STAT 513 fa 2020 Lec 11 slides

Censored data and survival analysis

Karl B. Gregory

University of South Carolina

These slides are an instructional aid; their sole purpose is to display, during the lecture, definitions, plots, results, etc. which take too much time to write by hand on the blackboard. They are not intended to explain or expound on any material.

Form of censored time-to-event data

Let $T_1, \ldots, T_n$ be indep. time-to-event rvs and $C_1, \ldots, C_n$ be censoring times.

We observe $(Y_1, \delta_1), \ldots, (Y_n, \delta_n)$, where $\delta_1, \ldots, \delta_n \in \{0, 1\}$ are indicators st

$$Y_i = \begin{cases} T_i & \text{if } \delta_i = 1 \\ C_i & \text{if } \delta_i = 0 \end{cases} \qquad \text{for } i = 1, \ldots, n.$$

So $\delta_i = 0$ means "censored" and $\delta_i = 1$ means "observed."

Right-censoring

If the event occurs *after* the censoring time, the censoring is called *right-censoring*.

Under right-censoring we have

$$\delta_i = \begin{cases} 1 & \text{if } T_i \leq C_i \\ 0 & \text{if } T_i > C_i \end{cases} \qquad \text{for } i = 1, \ldots, n,$$

so that

$$Y_i = \min\{T_i, C_i\} \quad \text{for } i = 1, \ldots, n.$$

Left-censoring

If the event occurs *before* the censoring time, the censoring is called *left-censoring*.

Under left-censoring we have

$$\delta_i = \begin{cases} 1 & \text{if } T_i \geq C_i \\ 0 & \text{if } T_i < C_i \end{cases} \qquad \text{for } i = 1, \ldots, n,$$

so that

$$Y_i = \max\{T_i, C_i\} \quad \text{for } i = 1, \ldots, n.$$

Example of right censoring: Times in remission (weeks) for two groups of leukemia patients; the event of interest is coming out of remission.

Group 1 (Treatment)	Group 2 (Placebo)
6, 6, 6, 7, 10,	1, 1, 2, 2, 3,
13, 16, 22, 23,	4, 4, 5, 5,
6+, 9+, 10+, 11+,	8, 8, 8, 8,
17+, 19+, 20+,	11, 11, 12, 12,
25+, 32+, 32+,	15, 17, 22, 23
34+, 35+	

The "+" denotes right-censoring.

Survival function

If T has cdf $F(\cdot)$, the *survival function* of T is defined as $S(\cdot) = 1 - F(\cdot)$.

In time-to-event contexts, we like to estimate/study the survival function.

We have $P(T > t) = S(t)$.

Likelihood under right-censoring with random censoring times

Let $C_1, \ldots, C_n \overset{\text{ind}}{\sim} G$, indep. of $T_1, \ldots, T_n$, and suppose G does not depend on θ.

Likelihood can be written as

$$L(\theta; (Y_1, \delta_1), \ldots, (Y_n, \delta_n)) = \prod_{i \in \mathcal{U}} f(Y_i; \theta) \prod_{i \in \mathcal{C}} S(Y_i; \theta) \times K((Y_1, \delta_1), \ldots, (Y_n, \delta_n)),$$

where

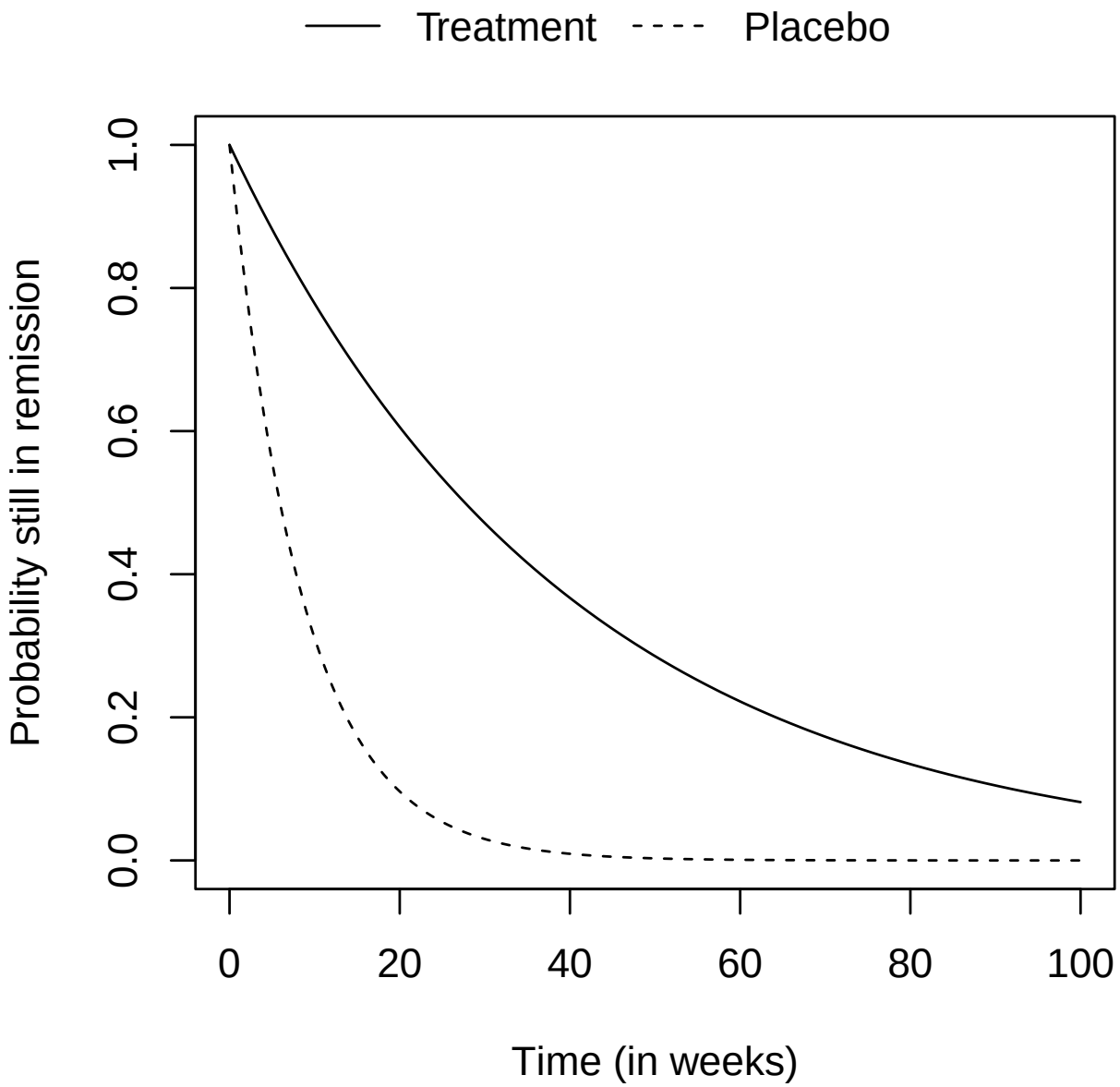
- K some function free of θ
- $\mathcal{U}$ the indices of uncensored obs
- $\mathcal{C}$ the indices of censored obs

Exercise: Derive the above.

Exercise: Observe $(Y_1, \delta_1), \ldots, (Y_n, \delta_n)$ such that $Y_i = \min\{T_i, C_i\}$, where

- $T_1, \ldots, T_n \overset{\text{ind}}{\sim} \text{Exponential}(\lambda)$
- $C_1, \ldots, C_n$ indep. random censoring times indep. of $T_1, \ldots, T_n$
- $\delta_i = 1$ if $T_i \leq C_i$ and $\delta_i = 0$ otherwise.

1. Find the maximum likelihood estimator of λ based on the censored data.
2. Compute this on the times in remission of the treatment group of the leukemia data (assumes an exponential distribution for times in remission).
3. Plot estimated survival functions for the treatment and placebo group of the leukemia data assuming exponentially distributed times in remission.


Treatment
Placebo
Probability still in remission
0.0
0.2
0.4
0.6
0.8
1.0
0
20
40
60
80
100
Time (in weeks)

Life table estimator of survival function

Break observation period into K intervals

$$[t_0, t_1), \ldots, [t_{K-1}, t_K), \quad \text{with} \quad 0 = t_0 < t_1 < \cdots < t_K,$$

$d_k = \#$ observed "deaths" in the interval $[t_{k-1}, t_k)$,

$c_k = \#$ subjects censored during the interval $[t_{k-1}, t_k)$, and

$n_k = \#$ subjects "alive" and uncensored at beginning of $[t_{k-1}, t_k)$, "at risk".

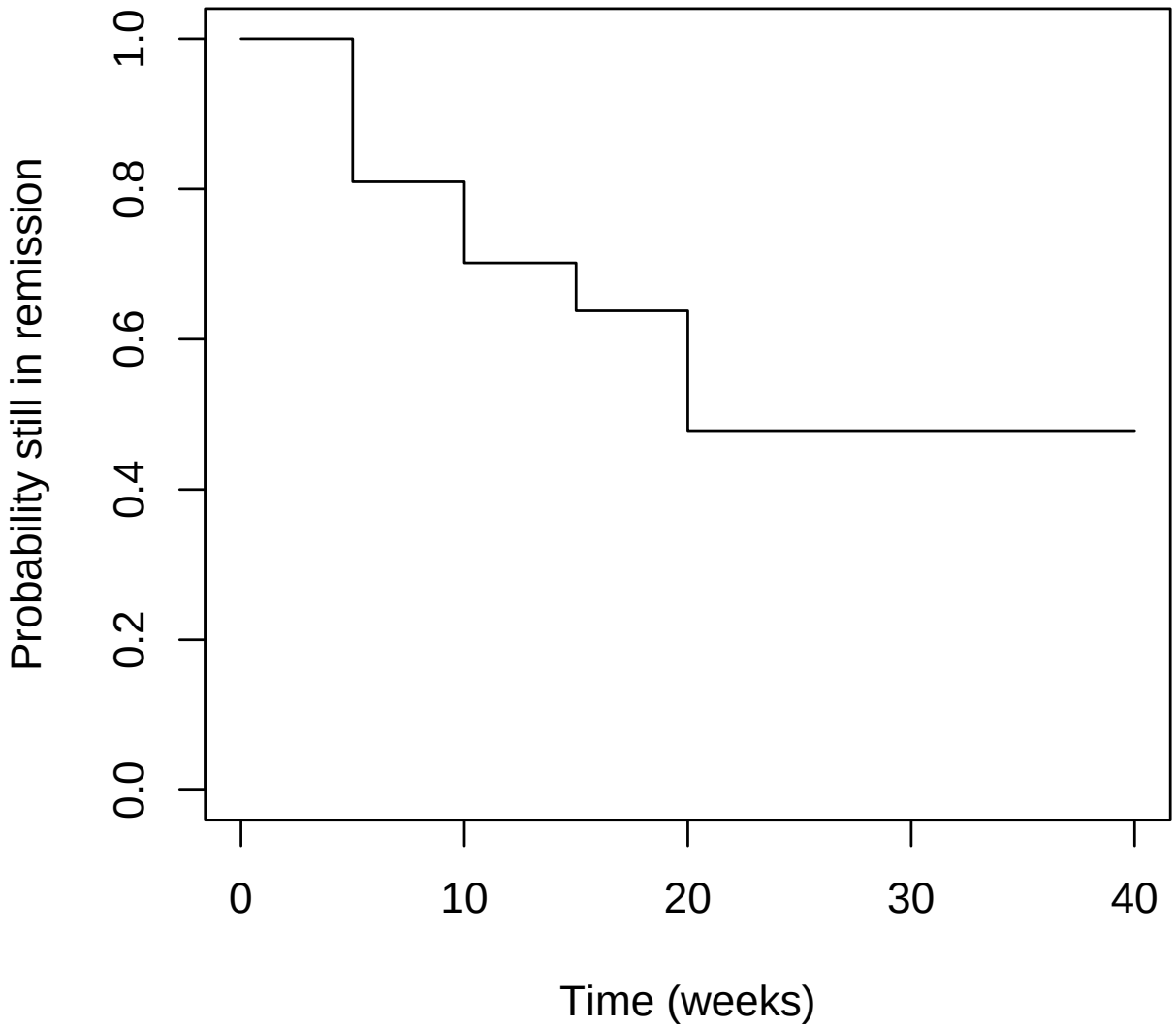
The *life table* is

Interval	# deaths	# censored	# at risk	$\hat{h}$	$\hat{S}$
$[t_0, t_1)$	d_1	c_1	n_1	d_1/n_1	$1 - d_1/n_1$
$[t_1, t_2)$	d_2	c_2	n_2	d_2/n_2	$(1 - d_2/n_2)(1 - d_1/n_1)$
$[t_2, t_3)$	d_3	c_3	n_3	d_3/n_3	$\prod_{j=1}^{3}(1 - d_j/n_j)$
	$\vdots$				
$[t_{K-1}, t_K)$	d_K	c_K	n_K	d_K/n_K	$\prod_{j=0}^{K}(1 - d_j/n_j)$

Can write as $\hat{S}(t) = \prod_{j=1}^{k}(1 - d_j/n_j)$ for $t \in [t_{k-1}, t_k)$.

Life table for treatment group of leukemia data base on 5-week intervals:

Interval	# out of remission	# cens.	# uncens. at beginning of int.	$\hat{h}$	$\hat{S}$
$[0, 5)$	0	0	21	0/21	1.000
$[5, 10)$	4	2	21	4/21	0.810
$[10, 15)$	2	2	15	2/15	0.702
$[15, 20)$	1	2	11	1/11	0.638
$[20, 25)$	2	1	8	2/8	0.478
$[25, 30)$	0	1	5	0/5	0.478
$[30, 35)$	0	3	4	0/4	0.478
$[35, 40)$	0	1	1	0/1	0.478


Probability still in remission
0.0
0.2
0.4
0.6
0.8
1.0
0
10
20
30
40
Time (weeks)

Hazard function

The *hazard function* for a random variable T is defined as

$$h(t) = \lim_{\delta \downarrow 0} \frac{P(T \leq t + \delta | T > t)}{\delta}.$$

Exercise: Show that if T has pdf f and survival function S, then

$$h(t) = \frac{f(t)}{S(t)}.$$

Discuss: How we get the estimator $\hat{S}_k = \prod_{j=0}^{k}(1 - \hat{h}_k)$.

Kaplan-Meier estimator of survival function

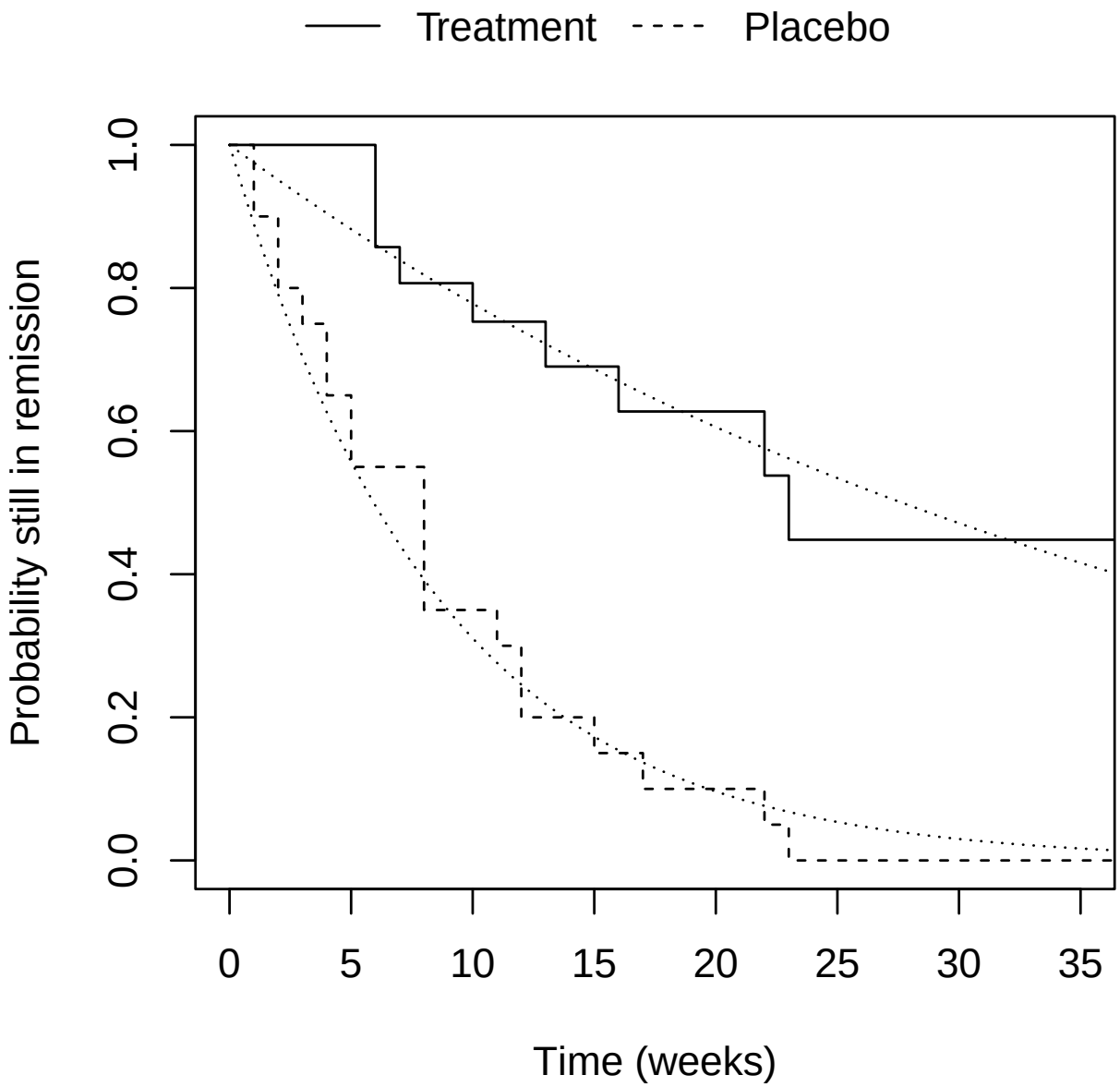
Given unique uncensored event times $U_{(1)} < \cdots < U_{(K-1)}$ the *Kaplan-Meier estimator* is the life table estimator based on the intervals defined by

$$t_0 = 0 \quad < \quad t_1 = U_{(1)} \quad < \quad \ldots \quad < \quad t_{K-1} = U_{(K-1)} \quad < \quad t_K = \infty.$$

This results in the life table

Interval	# deaths	# censored	# at risk	$\hat{h}$	$\hat{S}$
$[0, U_{(1)})$	d_1	c_1	n_1	d_1/n_1	$1 - d_1/n_1$
$[U_{(1)}, U_{(2)})$	d_2	c_2	n_2	d_2/n_2	$(1 - d_2/n_2)(1 - d_1/n_1)$
$[U_{(2)}, U_{(3)})$	d_3	c_3	n_3	d_3/n_3	$\prod_{j=1}^{3}(1 - d_j/n_j)$
	$\vdots$				
$[U_{(K-1)}, \infty)$	d_K	c_K	n_K	d_K/n_K	$\prod_{j=0}^{K}(1 - d_j/n_j)$

Exercise: Use the `Surv()` and `survfit()` functions from the `survival` library to compute the Kaplan-Meier estimator on the Leukemia data.


Treatment
Placebo
Probability still in remission
0.0
0.2
0.4
0.6
0.8
1.0
0
5
10
15
20
25
30
35
Time (weeks)

Greenwood's formula

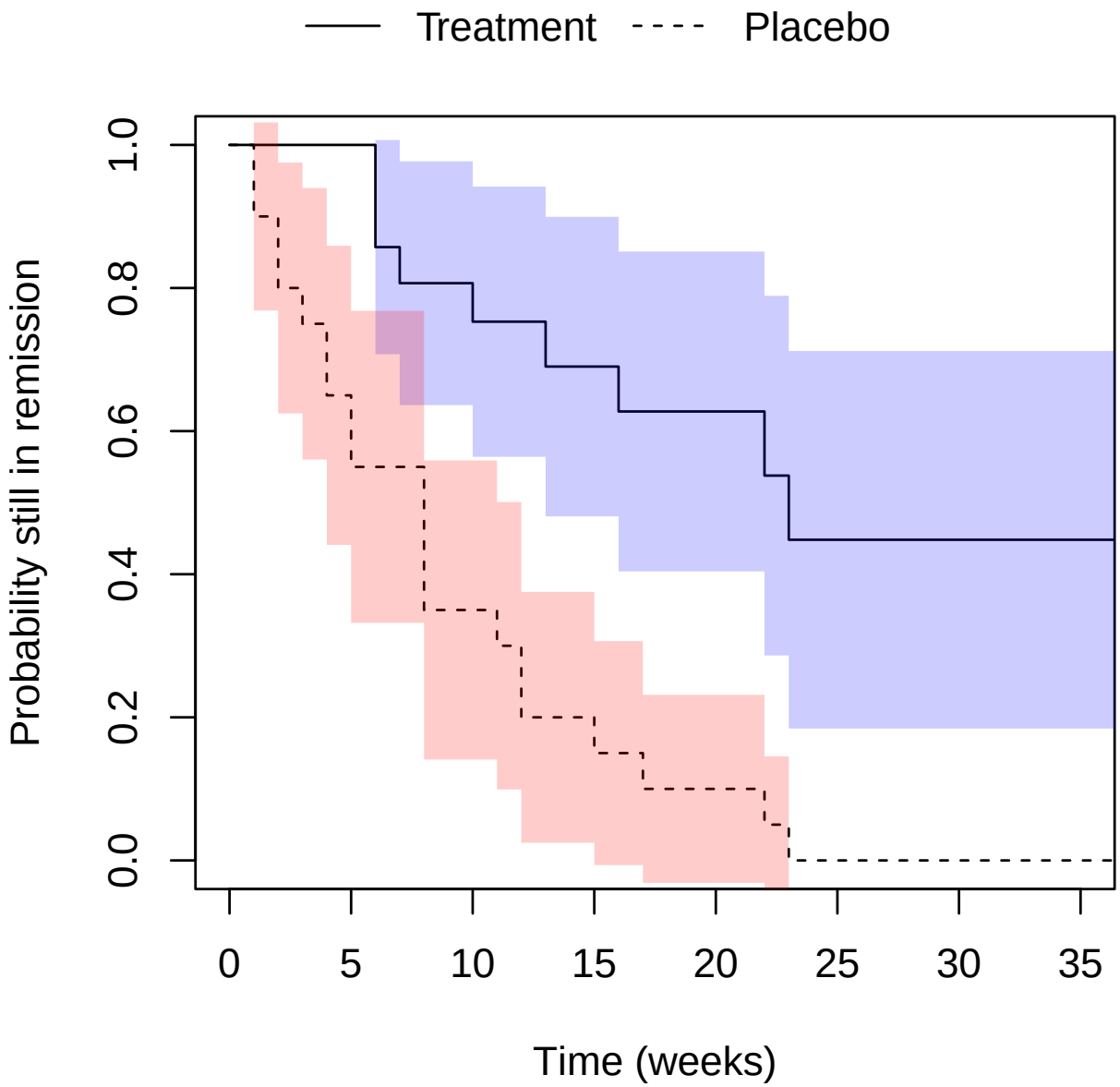
For the life-table estimator $\hat{S}(t) = \prod_{j=1}^{k}(1 - d_j/n_j)$ for $t \in [t_{k-1}, t_k)$, we have

$$\text{Var}\,\hat{S}(t) \approx [\hat{S}(t)]^2 \cdot \sum_{j=1}^{k} \frac{d_j}{n_j(n_j - d_j)} \qquad \text{for } t \in [t_{k-1}, t_k)$$

for large n. Can heuristically derive with two applications of the delta method.

An approximate $(1-\alpha)\%$ CI for $S(t)$ is therefore given by

$$S(t) \pm z_{\alpha/2} \cdot \hat{S}(t) \sqrt{\sum_{j=1}^{k} \frac{d_j}{n_j(n_j - d_j)}} \qquad \text{for } t \in [t_{k-1}, t_k).$$


Treatment
Placebo
Probability still in remission
0.0
0.2
0.4
0.6
0.8
1.0
0
5
10
15
20
25
30
35
Time (weeks)