ACKNOWLEDGEMENT

I would like to express my gratitude to Drs. Jianwen Cai and Ming-Hui Chen who generously provided their course materials for the preparation of this course.

CHAPTER 1: INTRODUCTION TO SURVIVAL ANALYSIS AND TIME-TO-EVENT DATA

Topics in this section:

- Survival / Failure time
- Censoring / Truncation
- Notation
- Functions used in survival analysis
- Counting process

References: TP Chapter 1; FH Chapter 0; KP Chapter 1

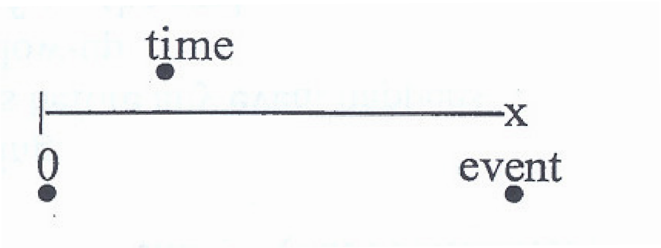
Note: TP: Therneau and Grambsch, FH: Fleming and Harrington, KP: Kabfleisch and Prentice

What is Survival Analysis?

1. The analysis of *time-to-event* data, generally called *survival analysis*, arises in many fields of study, including medicine, public health, epidemiology, biology, engineering (reliability), social studies (event history analysis), economics (duration time), sports and so on.

2. In survival analysis, interest centers on a group or groups of individuals for each of whom (or which) there is defined a *point event*, often called *failure*, occurring after a length of time called the failure time.

Outcome (Response) variable: ***time*** until an ***event*** occurs.



Example of failure times:

- time from *cancer incidence* to **death**
- time from *HIV infection* to **AIDS onset**
- time from *committee approval* to **policy implementation**
- time from *inception of a hedge fund* to **exiting the database**
- Any other examples?

NOTE: *Time origin* and **event (end-point)** should be clearly defined.

Assumption for the course (99%): at most 1 failure/individual.

Note: If # of failure $> 1 \Rightarrow$ competing risks or multivariate failure times (special topic section)

Our interest:

- Distribution of failure times
- Comparison of the failure times
- Effect of explanatory variables on survival

If we know the **EXACT** failure time:

– apply all the techniques we have learned in statistics.

Challenges in Survival Analysis

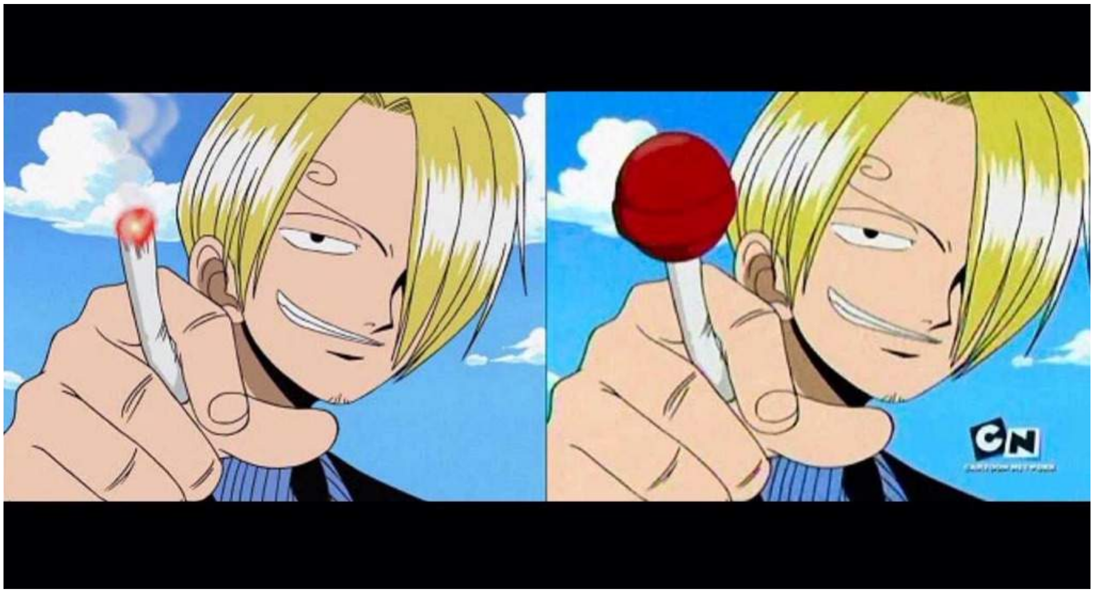
1. **Time domain**: The main outcome in survival analysis is the *time-to-event*, which is almost always **nonnegative**.
2. **Shape of time distribution**: The distribution of the time-to-event is typically **skewed**.
3. **Censoring**: Survival data are often **right censored**:

   – Survival times are known for only a portion of the individuals under study, and the remainder of the survival times are known only to exceed certain values.
4. **Tail probability**: Typically, there are few subjects at risk in the tail of the survival curve after sufficient follow-up:

   – Estimation of the tail of the survival curve can be quite difficult.

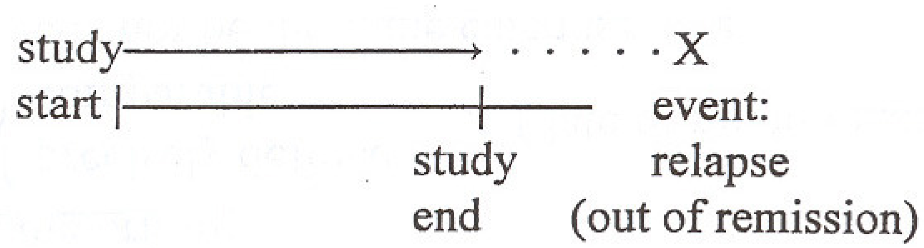
Censoring

What is censoring? incomplete observation of failure time



https://aminoapps.com/c/anime/page/blog/whats-the-worst-anime-censorship/Lkt8_ueWeqe38Drw6mZoNw1r7lxw57

ex): Leukemia patients in remission



Challenges :

1. Since the duration of any study is finite, event of interest may not be observed for all subjects.
2. In general, when failure times for some subjects are not observed (usual case in survival analysis), standard statistical techniques cannot be directly applied to survival data.

3. Subjects whose failure times are unobserved are said to be *censored.*
4. Estimation methods of survival analysis are built to extract information from all subjects including subjects with censored observations

Sources of Censoring: Subjects may be censored for different reasons!

- ***Administrative* censoring**:
  - study ends before the event has occurred
  - often independent of failure time
- **Loss to follow-up**:
  - A subject cannot be traced any longer; no longer under observation
  - ex) In a follow-up study, an individual leaves the country, so can no longer be followed through the national mortality database
  - Censoring may be related (indirectly) to the failure time
- **Withdrawal from study**:
  - ex 1) A patient drops out of clinical trial because he is too sick to participate.
  - ex 2) A subject discontinues participation in trial because her symptoms have subsided.
  - *Dependent* censoring ("informative drop-out"), i.e., censoring is related to failure time, will be a concern.

ex) made-up data (AIDS patients)

steroid:	$1, 1, 1, 1^+, 4^+, 5$	
placebo:	$1^+, 2^+, 3, 3^+$	(in years)

'+' means right censored

Some quick fixes for censored failure time data

1) Count censored observations as survival times (Ignore +'s) :

- Sample mean ($\bar{X}$) for the "steroid" group: $(1+1+1+1+4+5)/6 = 13/6$
- Any problem?
- $\bar{X}$ is likely to *underestimate* the true mean.

2) Delete censored observations (delete +'s) :

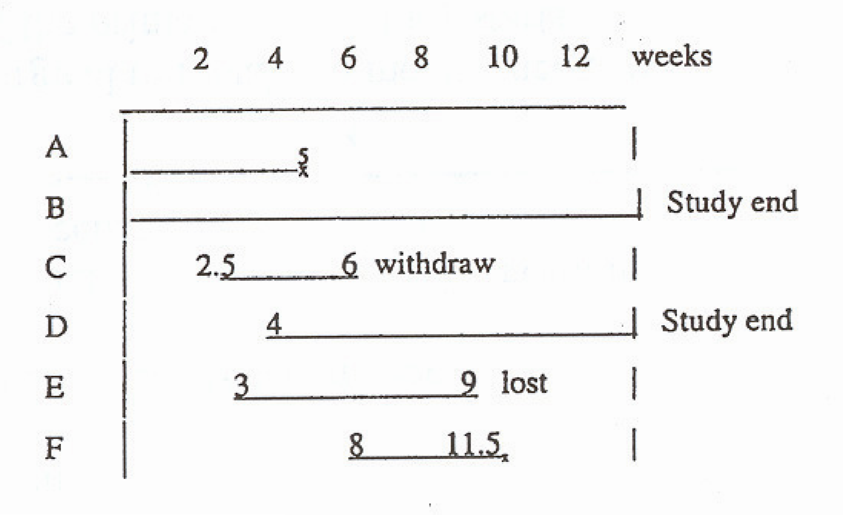
- Sample mean ($\bar{X}$) for the "steroid" group: $(1+1+1+5)/4 = 2$
- Any problem?
- Loss of information

Regular techniques do not work here!

Types of Censoring

- i: subject
- t: specific time value
- T_i: potential failure time for the i^{th} subject
- C_i: potential censoring time for the i^{th} subject
- $X_i = \min(T_i, C_i)$ observed time
- $\delta_i = \begin{cases} 1 & \text{if} \quad T_i \leq C_i \text{ (uncensored)} \\ 0 & \text{if} \quad T_i > C_i \text{ (censored)} \end{cases}$

1. Right Censoring: (most of the course)



person	X_i Observed Time	δ_i Failed (1) Censored (0)
A	5	1
B	12	0
C	3.5	0
D	8	0
E	6	0
F	3.5	1

- Events: A and F (2), Censored: B, C, D and E (4)
- Type of data to be analyzed in survival analysis

(1) Type I Censoring:

Study ends when a certain time point is reached.

ex): Consider a large scale animal experiment conducted in which mice were fed a particular dose of a carcinogen.

- Goal: Assess the effect of the carcinogen on survival
- Mice were followed from the beginning of the experiment until death or until a prespecified censoring time was reached, when all those still alive were sacrificed (censored).

Other forms of Type I censoring: *progressive Type I* censoring and *generalized Type I* censoring

(i) Progressive Type I censoring: subjects have different, **fixed**-sacrifice times

ex): Mice were randomly divided into 4 dose-level groups and each mouse was followed until death or until a prespecified sacrifice time (42 or 104 weeks) was reached.

- Shorter sacrifice times may be applied to the lower dose groups, while longer sacrifice times may be applied to the higher dose groups.
- The two sacrifice times were chosen to reduce the cost of maintaining the animals while allowing for (limited) information on the survival experience of longer lived mice.

(ii) Generalized Type I censoring: subjects enter the study at different times and the terminal point of the study is predetermined by the investigator.

- Censoring times are **known** when an individual is entered into the study.

(2) **Type II Censoring**: common in reliability analysis

Study ends when a certain number of failures occur.

Ex): Study of the life time of light bulbs:
study ends when the 10th light bulb
burns out.

$$\sum \delta_i = d \quad \text{fixed}$$

Note: *Right Censoring* - the person's exact survival time becomes incomplete at the right side of the follow-up period.

(3) **Random censoring**:

Censoring times are random.

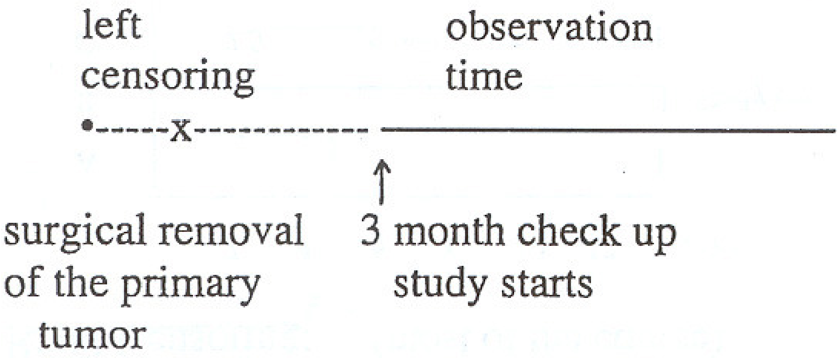
Ex): loss to follow-up, withdrawal from the study

Note: Let's focus on right censoring. Suppose $T_1, T_2, \ldots, T_n \sim f(t)$ and $C_1, C_2, \ldots, C_n \sim g(c)$. Then, we observe $X_i = \min(T_i, C_i)$ for $i = 1, 2, \ldots, n$. In type I censoring, C_i is fixed (at C_r or C_{r_i}). In random censoring, C_i is random.

2. Left Censoring: less common in practice

A person's survival time becomes incomplete at the left side of the follow-up period for that person.

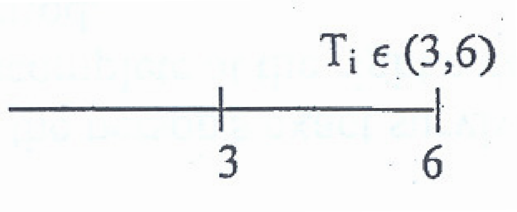
- Event: recurrence of cancer


left censoring
observation time
surgical removal of the primary tumor
3 month check up study starts

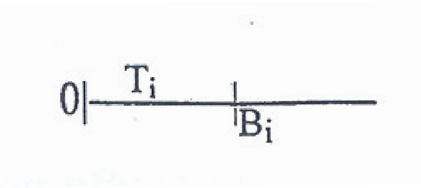
Note: Sometimes, left censoring occurs due to the measuring device having a minimal lower limit of detection.

(3) Interval Censoring:

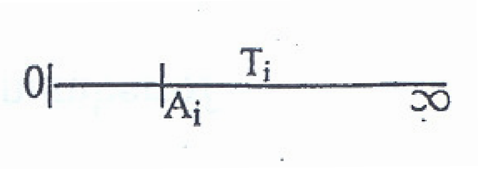
$T_i \in (A_i, B_i)$



$A_i = 0$
$B_i \neq \infty$ ⇒ left censoring

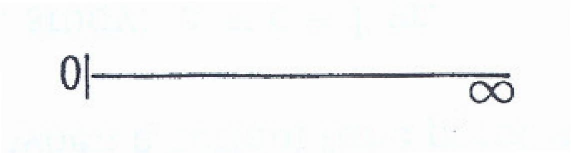


$B_i = \infty$
$A_i \neq 0$ ⇒ right censoring



Note:

− If ($A_i = 0$ and $B_i = \infty$), then there is no info. contained.



Doubly censoring:

- If left censoring occurs in a study, right censoring may also occur, and the lifetimes are considered *doubly censored.*
- *Doubly censoring* is a special case of interval censoring.

Truncation

What is truncation?

- *Truncation* is defined as a condition which screens certain subjects so that the investigator will not be aware of their existence.
- For truncated data, only individuals who experience some event are observed by the investigator.

1. Left Truncation:

If Y is the time of the event which truncates individuals, then, for left-truncated samples, only individuals with $T \geq Y$ are observed.

ex): Study of the Channing House retirement center:

- Ages at death (in months) and
- Ages at which individuals entered the retirement community are recorded.
- Since an individual must survive to a sufficient age to enter the retirement center, all individuals who died earlier will not enter the center (out of the investigator's cognizance).
- Such individuals have no chance to be in the study and are considered *left truncated.*

Note: As opposed to left censoring, where we have partial information on individuals who experience the event of interest prior to their age at entry, for left truncation, these individuals were never considered for inclusion in the study.

(2) Right Truncation:

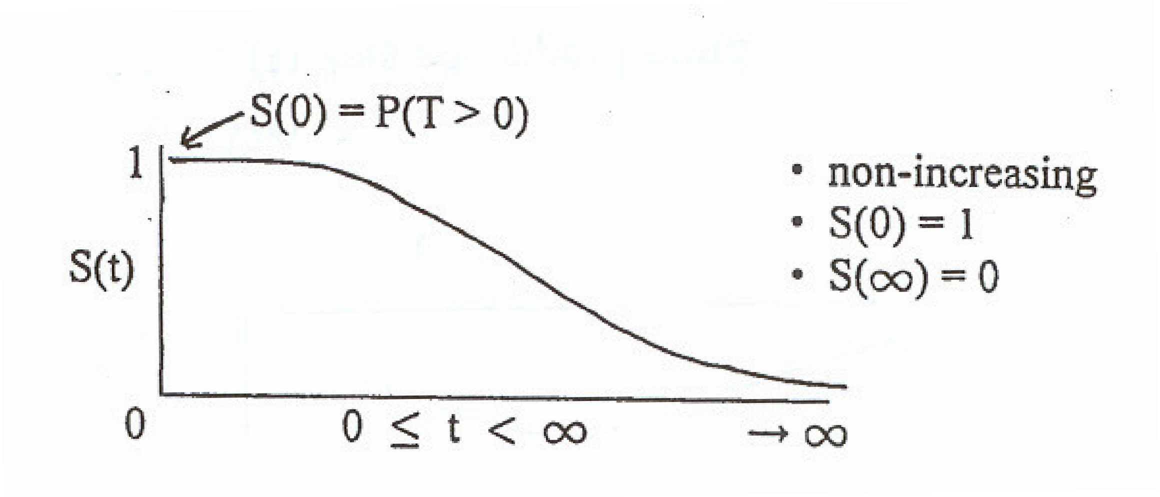
Right truncation occurs when only individuals who have experienced the event are included in the sample, and any individuals who have yet to experience the event is not observed.

ex): Estimating the distribution of stars from the earth:

- Stars too far away are not visible and are right truncated

Survival function

- $T_1, \ldots, T_n$ are *iid* (independent and identically distributed) non-negative random variates. We assume that T_i is continuous (for now).
- $S(t) = P(T_i > t) \left\{ \begin{array}{l} \text{Prob. that a person survives longer} \\ \text{than some specified time } t \end{array} \right.$
- Theoretically, (smooth curve)



Note: A survival function $S(t)$ that satisfies the above two properties ($S(0) = 1, S(\infty) = 0$) is termed as a *proper* survival function. In the context of equipment or manufactured item failures, $S(t)$ is referred to as the reliability function, which measures how reliable an equipment or manufactured item is at time t.

Density and Cumulative Distribution Function

- Density and cumulative distribution function (CDF) are defined as usual:

$$\begin{aligned} F(t) &= P(T_i \leq t) = 1 - S(t) \\ f(t) &= \lim_{dt \to 0} \frac{1}{dt} P(t \leq T_i < t + dt) \\ &= \frac{dF(t)}{dt} = -\frac{dS(t)}{dt} \\ F(t) &= \int_0^t f(s) ds \end{aligned}$$

- If $F(t)$ is continuous and differentiable , the expression $f(t) = F'(t) = dF(t)/dt$ is well-defined.
- In case F has *discontinuity points*, we can use Stieltjes integration.

Stieltjes Integration

- Continuity:
  - A function, $G(t)$, is *right-continuous* at $t = a$ if:

$$\lim_{t \to a^+} G(t) = G(a)$$

  - $G(t)$, is *left-continuous* at a if:

$$\lim_{t \to a^-} G(t) = G(a)$$

  - $G(t)$, is *continuous* at a if:

$$\lim_{s \to a^+} G(s) = G(a) = \lim_{t \to a^-} G(t)$$

- Stieltjes Integration:
  - Let $G(t)$ be a non-decreasing, right-continuous function with left-hand limits (*cadlag* function)

- Set $g(t) = G'(t)$, but suppose that G has discontinuities at: $t_1 < t_2 < \ldots < t_J$
- Set $\Delta G(t) = G(t) - G(t^-)$; we can write:

$$G(t) = \int_0^t g(s)ds + \sum_{j:t_j \leq t} \Delta G(t_j)$$

- note: if there were no discontinuities over $(0, t]$, we could write:

$$G(t) = \int_0^t g(s)ds$$

- setting $dG(t) = g(t)dt + \Delta G(t)$, we can always write

$$G(t) = \int_0^t dG(s),$$

  whether or not there are discontinuity points in $(0, t]$

Hazard Function

- **Definition**:

$$\lambda(t) = \lim_{dt \to 0^+} \frac{1}{dt} P(t \leq T_i < t + dt | T_i \geq t)$$

- Only requirement: $\lambda(t) \geq 0$ for $t \geq 0$
- Conditional failure *rate*; not a *probability*
- **Interpretation**: conditional rate of failure in $[t, t+dt)$, given that failure has not occurred as of time t
- Informally, $f(t)$ is related to Pr{die at t}

$$\begin{aligned} P\{\text{die at } t\} &= \Pr\{\text{die at } t, \text{ survive until } t\} \\ &= \Pr\{\text{die at } t | \text{survive until } t\} \Pr\{\text{survive until } t\} \end{aligned}$$

  which is related to $\lambda(t)S(t)$

- Formal derivation incorporating continuity of $F(t)$? **HW**

Cumulative Hazard Function

- **Definition**: $\Lambda(t) = \int_0^t \lambda(s)ds$
- **Properties**:

$$\begin{aligned} \Lambda(0) &= 0 \\ \Lambda(t) &\geq 0 \\ \lim_{t\to\infty} \Lambda(t) &= \infty \end{aligned}$$

- Relationship between the survival and cumulative hazard functions:

$$\lambda(t) = -\frac{d}{dt} \log S(t) \quad \text{(Why? } \textbf{HW}\text{)}$$

- Integrate both sides over $(0, t]$,

$$\begin{aligned} \Lambda(t) &= -\int_0^t d \log S(u) \\ &= -\log S(u)|_0^t \\ &= -\log S(t) \\ \exp\{-\Lambda(t)\} &= S(t). \end{aligned}$$

Summary of Functions Characterizing Survival Distribution

$$\begin{aligned} F(t) &= \int_0^t f(s)ds = 1 - S(t) \\ S(t) &= \int_t^\infty f(s)ds = e^{-\Lambda(t)} \\ \lambda(t) &= \frac{f(t)}{S(t)} = \frac{dF(t)/dt}{S(t)} \\ \Lambda(t) &= \int_0^t \lambda(s)ds = -\log S(t) \\ f(t) &= \frac{d}{dt}F(t) = -\frac{d}{dt}S(t) = S(t)\lambda(t) \end{aligned}$$

- Any one of the five functions $f(t)$, $S(t)$, $F(t)$, $\lambda(t)$, or $\Lambda(t)$ is sufficient to define the failure time distribution.

Some special distributions

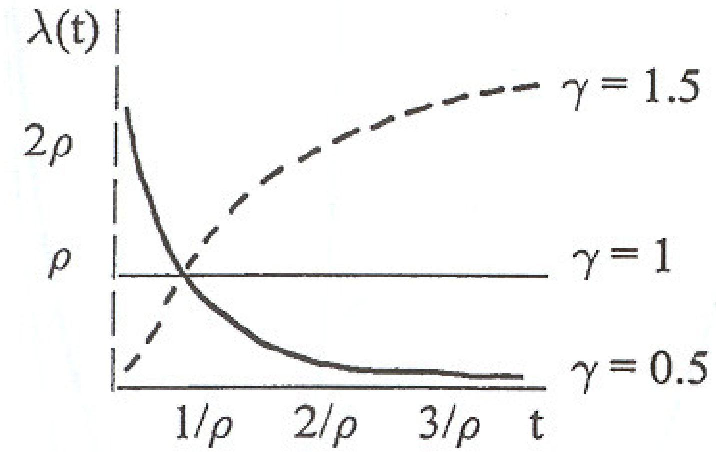
1. **exponential distribution** (w/parameter $\rho > 0$)

$$\begin{array}{lll} f(t) = \rho \exp(-\rho t) & & E(T) = \rho^{-1} \\ S(t) = \exp(-\rho t) & & F(t) = 1 - \exp(-\rho t) \\ \lambda(t) = \rho & \begin{array}{c}\text{constant hazard} \\ \text{(exponential model)}\end{array} & \Lambda(t) = \rho t \end{array}$$

- lack of memory ($\forall t_0 > 0, T - t_0 | T > t_0 \sim T$) (**Exercise**)
- coef. of variation $= \sqrt{\mathrm{Var}(T)} / \, \mathrm{E}(T) = 1$ (reference for dispersion)
- empirical check of the data plot $\log(\hat{S}(t))$ vs. t (should approximate a straight line through origin) (Why? **HW**)
  - $\hat{S}(t)$: estimated survival function (usually Kaplan-Meier estimator)
- If T has an arbitrary continuous distribution, then $\Lambda(T)$ has an exponential distribution with unit parameter (Why? **HW**)

2. **Weibull distribution** (w/parameter ρ & $\gamma > 0$)

$$\begin{array}{ll} f(t) = \rho\gamma t^{\gamma-1}\exp(-\rho t^{\gamma}) & \\ S(t) = \exp(-\rho t^{\gamma}) & F(t) = 1 - \exp(-\rho t^{\gamma}) \\ \lambda(t) = \rho\gamma t^{\gamma-1} & \Lambda(t) = \rho t^{\gamma} \end{array}$$



- important generalization of the exponential distribution; allows for a power dependence of the hazard on time.

- $\lambda(t)$ is monotone decreasing for $\gamma < 1$
  monotone increasing for $\gamma > 1$
  reduces to the constant exponential hazard if $\gamma = 1$

- empirical check of the data
  - plot $\log(-\log\widehat{S}(t))$ vs $\log t$

  - plot should give approximately a straight line: slope $\sim \gamma$ and intercept $\sim \log\rho$ (Why? **HW**)

3. **Log-normal distribution** (w/parameter μ & σ)

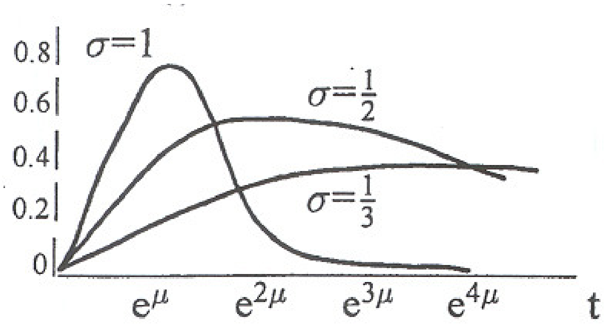
$$f(t) = \frac{1}{\sqrt{2\pi}\sigma t} \exp\left[-\frac{\{\log(t)-\mu\}^2}{2\sigma^2}\right], \sigma > 0, -\infty < \mu < \infty$$

$$S(t) = 1 - \Phi\left[\frac{1}{\sigma}\{\log(t)-\mu\}\right]$$

$\uparrow$ incomplete normal integral

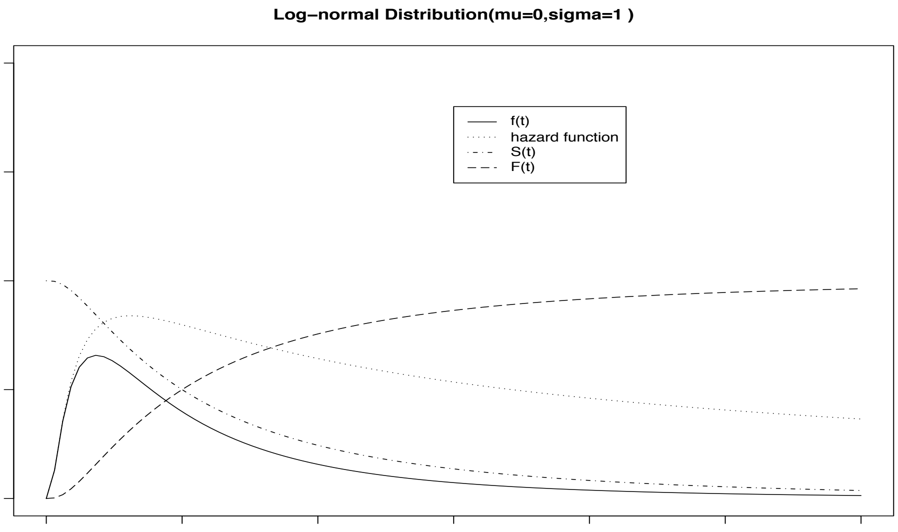
$$F(t) = \Phi\left[\frac{1}{\sigma}\{\log(t)-\mu\}\right]$$

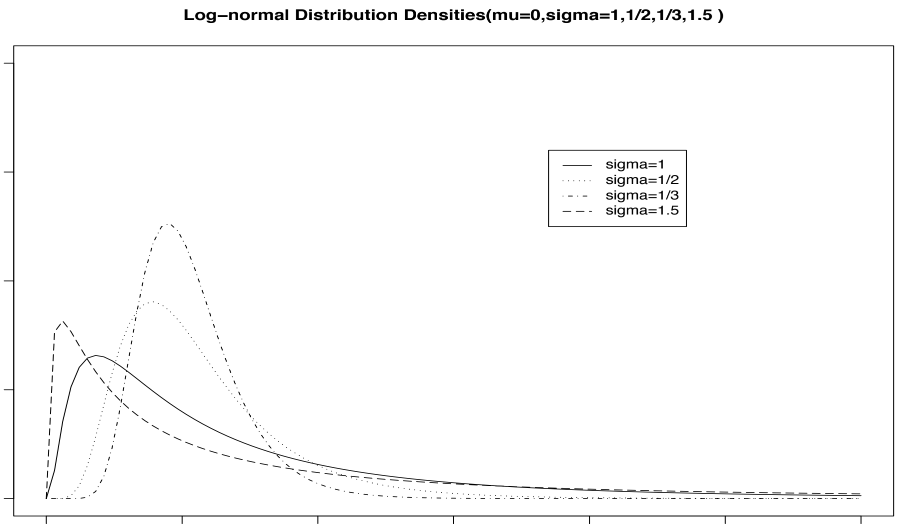
$$\lambda(t) = \frac{f(t)}{S(t)} : \text{ no closed form}$$

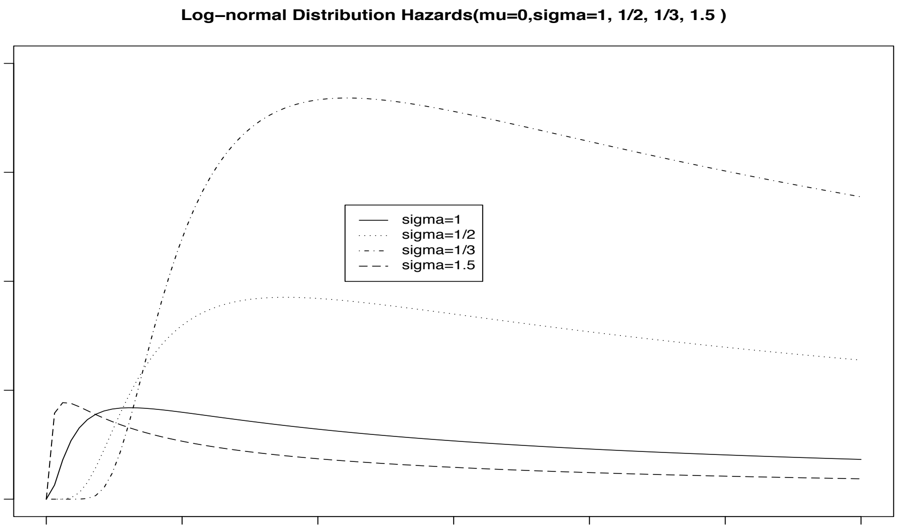


$$\lambda(0) = 0, \qquad \lambda(t) \to 0 \text{ as } t \to \infty$$

- simple to apply if no censoring
- sensitive to the small failure times
- log-logistic dist$\underline{^n}$ provides a good approximation to the log-normal distribution (may frequently be a preferable survival time model)
- $\log(T) \sim N(\mu, \sigma^2)$


Log−normal Distribution(mu=0,sigma=1)


Log−normal Distribution Densities(mu=0,sigma=1,1/2,1/3,1.5)


Log−normal Distribution Hazards(mu=0,sigma=1, 1/2, 1/3, 1.5)

4. **log-logistic distribution** (with parameters θ & κ):
logistic density with location parameter ν and scale parameter τ

$$f(x) = \tau^{-1}\exp(\frac{x-\nu}{\tau})\Big/\left\{1+\exp(\frac{x-\nu}{\tau})\right\}^2 \text{ let } \theta = -\frac{\nu}{\tau}, \kappa = \frac{1}{\tau} \text{ then,}$$

$$f(x) = \frac{\kappa\exp(\theta+\kappa x)}{(1+\exp(\theta+\kappa x))^2} \text{ next, do a change of variable from } x \text{ to } \log(t), \text{ then:}$$

$$f(t) = \frac{\exp(\theta)\kappa t^{\kappa-1}}{(1+\exp(\theta)t^{\kappa})^2} \textbf{ (HW)}$$

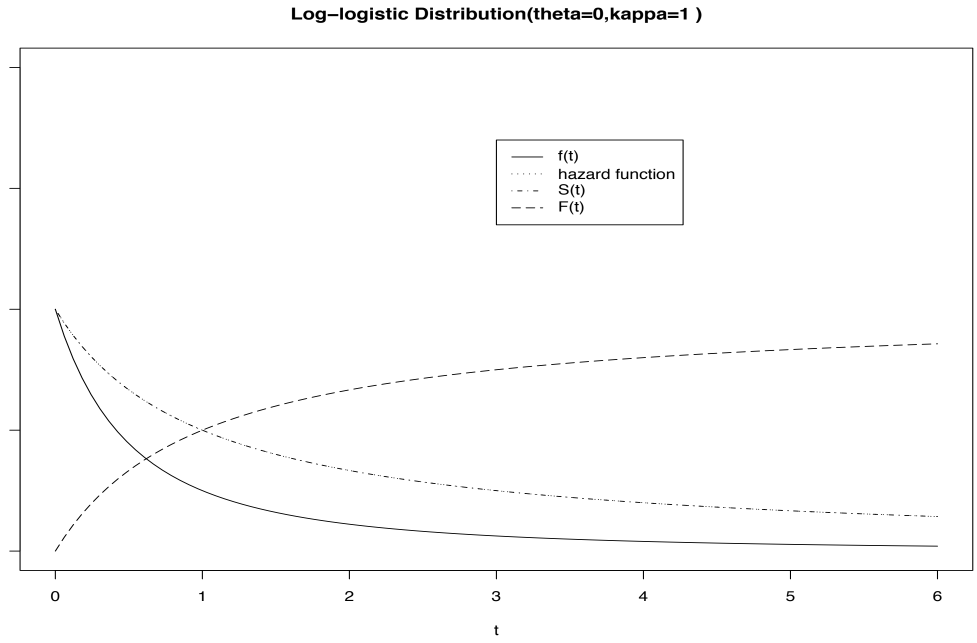
$$S(t) = \frac{1}{1+\exp(\theta)t^{\kappa}}, \text{ and } \lambda(t) = \frac{\exp(\theta)\kappa t^{\kappa-1}}{1+e^{\theta}t^{\kappa}}$$

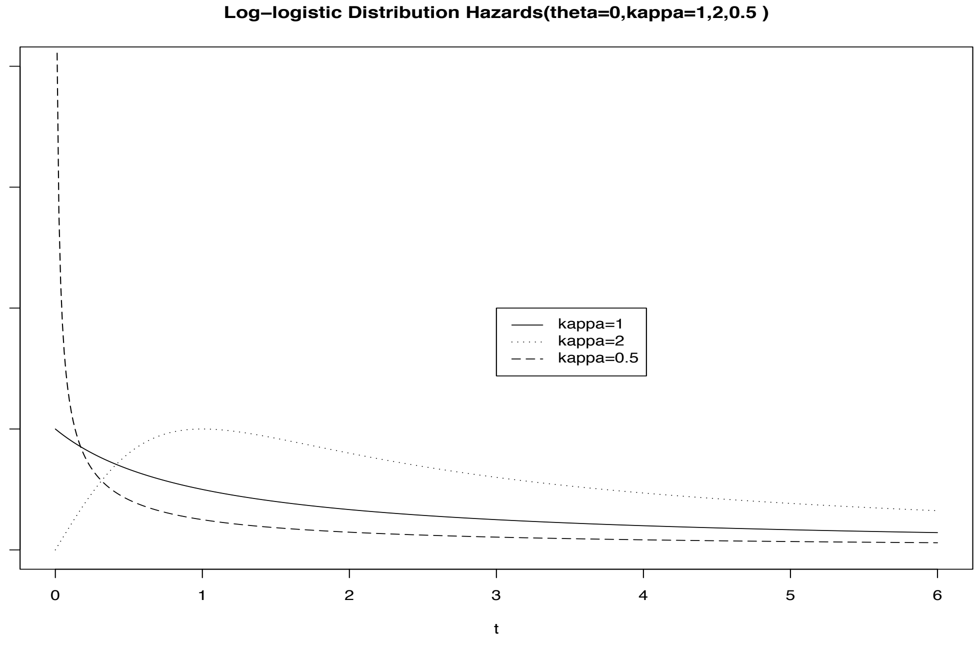
- $\log(T) \sim$ logistic distribution

- relatively simple explicit forms for $S(t), f(t)$ & $\lambda(t)$(vs. log-normal)

- $\lambda(t)$ has a single maximum if $\kappa > 1$
  is decreasing if $\kappa < 1$ (from ∞)
  is decreasing if $\kappa = 1$ (from $\exp(\theta)$)

- provides a good approximation to the log-normal distribution except in the extreme tails.


Log−logistic Distribution(theta=0,kappa=1)


Log−logistic Distribution Hazards(theta=0,kappa=1,2,0.5)

5. Other distributions

(i) Extreme Value or Gompertz Distribution

(with parameters ρ_1 & ρ_2)

$$\begin{aligned}\lambda(t) &= \rho_1 \exp(\rho_2 t), \rho_1 > 0, \rho_2 > 0,\\ \Lambda(t) &= \frac{\rho_1}{\rho_2}\left\{\exp(\rho_2 t) - 1\right\},\\ S(t) &= \exp\left(\frac{\rho_1}{\rho_2}\right)\left\{1 - \exp(\rho_2 t)\right\}.\end{aligned}$$

Gompertz-Makeham dist$\underline{^n}$ $\left(\begin{array}{c}\text{rapidly increasing}\\ \text{hazard}\end{array}\right)$

: $\lambda(t) = \rho_0 + \rho_1 \exp(\rho_2 t)$ ($\rho_0 = 0$ then Gompertz)

(ii) A Piecewise Exponential Distribution

Let $s_0 = 0 < s1 < s2 < \cdots < s_J = \infty$ denote a partition of the time axis ($R^+ = [0, 1)$). We define the hazard function by

$$\lambda(t) = \lambda_j \text{ for } s_{j-1} \le t < s_j.$$

Then, for $s_{j-1} \le t < s_j$,

$$\begin{aligned}S(t) &= \exp\{-\lambda_j(t - s_{j-1}) - \sum_{g=1}^{j-1} \lambda_g(s_g - s_{g-1})\},\\ f(t) &= \lambda_j \exp\{-\lambda_j(t - s_{j-1}) - \sum_{g=1}^{j-1} \lambda_g(s_g - s_{g-1})\},\end{aligned}$$

This is a very popular semi-parametric model for survival data. The piecewise exponential model is useful in the construction of a life table.

Product Integration

- Partition $(0, t]$ into m disjoint subintervals, $(s_{j-1}, s_j]$, such that

  $$0 \equiv s_0 < s_1 < s_2 < \ldots < s_{m-1} < s_m \equiv t$$

  such that

  $$\lim_{m\to\infty} (s_j - s_{j-1}) = 0$$

- Consider $s < t$:

  $$P(T > t) = P(T > t, T > s) = P(T > t|T > s)P(T > s)$$

- Applying this idea repeatedly and then

  $$\begin{aligned} S(t) = P(T > t) &= P(T > s_0)P(T > s_1|T > s_0)\cdots P(T > s_m|T > s_{m-1}) \\ &= \prod_{j=1}^{m} P(T > s_j|T > s_{j-1}) \\ &= \prod_{j=1}^{m} \{1 - P(T \leq s_j|T > s_{j-1})\} \end{aligned}$$

  If the partition is sufficiently fine, i.e., $m \to \infty$,

  $$\begin{aligned} \lim_{m\to\infty} \prod_{j=1}^{m} \{1 - P(T \leq s_j|T > s_{j-1})\} &= \lim_{m\to\infty} \prod_{j=1}^{m} \{1 - \lambda(s_j)ds_j\} \\ &= \prod_{s\in(0,t]} \{1 - d\Lambda(s)\}. \end{aligned}$$

Counting Process: Definition

- Stochastic process:
  A collection of random variables $X = \{X(t) : t \in \Gamma\}$ indexed by a set Γ.
  - $X(t)$ is a random variable for each t.
  - Γ is typically time and either discrete ($\{0, 1, 2, \ldots\}$) or continuous ($[0, \infty)$).
  - The realization of $X(t)$ (a function of t): *sample path.*
- Counting process, $\{N(t);\ t \geq 0\}$:
  - Stochastic process
  - $N(0) = 0$
  - counts the number of failures observed in the interval $(0, t]$
  - $N(t) < \infty$ for $t > 0$
  - $N(t)$ is right continuous, with left-hand limits; a *cadlag* process
  - $N(t) \geq N(s)$, for $t \geq s$
  - $P\{dN(t) > 1\} = 0$
    where $dN(t) = N(t^- + dt) - N(t^-)$, with t^- being the time instant immediately preceding t
- Why do we study counting processes? a motivation perspective:
  - Notation: cleaner representation of estimators
  - Asymptotic properties of estimators are much easier to derive
    (via Martingale theory)
  - Framework for univariate survival readily extends to multivariate survival
  - Methodological papers in survival analysis typically use this notation

Counting Process: Failure Time Models

- $N_i^*(t)$ = number of events for subject i, as of time t
- For univariate survival data, $N_i^*(t)$ equals 0 or 1
- $N_i^*(t)$ is a right-continuous step function
  - $N_i^*(t)$ takes the value 0 until $t = T_i$, then jumps to 1 and remains at 1 thereafter
  - $dN_i^*(t)$ takes the value 0, except at $t = T_i$, where it jumps to 1
- Note: we can always write: $N_i^*(t) = \int_0^t dN_i^*(s)$

  i.e., total number of events equals cumulation of event increments
- Referring to previous notation, $N_i^*(t) = I(T_i \le t)$

At Risk Process

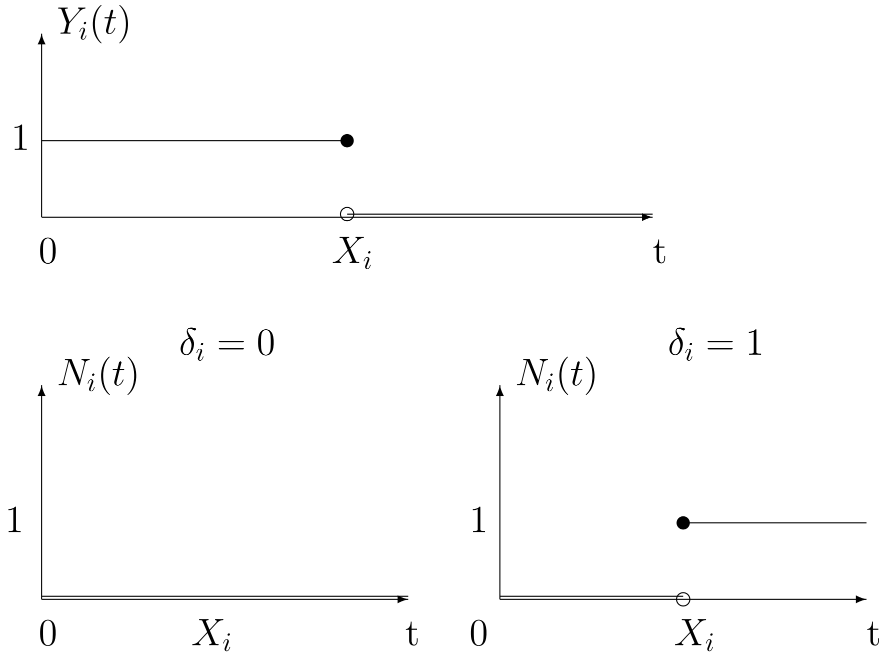
- Definition: $Y_i(t) = I(X_i \ge t)$
  - takes value 1 when subject is alive and uncensored (i.e., under observation)
  - is a left-continuous process
- To understand the left continuity of $Y(t) = \sum_{i=1}^n Y_i(t)$, consider the case where the largest observation time corresponds to a death ...
  - i.e., under left continuity, a subject has to be at risk for their own event

Counting Processes: Observed Data

- Define $dN_i(s) = Y_i(s)dN_i^*(s)$ (observed counting process)
- Set $N_i(t) = \int_0^t dN_i(s)$; $N_i(t) = \Delta_i I(X_i \le t)$
- Further, define

$$
\begin{aligned}
N(t) &= \sum_{i=1}^{n} N_i(t) \\
\overline{N}(t) &= n^{-1}N(t) \\
Y(t) &= \sum_{i=1}^{n} Y_i(t) \\
\overline{Y}(t) &= n^{-1}Y(t)
\end{aligned}
$$

- Graphs of $N_i(t)$, $dN_i(t)$ and $Y_i(t)$:



- $N_i(t)$ starts off at 0 and jumps to 1 at X_i (and stays there) if $\Delta_i = 1$; otherwise, it remains at 0
- $dN_i(t)$ starts off at 0; if $\Delta_i = 1$, then $dN_i(t)$ jumps to 1 at X_i then goes back to 0. if $\Delta_i = 0$, it remains at 0
- $Y_i(t)$ starts at 1 and drops to 0 at $t = X_i$

Discrete Survival Times

- T_i may be a discrete random variate
  - i.e., suppose T_i can only take values: $t_1 < t_2 < \ldots$
  - **ex**) if death times are grouped into small intervals (e.g., days, weeks, etc)
- In reality, T_i is almost always discrete; but, the continuous time framework is usually employed if the underlying time is believed to be continuous.
- Probability function: $f_j = f(t_j) = P(T_i = t_j)$
- Survival function: $S(t_j) = P(T_i > t_j) = \sum_{k:t_k>t_j} f(t_k)$

$$(\textbf{note}: f_j = S(t_{j-1}) - S(t_j))$$

- Hazard function:

$$\begin{aligned}\lambda_j = \lambda(t_j) &= P(T_i = t_j | T_i \geq t_j) \\ &= \frac{f(t_j)}{S(t_{j-1})}\end{aligned}$$

- combining interval-specific probabilities,

$$S(t_j) = \prod_{k=1}^{j} \{1 - \lambda_k\}$$