QMWS - Survival Analysis
Introduction to survival analysis

Instructor: Kevin McGregor

York University
Department of Mathematics and Statistics

Survival data

In survival analysis, we are interested in analyzing the time to an event, rather than the actual number of events.

Examples:

- Time to a myocardial infarction for a person put on a particular drug.
- Time to discharge for a patient in a hospital ward.
- Time to death after onset of a disease.

We might also wish to compare the times in different treatment groups.

- Time to a myocardial infarction for a person put on drug A vs the time for a person put on drug B.

Time to event

For each individual i in the study, the survival time T_i is the amount of time that elapses before the event of interest occurs.

- Each T_i is a **random variable**. The support of the random variable is $T_i > 0$.

Alternative names for survival time include **failure time** and **time to event**.

T_i is the **outcome** in survival analysis.

Entry time

Each individual has an **entry time**; that is, when $T_i = 0$.

Entry times depend on the context of the study. The entry time might be the same for each individual in the study or it might not.

- Entry times the same: estimating time to lung disease diagnosis after a gas leak at a factory exposes workers to a toxin.
    - The entry time is the time that the gas leak happens.
- Entry times different: estimating survival times after heart transplant. Event of interest is death.
    - The entry time is the time that the heart transplant takes place.

Survival function

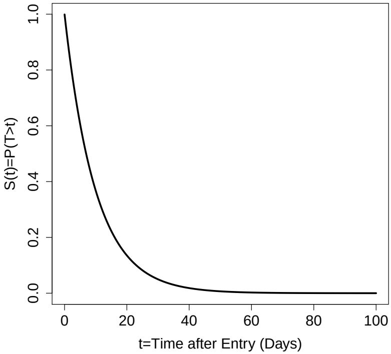
The main function of interest in survival analysis is the survival function. For subject i we denote this as $S_i(t)$. This is the probability that individual i does **not** have an event before time t.

$$S_i(t) = P(T_i > t)$$

$S_i(t)$ is the probability that i is a "survivor" at time t.

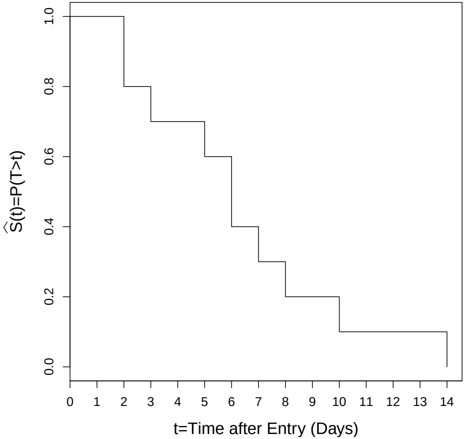
Example of a Survival Curve

- Note $S(0)=1$, we assume that the event occurs at $t > 0$.
- The curve is decreasing: e.g. probability of surviving to 10 days must be less than or equal to the probability of surviving to 5 days.
- $S(t)$ without a subscript i refers to a **population**. It's the proportion of individuals who have survived to time t (i.e. who have not yet had an event).



Estimating Survival Curves from Data

- In real data, the survival curve won't be smooth.
- We can plot it as a step function.
- A drop in the step function at time t represents one or more observed death(s) before time t (but after the last drop).



Simple estimate for survival curve

If every individual in our study had the event of interest (and we observed it), then estimating the survival curve is easy:

$$S(t) = \frac{\texttt{\# individuals surviving to time } t}{\texttt{Total individuals in study}}$$

- Then $S(0) = 1$, since nobody has had an event when $t = 0$.
- $S(t) = 0$ from the last event onwards... since every individual had an event. (i.e. nobody survives after this time)

Censoring

Here is the fundamental problem in analyzing time to event data: **we usually don't observe all events**. This could be because:

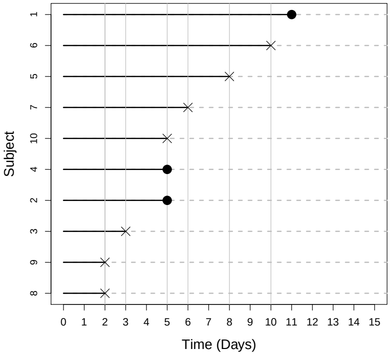
- an individual survives past the end of the study period
- an individual drops out of the study
- competing risk: an alternative event occurred which makes the event of interest impossible. E.g. death in the hospital ward example where discharge is the event of interest.

This problem is called **censoring**.

In this course we will begin with the most common type of censoring, which is called right censoring.

Right Censoring

- Filled dots represent censoring times.
- Crosses represent the events of interest (e.g. death, myocardial infarction, patient discharge, etc.).
- Problem: if we only consider individuals who had an observed event, then we will **underestimate** the survival times.



Accounting for censoring

Censoring **must** be accounted for in time to event data.

If we ignore censoring, we will introduce severe bias into our estimates.

Luckily, censored observations still give us useful information. That is, if an individual is censored at time t_i, we know they did not have an event before time t_i... so $T_i > t_i$.

- Survival analysis methods are designed to handle censoring.

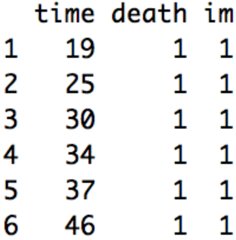
Data example

Data from (Sedmak et al., 1989). Survival times of women with breast cancer along with their immunohistochemical responses.

`time` Time (months) to death or censoring

`death` (1 if dead, 0 if alive)

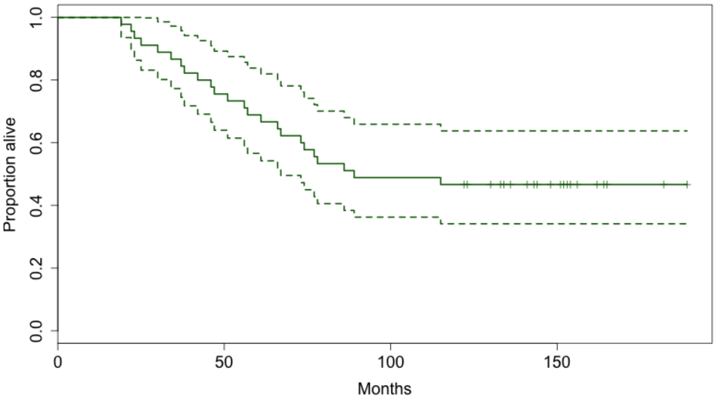
`im` Immunohistochemical response (1=negative, 2=positive)



```
  time death im
1   19     1  1
2   25     1  1
3   30     1  1
4   34     1  1
5   37     1  1
6   46     1  1
```

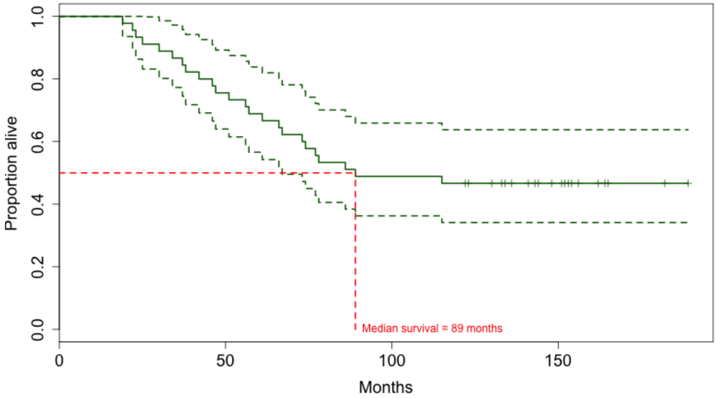
Data example curve

Estimated Kaplan-Meier curve along with 95% confidence bands. Vertical tick marks on curve represent censored women.



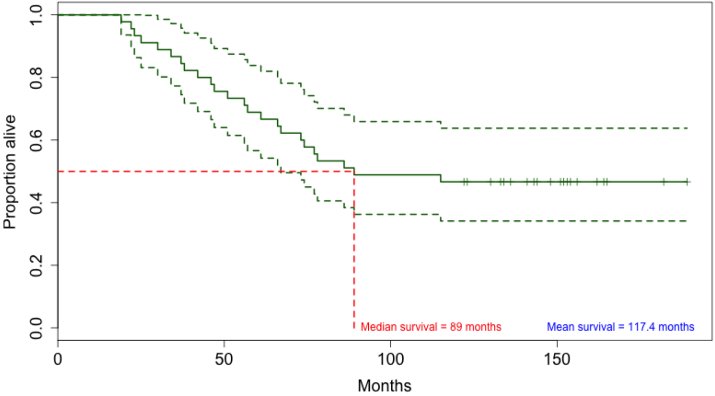
Median survival time

To estimate the **median** survival time, just look at where the survival curve hits 0.5 on the vertical axis.



Mean survival time

To estimate the **mean** survival time, it turns out we just need to take the area under the survival curve.



In this case, the area under the survival curve gives an estimated mean survival time of 117.4 months.

Hazard rate

In survival analysis we often work with something called the **hazard rate**.

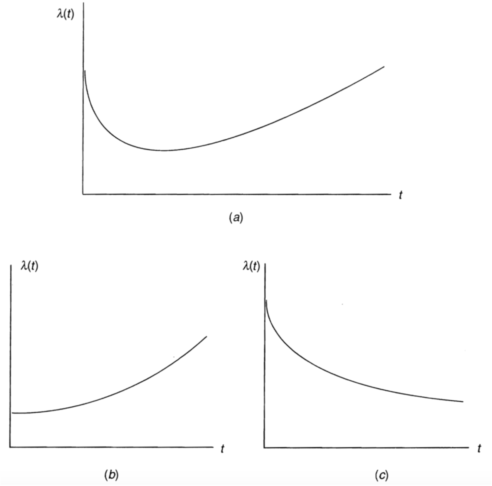
The hazard rate is kind of an abstract quantity. It's defined as the probability that an event happens in a very small interval of time, given that an individual has survived to time t.

$$h_i(t) = P(t < T_i < t + \delta | T_i > t),$$

for an arbitrarily small δ.

The hazard rate is very closely related to the survival function.

Hazard function examples

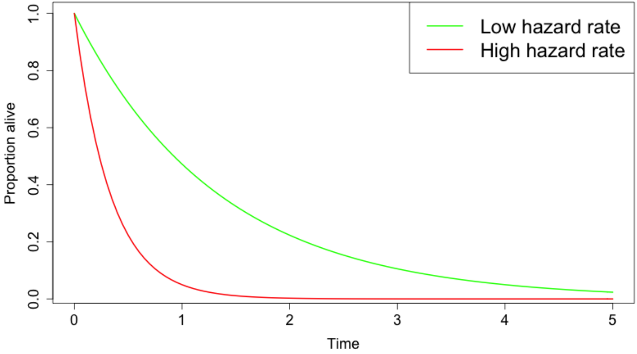


Source: Kalbfleisch, John D., and Ross L. Prentice. The statistical analysis of failure time data. John Wiley & Sons, 2011.

Hazard rate and the survival curve

A lower hazard rate implies a lower chance of an event happening... i.e. a flatter survival curve.

A higher hazard rate implies a higher chance of an event happening... i.e. a steeper survival curve.


Low hazard rate
High hazard rate
Proportion alive
Time