Survival Models

Lecture I. Introduction to survival theory

Study materials

Main study materials:

- Tableman & Kim (2003). *Survival Analysis Using S - Analysis of Time-To-Event Data.*
- Meira-Machado, Una-Alvarez, Cardarso-Suarez, Andersen (2007). *Multi-state models for the analysis of time to event data.*

More materials:

- Aalen, Borgan, Gjessing (2008). *Survival and event history analysis. A process point of view.*
- Beyersmann, Schumacher & Allignol (2012). *Competing risks and multistate models with R.*
- Dickson, Hardy, Waters (2013). *Actuarial mathematics for life contingent risks.*
- Hougaard (2001). *Analysis of multivariate survival data.*

Introduction to survival theory

The primary purpose of a survival analysis is to model and analyse time-to-event data, that is, data that have as a principal endpoint the time when an event occurs. Such events are generally called failures (or deaths).

Some examples:

- the time at which an electrical component fails
- time of first recurrence of a tumour after initial treatment
- time to death
- time to the learning of a new skill
- promotion time for employees

Survival analysis is the modern name given to the collection of statistical procedures which accommodate time-to-event censored data.

Motivating example (1). The AML data

Data: from a clinical trial (*Embury, S.H., Elias, L., Heller, P.H., Hood, C.E., Greenberg, P.L., and Schrier, S.L. (1997). Remission maintenance therapy in acute myelogenous leukaemia. Western Journal of Medicine* **126**, *267–272*) to evaluate the efficacy of maintenance chemotherapy for acute myelogenous leukemia (AML)

Goal of the trial: to see if maintenance chemotherapy prolonged the time until relapse

Setup. After reaching a status of remission through treatment by chemotherapy, patients are randomly assigned to two groups:

- first group received maintenance chemotherapy
- second group did not

Group	Length of complete remission (in weeks)
Maintained	9, 13, 13+, 18, 23, 28+, 31, 34, 45+, 48, 161+
Nonmaintained	5, 5, 8, 8, 12, 16+, 23, 27, 30, 33, 43, 45

A + indicates a censored value

Motivating example (2). Throwing out censored obs.

Measure	Maintained	Nonmaintained
Mean	25.1	21.7
Median	23.0	23.0

We observe that

- the mean for maintained group is slightly larger than that for nonmaintained group while their medians are the same
- the distribution of maintained group is slightly more skewed to the right
- the difference between the two groups is negligible

Motivating example (3). Treating censored obs. as exact

Measure	Maintained	Nonmaintained
Mean	38.5	21.3
Median	28.0	19.5

We observe that

- both the mean and median for maintained group are larger
- the difference between the two groups seems to be nonnegligible in terms of both mean and median
- the skewness of the maintained group is even more pronounced

NB! These estimates are biased in that they underestimate the true mean and median. The censored times are smaller than the true unknown failure times.

Motivating example (4). Taking censoring into account

Measure	Maintained	Nonmaintained
Mean	52.6	22.7
Median	31.0	23.0

We observe that

- both the mean and median for maintained group are larger than those for nonmaintained group
- further, the mean of the maintained group is much larger than that of the nonmaintained group
- the distribution of maintained group is much more skewed to the right than the nonmaintained group's distribution
- the difference between the two groups seems to be huge

Basic definitions. Survival function, I

Let us denote

- T – a (continuous) nonnegative rv (lifetime of individuals in a population)
- $F(t) = \mathbb{P}(T \leq t)$ – (cumulative) distribution function (cdf) of T
- $f(t) = F'(t)$ – probability density function (pdf) of T

In survival analysis, the following function plays central role.

Definition (Survival function/survivor function/reliability function)

The probability that an individual survives to time t is given by the survival function:

$$S(t) = \mathbb{P}(T \geqslant t) = 1 - F(t).$$

From the properties of pdf, it immediately follows that

- $S(0) = 1$, $S(\infty) = \lim_{t\to\infty} S(t) = 0$
- $S(t)$ is continuous and non-increasing

Basic definitions. Survival function, II

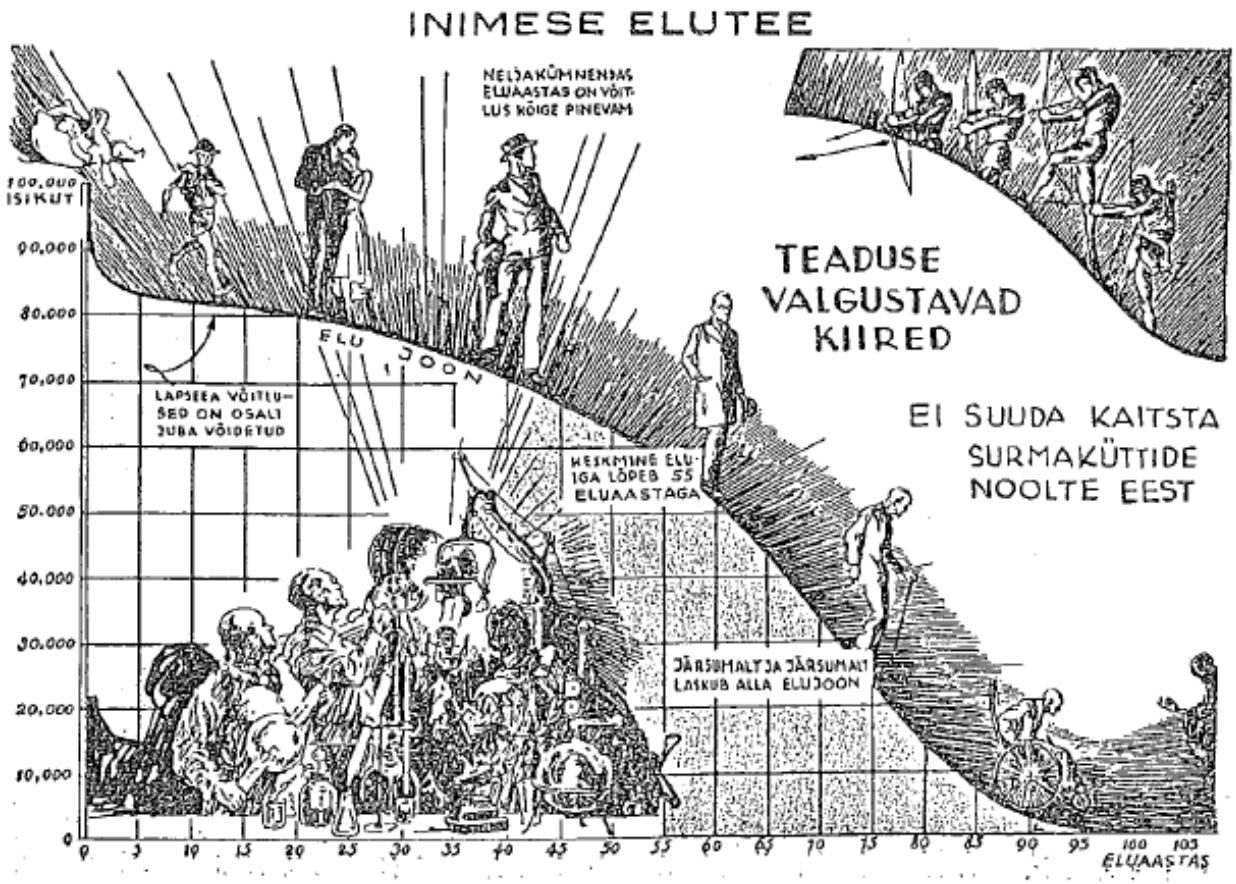


Figure: Survival curve (the graph of the survival function).
Taken from the calendar of insurance company "Eesti" (1935)

Basic definitions. Hazard function

Definition (Hazard function)

The hazard function specifies the instantaneous rate of failure at $T = t$ given that the individual survived up to time t

$$h(t) = \lim_{\Delta t \to 0+} \frac{\mathbb{P}\{t \leqslant T < t + \Delta t | T \geqslant t\}}{\Delta t} = \frac{f(t)}{S(t)}$$

In terms of survival function, we can write

$$h(t) = \lim_{\Delta t \to 0+} \frac{S(t) - S(t + \Delta t)}{\Delta t \cdot S(t)}$$

The hazard function is not a density or a probability. We can think of it as the probability of failure in an infinitesimally small time period between t and $t + \Delta t$ given that the subject has survived up till time t. In this sense, the hazard is a measure of risk: the greater the hazard between times t_1 and t_2, the greater the risk of failure in this time interval

Basic definitions. Cumulative hazard function

Definition (Cumulative hazard function/integrated hazard function)

Integrating the hazard function $h(t)$ over interval $(0, t)$ gives the cumulative hazard function

$$H(t) = \int_0^t h(u)\, du$$

Similarly to the hazard function, the cumulative hazard function is not a probability but is a measure of risk: the greater the value of $H(t)$, the greater the risk of failure by time t.

NB!

Both hazard function and cumulative hazard function define the corresponding distribution uniquely, thus they can be considered as alternative tools for defining a distribution.

Important relations

The following relations between the mentioned functions can be proved:

$$h(t) = -(\ln S(t))'$$

$$H(t) = -\ln(S(t))$$

$$S(t) = \exp(-H(t)) = \exp(-\int_0^t h(u)du)$$

$$f(t) = h(t)S(t) = h(t)\exp(-H(t)) = h(t)\exp(-\int_0^t h(u)du)$$

It can also be shown that

$$E(T) = \int_0^\infty S(t)dt$$

Nature of survival data. Censoring

We saw from the motivating example that the survival data is censored by nature.

Censoring can arise in different situations and can be classified in several ways:

- right censoring – a data point is above a certain value but it is unknown by how much, e.g., indemnity limits or re-insurance agreements in non-life insurance
- left censoring – a data point is below a certain value but it is unknown by how much, e.g. time spent for acquiring a skill in a study
- interval censoring – a data point is somewhere in an interval between two values, e.g., time between visits to a doctor

Censoring (2)

We can also divide the censoring types into categories based on the nature of the experiment:

- Type I censoring – the experiment is terminated at a prespecified time and the number of observed failure times is a rv (engineering applications, environmental data)
- Type II censoring – experiment is run until a prespecified number of failures has occured (also used in engineering)
- Random censoring – subjects enter a study at different times. Consider, e.g., clinical trials. Then censoring can occur in one of the following ways:
  - loss to follow-up – patient moves away and is not seen anymore
  - drop out – patient refuses to continue treatment (or is forced to discontinue because of side effects)
  - termination of study

In this course we

- focus on right censored data (the most common case)
- assume that censoring is random

Censoring (3)

Let us denote

- T_i – lifetime of individual i, $i = 1, \ldots, n$ (with cdf F, pdf f, and survival function S)
- C_i – random censor time with cdf G and pdf g

Thus we can model the observed time Y_i as

$$Y_i = \min\{T_i, C_i\},$$

Let us also define the failure indicator

$$D_i = \begin{cases} 1 & \text{if } T_i \leq C_i \\ 0 & \text{if } C_i < T_i \end{cases}$$

and observe random pairs (Y_i, D_i)

For simplicity, **assume that T_i and C_i are independent**

Censoring (3). Derivation of likelihood

Consider a sample $(y_1, \ldots, y_n, \delta_1, \ldots, \delta_n)$ of observed lifetimes and failure indicators

Then the corresponding likelihood function is

$$L = \prod_{i=1}^{n} [f(y_i)(1 - G(y_i))]^{\delta_i} [S(y_i)g(y_i)]^{1-\delta_i}$$

For derivation, consider two cases separately:

- if $T_i \leqslant C_i$ then the "contribution" of i-th subject to the likelihood is $\mathbb{P}(T_i = y_i, C_i \geqslant y_i)$
- if $T_i > C_i$ then the "contribution" of i-th subject to the likelihood is $\mathbb{P}(T_i > y_i, C_i = y_i)$

Home assignment

Home assignment 1.
Consider a censoring problem with constant censoring time (i.e. $C = c$). This means that the length of the time period when a subject is monitored at most c and if there is no event in that period, the censoring occurs. In other words, we can write

$$Y_i = \min\{T_i, c\}.$$

Derive the formula for likelihood function L in that case.

Empirical estimation of survival function

Now, if the distribution of C does not involve any parameters of interest, we can omit the corresponding terms in the maximization process and focus on the following likelihood function:

$$L = \prod_{i=1}^{n} f(y_i)^{\delta_i} S(y_i)^{1-\delta_i}$$

Now, order the sample (non-decreasing), obtaining $(y_{(1)}, \ldots, y_{(n)})$

Denote

- $I_1 = (0, y_{(1)}]$, $I_i = (y_{(i-1)}, y_{(i)}]$, $i = 2, \ldots, n$ – intervals between deaths/failures
- n_i – number of subjects at risk right before time $y_{(i)}$ (i.e., they did survive up to that time)
- d_i – number of deaths/failures at time $y_{(i)}$
- p_i – probability to survive the interval I_i
- q_i – probability to die in interval I_i, $q_i = 1 - p_i$

Kaplan-Meier estimator

By conditioning we obtain that

$$S(t) = \mathbb{P}(T \geqslant t) \approx \prod_{i:y_{(i)} \leq t} p_i$$

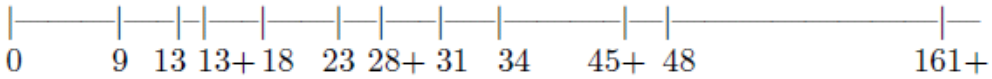
Obvious estimates for q_i and p_i are $\frac{d_i}{n_i}$ and $\frac{n_i - d_i}{n_i}$, implying

$$\hat{S}(t) = \prod_{i:y_{(i)} \leq t} \frac{n_i - d_i}{n_i},$$

which is known as the Kaplan-Meier product limit estimator

Example. Calculation of the K-M estimator for AML data

The calculation of K-M estimator for AML example can be done as follows:



$$
\begin{array}{lcl}
\widehat{S}(0) & = & 1 \\
\widehat{S}(9) & = & \widehat{S}(0) \times \frac{11-1}{11} = .91 \\
\widehat{S}(13) & = & \widehat{S}(9) \times \frac{10-1}{10} = .82 \\
\widehat{S}(13+) & = & \widehat{S}(13) \times \frac{9-0}{9} = \widehat{S}(13) = .82 \\
\widehat{S}(18) & = & \widehat{S}(13) \times \frac{8-1}{8} = .72 \\
\widehat{S}(23) & = & \widehat{S}(18) \times \frac{7-1}{7} = .61 \\
\widehat{S}(28+) & = & \widehat{S}(23) \times \frac{6-0}{6} = \widehat{S}(23) = .61 \\
\widehat{S}(31) & = & \widehat{S}(23) \times \frac{5-1}{5} = .49 \\
\widehat{S}(34) & = & \widehat{S}(31) \times \frac{4-1}{4} = .37 \\
\widehat{S}(45+) & = & \widehat{S}(34) \times \frac{3-0}{3} = \widehat{S}(34) = .37 \\
\widehat{S}(48) & = & \widehat{S}(34) \times \frac{2-1}{2} = .18 \\
\widehat{S}(161+) & = & \widehat{S}(48) \times \frac{1-0}{1} = \widehat{S}(48) = .18
\end{array}
$$

Home assignment

Home assignment 2. Let us have survival data

$$1, 1, 1+, 2, 2+, 3, 5+, 6, 7, 11, 12+, 16,$$

with "+" representing censoring.

Estimate the values $S(3)$ and $S(5)$ of survival function using different methods:

1. leaving out the censored observations (new sample size $n = 8$);
2. using all observations ($n = 12$), but treat censored observations as exact;
3. using Kaplan-Meier method.

Nelson-Aalen estimator, Fleming-Harrington estimator

Note also that $\hat{h}(y_i) = \frac{d_i}{n_i}$, which implies that

$$\hat{H}(t) = \sum_{i:y_{(i)} \leq t} \hat{h}(y_i) = \sum_{i:y_{(i)} \leq t} \frac{d_i}{n_i},$$

which is known as the Nelson-Aalen estimator

Now, using the last result and taking into account the relation between $S(t)$ and $H(t)$, we obtain

$$\hat{S}(t) = \exp(-\hat{H}(t)) = \exp(-\sum_{i:y_{(i)} \leq t} \frac{d_i}{n_i}) = \prod_{i:y_{(i)} \leq t} \exp(-\frac{d_i}{n_i}),$$

which is known as the Fleming-Harrington estimator