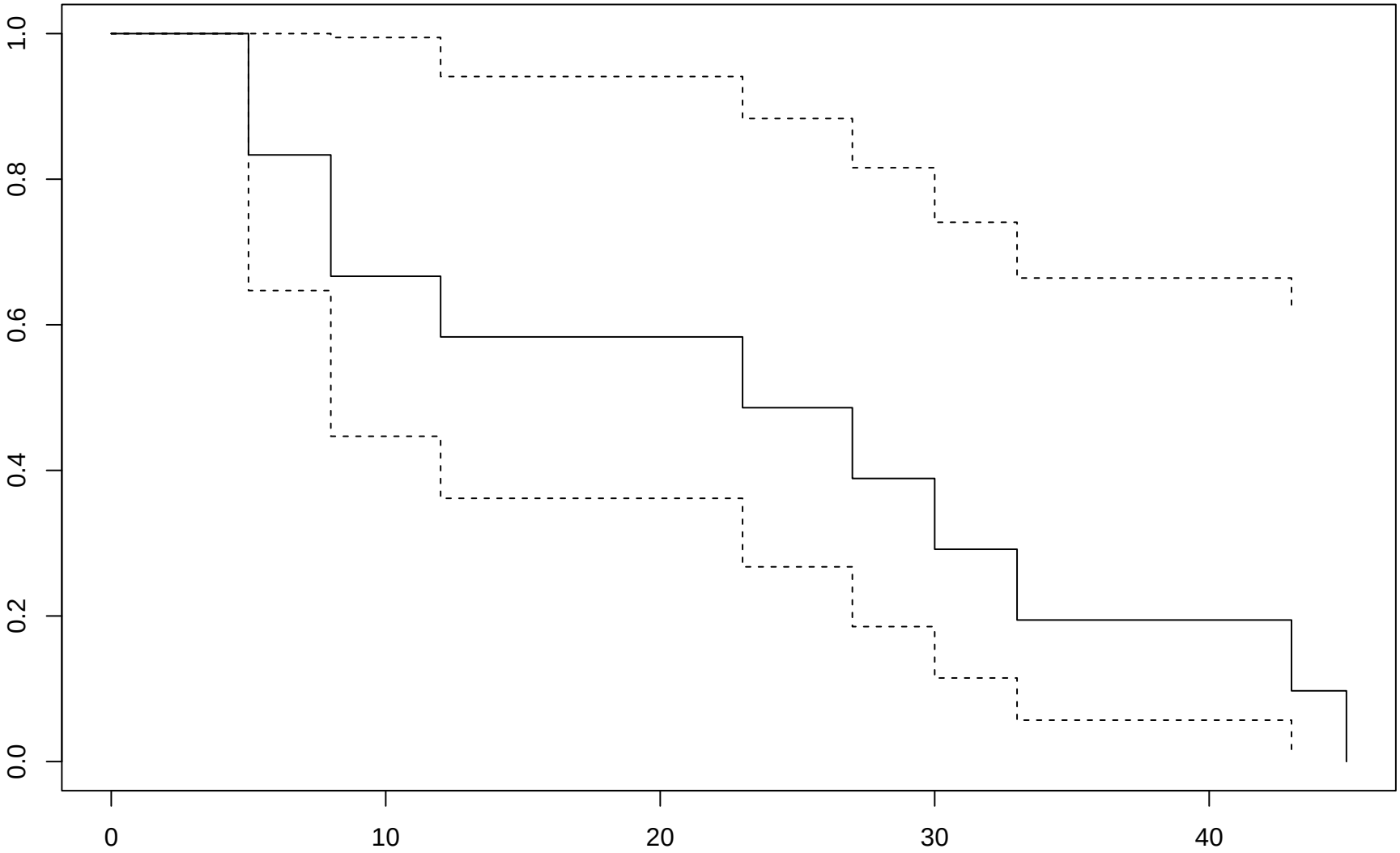
Subsection 5

`survival` package in R

survival package in R

```r
library(survival)
dat <- tibble(
  time = c(5, 8, 12, 5, 30, 33, 8, 16, 23, 27,  43, 45),
  status = c(rep(1, 7), 0, rep(1, 4))
)
surv_model <- survfit(Surv(time, status) ~ 1, data = dat)
```

survival package in R



survival package in R

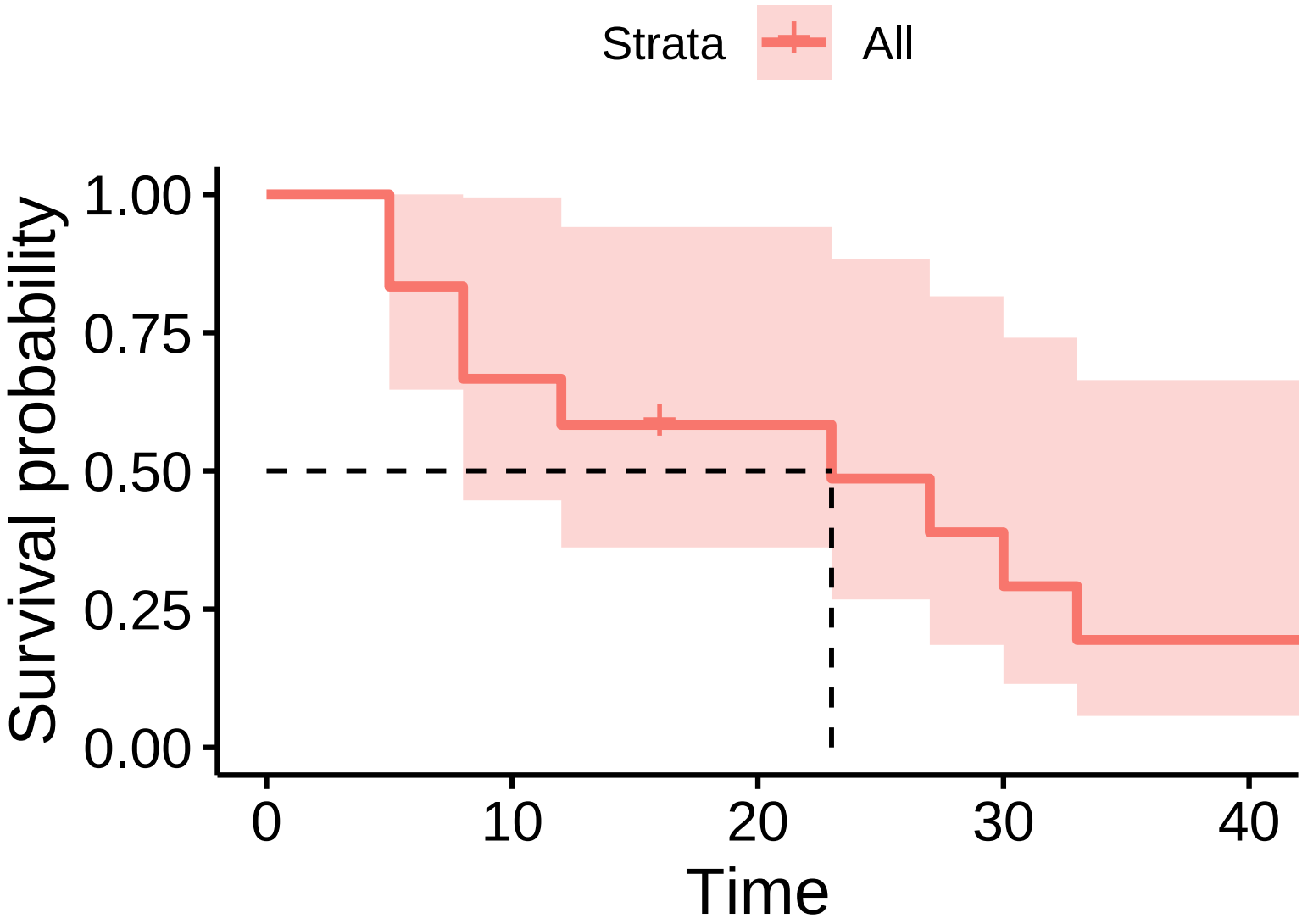
```
print(summary(surv_model), digits = 2)
```

```
Call: survfit(formula = Surv(time, status) ~ 1, data = dat)

 time n.risk n.event survival std.err lower 95% CI upper 95% C
    5     12       2    0.833   0.108        0.647          1.0
    8     10       2    0.667   0.136        0.447          0.9
   12      8       1    0.583   0.142        0.362          0.9
   23      6       1    0.486   0.148        0.268          0.8
   27      5       1    0.389   0.147        0.185          0.8
   30      4       1    0.292   0.139        0.115          0.7
   33      3       1    0.194   0.122        0.057          0.6
   43      2       1    0.097   0.092        0.015          0.6
   45      1       1    0.000     NaN           NA            N
```

survival package in R

```
survminer::ggsurvplot(surv_model, data = dat, surv.median.line
```



Subsection 6

Nelson-Aalen estimator

Estimator of $H(t)$

- Cumulative hazard function

$$H(t) = \int_0^t h(s)\, ds = \int_0^t dH(s) \tag{26}$$

  - $dH(t) \rightarrow$ increment of cumulative hazard function over $[t, t + dt)$

Nelson-Aalen estimator

- The following estimator of cumulative hazard function is known as Nelson-Aalen (NA) estimator (Nelson 1969; Aalen 1975)

$$\hat{H}(t) = \int_0^t d\hat{H}(s) = \int_0^t \frac{dN(s)}{Y(s)}, \tag{27}$$

  - $Y(s) > 0$ for $0 \leq s \leq t$

- In the notations used for Kaplan-Meier, NA estimator looks like

$$\hat{H}(t) = \sum_{j:\, t_j \leq t} \frac{d_j}{n_j} \tag{28}$$

Nelson-Aalen estimator

- The variance of $\hat{H}(t)$ can be obtained from the information matrix derived for the non-parametric likelihood function

$$\begin{aligned}\widehat{\text{Var}}\big[\hat{H}(t)\big] &= \sum_{j:\, t_j \le t} \text{Var}\Big(\frac{d_j}{n_j}\Big) \\ &= \sum_{j:\, t_j \le t} \Big(\frac{d_j}{n_j}\Big)\Big(1 - \frac{d_j}{n_j}\Big)\Big(\frac{1}{n_j}\Big) \\ &= \sum_{j:\, t_j \le t} \frac{d_j(n_j - d_j)}{n_j^3} \end{aligned} \tag{29}$$

Nelson-Aalen estimator

- Censored sample

$$5, 8, 12, 5, 30, 33, 8, 16^{+}, 23, 27, 43, 45$$

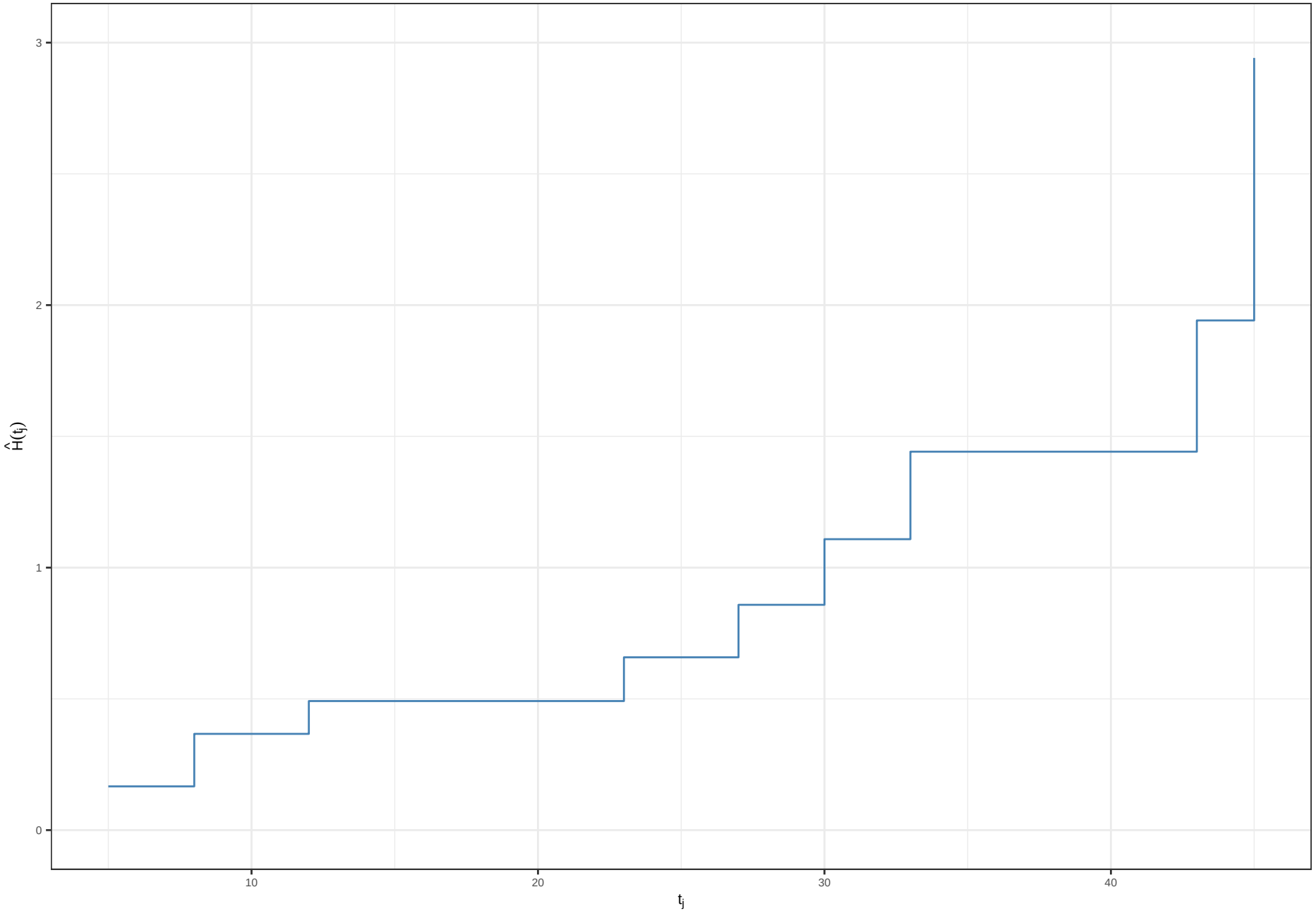
t_j	n_j	d_j
5	12	2
8	10	2
12	8	1
16	7	0
23	6	1
27	5	1
30	4	1
33	3	1
43	2	1
45	1	1

Nelson-Aalen estimator

$\hat{h}_j = P(T = t_j \,|\, T \geq t_j) = \frac{d_j}{n_j}$ and $\hat{H}(t) = \sum_{j:\, t_j \leq t} \hat{h}_j$

t_j	n_j	d_j	$\hat{h}_j$	$\hat{H}(t_j)$
5	12	2	0.167	0.167
8	10	2	0.200	0.367
12	8	1	0.125	0.492
16	7	0	0.000	0.492
23	6	1	0.167	0.658
27	5	1	0.200	0.858
30	4	1	0.250	1.108
33	3	1	0.333	1.442
43	2	1	0.500	1.942
45	1	1	1.000	2.942

Nelson-Aalen estimator



Nelson-Aalen estimator

$$\text{se}(\hat{H}(t)) = \sqrt{\sum_{j: t_j \leq t} \frac{d_j(n_j - d_j)}{n_j^3}}$$

t_j	n_j	d_j	$\hat{h}_j$	$\hat{H}(t_j)$	$se(\hat{H}(t_j))$
5	12	2	0.167	0.167	0.108
8	10	2	0.200	0.367	0.166
12	8	1	0.125	0.492	0.203
16	7	0	0.000	0.492	0.203
23	6	1	0.167	0.658	0.254
27	5	1	0.200	0.858	0.310
30	4	1	0.250	1.108	0.379
33	3	1	0.333	1.442	0.466
43	2	1	0.500	1.942	0.585
45	1	1	1.000	2.942	0.585

Nelson-Aalen estimator

- Both $\hat{S}(t)$ and $\hat{H}(t)$ are nonparametric m.l.e.'s, and are connected by the relationship between survivor and cumulative hazard function

$$S(t) = P(T \geqslant t) = \prod_{(0,t)} [1 - dH(u)]$$

$$S(t^+) = P(T > t) = \prod_{(0,t]} [1 - dH(u)]$$

- Note $\hat{S}(t)$ and $\hat{H}(t)$ are discrete and don't satisfy the relationship $H(t) = -\log S(t)$, which is true for the continuous distributions

$$\hat{S}_{NA}(t) = \exp\left(-\hat{H}(t)\right)$$

$$\hat{H}_{KM}(t) = -\log \hat{S}(t)$$

Subsection 7

Interval estimators of survival probabilities or quantiles

Interval estimators of survival probabilities or quantiles

- Nonparametric methods can also be used to construct confidence intervals for different lifetime distribution characteristics, such as
  - Survival probabilities $S(t)$
  - Quantiles t_p
- The methods of constructing confidence intervals are based on the following property of MLE $\hat{\theta}$

$$\sqrt{n}(\hat{\theta}-\theta) \sim N(0,\sigma^2) \;\Rightarrow\; (\hat{\theta}-\theta) \sim N(0, \text{var}(\hat{\theta}))$$

CIs for survival probabilities

- The PL estimator $\hat{S}(t)$ is an MLE of $S(t)$

$$\left(\hat{S}(t) - S(t)\right) \sim N(0, \sigma_s^2(t))$$

  - $\hat{\sigma}_s^2(t) = \widehat{\text{Var}}(\hat{S}(t)) \rightarrow$ Greenwood's variance estimator

- A pivotal quantity can be defined as

$$Z_1 = \frac{\hat{S}(t) - S(t)}{\hat{\sigma}_s(t)} \sim \mathcal{N}(0, 1) \tag{30}$$

CIs for survival probabilities

- The $100(1-\alpha)\%$ confidence interval for $S(t^+)$ can be obtained from the following expression

$$P(a \leq Z_1 \leq b) = 1 - \alpha$$

  - $b = -a = z_{(1-\alpha/2)}$
  - $z_p \rightarrow p$th quantile of the standard normal distribution, i.e. $P(Z < z_p) = p$

- $100(1-\alpha)\%$ confidence interval for $S(t^+)$

$$\hat{S}(t) \pm z_{(1-\alpha/2)}\, \hat{\sigma}_s(t) \tag{31}$$

CIs for survival probabilities

t_j	n_j	d_j	$\hat{S}(t_j^+)$	$\widehat{\text{Var}}(\hat{S}(t^+))$
5	12	2	0.833	0.012
8	10	2	0.667	0.019
12	8	1	0.583	0.020
16	7	0	0.583	0.020
23	6	1	0.486	0.022
27	5	1	0.389	0.022
30	4	1	0.292	0.019
33	3	1	0.194	0.015
43	2	1	0.097	0.008
45	1	1	0.000	NaN

- Find the 95% confidence interval of $S(15^+)$

$$\hat{S}(15^+) \pm z_{.975}\, \hat{\sigma}_s(15^+)$$

CIs for survival probabilities

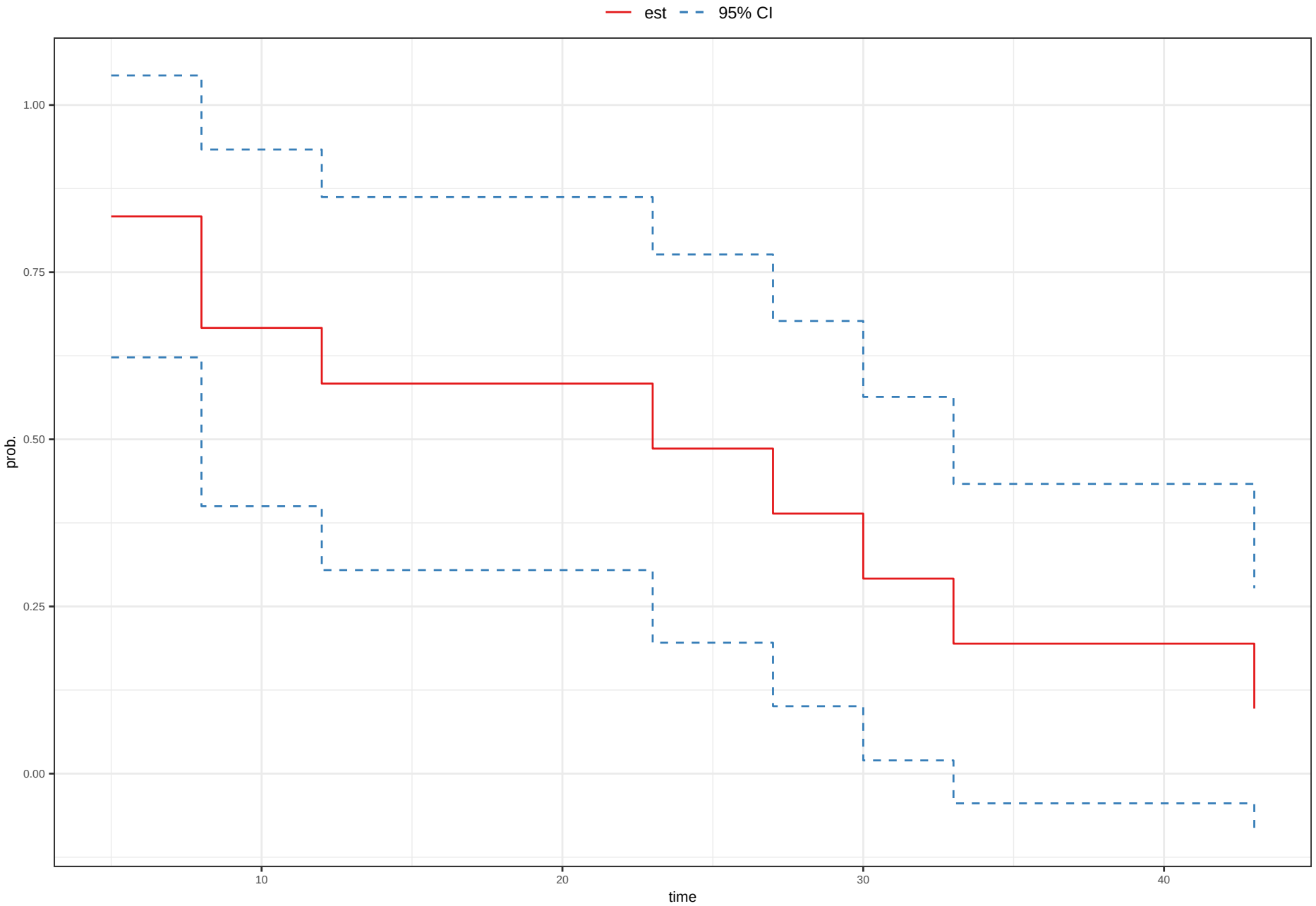
t_j	n_j	d_j	$\hat{S}(t_j^+)$	$\widehat{\text{Var}}(\hat{S}(t^+))$
5	12	2	0.833	0.012
8	10	2	0.667	0.019
12	8	1	0.583	0.020
16	7	0	0.583	0.020
23	6	1	0.486	0.022

- 95% confidence interval of $S(15^+)$

$$\begin{aligned}\hat{S}(15^+) \pm z_{.975}\,\hat{\sigma}_s(15^+) &= 0.583 \pm (1.960)(\sqrt{0.020}) \\ &= 0.583 \pm 0.279 \\ &= (0.304, 0.862)\end{aligned}$$

CIs for survival probabilities

CIs for survival probabilities



CIs for survival probabilities

- Z_1-based confidence interval

$$\hat{S}(t) \pm z_{(1-\alpha/2)}\, \hat{\sigma}_s(t) \tag{31}$$

- Limitations
  - When the number of uncensored lifetimes is small or when $S(t)$ is close to 0 or 1, the distribution of Z_1 may not be well approximated by $\mathcal{N}(0,1)$
  - The expression (31) may contain values outside of the interval $(0,1)$

CIs for survival probabilities

- Consider a function of $S(t)$ that takes values on $(-\infty, \infty)$

$$\psi(t) = g[S(t)] \tag{32}$$

- Examples of the function $g(\cdot)$, for $p \in (0, 1)$

$$g(p) = \begin{cases} \log(-\log(p)) \\ \log\left(\frac{p}{1-p}\right) \\ \log(p) \end{cases}$$

CIs for survival probabilities

- MLE of $\psi(t)$

$$\hat{\psi}(t) = g[\hat{S}(t)] \tag{33}$$

  - $\hat{S}(t) \rightarrow$ PL estimate of $S(t)$

- Asymptotic variance of $\hat{\psi}(t)$

$$\widehat{\text{Var}}[\hat{\psi}(t)] = \hat{\sigma}^2_{\psi}(t) = \{g'[\hat{S}(t)]\}^2 \widehat{\text{Var}}[\hat{S}(t)] \tag{34}$$

  - $g'(p) = \frac{dg(p)}{dp}$

CIs for survival probabilities

- We can define a pivotal quantity based on the distribution of the sampling distribution of $\hat{\psi}(t)$

$$Z_2 = \frac{\hat{\psi}(t) - \psi(t)}{\hat{\sigma}_{\psi}(t)} \tag{35}$$

  - Compare to Z_1, $Z_2 \sim \mathcal{N}(0, 1)$ is closer to standard normal distribution
  - Confidence intervals based on Z_2 are better performing compared to that of Z_1

CIs for survival probabilities

- Using the distribution of Z_2, confidence interval of $S(t)$ can be obtained in two steps
- Obtain the $(1-\alpha)\%$ CI of $\psi(t)$

$$\psi_L \leq \psi(t) \leq \psi_U \tag{36}$$

  - $\psi_L = \hat{\psi}(t) - z_{(1-\alpha/2)}\,\hat{\sigma}_{\psi}(t)$
  - $\psi_U = \hat{\psi}(t) + z_{(1-\alpha/2)}\,\hat{\sigma}_{\psi}(t)$

- Using inverse transformation, obtain the CI of $S(t)$ from that of $\psi(t)$

CIs for survival probabilities

- Using inverse transformation, obtain the CI of $S(t)$ from that of $\psi(t)$

$$\psi_L \leqslant \psi(t) \leqslant \psi_U$$
$$\psi_L \leqslant g\big[S(t)\big] \leqslant \psi_U$$
$$g^{-1}(\psi_L) \leqslant S(t) \leqslant g^{-1}(\psi_U)$$

  - $g^{-1}(\cdot) \rightarrow$ inverse function of $g(\cdot)$

Inverse functions

- Log function

$$g(p) = \log(p) = u \;\Rightarrow\; g^{-1}(u) = p = e^{u}$$

- Logit function

$$g(p) = \log\left(\frac{p}{1-p}\right) = u \;\Rightarrow\; g^{-1}(u) = p = \frac{\exp(u)}{1+\exp(u)}$$

- Log-log function

$$g(p) = \log\left(-\log(p)\right) = u \;\Rightarrow\; g^{-1}(u) = p = \exp\left(-e^{u}\right)$$

Inverse functions

- 95% CI of $\psi(t) = g[S(t)] = \log\left[-\log\left(S(t)\right)\right]$

$$\hat{\psi}(t) \pm \hat{\sigma}_{\psi}\, z_{.975}$$

- $\hat{\psi}(t) = \log\left[-\log\left(\hat{S}(t)\right)\right]$
- $\hat{\sigma}_{\psi}(t) = \sqrt{\left[\frac{1}{\hat{S}(t)\,\log\left(\hat{S}(t)\right)}\right]^2 \hat{\sigma}_S^2(t)}$

Inverse functions

t_j	n_j	d_j	$\hat{S}(t_j^+)$	$\widehat{\text{Var}}(\hat{S}(t^+))$
8	10	2	0.667	0.019
12	8	1	0.583	0.020
16	7	0	0.583	0.020

$$\hat{\psi}(15^+) = \log\left[-\log\left(\hat{S}(15)\right)\right] = \log\left[-\log\left(0.583\right)\right] = -0.618$$

$$\hat{\sigma}_{\psi}(15^+) = \sqrt{\left[\hat{S}(15)\,\log\left(\hat{S}(15)\right)\right]^{-2}\hat{\sigma}_S^2(15)}$$

$$= \sqrt{\left[(0.583)\log(0.583)\right]^{-2}(0.020)}$$

$$= 0.453$$

Inverse functions

- 95% CI of $\psi(15^+) = \log\left[-\log\left(S(15^+)\right)\right]$

$$\hat{\psi}(15^+) \pm \hat{\sigma}_{\psi}(15^+)\, z_{.975}$$

$$-0.618 \pm (0.453)(1.960)$$

$$-0.618 \pm 0.887$$

$$-1.505 \leqslant \psi(15^+) \leqslant 0.269$$

- $\hat{\psi}(15^+) = -0.618$
- $\hat{\sigma}_{\psi}(15^+) = 0.453$

Inverse functions

- 95% CI of $\psi(15^+) = \log\left[-\log\left(S(15^+)\right)\right]$

$$-1.505 \leqslant \psi(15^+) \leqslant 0.269$$

- 95% CI of $S(15^+)$

$$-1.505 \leqslant \psi(15) \leqslant 0.269$$

$$-1.505 \leqslant \log\Big(-\log\left[S(15^+)\right]\Big) \leqslant 0.269$$

$$-e^{0.269} \leqslant \log\left[S(15^+)\right] \leqslant -e^{-1.505}$$

$$\exp\left(-e^{0.269}\right) \leqslant S(15^+) \leqslant \exp\left(-e^{-1.505}\right)$$

$$0.270 \leqslant S(15^+) \leqslant 0.801$$

Inverse functions

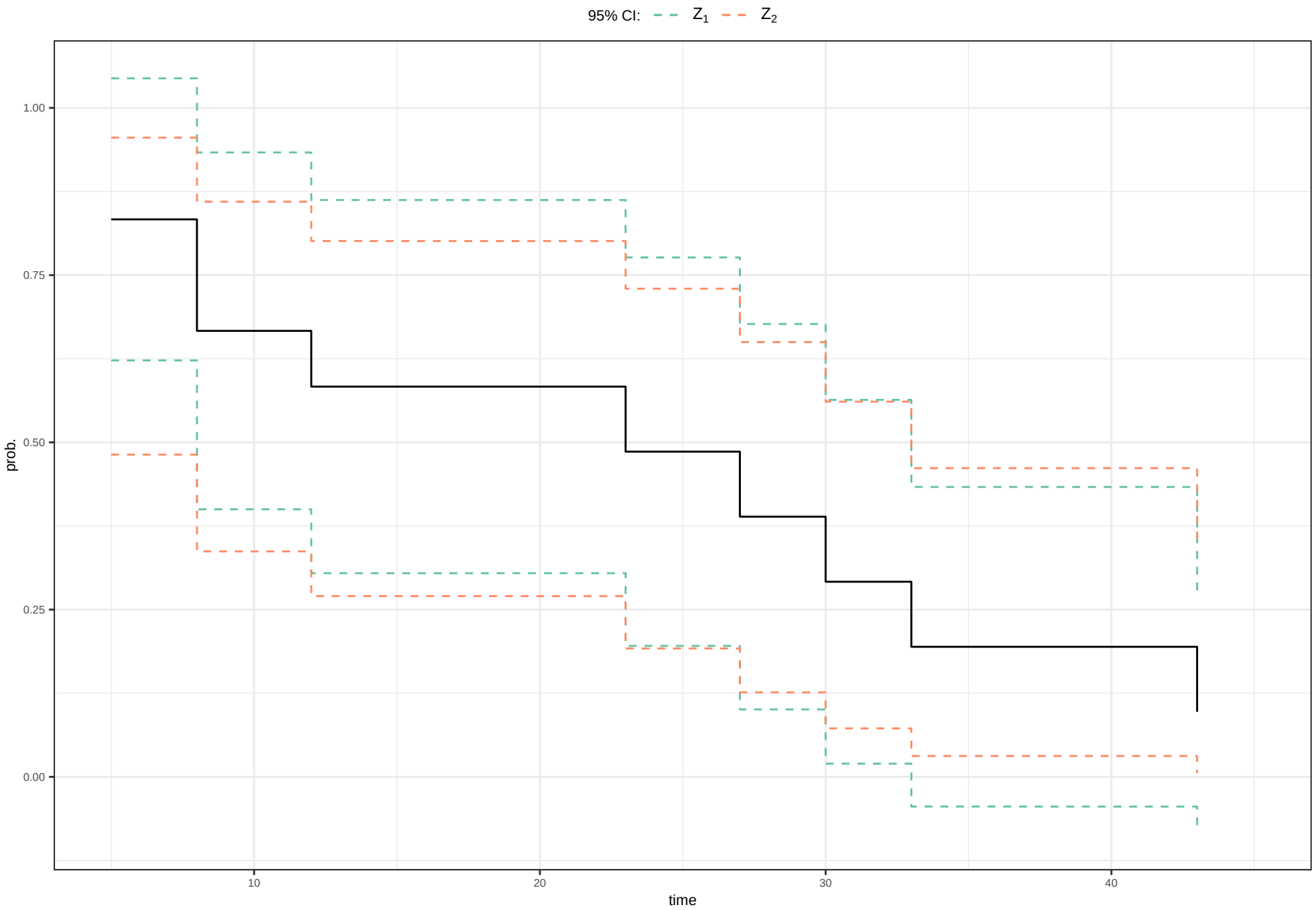
- 95% CI of $S(15^+)$
  - Using the distribution of Z_1

$$0.304 \leqslant S(15^+) \leqslant 0.862 \tag{37}$$

  - Using the distribution of Z_2

$$0.270 \leqslant S(15^+) \leqslant 0.801 \tag{38}$$

Inverse functions



Homework

- Obtain the 95% CI of $S(15^+)$ using the following transformations
  1. $g(p) = \log\left(\frac{p}{1-p}\right)$
  2. $g(p) = \log p$

Aalen, O. O. 1975. *Statistical Inference for a Family of Counting Processes*. Institute of Mathematical Statistics, University of Copenhagen. https://books.google.com.bd/books?id=sn1QAQAAMAAJ.

Kaplan, Edward L, and Paul Meier. 1958. "Nonparametric Estimation from Incomplete Observations." *Journal of the American Statistical Association* 53 (282): 457–81.

Nelson, Wayne. 1969. "Hazard Plotting for Incomplete Failure Data." *Journal of Quality Technology* 1 (1): 27–52.