Survival Analysis

Chao Cheng

November 28, 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

$$h(t) = \lim_{\Delta\to 0} \frac{P(T \le t+\Delta | T > t)}{\Delta} = \lim_{\Delta\to 0} \frac{F_T(t+\Delta) - F_T(t)}{\Delta \cdot S_T(t)} = \frac{f_T(t)}{S_T(t)}. \tag{1}$$

$h(t)$ 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

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

  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.

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

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

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

$$L(F) = \prod_{i=1}^{n} L_i(F) = \prod_{i=1}^{n} \left(f(u_i)^{\delta_i} (1 - F(u_i))^{1-\delta_i} \right) = \prod_{i=1}^{n} \left(h(u_i)^{\delta_i} S(u_i) \right). \tag{2}$$

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

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

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

$$\begin{aligned} L(F, G) &= \prod_{i=1}^{n} \left\{ (f(u_i)(1 - G(u_i)))^{\delta_i} ((1 - F(u_i)) g(u_i))^{1-\delta_i} \right\} \\ &= \prod_{i=1}^{n} \left\{ f(u_i)^{\delta_i} (1 - F(u_i))^{1-\delta_i} \right\} \cdot \prod_{i=1}^{n} \left\{ g(u_i)^{1-\delta_i} (1 - G(u_i))^{\delta_i} \right\} \end{aligned} \tag{3}$$

So the core to maximize $L(F, G)$ with respect to F in (3) is the same as that in (2).

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 (2) 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

$$\begin{cases} \mathrm{E}\dfrac{\partial \log L}{\partial \lambda} = \dfrac{\mathrm{E}r}{\lambda} - \mathrm{E}W & = 0 \\ I(\lambda) = -\mathrm{E}\dfrac{\partial^2 \log L}{\partial \lambda^2} & = \dfrac{\mathrm{E}r}{\lambda^2} \\ I^\star(\lambda) = \dfrac{1}{n} I(\lambda) & = \dfrac{\mathrm{E}r}{n\lambda^2} \end{cases}. \tag{4}$$

Note that in (4), r and W are random variables. And the probability to observe an event is

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

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

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

which means approximately

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

Unfortunately, both λ and p (essentially $G(\cdot)$) 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

$$\begin{cases} \hat{\lambda} = \dfrac{r}{W} \\ I(\hat{\lambda}) = \dfrac{r}{\hat{\lambda}^2}, \quad I^\star\hat{(\lambda)} = \dfrac{r}{n\hat{\lambda}^2} \end{cases} \tag{5}$$

Not that unlike (4), here in (5), r and W are observations. And we have

$$\hat{\lambda} \overset{\text{apx}}{\sim} N\left(\lambda, \frac{r}{W^2}\right). \tag{6}$$

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

$$\log\hat{\lambda} \overset{\text{apx}}{\sim} N\left(\log\lambda, \frac{1}{np}\right) \approx N\left(\log\lambda, \frac{1}{r}\right). \tag{7}$$

Now based on (6) or (7), 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

The NPMLE of survivor function $S(\cdot)$ based on i.i.d. survival time and non-informative right censoring is often known as Kaplan-Meier estimator or the Product-Limit Estimator. Here we provide some heuristic development, but formal proofs will be deferred to other notes. With the same notation as before, the observed data is

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

where T_is are i.i.d survival times and C_is are i.i.d non-informative censoring time. The full likelihood is already shown in (3).

2.2.1 Discrete time points

To begin with, let's assume $F(\cdot)$ takes discrete values with mass points at $\{v_i\}$s: $0 \le v_1 < v_2 < \cdots < \cdots$, and define the discrete hazard functions as

$$\begin{aligned} h_1 &= P(T = v_1) \\ h_j &= P(T = v_j | T > v_{j-1}) \qquad j > 1. \end{aligned} \tag{8}$$

Note that (8) can be seen as discrete version of (1). And for $t \in [v_j, v_{j+1})$,

$$\begin{aligned} S(t) &\overset{\text{def}}{=} P(T>t) = P(T>v_j) \\ &= P(T > v_j | T > v_{j-1}) P(T > v_{j-1}) \\ &= P(T > v_j | T > v_{j-1}) P(T > v_{j-1} | T > v_{j-2}) P(T > v_{j-2}) \\ &= \cdots \\ &= P(T > v_1) \prod_{i=1}^{j-1} P(T > v_{i+1} | T > v_i) \\ &= \prod_{i=1}^{j} (1 - h_i) \qquad j > 1. \end{aligned}$$

For discrete case, the p.m.f $f(\cdot)$ is

$$\begin{aligned} f(v_1) &= P(T = v_1) = h_1 \\ f(v_j) &= P(T = v_j) = P(T = v_j | T > v_{j-1}) P(T > v_{j-1}) = h_j \prod_{i=1}^{j-1} (1 - h_i). \end{aligned}$$

Then if we want to estimate $F(\cdot)$ from likelihood, either (2) or (3), we are just trying to maximizing

$$\begin{aligned}
L(F) &= \prod_{i=1}^{n}\left\{f(u_i)^{\delta_i}(1-F(u_i))^{1-\delta_i}\right\} \\
&= \prod_{\{u_i|\delta_i=1\}} f(u_i) \prod_{\{u_i|\delta_i=0\}} S(u_i).
\end{aligned}$$

Let $I(\cdot)$ be an index mapping function that returns the index in v_is that matches u_i, i.e. $I(u_i)=j$ if and only if $u_i \in [v_j, v_{j+1})$. Then we know that $u_i = v_{I(u_i)}$ and we can write

$$\begin{aligned}
L(F) &= \prod_{\{u_i|\delta_i=1\}} f\left(v_{I(u_i)}\right) \prod_{\{u_i|\delta_i=0\}} S\left(v_{I(u_i)}\right) \\
&= \left[\prod_{\{u_i|\delta_i=1,u_i=v_1\}} f(v_1)\right]\left[\prod_{\{u_i|\delta_i=1,u_i\neq v_1\}} f\left(v_{I(u_i)}\right)\right]\left[\prod_{\{u_i|\delta_i=0\}}\prod_{k=1}^{I(u_i)}(1-h_k)\right] \\
&= \left[\prod_{\{u_i|\delta_i=1,u_i=v_1\}} h_1\right]\left[\prod_{\{u_i|\delta_i=1,u_i\neq v_1\}}\left(h_{I(u_i)}\prod_{k=1}^{I(u_i)-1}(1-h_k)\right)\right]\left[\prod_{\{u_i|\delta_i=0\}}\prod_{k=1}^{I(u_i)}(1-h_k)\right] \\
&= \left[\prod_{\{u_i|\delta_i=1\}} h_{I(u_i)}\right]\left[\prod_{\{u_i|\delta_i=1,u_i\neq v_1\}}\prod_{k=1}^{I(u_i)-1}(1-h_k)\right]\left[\prod_{\{u_i|\delta_i=0\}}\prod_{k=1}^{I(u_i)}(1-h_k)\right].
\end{aligned} \tag{9}$$

Note that in (9), the first part is

$$\prod_{\{u_i|\delta_i=1\}} h_{I(u_i)} = \prod_{j=1}^{\infty} h_j^{d_j}, \tag{10}$$

where $d_j = \sum_{i=1}^{n}\delta_i \cdot 1(u_i = v_j)$ is the number of event at v_j. The second and third part in (9) is

$$\begin{aligned}
&\left[\prod_{\{u_i|\delta_i=1,u_i\neq v_1\}}\prod_{k=1}^{I(u_i)-1}(1-h_k)\right]\left[\prod_{\{u_i|\delta_i=0\}}\prod_{k=1}^{I(u_i)}(1-h_k)\right] \\
&= \left[\prod_{k=1}^{\infty}\prod_{\{i|\delta_i=1,I(u_i)-1\geq k\}}(1-h_k)\right]\left[\prod_{k=1}^{\infty}\prod_{\{i|\delta_i=0,I(u_i)\geq k\}}(1-h_k)\right] \\
&= \left[\prod_{k=1}^{\infty}(1-h_k)^{\sum_{i=1}^{n}\delta_i\cdot 1(I(u_i)-1\geq k)}\right]\left[\prod_{k=1}^{\infty}(1-h_k)^{\sum_{i=1}^{n}(1-\delta_i)\cdot 1(I(u_i)\geq k)}\right] \\
&= \left[\prod_{k=1}^{\infty}(1-h_k)^{\sum_{i=1}^{n}\delta_i\cdot[1(I(u_i)\geq k)-1(I(u_i)=k)]}\right]\left[\prod_{k=1}^{\infty}(1-h_k)^{\sum_{i=1}^{n}(1-\delta_i)\cdot 1(I(u_i)\geq k)}\right] \\
&= \prod_{k=1}^{\infty}(1-h_k)^{Y(v_k)-d_k},
\end{aligned} \tag{11}$$

where

$$Y(v_k) = \sum_{i=1}^{n} 1(I(u_i)\geq k) = \sum_{i=1}^{n} 1(u_i \geq v_k)$$

is the number of subjects that are "at risk" at time v_k. **Note:** by the word "at risk", we also count the subjects that died just at v_k, which means $Y(v_j) \geq d_j$.

Then from (10) and (11) we know that (9) can be written as

$$L(F) = \prod_{j=1}^{\infty} h_j^{d_j} (1 - h_j)^{Y(v_j) - d_j}. \tag{12}$$

And the NPMLE is just

$$\hat{h}_j = \frac{d_j}{Y(v_j)} \tag{13}$$

for $j = 1, \cdots, \infty$ and $Y(v_j) > 0$. (13) implies some properties of this discrete NPMLE:

1. This estimation makes sense: the probability of dying at v_j given the fact you live past v_{j-1} can be estimated by the proportion of subjects die at v_j over the number of "at risk" at v_j.

2. $\hat{h}_j$ is only defined at time points where $Y(v_j) > 0$. Therefore, for large enough v_j, there will be no observation, no matter event or censoring, resulting inability to make estimation about hazard at those time points.

3. For time points where $Y(v_j) > 0$ but no event occurs, the hazard is estimated to be 0.

This means

$$\hat{S}(t) = \begin{cases} 1 & t < v_1 \\ \prod_{j=1}^{k} \left(1 - \hat{h}_j\right) & v_k \leq t < v_{k+1} \end{cases}$$

Note: $S(\cdot)$ is defined to be right-continuous.

Let v_g denotes the largest time point with observation, which means $Y(v_g) > 0$ and $Y(v_{g+1}) = 0$. Then either $d_g = Y(v_g)$ or $d_g < Y(v_g)$. If $d_g = Y(v_g)$, then $\hat{h}_g = 1$ and $\hat{S}(t) = 0$ for $t \geq v_g$. But if $d_g < Y(v_g)$, then $\hat{S}(t) > 0$ for $v_g \leq t < v_{g+1}$ and $\hat{S}(t)$ is undefined on $t \in [v_{g+1}, \infty)$.

Here one might say that the KM estimator is undefined on $t \in [v_{g+1}, \infty)$. Or another explanation is that NPMLE is not unique and any survival function that is identical to $\hat{S}$ at previous time is the NPMLE.

References