Survival Analysis

Let $T \geq 0$ have density function $f(t)$ and distribution function $F(t)$. The survival function $S(t)$ is

$$S(t) = 1 - F(t) = P(T > t)$$

and the hazard rate or hazard function $\lambda(t)$ is

$$\lambda(t) = \frac{f(t)}{1 - F(t)}$$

interpreted as

$$\lambda(t) \approx P(t < T < t + dt \mid T > t) = P(\text{expiring in the interval } (t, t + dt) \mid \text{survived past time } t).$$

Integrating $\lambda(t)$ gives

$$\int_0^t \lambda(u)du = -\log S(t)$$
$$\Rightarrow S(t) = e^{-\int_0^t \lambda(u)du}$$

Censoring

What distinguishes survival analysis from other fields of statistics is censoring. Vaguely speaking, a censored observation contains only partial information about the random variable of interest.
Let $T_1, T_2, \ldots, T_n \overset{iid}{\sim} F(t)$.

Type I

Let t_c be a preassigned fixed number which we call the fixed censoring time. We observe $Y_1, \ldots, Y_n$

$$Y_i = \begin{cases} T_i & \text{if } T_i \leq t_c, \\ t_c & \text{if } T_i > t_c. \end{cases}$$

Type II

Let $r < n$ be fixed, and let $T_{(1)} < T_{(2)} < \ldots < T_{(n)}$ be the order statistics of $T_1, T_2, \ldots, T_n$. Observation ceases after the r^{th} failure, so we can observe $T_{(1)}, T_{(2)}, \ldots, T_{(r)}$. The full observed sample is

$$\begin{aligned} Y_{(1)} &= T_{(1)} \\ Y_{(2)} &= T_{(2)} \\ &\vdots \\ Y_{(r)} &= T_{(r)} \\ Y_{(r+1)} &= T_{(r)} \\ &\vdots \\ Y_{(n)} &= T_{(r)} \end{aligned}$$

Type III (Random Censoring)

Let $C_1, C_2, \ldots, C_n$ be iid each with distribution function G. Each C_i is the censoring time associated with T_i. We can only observe $(Y_1, \delta_1), (Y_2, \delta_2), \ldots, (Y_n, \delta_n)$ where

$$Y_i = \min(T_i, C_i) = T_i \wedge C_i$$

$$\delta_i = I(T_i \leq C_i) = \begin{cases} 1 & \text{if } T_i \leq c_i \text{ (not censored)}, \\ 0 & \text{if } T_i > c_i \text{ (censored)}. \end{cases}$$

notice $Y_1, \ldots, Y_n$ are iid with some distribution function H.

Example (p.185)

From February 1972 to February 1975, 29 severe vival hepatitis patients satisfied the admission criteria for a 16 week study of the effects of steroid therapy at the Stanford, Veterans Administration, and Santa Clara Valley Hospitals. These patients were randomized into either the steroid or control group. The survival times (in weeks) of the 14 patients in the steroid group were
1, 1, 1, 1+, 4+, 5, 7, 8, 10, 10+, 12+, 16+, 16+, 16+.

Assume an exponential distribution $S(t) = e^{-\lambda t}$

a) Estimate λ by maximum likelihood and construct an approximate 95% CI for λ.

b) Estimate $S(16)$ and construct an approximate 95% CI for $S(16)$.

c) Estimate the median survival time and construct an approximate 95% CI for the median.

Solutions:

a) Assume random censoring, the pair (y_i, δ) has likelihood

$$L(y_i, \delta_i) = \begin{cases} f(y_i) & \text{if } \delta_i = 1 \text{ (uncensored)} \\ S(y_i) & \text{if } \delta_i = 0 \text{ (censored)} \end{cases}$$

$$= f(y_i)^{\delta_i} S(y_i)^{1-\delta_i}.$$

The likelihood for the full sample

$$\begin{aligned} L &= L(y_1, \ldots, y_n; \delta_1, \ldots, \delta_n) \\ &= \prod_{i=1}^{n} L(y_i, \delta_i) \\ &= \left[\prod_u f(y_i)\right]\left[\prod_c S(y_i)\right] \\ &= \left[\lambda^{n_u} e^{-\lambda \sum_u t_i}\right]\left[e^{-\lambda \sum_c c_i}\right] \\ &= \lambda^{n_u} e^{-\lambda \sum_{i=1}^{n} y_i} \end{aligned}$$

$$\log L = n_u \log \lambda - \lambda \sum_{i=1}^{n} y_i$$

$$\frac{\partial}{\partial \lambda} \log L = \frac{n_u}{\lambda} - \sum_{i=1}^{n} y_i$$

Setting $\dfrac{\partial}{\partial \lambda} \log L = 0$ gives $\qquad \hat{\lambda} = \dfrac{n_u}{\sum_{i=1}^{n} y_i}$

$$\frac{\partial^2}{\partial \lambda^2} \log L = -\frac{n_u}{\lambda^2}$$

$$I(\lambda) = \frac{n_u}{\lambda^2}$$

So

$$\frac{\hat{\lambda} - \lambda}{\sqrt{\frac{\lambda^2}{n_u}}} \xrightarrow{d} N(0, 1)$$

Note that the normality approximation can be improved by transforming the estimate. By the δ-method,

$$\hat{\lambda} \xrightarrow{d} N\left(\lambda, \frac{\lambda^2}{n_u}\right)$$

then

$$\log \hat{\lambda} \xrightarrow{d} N\left(\log \lambda, \frac{1}{n_u}\right)$$

Notice that $1/n_u$, the asymptotic variance of $\log \hat{\lambda}$, does not depend on the unknown parameter λ. It is an empirical fact that transforming an estimate to remove the dependence of the variance on the unknown parameter tends to improve the convergence to normality by reducing the skewness.

$$\log \hat{\lambda} \xrightarrow{d} N\left(\log \lambda, \frac{1}{n_u}\right)$$

The 95% CI becomes

$$e^{\log \hat{\lambda} \pm z_\alpha / \sqrt{n_u}}$$

$$\left(\hat{\lambda} e^{-z_\alpha/\sqrt{n_u}}, \hat{\lambda} e^{z_\alpha/\sqrt{n_u}}\right)$$

$$(0.031, 0.136)$$

b) $\widehat{S(16)} = e^{-\hat{\lambda}(16)}$

95% CI for $S(16)$ is

$$\left(e^{-0.136(16)}, e^{-0.031(16)}\right)$$

$$(0.113, 0.609)$$

c) $\hat{t}_{med} = \log 2/\hat{\lambda} = 10.69$

95% CI for the median is

$$\left(\frac{\log 2}{0.136}, \frac{\log 2}{0.031}\right)$$

$$(5.097, 22.36)$$

Reference:

Miller Jr., Rupert G., *Survival Analysis*, Wiley, 1981.