Introduction to survival analysis

Jorge N. Tendeiro

Department of Psychometrics and Statistics
Faculty of Behavioral and Social Sciences
University of Groningen

j.n.tendeiro@rug.nl
www.jorgetendeiro.com
jorgetendeiro/Seminar-2020-Survival-Analysis

Plan for today

Gentle introduction to survival analysis.

Source:
Harrell, F. E., Jr. (2015). *Regression Modeling strategies*, 2nd edition. Springer

Chapters:
17, 18, and 20.

Survival analysis (SA)

Data:
For which the *time until the event* is of interest.

- This goes beyond *logistic regression*, which focuses on the *occurrence* of the event.

Outcome variable:

- T = Time until the event.
- Often referred to as *failure time*, *survival time*, or *event time*.

Examples

Survival time: Time until...

- death, desease, relapse.

Failure time: Time until...

- product malfunction.

Event time: Time until...

- graduation, marriage, divorce.

Advantages of SA over typical regression models

- SA allows modeling units that did not fail up to data collection (*censored on the right* data).

- Regression could be considered to model the expected survival time. *But*:
  - ✓ Survival time is often not normally distributed.
  - ✓ $P(\text{survival} > t)$ is often more interesting than $\mathbb{E}(\text{survival time})$.

Censoring

- For some subjects, the event did not occur up to the end of data collection.
- These data are right-censored.

Define random variables for the ith subject:

- $T_i =$ time to event
- $C_i =$ censoring time
- $e_i =$ event indicator = $\begin{cases} 1 & \text{if event is observed } (T_i \leq C_i) \\ 0 & \text{if event is not observed } (T_i > C_i) \end{cases}$
- $Y_i = \min(T_i, C_i) =$ what occurred first (failure or censoring)

Variables $\{Y_i, e_i\}$ include all the necessary information.

Typical data set

T_i	C_i	Y_i	e_i
5	10	5	1
4	12	4	1
13+	13	13	0
5	10	5	1
15+	15	15	0

Observe the flexibility of SA data:

- Subjects may join the study at different moments.
- Censoring times may differ among subjects.

$\{Y_i, e_i\}$ does include all the necessary information.

Three main functions

Recall that the outcome variable is $T =$ time until event.

- Survival function:

$$S(t) = P(T > t) = 1 - F(t),$$

  where $F = P(T \leq t)$ is distribution function of T.

- Cumulative hazard function:

$$\Lambda(t) = -\log(S(t))$$

- Hazard function:

$$\lambda(t) = \Lambda'(t)$$

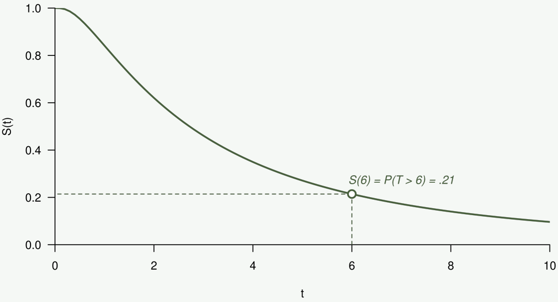
Survival function

$$\boxed{S(t) = P(T > t) = 1 - F(t)}$$

Example:
If event = death, then $S(t)$ = prob. that death occurs after time t.

Properties:

- $S(0) = 1, S(\infty) = 0$.
- Non-increasing function of t.



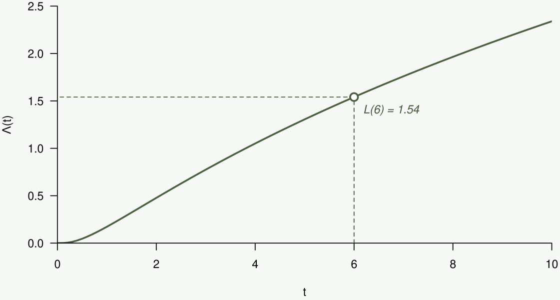
Cumulative hazard function

$$\boxed{\Lambda(t) = -\log(S(t))}$$

Idea:
Accumulated risk up until time t.

Properties:

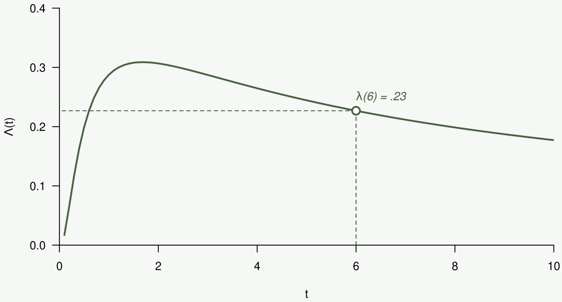
- $\Lambda(0) = 0$.
- Non-decreasing function of t.



Hazard function

$$\boxed{\lambda(t) = \Lambda'(t)}$$

Idea:
Instantaneous event rate at time t.



Relation between the three functions

All functions are related:
Any two functions can be derived from the third function.

- The three functions are equivalent ways of describing the same random variable ($T =$ time until event).

More generally, all the following functions give mathematically equivalent specifications of the distribution of T:

- $F(t)$: Distribution function
- $f(t)$: Density function
- $S(t)$: Survival function
- $\lambda(t)$: Hazard function
- $\Lambda(t)$: Cumulative hazard function.

Examples

Next are two primary examples of parametric survival distributions:

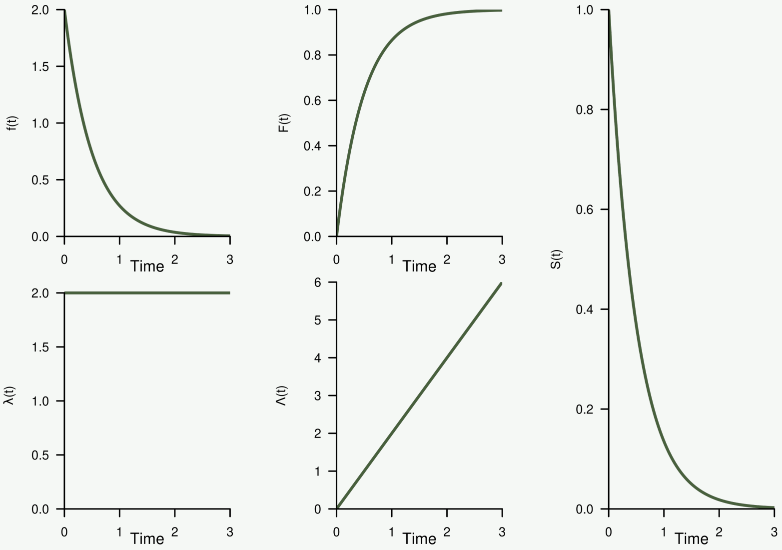
- the exponential distribution;
- the Weibull distribution.

These models (still) include no covariates, thus:

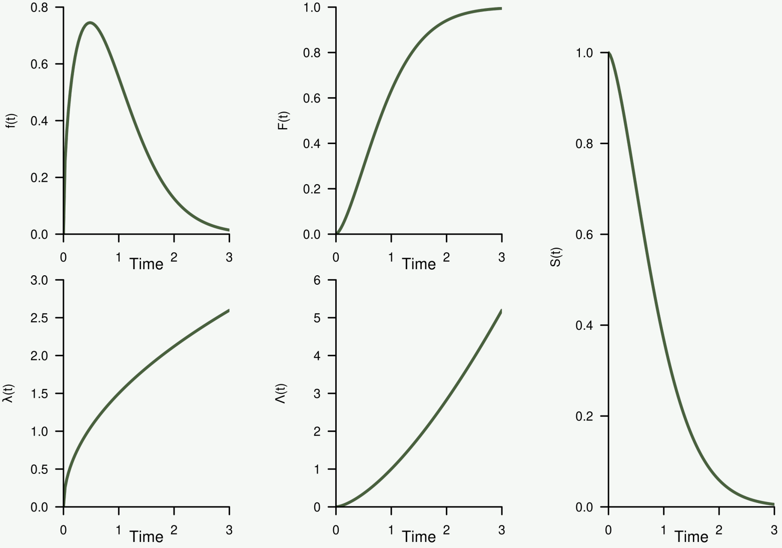
- Each subject in the sample is assumed to have the same distribution of T.

No formulas.
Instead: Let's plot.

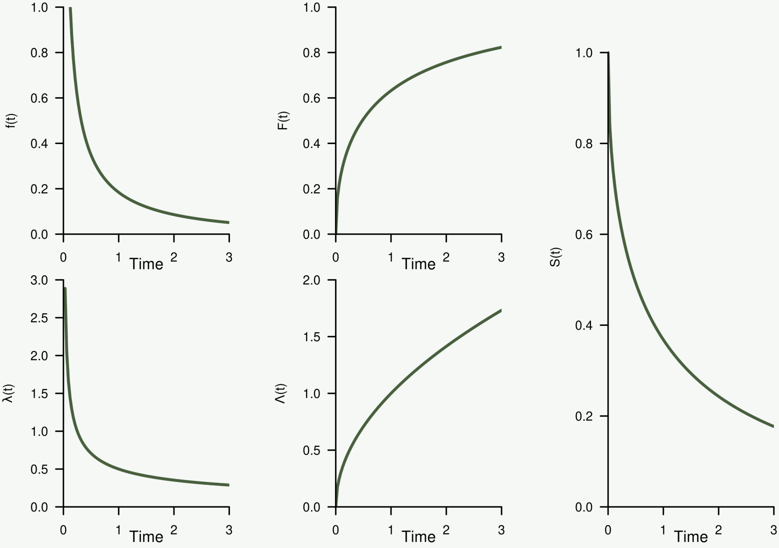
Exponential survival distribution



Weibull survival distribution (I)



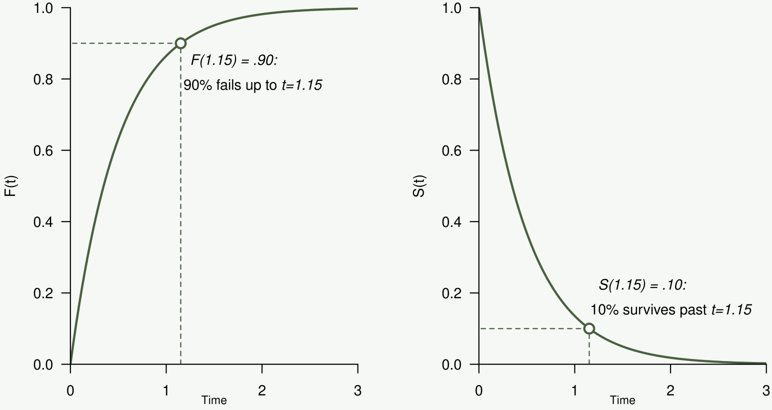
Weibull survival distribution (II)



Quantiles

Q: What is the time by which $(100q)\%$ of the population will fail?

A: Value t_q such that $F(t_q) = q$, or, equiv., $S(t_q) = 1 - q$.



In particular, median survival time = $t_{.50}$.

Expected failure time

(Note: T is skewed, so the mean is not the best summary. Better use median.)

Q: What is the expected failure time?

A: It is the area under the survival function.