Lecture13 Handout

Elizabeth Colantuoni

5/10/2021

I. Objectives

Upon completion of this session, you will be able to do the following:

- Understand and explain the relationships among the hazard, survival, density, and distribution functions for random variables
- Understand and explain the motivation for modeling the hazard/survival rather than density/distribution functions for censored time-to-event data
- Understand and explain the Kaplan-Meier estimate of the survival function
- Understand the log-rank test of equality for two survival functions and its connection to the analysis of data from nested 2x2 tables

I. Survival endpoints

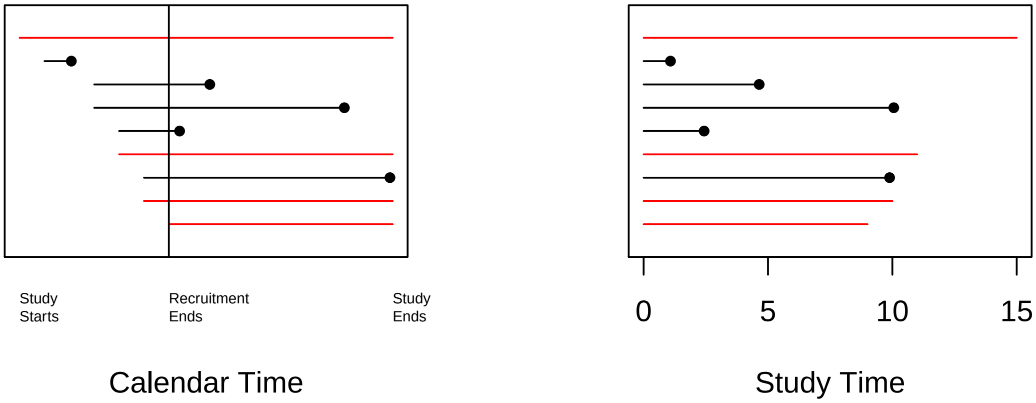
In the Lecture11-Handout, we discussed the application of log-linear models for survival outcomes. In this application, we are required to "bin" the survival time. "Binning" the survival time is somewhat arbitrary; i.e. how wide or narrow to make the bins may depend on the analyst.

We now move to thinking about analysis of survival outcomes where we let time be continuous!

First, let's revisit survival endpoints.

Take the example we disucssed in *Lecture11-Handout*.

1. The data contains information about *time to death* for inpatients hospitalized for a severe mental disorder. Survival time from hospitalization is in years.
2. In most studies measuring survival time of patients, we don't get to follow patients long enough to see when the event occurs for all patients. In fact, within a given study, patients are recruited to particapte over the course of an enrollment period, which is represented by calendar time. Then our goal is to follow the patients until a particular event occurs; however, the study may end prior to the event occurring. What information do we have for these patients? We know they didn't have the event up until a particular time but nothing else. These patient's outcomes are "censored".
3. The figure below on the left displays a hypothetical example where we see patients recruited over a particular time window and the follow patients to study termination. Patients still enrolled but whom did not have the event by study termination are censored (these are displayed in red). The figure on the right displays the same data but where we have scaled the time axis to reflect study time, not calendar time.



4. Patients are often *censored* because the study ends. We refer to this as administrative censoring. However, patients may be censored for other reasons including drop-out and becoming ineligible for the event of interest (perhaps a competing event like mortality occurs prior to observing the event of interest).
5. Absent censoring, the survival outcome Y_i is the time from start of an at risk period to when the event of interest occurs.
6. In the presence of censoring, we get to see $\delta_i = 1$ if the event occurs, and 0 if the event is censored and $T_i = min(D_i, C_i)$ where D_i is the time when the event occurs and C_i is the time of censoring. So, in the presence of censoring the data for patient i is (T_i, δ_i).
7. We often want to understand if the survival experience is different across exposure groups OR to predict the outcome. In this lecture, we focus on methods to compare the survival experience across exposure groups.

III. Definitions

For now, we will provide essential definitions and relationships for survival outcomes ignoring censoring.

Let T be a time to event random variable, $T \geq 0$.

Then we will define a series of quantities that can be used to describe the distribution of T.

- Cumulative Distribution Function: $F(t) = Pr(T \leq t)$
- Survival Function: $S(t) = Pr(T > t) = 1 - F(t)$
- Density function: $f(t) = \frac{d}{dt}F(T)$

$$\begin{aligned}
f(t)d(t) &= Pr(t < T < t + dt) \\
&= S(t) - S(t - dt) \\
&= (1 - S(t)) - (1 - S(t - dt)) \\
&= F(t + dt) - F(t)
\end{aligned}$$

- Hazard function: $h(t) = \lim_{dt \to 0} \frac{Pr(t < T \leq t + dt | T > t)}{dt}$

$$\begin{aligned}
h(t) &= \lim_{dt \to 0} \frac{Pr(t < T \leq t + dt | T > t)}{dt} \\
&= \lim_{dt \to 0} \frac{Pr(t < T \leq t + dt \text{ and } T > t)}{Pr(T > t)dt} \\
&= \lim_{dt \to 0} \frac{Pr(t < T \leq t + dt \text{ and } T > t)}{dtS(t)} \\
&= \tfrac{f(t)}{S(t)} \\
&= \tfrac{f(t)}{1-F(t)} \\
&= \tfrac{dF(t)}{dt}/[1 - F(t)] \\
&= -\tfrac{d}{dt}[1 - F(t)]/[1 - F(t)] \\
&= -\tfrac{d}{dt}S(t)/S(t) \\
&= -\tfrac{d}{dt}log_e S(t)
\end{aligned}$$

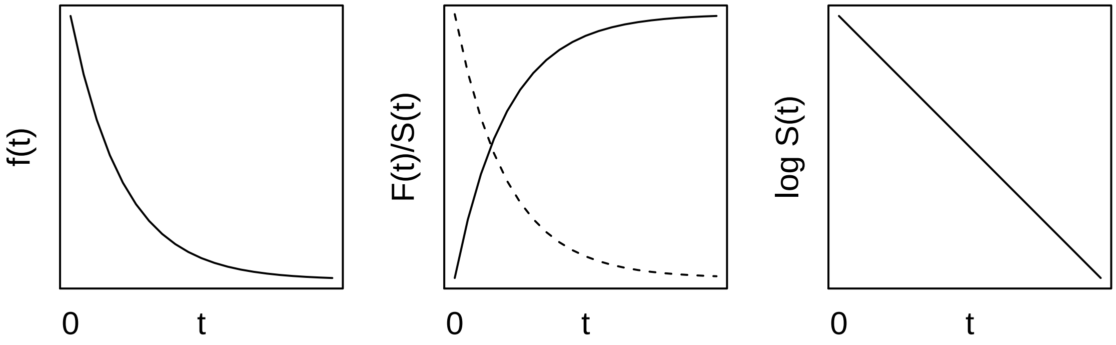
- Cumulative hazard function: $H(t) = \int_0^t h(u)du = log_e S(t)$. This implies: $S(t) = e^{-\int_0^t h(u)du} = e^{-H(t)}$.

IV. Common, Well known Parametric Models

A. Exponential Model

Assume $T \sim Exponential(\lambda)$ then

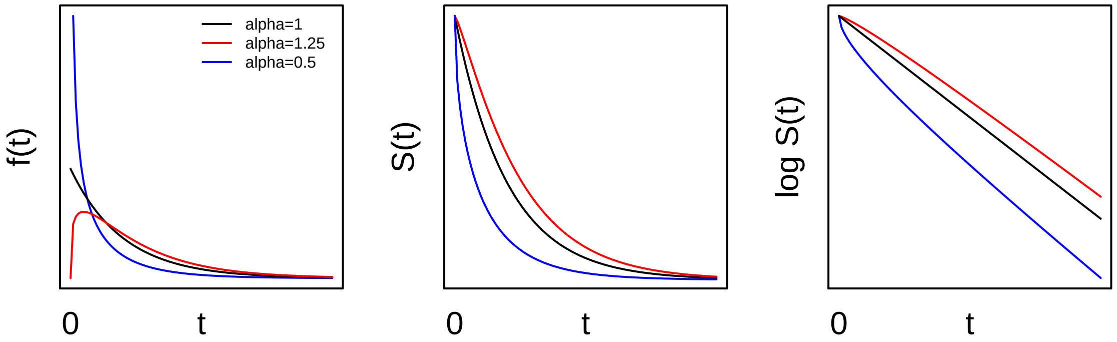
- $F(t) = 1 - e^{-\lambda t}$, $S(t) = e^{-\lambda t}$
- $f(t) = \frac{d}{dt}(1 - e^{-\lambda t}) = \lambda e^{-\lambda t}$
- $E(T) = 1/\lambda$, $Var(T) = 1/\lambda^2$
- $h(t) = f(t)/S(t) = \lambda e^{-\lambda t}/e^{-\lambda t} = \lambda$, i.e. a constant hazard model



B. Gamma Distribution

Assume $T \sim Gamma(\alpha, \lambda)$, then

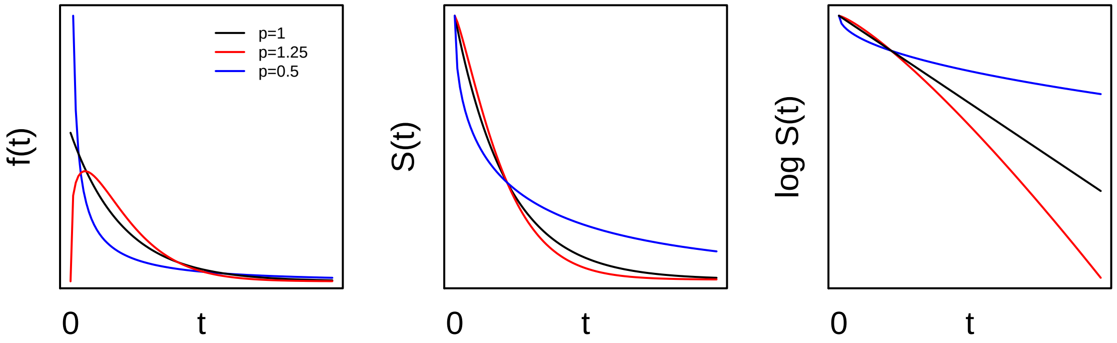
- $f(t) = \frac{\lambda^\alpha t^{\alpha-1} e^{-\lambda t}}{\Gamma(\alpha)}$, $t > 0$, $\Gamma(\alpha) = \int_0^\infty t^{\alpha-1} e^{-t} dt$
- $F(t)$, $S(t)$ and $h(t)$ have to be solved by numerical integration; there are no closed form solutions.



C. Weibull Distribution

Assume $T \sim Weibull(\lambda, p)$, then

- $f(t) = p\lambda t^{p-1} e^{-(\lambda t)^p}$
- $F(t) = 1 - e^{-(\lambda t)^p}$, $S(t) = e^{-(\lambda t)^p}$
- $h(t) = p\lambda^p t^{p-1}$
- When $p = 1$, $Weibull(\lambda, 1) = Exponential(\lambda)$.



V. Analysis of Survival Outcomes

Methods for the analysis of survival outcomes are extensive. Here we will review 3 **must know** methods for analyzing survival outcomes.

1. Estimating $S(t)$ via Kaplan-Meier survival function estimate (Lecture13-Handout)
2. Testing whether $S_1(t) = S_2(t)$ via the log-rank test (Lab 7)
3. Regression of survival outcomes on exposures via Cox Proportional Hazards regression models (Lecture14-Handout)

A. Estimating $S(t)$

The Kaplan-Meier estimate of the survival function $S(t)$ is also known as the **Product-limit** estimator.

This estimator for the survival function assumes that:

- censoring is unrelated to prognosis, i.e. event process and censoring process are independent
- the survival probabilities are the same for subjects recruited early and late in the study
- the events happened at the times specified

To construct the Kaplan-Meier estimator, you need to order the unique event times and compute:

Event times:	t_1	$<$	t_2	$<$	...	$<$	t_J
No. at risk:	N_1	$>$	N_2	$>$	...	$>$	N_J
No. of events:	y_1		y_2		...		y_J

The estimate of $S(t)$ is 1 if $t < t_1$ and

$$\hat{S}(t) = \prod_{j:t_j \leq t} \left(\frac{N_j - y_j}{N_j} \right)$$

1. Greenwood's formula

An estimate of the variance of $\hat{S}(t)$ based on Greenwood's formula (application of Delta method) is:

$$\hat{V}ar(\hat{S}(t)) = \hat{S}(t)^2 \sum_{j:t_j \leq t} \frac{y_j}{N_j(N_j - y_j)}$$

A 95% confidence interval for $S(t)$ can be derived as:

$$\hat{S}(t) \pm 1.96 \sqrt{\hat{V}ar(\hat{S}(t))}$$

with imposing the constraint that the confidence interval lies in $[0, 1]$, i.e. if the bounds of the confidence interval go outside $[0, 1]$, set the values to 0 or 1, respectively. This is unappealing in many respects!

2. Variance based on the complementary Log-Log transformation

An alternative to Greenwoods formula for the variance, a variance estimate can be derived based on the complementary Log-Log transformation.

Let $v(t) = log[-logS(t)]$. Note that $S(t) \in [0, 1]$ and $v(t) \in [-\infty, \infty]$.

$$\hat{V}ar(\hat{v}(t)) = \sum_{j:t_j \leq t} \frac{y_j}{N_j(N_j - y_j)} \left[\sum_{j:t_j \leq t} log\left(\frac{N_j - y_j}{N_j}\right) \right]^{-2}$$

The 95% confidence interval for $v(t)$ is given by:

$$\hat{v}(t) \pm 1.96\sqrt{\hat{V}ar(\hat{v}(t))}$$

where we can define the upper and lower bound as $\hat{v}_L(t)$ and $\hat{v}_U(t)$.

NOTE: $S(t) = exp(-exp(v(t)))$, so the 95% confidence interval for $S(t)$ is:

$$[exp(-exp(\hat{v}_U(t))), exp(-exp(\hat{v}_L(t)))]$$

3. Example

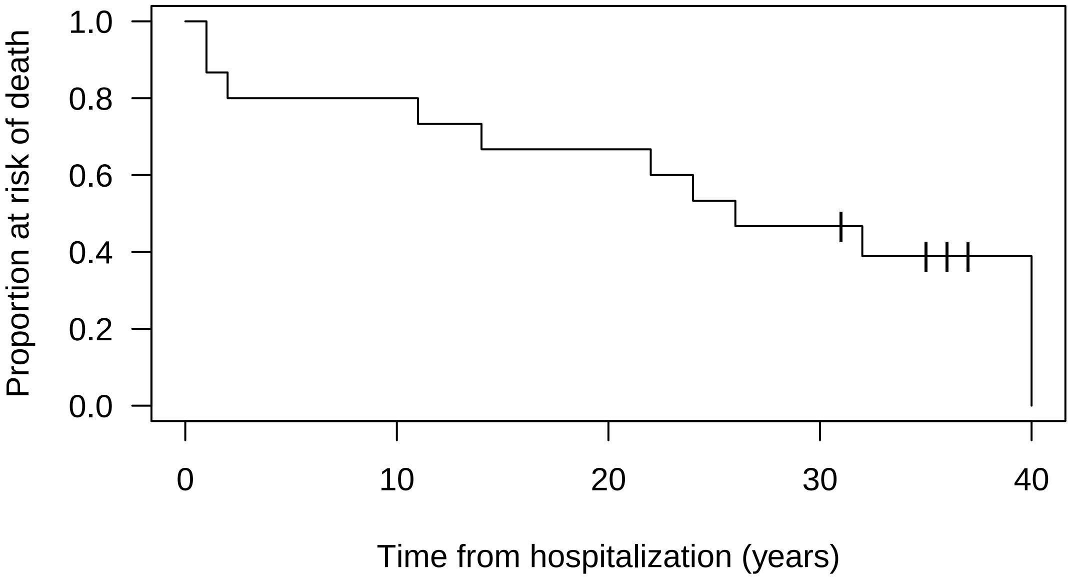
Using the data from Lecture11-Handout for inpatients hospitalized for a severe mental disorder, we will be computing the Kaplan-Meier estimate of the survival function for the female patients. Survival time from hospitalization is in years.

Recall the survival data for females is: 1, 1, 2, 11, 14, 22, 24, 26, 31+, 32, 35+, 35+, 36+, 37+, 40.

NOTE: the + indicates that the patient was censored at that time.

% latex table generated in R 3.6.0 by xtable 1.8-4 package % Mon May 10 17:50:43 2021

	1	1	2	11	14	22	24	26	31+	32	35+	35+	36+	37+	40	Ni	yi	(Ni-yi)/Ni	S(t)
1	1	1	0	0	0	0	0	0	0	0	0	0	0	0	0	15	2	0.867	0.867
2			1	0	0	0	0	0	0	0	0	0	0	0	0	13	1	0.923	0.800
3				0	0	0	0	0	0	0	0	0	0	0	0	12	0	1.000	0.800
4				0	0	0	0	0	0	0	0	0	0	0	0	12	0	1.000	0.800
5				0	0	0	0	0	0	0	0	0	0	0	0	12	0	1.000	0.800
6				0	0	0	0	0	0	0	0	0	0	0	0	12	0	1.000	0.800
7				0	0	0	0	0	0	0	0	0	0	0	0	12	0	1.000	0.800
8				0	0	0	0	0	0	0	0	0	0	0	0	12	0	1.000	0.800
9				0	0	0	0	0	0	0	0	0	0	0	0	12	0	1.000	0.800
10				0	0	0	0	0	0	0	0	0	0	0	0	12	0	1.000	0.800
11				1	0	0	0	0	0	0	0	0	0	0	0	12	1	0.917	0.733
12					0	0	0	0	0	0	0	0	0	0	0	11	0	1.000	0.733
13					0	0	0	0	0	0	0	0	0	0	0	11	0	1.000	0.733
14					1	0	0	0	0	0	0	0	0	0	0	11	1	0.909	0.667
15						0	0	0	0	0	0	0	0	0	0	10	0	1.000	0.667
16						0	0	0	0	0	0	0	0	0	0	10	0	1.000	0.667
17						0	0	0	0	0	0	0	0	0	0	10	0	1.000	0.667
18						0	0	0	0	0	0	0	0	0	0	10	0	1.000	0.667
19						0	0	0	0	0	0	0	0	0	0	10	0	1.000	0.667
20						0	0	0	0	0	0	0	0	0	0	10	0	1.000	0.667
21						0	0	0	0	0	0	0	0	0	0	10	0	1.000	0.667
22						1	0	0	0	0	0	0	0	0	0	10	1	0.900	0.600
23							0	0	0	0	0	0	0	0	0	9	0	1.000	0.600
24							1	0	0	0	0	0	0	0	0	9	1	0.889	0.533
25								0	0	0	0	0	0	0	0	8	0	1.000	0.533
26								1	0	0	0	0	0	0	0	8	1	0.875	0.467
27									0	0	0	0	0	0	0	7	0	1.000	0.467
28									0	0	0	0	0	0	0	7	0	1.000	0.467
29									0	0	0	0	0	0	0	7	0	1.000	0.467
30									0	0	0	0	0	0	0	7	0	1.000	0.467
31									0	0	0	0	0	0	0	7	0	1.000	0.467
32										1	0	0	0	0	0	6	1	0.833	0.389
33											0	0	0	0	0	5	0	1.000	0.389
34											0	0	0	0	0	5	0	1.000	0.389
35											0	0	0	0	0	5	0	1.000	0.389
36													0	0	0	3	0	1.000	0.389
37														0	0	2	0	1.000	0.389
38															0	1	0	1.000	0.389
39															0	1	0	1.000	0.389
40															1	1	1	0.000	0.000



In Lab 7, you will compute the Kaplan-Meier estimate of the survival curve for the male patients!

Compute the 95% confidence interval for $S(2)$:

1. Using Greenwood's formula:

$$\begin{aligned}
\hat{V}ar(\hat{S}(2)) &= \hat{S}(2)^2 \sum_{j:t_j \leq 2} \frac{y_j}{N_j(N_j - y_j)} \\
&= \hat{S}(2)^2 \left[\tfrac{y_1}{N_1(N_1-y_1)} + \tfrac{y_2}{N_2(N_2-y_2)}\right] \\
&= 0.8^2 \left[\tfrac{2}{15\times(15-2)} + \tfrac{1}{13\times(13-1)}\right] \\
&= 0.0107
\end{aligned}$$

95% CI for $S(2)$: $0.8 \pm 1.96 * \sqrt{0.0107} \rightarrow (0.598, 1.003)$

2. Using the Complementary Log-Log transformation

$$\begin{aligned}
\hat{v}(2) &= log(-log(\hat{S}(2))) \\
&= log(-log(0.8)) \\
&= -1.50
\end{aligned}$$

$$\begin{aligned}
\hat{V}ar(\hat{v}(2)) &= \sum_{j:t_j \leq 2} \frac{y_j}{N_j(N_j - y_j)} \left[\sum_{j:t_j \leq 2} log\left(\frac{N_j - y_j}{N_j}\right) \right]^{-2} \\
&= \left[\tfrac{y_1}{N_1(N_1-y_1)} + \tfrac{y_2}{N_2(N_2-y_2)} \right] \left[log\left(\tfrac{N_1-y_1}{N_1}\right) + log\left(\tfrac{N_2-y_2}{N_2}\right) \right]^{-2} \\
&= \left[\tfrac{2}{15 \times 13} + \tfrac{1}{13 \times 12} \right] [log(13/15) + log(12/13)]^{-2} \\
&= 0.335
\end{aligned}$$

95% CI for $v(2)$ is: $\hat{v}(2) \pm 1.96\sqrt{\hat{V}ar(\hat{v}(2))}$ is $-1.50 \pm 1.96\sqrt{0.335}$ is $(-2.63, -0.36)$.

95% CI for $S(2)$ is: $(exp(-exp(-0.36)), exp(-exp(-2.63)))$ is $(0.50, 0.93)$.

Now, the same analysis using R!

```
library(survival)
```

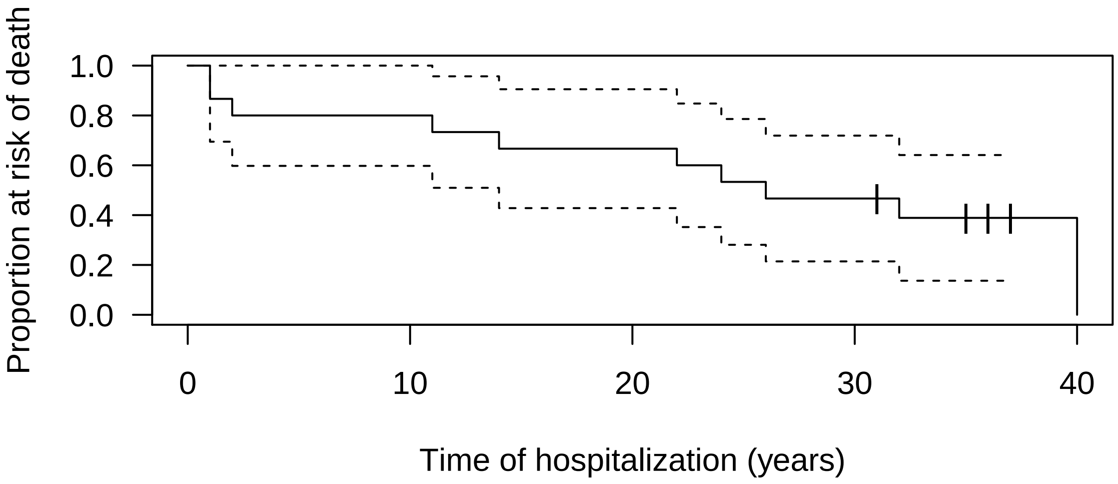
```
## Warning: package 'survival' was built under R version 3.6.3
```

```
St.green = survfit(Surv(survive,event) ~ 1, data = d.female,
                type = "kaplan-meier",
                conf.type = "plain")
St.cll = survfit(Surv(survive,event) ~ 1, data = d.female,
                type = "kaplan-meier",
                conf.type = "log-log")
summary(St.green)
```

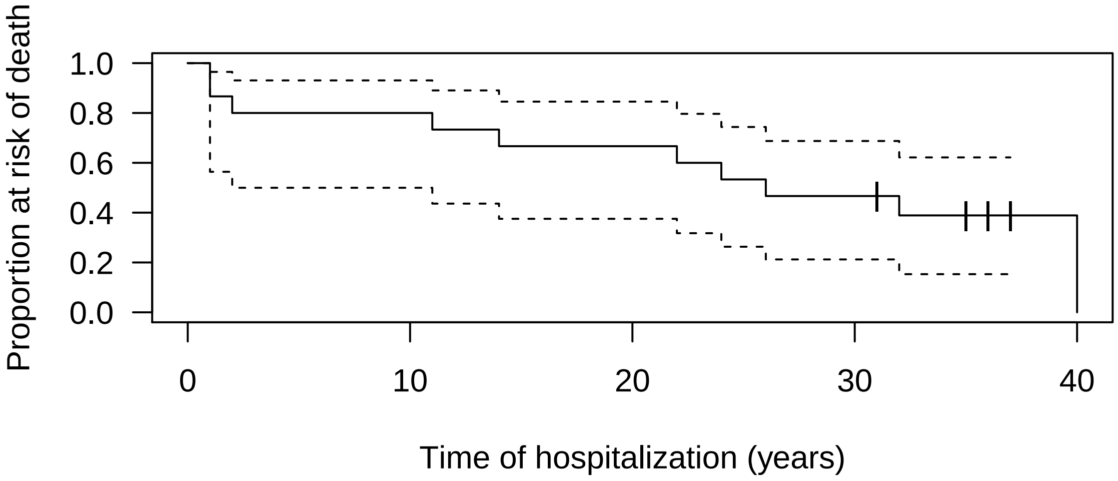
```
## Call: survfit(formula = Surv(survive, event) ~ 1, data = d.female,
##     type = "kaplan-meier", conf.type = "plain")
##
##  time n.risk n.event survival std.err lower 95% CI upper 95% CI
##     1     15       2    0.867  0.0878        0.695        1.000
##     2     13       1    0.800  0.1033        0.598        1.000
##    11     12       1    0.733  0.1142        0.510        0.957
##    14     11       1    0.667  0.1217        0.428        0.905
##    22     10       1    0.600  0.1265        0.352        0.848
##    24      9       1    0.533  0.1288        0.281        0.786
##    26      8       1    0.467  0.1288        0.214        0.719
##    32      6       1    0.389  0.1287        0.137        0.641
##    40      1       1    0.000     NaN          NaN          NaN
```

```
summary(St.cll)
```

```
## Call: survfit(formula = Surv(survive, event) ~ 1, data = d.female,
##     type = "kaplan-meier", conf.type = "log-log")
##
##  time n.risk n.event survival std.err lower 95% CI upper 95% CI
##     1     15       2    0.867  0.0878        0.564        0.965
##     2     13       1    0.800  0.1033        0.500        0.931
##    11     12       1    0.733  0.1142        0.436        0.891
##    14     11       1    0.667  0.1217        0.375        0.846
##    22     10       1    0.600  0.1265        0.318        0.797
##    24      9       1    0.533  0.1288        0.263        0.744
##    26      8       1    0.467  0.1288        0.212        0.687
##    32      6       1    0.389  0.1287        0.153        0.622
##    40      1       1    0.000     NaN           NA           NA
```



Confidence intervals: Greenwoods formula



Confidence intervals: Log(− Log S(t))