MATH 60604A
Statistical modelling
§ 7a -Censoring

HEC Montréal
Department of Decision Sciences

Survival data

- In survival analysis, we're interested in the time until an event occurs, a non-negative response variable.
- Let T_i denote the survival time for subject i ($i = 1, \ldots, n$).
- The survival time is the amount of time that elapses before the event of interest occurs.
  - Usually continuous, but often measured at discrete time values.
  - Survival time is also referred to as failure time or time-to-event.
- Survival data itself is quite particular and it's critical to have a good understanding of the actual data in order to carry out an appropriate analysis.

Examples of survival data

Examples include

- time until death for a patient diagnosed with cancer,
- time until a patient is discharged from a hospital,
- time until a customer cancels their gym membership,
- time until an unemployed person finds a job,
- time until a system fails.

Right-censoring

The biggest difficulty that arises when analyzing survival data is that **we don't necessarily observe all events**.

- the subject "survives" past the end of the study period
    - T is the time until death after a patient is diagnosed with terminal cancer
        - could be that some patients are still alive at the end of the study period.
- the subject drops out of the study
    - T is the time it takes students to finish their degree
        - it could be that some students drop out of their program.
- a different event (or "competing risk") occurs which makes the event of interest impossible
    - T is the time until employees retire
        - it could be that an individual dies before being able to retire.

Censoring

With censoring, the general idea is that we don't know the exact value of T, but we still have some information regarding some interval in which it could fall, e.g., $T > t$, or $T < t$, or $T \in [t_1, t_2]$

- **right censoring**: the event happens after some time t, i.e., $T_i \geq t$.
- **left censoring**: the event of interest has already occurred before the individual enters the study. That is, all we know is that $T < t$.
- **interval censoring**: the event of interest occurs somewhere in an interval, but we don't know where exactly. That is, all we know is that $T \in [t_1, t_2]$

We will focus on the most common case, that of right censoring.

Examples of censoring

- Suppose a researcher is interested in studying the age at which children are able to write their names.
- Here, T is the time (in years) until a child can write their name.
- The researcher follows children in a kindergarten class over the course of the school year.
    - When the researcher arrives, some children are already able to spell their name, in which case their time T is **left censored**.
    - Some children learn to write their name while away during the Christmas Holidays, in which case their T is **interval censored**.
    - Some children still don't know how to spell their name by the end of the school year, in which case their T is **right censored**.

Non-informative censoring

We typically assume that censoring is non-informative.

- That is, the censoring time is independent of the survival time: it doesn't give us any information on what the survival time could be.

An example of **informative** censoring:

- Suppose a group of patients who are terminally ill are on an experimental treatment regime in which they are given some drug that could potentially have harmful side effects. However, for ethical reasons, patients who become very sick are taken out of the study. The patients who drop out will have T that is right censored. However, those patients who drop out probably have poorer health to begin with and might be more likely to die sooner.

Non-informative right-censoring schemes

We distinguish between different types of right-censoring schemes:

- type 1 censoring: observations are collected until time C; all remaining observations are right-censored.
- type 2 censoring: we continue collection until k events have been observed.
- **random censoring**: we observe $T_i = \min\{T_i^0, C_i\}$, where the time-to-event T_i^0 is independent of the censoring time C_i.

Troncation

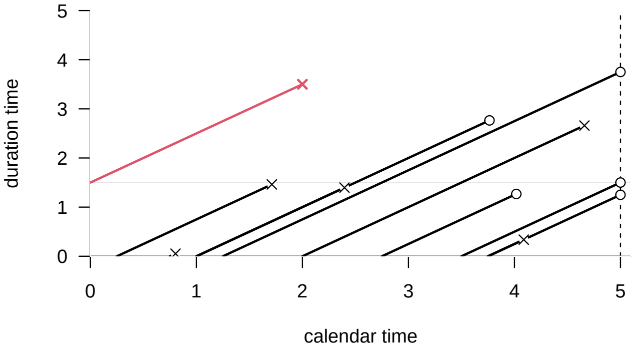
In certain studies, we collect data only during a time interval $[a, b]$.

- Survival time is **left-troncated** if survival time exceeds zero at time a

For example, during a unemployment survey, we could consider all people registered at the unemployment office between January and March.

- Some individuals lost their jobs before the year started and are already unemployed (left-truncation).
- If a person is still looking for a job at time b, the corresponding survival time will be right-censored (type I administrative censoring).

Lexis diagram



A Lexis diagram represents the temporal trajectories of lifetime, with observed failure times denoted by x and right-censored observations by ○. Residual observations at time 5 are all censored. The trajectory in red corresponds to an individual whose survival time is left-truncated.