Part III: Survival Analysis

Outline

- Basic concepts (time-to-event, censoring, hazard)
- Kaplan-Meier curve and survival function
- Cox proportional hazards model
- Application examples

Survival Analysis

- Survival analysis is a method for analyzing survival data or failure (death) time data, that is *time-to-event* data.
- It arises in a number of applied fields, such as medicine, biology, public health, epidemiology, engineering, economics, and demography.
- The time-to-event (or failure time) variable T is a non-negative random variable.
- T is the time from a well-defined time origin to a failure event.

Examples

- Times to death of patients with certain disease
- Remission duration of certain disease in clinical trials
- Incubation times of certain disease, such as AIDS
- Failure times of certain manufactured products
- Life times of elderly in particular social programs
- Patience time of call center customers
- ...

Incomplete Observations

Times-to-event are not always completely observable. These times are subject to censoring and truncation. For a censored or a truncated time-to-event, only partial information is available.

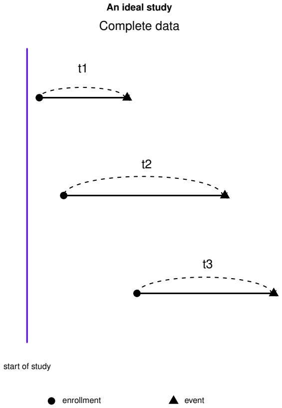
- Censoring: When an observation is incomplete due to some random cause.
- Truncation: When the incomplete nature of the observation is due to a systematic selection process inherent to the study design.

Censoring

- Right censoring: some individuals do not fail or lost-to-follow-up during the observed period; instead of knowing the failure time T, all we know about these individuals is that their time-to-event exceeds some observed value Y (type I, type II, random, etc).
- Left censoring: we only know the event happens before an observed time
  e.g., time to first use of marijuana: used it but forgot when
- Interval censoring: when time-to-event is only known to fall within an interval
  e.g., in clinical trials where patients have periodic follow-ups

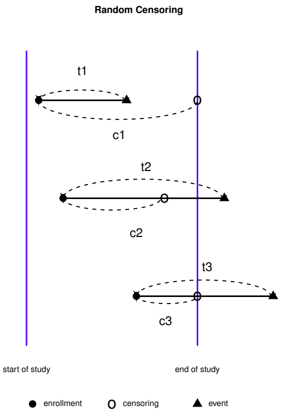
Right Censoring

- events occur after the end of study
- subjects drop out of study
- subjects are lost to follow-up during the study period



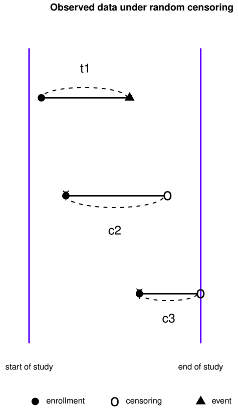
Right Censoring

- events occur after the end of study
- subjects drop out of study
- subjects are lost to follow-up during the study period



Right Censoring

- events occur after the end of study
- subjects drop out of study
- subjects are lost to follow-up during the study period



Example: A set of observed survival data is

i	y_i	δ_i
1	25	1
2	18	0
3	17	1
4	22	0
5	27	1

where y_i is the observed time, and δ_i is the indicator of event. The data can also be presented as

$$25 \quad 18^{+} \quad 17 \quad 22^{+} \quad 27$$

Why Survival Analysis

- A special course of difficulty in the analysis of survival data is the possibility that some individuals may not be observed for the full time to failure.
- The goal of survival analysis is to make inferences about the underlying survival time random variable T based on the observed, incomplete data (y_1, δ_1), $(y_2, \delta_2), \ldots, (y_n, \delta_n)$.
- Special methods are needed to characterize the distribution of the time-to-event variable, and its association with other factors.

In survival analysis, the time origin and end event must be clearly defined based on the research question of interest. For example,

- If we want to study the disease-specific survival (DSS) after a surgical procedure, the time origin is the surgery completion time, and the end event is disease-specific death.

Basic Functions and Quantities

Let T denote the time-to-event random variable. Assume T is continuous for now.

- Cumulative distribution function
- Survival function
- Hazard function
- Cumulative hazard function

Survival Function

Cumulative distribution function of T is defined as

$$F(t) = \mathbb{P}(T \leq t) = \int_0^t f(x)dx$$

Survival function is defined as

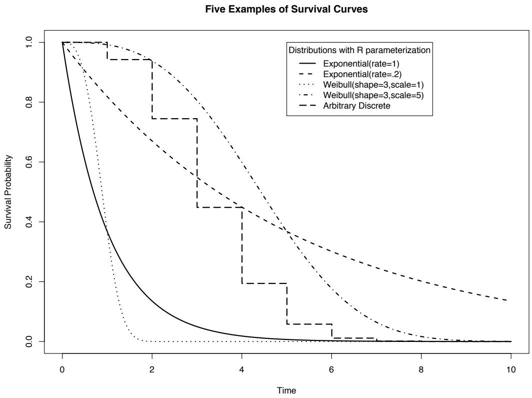
$$S(t) = \mathbb{P}(T > t) = 1 - F(t)$$

- $S(t) = 1$ if $t \leq 0$; $S(\infty) = 0$.
- $S(t)$ is continuous and decreasing (for continuous T).
- $S(t)$ provides useful summary information, such as median survival time ($S^{-1}(0.5)$), 5-year survival rate ($S(5)$), etc.

Parametric Survival Functions

Examples of parametric distribution families for survival analysis:

- Exponential distribution:
  - $f(x) = \lambda e^{-\lambda x}$
  - $S(x) = e^{-\lambda x}$
- Weibull distribution:
  - $f(x) = \lambda \alpha x^{\alpha - 1} e^{-\lambda x^{\alpha}}$
  - $S(x) = e^{-\lambda x^{\alpha}}$
- Log-normal, Gamma, Pareto, etc


Five Examples of Survival Curves
Distributions with R parameterization
Exponential(rate=1)
Exponential(rate=.2)
Weibull(shape=3,scale=1)
Weibull(shape=3,scale=5)
Arbitrary Discrete
Survival Probability
0.0
0.2
0.4
0.6
0.8
1.0
0
2
4
6
8
10
Time

Hazard Rate Function

Hazard rate captures the instantaneous failure rate at time t, given survival up to time t. Hazard rate function is defined as

$$h(t) = \lim_{\Delta t \to 0} \frac{\mathbb{P}(t \leq T < t + \Delta t | T \geq t)}{\Delta t}$$

Cumulative hazard function is defined as

$$H(t) = \int_0^t h(x) dx$$

Characteristics of $h(t)$:

- $0 \leq h(t) < \infty$
- $h(t) \cdot \Delta t$ is approximately the proportion of individuals experiencing failure in $[t, t + \Delta t)$ among those surviving up to t.

By definition, $f(t) = \lim_{\Delta t \to 0} \mathbb{P}(t \leq T < t + \Delta t)/\Delta t$. Thus, we have

$$h(t) = \frac{f(t)}{S(t)} = -\frac{\partial \log(S(t))}{\partial t}$$

and

$$S(t) = \exp\{-H(t)\}$$

- This is a well know relation among the density, hazard and survival functions.
- **The distribution of T can be fully defined by a hazard function!**

Parametric Hazard Functions

- Exponential distribution: $h(x) = \lambda$ (constant!)
- Weibull distribution: $h(x) = \lambda\alpha x^{\alpha-1}$