Introduction to Survival Analysis

Survival Data

Typical examples of survival data are

- time to disease progression or death
- time to remission or recovery from a disease
- time from transplant surgery until new organ failure
- time to experience a certain event (birth, menopause, failure of equipment, first employment, divorce, …)

Survival data are also known as *time-to-event data*, duration data, *failure-time data*, reliability data, *event history data* according to different disciplines of fields of application.

The event of interest is called a failure (even if it is a good thing).

The time interval from a starting point and the failure is known as the survival time and is indicated by t.

Survival Data cont’d

Survival data occurs in longitudinal studies used to measure changes that follow research participants over a period of time measuring one or more variables being collected on a group of units.

Typical studies in health and medicine are:

- Cohort studies that recruit and follow participants who share a common characteristic, such as a particular occupation or demographic similarity.

  Example: BCS70 (1970 British Cohort Study) is following the lives of around 17,000 people born in England, Scotland and Wales in a single week of 1970. An example of time-to-event data is the time to the first birth or pregnancy to investigate patterns in early pregnancy (before 19 or earlier) or parenthood.

- Experiential and clinical trials i.e. prospective biomedical or behavioural research studies on (human) participants designed to answer specific questions about biomedical or behavioural interventions, including new treatments and known interventions that warrant further study and comparison.

Aim of survival analysis

Survival Analysis is a set of statistical techniques for analysing the time until a particular event occurs. The aim of survival analysis are

- to estimate the time-to-event (survival) distributions: ✓

  estimation
- to compare time-to-event distributions in different sub-populations: ✓

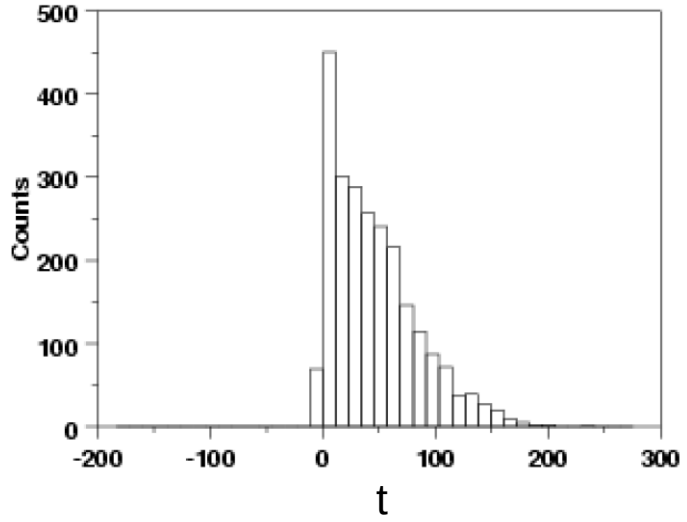
  hypothesis testing
- to determine which factors/covariates influence these distributions: ✗

  regression

Characteristics of survival data

- Survival data are always nonnegative since the variable of interest is time. For this reason, the empirical distribution tends to be right (positively) skewed.

  Data are not normally distributed.

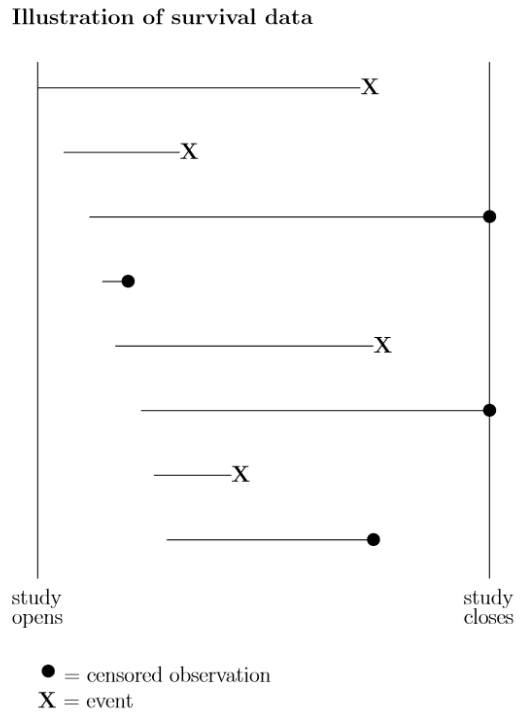


- Typically, some units have censored survival times: the survival times of some subjects are not observed, for example, because the event of interest does not take place for these subjects before the termination of the study.

Censoring

Typically, not all the individuals are observed until their times of failure:

- an organ transplant recipient may die in an automobile accident before the new organ fails
- not everyone gets divorced
- a pancreatic cancer patient may move to Fiji instead of choosing to undergo further treatment
- …



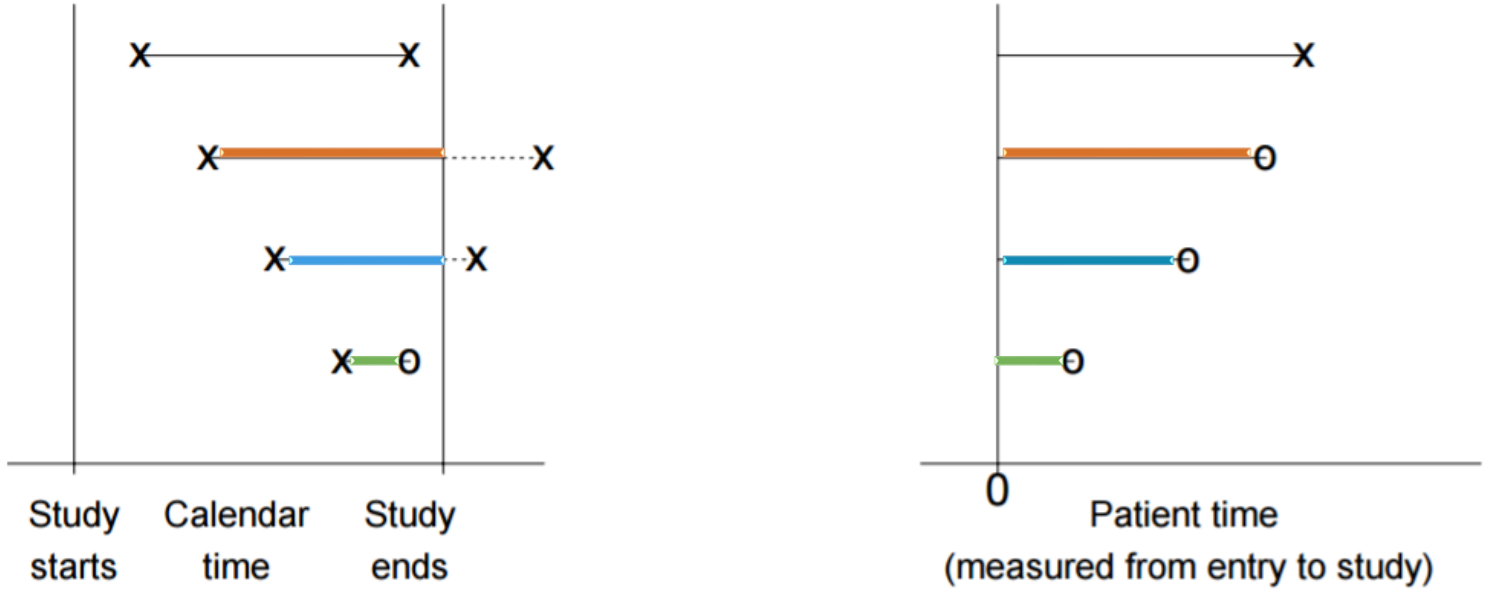
Censored data: the event of interest is not observed for all units in the sample.

Right censoring:

- the event has not occurred when study ended, or data analysis was performed;
- the event of interest has not occurred when the unit left the study. These are known as Loss to follow-up or drop out. This is a kind of missing data.

Study time and unit time

It is important to distinguish between study time and unit (patient) time.



A study may start enrolling patients the 1st of May and continue until 200 patients have been enrolled.

This may take months.

Time is converted to unit time (also called *duration*) i.e. the time between enrolment and failure or censoring.

Survival Function

Assume that

- the time to failure can be represented by a continuous random variable T, $T > 0$
- T is distributed according to a density function $f(t)$ i.e. $T \sim f(t)$

This is the model i.e. the «population» of interest

The survival function is the probability of not experiencing event of interest (“surviving”) up to time t.

$$S(t) = P(T > t) \quad t > 0$$

If $F(t)$ is the cdf of T then

$$S(t) = P(T > t) = 1 - P(T \leq t) = 1 - F(t).$$

Example

Suppose that T is a negative exponential random variable

$$T \sim f(t) = \theta e^{-\theta t} \quad \theta > 0$$

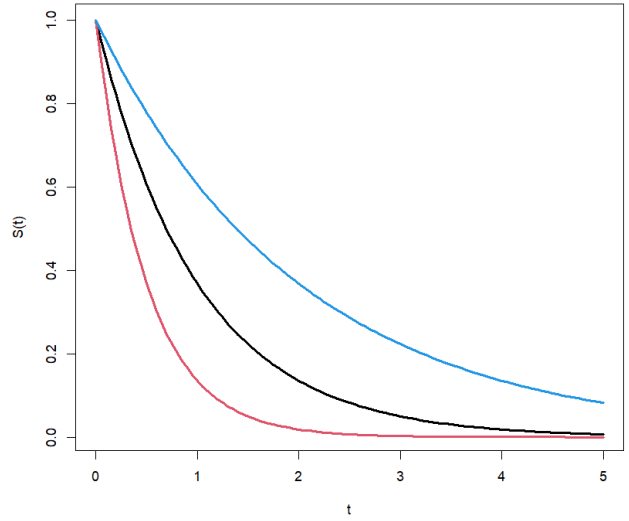
where θ is an unknown parameter.

It holds true that $E(T) = \theta^{-1}$ (expected time to failure)

Hence, θ is the rate at which the failure occurs in time

The survival function is $S(t) = e^{-\theta t}$

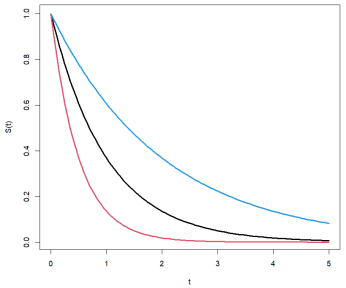
Exponential survival curve for 3 values of θ



Common models for survival time include the Weibull model, the (generalized) Gamma model, the Log-normal model, the Loglogistic model …. . Each of these models has distinct properties and accommodate different characteristics regarding the manner in which units 'fail' over time.

Survival Function and the survival curve

The survival curve represents the survival function graphically i.e. it is the graph of $S(t)$ versus t.



The survival function is a decreasing function such that

$$S(0) = 1 \qquad \text{and} \quad S(t) \to 0 \text{ when } t \to +\infty$$

The survival curve shows the probability of surviving at any given time. Therefore, it can be interpreted as the proportion of those units in the population surviving at any given time.

The median of T is the median survival time and represents the time at which the surviving probability is 0.5.

Example

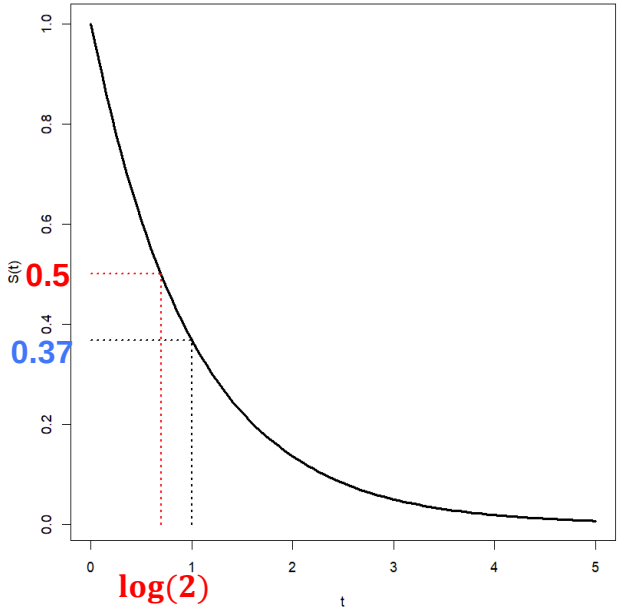
Suppose that T is a negative exponential random variable with $\theta = 1$ i.e. $f(t) = e^{-t}$

The survival function is $S(t) = e^{-t}$.

If T represents time in weeks, the probability of a unit surviving past one week under this model is 0.37.

The median survival time is $\mathbf{\log(2) = 0.6931472.}$

Approximately after 5 days, the considered population is halved.



Estimate of the survival function

However, since θ is unknown, we need to estimate it using data.

Instead of estimating the parameter, we can estimate the entire curve from the data.

In this manner, we do not need to specified a model for T.

This approach is called **nonparametric**.

We consider the **Kaplan-Meier (K-M)** estimator of the survival function.

The Kaplan-Meyer estimator of the survival function

Let's indicate by $\hat{S}(t)$ the estimate of the survival function at time t

Rationale: $\hat{S}(t)$ represents the (estimate of the) probability of not being died up to time t.

Therefore, at the time of the first failure, we can estimate by

$$1 - \frac{\#\, failures}{\#\, at\ risks}$$ (remember the classical definition of probability)

where # denots «number of» and *units at risks* refers to all those units under observation.

Before any failure occurs in the data we have $\hat{S}(t) = 1$.

The Kaplan-Meyer estimator of the survival function cont'd

After the **first failure**, we need to adjust the estimate to account for the reduction in the number of units at risks due to the previous failures.

In fact, we are calculating the conditional probability of surviving beyond the time of the first failure.

When we calculate the survival at the second failure we get

$$\hat{S}(t) = P(survived\ past\ 1st\ \boldsymbol{and}\ 2nd\ times) =$$

$$P(survive\ past\ 1st\ time) \times P(survive\ past\ 2nd\ time \mid survive\ past\ 1st\ time) =$$

$$\left(1 - \frac{\#\ failures, failure\ time\ 1}{\#\ at\ risks, failure\ time\ 1}\right) \times \left(1 - \frac{\#\ failures, failure\ time\ 2}{\#\ at\ risks, failure\ time\ 2}\right)$$

If someone is censored, they are no longer at risk of failing at the next failure time and are taken out of the calculation.

The Kaplan-Meyer estimator of the survival function cont'd

The procedure is repeated iteratively for each subsequent failure, always adjusting the number of units at risk to account for both previous failures and censored observations.

After the second Failure

$$\hat{S}(t) = P(survived\ past\ 1st\ \textbf{and}\ 2nd\ \textbf{and}\ 3rd\ times) =$$

$$= P(survive\ past\ 1st\ time) \times \quad P(survive\ past\ 2nd\ and\ 3rd\ time \mid survive\ past\ 1st\ time) =$$

$$= P(survive\ past\ 1st\ time) \times \quad P(survive\ past\ 2nd\ time \mid survive\ past\ 1st\ time)$$

$$\times \quad P(survive\ past\ 3nd\ time \mid survive\ past\ 1st\ and\ 2nd\ time)$$

$$= \left(1 - \frac{\#\ failures, failure\ time\ 1}{\#\ at\ risks, failure\ time\ 1}\right) \times \left(1 - \frac{\#\ failures, failure\ time\ 2}{\#\ at\ risks, failure\ time\ 2}\right) \times \left(1 - \frac{\#\ failures, failure\ time\ 3}{\#\ at\ risks, failure\ time\ 3}\right)$$

The Kaplan-Meyer estimator of the survival function cont'd

The Kaplan-Meyer (K-M) estimator is defined by

$$\hat{S}(t) = \prod_{t_i \leq t} \left(1 - \frac{d_{t_i}}{R_{t_i}}\right)$$

t_i : time of the i-th event in the sample $i = 1, \ldots, n$ (assuming n failures in the sample)

d_{t_i}: number of events at time t_i ($\# \, failures, time \; t_i$)

R_{t_i}: units at risk at time t_i ($\# \, at \; risks, time \; t_i$)

Note: $\frac{d_{t_i}}{R_{t_i}}$ is the estimated probability of failure at time t_i.

The Kaplan-Meyer estimator of the survival function
Example

At start

At failure 1 (time 4.5)

t	dt (# failures)	# Censored	Rt # Risks	$\hat{S}(t)$
0	0	0	10	
4.5	1	0	9	
7.5	1	0	8	
8.5	0	1	7	
11.5	1	0	6	
13.5	0	1	5	
15.5	1	0	4	
16.5	1	0	3	
17.5	0	1	2	
19.5	1	0	1	
21.5	0	1	0	

t	dt (# failures)	# Censored	Rt # Risks	$\hat{S}(t)$
0	0	0	10	1
4.5	1	0	9	$0.9 = 1 - \frac{1}{10}$
7.5	1	0	8	
8.5	0	1	7	
11.5	1	0	6	
13.5	0	1	5	
15.5	1	0	4	
16.5	1	0	3	
17.5	0	1	2	
19.5	1	0	1	
21.5	0	1	0	

Toy example: sample of 10 units: all units entered the study at 0, all exited the study by time 21.5,

4 were censored at different times and 6 experienced the event (failure)

The Kaplan-Meyer estimator of the survival function
Example cont'd

At failure 2 (time 7.5)

At the end of the study

t	dt (# failures)	# Censored	Rt # Risks	$\hat{S}(t)$
0	0	0	10	1
4.5	1	0	9	0.9
7.5	1	0	8	**0.8**
8.5	0	1	7	**0.8**
11.5	1	0	6	
13.5	0	1	5	
15.5	1	0	4	
16.5	1	0	3	
17.5	0	1	2	
19.5	1	0	1	
21.5	0	1	0	

$$= 0.9 \times \left(1 - \frac{1}{9}\right)$$

$$= 0.8 \times \left(1 - \frac{0}{8}\right)$$

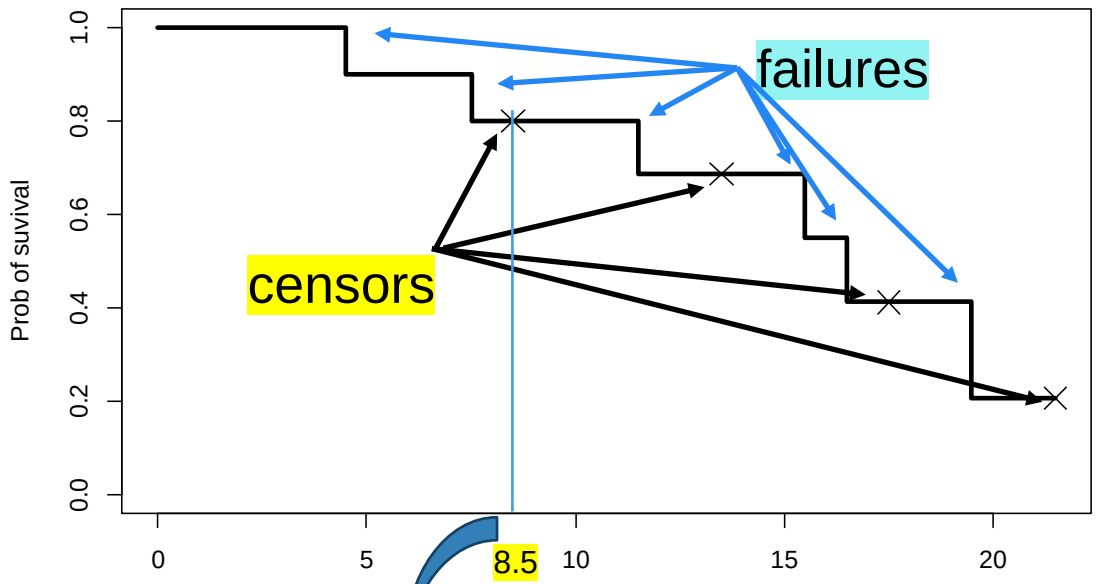
t	dt (# failures)	# Censored	Rt # Risks	$\hat{S}(t)$
0	0	0	10	1
4.5	1	0	9	0.9
7.5	1	0	8	0.8
8.5	0	1	7	0.8
11.5	1	0	6	0.69
13.5	0	1	5	0.69
15.5	1	0	4	0.552
16.5	1	0	3	0.414
17.5	0	1	2	0.414
19.5	1	0	1	0.207
21.5	0	1	0	0.207

The Kaplan-Meyer estimate of the survival function

In between failure times, the K-M estimate does not change but is constant.

This gives the estimated survival function its step-like appearance i.e. it is a step-function.

Step function is defined as a piecewise constant function, that has only a finite number of pieces.



t	dt (# failures)	# Censored
4.5	1	0
7.5	1	0
8.5	0	1
11.5	1	0
13.5	0	1
15.5	1	0
16.5	1	0
17.5	0	1
19.5	1	0
21.5	0	1

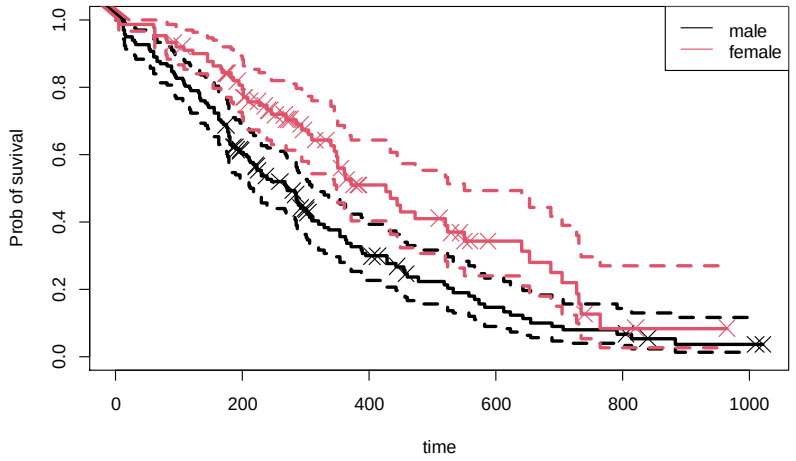
Comparing the K-M survival curves across groups

Assume that the time-to-event is measured on two different groups (i.e. treated and not treated patients, male and female, ….).

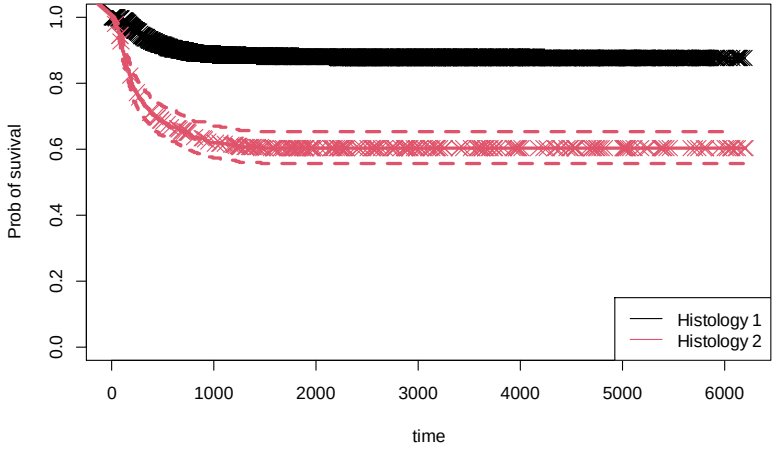
We can compare the survival functions of the two groups by calculating the K-M estimates for the two groups and ploting them on the same graph.

Adding confidence intervals (the maths behind is a bit complicate) of the two curves to the graph helps the comparison.

Survival in patients with advanced lung cancer by gender



Survival to cancer by histology



Comparing survival across groups

CI of the K-M estimates helps in understanding differences in the survival of the two groups. However, they are not a formal statistical test but a kind of graphical diagnostic that permits us to better address differences, accounting for the sampling variability.

A formal procedure would be obtained by using a Wilcoxon-Mann-Whitney test (in case of two groups) for testing differences in the median survival of the two groups.

However, this test is not completely appropriate for time-to-event data. A far better procedure is the so-called log-rank test.

Hazard Function

Let T be distributed according to a density function $f(t)$ i.e. $T \sim f(t)$.

The *hazard (or hazard rate)* is defined as the slope of the survival curve, and it is a measure of how rapidly subjects are dying. Hazard function, indicated by $h(t)$, describes how hazard varies over time

Formally: $$h(t) = \lim_{h \to 0} \frac{1}{h} P(t \leq T \leq t + h | T \geq t) = \frac{f(t)}{S(t)}$$

It is an instantaneous failure rate for observations, conditionally on already having survived to time t.

Useful interpretation.

Suppose T denotes time from surgery for breast cancer until recurrence. Then when a patient who had received surgery visits her physician, she would be more interested in conditional probabilities such as "Given that I haven't had a recurrence yet, what are my chances of having one in the next year?"

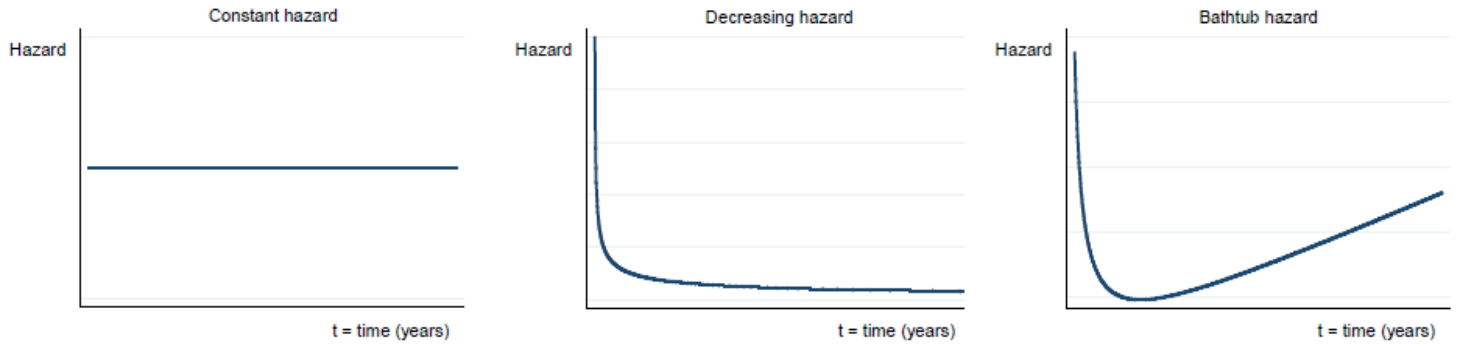
Hazard Function cont'd

$$h(t) = \lim_{h \to 0} \frac{1}{h} P(t \leq T \leq t + h | T \geq t) = \frac{f(t)}{S(t)}$$

- Not a probability (assuming continuous T) although

$$h(t)dt = \frac{f(t)\,dt}{S(t)} \approx P(fail\ in\ [t, t + dt]);$$

- Non-negative and unbounded for all *t*;
- One-to-one relationship between the survival and hazard functions (knowing one tells us exactly what the other is).



Example (exponential model)

Suppose that T is a negative exponential RV with pdf

$f(t) = \lambda e^{-\lambda t}$

and survival function

$S(t) = e^{-\lambda t}$.

Therefore

$$h(t) = \frac{\lambda e^{-\lambda t}}{e^{-\lambda t}} = \lambda$$

i.e. this model has a constant instantaneous failure rate at every time.

Comparing survival across two groups: the log-rank test

Let assume we have two groups and indicate by $S_1(t)$ and $S_2(t)$ the survival functions of the two groups.

The hypothesis we test is

$$\mathrm{H}_0\colon S_1(t) = S_2(t)$$

or equivalently

$$\mathrm{H}_0\colon h_1(t) = h_2(t)$$

Interpretation and procedure: (as in many other tests) we need to

- Calculate the observed value of the test statistics on the data at hand (the two set of survival times).
- Calculate p-value.
- Take a decision: if $p - value < \alpha\ (= 0.05)$ (for instance) reject the null hypothesis.

  In this case there is a statistically significant difference between the two groups: i.e. the survival in one of the two is higher than in the other (more effective treatment).

Comparing survival across two groups: the log-rank test cont'd

The log-rank test is based on the following procedure

Assume that the two subgroups are identified by a binary variable X. At each time point t

- calculate the number of events and units at risk for each of the two groups

	X=1	X=2
Observed	d_{t1}	d_{t2}
At risk	r_{t1}	r_{t2}

- calculate the number of events *expected* for each of the two groups under the null hypothesis that the two survivals are equal

	X=1	X=2
Expected	$\hat{d}_{t1}$	d_{t2}

$\hat{d}_{t1} = d_t \times \frac{r_{t1}}{r_t}$ (same for $\hat{d}_{t2} = d_t \times \frac{r_{t2}}{r_t}$)

where $d_t = d_{t1} + d_{t2}$ total number of observed events

$r_t = r_{t1} + r_{t2}$ total number of units at risk

i.e if the survival is the same in the two groups the expected failures are proportional to the exposed people

- calculate $E_1 = \sum_{t=start}^{end} \hat{d}_{t1}$ and $O_1 = \sum_{t=start}^{end} d_{t1}$ and $A_1 = \frac{(O_1 - E_1)^2}{E_1}$
- repeat for the group $X = 2$ and calculate A_2
- Sum the two: $A_1 + A_2$. This sum is approximately distributed as a χ_1 (under H_0)

Comparing survival across more groups and extensions

The log-rank test procedure can be extended to three or more groups i.e. $X = 1, 2, \ldots, k$

Procedure: at each time

- calculate the number of events and units at risk for each group;
- calculate the number of *expected* events for each group under H_0, assuming all k survivals are equal;
- calculate the previous Observed-Expected statistics for each time and group and sum the obtained values. The chi-square distribution has $k - 1$ df in this case.

Towards regression….

The test, however, cannot be used if there are 2 or more explanatory variables. To compare survival as a function of multiple variables (such as treatment, age, sex, ….) different methodologies need to be adopted (i.e. parametric survival regression models or semiparametric Cox model).