Survival Analysis

Chao Cheng

November 16, 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.

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

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

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

$h(t)$ 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.

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

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

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

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

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

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

$$U_i = \min(T_i, c_i)$$
$$\delta_i = 1_{T_i \leq c_i}.$$

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 $T_i \leq l_i$.

1.2.3 Interval censor

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

2 MLE

There is an i.i.d survival time sample $T_1, \cdots, T_n$ with common and unknown c.d.f. $F(\cdot)$ and the observated data is (U_i, δ_i) for $i = 1, \cdots, n$, where

$$U_i = \min(T_i, C_i)$$
$$\delta_i = 1(T_i \leq C_i)$$

and C_i is the (fixed or random) censoring time. Let $\perp$ denote "is independent of". We assume $T_i \perp C_i$ (Non-informative censoring, the key assumption) and (U_i, δ_i) are also i.i.d. The observed data consists of two parts. U_i is continuous while δ_i is binary.

$$\begin{aligned} (U_i, \delta_i) &= (u_i, 1) \qquad T_i \text{ is uncensored at } u_i \\ (U_i, \delta_i) &= (u_i, 0) \qquad T_i \text{ is censored at } u_i \end{aligned}$$

When C_is are known constants, the likelihood for (U_i, δ_i) is

$$\begin{aligned} L_i(F) = & \begin{cases} f(u_i) & \text{if } \delta_i = 1 \\ 1 - F(u_i) & \text{if } \delta_i = 0 \end{cases} \\ = & f(u_i)^{\delta_i} (1 - F(u_i))^{1-\delta_i} \end{aligned}$$

Therefore

$$L(F) = \prod_{i=1}^{n} L_i(F) = \prod_{i=1}^{n} \left(f(u_i)^{\delta_i} (1 - F(u_i))^{1-\delta_i} \right) = \prod_{i=1}^{n} \left(h(u_i)^{\delta_i} S(u_i) \right). \tag{1}$$

The last equality relies on the fact that $f(t) = h(t) S(t)$.

When C_is are i.i.d. $\sim G$, where G is continuous with p.d.f g. Then we have

$$P(U_i \leq u, \delta_i = 1) = P(T_i \leq u, T_i \leq C_i) = \int_0^u \int_t^\infty f(t) g(c) \, \mathrm{d}c\mathrm{d}t = \int_0^u f(t)(1 - G(t)) \, \mathrm{d}t$$

Therefore the likelihood for $\delta_i = 1$ is

$$L_i(F, G) = f(u_i)(1 - G(u_i)) \qquad \text{when } \delta_i = 1.$$

And similarly, for $\delta_i = 0$, the likelihood is

$$L_i(F, G) = g(u_i)(1 - F(u_i)) \qquad \text{when } \delta_i = 0.$$

Hence the full likelihood is

$$\begin{aligned} L(F, G) &= \prod_{i=1}^{n} \left\{ (f(u_i)(1 - G(u_i)))^{\delta_i} ((1 - F(u_i)) g(u_i))^{1-\delta_i} \right\} \\ &= \prod_{i=1}^{n} \left\{ f(u_i)^{\delta_i} (1 - F(u_i))^{1-\delta_i} \right\} \cdot \prod_{i=1}^{n} \left\{ g(u_i)^{1-\delta_i} (1 - G(u_i))^{\delta_i} \right\} \end{aligned} \tag{2}$$

So the core to maximize $L(F, G)$ with respect to F in (2) is the same as that in (1).

2.1 Parametric MLE

2.1.1 One-sample setting

Suppose $T_1, \cdots, T_n$ are i.i.d. $Exp(\lambda)$, and subject to noninformative right censoring. Then (1) becomes

$$L = L(\lambda) = \prod_{i=1}^{m} \left\{ \left(\lambda e^{-\lambda u_i}\right)^{\delta_i} \left(e^{-\lambda u_i}\right)^{1-\delta_i} \right\} = \lambda^{\sum_{i=1}^{n} \delta_i} e^{-\lambda \sum_{i=1}^{n} u_i} = \lambda^r e^{-\lambda W},$$

where $r=\sum_{i=1}^{n}\delta_i$ is the number of observed events and $W=\sum_{i=1}^{n}u_i$ is the total of observed time. Therefore $\log L = r\log\lambda - \lambda W$ and the MLE for λ is

$$\hat{\lambda}=\frac{r}{W}.$$

Furthermore, we know that

$$\begin{cases} \dfrac{\partial \log L}{\partial \lambda} = \dfrac{r}{\lambda} - W \\ \dfrac{\partial^2 \log L}{\partial \lambda^2} = -\dfrac{r}{\lambda^2} \end{cases}.$$

Based on properties of fisher information (See the notes about fisher information for more details.), we know that at the true underlying value λ, it must satisfy

$$\begin{cases} \mathrm{E}\dfrac{\partial \log L}{\partial \lambda} = \dfrac{\mathrm{E}r}{\lambda} - \mathrm{E}W & = 0 \\ I\left(\lambda\right) = -\mathrm{E}\dfrac{\partial^2 \log L}{\partial \lambda^2} & = \dfrac{\mathrm{E}r}{\lambda^2} \\ I^{\star}\left(\lambda\right) = \dfrac{1}{n} I\left(\lambda\right) & = \dfrac{\mathrm{E}r}{n\lambda^2} \end{cases}. \tag{3}$$

Note that in (3), r and W are random variables. And the probability to observe an event is

$$p = P\left(\delta_i = 1\right) = P\left(U_i \leq \infty, \delta_i = 1\right) = \int_0^{\infty} f\left(t\right)\left(1 - G\left(t\right)\right)\mathrm{d}t.$$

Therefore $r \sim \mathit{binomial}\left(n, p\right)$, $\mathrm{E}r = np$. And from property of MLE, we can write

$$\frac{\sqrt{n}\left(\hat{\lambda}-\lambda\right)}{\sqrt{I^{\star}\left(\lambda\right)^{-1}}} = \frac{\left(\hat{\lambda}-\lambda\right)}{\sqrt{I\left(\lambda\right)^{-1}}} \xrightarrow{D} N\left(0,1\right),$$

which means approximately

$$\hat{\lambda} \overset{\text{apx}}{\sim} N\left(\lambda, I\left(\lambda\right)^{-1}\right) = N\left(\lambda, \frac{\lambda^2}{np}\right).$$

Unfortunately, both λ and p (essentially $G\left(\cdot\right)$) are unknown. We plug in the estimation $\hat{\lambda} = r/W$ and $\hat{p} = r/W$ and apply Slutsky's theorem. This means for the purpose of estimation, we use

$$\begin{cases} \hat{\lambda} = \dfrac{r}{W} \\ I\left(\hat{\lambda}\right) = \dfrac{r}{\hat{\lambda}^2}, \quad I^{\star}\hat{\left(\lambda\right)} = \dfrac{r}{n\hat{\lambda}^2} \end{cases} \tag{4}$$

Not that unlike (3), here in (4), r and W are observations. And we have

$$\hat{\lambda} \overset{\text{apx}}{\sim} N\left(\lambda, \frac{r}{W^2}\right). \tag{5}$$

Note that it turns out that a better approximation is to assume $\log\hat{\lambda}$ is normal. Using the delta method, this gives

$$\log\hat{\lambda} \overset{\text{apx}}{\sim} N\left(\log\lambda, \frac{1}{np}\right) \approx N\left(\log\lambda, \frac{1}{r}\right). \tag{6}$$

Now based on (5) or (6), we can construct CI on λ, which also means we can perform hypothesis testing about λ.

2.1.2 Two-sample setting

For two samples $x_1, \cdots, x_n$ and $y_1, \cdots, y_m$, both follow exponential distribution with parameters λ_1 and λ_2. Assume noninformative censoring in each group, using same tech in Section 2.1.1 we can get

$$Z = \frac{\log \hat{\lambda}_1 - \log \hat{\lambda}_2}{\sqrt{\frac{1}{r_1} + \frac{1}{r_2}}} \overset{\text{apx}}{\sim} N(0, 1).$$

2.2 Nonparametric MLE

References