Introduction to survival analysis

What is survival analysis?

Survival analysis is a class of statistical methods for studying the occurrence and timing of events. These are most often applied to the study of deaths. In fact, they were originally designed for that purpose, which explains the name survival analysis.

The methods of survival analysis have been adapter by researchers in several different fields, they also go by several different names: event history analysis (sociology), reliability analysis (engineering), failure time analysis (engineering), duration analysis (economics), and transition analysis (economics).

What is survival data?

Survival analysis was designed for longitudinal data on the occurrence of events. An event is defined as a qualitative change that can be situated in time. By a qualitative change, it’s meant a transition from one discrete state to another. To apply survival analysis, it is necessary to know when the change occurred.

For survival analysis, the best observation plan is prospective. A set of individuals is being observed at some well-defined point in time, and followed for some substantial period of time, recording the times at which the events of interest occur. It's not necessary that every individual experience the event.

Why use survival analysis?

Survival data have two common features that are difficult to handle with conventional statistical methods:

- censoring
- time-dependent covariates (also called time-varying explanatory variables)

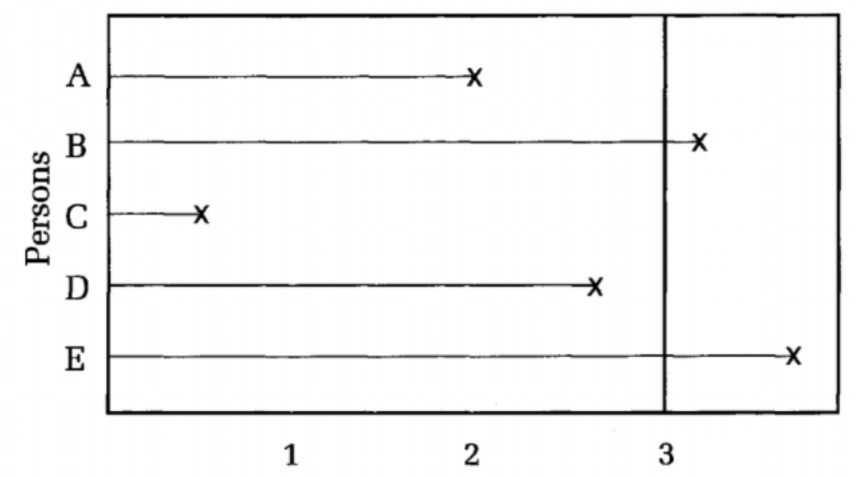
In the case of censoring, the the aim is to combine the information in the censored and uncensored cases in a way that produces consistent estimates of the parameters of interest.

Censoring

In both the natural and social sciences, right censoring is the most common type of censoring.

An observation on a variable T is right censored if all that is known about T is that it is greater than some value c. In survival analysis, T is typically the time of occurrence for some event, and cases are right censored because observation is terminated before the event occurs.

Example



The figure depicts the observation window. An X indicates that an event occurred at that point in time. The vertical line at 3 is the point at which we stop following the observation units.

Persons A, C, and D have uncensored event times, while persons B and E have right-censored event times.

Describing survival distributions

Cumulative distribution function

$$F(t) = \Pr(T \leq t)$$

The CDF gives the probability that the variable T will be less than or equal to any value t.

Survivor function

$$S(t) = \Pr(T > t) = 1 - F(t)$$

The S gives the probability of surviving beyond t.

Properties:

- is bounded by 0 and 1
- S(0) = 1 (as T can not be negative)
- with t getting larger, S never increases (and usually decreases)

Describing survival distributions

Probability density function

$$f(t) = \frac{dF(t)}{dt} = -\frac{dS(t)}{dt}$$

The PDF is a derivative or slope of CDF. The PDF most directly corresponds to notions of distributional shape.

Hazard function

$$h(t) = \lim_{\Delta t \to 0} \frac{\Pr(t \leq T < t + \Delta t | T \geq t)}{\Delta t}$$

The aim is to quantify the instantaneous risk thet an event will occur at time t. Since time is continuous, the probability that an event will occur at exactly time t is 0. Thus consider the probability that an event occurs in the small interval between t and t+Δt. Additionaly the probability is conditional on the individual surviving t time t.

Properties of hazard

Hazard is not a probability since the hazard can be greater than 1. Although the hazard has no upper bound, it can not be less than 0.

Hazard is an unobserved qunatity. It may be estimated with data, but that’s only an estimate.

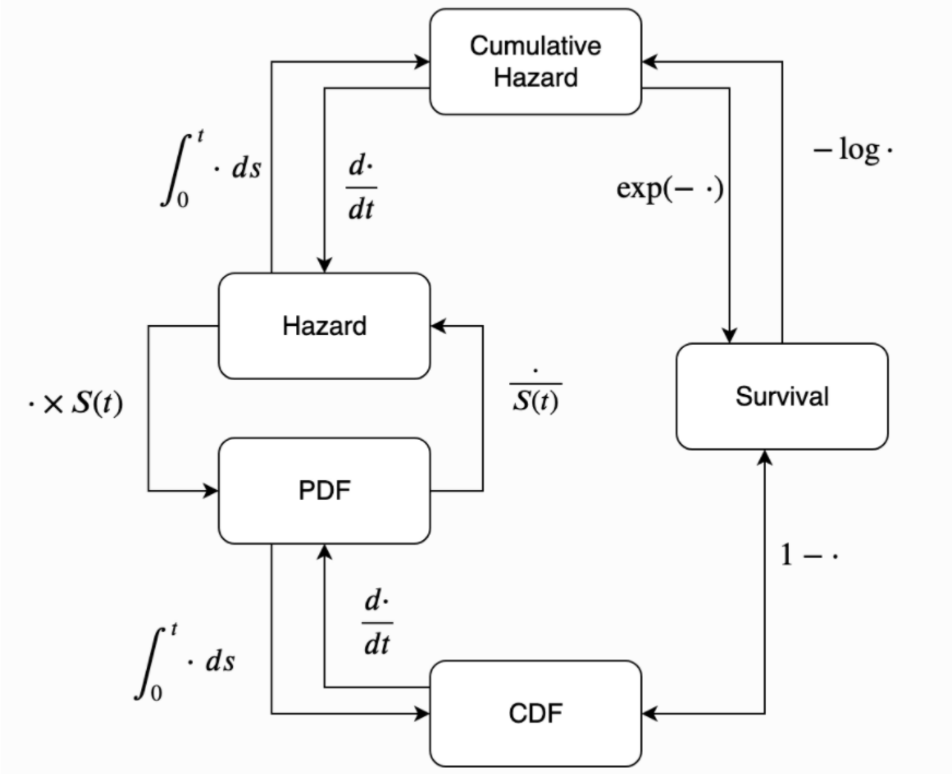
Hazard should be thought as a characteristic of individuals, not of populations or samples. Each individual may have a hazard function that is completely different from anyone’s else.

Describing survival distributions

The survivor function, the probability density function and the hazard function are equivalent ways of decribing a continuous probability distribution.

Given any one of them, the other one can be recovered.

Map of the mathematical entities used in survival analysis and the transforms between them



Source: https://lifelines.readthedocs.io/en/latest/Survival%20Analysis%20intro.html

Data structure

Basic data structure for survival analysis:

- **duration/time variable** – contains either the time that an event occured or, for censored cases, the last time at which that case was observed, both measured from the chosen origin
- **censoring variable** – its arbitrary values indicate the status of the individual at the time recorded in the time variable; it is common to have value 1 for the uncensored cases and 0 for censored cases (in case of one type of event)

BMT dataset in SASHELP library

At the time of bone marrow transplant (BMT), each patient is classified into one of three risk categories: ALL (acute lymphoblastic leukemia), AML-Low Risk (acute myelocytic leukemia, low risk), and AML-High Risk. The endpoint of interest is the disease-free survival time, which is the time in days to death, relapse, or the end of the study.

Variables:

- Group - the patient’s risk category, the variable
- T - the disease-free survival time
- Status - the censoring indicator (1 - an event time, 0 - a censored time)

Life-table method (actuarial method)

With LT method event Times are grouped into intervals of default or set size. In addition, it can produce esimates and plots of the hazard function, which are not available with KM method.

The downside is that the choice of intervals is usually arbitrary, leading to arbitrariness in the results and possible uncertainty about how to choose the intervals. There is inevitably some loss of information as well.

Effective sample size:

$$n_i' = n_i - w_i/2$$

where w_i is the number of units censored in the interval

Kaplan-Meier method (product-limit method)

KM estimator is defined for any time between 0 and the largest event or censoring time. It only changes at an observed event time.

The Kaplan-Meier method is most suitable for smaller data sets with precisely measured event times.

The life-table or actuarial method may be better for large data sets or when measurement of event times is crude.

Testing for differences in survivor functions

Homogeneity tests allow to examine whether a given variable has an influence on the shape of the survival function of the studied individuals. Thanks to this, non-parametric methods (actuarial and K-M) are also used in preliminary, descriptive data analysis and pre-selection of variables used in further analysis, e.g. in parametric models.

Available test:

- the log-rank test (the Mantel-Haenszel test)
- the Wilcoxon test
- the likelihood-ratio statistic (under the additional assumption – the event times have an exponential distribution)

- Null hypothesis: The survivor functions are the same in the analysed groups (for all t: $S_1(t) = S_2(t)$)

Stratification in SAS

Using PROC LIFETEST, add STRATA statement after TIME statement. In case of more than two categories add ADJUST statement. TEST statement allows to specify which test should be used.

As a result:

- Separate tabels with estimates for each of the group
- The survivor funtions of each group superimposed on the same axes
- Additional statistics related to testing for differences between the groups