STAT 7780: Survival Analysis

2. Basic Quantities and Models

Peng Zeng

Department of Mathematics and Statistics
Auburn University

Fall 2017

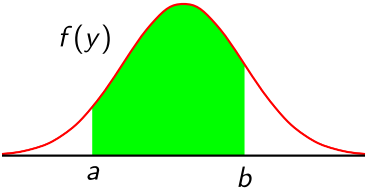
Outline

1. Basic quantities and models
   - Survival function
   - Hazard function
   - Mean residual life
   - Median life
   - Common parametric models
2. Censoring and truncation
   - Censoring
   - Truncation
3. Likelihood function

Reference: Chapter 2 and 3 in Klein and Moeschberger (2003).

Continuous Random Variables

In general, a continuous random variable X can have values from an interval. Its probability density function $f(x)$ satisfying



$$f(x) \geq 0, \quad \int_{-\infty}^{\infty} f(x)dy = 1.0$$

$$P(a \leq X \leq b) = \int_a^b f(x)dx$$

The cumulative distribution function (CDF) is

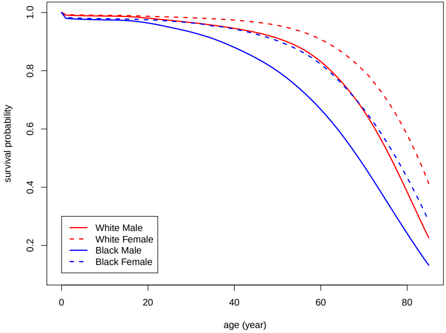
$$F(x) = P(X \leq x) = \int_{\infty}^{x} f(u)du.$$

Survival Function of US Population

How likely a person can live beyond the age of 70 in US?

According to US Department of Health and Human Services, in 1989, the survival probability is

White Male	0.66166
White Female	0.79839
Black Male	0.47312
Black Female	0.66730



Survival Function: Definition

Let X be the time until some specified event.

The survival function is the probability of an individual surviving beyond time x (experiencing the event after time x).

$$S(x) = P(X > x).$$

Some comments.

- $S(x)$ is called the reliability function in engineering applications.
- $S(x)$ is a monotone, non-increasing function.
- $S(0) = 1$ and $S(\infty) = 0$.

Survival Function: Continuous Random Variable

When X is a continuous random variable with density $f(x)$ and cumulative distribution function $F(x) = P(X \leq x)$,

$$S(x) = \int_x^\infty f(t)dt = 1 - F(x).$$

and

$$f(x) = -\frac{dS(x)}{dx}.$$

Example: The density of Weibull distribution is

$$f(x) = \alpha\lambda x^{\alpha-1}\exp(-\lambda x^\alpha), \quad \lambda > 0, \alpha > 0.$$

Its survival function is $S(x) = \exp(-\lambda x^\alpha)$.

Hazard Function

The hazard function is the instantaneous rate of failure at time x,

$$h(x) = \lim_{\Delta \to 0} \frac{P(x \leq X < x + \Delta \mid X \geq x)}{\Delta}$$

Some comments

- $h(x)$ is nonnegative, $h(x) \geq 0$.
- $h(x)\Delta$ is the approximate probability of an individual of age x experiencing the event in $[x, x + \Delta)$.
- Other names: conditional failure rate (reliability), force of mortality (demography), age-specific failure rate (epidemiology), inverse of the Mills ratio (economics).

Cumulative Hazard Function

The cumulative hazard function is the probability of death up to time x.

$$H(x) = \int_0^x h(u)du$$

Hazard Function: Continuous Random Variable

For a continuous random variable X with density $f(x)$ and survival function $S(x)$, we have

$$
\begin{aligned}
h(x) &= \frac{f(x)}{S(x)} = -\frac{d \ln S(x)}{dx} \\
H(x) &= -\ln S(x) \\
S(x) &= \exp[-H(x)] = \exp[-\int_0^x h(u) du]
\end{aligned}
$$

Example: Weibull Distribution

For Weibull distribution, its density is

$$f(x) = \alpha\lambda x^{\alpha-1} \exp(-\lambda x^{\alpha}), \quad \lambda > 0, \alpha > 0.$$

Its survival function is

$$S(x) = \exp(-\lambda x^{\alpha})$$

Its hazard function is

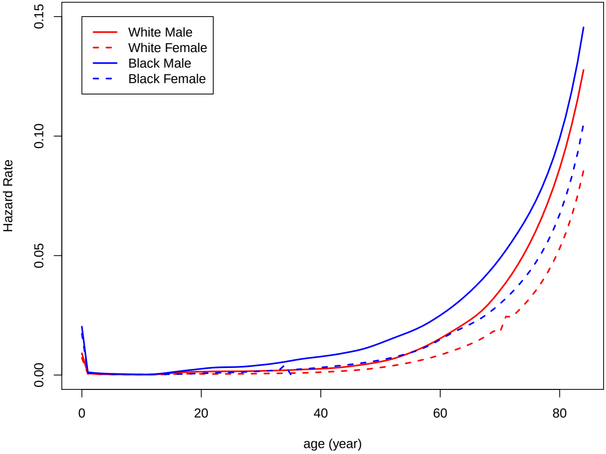
$$h(x) = \alpha\lambda x^{\alpha-1}$$

Its cumulative hazard function is

$$H(x) = \lambda x^{\alpha}$$

Hazard Rate for US Population

Hazard functions for all cause mortality for the US population in 1989.



Mean Residual Life

Mean residual life is the expected remaining life.

$$\mathrm{mrl}(x) = E(X - x \mid X > x)$$

It is the area under $S(x)$ to the right of x divided by $S(x)$.

$$\mathrm{mrl}(x) = \frac{\int_x^\infty (t-x) f(t) dt}{S(x)} = \frac{\int_x^\infty S(t) dt}{S(x)}.$$

(Hint: integration by parts and $f(t)dt = -dS(t)$.)

Mean and Variance

The mean and variance of life can be expressed in terms of survival function.

The mean life is $\mu = \text{mrl}(0)$.

$$\mu = E(X) = \int_0^\infty tf(t)dt = \int_0^\infty S(t)dt.$$

The variance of X is

$$var(X) = 2\int_0^\infty tS(t)dt - \Big[\int_0^\infty S(t)dt\Big]^2$$

Median Lifetime

The pth quantile (also 100pth percentile) of X is the smallest x_p so that

$$S(x_p) \leq 1 - p, \quad \text{or} \quad x_p = \inf\{t : \ S(t) \leq 1 - p\}$$

The median lifetime is $x_{0.5}$, the 50th percentile. When X is continuous, we can find $x_{0.5}$ from

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

Question: How to find the median lifetime from a plot of survival function?

Example: Exponential Distribution

The density of exponential distribution is

$$f(x) = \lambda e^{-\lambda x}, \quad \lambda > 0.$$

Recall that its survival function is $S(x) = e^{-\lambda x}$. Its mean and median life are

$$\mu = \frac{1}{\lambda}, \quad \text{median} = \frac{\ln 2}{\lambda}.$$

Common Parametric Models

Common parametric models include

- exponential distribution
- Weibull distribution
- gamma distribution
- . . .

Refer to Table 2.2 of the textbook.

Exponential and Weibull Distribution

- Exponential distribution is memoryless.

$$P(X \geq x + z \mid X \geq x) = P(X \geq z).$$

  The mean residual life is a constant, and the hazard function is also a constant.

- Weibull distribution has a hazard function

$$h(x) = \lambda \alpha x^{\alpha - 1}.$$

  It can accommodate increasing $(\alpha > 1)$, decreasing $\alpha < 1$, or constant $\alpha = 1$.

Right Censoring: Type I Censoring

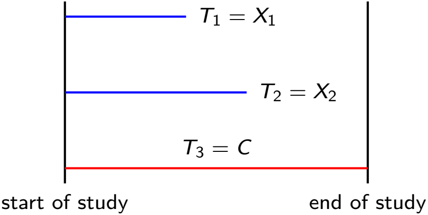
For a specific individual under study, assume that

- lifetime X, (iid with density $f(x)$ and survival function $S(x)$)
- fixed censoring time C_r.

We actually observe (T, δ), where

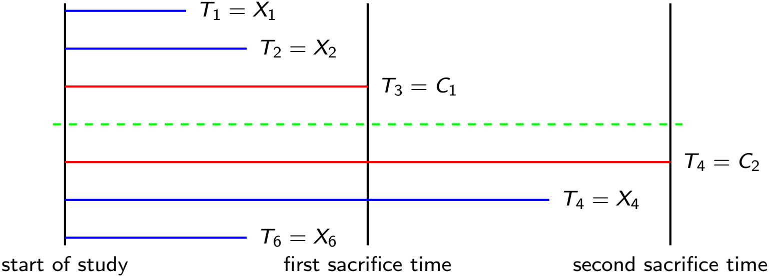
$$T = \min\{X, C_r\}$$

$$\delta = \begin{cases} 1, & \text{actual lifetime} \\ 0, & \text{censoring time} \end{cases}$$



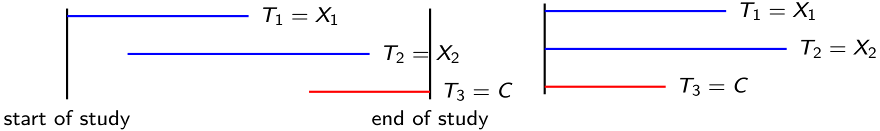
Progressive Type I Censoring

For progressive type I censoring, we may have different and fixed censoring times. For example, sacrifice the first batch of animals at time C_1 and sacrifice the second batch time of animals at time C_2.



Generalized Type I Censoring

For generalized type I censoring, individuals enter the study at different times and the terminal point of the study is predetermined by the investigator.



We can shift each individual's starting time to 0. Note that different individuals may have different censoring time.

Type II Censoring

For type II censoring, the study continues until the failure the first r individuals, where r is some predetermined integer ($r < n$).

- For example: In testing equipment life, the test is terminated when r of the total n items have failed.

For progressive Type II censoring, when the first r_1 items fail, remove $n_1 - r_1$ of the remaining $n - r_1$ unfailed items. Continue the experiment until the next r_2 items fail.

Left Censoring

For left censoring, the event of interest has already occurred for the individual before that person is observed in the study.

- In early childhood learning center, an investigator is interested in when a child learns to accomplish certain tasks. It is quite often that some children can already perform the task when they start in the study.

Let X be the exact lifetime and C_l be the left censoring time. We actually observe (T, ε), where

$$T = \max\{X, C_l\}, \quad \varepsilon = \begin{cases} 1, & T \text{ is the actual lifetime} \\ 0, & T \text{ is the censoring time} \end{cases}$$

Double Censoring and Interval Censoring

For double censoring, we are able to observe the actual lifetime X only when X falls within (C_l, C_r). We actually observe (T, δ), where

$$T = \max\{\min\{X, C_r\}, C_l\},$$

$$\delta = \begin{cases} 1, & T \text{ is the actual lifetime} \\ 0, & T \text{ is the right censoring time} \\ -1, & T \text{ is the left censoring time} \end{cases}$$

For interval censoring, we only know that the individual's event time falls in an interval $(L_i, R_i]$.

Truncation

Truncation occurs when only those individuals whose event time lies within a certain observational window (Y_L, Y_R) are observed.

- (left truncation) A study involved residents of a retirement center. Since an individual must survive to a sufficient age to enter the retirement center, all individuals who died earlier are out of the investigator's cognizance.
- (right truncation) A study considered patients with transfusion induced AIDS. The registry was sampled on June 30, 1986. So only those whose waiting time from transfusion to AIDS was less than the time from transfusion to June 30, 1986 were available for observation. Patients transfused prior to June 30, 1986, who developed AIDS after June 30, 1986, were not observed.

censoring: partial information. truncation: no information.

Likelihood Function

Assume that $X_1, \ldots, X_n$ are iid with density $f(x;\theta)$. Then the joint density function is

$$f(x_1, \ldots, x_n; \theta) = \prod_{i=1}^{n} f(x_i; \theta).$$

The likelihood function is essentially the joint density function, but we treat it as a function of θ.

$$L(\theta) = f(x_1, \ldots, x_n; \theta) = \prod_{i=1}^{n} f(x_i, \theta).$$

Example: Write out the likelihood function for $X_1, \ldots, X_n$, which are iid exponential(λ).

Maximum Likelihood Estimate

The maximum likelihood estimate (MLE) maximizes the likelihood function.

$$\hat{\theta} = \arg\max L(\theta; x_1, \ldots, x_n) = \arg\max \ell(\theta; x_1, \ldots, x_n),$$

where $\ell(\theta) = \ln L(\theta)$ is the log-likelihood.

$$\ell(\theta; x_1, \ldots, x_n) = \sum_{i=1}^{n} \ln f(x_i; \theta).$$

Example: Suppose that $X_1, \ldots, X_n$ are iid exponential(λ). Find MLE of λ.

Properties of Maximum Likelihood Estimate

Maximum likelihood estimate is the most popular method for estimating unknown parameters in a statistical model.

When the sample size is large, the MLE $\hat{\theta}$ approximately follows a normal distribution.

$$E(\hat{\theta}) \approx \theta, \quad \text{and in many cases, } E(\hat{\theta}) = \theta.$$
$$var(\hat{\theta}) \approx \left\{ n\, E\left(\frac{\partial \ln f(X;\theta)}{\partial \theta} \right)^2 \right\}^{-1}$$

where $f(x;\theta)$ is the density or probability mass function of a single observation.

Example: Suppose that $X_1, \ldots, X_n$ are iid exponential(λ). Find the asymptotic distribution of MLE of λ.

Delta Method

If $\hat{\theta}$ is the MLE of θ, then $\tau(\hat{\theta})$ is the MLE of $\tau(\theta)$ and $\tau(\hat{\theta})$ is approximately normal with

$$E[\tau(\hat{\theta})] \approx \tau(\theta), \quad var[\tau(\hat{\theta})] \approx [\tau'(\theta)]^2\, var(\hat{\theta})$$

where τ is any differentiable function and τ' is the first-order derivative of τ.

Example: Suppose that $X_1, \ldots, X_n$ are iid exponential(λ). Find the MLE of λ^2 and the asymptotic distribution of the estimate.

Likelihood for Data With Censoring and Truncation

Assemble the likelihood using the following components

exact lifetime	$f(x)$
right-censored observations	$S(C_r)$
left-censored observations	$1 - S(C_l)$
interval-censored observations	$S(L) - S(R)$
left-truncated observations	$f(x)/S(Y_L)$
right-truncated observations	$f(x)/[1 - S(Y_R)]$
interval-truncated observations	$f(x)/[S(Y_L) - S(Y_R)]$

Example: Write out the likelihood function for the following observations

$$(t_1, 1), (t_2, 0), (t_3, 1), (t_4, 0), (t_5, 1).$$

Likelihood for Right Censoring

For right censoring, assume that the observations are $\{(t_i, \delta_i), i = 1, \ldots, n\}$. The density function can be written as

$$f(t, \delta) = [f(t)]^{\delta}[S(t)]^{1-\delta}$$

Example: Assume that the observations are $\{(t_i, \delta_i), i = 1, \ldots, n\}$, where the lifetime follows exponential distribution. The likelihood function is

$$L(\lambda) = \lambda^r e^{-\lambda S_T},$$

where $r = \sum \delta_i$ is the observed number of events and $S_T = \sum t_i$ is the total time for n individuals under study.

Calculation of MLE

The MLE is the maximizer of $L(\theta)$ or $\ell(\theta)$, or equivalently, the root of

$$U(\theta) = \frac{\partial \ell(\theta)}{\partial \theta} = \begin{pmatrix} \frac{\partial \ell(\theta)}{\partial \theta_1} \\ \vdots \\ \frac{\partial \ell(\theta)}{\partial \theta_p} \end{pmatrix} = \sum_{i=1}^{n} \frac{\partial \log f(x_i; \theta)}{\partial \theta} = 0.$$

In many applications, there exists no explicit solution for MLE. Hence, a numerical approach has to be applied.

- When θ is a vector, the partial derivative is also a vector.
- $U(\theta)$ is referred to as the score function.

Linearization

Suppose that θ_0 is the maximizer. We approximate $U(\theta_0)$ using Taylor expansion in a neighborhood of θ,

$$0 = U(\theta_0) \approx U(\theta) + H(\theta)(\theta_0 - \theta),$$

where H is usually called a Hessian matrix

$$H(\theta) = \frac{\partial U(\theta)}{\partial \theta^T} = \frac{\partial^2 \ell(\theta)}{\partial \theta \partial \theta^T} = \left(\frac{\partial^2 \ell(\theta)}{\partial \theta_i \partial \theta_j} \right)_{p \times p}$$

Therefore, we can calculate θ_0 if θ is close to θ_0.

$$\theta_0 \approx \theta - [H(\theta)]^{-1} U(\theta).$$

Newton-Raphson Algorithm

Starting from an initial guess of θ, say $\theta^{(0)}$, we iteratively apply the following formula,

$$\theta^{(k+1)} = \theta^{(k)} - [H(\theta^{(k)})]^{-1} U(\theta^{(k)}).$$

Stop if one of the following two criteria is satisfied.

- k is too large
- $\|\theta^{(k+1)} - \theta^{(k)}\|_2$ is small enough.

The second criterion can be replaced by $\|U(\theta^{(k+1)})\|_2$ is small enough or $|\ell(\theta^{(k+1)}) - \ell(\theta^{(k)})|$ is small enough.

Example

Suppose that $X_1, \ldots, X_n$ are iid Weibull(α, λ) random variables. Find the MLE of α and λ.

The log-likelihood function is

$$\begin{aligned}\ell(\alpha, \lambda) &= \sum_{i=1}^{n} \ln(\alpha \lambda x_i^{\alpha-1} e^{-\lambda x_i^{\alpha}}) \\ &= n \ln \alpha + n \ln \lambda + (\alpha - 1) \sum \ln(x_i) - \lambda \sum x_i^{\alpha}\end{aligned}$$

Let $\theta = (\alpha, \lambda)^T$. The score function $U(\theta)$ is

$$U(\theta) = \begin{pmatrix} n\alpha^{-1} + \sum \ln(x_i) - \lambda \sum x_i^{\alpha} \ln(x_i) \\ n\lambda^{-1} - \sum x_i^{\alpha} \end{pmatrix}$$

Hessian Matrix

The Hessian matrix is

$$H(\theta) = \begin{pmatrix} -n\alpha^{-2} - \lambda \sum x_i^{\alpha} [\ln(x_i)]^2 & -\sum x_i^{\alpha} \ln(x_i) \\ -\sum x_i^{\alpha} \ln(x_i) & -n\lambda^{-2} \end{pmatrix}$$

Therefore, the updating formula is

$$\theta^{(k+1)} = \theta^{(k)} - [H(\theta^{(k)})]^{-1} U(\theta^{(k)})$$

Simulation

Randomly generate 200 samples from Weibull distribution with $\alpha = 2.0$ and $\lambda = 4.0$.

	$\hat{\alpha}$	$\hat{\lambda}$	$\hat{\alpha}$	$\hat{\lambda}$
step 0	1.000000	2.249427	0.5000000	1.000000
step 1	1.581915	3.137301	0.9024095	1.565138
step 2	1.965795	3.869274	1.4370040	2.402186
step 3	2.049438	4.086029	1.8644937	3.343296
step 4	2.052623	4.099165	2.0254037	3.939896
step 5	2.052630	4.099215	2.0514540	4.091416
step 6	2.052630	4.099215	2.0526276	4.099196
step 7			2.0526305	4.099215
step 8			2.0526305	4.099215

Some Comments

- A good initial guess usually leads to faster convergence.
- local minimum.
- Change step length if necessary. Use a value $s \in [0, 1]$.

$$\theta^{(k+1)} = \theta^{(k)} - s[H(\theta^{(k)})]^{-1} U(\theta^{(k)})$$

- The estimate of variance of $\hat{\theta}$ is $I(\hat{\theta})^{-1}$, the inverse of observed information matrix,

$$I(\theta) = -\frac{\partial U(\theta)}{\partial \theta^T} = -\frac{\partial^2 \ell(\theta)}{\partial \theta \partial \theta^T}$$

Multivariate Normal Distribution

A random vector Y follows multivariate normal distribution $MVN(\mu, \Sigma)$ if and only if for any vector a, $a^T Y$ follows a univariate normal distribution.

- $E(Y) = \mu$ and $var(Y) = \Sigma$.
- $Y + a \sim N(\mu + a, \Sigma)$ for any constant vector a.
- $a^T Y \sim N(a^T \mu, a^T \Sigma a)$ for any constant vector a.
- $AY \sim N(A\mu, A\Sigma A^T)$ for any constant matrix A.
- $(Y - \mu)^T \Sigma^{-1} (Y - \mu) \sim \chi^2_p$, where p is the dimension of Y.

MLE and Score Function

The MLE $\hat{\theta}$ follows a normal distribution

$$\hat{\theta} \sim MVN(\theta, E(I(\theta))^{-1}),$$

where $I(\theta)$ is the Fisher's information matrix.

The score function $U(\theta)$ follows a normal distribution

$$U(\theta) \sim MVN(0, E(I(\theta))).$$

Hypothesis Testing

Three different approaches to test

$$H_0: \ \theta = \theta_0, \quad H_a: \ \theta \neq \theta_0,$$

where θ_0 is a given value (or vector).

- Wald's test. Notice that $\hat{\theta}$ asymptotically follows a normal distribution with mean θ_0 and variance $E[I(\theta_0)]^{-1}$ under H_0.

$$X_W^2 = (\hat{\theta} - \theta_0)^T I(\hat{\theta})(\hat{\theta} - \theta_0) \sim \chi_p^2$$

- Likelihood ratio test.

$$X_{LR}^2 = 2[\ell(\hat{\theta}) - \ell(\theta_0)] \sim \chi_p^2$$

- Scores test. Notice that $U(\theta_0)$ asymptotic follows a normal distribution with mean 0 and covariance $E[I(\theta_0)]$ under H_0.

$$X_{SC}^2 = U(\theta_0)^T [I(\theta_0)]^{-1} U(\theta_0) \sim \chi_p^2$$

Some Comments

For all three tests,

- The number of degrees of freedom is p, the dimension of θ.
- The p-value is $P(\chi^2_p > X^2)$, where X^2 can be X^2_W, X^2_{LR}, or X^2_{SC}.
- Reject H_0 if $X^2 > \chi^2_{p,\alpha}$.
- The tests can be used for more complicated hypothesis.

$$H_0:\ \theta \in \Theta_0, \quad H_a:\ \theta \notin \Theta_0,$$

The number of degrees of freedom becomes the difference of the number of free parameters in H_0 and H_a.