Chapter 1. A simple introduction to survival analysis

1.1 Introduction

Survival analysis is a method for analyzing survival data or failure (death) time data, that is time-to-event data, which arises in a number of applied fields, such as medicine, biology, public health, epidemiology, engineering, economics, and demography. The interested random variable (r.v.) T is non-negative. In general, T will be the time from some origin event till occurrence of some other event.

$$\boxed{0 \text{ (for example, birth)}} \longrightarrow \boxed{1 \text{ (for example, death)}}$$

The following are four examples of survival analysis problems.

(i) Clinical trial:

Remission Duration of Leukaemia patients from a clinical trial. The failure T is defined to be the time from remission to relapse. Patients were randomized to either 6MP or placebo. The study faced to a problem that 12 patients did not returned (relapse) at the end of the study. Thus we do not know the exact failure time T of these 12 patients. There are two reasons that we do not want to throw away these 12 incomplete observations. First, it will loss efficiency. Second, it will cause bias in analysis.

(ii) Time of first use of Marijuana:

In the study, 191 high school boys were asked "when did you first use marijuana?". We can know the exact time if the boy can remember and tell. However, even if we assume all the boys will tell the truth, we may still have two kinds of incomplete answers. First, "I never use it". Second, "I have used it but cannot recall just when the first time I used."

(iii) Basic time variables: Duration from infection to AIDS.

For the pediatric population, the infecting time is assumed at age 0 (birth). The sampling scheme only individuals who have developed AIDS prior to the end of the study period are included in the study. Therefore, infected individuals who have yet to develop AIDS are not included in the sample.

(iv) Channing house data (retired faculty or staff of Stanford University):

The interested failure time is the age at death. The problem is that the individuals who die before retirement are not included in the sample. This such that the younger individuals are not included in the data.

It is important to note that in a survival analysis

(i) The time origin should be precisely defined for each individual. (for example, birth)

(ii) The end event (failure) must be clearly defined. (for example, death)

(iii) All individuals should be as comparable as possible at their time origin. (for example, the date of randomization)

1.2 Censoring mechanism through the discussion of right censoring

A special course of difficulty in the analysis of survival data is the possibility that some individual may not be observed for the full time to failure. In some circumstance, 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-failure exceeds some value y where y is the follow-up time of these individuals in the study.

The following are three different types of censoring

(i) Type I censoring: Observed count for a fixed period of time. At the end of the study, any subject that has not yet fail is censored.

(ii) Type II censoring: Similar to Type I censoring, but instead of fixing length of the study, the total number of failure is fixed in advance.

(iii) Random censoring: The total period of observation is fixed, but subjects enter the study at different time points. Some individuals fail, some individual lost-to-follow-up, some individual still alive at the end of the study.

In random censoring, we assume that each individual has his/her own failure time T and censoring time C, however, we can only observe the random vector (Y, δ) where

$$\begin{aligned} Y &= \min(T, C) = T \wedge C, \\ \delta &= \begin{cases} 1, & \text{if } T = Y \\ 0, & \text{otherwise} \end{cases}. \end{aligned}$$

Usually, we assume $T \perp C$, that is T and C are independent to each other. This implies that the censoring time C is non-informative in analyzing the failure time T. To have this assumption, we have to make sure that individuals are not lost-to-follow-up for reasons related to failure time.

The following are some others possible incomplete mechanism in survival data.

(i) Left censoring: We observe (Y, δ) where $Y = \max(T, C)$ and $\delta = I(T = Y)$.

(ii) Interval censoring: We observe (L, R) where $L \leq T \leq R$.

(iii) Random Truncation: In random truncation, we assume that each individual has his/her own failure time T and truncation time K. Usually, we assume $T \perp K$.

(iiia) Left truncation: We observe (T^*, K^*) where $T \geq K, T^* = T, K^* = K$.

(iiib) Right truncation: We observe (T^*, K^*) where $T \leq K, T^* = T, K^* = K$.

(iiic) Interval (Doubly) truncation: We observe (T^*, K_1^*, K_2^*) where $K_1 \leq T \leq K_2, T^* = T, K_1^* = K_1, K_2^* = K_2$.