BSTT536: Survival Data Analysis

Instructor: Hua Yun Chen, PhD

Division of Epidemiology and Biostatistics
School of Public Health
University of Illinois at Chicago

Table of Content

Survival Data

1. Measurement: Time to the occurrence of an event.
2. A well-defined event outcome.
3. A well-defined time origin leading to the event.
4. Examples
   4.1 Age in years at death.
   4.2 Time in days to recover from a common flu infection since the symptoms began.
   4.3 Time in waiting for a kidney transplant.

Survival Time Distribution

1. Let T denote the time to the event. The distribution of T is

$$P(T \leq t) = F(t).$$

   The density of the distribution is $f(t) = \frac{dF(t)}{dt}$.

2. In survival analysis, it is often more convenient to work with survival function defined as

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

3. The hazard density function is defined as

$$h(t) = \frac{f(t)}{S(t)}.$$

Interpretation of Hazard

1. Hazard density (or rate) at time t is the rate of the failure probability change just beyond time t given that the subject has not failed at time t.
2. More precisely,

$$h(t)\Delta t = P(T < t + \Delta t | T \geq t).$$

   The right-hand side is the probability of a subject failed by time $t + \Delta t$ given that the subject had survived at time t.

Cumulative Hazard

1. Define cumulative hazard as

$$H(t) = \int_0^t h(t)dt.$$

2. The cumulative hazard relates to the survival function.

$$S(t) = \exp\{-H(t)\}$$

or

$$H(t) = -\log S(t) = -\log\{1 - F(t)\}.$$

3. Distribution expressed in terms of the hazard function

$$\begin{aligned} S(t) &= \exp\{-H(t)\}, \\ f(t) &= h(t)S(t) = h(t)\exp\{-H(t)\}, \\ F(t) &= 1 - S(t) = 1 - \exp\{-H(t)\}. \end{aligned}$$

Exponential Distribution with Intensity λ

1. Distribution function and density

$$\begin{aligned} F(t) &= 1 - \exp(-\lambda t), \lambda > 0, \\ f(t) &= \lambda \exp(-\lambda t). \end{aligned}$$

2. Survival function and hazard

$$\begin{aligned} S(t) &= \exp(-\lambda t), \\ h(t) &= \lambda, \\ H(t) &= \lambda t. \end{aligned}$$

3. Relationship

$$\begin{aligned} f(t) &= h(t)S(t) = h(t)\exp\{-H(t)\} = \lambda \exp(-\lambda t), \\ F(t) &= 1 - S(t) = 1 - \exp\{-H(t)\} = 1 - \exp(-\lambda t). \end{aligned}$$

Weibull Distribution

1. Distribution function and density

$$\begin{aligned} F(t) &= 1 - \exp(-\lambda t^b), \lambda > 0, b > 0, \\ f(t) &= \lambda b t^{b-1} \exp(-\lambda t^b). \end{aligned}$$

2. Survival function and hazard

$$\begin{aligned} S(t) &= \exp(-\lambda t^b), \\ h(t) &= \lambda b t^{b-1}, \\ H(t) &= \lambda t^b. \end{aligned}$$

3. Relationship

$$\begin{aligned} f(t) &= h(t)S(t) = h(t)\exp\{-H(t)\} = \lambda b t^{b-1} \exp(-\lambda t^b), \\ F(t) &= 1 - S(t) = 1 - \exp\{-H(t)\} = 1 - \exp(-\lambda t^b). \end{aligned}$$

Complications in Analysis of Survival Data

1. Not followed up long enough to see the event: right censoring.
2. The event is known to have occurred before the follow-up started: left censoring.
3. The event is only known to have occurred in a time interval: interval censoring.
4. The event is not known to have occurred before the start of follow-up. There is no record for such subjects: left truncation.

Right censoring is mostly frequently occurred in practice.

Modeling Right Censoring

1. Treat censoring as another type of events.
2. The time from the time origin to the occurrence of a censoring event is called censoring time.
3. The survival time and the censoring time cannot be observed in the same time.
4. If the failure occurred before the censoring event, the survival time is observed. Otherwise, the censoring time is observed.
5. The right censored survival time data is a problem with incompletely observed data.

Two Types of Observed Data

1. Observed event time: Use X to denote the observed time and $\delta = 1$ to indicate a failure occurred. The observed X is the time it took for the event to happen.
2. Observed censoring time: Again, use X to denote the observed time and $\delta = 0$ to indicate a censoring occurred. The observed X is the time until the censoring occurred.
3. In either case, the time can be viewed as the time under observation, usually called follow-up time.

Likelihood for Survival Data Subject to Right Censoring

1. Assume the censoring and the failure risks are independent.
2. When $\delta = 1$, the event time $T = X$. The contribution of the observed data to the likelihood is

$$f(X) = h(X)\exp\{-H(X)\}.$$

3. When $\delta = 0$, the event time $T > X$. The contribution of the observed data to the likelihood is

$$S(X) = \exp\{-H(X)\}.$$

Likelihood for Survival Data Subject to Right Censoring (Continuing)

1. For a set of observed data (X_i, δ_i), $i = 1, \cdots, n$, the likelihood is
$$\prod_{i=1}^{n} f^{\delta_i}(X_i) S^{1-\delta_i}(X_i)$$
2. The likelihood expressed in terms of the hazard functions is
$$\prod_{i=1}^{n} h^{\delta_i}(X_i) \exp\{-H(X_i)\}.$$

The Maximum Likelihood Estimator for the Exponential Model

1. The likelihood under the exponential model

$$\prod_{i=1}^{n} \lambda^{\delta_i} \exp(-\lambda X_i).$$

2. The log-likelihood

$$l(\lambda) = \sum_{i=1}^{n} \delta_i \log \lambda - \lambda \sum_{i=1}^{n} X_i.$$

3. The first derivative (likelihood score) is

$$\frac{\partial l}{\partial \lambda} = \frac{1}{\lambda} \sum_{i=1}^{n} \delta_i - \sum_{i=1}^{n} X_i.$$

4. The second derivative is

$$\frac{\partial^2 l}{\partial \lambda^2}(\lambda) = -\frac{1}{\lambda^2} \sum_{i=1}^{n} \delta_i.$$

The Maximum Likelihood Estimator

1. The maximum likelihood estimator for λ is

$$\hat{\lambda} = \frac{\sum_{i=1}^n \delta_i}{\sum_{i=1}^n X_i}.$$

2. The observed information

$$-\frac{\partial^2 l}{\partial \lambda^2}(\hat{\lambda}) = \frac{1}{\hat{\lambda}^2} \sum_{i=1}^n \delta_i = \frac{\{\sum_{i=1}^n X_i\}^2}{\sum_{i=1}^n \delta_i}.$$

3. The estimated variance for $\hat{\lambda}$,

$$\hat{V} = \left\{-\frac{\partial^2 l}{\partial \lambda^2}(\hat{\lambda})\right\}^{-1} = \frac{\sum_{i=1}^n \delta_i}{\{\sum_{i=1}^n X_i\}^2}$$

Inference on the Survival Function

1. Estimated survival function

$$\hat{S}(t) = \exp(-\hat{\lambda} t).$$

2. Variance of the estimated survival function

$$\mathrm{Var}\{\hat{S}(t)\} = \hat{S}^2(t) t^2 \hat{V} = t^2 \exp(-2\hat{\lambda} t)\hat{V}.$$

3. 95% confidence interval (band) for $S(t)$

$$\hat{S}(t) \pm 1.96\sqrt{\mathrm{Var}\{\hat{S}(t)\}} = \hat{S}(t)\left[1 \pm 1.96 t\sqrt{\hat{V}}\right].$$

Derivation of the Variance Formula

1. The log-transformed survival function
$$\log \hat{S}(t) = -\hat{\lambda} t.$$
2. Variance for the log-transformed survival function
$$\text{Var}\left\{\log \hat{S}(t)\right\} = t^2 \text{Var}(\hat{\lambda}) = t^2 \hat{V}.$$
3. Variance for the survival function
$$\text{Var}\left\{\log \hat{S}(t)\right\} \approx \frac{\text{Var}\left\{\hat{S}(t)\right\}}{\hat{S}^2(t)}.$$
It follows that
$$\text{Var}\left\{\hat{S}(t)\right\} = \hat{S}^2(t) t^2 \hat{V}.$$
4. Confidence interval for $S(t)$
$$\hat{S}(t) \pm z_{1-\alpha/2} \hat{S}(t) t \sqrt{\hat{V}} = \hat{S}(t) \left\{1 \pm z_{1-\alpha/2} t \sqrt{\hat{V}}\right\}.$$

Confidence interval based on the log transformation

1. To ensure the confidence interval falls in $[0, 1]$, the confidence interval may be constructed based on the log-log transformation.
2. The $1 - \alpha$ confidence interval for $logS(t)$ is

$$\log \hat{S}(t) \pm z_{1-\alpha/2}\sqrt{\hat{Var}\left\{\log \hat{S}(t)\right\}}$$

   where

$$\widehat{Var}\left\{\log \hat{S}(t)\right\} = t^2\widehat{Var}(\hat{\lambda}) = t^2\hat{V}.$$

3. The confidence interval for $S(t)$ derived from the bove construction is

$$\left(\hat{S}(t)e^{-z_{1-\alpha/2}\sqrt{\widehat{Var}\{\log \hat{S}(t)\}}}, \hat{S}(t)e^{z_{1-\alpha/2}\sqrt{\widehat{Var}\{\log \hat{S}(t)\}}}\right).$$

Confidence interval based on the log-log transformation

1. To avoid the confidence interval including negative values, the confidence interval may be constructed based on the log-transformed survival function.
2. The $1-\alpha$ confidence interval for $log(-logS(t))$ is

$$\log(-\log \hat{S}(t)) \pm z_{1-\alpha/2}\sqrt{\hat{Var}\left\{\log(-log\hat{S}(t))\right\}}$$

   where

$$\widehat{Var}\left\{\log(-\log \hat{S}(t))\right\} = \widehat{Var}(\log \hat{\lambda}) = \frac{\widehat{Var}(\hat{\lambda})}{\hat{\lambda}^2}.$$

3. The confidence interval for $S(t)$ derived from the bove construction is

$$\hat{S}(t)^{\exp\{\mp z_{1-\alpha/2}\sqrt{\widehat{Var}\left\{\log(-log\hat{S}(t))\right\}}\}}$$

Inference on the mean survival time

1. The mean survival time

$$\mu = E(T) = \int_0^\infty t \exp(-\lambda t) dt = \frac{1}{\lambda}.$$

2. The maximum likelihood estimator of the mean survival time is

$$\hat{\mu} = \frac{1}{\hat{\lambda}} = \frac{\sum_{i=1}^n X_i}{\sum_{i=1}^n \delta_i}.$$

3. The variance of the mean survival time estimator is

$$var(\hat{\mu}) \approx \frac{1}{\hat{\lambda}^4} var(\hat{\lambda}) \approx \left(\frac{\sum_{i=1}^n X_i}{\sum_{i=1}^n \delta_i}\right)^4 \frac{\sum_{i=1}^n \delta_i}{(\sum_{i=1}^n X_i)^2} = \frac{(\sum_{i=1}^n X_i)^2}{(\sum_{i=1}^n \delta_i)^3}.$$

4. The 95% confidence interval for $\mu = E(T)$ is

$$\left[\hat{\mu} - 1.96\sqrt{var(\hat{\mu})}, \hat{\mu} + 1.96\sqrt{var(\hat{\mu})}\right].$$

Inference on the median survival time

1. The median survival time $t_0.5$ satisfies

$$S(t_{0.5}) = 0.5.$$

   For the exponential model, $t_{0.5} = \frac{1}{\lambda}\log(2)$. The maximum likelihood estimator is

$$\hat{t}_{0.5} = \frac{1}{\hat{\lambda}}\log(2).$$

2. The variance of the median survival time estimator is

$$var(\hat{t}_{0.5}) \approx \frac{\{\log(2)\}^2}{\hat{\lambda}^4} var(\hat{\lambda}) \approx \{\log(2)\}^2 \frac{(\sum_{i=1}^n X_i)^2}{(\sum_{i=1}^n \delta_i)^3}.$$

3. The 95% confidence interval for the median survival time $t_{0.5}$ is

$$\left[\hat{t}_{0.5} - 1.96\sqrt{var(\hat{t}_{0.5})}, \hat{t}_{0.5} + 1.96\sqrt{var(\hat{t}_{0.5})}\right].$$

Example 1: Survival Times of Patients in a Study on Multiple Myeloma

Observed events times

13,52*,6,40,10,7*,66,10*,10,14,16,4,65,5,11*,10,15*,5,
76*,56*,88,24,51,4,40*,8,18,5,16,50,40,1,36,5,10,91,
18*,1,18*,6,1,23,15,18,12*,12,17,3*

A star following a number indicates that number is a censoring time.

Exponential Model Fit

1. Key statistics

$$\sum_{i=1}^{n} d_i = 36, \qquad \sum_{i=1}^{n} X_i = 1122.$$

2. λ estimate

$$\hat{\lambda} = \frac{\sum_{i=1}^{n} \delta_i}{\sum_{i=1}^{n} X_i} = \frac{36}{1122} = 0.0321.$$

3. Variance estimate

$$\hat{V} = \frac{\sum_{i=1}^{n} \delta_i}{\{\sum_{i=1}^{n} X_i\}^2} = \frac{36}{1122^2} = 2.86 \times 10^{-5}.$$

Exponential Model Fit (continuing)

1. 95% confidence interval for $S(t)$:

$$\exp(-0.0321t)\left\{1 \pm 1.96 \times \sqrt{2.86 \times 10^{-5}}\ t\right\}$$

$$= \exp(-0.0321t)\left\{1 \pm 0.0105t\right\}.$$

2. 95% confidence interval for the mean survival time is

$$\left[31.17 - 1.96\sqrt{26.98}, 31.17 + 1.96\sqrt{26.98}\right] = [20.99, 41.35].$$

3. 95% confidence interval for the median survival time is

$$\left[21.60 - 1.96\sqrt{12.96}, 21.60 + 1.96\sqrt{12.96}\right] = [14.54, 28.66].$$

4. Test the hypothesis $H_0 : \lambda = 0.05$.

$$\frac{\hat{\lambda} - 0.05}{\sqrt{\hat{V}}} = \frac{0.0321 - 0.05}{\sqrt{2.86 \times 10^{-5}}} = -3.35.$$

Survival Function Estimate

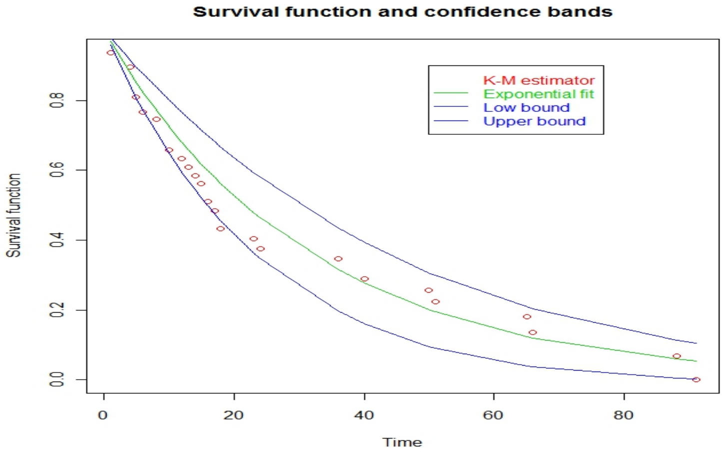

Figure: Survival function estimate for the data on multiple myeloma.

Confidence band I*

1. The band in the previous figure are composed of individual confidence intervals and should not be interpreted as confidence band.
2. The 95% confidence band refers to having 0.95 probability that the true survival curve contained entirely in the band, i.e.,

$$P(\sup_{t\in[0,\tau]} |\sqrt{n}\{\exp(-\hat{\lambda}t) - \exp(-\lambda t)\}| < z) = 0.95.$$

3. Note that

$$\sup_{t\in[0,\tau]} \{t\exp(-\lambda t)\} = \begin{cases} \tau\exp(-\lambda\tau) & \text{if } \tau < 1/\lambda, \\ \exp(-1)/\lambda & \text{if } \tau \geq 1/\lambda. \end{cases}$$

Confidence band II*

1. Since

$$\begin{aligned} B(\lambda) &= \sup_{t \in [0,\tau]} |\sqrt{n}\{\exp(-\hat{\lambda} t) - \exp(-\lambda t)\}| \\ &= \sup_{t \in [0,\tau]} \{t \exp(-\lambda t)\}|\sqrt{n}(\lambda - \lambda)|, \end{aligned}$$

the 95% confidence band for $S(t)$ on $[0, \tau]$ is

$$\exp(-\hat{\lambda} t) \pm 1.96 B(\hat{\lambda})\sqrt{V}$$

where V is the estimated variance of $\hat{\lambda}$.