生存分析

2020春季本科课程

严颖

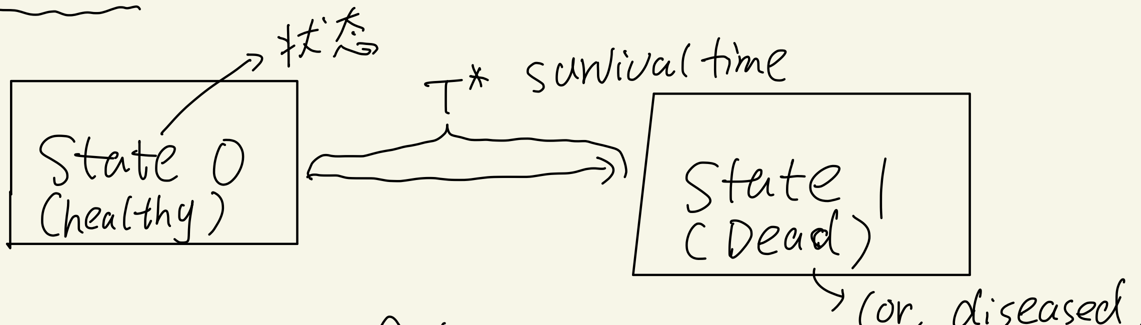
Survival Analysis

Chp 1. Introduction

§1.1 Definition

Def. Survival Analysis is a collection of statistical procedure for which the outcome of interest is time until a single event occurs.

e.g.



T^*: survival time (failure time; lifetime): from beginning of followup of an individual until an event occurs.

e.g.

医学：

$$\underset{\text{心脏移植手术}}{\overset{0}{\boxed{\text{Heart transplant surgery}}}} \xrightarrow{T^*} \overset{1}{\boxed{\text{Dead}}}$$

银行：

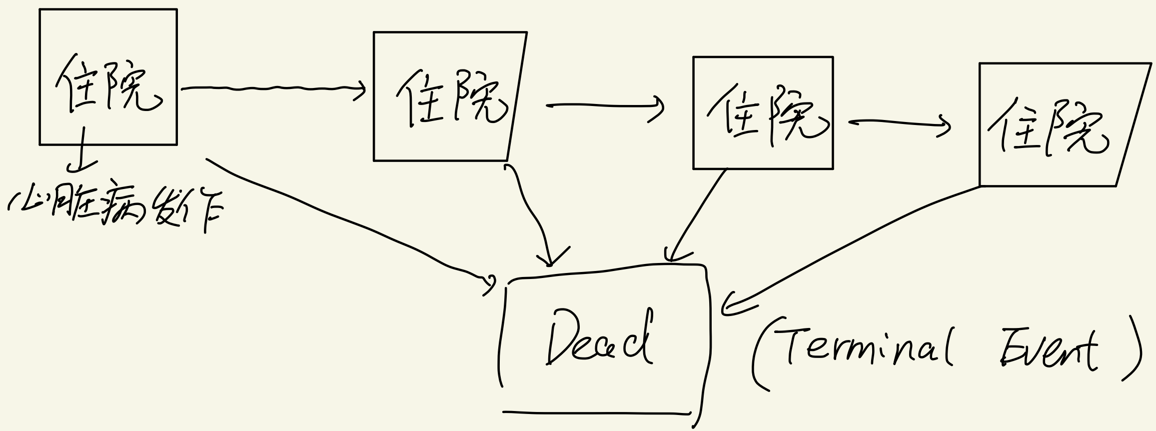
$$\overset{0}{\boxed{\text{Approval of loan}}} \xrightarrow{T^*} \underset{\text{破产}}{\overset{1}{\boxed{\text{Bankcruptcy}}}}$$

Def. Event History Analysis : We are interested in multiple events

事件历史分析

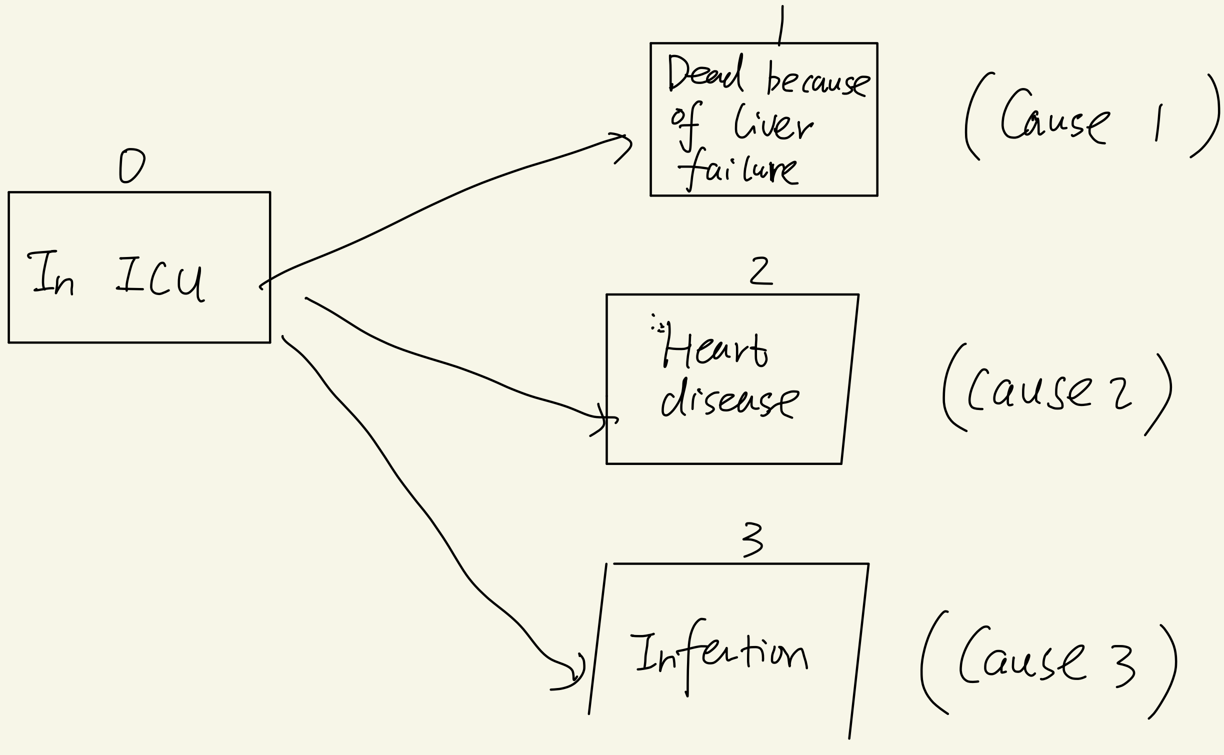
① Recurrent event analysis.(复发事件)

e.g.



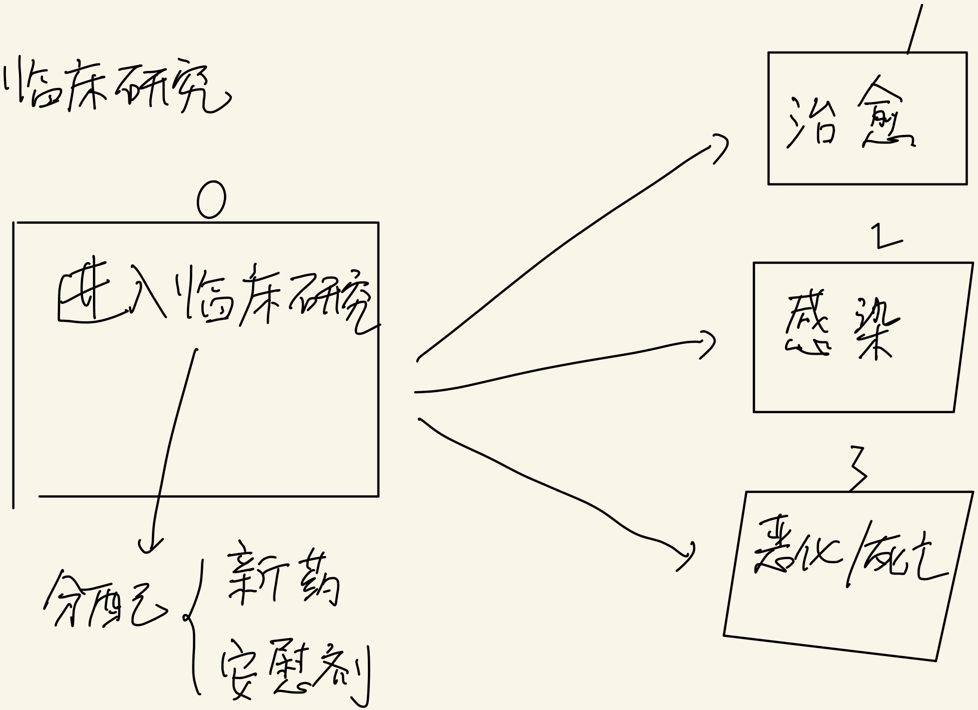
② Competing risk Analysis (竞争风险)

e.g.



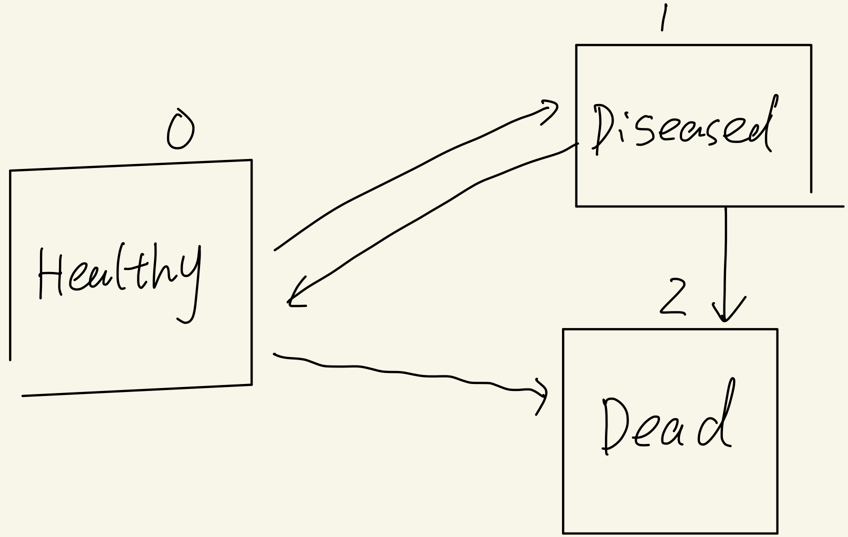
临床研究

e.g.



③ Multistate model (多状态模型)

e.g.

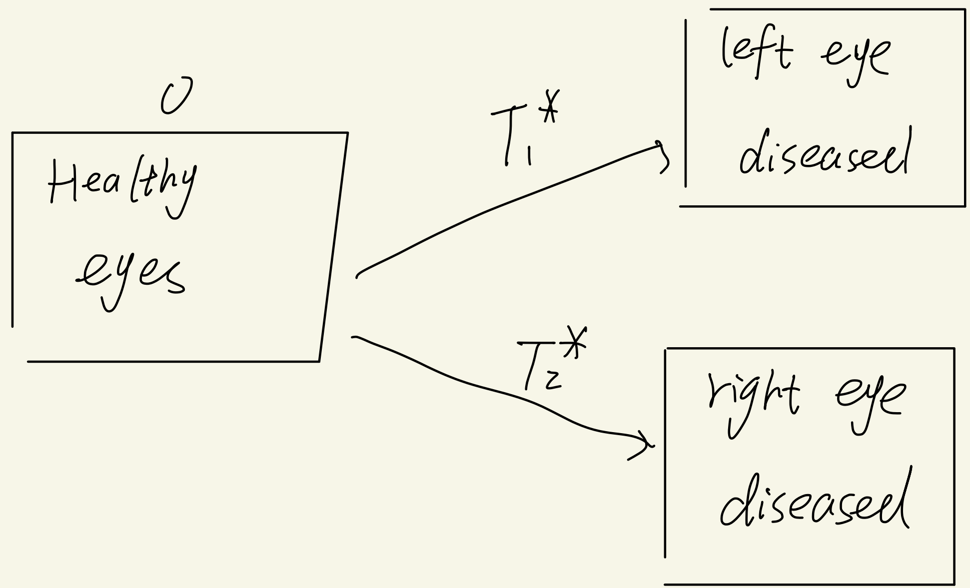


illness-death model

Note: Survival, recurrent event, competing risk are special cases of multistate models.

④ Multivariate Survival Analysis (多元生存分析)

e.g.



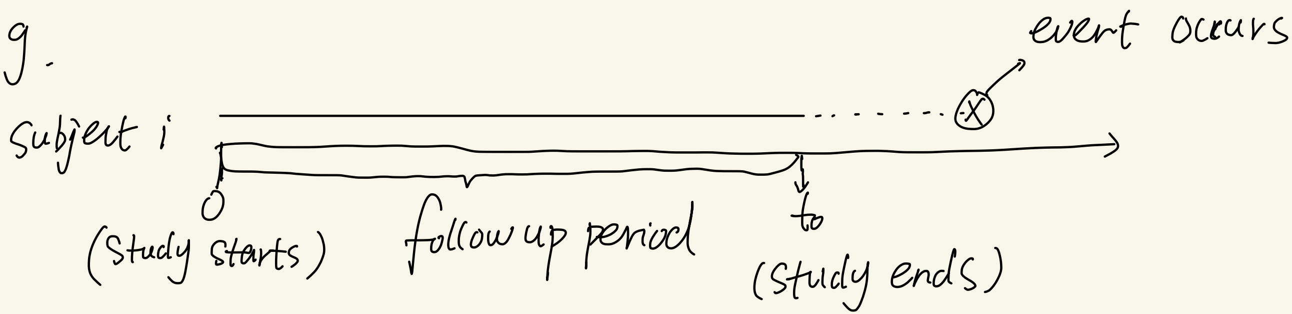
$$T^* = \begin{pmatrix} T_1^* \\ T_2^* \end{pmatrix}$$

Note: it's different from competing risk.

§1.2 Censoring (删失) and Truncation (截断)

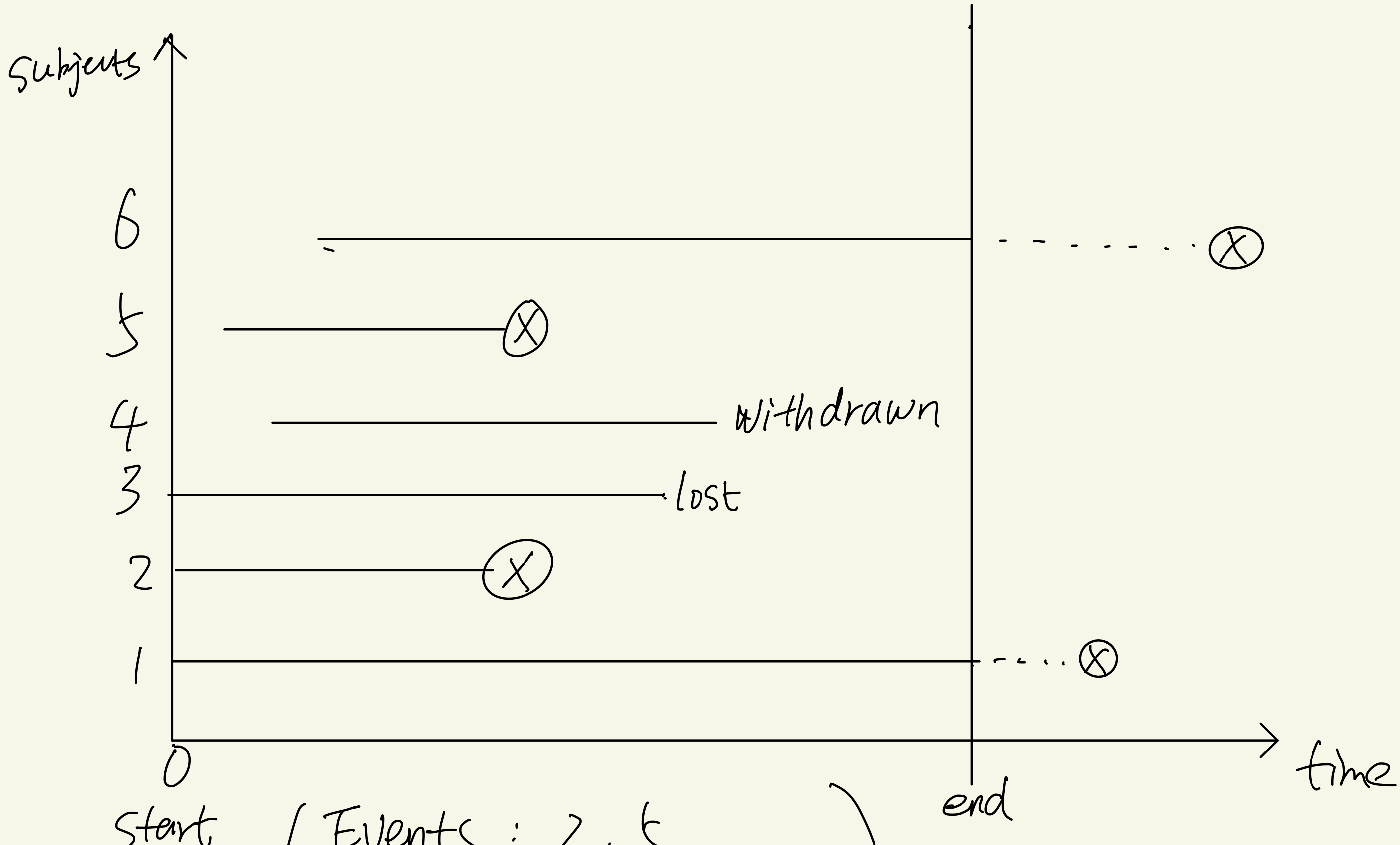
Def. <u>Censoring</u> occurs when we have some information of survival time, but we don't know the survival time <u>exactly</u>.

e.g.



we know $T_i^* > t_0$ only.

missing data



Events : 2, 5

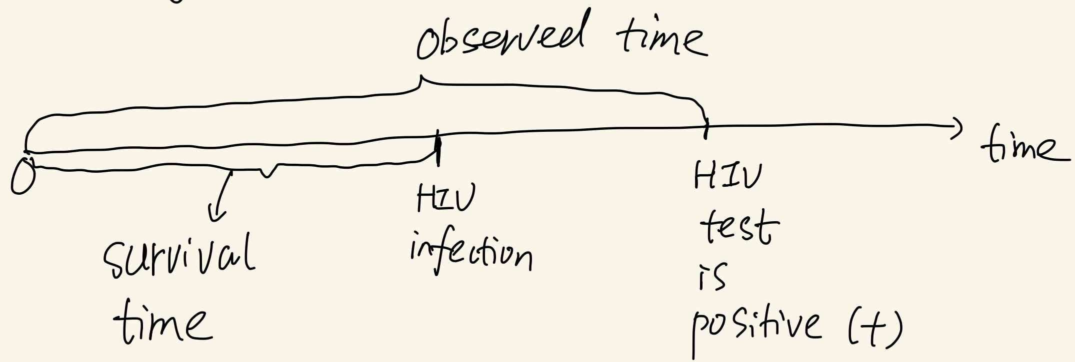
Censored : 1, 3, 4, 6

Def. Right-censored (右删失): Survival time > observed time

left-censored (左删失): <

⇓

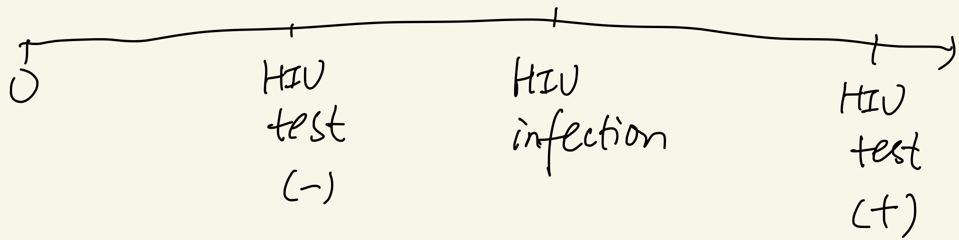
e.g.



核酸检测

Interval-censored (区间删失): survival time is unknown, but we know it's <u>within a time interval</u>.

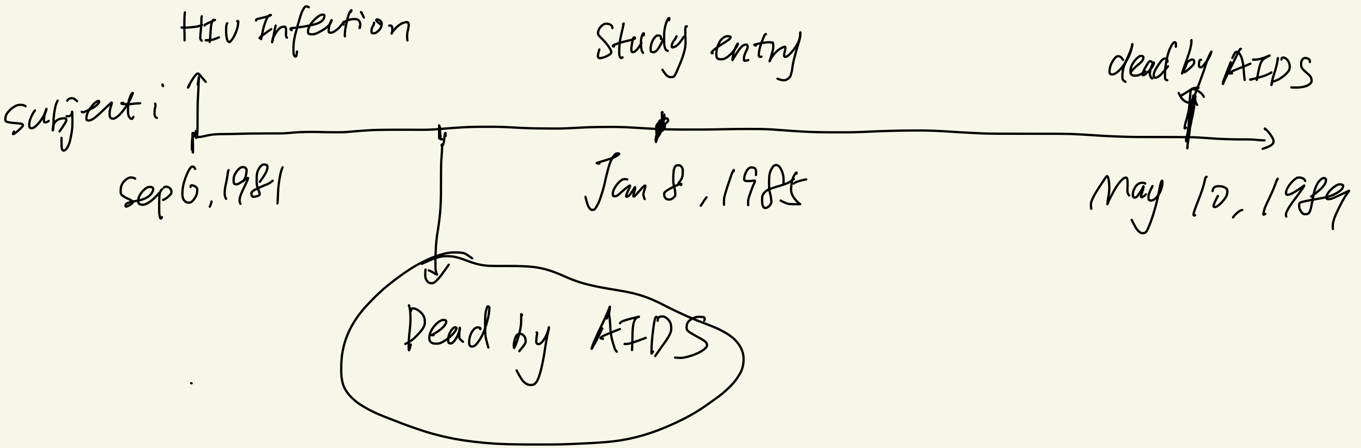
e.g.



Note: left & right censoring are special cases of interval censoring.

Left truncation (左截断):

e.g. First test of HIV: 1984年



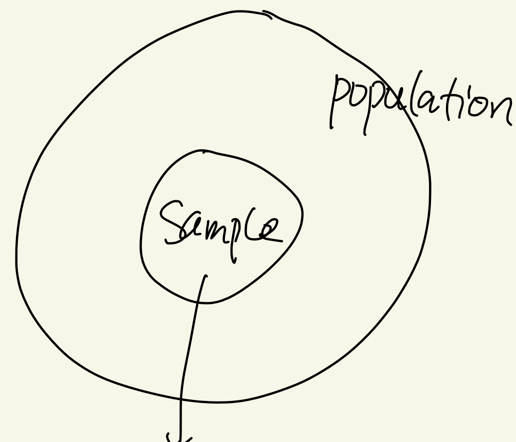
Left truncion: subjects that become HIV infected and have a short time to Dead are likely to be missed by the study. Those who are missed are called left-truncated.

结果: $\boxed{\text{biased sample}}$

(有偏样本)

survivor bias

(幸存者偏差)



随机抽样(无偏样本)

Note: In this course, we mainly focus on right-censored data.

§1.3 Survival function and Hazard function

生存函数 风险函数

T^* is survival time, i.e. time from initiation (e.g. start of study; HIV Infection) to event occurrence.

Note: We mainly focus on continuous T^*.

$S(t) = P(T^* > t)$ is survival function $= 1 - P(T^* \le t)$ → 分布函数

$$\lambda(t) = \lim_{h \to 0^+} \frac{P(t \le T^* < t+h \mid T^* \ge t)}{h}$$ is hazard function $[t, t+h)$

$$\Rightarrow \lambda(t) = \lim_{h \to 0^+} \frac{P(t \le T^* < t+h)}{h} \Big/ S(t)$$

$$P(t \le T^* < t+h \mid T^* \geq t)$$

$$= \frac{P(t \le T^* < t+h \cap T^* \geq t)}{P(T^* \geq t)}$$

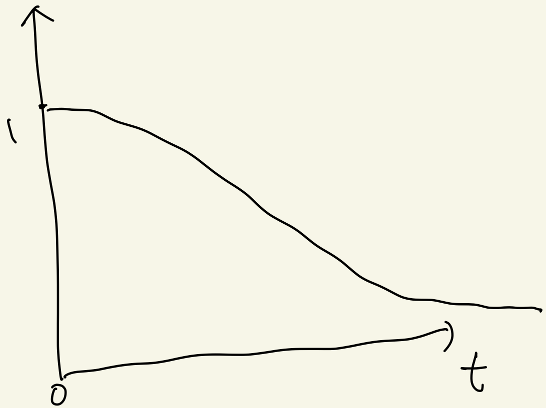
$$= \frac{P(t \le T^* < t+h)}{P(T^* \geq t) \to S(t)}$$

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

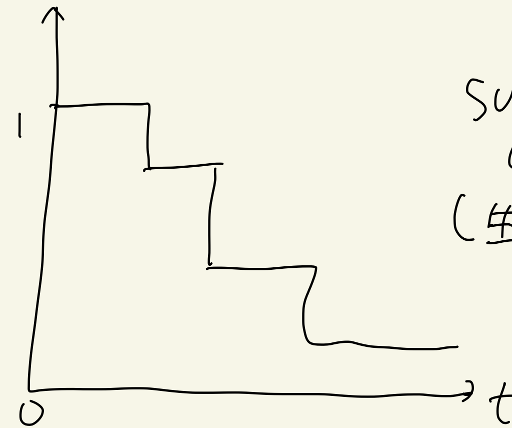
where $f(t)$ is density of T^*.

C^* is censoring time. i.e. time from initiation until end of observation (e.g. end of study; lost of followup; withdrawn

Theoretical $S(t)$



Estimated $S(t)$

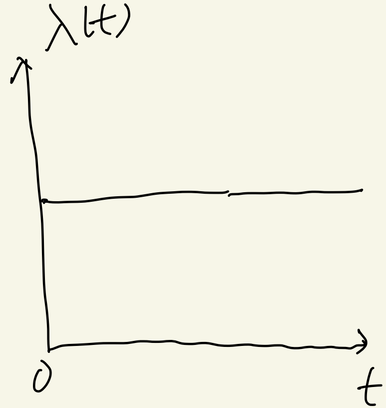


Survival curve (生存曲线)

Theoretical $\lambda(t)$:

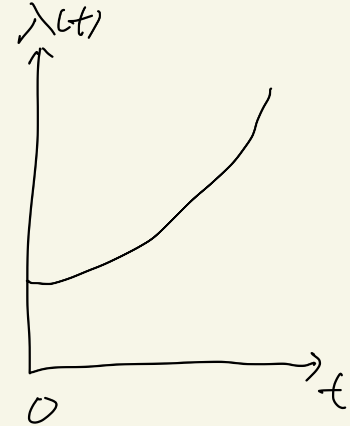
① Healthy persons

event : dead

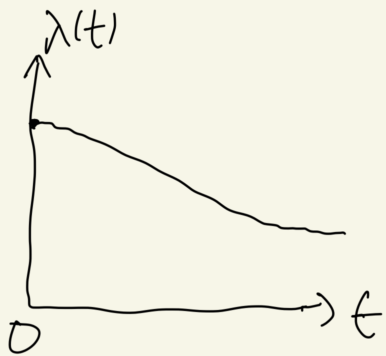


② Cancer patients

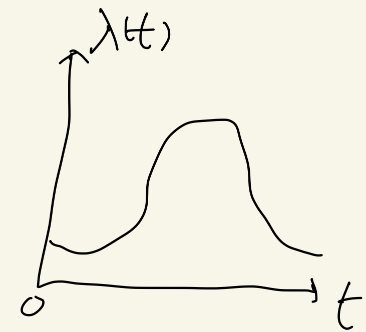
event : dead



③ patients recovering from surgery



④ 肺结核



One to one Correspondence of $S(t)$ and $\lambda(t)$:

$$\begin{cases} S(t) = \exp\left\{-\int_0^t \lambda(u)\,du\right\} \\ \lambda(t) = -\left\{\dfrac{dS(t)/dt}{S(t)}\right\} \end{cases}$$

为什么?

① $S(t)$ is more informative. It directly describes survival experience (survival curves).

② $\lambda(t)$ is easier to identify a model from survival data (hazard models. e.g. Cox model)

§1.4 Censoring Assumption

We assume **Random censoring**: ✓ $T^* \perp\!\!\!\perp C^* \mid X$ → covariate

(remove X if there is no covariates in the data)

A more general assumption is

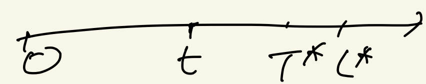
independent censoring assumption:

$$\lim_{h\to 0^+} \frac{P(t\le T^* < t+h \mid T^*\geq t,\ C^*\geq t,\ X)}{h}$$

→ this smaller group (under observation at time t) is

$$= \lim_{h\to 0^+} \frac{P(t\le T^* < t+h \mid T^*\geq t,\ X)}{h}$$

→ representative of this larger group

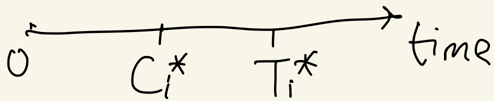


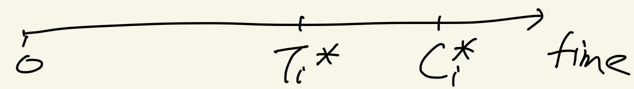
§1.5. Math notions of observed data.

$T_i = \min(T_i^*, C_i^*)$ is the observed time of subject i.
观测值

$\delta_i = I(T_i^* \le C_i^*)$ is the status of subject i

$$= \begin{cases} 1, & \text{event occurred, } T_i^* \text{ is observed (i.e. } T_i = T_i^*) \\ 0, & \text{censored} \end{cases}$$





The typical observed right-censored data are e.g. KK2012 P625

$\{ (T_i, \delta_i, X_i(t));\ 0 \le t \le T_i,\ \ i = 1, 2, \cdots, n \}$ (we assume i.i.d. 独立同分布)

§1.6 Counting process and Martingale
(计数过程) (鞅)

Def Let $\{M(t), t \geq 0\}$ be a stochastic process. It's a martingale if $E[M(t) \mid \mathcal{F}_s] = M(s)$ for all $t \geq s$ (*)

where $\mathcal{F}_s$ is the history information in $[0, s]$.

or $E[dM(t) \mid \mathcal{F}_{t-}] = 0$ (*)'

or simply $E[dM(t) \mid \text{past}] = 0$

increment of $M(t)$ in $[t, t+dt)$

history prior to t.

* (打星号代表不作要求)

Note: $\mathcal{F}_s = \sigma\{(N_i(u), Y_i(u), X_i(u)) ; i=1, \cdots, n, 0 \leq u \leq s\}$. generated by A.

$\sigma(A)$ is σ-algebra of a set A. 意义: $\sigma(A)$ is the history ~~information~~ ✓

$N_i(t) =$ # of observed events in $[0,t]$ for subject i;

counting process

$Y_i(t) = \begin{cases} 1, & \text{if subject } i \text{ is under observation and at risk of event occurrence at time } t; \\ 0, & \text{otherwise.} \end{cases}$

at-risk process

For right-censored survival data,

$$\begin{cases} N_i(t) = I(T_i^* \le t, \ \delta_i = 1) = I(T_i \le t, \ \delta_i = 1) \\ Y_i(t) = I(T_i^* \geq t, \ C_i^* \geq t) = I(T_i \geq t) \end{cases}$$

$\Downarrow$

Note: Data $\{(T_i, \delta_i), i = 1, \cdots, n\}$ 等价于 $\{(N_i(t), Y_i(t)), i = 1, \cdots, n, t \geq 0\}$

*(不要求) 由 $\lambda(t)$ 定义及 independent censoring 假设

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

$$\Rightarrow E[dN(t) \mid \mathcal{F}_{t-}) = Y(t)\lambda(t)dt$$

$$\Rightarrow E[dM(t) \mid \mathcal{F}_{t-}) = 0 .$$

where $M(t) = N(t) - \int_0^t Y(u)\lambda(u)du$

$$dM(t) = dN(t) - Y(t)\lambda(t)dt.$$

$$dN(t) = \underset{\Downarrow}{Y(t)\lambda(t)dt} + dM(t)$$

解释: "observation" = "signal" + "noise"

Key: 如何从 "observation" 中 recover "$\underline{\underline{\lambda(t)}}$" ?? (第2章)

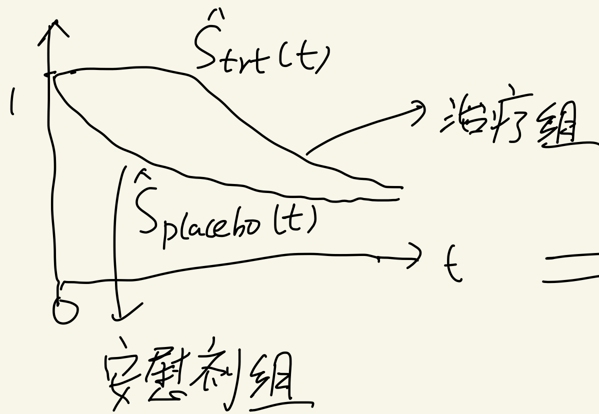
§ 1.7 Main tasks of survival Analysis

① Estimate survival function and hazard function from data

survival function → Kaplan-Meier estimator (K-M)

hazard function → Nelson–Aalen Estimator (N-A)

② Compare survival curves
- i) by plot (by K-M or Cox regression)
- ii) by test (log-rank test)
  对数秩



⟹ This is a plot of two survival curves. It implies subjects in the treatment group tends to survive longer than those in the placebo group.

③ Regression models that links T^* with covariates $X(t)$

e.g. Cox model: $\lambda(t|X(t)) = \lambda_0(t) e^{\beta^T \cdot X(t)}$ (T 代表列向量 β 转为行向量 β^T)

识别哪些因素(自变量)
i) 增加
ii) 降低
iii) 不改变
发生事件的风险

④ Survival prediction
- C-Index
- Cox model
- Random survival forest