Survival Analysis

Chao Cheng

November 1, 2022

Contents

1 Basic knowledge

1.1 Survival and hazard

Let T denote the time to an event that we are interested in. Then we know the c.d.f.

$$F_T(t) = P(T \leq t),$$

and the corresponding p.d.f.

$$f_T(t) = \frac{\mathrm{d}}{\mathrm{d}t} F_T(t).$$

Here to simplify the discussion, we assume T is a continuous random variable. In the context of survival analysis, the *event* often refers to death. Then T represents the lifespan of the subject. So $F_T(t)$ represents the probability that the death occurs before t. In another word, we know the probability that the subject survives passes t is

$$S_T(t) = 1 - F_T(t) = P(T > t).$$

$S_T(t)$ is often called the survival function? and clearly

$$f_T(t) = -\frac{\mathrm{d}}{\mathrm{d}t} S_T(t).$$

The **hazard function** $h(t)$ is defined as

$$h(t) = \lim_{\Delta \to 0} \frac{P(T \leq t + \Delta | T > t)}{\Delta} = \lim_{\Delta \to 0} \frac{F_T(t+\Delta) - F_T(t)}{\Delta \cdot S_T(t)} = \frac{f_T(t)}{S_T(t)}.$$

$h(t)$ represents the instant hazard? unified probability? that the subject will be dead instantly after t given the fact that it's alive at t. And the **cummulative hazard function** is

$$H(t) = \int_0^t h(x)\,\mathrm{d}x = \int_0^t \frac{f_T(x)}{S_T(x)}\mathrm{d}x = \int_0^t \frac{-\mathrm{d}S_T(x)}{S_X(t)} = -\log(S_T(x))|_0^t = -\log(S_T(t)).$$

Proposition 1. *The random variable $H(T)$ follows unit exponential distribution $EXP(1)$.*

Proof.

$$\begin{aligned}P(H(T) \leq t) =& P(-\log S(T) \leq t) \\ =& P\left(1 - F(T) \geq e^{-t}\right) \\ =& P\left(T \leq F^{-1}\left(1 - e^{-t}\right)\right) \\ =& F\left(F^{-1}\left(1 - e^{-t}\right)\right) \\ =& 1 - e^{-t},\end{aligned}$$

which is the c.d.f of $EXP(1)$. Here to simplify the deduction we make some assumptions that

- $F(t)$ is continuous.
- $F^{-1}(t)$ is well defined.

Also to simplify the notation and avoid confusion, we use $S(\cdot)$ and $F(\cdot)$ instead of $S_T(\cdot)$ and $F_T(\cdot)$ like before. □

1. **Exponential distribution:** Denote $T \sim EXP(\lambda)$. Then

$$\begin{aligned}&f(t) = \lambda e^{-\lambda t} \\ &F(t) = 1 - e^{-\lambda t} \qquad S(t) = e^{-\lambda t} \\ &h(t) = \lambda \qquad \text{constant hazard} \\ &H(t) = \lambda t \\ &\mathrm{E}(T) = 1/\lambda \qquad \mathrm{Var}(T) = 1/\lambda^2\end{aligned}$$

2. **Weibull distribution:** Denote $T \sim W(p, \lambda)$. Then

$$\begin{aligned}&f(t) = p\lambda^p t^{p-1} \exp\left(-(\lambda t)^p\right) \\ &F(t) = 1 - \exp\left(-(\lambda t)^p\right) \qquad S(t) = \exp\left(-(\lambda t)^p\right) \\ &h(t) = p\lambda^p t^{p-1} \\ &H(t) = (\lambda t)^p \\ &\mathrm{E}(T) = \frac{1}{\lambda} \cdot \Gamma\left(1 + \frac{1}{p}\right) \qquad \mathrm{Var}(T) = \frac{1}{\lambda^2}\left(\Gamma\left(1 + \frac{2}{p}\right) - \Gamma\left(1 + \frac{1}{p}\right)\right) \\ &\mathrm{E}(T^m) = \frac{1}{\lambda^m}\Gamma\left(1 + \frac{m}{p}\right)\end{aligned}$$

1.2 Censor

1.2.1 Right censor

- Type I: an i.i.d sample $T_1, \cdots, T_n \sim F$ and a fixed constant c. And the observed data is (U_i, δ_i) for $i = 1, \cdots, n$ where

$$\begin{aligned}U_i &= \min(T_i, c) \\ \delta_i &= 1_{T_i \leq c}.\end{aligned}$$

  So the observed data consists of a random number, r, of uncensored observations, all of which are less than c. And $n - r$ censored observations, all are c.

- Type II: an i.i.d sample $T_1, \cdots, T_n \sim F$ and a pre-defined number of failure r. The observation is stopped when r failure occures and the stopping time is c. The observed data is still the form (U_i, δ_i) for $i = 1, \cdots, n$, the same as that in Type I censor. But in actuality, we observe the first r order statistics

$$T_{(1,n)}, \cdots, T_{(r,n)}.$$

Note that here $(U_1, \delta_1), \cdots, (U_n, \delta_n)$ are dependent whereas they are independent for Type I.

- Type III (Random censor): The underlying data is

$$\begin{aligned} &c_1, \cdots, c_n \quad \text{constant} \\ &T_1, \cdots, T_n \sim F. \end{aligned}$$

And the observed data is (U_i, δ_i) for $i = 1, \cdots, n$, where

$$\begin{aligned} U_i &= \min(T_i, c_i) \\ \delta_i &= 1_{T_i \le c_i}. \end{aligned}$$

Note: for inference, c_i is often treated as constant. For study design or studying the asymptotic property, they are often treated as i.i.d random variables $C_1, \cdots, C_n$.

1.2.2 Left censor

T_i is censored when $T_i \le l_i$.

1.2.3 Interval censor

$l_i \le T_i \le u_i$, but only l_i and u_i are observed.

References