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:

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

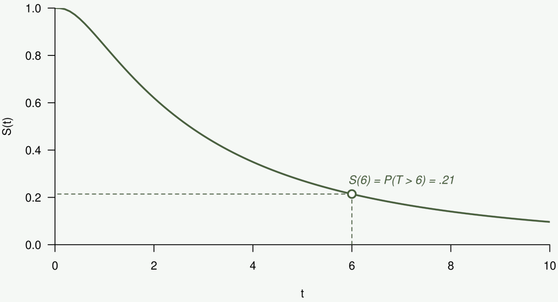
Survival function

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

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.



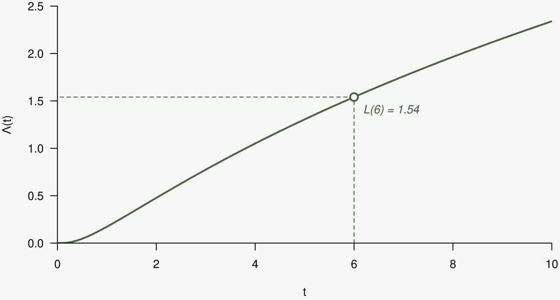
Cumulative hazard function

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

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:

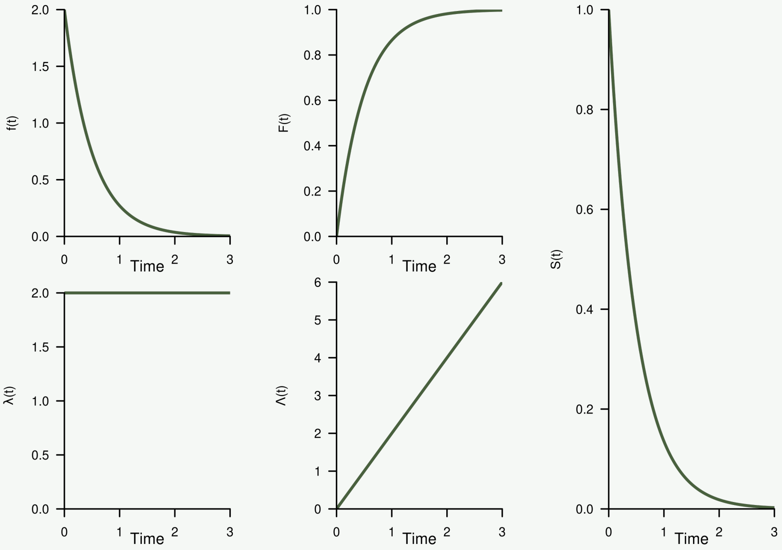
- the exponential distribution;
- the Weibull distribution.

These models (still) include no covariates, thus:

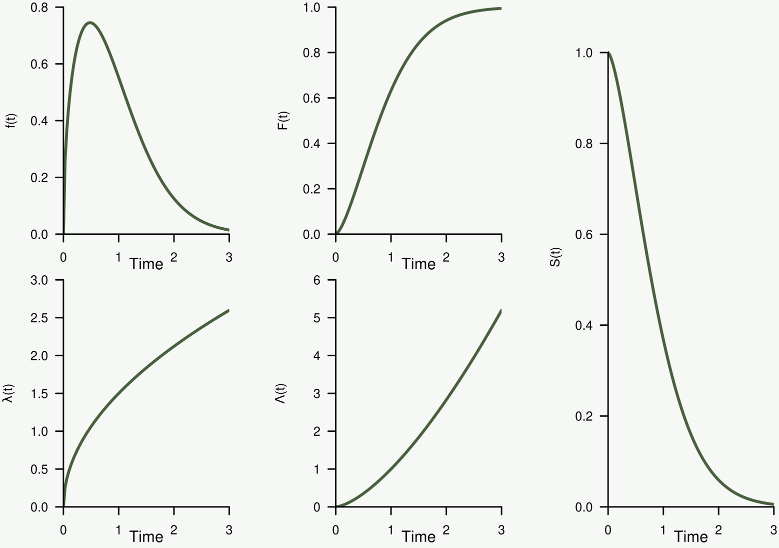
- Each subject in the sample is assumed to have the same distribution of T.

No formulas.
Instead: Let's plot.

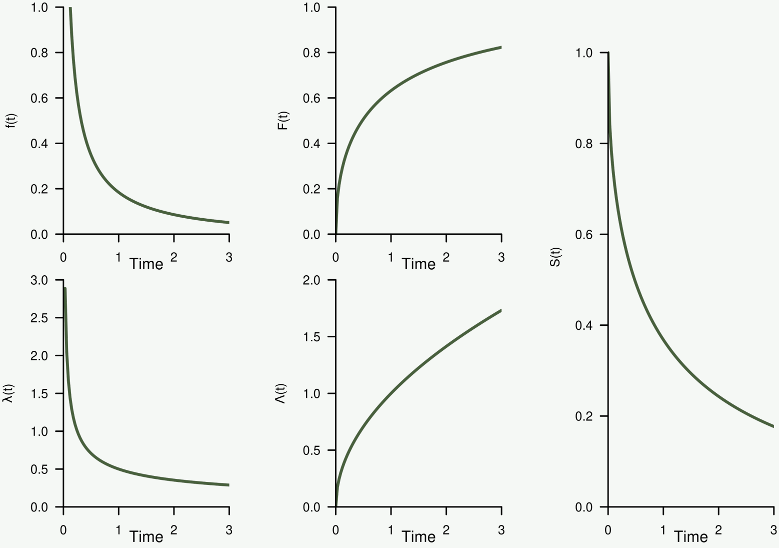
Exponential survival distribution



Weibull survival distribution (I)



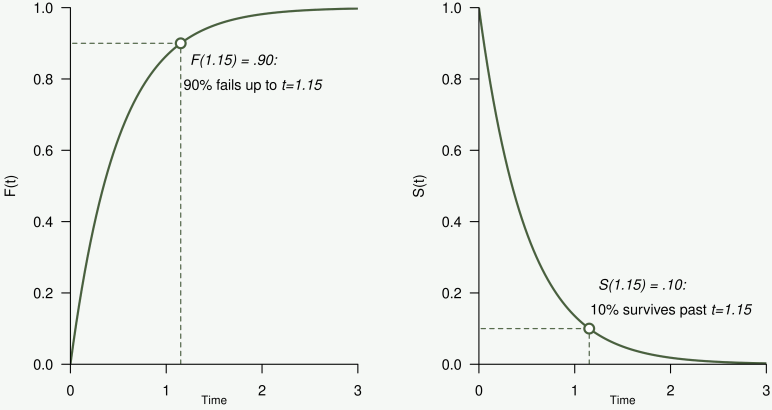
Weibull survival distribution (II)



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., $S(t_q) = 1 - q$.



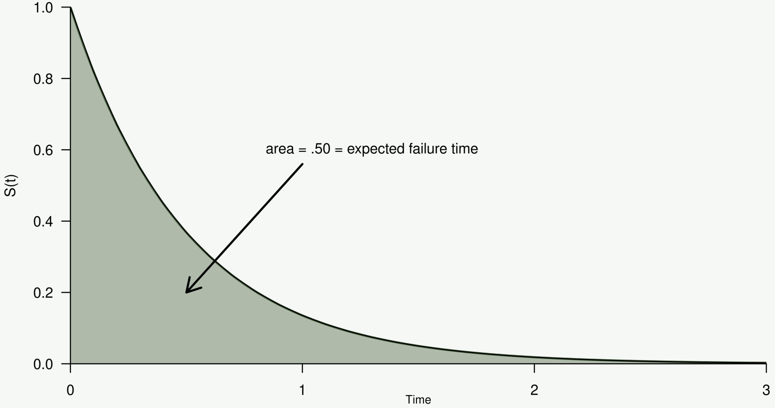
In particular, median survival time = $t_{.50}$.

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.



Various estimation approaches

There are several options available to estimate the survival function (and friends).

Here we will briefly go through only a few:

- Not parametric:
  - ✓ Kaplan-Meier estimator
  - ✓ Altschuler-Nelson estimator
- Parametric:
  - ✓ Proportional hazards models
- Semi-parametric:
  - ✓ Cox proportional hazards regression model

After a brief intro to each, I will use them all on an empirical dataset.

Kaplan-Meier estimator

- Also known as the *product-limit* estimator.
- Non parametric, and super simple to do even manually.
- Key ingredient: *Conditional probabilites*.

Assume $t = 0, 1, 2, \ldots$
We have that $S(0) = P(T > 0) = 1$. For $t \geq 1$ we then have that

$$P(T > t | T > t-1) = \frac{P(T > t, T > t-1)}{P(T > t-1)} = \frac{P(T > t)}{P(T > t-1)}$$

and so

$$P(T > t) = P(T > t-1) \times P(T > t | T > t-1),$$

or in terms of the survival function,

$$S(t) = S(t-1) \times P(T > t | T > t-1)$$

$$\boxed{S(t) = S(t-1) \times (1 - P(T \leq t | T > t-1))}$$

Kaplan-Meier estimator – Example

Data: Seven subjects; failure times $T = 1, 3, 3, 3+, 6+, 9, 10+$.

Day	No. subjects at risk	Deaths	Censored	$S(t) = S(t-1)\times$ $\times(1 - P(T \le t \mid T > t-1))$
1	7	1	0	$1 \times (1 - 1/7) = 6/7$
3	$7 - (1+0) = 6$	2	1	$6/7 \times (1 - 2/6) = 4/7$
6	$6 - (2+1) = 3$	0	1	$4/7 \times (1 - 0/3) = 4/7$
9	$3 - (0+1) = 2$	1	0	$4/7 \times (1 - 1/2) = 2/7$
10	$2 - (1+0) = 1$	0	1	$2/7 \times (1 - 0/1) = 2/7$

Hence:

$$S(t) = \begin{cases} 1, & 0 \le t < 1 \\ 6/7 = .86, & 1 \le t < 3 \\ 4/7 = .57, & 3 \le t < 9 \\ 2/7 = .29, & 9 \le t < 10 \end{cases}.$$

Kaplan-Meier estimator – Example