Survival Analysis typically focuses on **time to event** data. In the most general sense, it consists of techniques for positive-valued random variables, such as

- time to death
- time to onset (or relapse) of a disease
- length of stay in a hospital
- duration of a strike
- money paid by health insurance
- viral load measurements
- time to finishing a doctoral dissertation!

Kinds of survival studies include:

- clinical trials
- prospective cohort studies
- retrospective cohort studies

Typically, survival data are not fully observed, but rather are *censored*.

In this course, we will:

- **describe survival data**
- **compare survival of several groups**
- **explain survival with covariates**

Some knowledge of discrete data methods will be useful, since analysis of the "time to event" uses information from the discrete (i.e., binary) outcome of whether the event occurred or not.

Some Definitions and notation

Failure time random variables are always **non-negative**. That is, if we denote the failure time by T, then $T \geq 0$.

T can either be **discrete** (taking a finite set of values, e.g. a_1, a_2, ..., a_n) or **continuous** (defined on $(0, \infty)$).

A random variable X is called a **censored failure time random variable** if $X = \min(T, U)$, where U is a non-negative censoring variable.

In order to define a failure time random variable, we need:

(1) an unambiguous **time origin**
(e.g. randomization to clinical trial, purchase of car)

(2) a **time scale**
(e.g. real time (days, years), mileage of a car)

(3) definition of the **event**
(e.g. death, need a new car transmission)

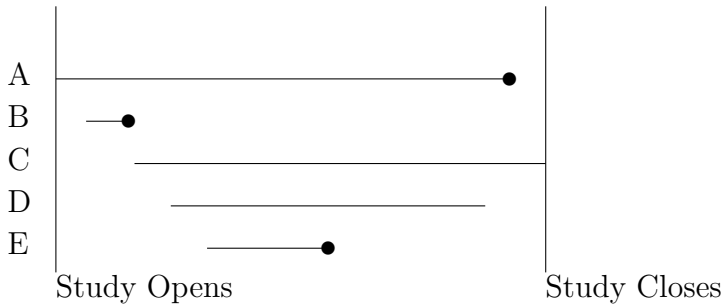


Figure 1: Some censored data. Dots denote events.

The illustration of survival data in Figure 1 shows several features which are typically encountered in analysis of survival data:

- individuals do not all enter the study at the same time (B,C,D,E)
- when the study ends, some individuals still haven't had the event yet (C)
- other individuals drop out or get lost in the middle of the study, and all we know about them is the last time they were still "free" of the event (D)

The first feature is referred to as "**staggered entry**"

The last two features relate to "**censoring**" of the failure time events

Reasons of censoring:

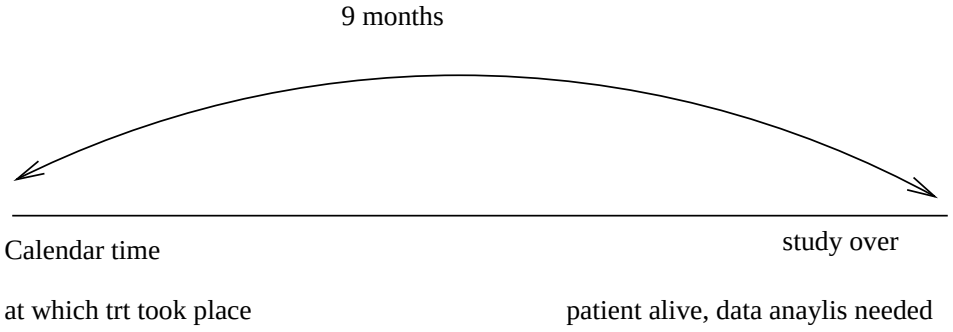
only the random variable (r.v.) $X_i = \min(T_i; U_i)$ is observed due to

1. loss to follow-up
2. drop-out
3. study termination

Types of censoring:

- Right-censoring (RC)

Fig 1: Illustration for Right Censoring



Do we observe the event time T_i? What is known about T_i? We can say that "T_i is right censored at 9 months".

We call this right-censoring because the true unobserved event is to the right of our censoring time; i.e., all we know is that the event has not happened at the end of follow-up.

In addition to observing X_i, we also get to see the **failure indicator**:

$$\delta_i = \begin{cases} 1 & \text{if } T_i \leq U_i \\ 0 & \text{if } T_i > U_i \end{cases}$$

Some software packages instead assume we have a **censoring indicator**:

$$c_i = \begin{cases} 0 & \text{if } T_i \leq U_i \\ 1 & \text{if } T_i > U_i \end{cases}$$

Right-censoring is the most common type of censoring assumption we will deal with in survival analysis.

- Left-censoring (LC): all we known is that the individual has experienced the event of interest prior to the start of the study.
  We can only observe $Y_i = \max(T_i, U_i)$ and the failure indicators:

$$\delta_i = \begin{cases} 1 & \text{if } U_i \leq T_i \\ 0 & \text{if } U_i > T_i \end{cases}$$

  Example (Kleinbaum & Klein): If we are following persons until they become HIV positive, we may record a failure when a subject firsts tests positive for the virus. However, we may not know exactly the time of first exposure to the virus, and therefore do not know exactly when the failure occurred. Thus, the survival time is censored on the left side since the true survival time, which ends at exposure, is shorter than the follow-up time, which ends when the subject tests positive.

- Double censoring (DC): both left censoring and right censoring present in the data

Example 1: Study conducted by Stanford per consulting program in 1978, where 191 California teenage boys were asked, "When did you first use marijuana?"

- There are several possible responses:
  (1) "I never used it"; Is it a complete or right censored or left censored observation?
  (2) "I have used it but cannot recall just when the first time was"; Which observation is this?
  (3) The answers were exact ages; complete data.
- This is an example of "double censored" data with a mixture of RC, LC and complete data.

Example 2: study of age at which African children learn a task. Some already knew (left-censored), some learned during study (exact), some had not yet learned by end of study (right-censored).

- Interval-censoring (IC)
  Observe (L_i, R_i) where $T_i \in (L_i, R_i)$
  Example 1: Time to prostate cancer, observe longitudinal PSA measurements
  Example 2: Time to undetectable viral load in AIDS studies, based on measurements of viral load taken at each clinic visit
  Example 3: Detect recurrence of colon cancer after surgery. Follow patients every 3 months after resection of primary tumor.

Summary: Censoring: T_i may not be observed directly, and it is known to be $>$ or $<$ some value or in an interval.

Independent vs informative censoring

- We say censoring is **independent** (non-informative) if U_i is independent of T_i.

    1. Ex. 1 If U_i is the planned end of the study (say, 2 years after the study opens), then it is usually independent of the event times.
    2. Ex. 2 If U_i is the time that a patient drops out of the study because he/she got much sicker and/or had to discontinue taking the study treatment, then U_i and T_i are probably not independent.

    An individual censored at U should be representative of all subjects who survive to U.

    This means that censoring at U could depend on prognostic characteristics measured at baseline, but that among all those with the same baseline characteristics, the probability of censoring prior to or at time U should be the same.

- Censoring is considered informative if the distribution of U_i contains any information about the parameters characterizing the distribution of T_i.

Different types of Right Censoring

Suppose we have a sample of observations on n people:

$$(T_1, U_1), (T_2, U_2), ..., (T_n, U_n)$$

There are three main types of (right) censoring times:

1. **Type I:** All the U_i's are the same

   e.g. animal studies, all animals sacrificed after 2 years

2. **Type II:** $U_i = T_{(r)}$, the time of the rth failure.

   e.g. animal studies, stop when 4/6 have tumors

3. **Type III:** the U_i's are random variables, δ_i's are failure indicators:

$$\delta_i = \begin{cases} 1 & \text{if } T_i \leq U_i \\ 0 & \text{if } T_i > U_i \end{cases}$$

   Type I and **Type II** are called *singly* censored data, **Type III** is called *randomly* censored (or sometimes *progressively* censored).

Datasets used for this class – demonstration and homework

Dowload the datasets from

http://www.mcw.edu/biostatistics/FacultyStaff/JohnPKleinPhD/SurvivalAnalysisBook.htm#.UDPGzmzq1hF

Or google John P. Klein and follow the link on his website

1. Remission Duration from a Clinical Trial for Acute Leukemia

   Freireich et al. (1963) report the results of a clinical trial of a drug 6-mercaptopurine (6-MP) versus a placebo in 42 children with acute leukemia. The trial was conducted at 11 American hospitals. Patients were selected who had a complete or partial remission of their leukemia induced by treatment with the drug prednisone. (A complete or partial remission means that either most or all signs of disease had disappeared from the bone marrow.) The trial was conducted by matching pairs of patients at a given hospital by remission status (complete or partial) and randomizing within the pair to either a 6-MP or placebo maintenance therapy. Patients were followed until their leukemia returned (relapse) or until the end of the study (in months). The data is reported in Table 1.

   This dataset will be used to demonstrate:

   (a) calculation of the estimated probability of survival using the product-limit estimator;

   (b) the calculation of the Nelson-Aalen estimator of the cumulative hazard function;

   (c) calculation of the mean survival time, along with their standard errors;

   (d) estimate the survival function using Bayesian approaches (optional);

   (e) matched pairs tests for differences in treatment efficacy are performed using the stratified log rank test;

Table 1: Remission duration of 6-MP versus placebo in children with acute leukemia

Pair	Remission Status at Randomization	Time to Relapse for Placebo Patients	Time to Relapse for 6-MP Patients
1	Partial Remission	1	10
2	Complete Remission	22	7
3	Complete Remission	3	32^+
4	Complete Remission	12	23
.	.	.	.
.	.	.	.
.	.	.	.
19	Complete Remission	4	9^+
20	Complete Remission	1	6^+
21	Complete Remission	8	10^+

$^+$ Censored observation

(f) stratified proportional hazards model.