Introduction to Event History/Survival Analysis

Janez Stare

Faculty of Medicine, Ljubljana, Slovenia

Ljubljana, 2017

Some literature

1. David Collett. Modelling Survival Data in Medical Research. Chapman and Hall 2003.
2. David W. Hosmer, Stanley Lemeshow , Susanne May. Applied Survival Analysis. Wiley-Interscience 2008.

Characterization of processes we are interested in

Coleman (1981):

1. there is a collection of units, each moving among a finite number of states;
2. changes (events) may occur at any point in time;
3. there are factors, possibly time-dependent, influencing the events.

We should add

1. effects of covariates may change in time;
2. measurements are often (almost always) censored.

Examples

Source: Blossfeld, Golsch, Rohwer (2007)

medical studies duration of life after diagnosis;

labour market studies workers move between unemployment and employment, full-time and part-time jobs, or among various kind of jobs;

demographic studies durations of marriages or consensual unions;

studies of organizational ecology durations of existence of firms, unions, organizations;

Other names for Event History/Survival Analysis are

- Failure Time Data Analysis
- Reliability Analysis

When would one use the methods of Survival Analysis?

When would one use the methods of Survival Analysis?

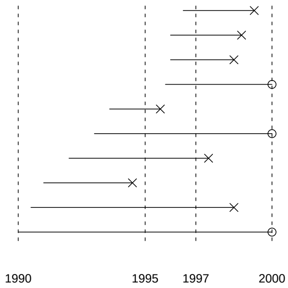
When the outcome of interest is time to some event.

When would one use the methods of Survival Analysis?

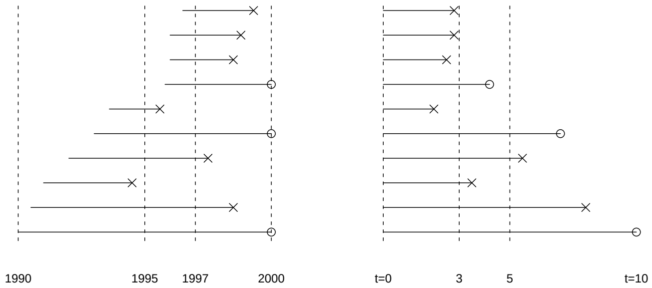
When the outcome of interest is time to some event.

This answer is rather obvious, but **why do we need special methods**?

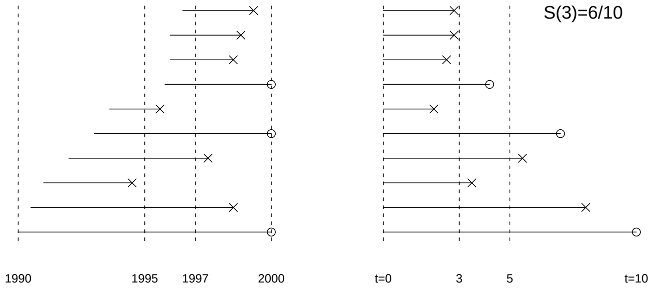
A typical situation



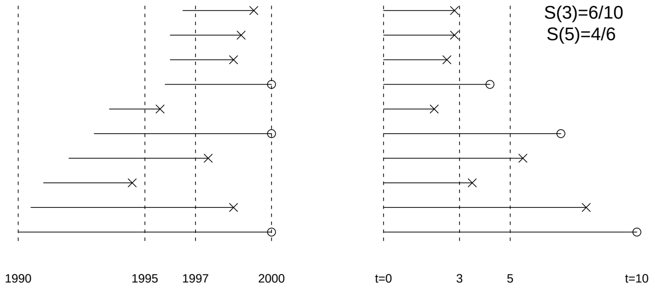
A typical situation



A typical situation



A typical situation



The need for special methods comes from **censoring.** There may be different reasons for censored data:

- lost to follow up
- event of a different type (like death for other reasons)
- end of study (most common)

With censored data we can't even calculate a simple arithmetic mean (in the usual way) or draw a histogram.

With censored data we can't even calculate a simple arithmetic mean (in the usual way) or draw a histogram.

So, the situation seems pretty much hopeless.

With censored data we can't even calculate a simple arithmetic mean (in the usual way) or draw a histogram.

So, the situation seems pretty much hopeless.

Luckily, it is not, although it took some time to come up with methods that deliver want we want.

With censored data we can't even calculate a simple arithmetic mean (in the usual way) or draw a histogram.

So, the situation seems pretty much hopeless.

Luckily, it is not, although it took some time to come up with methods that deliver want we want.

What do we want?

The Goals of Survival Analysis

The Goals of Survival Analysis

1. Estimation of the distribution (survival) function.

The Goals of Survival Analysis

1. Estimation of the distribution (survival) function.
2. Comparison of distribution (survival) functions.

The Goals of Survival Analysis

1. Estimation of the distribution (survival) function.
2. Comparison of distribution (survival) functions.
3. Finding association between the outcome (survival time) and prognostic variables.

Survival function

Formally:

If T is a continuous non-negative random variable with density $f(t)$, then its survival function is

$$S(t) = P(T > t) = 1 - F(t) = \int_t^{\infty} f(x)dx,$$

Survival function

Formally:

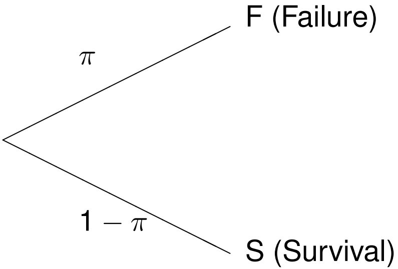
If T is a continuous non-negative random variable with density $f(t)$, then its survival function is

$$S(t) = P(T > t) = 1 - F(t) = \int_t^\infty f(x)dx,$$

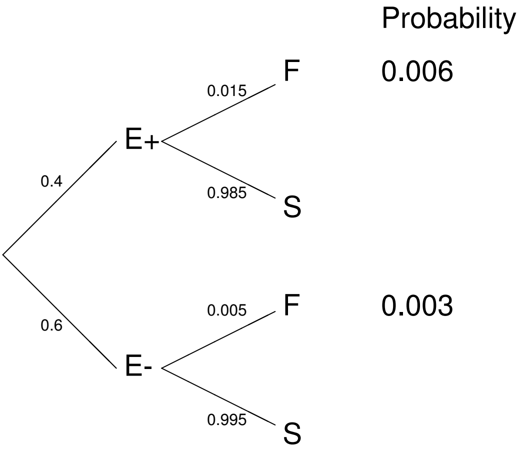
It means:

The value of the survival function at any given time t is the **proportion of people still not experiencing the event (e.g. still alive, still working)** at that time.

Estimating the survival function



Estimating the survival function



Estimating the survival function

More formally, we are using the formula for the probability of a product of events.

Estimating the survival function

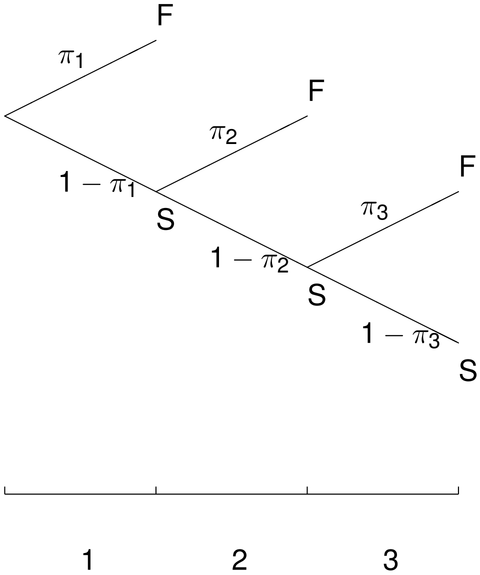
More formally, we are using the formula for the probability of a product of events.

If A and B are two events, then the probability of the product AB is

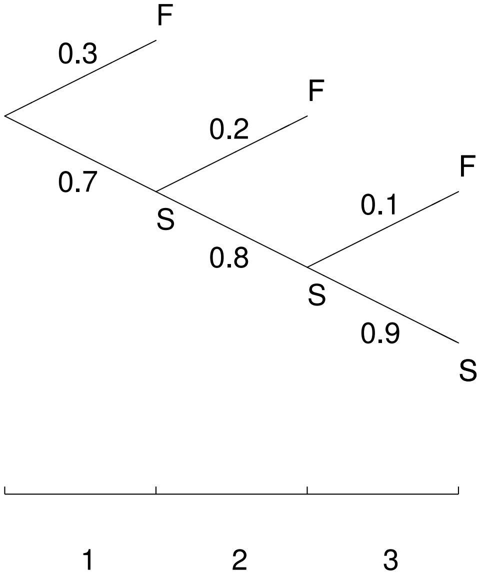
$$P(AB) = P(A)P(B|A)$$

where $P(B|A)$ is the conditional probability of B given A.

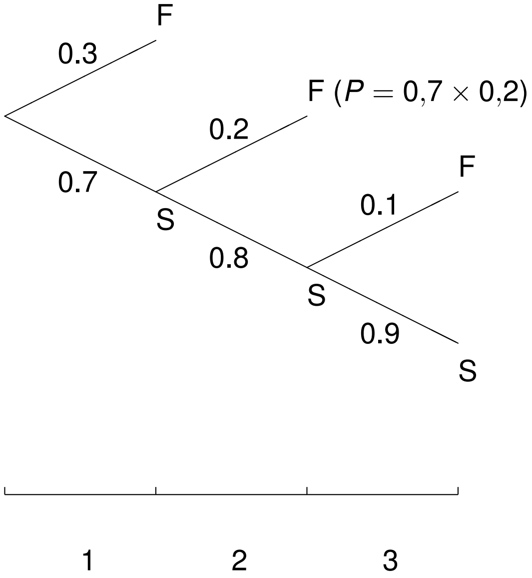
Estimating the survival function



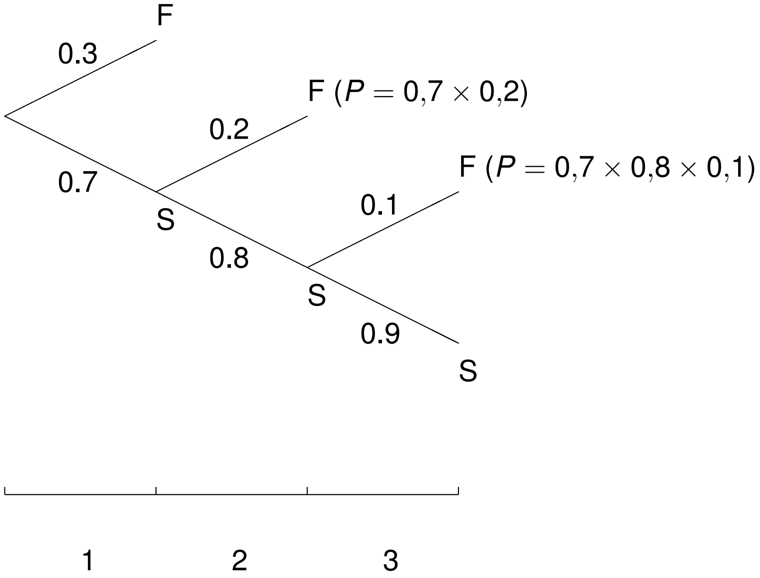
Estimating the survival function



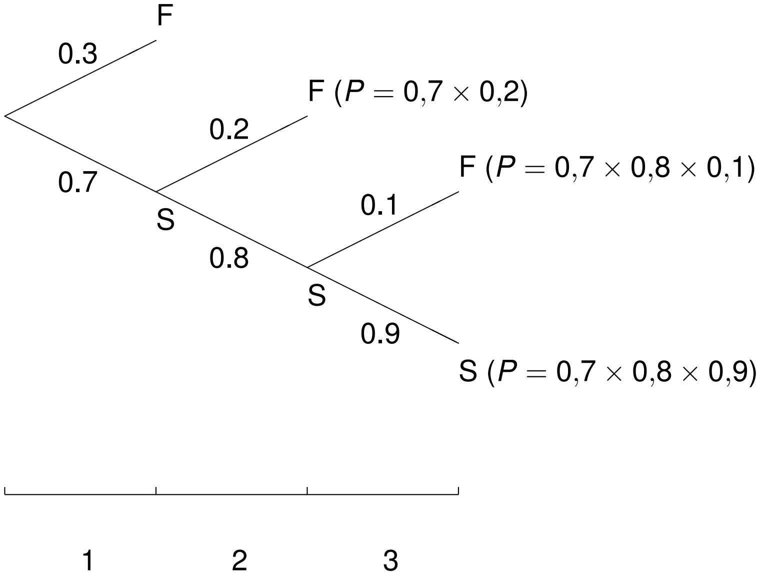
Estimating the survival function



Estimating the survival function



Estimating the survival function



Estimating the survival function

We can use this principle in calculating survival even with **censored data**.

Estimating the survival function

We can use this principle in calculating survival even with **censored data.**

We first divide the time scale into intervals in such a way that events or censorings occur on the boarders of the intervals.

Estimating the survival function

We can use this principle in calculating survival even with **censored data**.

We first divide the time scale into intervals in such a way that events or censorings occur on the boarders of the intervals.

Then we calculate (conditional) probabilities of surviving each interval and obtain probability of surviving any time by simply multiplying the probabilities of survival up to the given point in time.

Estimating the survival function

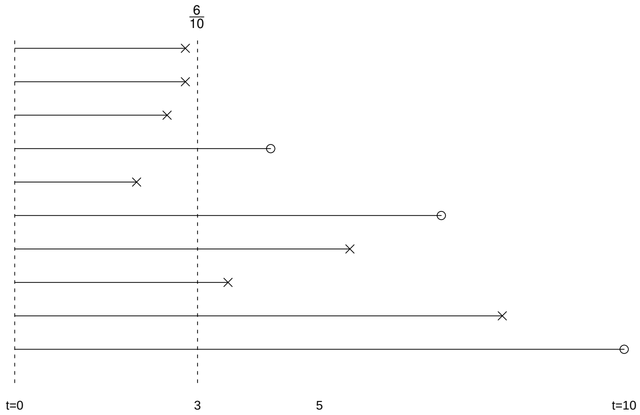
We can use this principle in calculating survival even with **censored data**.

We first divide the time scale into intervals in such a way that events or censorings occur on the boarders of the intervals.

Then we calculate (conditional) probabilities of surviving each interval and obtain probability of surviving any time by simply multiplying the probabilities of survival up to the given point in time.

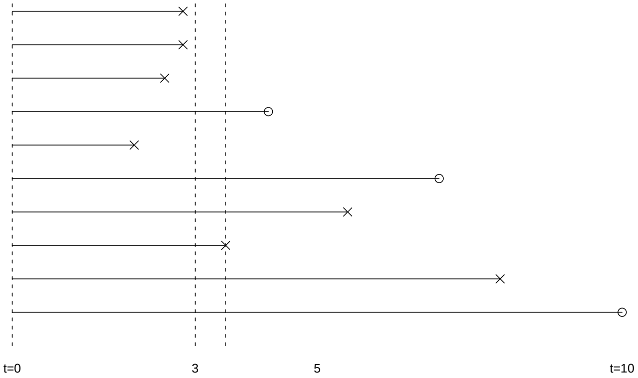
The method is named after **Kaplan and Meier**.

The Kaplan-Meier method



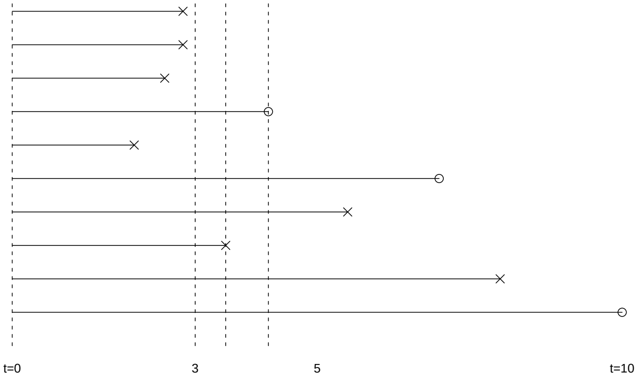
The Kaplan-Meier method

$$\frac{6}{10} \cdot \frac{5}{6} = \frac{5}{10}$$

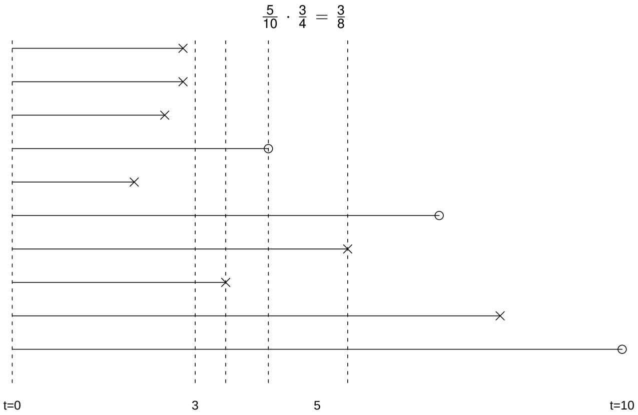


The Kaplan-Meier method

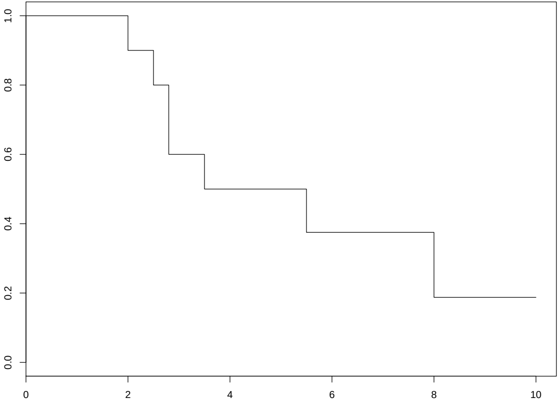
$$\frac{5}{10} \cdot \frac{5}{5} = \frac{5}{10}$$



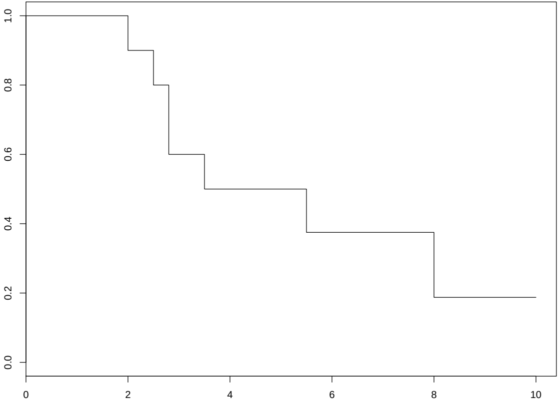
The Kaplan-Meier method



The Kaplan - Meier curve for our example



The Kaplan - Meier curve for our example



What do flat regions on the curve mean?

Another example

Data

Time	At Risk
55	12
61+	11
74	10
81	9
93+	8
122+	7
138	6
151	5
168	4
202+	3
220+	2
238	1

Another example

Data

Time	At Risk
55	12
61+	11
74	10
81	9
93+	8
122+	7
138	6
151	5
168	4
202+	3
220+	2
238	1

Calculation

$$\hat{S}(55) = \frac{11}{12} = 0{,}917$$

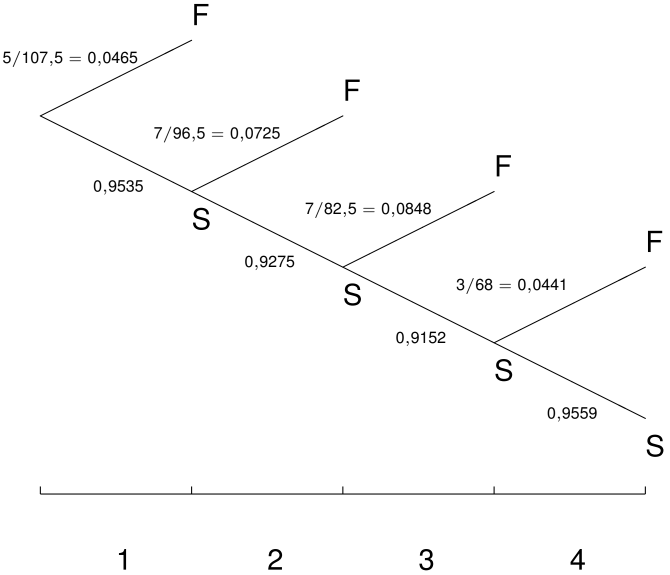
$$\hat{S}(61) = \frac{11}{12} \cdot \frac{11}{11} = 0{,}917$$

$$\hat{S}(74) = \frac{11}{12} \cdot \frac{9}{10} = 0{,}825$$

Life tables

Year	N	E	L
1	110	5	5
2	100	7	7
3	86	7	7
4	72	3	8
5	61	0	7
6	54	2	10
7	42	3	6
8	33	0	5
9	28	0	4
10	24	1	8

Life tables


F
5/107,5 = 0,0465
F
7/96,5 = 0,0725
0,9535
S
F
7/82,5 = 0,0848
0,9275
S
F
3/68 = 0,0441
0,9152
S
0,9559
S
1
2
3
4

Plotting survival curves from life tables

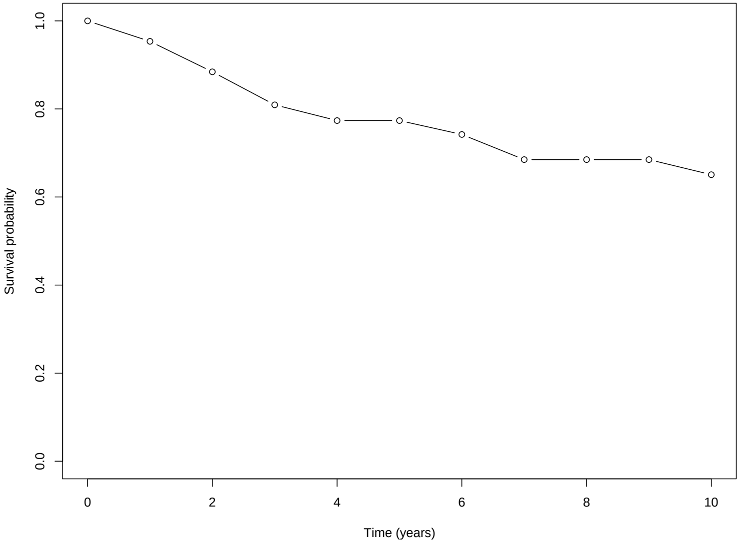


Illustration - survival after myocardial infarction

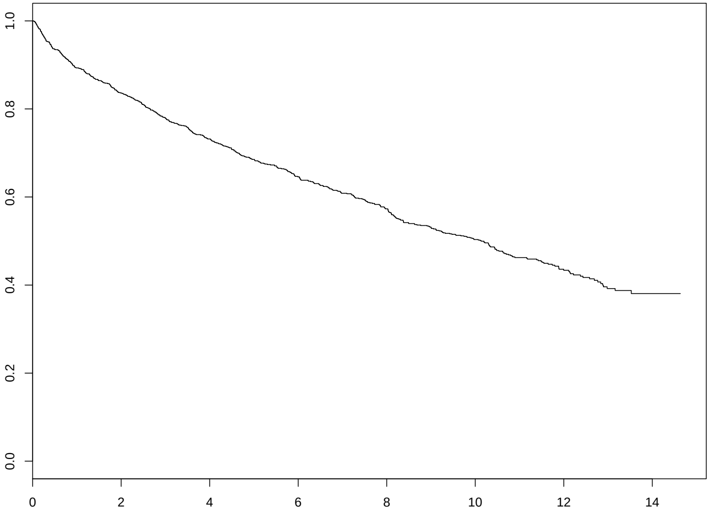


Illustration - survival after myocardial infarction

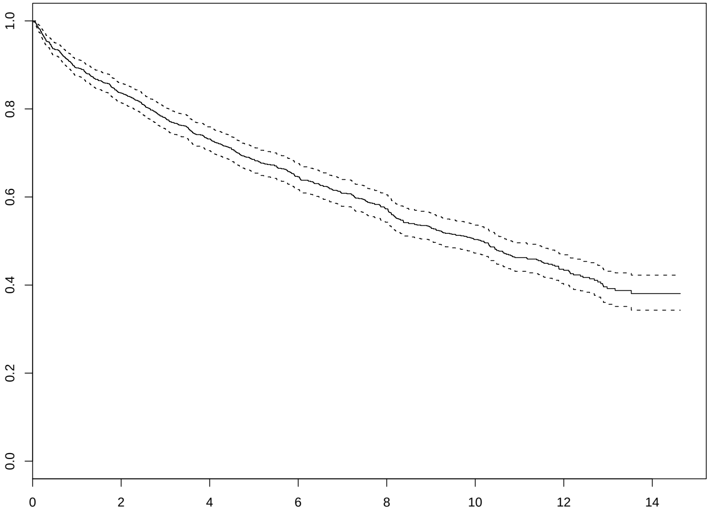
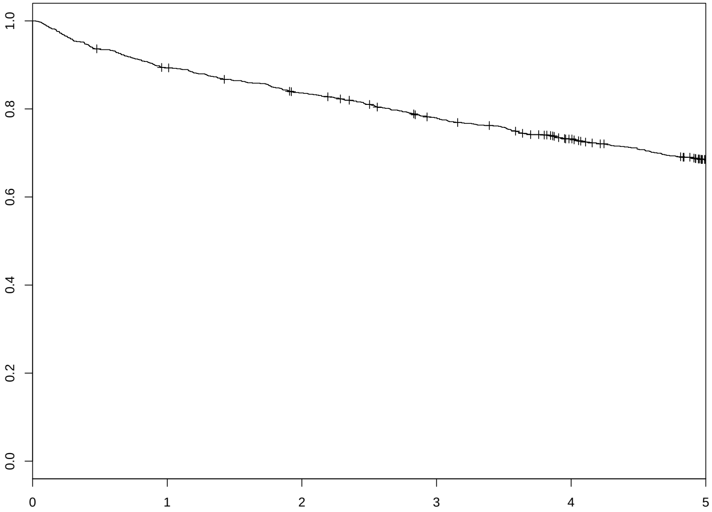
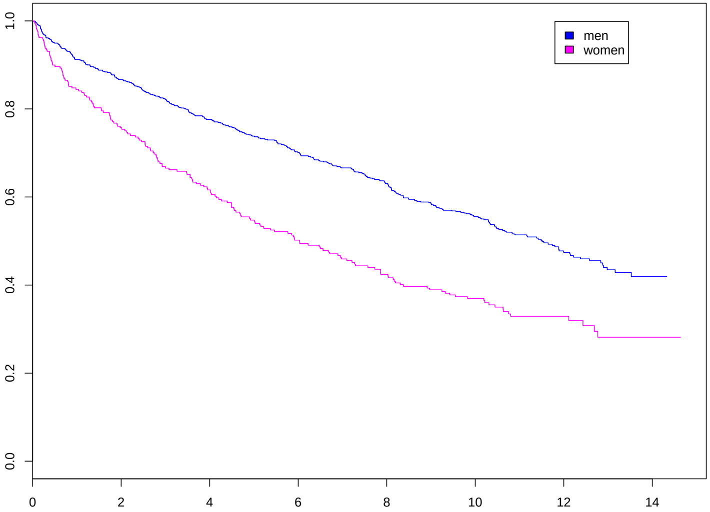


Illustration - survival after myocardial infarction



Comparison of survival curves



The statistical test for the null hypothesis (that the two samples come from the same population) is based on the usual idea:

The statistical test for the null hypothesis (that the two samples come from the same population) is based on the usual idea:

Under the null hypothesis we expect that people will be dying proportionally to the group size.

The statistical test for the null hypothesis (that the two samples come from the same population) is based on the usual idea:

Under the null hypothesis we expect that people will be dying proportionally to the group size.

Based on this we calculate the **expected number of deaths** in each group and compare it to the **observed number of deaths.**

The statistical test for the null hypothesis (that the two samples come from the same population) is based on the usual idea:

Under the null hypothesis we expect that people will be dying proportionally to the group size.

Based on this we calculate the **expected number of deaths** in each group and compare it to the **observed number of deaths.**

The name of the test is **log rank test** for some strange reasons.

The statistical test for the null hypothesis (that the two samples come from the same population) is based on the usual idea:

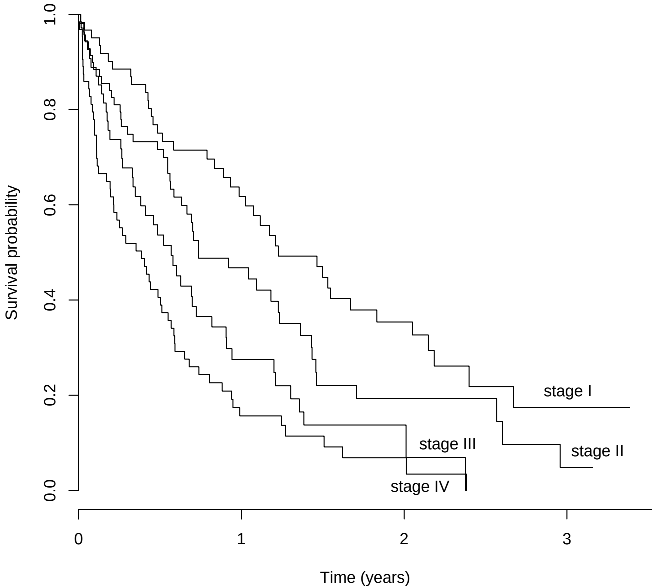
Under the null hypothesis we expect that people will be dying proportionally to the group size.

Based on this we calculate the **expected number of deaths** in each group and compare it to the **observed number of deaths.**

The name of the test is **log rank test** for some strange reasons.

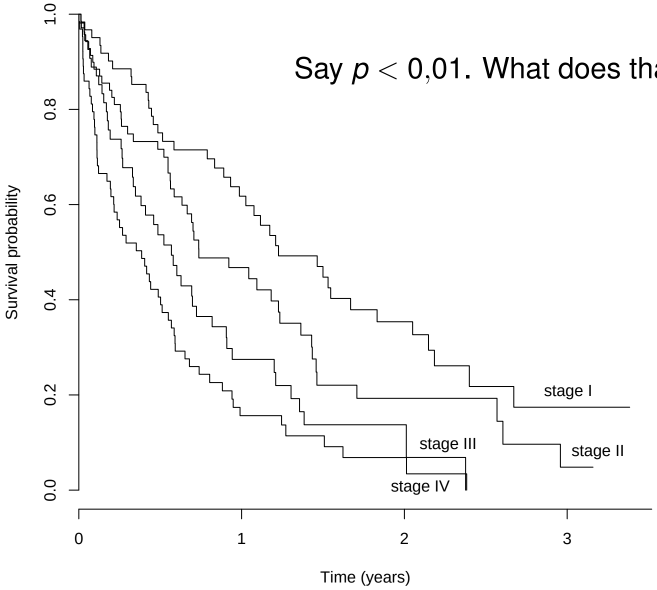
The *p*-value for the log rank test for the previous example is $3{,}1 \cdot 10^{-9}$.

Log rank test



Log rank test

Say $p < 0{,}01$. What does that mean?



How to prepare data

Start date	End date	Status (failure, no failure)

or

Survival time	Status (failure, no failure)

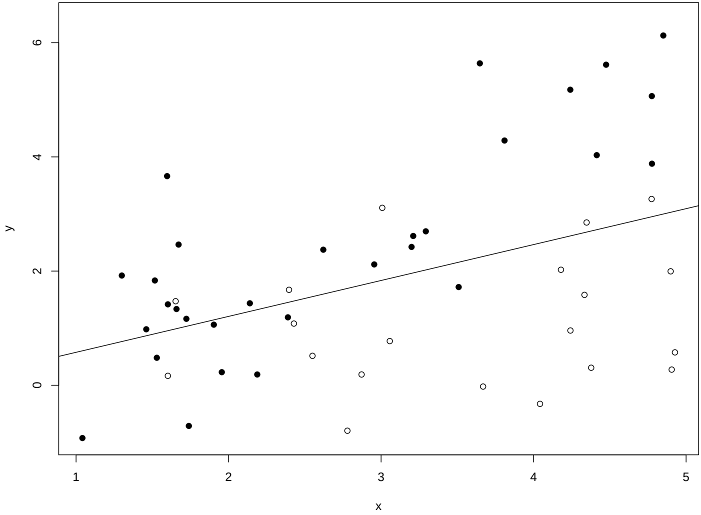
Regression models

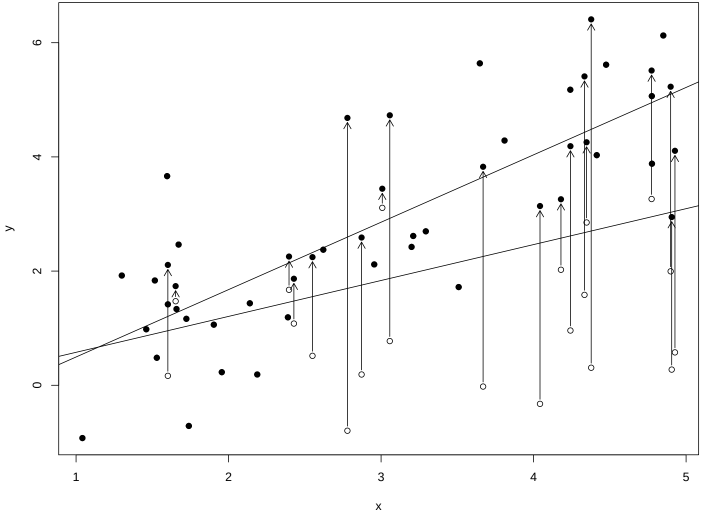
Linear regression model says

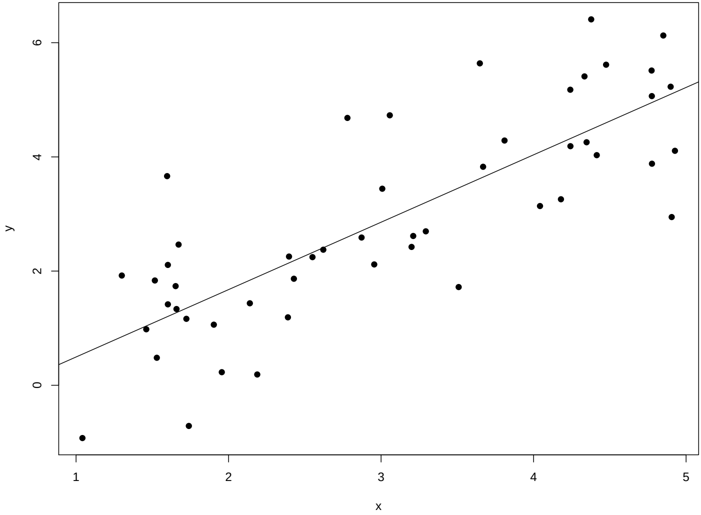
$$Y \sim \mathcal{N}(\alpha + \sum \beta_i X_i,\ \sigma^2)$$

This relates the values of Y to the values of X_i. We can not do this in survival because of censoring.

The problem can be solved by using the **hazard function.**


y
6
4
2
0
x
1
2
3
4
5


y
x
0
2
4
6
1
2
3
4
5


y
0
2
4
6
x
1
2
3
4
5

The hazard function

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

The hazard function

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

$$S(t) = e^{-\int_0^t h(u) du}$$

Parametric regression models

We assume some parametric model for the hazard.

Since in biological research it seems to be difficult to assume a certain parametric model, we rarely see parametric models being used.

We rather use the (semiparametric) **Cox model.**

The Cox model

$$h(t, x) = h_0(t)e^{\beta x}$$

The Cox model

$$h(t, x) = h_0(t)e^{\beta x}$$

$$\frac{h_1(t, x_1)}{h_2(t, x_2)} = e^{\beta(x_1 - x_2)}$$

The Cox model

$$h(t,x) = h_0(t)e^{\beta x}$$

$$\frac{h_1(t,x_1)}{h_2(t,x_2)} = e^{\beta(x_1 - x_2)}$$

$$\frac{h(t,x+1)}{h(t,x)} = e^{\beta}$$

The Cox model

$$h(t,x) = h_0(t)e^{\beta x}$$

$$\frac{h_1(t,x_1)}{h_2(t,x_2)} = e^{\beta(x_1 - x_2)}$$

$$\frac{h(t,x+1)}{h(t,x)} = e^{\beta}$$

Cox model is often called the **proportional hazards model.**

The Cox model

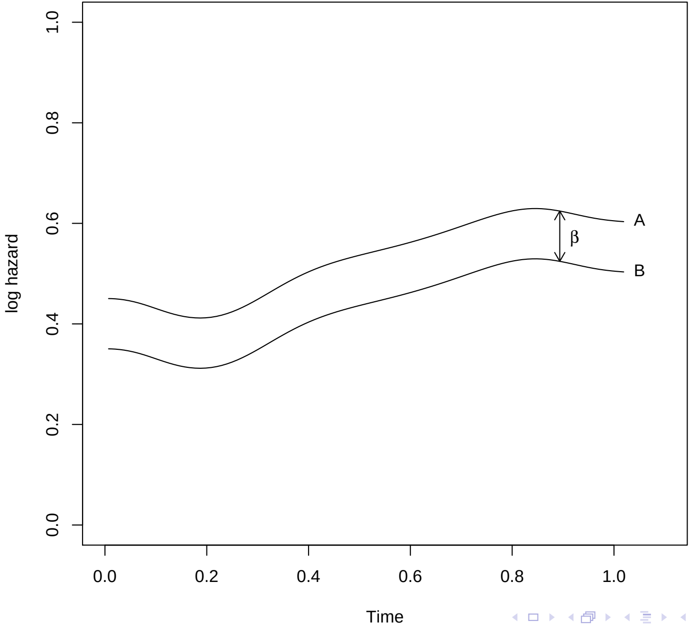
$$h(t, x) = h_0(t)e^{\beta x}$$

$$\frac{h_1(t, x_1)}{h_2(t, x_2)} = e^{\beta(x_1 - x_2)}$$

$$\frac{h(t, x+1)}{h(t, x)} = e^{\beta}$$

Cox model is often called the **proportional hazards model.**

Important: the baseline hazard stays unspecified! This is why we sometimes say that the model is semiparametric.



log hazard
1.0
0.8
0.6
0.4
0.2
0.0
A
B
β
0.0
0.2
0.4
0.6
0.8
1.0
Time

A typical description of the methods of survival analysis

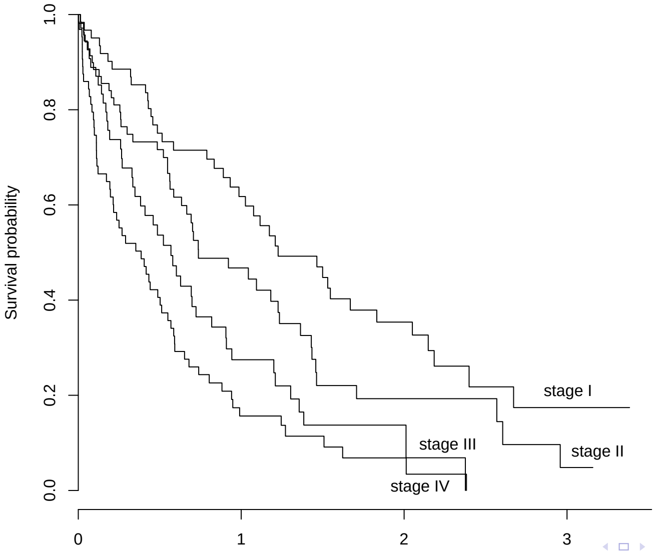
Survival curves were constructed with the Kaplan-Meier method and compared with the log-rank test. Analyses requiring adjustments for potential confounding factors were conducted using the Cox proportional hazards method. The proportional hazards assumption was tested and satisfied for each mathematical model using Cox analysis.

Example: Cox model fit to MI data

	coef	exp(coef)	se(coef)	z	p
age	0.056	1.057	0.004	12.554	0.000
sex	0.004	1.004	0.102	0.036	0.970
year	-0.081	0.922	0.035	-2.295	0.022
diabetes	0.488	1.630	0.102	4.781	0.000
aspirin	-0.335	0.716	0.094	-3.568	0.000
reinfarct	0.503	1.653	0.125	4.025	0.000

Likelihood ratio test $= 289$ on 6 df, $p = 0$ $n = 1017$ (23 observations deleted due to missing)

Example: using the Cox model to compare survival curves



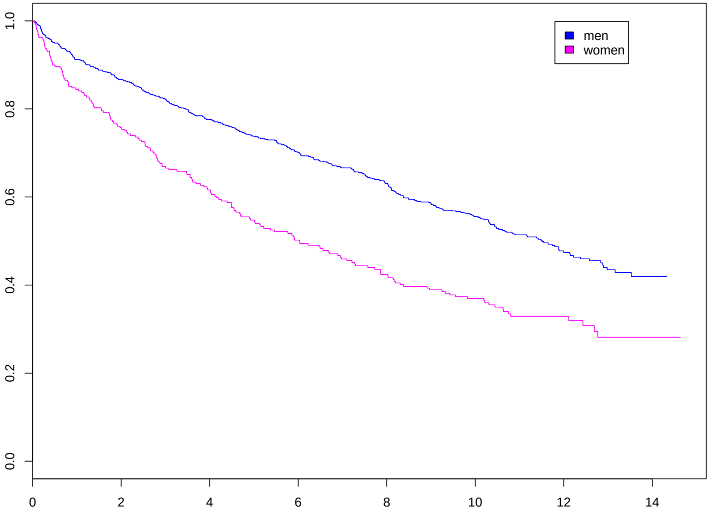
Example: using the Cox model to compare survival curves

We take stage IV to be the reference category.

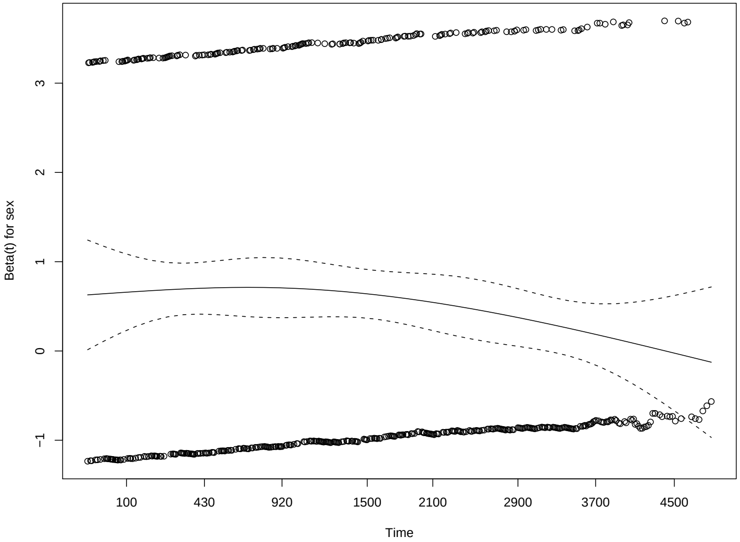
Stage	Stage I	Stage II	Stage III
I	1	0	0
II	0	1	0
III	0	0	1
IV	0	0	0

	coef	exp(coef)	se(coef)	z	p
Stage III	-0.316	0.729	0.202	-1.57	0.120
Stage II	-0.779	0.459	0.199	-3.92	< 0.001
Stage I	-1.203	0.300	0.213	-5.64	< 0.001

Example: checking the fit



Example: checking the fit



Time dependent covariates and effects

In its most general form the Cox model can be written as

$$h(t, x(t)) = h_0(t)e^{\beta(t)x(t)}$$

The model easily incorporates time dependent covariates, time dependent effects are more difficult (as they would be in any model). We'll have a look at an easy method to estimate such effects.

Frailties

Basically, we talk about frailties when the true model is, say

$$h(t, x) = h_0(t)e^{\beta_1 x_1 + \beta_2 x_2}$$

but we only measure X_1 and omit X_2 in the model. Even if X_1 and X_2 are independent, the result changes, sometimes by a lot (unlike in linear regression).

We can have individual frailties, or frailties pertaining to a group, in which case we talk of **shared frailties**.

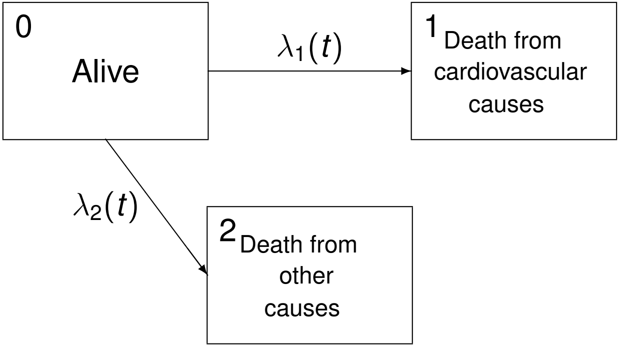
Repeated events

There are different approaches to such data:

1. assuming independence (in the past common, not recommended)
2. shared frailties
3. stratified model

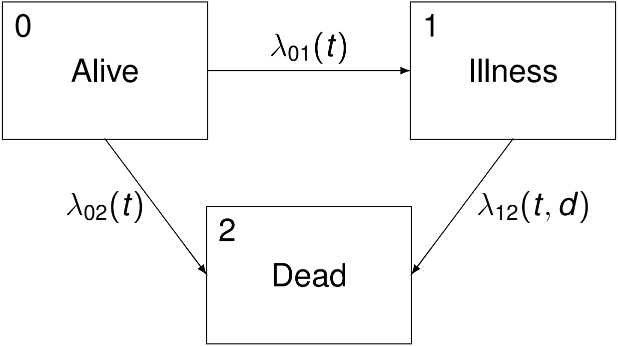
Competing risks

An individual can experience different kinds of events, but only one of them (experiencing one prevents him/her to experience another).



Multi state models

Here we also have states that are transitional, i.e. states from which an individual can exit.



Our published research in Survival Analysis

1. Explained variation in survival analysis
2. Linear model of Buckley and James
3. Frailties
4. Goodness of fit of survival models
5. Relative survival
6. Multi state models