Lecture 1
INTRODUCTION TO SURVIVAL ANALYSIS

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
- time to failure of a machine part
- length of stay in a hospital
- duration of unemployment
- accumulated medical costs in chronic disease
- HIV viral load measurements

One main feature of survival analysis is that data are not fully observed, but rather some are **censored**.

Types of survival studies include:

- clinical trials
- observational studies
- labor/economics
- engineering (reliability analysis)

In this course, we will:

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

Survival analysis relates to some of the binary data methods, since analysis of the "time to event" uses information from the binary outcome of whether the event occurred or not.

Some useful reference books:

- Cox and Oakes: *Analysis of Survival Data*, Chapman & Hall,1984
- Fleming and Harrington, *Counting Processes and Survival Analysis*, Wiley, 1991
- O'Quigley, *Proportional Hazards Regression*, Springer, 2008
- Allison: *Survival Analysis Using the SAS System*

Some important concepts

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

T can either be discrete or continuous (defined on $(0, \infty)$). In this class we mainly focus on the latter.

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

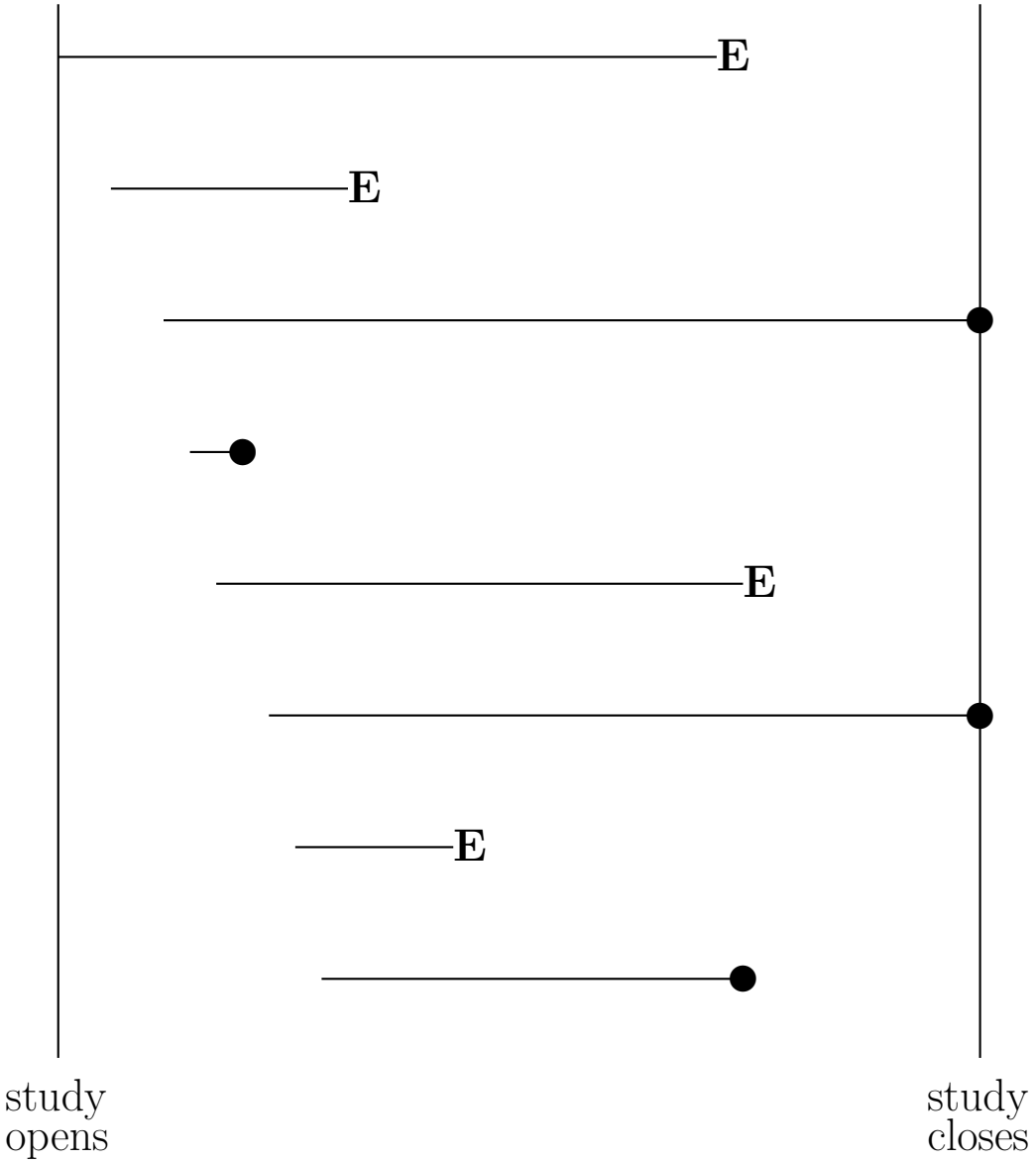
(1) a **time origin**
(e.g. randomization to clinical trial, purchase of car)

(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)

Sometimes instead of observing the actual failure time T, we observe $X < T$. In this case there is a **censoring mechanism** (eg. end of trial, patient drop out), and a corresponding **censoring random variable** C. In general we only observe $X = \min(T, C)$, and X is called a **censored failure time** random variable.

Illustration of survival data



● = censored observation

E = event

The illustration of survival data on the previous page shows several features which are typically encountered in analysis of survival data:

- individuals do not all enter the study at the same time;
- when the study ends, some individuals still haven't had the event yet;
- 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.

The first feature is referred to as **"staggered entry"**. In data analysis, staggered entry is usually not a concern because we only use the length of the observation time.

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

Types of censoring:

- **Right-censoring**:

  only the r.v. $X_i = \min(T_i, C_i)$ is observed due to

  – loss to follow-up

  – study termination

  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 have the **event indicator**:

$$\delta_i = \begin{cases} 1 \text{ if } T_i \leq C_i \text{ (or } X_i = T_i) \\ 0 \text{ if } T_i > C_i \text{ (or } X_i = C_i) \end{cases}$$

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

  Ex: think about how you would simulate right-censored data.

- **Left-censoring**

  Can only observe $Y_i = \max(T_i, U_i)$ and the failure indicators:

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

  e.g. (Miller) 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).

  Note: left-censoring can often be handled as right-censoring after a transformation like $X = c - Y$ where c is a large enough constant. We will not discuss left-censoring in the following.

- **Interval-censoring**

  Observe (L_i, R_i) where $T_i \in (L_i, R_i)$

  Eg.1: Time to onset of dementia

  Eg.2: Time to undetectable viral load in AIDS studies, based on measurements of viral load taken at each clinic visit.

  Note: This can be a useful topic to discuss, but we may not have time in this course.

Some example datasets:

Example A. Duration of nursing home stay

(Morris et al., *Case Studies in Biometry*, Ch. 12)

The National Center for Health Services Research studied 36 for-profit nursing homes to assess the effects of different financial incentives on length of stay. "Treated" nursing homes received higher per diems for Medicaid patients, and bonuses for improving a patient's health and sending them home.

Study included 1601 patients admitted between May 1, 1981 and April 30, 1982.

Variables include:

LOS - Length of stay of a resident (in days)
AGE - Age of a resident
RX - Nursing home assignment (1:bonuses, 0:no bonuses)
GENDER - Gender (1:male, 0:female)
MARRIED - (1: married, 0:not married)
HEALTH - health status (2:second best - 5:worst)
CENSOR - Censoring indicator (1:censored, 0:discharged)

First few lines of data:

37 86 1 0 0 2 0
61 77 1 0 0 4 0

Example B. Fecundability

Women who had recently given birth were asked to recall how long it took them to become pregnant, and whether or not they smoked during that time. The outcome of interest (summarized below) is time to pregnancy (measured in menstrual cycles).

19 subjects were not able to get pregnant after 12 months.

Cycle	Smokers	Non-smokers
1	29	198
2	16	107
3	17	55
4	4	38
5	3	18
6	9	22
7	4	7
8	5	9
9	1	5
10	1	3
11	1	6
12	3	6
12+	7	12

Example C: EST 1582 lung cancer trial

E1582 was a randomized multicenter lung cancer clinical trial conducted by the Eastern Cooperative Oncology Group (ECOG) to study the effects of standard chemotherapy (CAV) versus alternating regimens (CAV-HEM).

- Primary endpoint: overall survival (time to death)
- Accrued 579 patients from 31 institutions
- Other covariates:
  - bone metastases
  - liver metastases
  - performance status (score 10-100)
  - weight loss at study entry

Example D: UMARU Impact Study (UIS)

This dataset comes from the University of Massachusetts AIDS Research Unit (UMARU) IMPACT Study, a 5-year collaborative research project comprised of two concurrent randomized trials of residential treatment for drug abuse.

(1) **Program A:** Randomized 444 subjects to a 3- or 6-month program of health education and relapse prevention. Clients were taught to recognize "high-risk" situations that are triggers to relapse, and taught skills to cope with these situations without using drugs.

(2) **Program B:** Randomized 184 participants to a 6- or 12-month program with highly structured life-style in a communal living setting.

Variables:

ID	Subject ID (1-628)
AGE	Age in years
BECKTOTA	Beck Depression Score
HERCOC	Heroin or Cocaine Use prior to entry
IVHX	IV Drug use at Admission
NDRUGTX	Number previous drug treatments
RACE	Subject's Race (0=White, 1=Other)
TREAT	Treatment Assignment (0=short, 1=long)
SITE	Treatment Program (0=A,1=B)
LOT	Length of Treatment (days)
TIME	Time to Return to Drug Use (days)
CENSOR	Indicator of Drug Use Relapse (1=yes,0=censored)

Example E: Atlantic Halibut Survival Times

One conservation measure suggested for trawl fishing is a minimum size limit for halibut (32 inches). However, this size limit would only be effective if captured fish below the limit survived until the time of their release. An experiment was conducted to evaluate the survival rates of halibut caught by trawls or longlines, and to assess other factors which might contribute to survival (duration of trawling, maximum depth fished, size of fish, and handling time).

There is an article by Smith *et al.* 'Survival analysis for size regulation of Atlantic Halibut' comparing parametric survival models to semi-parametric survival models as applied to this data.

Obs #	Survival *Time* (min)	*Censor*ing Indicator	*Tow Dur*ation (min.)	Diff in *Depth*	*Length* of Fish (cm)	*Handling* Time (min.)	Total *log(catch)* log(weight)
100	353.0	1	30	15	39	5	5.685
109	111.0	1	100	5	44	29	8.690
113	64.0	0	100	10	53	4	5.323
116	500.0	1	100	10	44	4	5.323
....							

More on censoring

- We say censoring is **independent** (non-informative) if C_i is independent of T_i.
  - **eg.1** If C_i is the planned end of the study (say, 2 years after the study opens), then it is usually independent of the event times
  - **eg.2** If C_i is the time that a patient drops out of the study because they've gotten much sicker and/or had to discontinue taking the study treatment, then C_i and T_i are probably not independent

- When the censoring is **dependent**, Tsiatis (1975) showed that it is impossible to identify the distribution of T from the data without further assumptions.

- Sometimes when there are covariates (eg. treatment), censoring may depend on the covariates. But if conditional on the covariates (same treatment arm), C and T are independent, it is called **conditional independent censoring**.

- As long as censoring does not contain information about the parameters of interest of the distribution of T, it is called **non-informative** censoring.

- Both independent and conditional independent censoring are usually non-informative (more discussion later).

- We will first and mostly focus on non-informative censorships.

Suppose we have a sample of n individuals:

$$(T_1, C_1), (T_2, C_2), ..., (T_n, C_n)$$

There are three main types of censoring times:

- **Type I:** All the C_i's are the same.
  e.g. engineering reliability experiments, stop the experiment after a fixed amount of time;
  Or animal studies, all animals sacrificed after 2 years.

- **Type II:** All $C_i = T_{(r)}$, the time of the rth failure, for given r.
  e.g. engineering reliability experiments, stop the experiment after r machine parts have failed;
  Or animal studies, stop when 4/6 have tumors.

  Inference under the type I, II censoring is detailed in Lawless' book '*Statistical Models and Methods for Lifetime Data*'.

- **Random:** the C_i's are random variables, recall δ_i's are failure indicators:

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

Truncation

- We will focus on **left truncation**, i.e. we do not observe a subject from time 0, but rather a later time point Q.

  This way we only observe subjects with $Q < T$.

  Those with $T < Q$ are 'truncated' out.

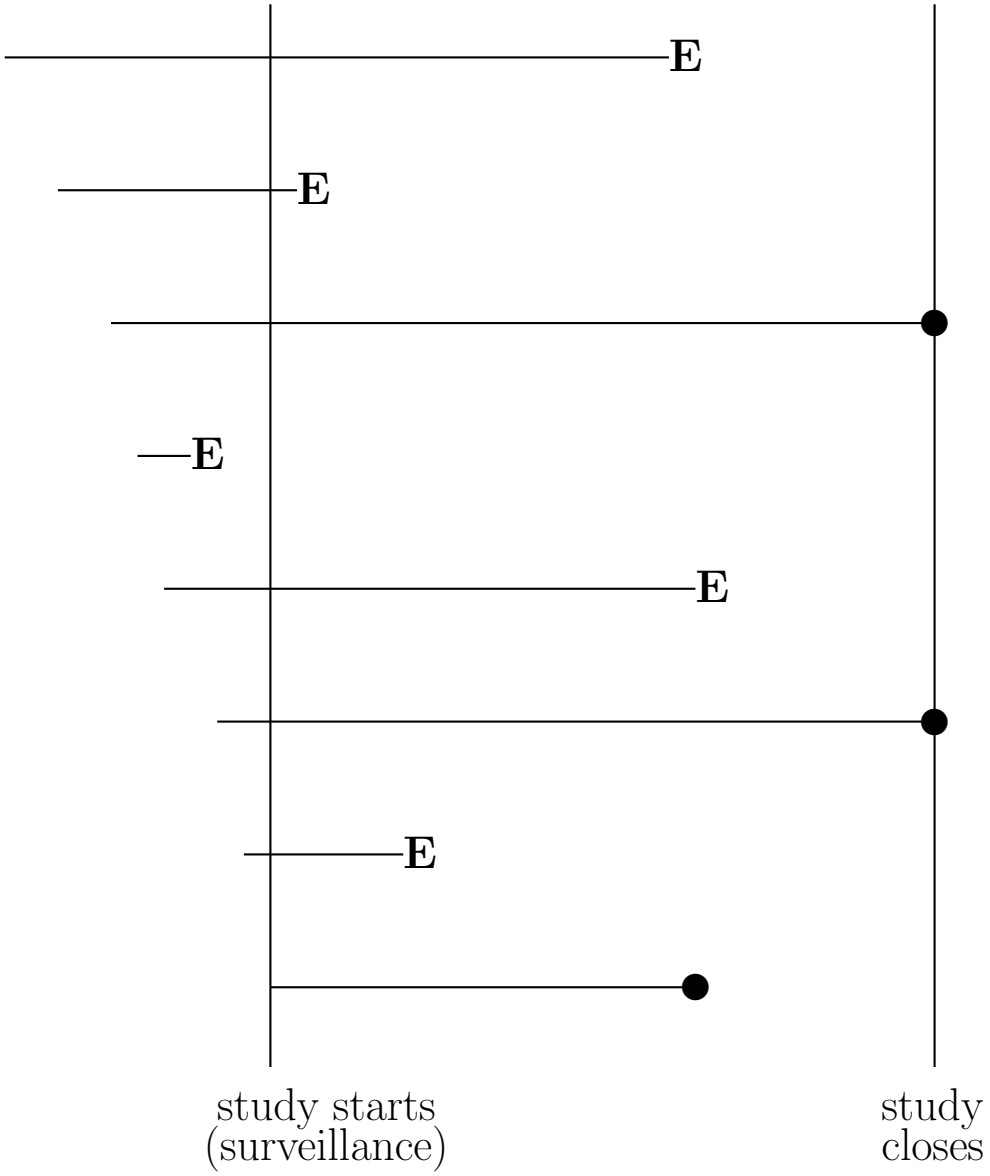
  Example: disease - surveillance - death model (next page)

  This is a biased sampling problem (Vardi, 1989), one also noted in the recent National Research Council report '*Frontiers in Massive Data Analysis*'.

  Note that this is a different problem from staggered entry mentioned earlier.

- **Right truncation**: we only observe subjects with T less than a certain time.

Illustration of left truncation



● = censored observation

E = event

Describing a survival distribution

There are several equivalent ways to characterize the probability distribution of a survival random variable T. Some of these are familiar; others are special to survival analysis. We will focus on the following terms:

- The density function $f(t)$
- The survivorship (survival) function $S(t)$
- The hazard function $\lambda(t)$
- The cumulative hazard function $\Lambda(t)$

- **Density function**

$$f(t) = \lim_{\Delta t \to 0} \frac{1}{\Delta t} Pr(t \leq T < t + \Delta t)$$

- **Survivorship (Survival) Function**:
  $S(t) = P(T \geq t)$.

  In other settings, the cumulative distribution function, $F(t) = P(T \leq t)$, is often used. In survival analysis, our interest tends to focus on the survival function, $S(t)$.

$$S(t) = \int_t^\infty f(u)du$$

Notes:

- For a continuous variable, $S(t) = 1 - F(t)$ (why).
- some books define
  $S(t) = P(T > t)$, or even define $F(t) = P(T < t)$.

- **Hazard Function** $\lambda(t)$

  Sometimes called an *instantaneous failure rate*, the *force of mortality*, or the *age-specific failure rate*.

$$\begin{aligned}
\lambda(t) &= \lim_{\Delta t \to 0} \frac{1}{\Delta t} Pr(t \leq T < t + \Delta t | T \geq t) \\
&= \lim_{\Delta t \to 0} \frac{1}{\Delta t} \frac{Pr([t \leq T < t + \Delta t] \bigcap [T \geq t])}{Pr(T \geq t)} \\
&= \lim_{\Delta t \to 0} \frac{1}{\Delta t} \frac{Pr(t \leq T < t + \Delta t)}{Pr(T \geq t)} \\
&= \frac{f(t)}{S(t)}
\end{aligned}$$

- **Cumulative Hazard Function** $\Lambda(t)$

$$\Lambda(t) = \int_0^t \lambda(u) du$$

[Read next 2 pages:] relationship between $S(t)$ and $\lambda(t)$

We've already shown that, for a continuous r.v.

$$\lambda(t) = \frac{f(t)}{S(t)}$$

We can also show:

$$f(t) \;=\; -S'(t) \qquad \text{or} \quad S'(t) \;=\; -\,f(t)$$

We can use this relationship to show that:

$$\begin{aligned} -\,\frac{d}{dt}[\log S(t)] \;&=\; -\,\left(\frac{1}{S(t)}\right) S'(t) \\ &=\; -\,\frac{-f(t)}{S(t)} \\ &=\; \frac{f(t)}{S(t)} \end{aligned}$$

So another way to write $\lambda(t)$ is as follows:

$$\lambda(t) \;=\; -\frac{d}{dt}[\log S(t)]$$

Relationship between $S(t)$ and $\Lambda(t)$:

- **Continuous case:**

$$\begin{aligned}
\Lambda(t) &= \int_0^t \lambda(u)du \\
&= \int_0^t \frac{f(u)}{S(u)}du \\
&= \int_0^t -\frac{d}{du}\log S(u)du \\
&= -\log S(t) + \log S(0)
\end{aligned}$$

$$\Rightarrow \underline{\mathbf{S(t) = e^{-\Lambda(t)}}}$$

Some hazard shapes seen in applications:

- **increasing**

  e.g. aging after 65

- **decreasing**

  e.g. survival after surgery

- **bathtub**

  e.g. age-specific mortality

- **constant**

  e.g. survival of patients with advanced chronic disease

Among f, S, λ and Λ, hazard is probably the one that describes the most of a survival distribution.

Measuring Central Tendency in Survival

- **Mean survival** - call this μ

$$\mu = \int_0^\infty u f(u) du$$

- **Median survival** - call this τ, is defined by

$$S(\tau) = 0.5$$

Similarly, any other quantile could be defined.

Because the distribution of a failure time r.v. is often not symmentric (eg. Exponential), we often use median survival. Also, median survival is usually better estimated than mean survival (especially nonparametrically, as will be discussed later).