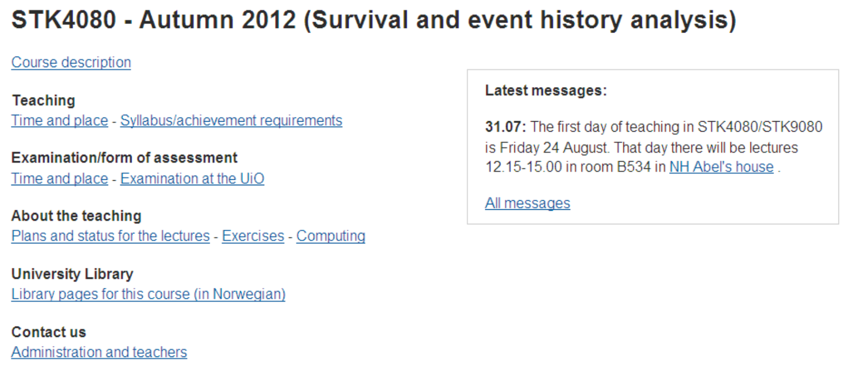
STK4080/9080: Survival and event history analysis



STK4080 - Autumn 2012 (Survival and event history analysis)

Course description

Teaching
Time and place - Syllabus/achievement requirements

Examination/form of assessment
Time and place - Examination at the UiO

About the teaching
Plans and status for the lectures - Exercises - Computing

University Library
Library pages for this course (in Norwegian)

Contact us
Administration and teachers

Latest messages:

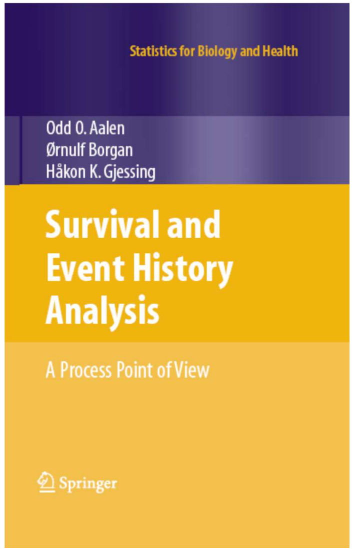
31.07: The first day of teaching in STK4080/STK9080 is Friday 24 August. That day there will be lectures 12.15-15.00 in room B534 in NH Abel's house .

All messages

The webpage for STK4080 contain all information about the course and will be updated regularly

The webpage for STK9080 will not be updated

Text book



Curriculum:

Selected parts of chapters 1-8 (details will be given along the way)

The book’s webpage http://folk.uio.no/borgan/abg-2008/ contains a list of corrections (and more)

What is survival and event history analysis?

Survival and event history analysis is a set of statistical concepts, models and methods for studying the occurrences of events over time for a number of subjects

The subjects under study my be humans, animals, engines, etc.

The events of interest may be deaths, cancer diagnoses, divorces, child births, engine failures, etc.

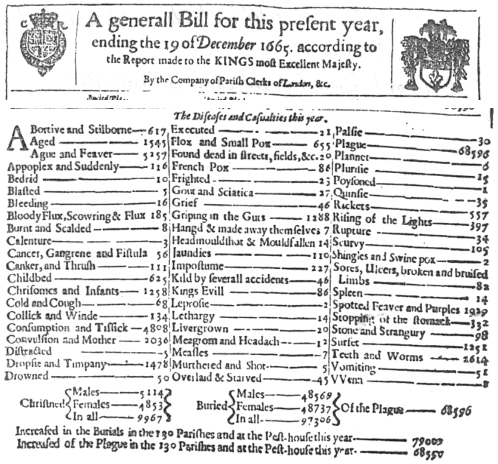
The aim of a study may be to study the effect of a medical treatment, to establish risk factors for a disease, to monitor a demographic or social phenomenon, to make predictions, etc

The scientific and professional fields using event history methodology are clinical medicine, epidemiology, demography, actuarial science, econometrics, technical reliability, sociology, etc

Traditionally most research in event history analysis has focused on situations where the interest is in a single event for each subject under study. This is called survival analysis

A very brief history

Bills of mortality and Graunt's life table



A generall Bill for this preſent year,
ending the 19 of *December* 1665. according to
the Report made to the KINGS moſt Excellent Majeſty.

By the Company of Pariſh Clerks of *London*, &c.

The Diſeaſes and Caſualties this year.

Abortive and Stilborne	617	Executed	21	Palſie	30
Aged	1545	Flox and Small Pox	655	Plague	68596
Ague and Feaver	5257	Found dead in ſtreets, fields, &c.	20	Plannet	6
Appoplex and Suddenly	116	French Pox	86	Pluriſie	15
Bedrid	10	Frighted	23	Poyſoned	1
Blaſted	5	Gout and Sciatica	27	Quinſie	35
Bleeding	16	Grief	46	Rickets	557
Bloody Flux, Scowring & Flux	185	Griping in the Guts	1288	Riſing of the Lights	397
Burnt and Scalded	8	Hangd & made away themſelves	7	Rupture	34
Calenture	3	Headmouldſhot & Mouldfallen	14	Scurvy	105
Cancer, Gangrene and Fiſtula	56	Jaundies	110	Shingles and Swine pox	2
Canker, and Thruſh	111	Impoſtume	227	Sores, Ulcers, broken and bruiſed	
Childbed	625	Kild by ſeverall accidents	46	Limbs	82
Chriſomes and Infants	1258	Kings Evill	86	Spleen	14
Cold and Cough	68	Leproſie	2	Spotted Feaver and Purples	1929
Collick and Winde	134	Lethargy	14	Stopping of the ſtomack	332
Conſumption and Tiſſick	4808	Livergrown	20	Stone and Strangury	98
Convulſion and Mother	2036	Meagrom and Headach	12	Surfet	1251
Diſtracted	5	Meaſles	7	Teeth and Worms	2614
Dropſie and Timpany	1478	Murthered and Shot	5	Vomiting	51
Drowned	50	Overlaid & Starved	45	VVenn	8

Chriſtned	Males	5114	Buried	Males	48569
	Females	4853		Females	48737
	In all	9967		In all	97306

Of the Plague 68596

Increaſed in the Burials in the 130 Pariſhes and at the Peſt-houſe this year 79009

Increaſed of the Plague in the 130 Pariſhes and at the Peſt-houſe this year 68550

John Graunt
(1620-1674)

Graunt's 1662 Life-Table	
Age Group	Probability of Survival to next age group as %
0-5	64.0%
6-15	62.5%
16-25	62.5%
26-35	64.0%
36-45	62.5%
46-55	60.0%
56-65	50.0%

Halley's life table and life annuities

Edmond Halley
(1656-1742)

Halley's 1693 Life-Table	
Age Group	Probability of Survival to next age group as %
0-5	71.0%
6-15	87.6%
16-25	90.0%
26-35	85.9%
36-45	80.5%
46-55	72.9%
56-65	64.5%
66-75	42.9%

Halley's comet

Throughout the 18th and 19th century and the first part of the 20th century

actuarial problems

and demography

were an inspiration for methodological developments in survival analysis

Today life tables are routinely computed by central offices of statistics around the world

Modern survival analysis has been developed over the last 50 years. The main motivation has come from *medical research*, but also problems in *econometrics* and *technical reliability* have been of importance

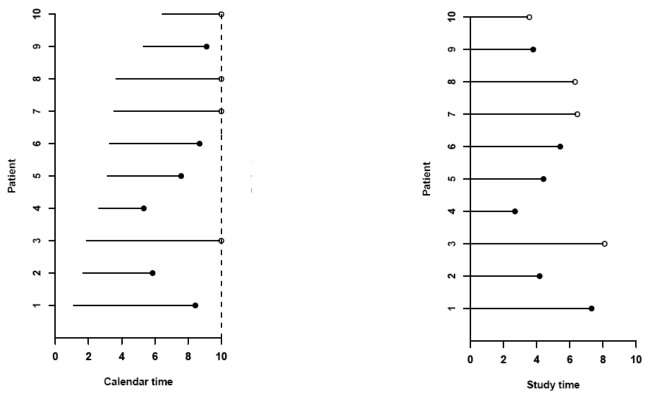
KAPLAN EL, **MEIER P (1958)**
NONPARAMETRIC-ESTIMATION FROM INCOMPLETE OBSERVATIONS
JOURNAL OF THE AMERICAN STATISTICAL ASSOCIATION
Times Cited: 36422

COX DR (1972)
REGRESSION MODELS AND LIFE-TABLES
JOURNAL OF THE ROYAL STATISTICAL SOCIETY SERIES B
Times Cited: 26449

Survival analysis: data

- A survival time is the time elapsed from an initial event to a well-defined end-point. E.g.
  - From birth to death (time=age)
  - From birth to breast cancer diagnosis (time=age)
  - From disease onset to death (time=disease duration)
  - From marriage to divorce (time=duration of marriage)
- A special feature of survival data is right censoring: we may only know that a true survival time is *larger* than (e.g.) 5 years

One reason for censoring is termination of a study:



Censoring may also be due to other reasons, e.g., withdrawals, lost to follow-up, deaths from another cause than the one under study

It is crucial that censoring does *not* selectively remove subjects at particular high or low risk of experiencing an event. This is the independent censoring assumption to be discussed later.

Due to right censoring, traditional statistical method (like t-tests and linear regression) cannot be used to analyze survival data (we cannot even compute a mean)

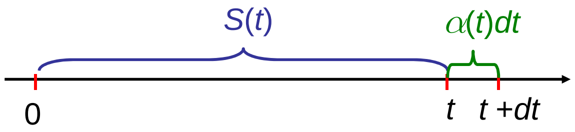
A further complication in some studies is that the subjects are not followed from time 0 (in the study time scale), but only from a later entry time. This is called delayed entry or left-truncation

Survival analysis: concepts

In order to analyse survival data, we need the right concepts

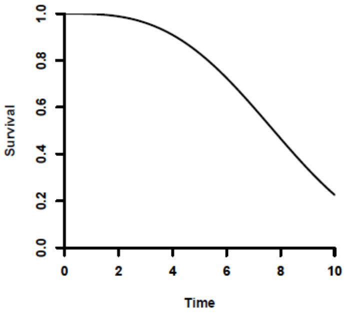
The *survival function* $S(t) = P(T > t)$ is the probability that the survival time *T* exceeds *t* (in the study time scale)

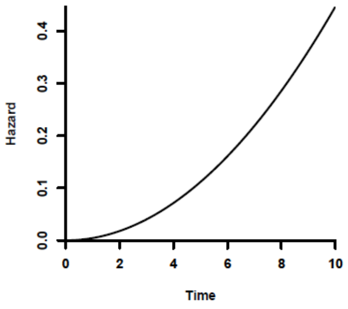
The *hazard rate* $\alpha(t)$ is the instantaneous probability of the event per unit of time, i.e. $\alpha(t)dt$ is the probability that the event will happen between time *t* and time $t + dt$ given that it has not happened earlier



The *survival function* describes the proportion of the population that has not experienced the event by time *t*

The *hazard rate* specifies the instantaneous risk of the event as function of time *t*

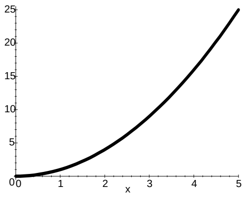


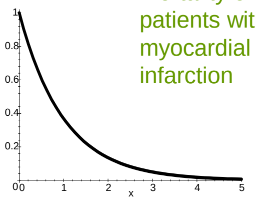


Other names for hazard rate: *intensity, incidence rate, mortality rate, etc.*

Shapes of hazard rates

General mortality. Incidence of most cancers

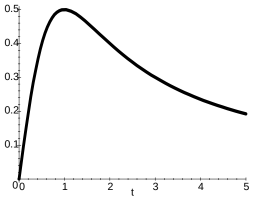




Mortality of patients with myocardial infarction

How can the decreasing hazards be interpreted?

Divorce rates. Mortality of cancer patients. Incidence of childhood leukemia



A reduced risk over time at the individual level or a selection effect?

We will discuss this in chapter 6

Formal definitions and relations:

$$S(t) = P(T > t)$$

$$\alpha(t) = \lim_{\Delta t \to 0} \frac{1}{\Delta t} P(t \leqslant T < t + \Delta t \mid T \geq t)$$

$$\alpha(t) = -\frac{S'(t)}{S(t)}$$

$$-\log\{S(t)\} = \int_0^t \alpha(s)ds$$

$$S(t) = \exp\left\{-\int_0^t \alpha(s)ds\right\}$$

Survival analysis: regression

- Usually one wants to study the effects on survival of a number of variables (covariates).
- Due to censoring the usual regression methods can not be applied
- The most common regression model for censored survival data is Cox's regression model:

$$\alpha(t|x_1,\ldots,x_p) = \alpha_0(t)\exp\{\beta_1 x_1 + \cdots + \beta_p x_p\}$$

- Another regression model is Aalen’s additive regression model:

$$\alpha(t|x_1,\ldots,x_p) = \beta_0(t) + \beta_1(t)x_1 + \cdots + \beta_p(t)x_p$$

- These will be considered in Chapter 4

Survival analysis: some examples

Example 1.1: Time between first and second births

for women whose

(i) first child dies within one year

(ii) survives at least one year

Aim: study the effect of the loss of a child on the likelihood of getting a new one

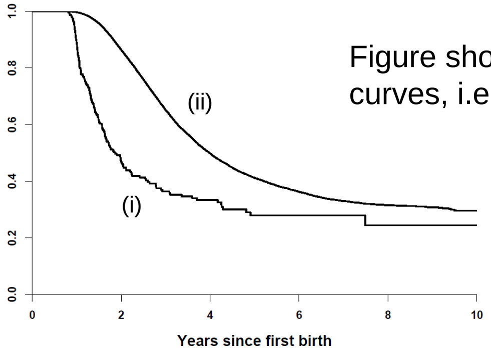
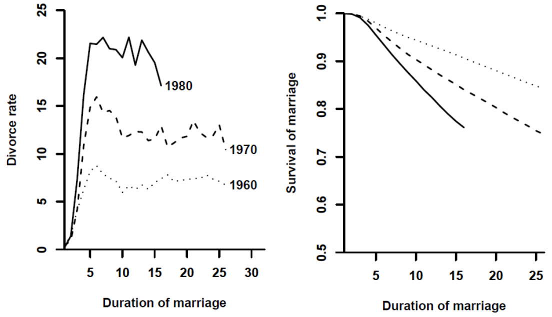


Figure show empirical survival curves, i.e. Kaplan-Meier estimates

We return to the example in Chapter 3

Example 1.2:
Divorce for couples married in 1960, 1970 and 1980

Aim: describe how the divorce rates (i.e. hazard rate for divorce) varies with the duration of the marriage and over calendar time



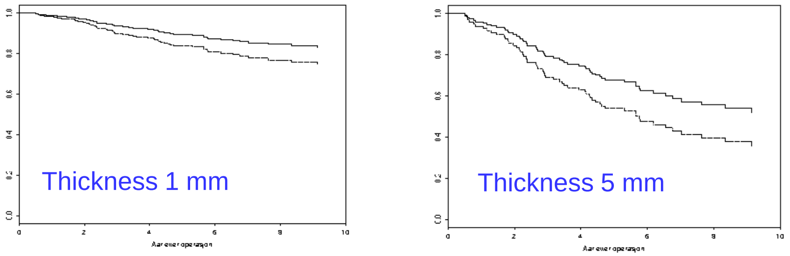
We return to the example in Chapter 5

Example: Survival with malignant melanoma

Patients operated for malignant melanoma. Many clinical variables recorded at operation (details later).

Aims: Study which clinical variables increase the risk of cancer death. Establish a model that can be used to predict survival probabilities for future patients

Illustration based on a Cox model with sex and tumor thickness as covariates: (females upper curves, males lower curves)



We return to such examples in Chapters 4

Example 1.9: Amalgam fillings

Have data on the duration of amalgam fillings in teeth for 32 patients with from 4 to 38 fillings

Aim: Study the duration of amalgam fillings and how it depends on patient properties

This is an example of clustered survival data, where the durations for one patient are dependent

We return to the example in Chapter 7

Event history analysis

Connecting together several events for a subject as they occur over time yields *event histories*

Events may be of the same type (recurrent events):

- Births for a woman
- Recurrent cancers
- Heart attacks

The events may be of different types:

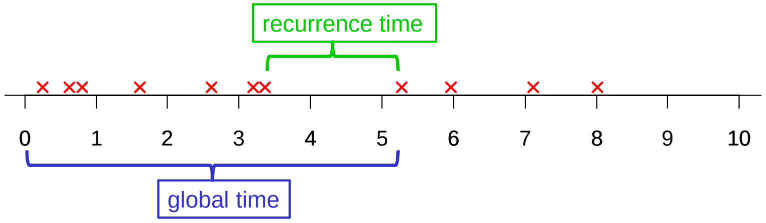
- Marriage, divorce, new marriage, etc.
- Cancer diagnosis, remission, relapse, death
- Employed, out of work, employed, out of work, on disability pension, etc

Such data may be modeled by multi-state models

Recurrent event data

For each individual in the study we observe repeated occurrences of an event (e.g epileptic seizures, heart attacks)

Data for one individual (events marked with x):



For modeling one may use global time (time since start) or recurrence time (time since last event)

The simplest model using global time is a Poisson process with intensity $\alpha(t)$

The simplest model using recurrence time is a renewal process, where the times between events are iid with hazard rate $h(u)$

For both types of models, one may obtain regression models by allowing the hazard rates to depend on covariates, e.g. as in Cox regression

Example 1.10: Bladder cancer

Study patients with superficial bladder cancer.

Tumors were removed, and the patients were randomized to placebo or active treatment

Register recurrences of new tumors

Aims: Study the effect of treatment and other covariates have on the recurrence of new tumors

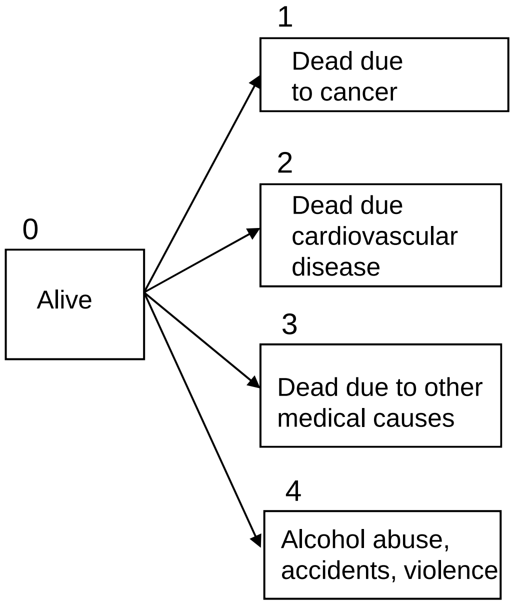
We return to the example in Chapter 7

Multistate models

Example 1.12: Competing causes of death

Data from the health screenings in three Norwegian counties 1974-78.

Followed-up to the end of 2000 by record linking to the cause of death registry at Statistics Norway.



0

Alive

1

Dead due to cancer

2

Dead due cardiovascular disease

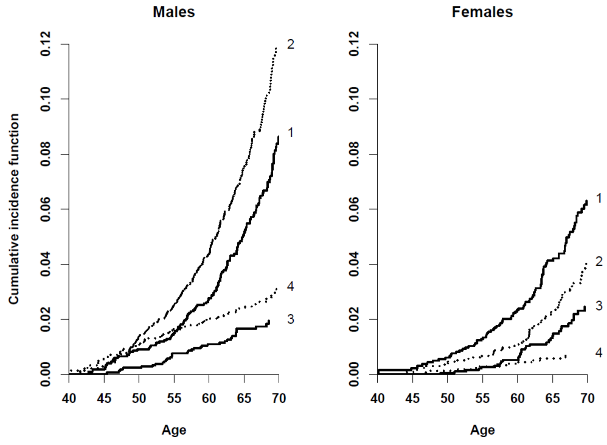
3

Dead due to other medical causes

4

Alcohol abuse, accidents, violence

The figures show estimated probabilities of death according to cause and sex:



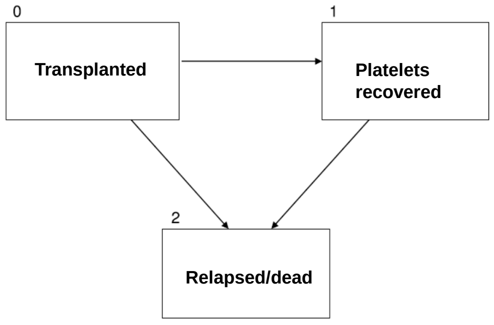
1) Cancer
2) Cardiovascular disease

3) Other medical
4) Alcohol abuse, violence, accidents

We return to the example in Chapter 3

Example 1.13: Platelet recovery, relapse and death for bone marrow transplant patients

137 patients with acute leukemia have had a bone marrow transplantation. Record the time of the events "platelet recovery" and "death/relapse"



0

Transplanted

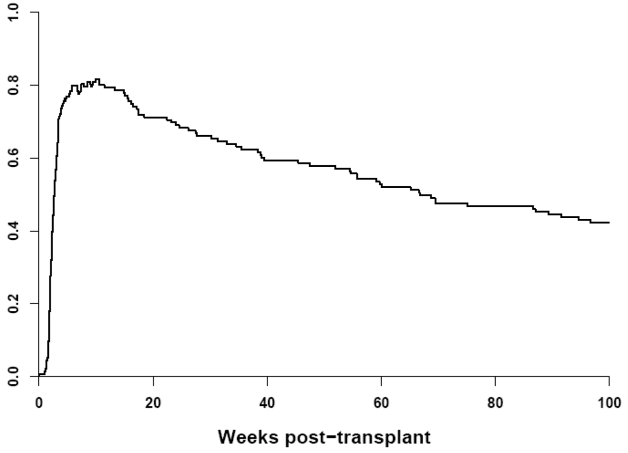
1

Platelets recovered

2

Relapsed/dead

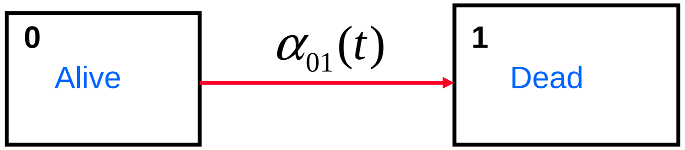
The figure shows estimated probabilities of "being in response", i.e. alive with platelets recovered



We return to the example in Chapter 3

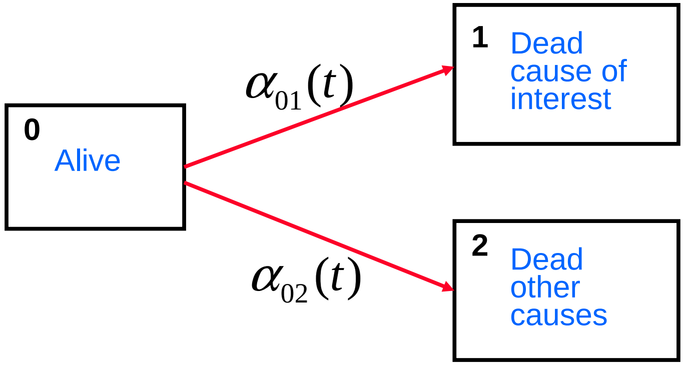
Multistate models: the Markov case

The survival analysis situation may be modelled by a Markov model with two states:



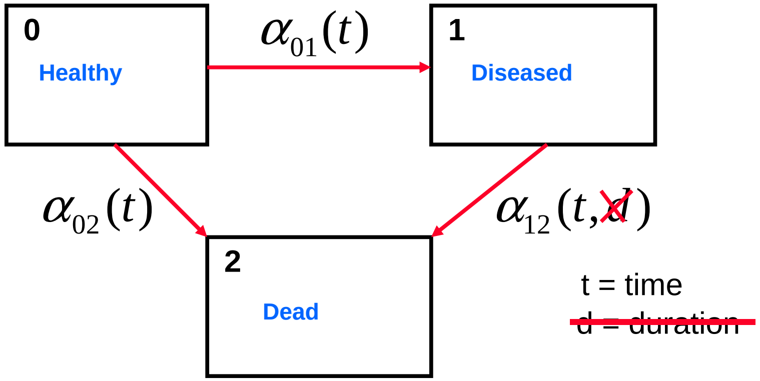
$\alpha_{01}(t)$ is the hazard rate or transition intensity.

With two or more causes of failure we get a model for competing risks:



$\alpha_{01}(t)$ and $\alpha_{02}(t)$ are the cause specific hazards or transition intensities (i.e. instantaneous probabilities of a transition per unit of time).

An illness-death model:



We have a Markov process if the transition intensities do not depend on duration in a state

In general we consider a stochastic *X*(*t*) process with state space $\mathscr{I} = \{0,1,2,\ldots,k\}$

The process is a Markov process if future transitions only depend on the current state

May define transition probabilities

$$P_{gh}(s,t) = P(X(t) = h \,|\, X(s) = g) \qquad s < t, \quad g,h \in \mathscr{S}$$

and transition intensities

$$\alpha_{gh}(t) = \lim_{\Delta t \to 0} \frac{1}{\Delta t} P(X(t+\Delta t) = h \,|\, X(t-) = g)$$

for $g \neq h$

In Chapter 3 we will see how the transition probabilities may be obtained from the transition intensities

Counting processes: an informal introduction

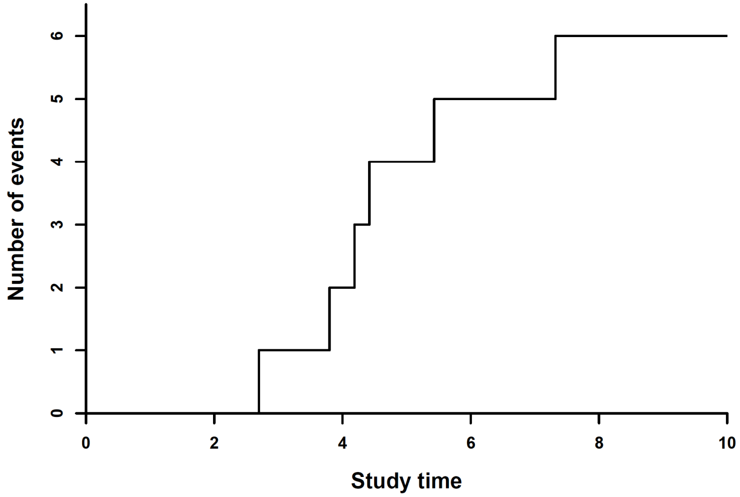
Counting processes will play a key role in formulating models for survival and event history data and in deriving estimators and test statistics

Consider the occurrence of a single type of event

Example data (with * corresponding to censoring)

$$2.70,\ 3.50^*,\ 3.80,\ 4.19,\ 4.42,$$
$$5.43,\ 6.32^*,\ 6.46^*,\ 7.32,\ 8.11^*$$

The counting process *N*(*t*) counts the number of that have occurred in the time interval [0, *t*]



Note that *N*(*t*) is continuous from the right

A well-known example of a counting process, is a (homogeneous) Poisson process with intensity λ

For a Poisson process, the events occur independently of each other and

$$P(\text{event between } t \text{ and } t+dt) = \lambda\, dt$$

For a counting process, the occurrence of future event will typically depend on “the past”

We may then (informally) define an intensity process $\lambda(t)$ by

$$\lambda(t)dt = P(dN(t) = 1 \mid \text{past})$$

where *dN*(*t*) is the number of jumps of the process in [*t*, *t* + *dt*), assumed to be 0 or 1

Since *dN*(*t*) is a binary variable we have

$$\lambda(t)dt = \mathrm{E}(dN(t) \mid \text{past})$$

which gives $\mathrm{E}(dN(t) - \lambda(t)dt \mid \text{past}) = 0$

We now define $M(t) = N(t) - \int_0^t \lambda(s)ds$

Then $\mathrm{E}(dM(t) \mid \text{past}) = 0$

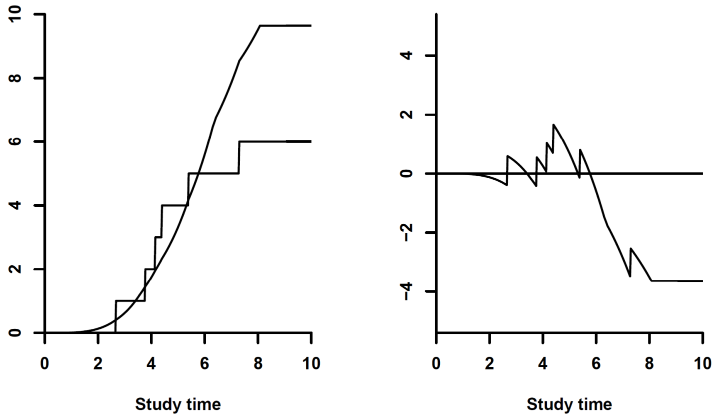
This shows (informally) that *M*(*t*) is a martingale

Note that

$$\boxed{dN(t) = \lambda(t)dt + dM(t)}$$

Martingales will be studied further in Chapter 2

Counting process *N*(*t*), cumulative intensity process $\Lambda(t) = \int_0^t \lambda(s)ds$ and martingale *M*(*t*) for the example:



Counting processes formulation of survival data

Example 1.16: One uncensored survival time

T survival time with hazard $\alpha(t)$

Counting process $N^c(t) = I\{T \le t\}$

Then

$$P(dN^c(t) = 1\,|\,\text{past}) = P(t \le T < t + dt\,|\,\text{past})$$
$$= \begin{cases} \alpha(t)dt & \text{for } T \ge t \\ 0 & \text{for } T < t \end{cases}$$

Intensity process:

$$\lambda^c(t) = \alpha(t)I\{T \ge t\}$$

Example 1.17: Uncensored survival times

$T_1, T_2, \ldots., T_n$ independent survival times

Hazard rate for T_i is $\alpha_i(t)$

Counting processes $N_i^c(t) = I\{T_i \le t\} \qquad i = 1, 2, \ldots., n$

Intensity process (due to independence):

$$\lambda_i^c(t) = \alpha_i(t)I\{T_i \ge t\}; \qquad i = 1, \ldots, n$$

Aggregated process $N^c(t) = \sum_{i=1}^n N_i^c(t)$ has intensity process

$$\lambda^c(t) = \sum_{i=1}^n \lambda_i^c(t) \; = \sum_{i=1}^n \alpha_i(t)I\{T_i \ge t\}$$

Examples of specific censoring schemes:

Type I censoring: Observe T_i if $T_i \le c_i$, otherwise just observe that $T_i > c_i$ for a fixed censoring time c_i

Type II censoring: Observe the *r* smallest survival times, for the *n* – *r* largest survival times we just know that they exceed $T_{(r)}$

Random censoring: Similar to Type I censoring, except that the c_i are observed values of random variables C_i that are independent of the survival times T_i

We will not assume any of these, but make the weakest possible assumption on the censoring that allows for valid inference. This is the independent censoring assumption

When we have censoring, we for each individual observe a (possibly) censored survival time $\tilde{T}_i$ together with an indicator D_i that takes the value 1 when $\tilde{T}_i = T_i$ and the value 0 when $\tilde{T}_i < T_i$

For survival data the independent censoring assumption takes the form (informally)

$$P(t \le \tilde{T}_i < t + dt, D_i = 1 \mid \tilde{T}_i \ge t, \text{past})$$

$$= P(t \le T_i < t + dt \mid T_i \ge t)$$

Introduce counting processes

$$N_i(t) = I\{\tilde{T}_i \le t, D_i = 1\} \qquad i = 1, 2, ..., n$$

The intensity process $\lambda_i(t)$ of $N_i(t)$ is given by

$$\lambda_i(t)dt = P(dN_i(t) = 1 \mid \text{past})$$

$$= P(t \le \tilde{T}_i < t + dt, D_i = 1 \mid \text{past})$$

$$= \begin{cases} 0 & \text{if} \quad \tilde{T}_i < t \\ \alpha_i(t)dt & \text{if} \quad \tilde{T}_i \ge t \end{cases}$$

Thus

$$\lambda_i(t) = \alpha_i(t) Y_i(t)$$

where $Y_i(t) = I\{\tilde{T}_i \ge t\}$ is an ”at risk” indicator

Aggregated process

$$N(t) = \sum_{i=1}^{n} N_i(t) = \sum_{i=1}^{n} I\{\tilde{T}_i \le t, D_i = 1\}$$

has intensity process

$$\lambda(t) = \sum_{i=1}^{n} \lambda_i(t) = \sum_{i=1}^{n} \alpha_i(t) Y_i(t)$$

In particular, when $\alpha_i(t) = \alpha(t)$ for all *i*, we have:

$$\lambda(t) = \alpha(t) Y(t)$$

where $Y(t) = \sum_{i=1}^{n} Y_i(t)$ is the number at risk