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 may 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

John Graunt (1620-1674)

Halley's life table and life annuities

Edmond Halley (1656-1742)

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.


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

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
- Divorce rates. Mortality of cancer patients. Incidence of childhood leukemia
- How can the decreasing hazards be interpreted?
- A reduced risk over time at the individual level or a selection effect?
- We will discuss this in chapter 6

---

**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)

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

We return to such examples in Chapters 4
**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 \( \times \)):

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

---

**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.
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.

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”

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

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:

$$
\begin{array}{c|c}
0 & 1 \\
\hline 
\text{Alive} & \text{Dead} \\
\end{array}
$$

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

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

$$
\begin{array}{c}
0 \quad \text{Alive} \\
\hline 
1 \quad \text{Dead cause of interest} \\
2 \quad \text{Dead other causes} \\
\end{array}
$$

$$\alpha_{01}(t)$$ 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:

$$
\begin{array}{c|c}
0 & 1 & 2 \\
\hline 
\text{Healthy} & \text{Diseased} & \text{Dead} \\
\end{array}
$$

$$\alpha_{01}(t)$$ and $$\alpha_{02}(t)$$

$$\alpha_{12}(t,d)$$

$$\alpha_{12}(t,d)$$

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 $$\mathcal{F} = \{0, 1, 2, \ldots, k\}$$.

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

May define transition probabilities

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

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 $\lambda$.

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 events will typically depend on “the past.”

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

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

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

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

which gives

$$E(dN(t) - \lambda(t)dt | \text{past}) = 0$$

We now define

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

Then

$$E(dM(t) | \text{past}) = 0$$

This shows (informally) that $M(t)$ is a martingale.

Note that

$$dN(t) = \lambda(t)dt + dM(t)$$

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:

Example 1.16: One uncensored survival time

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

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

Then

$$P(dN^c(t) = 1 | \text{past}) = P(t \leq T < t + dt | \text{past})$$

$$= \begin{cases} \alpha(t) dt & \text{for } T \geq t \\ 0 & \text{for } T < t \end{cases}$$

Intensity process:

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

Example 1.17: Uncensored survival times

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

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

Counting processes $N^c_i(t) = I\{T_i \leq t\}$ for $i = 1, 2, ..., n$

Intensity process (due to independence):

$$\lambda^c_i(t) = \alpha_i(t) I\{T_i \geq t\}; \quad i = 1, ..., n$$

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

$$\lambda^c(t) = \sum_{i=1}^n \lambda^c_i(t) = \sum_{i=1}^n \alpha_i(t) I\{T_i \geq t\}$$

Examples of specific censoring schemes:

**Type I censoring:** Observe $T_i$ if $T_i \leq c_i$, 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)

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

Introduce counting processes

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

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

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

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

Thus

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

where $Y_i(t) = I\{\tilde{T}_i \geq t\}$ is an "at risk" indicator.

Aggregated process

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

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.