Chapter 3

(AST305) Lifetime Data Analysis I

Md Rasel Biswas

Lecture Outline

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 $I_j = \left[a_{j-1}, a_j\right), j = 1, \ldots, k+1$, where $0 = a_0 < a_1 < \ldots < a_k < a_{k+1} = \infty$; 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 frequencies, d_j/n.
- A relative-frequency histogram, consisting of rectangles with bases on $\left[a_{j-1}, a_j\ \right)$ 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 Nonparametric Estimation of a Survivor Function and Quantiles

3.2 Nonparametric 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

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

Nonparametric estimation of a survivor function

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

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

Nonparametric 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

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

Nonparametric 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?

Nonparametric 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

Nonparametric 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 $\hat{S}(9^+)$

Nonparametric estimation of a survivor function

Sorted lifetimes

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

- Estimated survivor function

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

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

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

Nonparametric 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

Nonparametric estimation of a survivor function

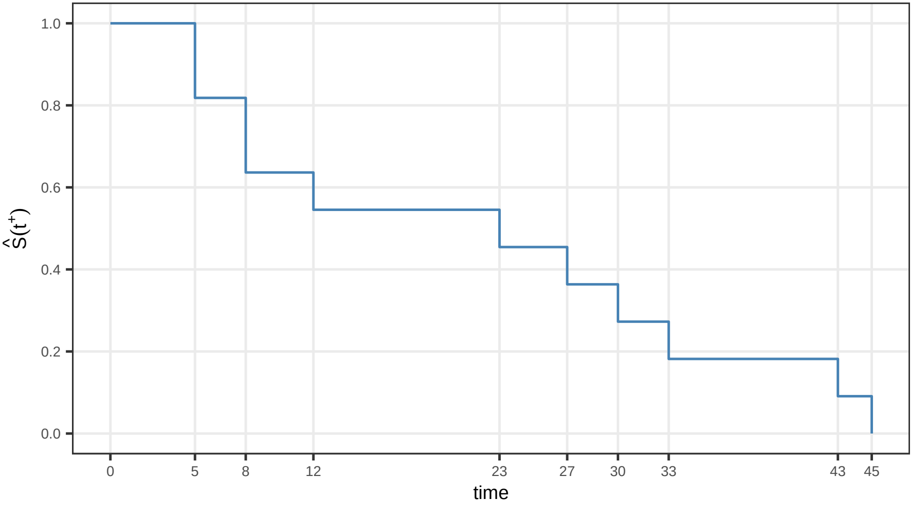


Figure 1: Nonparametric estimate of survivor function (Empirical Survivor Function - ESF)

Nonparametric estimation of a survivor function

Exercise

The following are lifetimes 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?

Nonparametric 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 at risk) just before time t_j

$d_j =$ number of subjects failed at time t_j

Nonparametric 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

$$\hat{p}_j = \widehat{Pr}(T > t_j \,|\, T \geq t_j)$$

$$= 1 - \frac{d_j}{n_j}$$

Nonparametric 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

Nonparametric estimation of a survivor function

- Sorted unique lifetimes

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

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

Nonparametric estimation of a survivor function

- Sorted unique lifetimes

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

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

Nonparametric 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)

Nonparametric estimation of a survivor function

Notations:

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

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

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

Nonparametric estimation of a survivor function

- Sorted unique lifetimes

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

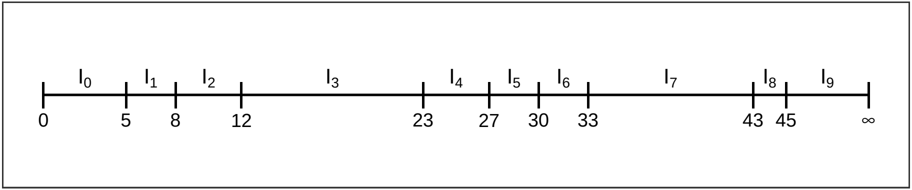


Figure 2

Nonparametric estimation of a survivor function

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

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

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

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:

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

- 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

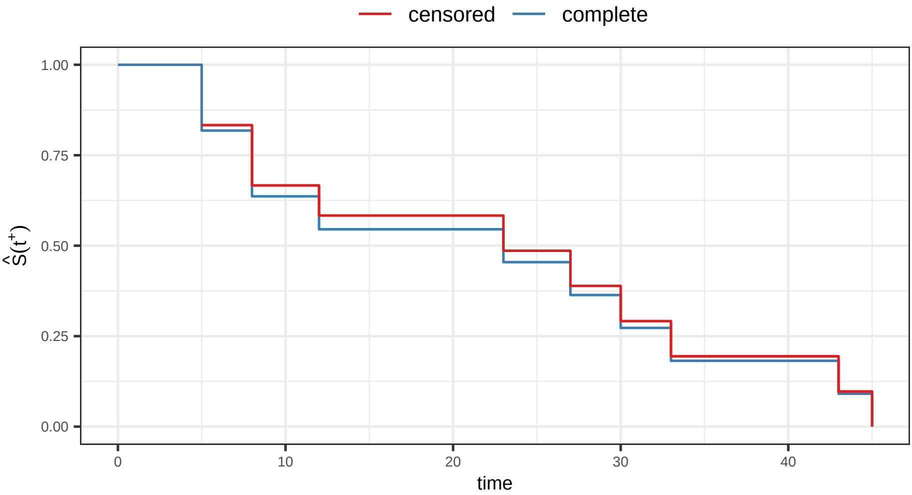


Figure 3: A comparison between estimated survivor function with and without censored observations

Subsection 3

Kaplan-Meier estimator

Kaplan-Meier estimator

- Let (t'_i, δ_i) be a censored random sample of lifetimes $i = 1, \dots, n$
- 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 $j = 1, \dots, k$
  - $d_j = \sum_i I(t_i' = t_j, \delta_i = 1) \rightarrow$ no. of deaths observed at t_j
  - $n_j = \sum_i I(t_i' \geq t_j) \rightarrow$ no. of individuals at risk at time t_j, i.e. number of individuals alive and uncensored just prior time t_j

Kaplan-Meier estimator

- A nonparametric estimator of survivor function $S(t)$

$$\hat{S}(t) = \prod_{j:\, t_j < t} \hat{p}_j = \prod_{j:\, t_j < t} \left(1 - \frac{d_j}{n_j}\right) = \prod_{j:\, t_j < t} \frac{n_j - d_j}{n_j} \qquad (4)$$

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

$$\hat{S}(t^+) = \prod_{j:\, t_j \leqslant t} \hat{p}_j \qquad (5)$$

Kaplan-Meier estimator

NONPARAMETRIC ESTIMATION FROM
INCOMPLETE OBSERVATIONS*

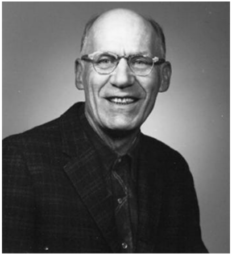
E. L. Kaplan
University of California Radiation Laboratory
AND
Paul Meier
University of Chicago

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

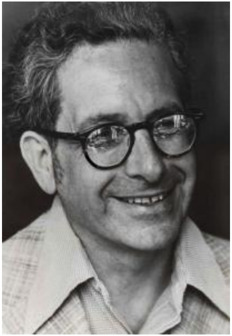
- The paper was published in *the Journal of the American Statistical Association* in 1958
- Number of citations 68,242 (Google Scholar, 27 September 2025)

Kaplan-Meier estimator

Edward L Kaplan (1920–2006)



Paul Meier (1924–2011)



Subsection 4

Kaplan–Meier estimator as an MLE

Kaplan–Meier estimator as an MLE

We only look at **event times** where at least one failure occurs.

- Let the ordered distinct event times be
  $t_1 < t_2 < \cdots < t_k$.
- At each event time t_j:
  - d_j = number of failures that happen exactly at t_j
  - n_j = number **at risk just before** t_j
    (alive and not yet censored, and have not failed earlier)
- **Discrete hazard at the event time**

$$h_j \;=\; P(T = t_j \mid T \geq t_j).$$

This is the conditional chance that someone who has survived up to t_j fails at t_j.

Likelihood at the event times

Given n_j people at risk at t_j, each either fails at t_j or survives past t_j. If the conditional probability of failing at t_j is h_j, then

$$d_j \sim \text{Binomial}(n_j,\, h_j).$$

Across the distinct event times the contributions multiply, so the likelihood is

$$L(\mathbf{h}) \;=\; \prod_{j=1}^{k} h_j^{\,d_j}\,(1-h_j)^{\,n_j-d_j},$$

where $\mathbf{h} = (h_1, \ldots, h_k)$.

MLE of the discrete hazards

Maximizing each binomial term separately gives

$$\hat{h}_j = \frac{d_j}{n_j}, \qquad j = 1, \ldots, k.$$

From hazards to the survivor function

The survivor function at a time t is the chance of getting **past** t. Between event times the survivor is constant. Just after t_j it equals the product of "survive this event time" factors:

$$\hat{S}(t) \;=\; \prod_{t_j<t}(1-\hat{h}_j) \;=\; \prod_{t_j<t}\Big(1-\frac{d_j}{n_j}\Big).$$

If you want the right-limit $\hat{S}(t^+)$ at or past t, include the factor at t as well:

$$\hat{S}(t^+) \;=\; \prod_{t_j\le t}\Big(1-\frac{d_j}{n_j}\Big).$$

These are exactly the **Kaplan–Meier (product-limit)** formulas.

Standard errors (Greenwood)

Greenwood Variance Formula

A widely used variance estimate for $\hat{S}(t^+)$ is

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

You can form Wald-type intervals for $S(t)$, or use a log–log transform for better behavior near 0 and 1.

Derivation

Variance of the KM estimator: Greenwood's formula

To understand the uncertainty in $\hat{S}(t^+)$, we derive an approximate variance estimator.
We start from the KM representation:

$$\hat{S}(t^+) \;=\; \prod \Big(1 - \frac{d_j}{\ldots}\Big).$$

Connection to continuous time

If lifetimes are continuous, failures occur at isolated points.
Treat those event times as above and define h_j as the **conditional jump** in failure probability at t_j.
Maximizing the same product of binomials yields the same KM estimator.

Back to the Example

- 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

Back to the Example

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

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

Back to the Example

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

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

Back to the Example

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

t_j	n_j	d_j	$\hat{S}(t_j^+)$	$\frac{d_j}{n_j(n_j-d_j)}$	$\sum_{j:t_j\le t} \frac{d_j}{n_j(n_j-d_j)}$	$\widehat{\mathsf{Var}}(\hat{S}(t^+))$
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

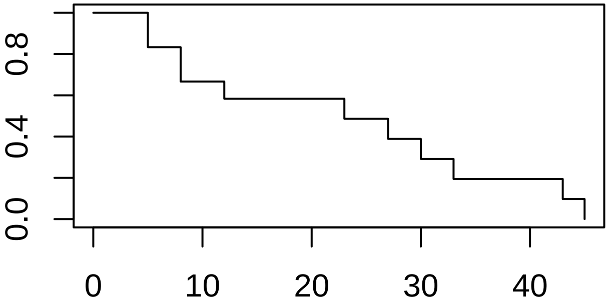
survival package in R

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

```
plot(surv_model, conf.int = FALSE)
```



survival package in R

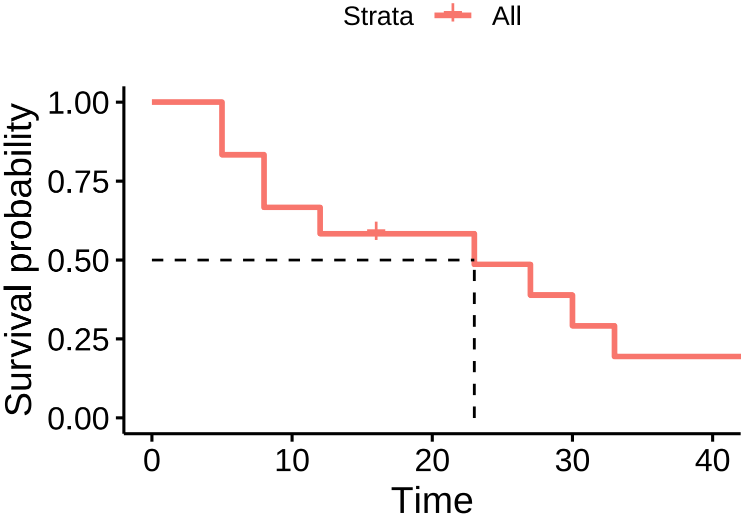
```r
print(summary(surv_model), digits = 2)
```

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

 time n.risk n.event survival std.err lower 95% CI upper 95% CI
    5     12       2    0.833   0.108        0.647         1.00
    8     10       2    0.667   0.136        0.447         0.99
   12      8       1    0.583   0.142        0.362         0.94
   23      6       1    0.486   0.148        0.268         0.88
   27      5       1    0.389   0.147        0.185         0.82
   30      4       1    0.292   0.139        0.115         0.74
   33      3       1    0.194   0.122        0.057         0.66
   43      2       1    0.097   0.092        0.015         0.62
   45      1       1    0.000     NaN           NA           NA
```

survival package in R

```r
survminer::ggsurvplot(surv_model, data = dat,
                      surv.median.line = "hv", conf.int = FALSE)
```



Subsection 6

Nelson-Aalen estimator

Cumulative hazard and its estimator

- Cumulative hazard:

$$H(t) = \int_0^t h(s)\, ds$$

- **Nelson-Aalen (NA) estimator** of $H(t)$ using the KM notations (event times t_j, risk set n_j, deaths d_j):

$$\hat{H}(t) = \sum_{t_j \le t} \frac{d_j}{n_j}, \qquad n_j > 0.$$

- A useful companion estimator of survival derived from NA is

$$\hat{S}_{\mathrm{NA}}(t) = \exp\{-\hat{H}(t)\}.$$

  In continuous-time settings with small event-time jumps, $\hat{S}_{\mathrm{NA}}(t)$ is often close to the KM curve.

Cumulative hazard and its estimator

- The variance of $\hat{H}(t)$ can be obtained as

$$\widehat{\mathsf{Var}}[\hat{H}(t)] = \sum_{j:\, t_j \le t} \mathsf{Var}\left(\frac{d_j}{n_j}\right) = \sum_{j:\, t_j \le t} \left(\frac{d_j}{n_j}\right)\left(1 - \frac{d_j}{n_j}\right)\left(\frac{1}{n_j}\right) = \sum_{j:\, t_j \le t} \frac{d_j(n_j - d_j)}{n_j^3} \quad (6)$$

Cumulative hazard and its 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

Cumulative hazard and its 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

Cumulative hazard and its estimator

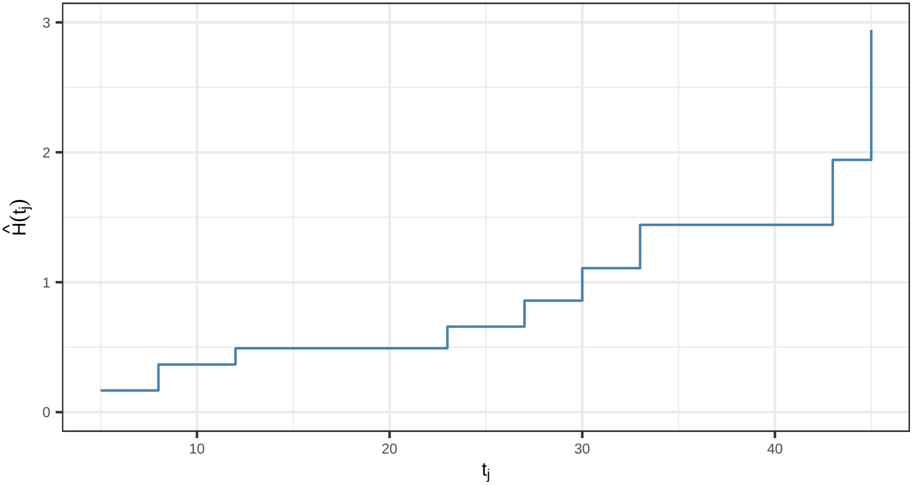


Figure 4: Nelson-Aalen estimator

Cumulative hazard and its estimator

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

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

Subsection 7

CIs for survival probabilities

CIs for survival probabilities

- 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 $\hat{\theta}$

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

Plain Wald CI (using Greenwood)

Let $\hat{S}(t^+)$ be the KM estimate and

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

Then a $(1-\alpha)100\%$ CI is

$$\hat{S}(t^+) \pm z_{1-\alpha/2} \sqrt{\widehat{\mathrm{Var}}(\hat{S}(t^+))}.$$

Plain Wald CI (using Greenwood)

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

Plain Wald CI (using Greenwood)

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

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

Drawback of Plain CI: for small samples or when $S(t)$ is near 0 or 1, Wald CIs may exceed $(0, 1)$.

Plain Wald CI (using Greenwood)

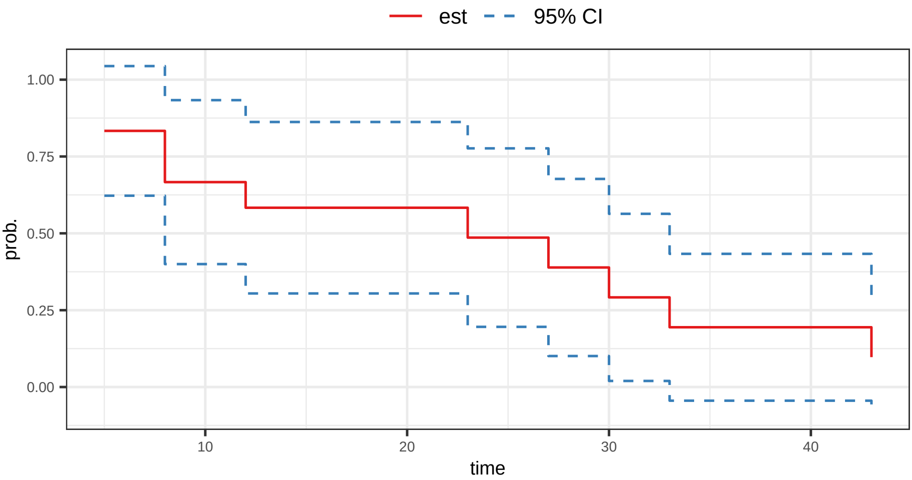


Figure 5: PL estimate and corresponding 95% CI of $S(15^+)$ using the censored sample

CI via variance-stabilizing transformations

- Let $\psi(t) = g\big[S(t)\big]$ where g maps $(0,1) \to \mathbb{R}$.

Common choices (with inverse g^{-1}):

- **Log–log:** $g(p) = \log[-\log p], \quad g^{-1}(u) = \exp(-e^u)$
- **Logit:** $g(p) = \log\big(p/(1-p)\big), \quad g^{-1}(u) = \exp(u)/[1+\exp(u)]$
- **Log:** $g(p) = \log p, \quad g^{-1}(u) = e^u$

CI via variance-stabilizing transformations

- Delta method:

$$\widehat{\text{Var}}(\hat{\psi}(t)) = \left\{g'(\hat{S}(t))\right\}^2 \widehat{\text{Var}}(\hat{S}(t)).$$

- Then $(1-\alpha)100\%$ CI for $\psi(t)$:

$$\hat{\psi}(t) \pm z_{1-\alpha/2}\sqrt{\widehat{\text{Var}}(\hat{\psi}(t))} \quad \Rightarrow \quad [\psi_L,\, \psi_U]$$

CI via variance-stabilizing transformations

- Finally, $(1-\alpha)100\%$ CI for $S(t)$ is $\left[g^{-1}(\psi_L),\ g^{-1}(\psi_U)\right]$ that stays inside $(0,1)$. That is,

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

Example of Log-log CI

- **95% CI of** $\psi(t) = g\big[S(t)\big] = \log\big[-\log\big(S(t)\big)\big]$ **is**

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

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

Example of Log-log CI

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

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

$$\begin{aligned}
\hat{\sigma}_{\psi}(15^+) &= \sqrt{\left[\hat{S}(15)\ \log\left(\hat{S}(15)\right)\right]^{-2}\hat{\sigma}_S^2(15)} \\
&= \sqrt{\left[(0.583)\log(0.583)\right]^{-2}(0.020)} \\
&= 0.453
\end{aligned}$$

Example of Log-log CI

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

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

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

$$-0.618 \pm 0.887$$

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

Example of Log-log CI

- **95% CI of $S(15^+)$ can be obtained as**

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

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

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

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

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

Example of Log-log CI

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

$$0.304 \leqslant S(15^+) \leqslant 0.862 \tag{7}$$

  - Log-log CI

$$0.270 \leqslant S(15^+) \leqslant 0.801 \tag{8}$$

Example of Log-log CI

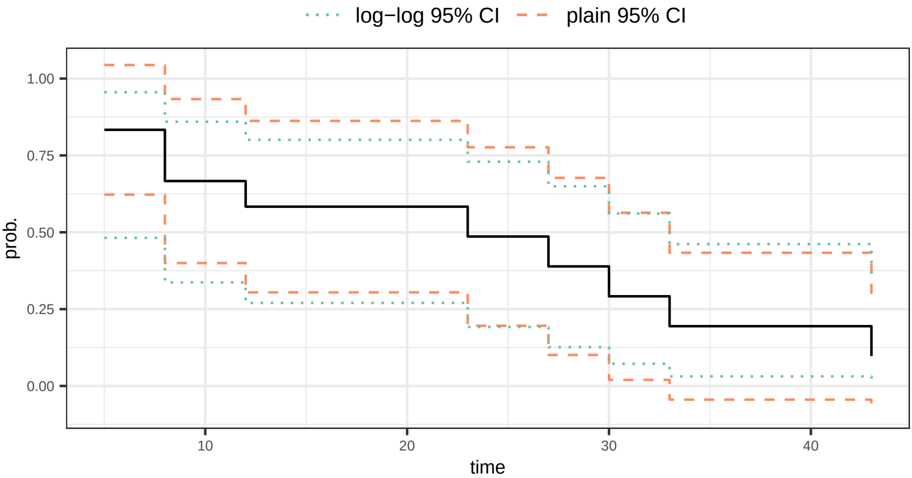


Figure 6: KM estimate and 95% CIs: Wald (plain) and log–log

Homework

Bootstrap CI

- The **nonparametric bootstrap** can be used to approximate the sampling distributions of the pivotal quantities Z_1 and Z_2.
- The $(\alpha/2)$- and $(1-\alpha/2)$-quantiles of the bootstrap distribution yield a $(1-\alpha)100\%$ confidence interval for $S(t)$.

Bootstrap CI

Steps for obtaining bootstrap CIs

0. Let the observed data be $\{(t_i, \delta_i),\, i = 1, \dots, n\}$.
   The Kaplan–Meier estimator $\hat{S}(t)$ and its standard error $\hat{\sigma}_s(t)$ are treated as MLEs of $S(t)$ and its SE.
1. **Resample:**
   Draw a bootstrap sample $\{(t_i^\star, \delta_i^\star),\, i = 1, \dots, n\}$ by sampling **with replacement** from the observed pairs $\{(t_i, \delta_i)\}$.
2. **Recompute estimates:**
   For each bootstrap sample, compute the product-limit estimate $\hat{S}^\star(t)$ and its SE $\hat{\sigma}_s^\star(t)$.
3. **Form pivotal statistic:**

$$Z_1^\star = \frac{\hat{S}^\star(t) - \hat{S}(t)}{\hat{\sigma}_s^\star(t)}.$$

4. **Repeat** Steps 1–3 a large number of times $(B > 1000)$ to obtain the bootstrap replicates

Bootstrap CI

5. **Quantiles of the bootstrap distribution:**
   Let $z^{\star}_{(qB)}$ denote the (qB)th smallest value of $\{Z^{\star}_{1,1}, \dots, Z^{\star}_{1,B}\}$.
   This estimates the q-quantile of the distribution of Z_1.

6. **Bootstrap confidence interval:**
   The $(q_2 - q_1)100\%$ CI of $S(t)$ is given by

$$\boxed{\hat{S}(t) - z^{\star}_{(q_2 B)}\,\hat{\sigma}_s(t) < S(t) < \hat{S}(t) - z^{\star}_{(q_1 B)}\,\hat{\sigma}_s(t)}$$

   For a 95% CI, use $q_2 = 0.975$ and $q_1 = 0.025$.

Bootstrap CI

Homework

- Obtain a **bootstrap confidence interval** for $S(15^+)$ using $B = 1000$ replications.
- Compare it to the intervals based on Z_1 (plain) and Z_2 (log–log) methods.

Example: Remission data

- The following data are on lengths of remission for two groups (placebo and 6-MP) of leukemia patients
- Objective was to examine whether the drug 6-MP is more effective than placebo

Table 1.3. Lengths of Remission (in weeks) for Two Groups of Patients[a]

6-MP	6, 6, 6, 6*, 7, 9*, 10, 10*, 11*, 13, 16, 17*, 19*, 20*, 22, 23, 25*, 32*, 32*, 34*, 35*
Placebo	1, 1, 2, 2, 3, 4, 4, 5, 5, 8, 8, 8, 8, 11, 11, 12, 12, 15, 17, 22, 23

[a] Stars denote censored observations.

Example: Remission data

Table 19: PL estimates and CIs (no transformation) for 6-MP group

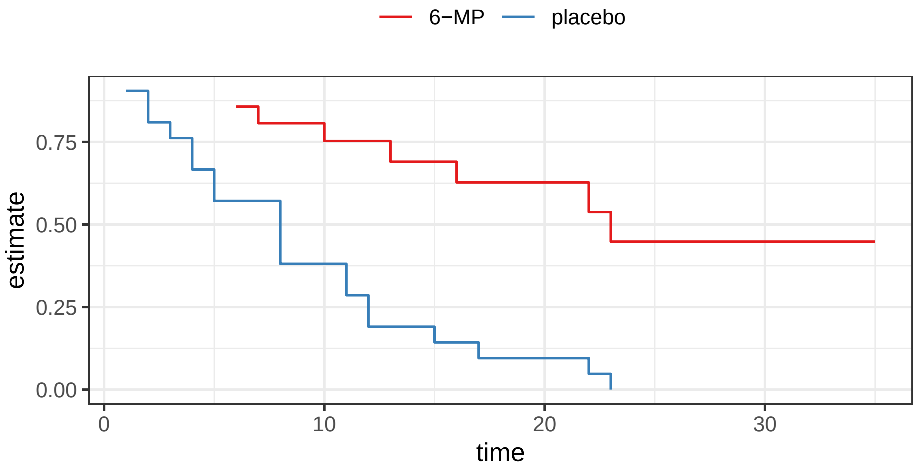
t_j	n_j	d_j	$\hat{S}(t_j^+)$	SE	lower	upper
6	21	3	0.857	0.076	0.707	1.000
7	17	1	0.807	0.087	0.636	0.977
10	15	1	0.753	0.096	0.564	0.942
13	12	1	0.690	0.107	0.481	0.900
16	11	1	0.627	0.114	0.404	0.851
22	7	1	0.538	0.128	0.286	0.789
23	6	1	0.448	0.135	0.184	0.712

Example: Remission data

Table 20: PL estimates and CIs (no transformation) for placebo group

t_j	n_j	d_j	$\hat{S}(t_j^+)$	SE	lower	upper
1	21	2	0.905	0.064	0.779	1.000
2	19	2	0.810	0.086	0.642	0.977
3	17	1	0.762	0.093	0.580	0.944
4	16	2	0.667	0.103	0.465	0.868
5	14	2	0.571	0.108	0.360	0.783
8	12	4	0.381	0.106	0.173	0.589
11	8	2	0.286	0.099	0.092	0.479
12	6	2	0.190	0.086	0.023	0.358
15	4	1	0.143	0.076	0.000	0.293
17	3	1	0.095	0.064	0.000	0.221
22	2	1	0.048	0.046	0.000	0.139
23	1	1	0.000	NaN	NaN	NaN

Example: Remission data



Example: Remission data

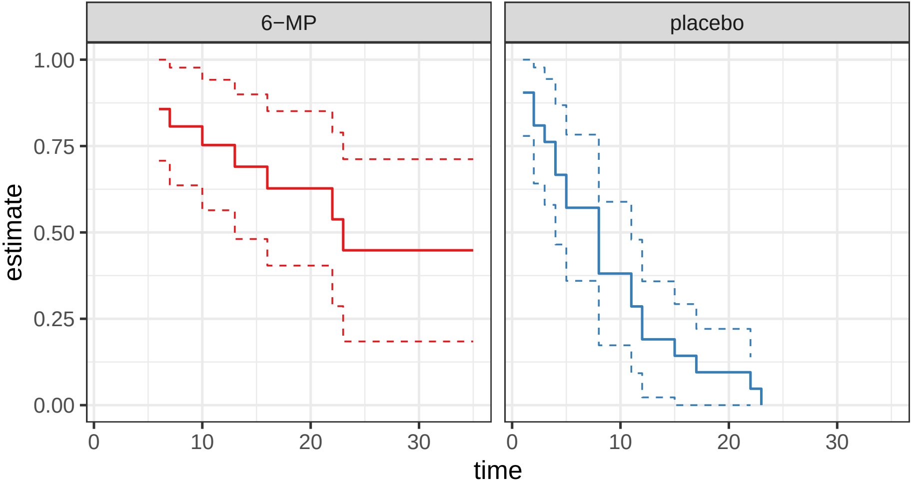


Figure 7: No transformation was used to obtain CIs

Example: Remission data

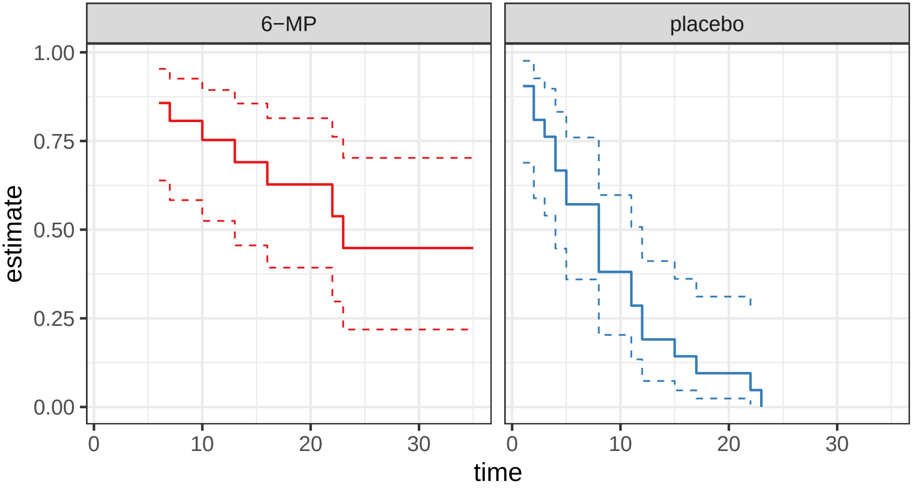


Figure 8: Logit transformation was used to obtain CIs

Example: Remission data

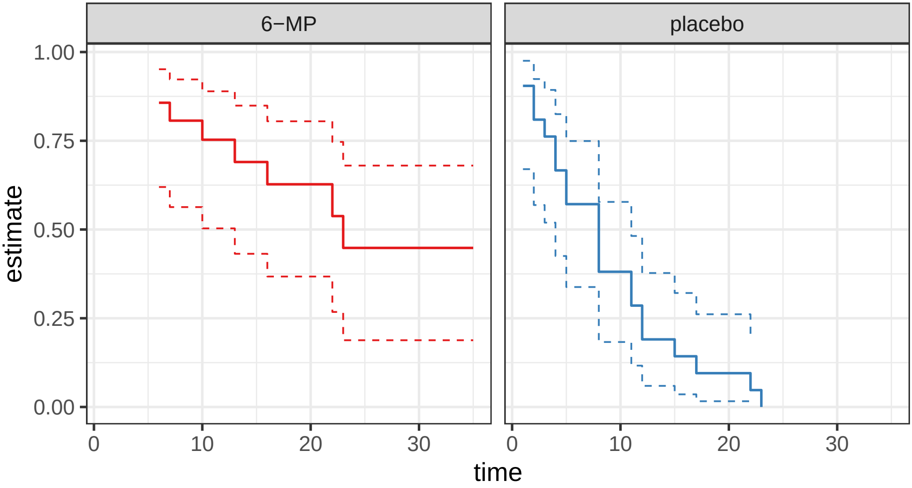


Figure 9: Log-log transformation was used to obtain CIs

Example: Remission data

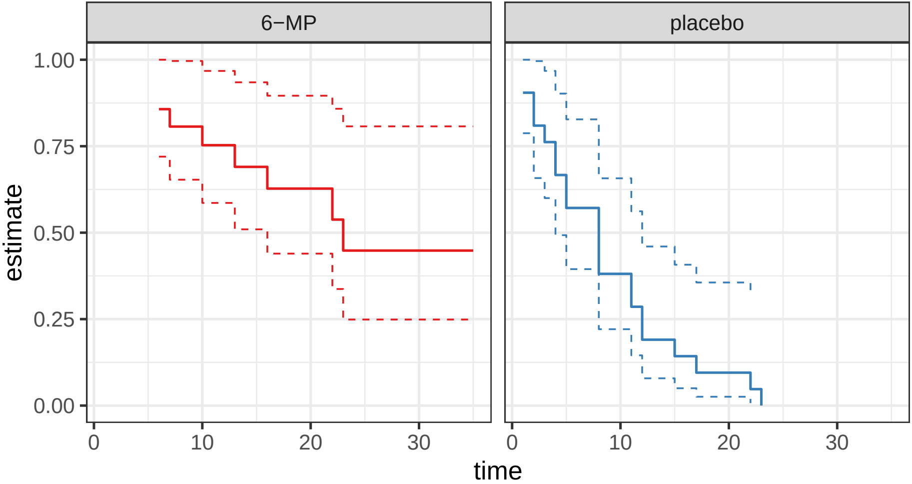


Figure 10: Log transformation was used to obtain CIs

Subsection 8

CIs for quantiles

CIs for quantiles

- For lifetime distribution, the quantiles t_p are of more interest than mean of the distribution, e.g., the median $t_{.50}$ is used as the measure of location for lifetime distribution
- Median has some advantages over mean as a measure of location for lifetime distributions
- Median always exist (provided $S(\infty) < .5$) and it is easier to estimate when data are censored

CIs for quantiles

Estimates of quantiles

- Nonparametric estimates of t_p can be obtained from the PL estimates $\hat{S}(t)$
- For a step function $\hat{S}(t)$, the corresponding inverse function $\hat{S}^{-1}(1-p)$ is not uniquely defined
- The estimate $\hat{t}_p$ could be either an intervals of times (t's) or a specific value of times (t's) depending on the point at which the line $\hat{S}(t) = 1 - p$ intersect the step function $\hat{S}(t)$

$$\hat{t}_p = \min\left\{t_i : \hat{S}(t_i) \leqslant 1 - p\right\} \tag{9}$$

CIs for quantiles

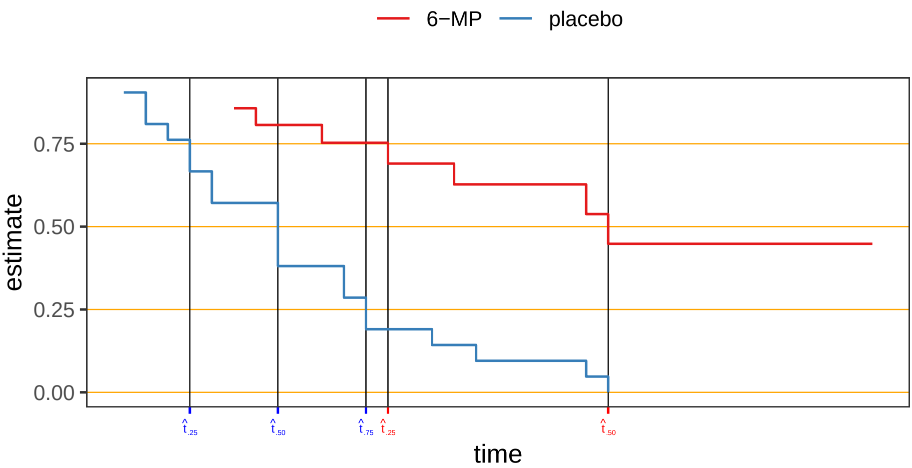


Figure 11: Estimated quartiles of the length of remission time for the two treatment groups

CIs for quantiles

Table 21: Estimated quartiles for two treatment groups

prob	6-MP	placebo
0.25	13	4
0.50	23	8
0.75	NA	12

CIs for quantiles

Confidence intervals of quantiles

- $100(1-\alpha)\%$ confidence interval of t_p can be obtained by inverting the corresponding confidence interval of survival function $S(t)$

CIs for quantiles

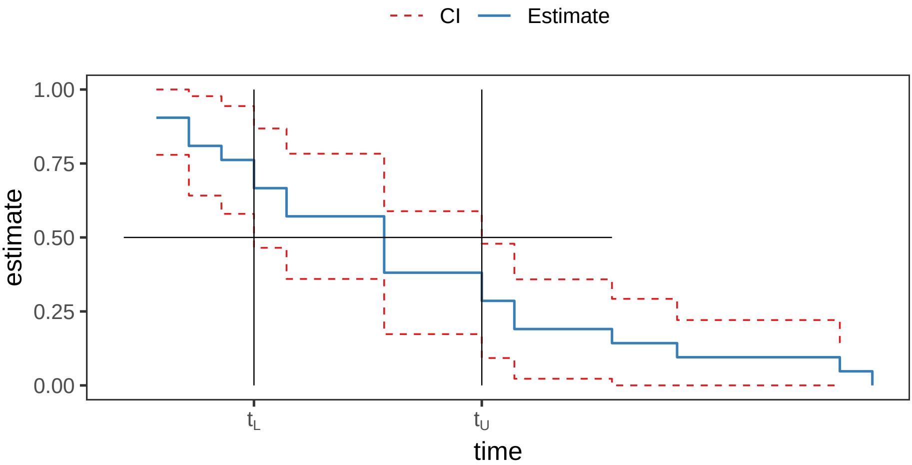


Figure 12: 95% CI of median $(t_{.5})$ remission time for placebo group of patients

CIs for quantiles

Table 22: Estimated two-sided 95% CI of different quartiles for two treatment groups

prob	6-MP	placebo
0.25	$(6, 23)$	$(2, 8)$
0.50	NA	$(4, 11)$
0.75	NA	$(8, 17)$

- Confidence intervals for second and third quartiles cannot be estimated from this data set for the "6-MP" group because $\hat{S}(\infty) < .5$

Subsection 9

3.3 Descriptive and diagnostic plots

3.3 Descriptive and diagnostic plots

- Plots of PL $(\hat{S}(t))$ or Nelson-Aalen $(\hat{H}(t))$ estimates can be used to provide good description of univariate lifetime data
- These estimates are useful to assess the appropriateness of a parametric model

Plots of survivor functions

- Data: $\{(t_i, \delta_i), i = 1, \ldots, n\}$
- $\hat{S}(t) \rightarrow$ PL estimate of survivor function $S(t)$
- Assume lifetimes follow a distribution with survivor function $S(t;\theta)$ and $F(t;\theta)$ is the corresponding distribution function, where θ is the parameter vector, e.g. for exponential distribution

$$S(t;\theta) = e^{-t/\theta} \qquad t \geqslant 0, \;\; \theta > 0$$

Plots of survivor functions

- Let $\hat{\boldsymbol{\theta}}$ is an estimate of $\boldsymbol{\theta}$ and $S(t; \hat{\boldsymbol{\theta}})$ is the estimate of survivor function $S(t; \boldsymbol{\theta})$, e.g. for exponential distribution

$$S(t; \hat{\theta}) = e^{-t/\hat{\theta}}$$

Plots of survivor functions

- If the model assumption (i.e. lifetimes follow a distribution with survivor function $S(t;\boldsymbol{\theta})$) is appropriate then $S(t;\hat{\boldsymbol{\theta}})$ should not be very far from $\hat{S}(t)$
- A comparison between $S(t;\hat{\boldsymbol{\theta}})$ and $\hat{S}(t)$ can be used as a model assessment tool
- A plot of $S(t;\hat{\boldsymbol{\theta}})$ and $\hat{S}(t)$ on the same graph can be used to compare graphically

Plots of survivor functions

Example 3.3.1: Ball bearing data

Table 23: The number of revolutions (in million) before failure for each of 23 ball bearings

17.88	41.52	48.40	54.12	68.64	84.12	105.12	128.04
28.92	42.12	51.84	55.56	68.64	93.12	105.84	173.40
33.00	45.60	51.96	67.80	68.88	98.64	127.92	NA

Plots of survivor functions

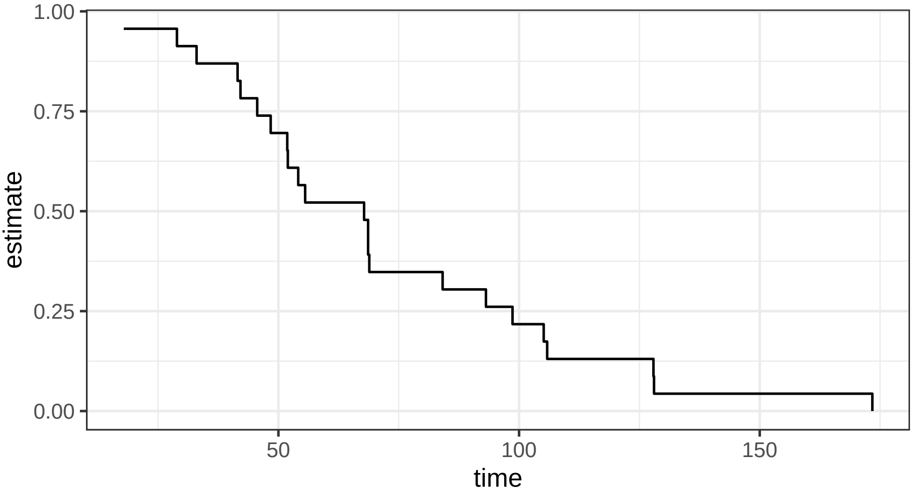


Figure 13: PL estimate of survival for ball bearing data

Plots of survivor functions

- Assume Weibull and log-normal models for analyzing ball bearing data and want to assess which of these two models is appropriate for the data
  - Weibull model

$$S(t;\alpha,\beta) = e^{-(t/\alpha)^\beta} \quad t \geqslant 0, \alpha > 0, \beta > 0$$

  - Log-normal model

$$S(t;\mu,\sigma) = 1 - \Phi\left(\frac{\log t - \mu}{\sigma}\right) \quad t \geqslant 0, -\infty < \mu < \infty, \sigma > 0$$

Plots of survivor functions

Estimated survivor functions

- Weibull model

$$S(t; \hat{\alpha}, \hat{\beta}) = e^{-(t/81.87)^{2.10}}$$

- Log-normal model

$$S(t; \hat{\mu}, \hat{\sigma}) = 1 - \Phi\left(\frac{\log t - 4.15}{0.52}\right)$$

Plots of survivor functions

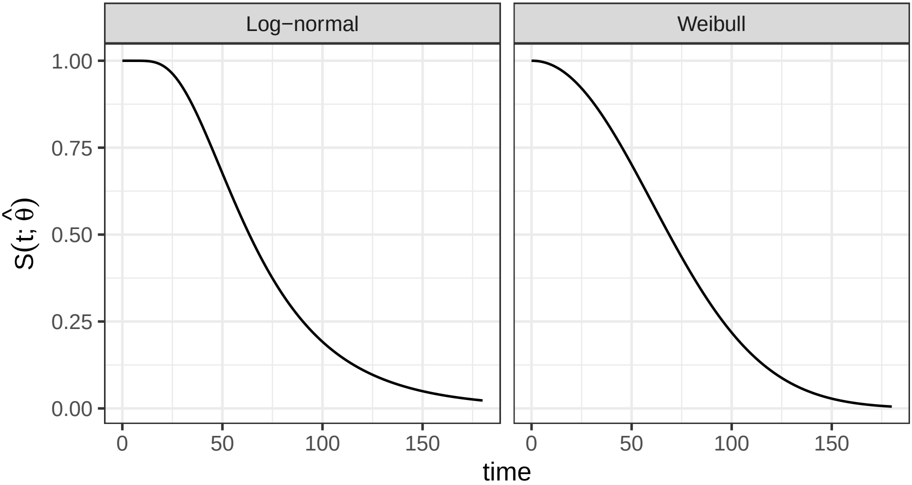


Figure 14: Estimated survivor functions using Weibull and log-normal models

Plots of survivor functions

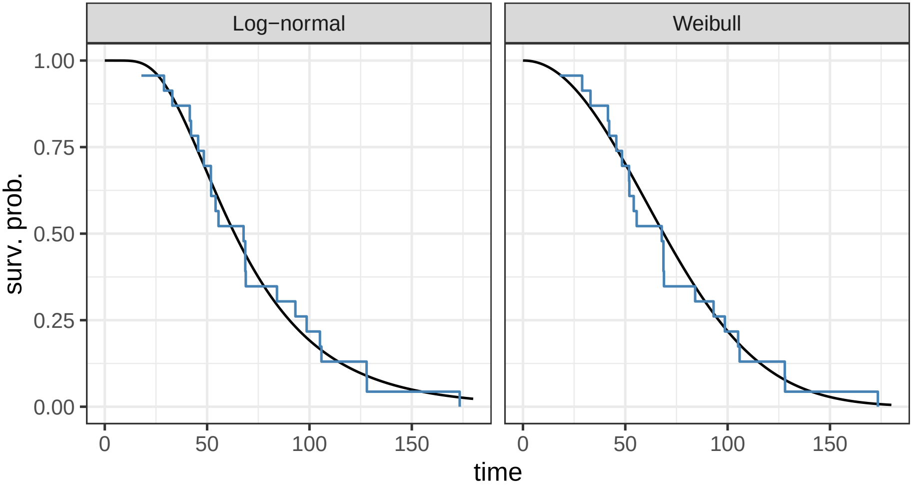


Figure 15: Comparison of Weibull and log-normal models with respect to PL estimates

Plots of survivor functions

- Figure 15 shows that there is a good agreement between nonparametric PL estimates $\hat{S}(t)$ and both the Weibull and log-normal models
- Either Weibull or log-normal can be assumed to analyze ball bearing data

Plots of survivor functions

Probability plots

- Let $t_1 < \cdots < t_k$ are k distinct failure times, and $S(t_j; \hat{\boldsymbol{\theta}})$ and $\hat{S}(t_j)$ are corresponding parametric and non-parametric estimates of survivor function
- Probability-probability (P-P) plot is defined as the scatter plot of

$$\{S(t_j; \hat{\boldsymbol{\theta}}),\ \hat{S}(t_j)\},\ \ j = 1, \ldots, k$$

- If the assumed parametric model $S(t; \boldsymbol{\theta})$ is appropriate then all the points in the resulting scatter plot should lie around a straight line with slope one

Plots of survivor functions

Table 24: Estimate of survivor function with nonparametric (PL) and parametric (Weibull, and log-normal) methods for smallest 10 failure times of ball bearing data

time	PL	Weibull	Log-normal
17.88	0.978	0.960	0.992
28.92	0.935	0.894	0.934
33.00	0.891	0.862	0.895
41.52	0.848	0.787	0.792
42.12	0.804	0.781	0.784
45.60	0.761	0.747	0.737
48.40	0.717	0.718	0.698
51.84	0.674	0.682	0.651
51.96	0.630	0.681	0.649
54.12	0.587	0.658	0.620

Plots of survivor functions

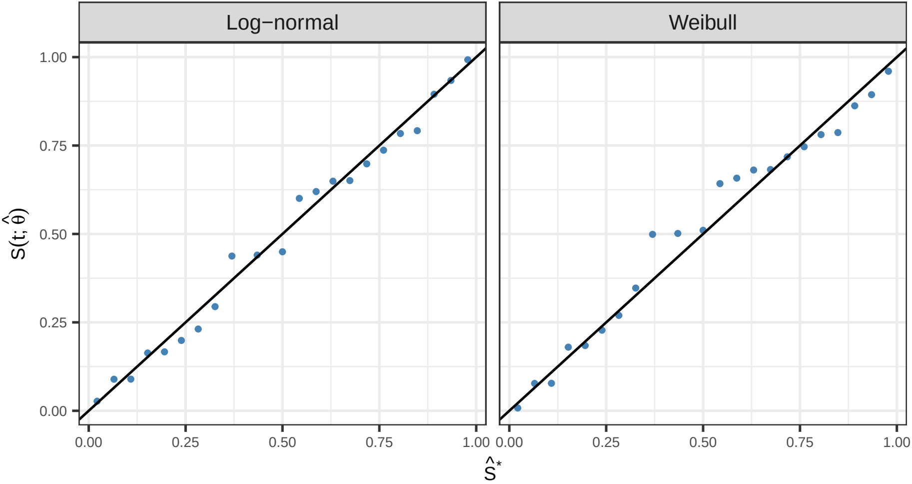


Figure 16: P-P plot for comparing Weibull and log-normal models

Plots of survivor functions

Quantile-Quantile (Q-Q) plot

- Quantile-quantile (Q-Q) plot compares observed quantiles and the corresponding quantiles estimated from the assumed parametric model $S(t; \boldsymbol{\theta})$
- Observed quantiles are observed distinct failure time t_j corresponding to the PL estimates $\hat{S}_j$

Plots of survivor functions

- $t(\hat{S}_j, \boldsymbol{\theta})$ are the quantiles corresponding to $\hat{S}_j$ obtained from the model $S(t; \boldsymbol{\theta})$

$$t(p, \boldsymbol{\theta}) = \begin{cases} \alpha\big[-\log(1-p)\big]^{1/\beta} & \text{Weibull} \\ \exp\left[\mu + \sigma\Phi^{-1}(p)\right] & \text{Log-normal} \end{cases}$$

- Q-Q plot is defined as the scatter plot of $t(\hat{S}_j, \hat{\boldsymbol{\theta}})$ versus t_j

Plots of survivor functions

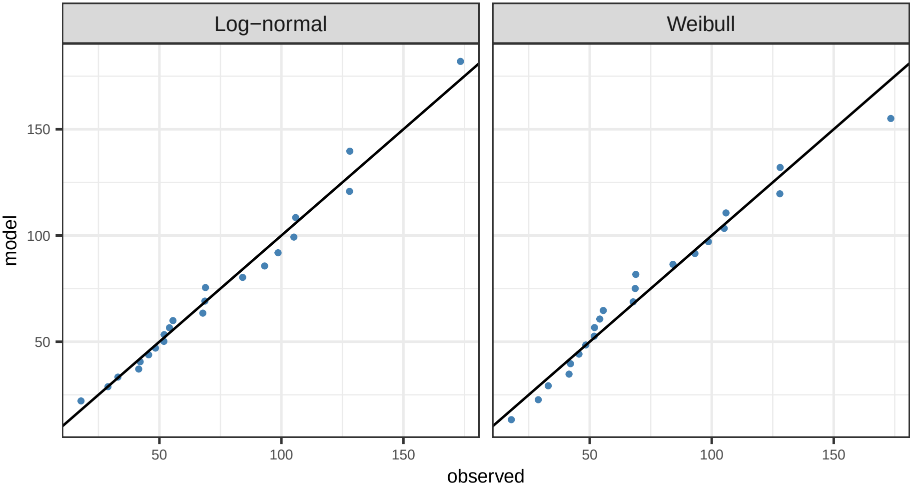


Figure 17: Q-Q plot for comparing Weibull and log-normal models

Plots of survivor functions

Linearization method

- This method is based on linearizing the survivor or distribution function of the assumed parametric model
- There exists functions $g_1(\cdot)$ and $g_2(\cdot)$ such that $g_1[S(t;\boldsymbol{\theta})]$ is a linear function of $g_2(t)$
- If the parametric model $S(t;\boldsymbol{\theta})]$ is appropriate, the plot of $g_1[\hat{S}(t)]$ versus $g_2(t)$ should be roughly linear, where $\hat{S}(t)$ is the PL estimate
- This method does not require to estimate the parameter $\boldsymbol{\theta}$

Plots of survivor functions

- Exponential distribution

$$S(t;\alpha) = e^{-t/\alpha} \Rightarrow \log\left[S(t;\alpha\right] = -t/\alpha \tag{10}$$

  - $\log\left[S(t;\alpha)\right]$ is a linear function of t, so a plot of $\log \hat{S}(t)$ versus t should be linear through the origin if the exponential model is appropriate
  - A graphical estimate of α can be obtained when the plot is roughly linear by fitting a straight line through the points

Plots of survivor functions

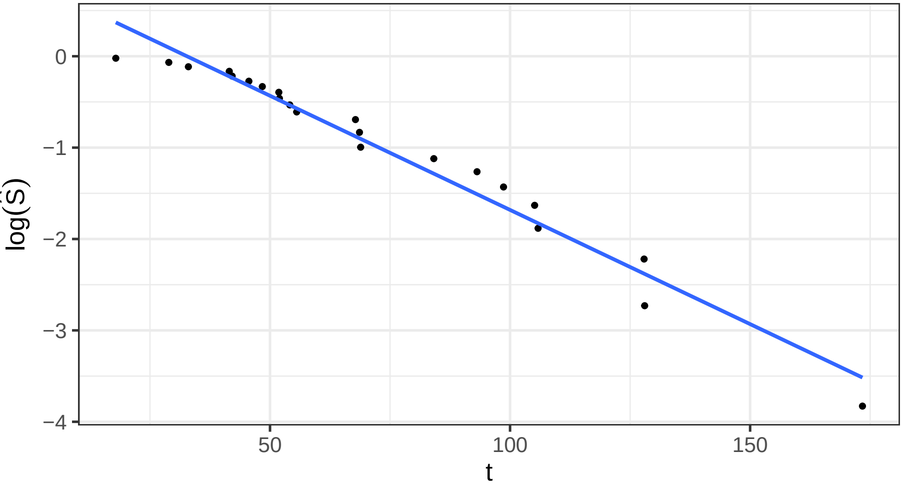


Figure 18: A plot of $\log(\hat{S}(t))$ versus t to assess whether exponential model is appropriate for ball bearing data

Plots of survivor functions

- Weibull distribution

$$
\begin{aligned}
S(t;\alpha,\beta) &= e^{-(t/\alpha)^{\beta}} \\
\Rightarrow \ \log\left[-\log S(t;\alpha,\beta)\right] &= \beta \log t - \beta \log \alpha
\end{aligned}
\tag{11}
$$

  - A plot of $\log\left[-\log \hat{S}(t)\right]$ versus $\log t$ should roughly linear if a Weibull model is appropriate
  - When the plot is approximately linear, graphical estimates of α and β can be obtained by intercept and slope of the fitted straight line through the points

Plots of survivor functions

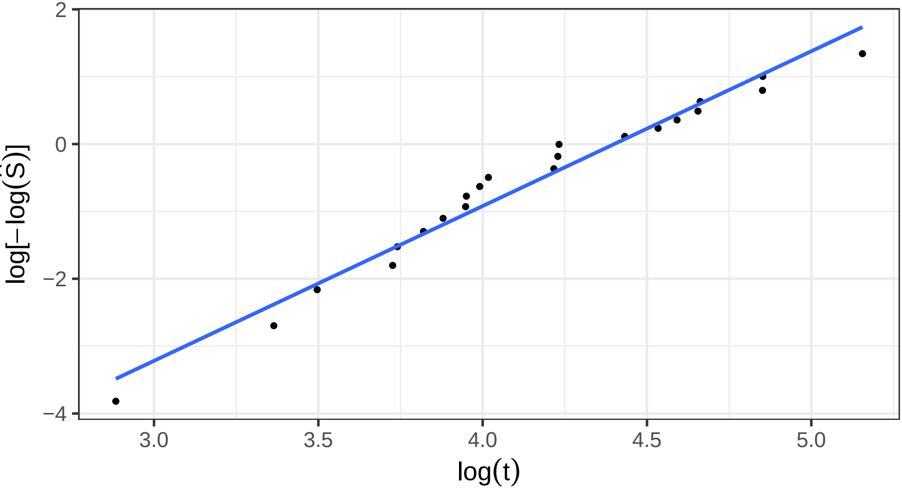


Figure 19: A plot of $\log[-\log(\hat{S}(t))]$ versus $\log(t)$ to asses whether Weibull model is appropriate for ball bearing data

Plots of survivor functions

Location-scale family

- In general, linearization method of assessing the appropriateness of the assumed parametric model can be defined for location-scale family of models
- Let transformed lifetime $Y = g(T)$ has a location-scale distribution, e.g. $g(\cdot) = \log(\cdot)$

Plots of survivor functions

- Survivor function of Y can be defined as

$$\begin{aligned} Pr(Y \geqslant y) &= Pr\left(\frac{Y-u}{b} \geqslant \frac{y-u}{b}\right) \\ &= Pr\left(Z \geqslant \frac{y-u}{b}\right) = S_0\left(\frac{y-u}{b}\right) \\ &= Pr(T \geqslant t) = S(t) \end{aligned} \tag{12}$$

  - $t = g^{-1}(y)$, $-\infty < u < \infty$, $b > 0$

Plots of survivor functions

- It can be shown that

$$S_0\left(\frac{y-u}{b}\right) = S(t)$$
$$\Rightarrow\ S_0^{-1}\Big[S(t)\Big] = \frac{y}{b} - \frac{u}{b}$$

  is a linear of $y = g(t)$

- A plot of $S_0^{-1}[\hat{S}(t)]$ versus $g(t)$ should be roughly linear if the assumed family of models is appropriate.

Plots of survivor functions

- Expressions of $S_0(\cdot)$ and $S_0^{-1}(\cdot)$ for log-location-scale family of models

$$S_0(z) = \begin{cases} \exp(-e^z) & \text{Extreme value} \\ (1+e^z)^{-1} & \text{Logistic} \\ 1-\Phi(z) & \text{Normal} \end{cases}$$

$$S_0^{-1}(p) = \begin{cases} \log\big(-\log(p)\big) & \text{Extreme value} \\ \log\frac{1-p}{p} & \text{Logistic} \\ \Phi^{-1}(1-p) & \text{Normal} \end{cases}$$

- For all these three distributions, $y = g(t) = \log(t)$

Plots of survivor functions

- In general, for distinct failure times, a plot of points

$$(\log(t_j), S_0^{-1}(\hat{S}_j)),\ j = 1, \dots, k$$

can be used to assess whether the assumed location-scale model is appropriate

Plots of survivor functions

- Graphical estimates of u and b can be obtained from the lines fitted to the points, where b^{-1} and u are estimated by the slope and $\log(t)$-intercept, respectively
- Graphical estimates can be used as initial values of the optimization routines that are used for estimating model parameters
- Since these plots are subject to sampling variations, these are often used for informal model assessment

Plots of survivor functions

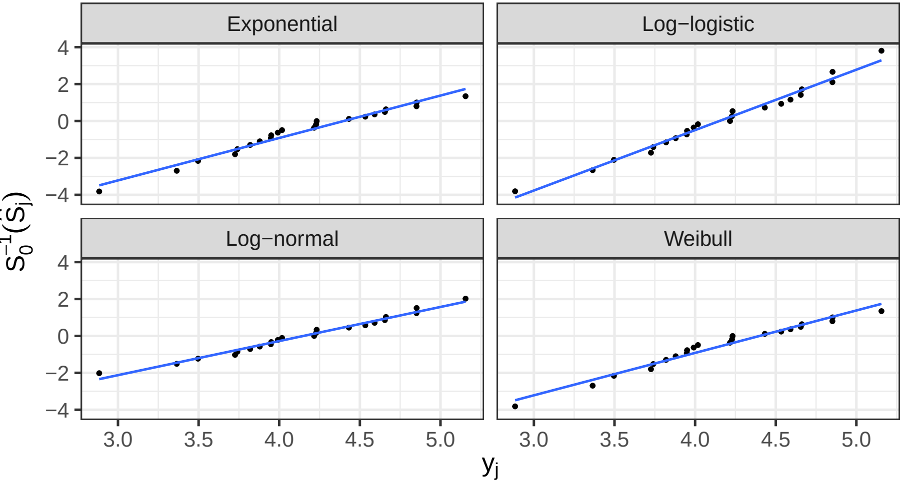


Figure 20: Comparison of models of log-location-scale family

Hazard plots

- The plotting procedures described for survivor functions $S(t)$ can also be described for cumulative hazards function $H(t)$
  - Survivor function of Weibull distribution

$$S(t;\alpha,\beta) = e^{-(t/\alpha)^\beta}$$
$$\Rightarrow \log[-\log S(t;\alpha,\beta)] = \beta\log(t) - \beta\log(\alpha)$$

  - Cumulative hazard function of Weibull distribution

$$H(t;\alpha,\beta) = \left(\frac{t}{\alpha}\right)^\beta$$
$$\log H(t;\alpha,\beta) = \beta\log(t) - \beta\log(\alpha)$$

Hazard plots

- An alternative of plotting $\log[-\log \hat{S}(t)]$ versus $\log(t)$ would be to plot $\log \hat{H}(t)$ versus $\log(t)$ can be used to assess the appropriateness of the Weibull model for the data at hand
  - $\hat{S}(t) \rightarrow$ PL estimate
  - $\hat{H}(t) \rightarrow$ Nelson-Aalen estimate
- The plots based on survivor function and cumulative hazard function could differ slightly, specially for a large t, because $-\log \hat{S}(t) \neq \hat{H}(t)$ for discrete data

References

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