Chapter 2
(AST305) Lifetime Data Analysis I

Md Rasel Biswas

Lecture Outline

Section 1

2. Observed Schemes, Censoring, and Likelihood

Subsection 1

2.1 Introduction

Preliminary Discussion of Likelihood

- Suppose that the probability distribution of potentially observable data in a study specified up to a parameter vector θ

- The **likelihood function** for θ is a function of θ that is proportional to the probability density/mass of the observed data, viewed as a function θ

$$L(\theta) \propto Pr(\text{Data};\ \theta) \qquad (2.1.1)$$

- Interpretation of notation:
    - Data $\rightarrow$ observed data
    - $Pr\ \rightarrow$ probability density or mass function from which the data are assumed to arise
    - $L(\theta; \text{Data})\ \rightarrow$ is a more formal notation for likelihood function

Asymptotic Results and Large Sample Methods

- Assume that the data consist of a random sample $y_1, \ldots, y_n$ from a distribution with density function $f(y; , \theta)$, where

  - $\theta = (\theta_1, \ldots, \theta_k)' \in \Omega$ is a vector of unknown parameters
  - For simplicity, each y_i is considered scalar, though it may be vector-valued in general

- The likelihood function is

$$L(\theta) = \prod_{i=1}^{n} f(y_i; \theta)$$

- If y_i's are independent but **not identically distributed** then $f(y_i; \theta)$ is replaced by $f_i(y_i; \theta)$ in the definition of likelihood function

Asymptotic Results and Large Sample Methods

- Let $\hat{\theta}$ be a value in Ω at which $L(\theta)$ is maximized

  - $\hat{\theta}$ is called the **maximum likelihood estimator (MLE)**.

  - In most simple settings $\hat{\theta}$ exists and is unique

- It is often convenient to work with the log-likelihood function, which is also maximized at $\hat{\theta}$

$$\ell(\theta) = \log L(\theta) = \sum_i \log f(y_i; \theta)$$

Asymptotic Results and Large Sample Methods

- The MLE $\hat{\theta}$ can be obtained by solving the system of score equations

$$U_j(\theta) = \mathbf{0} \quad (j = 1, ..., k)$$

  - $U_j(\theta) = \frac{\partial \ell(\theta)}{\partial \theta_j} \to j$th score or score function
  - $\mathbf{U}(\theta) = [U_1(\theta), ..., U_k(\theta)]' \to$ *score vector* of order $k \times 1$

Asymptotic Results and Large Sample Methods

- The score vector $\mathbf{U}(\theta)$ is asymptotically (k-variate) normal with

  - mean $\mathbf{0}$
  - variance–covariance matrix $\mathcal{I}(\theta)$, the **Fisher information matrix**

- The $(j, j')th$ entry of $\mathcal{I}(\theta)$ is

$$\mathcal{I}_{jj'}(\theta) = E\left(\frac{-\partial^2 \ell(\theta)}{\partial \theta_j \, \partial \theta_{j'}} \right) \quad j, j' = 1, ..., k$$

- Terminology:

  - $\mathcal{I}(\theta) \rightarrow$ Fisher or expected information matrix

  - $I(\hat{\theta}) \rightarrow$ observed information matrix

- Under mild regularity conditions

  - $\hat{\theta}$ is a **consistent** estimator of θ

  - $I(\hat{\theta})/n$ is a consistent estimator of $\mathcal{I}(\theta)/n$

Optimization Methods for Maximum Likelihood

- The MLE corresponds to the maximizer of $\ell(\theta)$:

$$\hat{\theta} = \arg\max_{\theta \in \Omega} \ell(\theta)$$

- There are different numerical approaches for optimizing the multiparameter log-likelihood function $\ell(\theta)$ for $\theta \in \Omega$

- Many likelihood functions have a unique maximum at $\hat{\theta}$, which is a stationary point satisfying $\partial\ell(\theta)/\partial\theta = \mathbf{0}$

- Numerical approaches involve a starting point θ_0 and an iterative procedure designed to give a sequence of points $\theta_1, \theta_2, \ldots$ converging to $\hat{\theta}$ (i.e. when $\theta_j \simeq \theta_{j-1}$)

Optimization Methods for Maximum Likelihood

- Three types of optimization methods are available in statistical software

  1. Methods that do not use derivatives (e.g., simplex algorithm, such as Nelder-Mead method)

  2. Methods that use only first derivative $U(\theta) = \partial\ell(\theta)/\partial\theta$ (e.g. steepest ascent, quasi-Newton, and conjugate gradient method)

  3. Methods that use both first derivative $U(\theta)$ and second derivative $H(\theta) = \partial^2\ell(\theta)/\partial\theta\,\partial\theta'$ (e.g. Newton-Raphson method)

     $H(\theta)$ is known as **Hessian matrix** and observed information matrix $I(\hat{\theta}) = -H(\hat{\theta})$

Optimization Methods for Maximum Likelihood

- **Newton-Raphson method** is commonly used for optimizing $\ell(\theta)$, which is based on the iteration scheme for the jth step $(j = 1, 2, ...)$ as

$$\theta_j = \theta_{j-1} - \left[H(\theta_{j-1})\right]^{-1} U(\theta_{j-1})$$

  ▶ $\theta_{j-1} \rightarrow$ value of θ at the $(j-1)th$ iteration

---

- Using Taylor series expansion, expanding $U(\theta)$ at θ_j

$$U(\theta) = U(\theta_j) + H(\theta_j)(\theta - \theta_j)$$

  ▶ Then

$$U(\theta) = \mathbf{0} \Rightarrow \theta = \theta_j - \left[H(\theta_j)\right]^{-1} U(\theta_j)$$

Optimization Methods for Maximum Likelihood

- In many situations, finding the MLE may be challenging, e.g. it may be on the boundary of Ω

- Likelihood function may also possess multiple stationary points, and optimization techniques are designed to obtain local maxima, so it may not converge to global maxima

- It is important to understand the shape of $\ell(\theta)$ before applying an optimization method

Example 2.1.1

- Suppose lifetimes $t_1, \dots, t_n$ are observed from a population with density $f(t)$ and distribution $F(t)$.

- Then the likelihood is (from Eq. 2.1.1)

$$Pr(\text{Data}) = \prod_{i=1}^{n} f(t_i) \qquad (2.1.2)$$

Parametric Approach of Inference

- Assume that $f(t)$ has a specific parametric form $f(t; \theta)$

- Likelihood function

$$L(\theta) = \prod_{i=1}^{n} f(t_i; \theta)$$

- By maximizing $L(\theta)$ or $\ell(\theta)$, we obtain $\hat{\theta}$ and consequently an estimate of the cumulative distribution function is $F(t; \hat{\theta})$.

- For example, if $T \sim \mathsf{Exp}(\theta)$ then $\hat{\theta} = \bar{t}$ and $F(t; \hat{\theta}) = 1 - \exp(-t/\hat{\theta})$

Nonparametric Approach of Inference

- Suppose $F(t)$ is discrete with jump probabilities
  $f(t) = F(t) - F(t-1)$ at $t = 1, 2, ...$

- The model parameters are $\mathbf{f} = (f(1), f(2), ...)$ and the likelihood function is

$$L = \prod_{i=1}^{n} f(t_i)$$

  ▶ Restrictions: $f(t) \geq 0 \ \forall t$ and $\sum_t f(t) = 1$

Nonparametric Approach of Inference

- The MLEs obtained by maximizing the corresponding likelihood function are:
$$\hat{f}(t) = \frac{1}{n} \sum_{i=1}^{n} I(t_i = t)$$

  ▶ $I(A)$ is an indicator function

$$I(A) = \begin{cases} 1 & \text{if } A \text{ is true} \\ 0 & \text{otherwise} \end{cases}$$

Nonparametric Approach of Inference

- Estimate of the cumulative distribution function $F(t)$:

$$\hat{F}(t) = \sum_{t_i \leq t} \hat{f}(t_i)$$

$$= \frac{1}{n} \sum_{i=1}^{n} I(t_i \leq t)$$

  ▸ $\hat{F}(t) \rightarrow$ **empirical distribution function**

Likelihood for a Truncated Sample

- Suppose that $t_1, \ldots, t_n$ are not from an unrestricted random sample of individuals, but a random sample of those with lifetimes one year or less

  - No information is available for those whose lifetimes are greater than one year

- The likelihood function for this truncated sample is given by

$$\prod_{i=1}^{n} Pr(t_i \,|\, T_i \leq 1) = \prod_{i=1}^{n} \frac{f(t_i)}{F(1)}$$

rather than (2.1.2)

Likelihood Based Inferences

Score Test

- Score vector $U(\theta)$ asymptotically follows k-variate normal distribution with mean vector $\mathbf{0}$ and variance-covariance matrix $\mathcal{I}(\theta)$
- Under the null hypothesis $H_0 : \theta = \theta_0$, the test statistic

$$W(\theta_0) = U(\theta_0)' \, \mathcal{I}(\theta_0)^{-1} \, U(\theta_0)$$

  asymptotically follows $\chi^2_{(k)}$ distribution.
- The statistic $W(\theta_0)$ can also be used to obtain confidence intervals for θ

Likelihood Based Inferences

Wald Test

- The MLE $\hat{\theta}$ follows a k-dimensional normal distribution with mean θ and variance-covariance matrix $\mathcal{I}(\theta)^{-1}$
- In other words, $\sqrt{n}(\hat{\theta} - \theta)$ follows a k-dimensional normal distribution with mean $\mathbf{0}$ and variance-covariance matrix $n\,\mathcal{I}(\theta)^{-1}$
- Under $H_0 : \theta = \theta_0$,

$$(\hat{\theta} - \theta_0)' \mathcal{I}(\theta_0)(\hat{\theta} - \theta_0)$$

asymptotically follows $\chi^2_{(k)}$.
- Since $I(\hat{\theta})/n$ is a consistent estimator of $\mathcal{I}(\theta_0)$, we can replace $\mathcal{I}(\theta_0)$ by $\mathbf{I}(\hat{\theta})$ in the test statistic.

Likelihood Based Inferences

Likelihood Ratio Test (LRT)

- Under $H_0 : \theta = \theta_0$,

$$\Lambda(\theta_0) = -2 \log \left[\frac{L(\theta_0)}{L(\hat{\theta})} \right]$$

$$= 2\ell(\hat{\theta}) - 2\ell(\theta_0)$$

asymptotically follows $\chi^2_{(k)}$

Subsection 2

2.2 Right Censoring and Maximum Likelihood

2.2 Right Censoring and Maximum Likelihood

- For *right censored data*, only the lower bounds on lifetime are available for some individuals

- Right censored lifetimes are observed for various reasons, such as termination of the study, lost-to-follow-up, etc.

- Contribution to the likelihood function would be different for right censored and complete lifetimes

- Construction of the likelihood function could differ for different types of censoring, such as left censoring, interval censoring, etc.

2.2 Right Censoring and Maximum Likelihood

- Let the random variables $T_1, \ldots, T_n$ represent the lifetimes of n individuals
  - Let $C_1, \ldots, C_n$ be the corresponding right censoring times (random variables)
- For the ith individual, we observe $\{(t_i, \delta_i), i = 1, \ldots, n\}$
  - t_i is a sample realization of $X_i = \min\{T_i, C_i\}$
  - $\delta_i = I(T_i \leq C_i)$ is known as censoring or **status indicator**, i.e.

$$\delta_i = \begin{cases} 1 & \text{if } t_i \text{ is observed failure time} \\ 0 & \text{if } t_i \text{ is observed censoring time} \end{cases}$$

2.2 Right Censoring and Maximum Likelihood

- There are three types of right censoring mechanism
  - Type I censoring (fixed end time)
  - Independent random censoring
  - Type II censoring (stop after r failures)

Type I Censoring

- In Type I censoring, potential censoring time C_i is assumed to be **fixed** for each individual

- Type I censoring often arises when a study is conducted over a specified period of time

  - For example, if termination of life test on electrical insulation specimens after 180 minutes, then $C_i = 180 \ \forall \, i$

Type I Censoring

- Likelihood function for the observed Type I censored sample $\{(t_i, \delta_i), i = 1, \dots, n\}$

$$L = \prod_{i=1}^{n} Pr(X_i = t_i, \delta_i)$$

Type I Censoring

- Likelihood function for the observed Type I censored sample $\{(t_i, \delta_i), i = 1, ..., n\}$

$$L = \prod_{i=1}^{n} Pr(X_i = t_i, \delta_i)$$

$$L = \prod_{i=1}^{n} \left[Pr(X_i = t_i, \delta_i = 1) \right]^{\delta_i} \left[Pr(X_i = t_i, \delta_i = 0) \right]^{1-\delta_i}$$

Type I Censoring

- Likelihood function for the observed Type I censored sample $\{(t_i, \delta_i), i = 1, ..., n\}$

$$L = \prod_{i=1}^{n} \left[Pr(X_i = t_i, \delta_i = 1) \right]^{\delta_i} \left[Pr(X_i = t_i, \delta_i = 0) \right]^{1-\delta_i}$$

- We can show that

$$
\begin{aligned}
&Pr(X_i = t_i, \delta_i = 1) \\
&= Pr(\min\{T_i, C_i\} = t_i, \delta_i = 1) \\
&= Pr(T_i = t_i) \\
&= f(t_i)
\end{aligned}
\qquad
\begin{aligned}
&Pr(X_i = t_i, \delta_i = 0) \\
&= Pr(\min\{T_i, C_i\} = t_i, \delta_i = 0) \\
&= Pr(T_i > C_i, C_i = t_i) \\
&= Pr(T_i > t_i) \\
&= S(t_i+)
\end{aligned}
$$

Type I Censoring

- Likelihood function for the observed Type I censored sample $\{(t_i, \delta_i), i = 1, ..., n\}$

$$L = \prod_{i=1}^{n} Pr(X_i = t_i, \delta_i)$$
$$= \prod_{i=1}^{n} \left[Pr(X_i = t_i, \delta_i = 1) \right]^{\delta_i} \left[Pr(X_i = t_i, \delta_i = 0) \right]^{1-\delta_i}$$
$$= \prod_{i=1}^{n} \left[f(t_i) \right]^{\delta_i} \left[S(t_i+) \right]^{1-\delta_i}$$

▶ If $S(t)$ is continuous at t_i, then $S(t_i+) = S(t_i)$

Example 2.2.1

- Suppose that lifetimes T_i are independent and follow an exponential distribution with the pdf $f(t) = \lambda\, e^{-\lambda t}$.

- Let $\{(t_i, \delta_i), i = 1, \ldots, n\}$ be a random sample (right censored, Type I) from the exponential distribution.

- Obtain the expression of the likelihood function for the given sample.

Example 2.2.1

- Given

$$f(t) = \lambda\, e^{-\lambda t} \;\; \Rightarrow \;\; S(t) = e^{-\lambda t}$$

- The likelihood function

$$
\begin{aligned}
L &= \prod_{i=1}^{n} \Big[f(t_i) \Big]^{\delta_i} \Big[S(t_i) \Big]^{1-\delta_i} \\
&= \prod_{i=1}^{n} \Big[\lambda e^{-\lambda t_i} \Big]^{\delta_i} \Big[e^{-\lambda t_i} \Big]^{1-\delta_i} \\
&= \Big[\lambda^{\sum_{i=1}^{n} \delta_i} \Big] \Big[e^{-\lambda \sum_{i=1}^{n} t_i} \Big] \\
&= \Big[\lambda^{r} \Big] \Big[e^{-\lambda \sum_{i=1}^{n} t_i} \Big]
\end{aligned}
$$

  ▶ $r = \sum_{i=1}^{n} \delta_i$

Independent Random Censoring

- Censoring time C is assumed to be continuous random variable with survivor function $G(t)$ and density function $g(t)$

- Lifetime T is also continuous random variable with survivor function $S(t)$ and density function $f(t)$

- Assumptions:

    - T and C are independent

    - $G(t)$ does not depend on any of the parameters of $S(t)$

Independent Random Censoring

- Likelihood function for the observed independent random censored sample $\{(t_i, \delta_i), i = 1, \dots, n\}$

$$L = \prod_{i=1}^{n} Pr(X_i = t_i, \delta_i)$$
$$= \prod_{i=1}^{n} \left[Pr(X_i = t_i, \delta_i = 1)\right]^{\delta_i} \left[Pr(X_i = t_i, \delta_i = 0)\right]^{1-\delta_i}$$

Independent Random Censoring

- We can show

$$
\begin{aligned}
Pr(X_i = t_i, \delta_i = 1) &= Pr(T_i < C_i, T_i = t_i) \\
&= Pr(T_i < C_i \,|\, T_i = t_i) \, Pr(T_i = t_i) \\
&= Pr(C_i > t_i) \, Pr(T_i = t_i) \\
&= G(t_i+) \, f(t_i)
\end{aligned}
$$

Independent Random Censoring

- We can show

$$
\begin{aligned}
Pr(X_i = t_i, \delta_i = 0) &= Pr(T_i > C_i, C_i = t_i) \\
&= Pr(T_i > C_i \,|\, C_i = t_i)\, Pr(C_i = t_i) \\
&= Pr(T_i > t_i)\, Pr(C_i = t_i) \\
&= S(t_i+)\, g(t_i)
\end{aligned}
$$

Independent Random Censoring

- Likelihood function for the observed independent random censored sample $\{(t_i, \delta_i), i = 1, \ldots, n\}$

$$
\begin{aligned}
L &= \prod_{i=1}^{n} Pr(X_i = t_i, \delta_i) \\
&= \prod_{i=1}^{n} \left[Pr(X_i = t_i, \delta_i = 1) \right]^{\delta_i} \left[Pr(X_i = t_i, \delta_i = 0) \right]^{1-\delta_i} \\
&= \prod_{i=1}^{n} \left[G(t_i+) \, f(t_i) \right]^{\delta_i} \left[S(t_i+) \, g(t_i) \right]^{1-\delta_i}
\end{aligned}
$$

- Since $G(t)$ and $g(t)$ don't involve any parameters of $f(t)$

$$
L = \prod_{i=1}^{n} \left[f(t_i) \right]^{\delta_i} \left[S(t_i+) \right]^{1-\delta_i}
$$

Type II censoring

- In Type II censoring, lifetest starts with n units and it stops when r number of failures are observed

  ► So r smallest lifetimes $t_{(1)} \leq \cdots \leq t_{(r)}$ in a random sample of n are observed

  ► $r \rightarrow$ a specified integer that lies between 1 and n

  ► The remaining $(n-r)$ units are considered as censored at the time $t_{(r)}$

Type II censoring

- For Type II censoring, the likelihood function is the probability of observing r smallest lifetimes $t_{(1)} \leq \cdots \leq t_{(r)}$ out of n lifetimes

$$L = \binom{n}{r} \left\{ \prod_{i=1}^{r} f(t_{(i)}) \right\} \left[S(t_{(r)}) \right]^{n-r}$$

- This expression is similar to the expression obtained for Type I and random independent censoring with all the censoring times equal to $t_{(r)}$

Example 2.2.2

- Let $t_{(i)}, \dots, t_{(r)}$ be a Type II random sample of n lifetimes $T_1, \dots, T_n$, where T_i follows an exponential distribution with rate λ

- The likelihood function

$$
\begin{aligned}
L &= \binom{n}{r} \left\{ \prod_{i=1}^{r} f(t_{(i)}) \right\} \left[S(t_{(r)}) \right]^{n-r} \\
&= \binom{n}{r} \left\{ \prod_{i=1}^{r} \lambda e^{-\lambda t_{(i)}} \right\} \left[e^{-\lambda t_{(r)}} \right]^{n-r} \\
&= \binom{n}{r} \lambda^r e^{-\lambda W}
\end{aligned}
$$

  ► $W = \sum_{i=1}^{r} t_{(i)} + (n-r) t_{(r)}$

A general formulation of right censoring

- The censoring process is often not any of the types discussed so far, and may be sufficiently complicated to make modeling it impossible.

- For example, a decision to terminate a life test or clinical trial at time t, or to withdraw certain individuals, might be based on failure information prior to time t,

- Fortunately it can be shown that under rather general conditions the observed likelihood is of the form (2.2.3) and can be used in the normal way to make inferences about the lifetime distribution under study.

$$L = \prod_{i=1}^{n} \left[f(t_i) \right]^{\delta_i} \left[S(t_i+) \right]^{1-\delta_i}$$

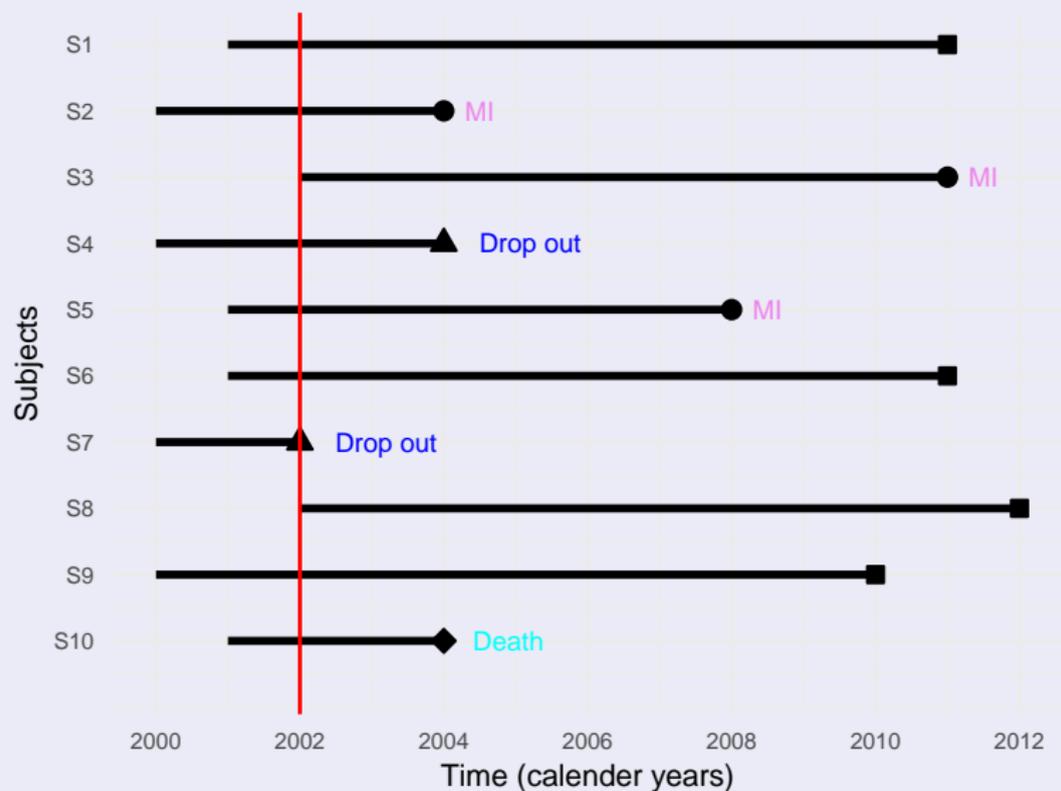- Read Section 2.2.2 of the textbook for details.

Subsection 3

A Hypothetical Study

A hypothetical study

- A small prospective study was run, where 10 participants were recruited to follow

- The event of interest was **the development of myocardial infarction** (MI, or heart attack) over a period of 10 years (follow-up period)

- Participants are recruited into the study over a period of two years and were then followed for up to 10 years

A hypothetical study

A hypothetical study
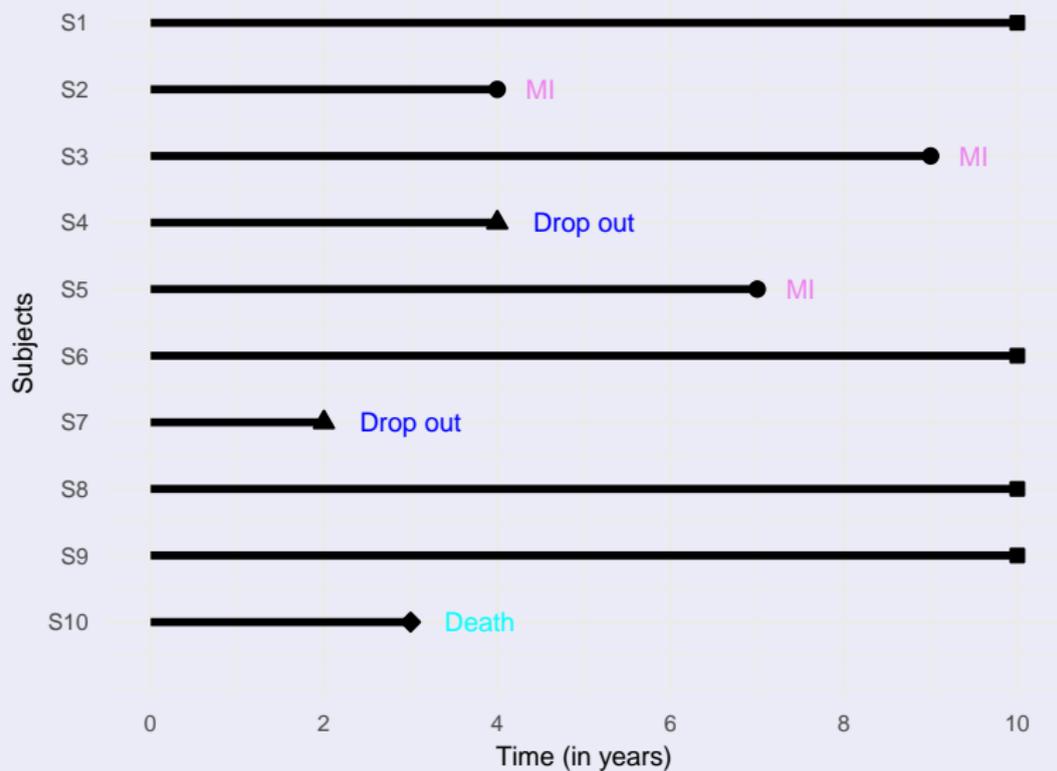
- $\{S2, S3, S5\} \rightarrow$ experienced MI

- $\{S4, S7\} \rightarrow$ dropped out

- $\{S10\} \rightarrow$ died from other causes

- $\{S1, S6, S8, S9\} \rightarrow$ completed 10-year follow-up without MI

A hypothetical study

Study in years

A hypothetical study

subject	t	outcome	status	δ
S1	10	alive	censored	0
S2	4	MI	failure	1
S3	9	MI	failure	1
S4	4	drop-out	censored	0
S5	7	MI	failure	1
S6	10	alive	censored	0
S7	2	drop-out	censored	0
S8	10	alive	censored	0
S9	10	alive	censored	0
S10	3	death	censored	0

A hypothetical study

- Times for the subjects who did not experience MI by the end of 10-year follow-up or dropped-out or died from causes not related to MI are known as **censored times**

- Time-to-MI (time to the event of interest) are known as **failure time** (or survival time)

- For survival data, **the pair (time, status)**, (t, δ), is considered as the response

- A sample of survival data can also be expressed as following, where $+$ sign indicates censored observations

$$10^+, 4, 9, 4^+, 7, 10^+, 2^+, 10^+, 10^+, 3^+$$

Subsection 4

2.3 Other Type of Incomplete Data

2.3 Other Type of Incomplete Data

Incomplete data in survival analysis is not limited to **right censoring**.

Depending on how and when information is missing, several other types of incomplete data can occur.

Understanding them is essential because the likelihood formulation and estimation procedures may differ.

Left Censoring

- An observation is *left censored* if the event of interest has already occurred **before the subject enters the study**, but the exact event time is unknown.

- In this case, we only know that the lifetime $T \leq t_i$, where t_i is the observed entry time.

- Contribution to likelihood:

$$Pr(T \leq t_i) = F(t_i)$$

where $F(t)$ is the cumulative distribution function.

Example: In a study of HIV infection, if a subject is already infected when first tested, the actual time of infection is left censored.

Interval Censoring

- **Definition**: The event of interest is only known to occur **between two observation times**.

- That is, for subject i, we know $L_i < T_i \leq R_i$ (with $0 < L_i < R_i < \infty$).

- Contribution to likelihood:

$$Pr(L_i < T_i \leq R_i) = S(L_i) - S(R_i)$$

Example: In medical screening studies, cancer may be detected at the second screening visit, but the exact time of onset is somewhere between the first and second visits.

Truncation

- Truncation occurs when individuals are included in the sample **only if their lifetimes satisfy certain conditions**. Unlike censoring, truncated individuals are never observed at all.

- **Types**:
  - **Left truncation (delayed entry)**: Only individuals with $T \geq t_0$ are observed.
  - **Right truncation**: Only individuals with $T \leq t_0$ are observed.

- Contribution to likelihood must be **conditional** on the truncation event.

Example:

- Left truncation: In a cohort study of heart disease, only people who survive long enough to be enrolled at age 50 are included.
- Right truncation: In a study of human longevity, only individuals who died before the study ended are observed.

Summary

Type	Observed information	Likelihood contribution L_i
Direct observation	$T = t_i$	$f(t_i)$
Right censoring	$T > t_i$	$S(t_i)$
Left censoring	$T < t_i$	$F(t_i)$
Interval censoring	$l_i < T \leq r_i$	$F(r_i) - F(l_i)$
Left truncation	$T = t_i \mid T > u$	$\dfrac{f(t_i)}{S(u)}$
Right truncation	$T = t_i \mid T < v$	$\dfrac{f(t_i)}{F(v)}$

Acknowledgements

This lecture is adapted from materials created by Mahbub Latif