ST 745, Spring 2006

Analysis of Survival Data

Lecture Notes

(Modified from Drs. Tsiatis and Zhang's Lecture Notes)

Wenbin Lu

Department of Statistics

North Carolina State University

1. Introduction of Survival Analysis

In many biomedical applications the primary endpoint of interest is time to a certain event. Examples include

- time to death;
- time it takes for a patient to respond to a therapy;
- time from response until disease relapse (i.e., disease returns); etc.

We may be interested in characterizing the distribution of "time to event" for a given population as well as comparing this "time to event" among different groups (*e.g.*, treatment vs. control in a clinical trial or an observational study), or modeling the relationship of "time to event" to other covariates (sometimes called prognostic factors or predictors). Typically, in biomedical applications the data are collected over a finite period of time and consequently the "time to event" may not be observed for all the individuals in our study population (sample). This results in what is called censored data. That is, the "time to event" for those individuals who have not experienced the event under study is **censored** (by the end of study). It is also common that the amount of follow-up for the individuals in a sample vary from subject to subject. The combination of censoring and differential follow-up creates some unusual difficulties in the analysis of such data that cannot be handled properly by the standard statistical methods. Because of this, a new research area in statistics has emerged which is called Survival Analysis or Censored Survival Analysis.

1.1 Some examples of survival data

- Times to relapse or death of breast cancer patients (Farewell 1986, right censoring)
- Remission durations of leukaemia patients (Gehan 1965; Freireich et al. 1963, right censoring)

- Time until first marijuana use among high school boys (left censoring, current status data); Time until a child learns to accomplish certain specified task (Turnbull 1974, doubly censoring); periodic follow-up or examination times (interval censoring)
- Death times of individuals entering the retirement community (example 1.16 on page 16 of Klien and Moeschberger, left truncation); Time to AIDS (example 1.19 on page 19 of Klien and Moeschberger, right truncation)
- Infections in catheters for patients on dialysis (multiple events)
- Bladder tumor recurrence data (Byar 1980, recurrent events, the data is given in Table 9.2, page 292 of Kalbfleisch and Prentice 2002)
- Familial disease study (eg. Australian twin study, clustered events)

1.2 Basic terminologies

- Study endpoint
- Censoring (right, left, doubly, interval, current status data)
- Truncation (left, right)
- Multivariate event times (multiple events, clustered events, recurrent events)

1.3 Notations

Right censoring: failure time T, censoring time C, censoring indicator Δ. The observed data $\tilde{T} = \min(T, C)$ and $\Delta = I(T \le C)$. Since right censoring is the most common censoring scheme, we will focus on this special case most of the time in this course.

Table 1.1: *Time to relapse or death data for patients with breast cancer*

Event	Censoring	Treatment	Clinical	Pathological	Histological	Number of
time	indicator	group	stage	stage	stage	lymph nodes
3992	0	2	1	1	0	0
3976	0	2	1	0	1	3
303	1	1	0	0	1	1
802	1	2	0	0	0	7
2403	1	2	0	0	1	8
3961	0	1	1	1	1	0
3974	0	3	1	0	1	1
.	.	.	.	.	.	.
.	.	.	.	.	.	.
.	.	.	.	.	.	.
.	.	.	.	.	.	.
.	.	.	.	.	.	.
.	.	.	.	.	.	.
663	0	3	0	0	0	3
227	1	2	0	0	0	3
460	1	1	0	0	1	6
498	0	2	0	0	1	1
549	0	1	0	0	1	2
447	0	1	1	1	0	0

Note: Event time is recorded in days. Treatment group A=2, Treatment group B=3, Control group=1.

Source : Farewell (1986)

Table 1.2: *Remission durations data for patients with leukaemia*

		Control Group		Treatment Group	
Pair	Remission Status	t_i	δ_i	t_i	δ_i
1	partial	1	1	10	1
2	complete	22	1	7	1
3	complete	3	1	32	0
4	complete	12	1	23	1
5	complete	8	1	22	1
6	partial	17	1	6	1
7	complete	2	1	16	0
8	complete	11	1	34	0
9	complete	8	1	32	0
10	complete	12	1	25	0
11	complete	2	1	11	0
12	partial	5	1	20	0
13	complete	4	1	19	0
14	complete	15	1	6	1
15	complete	8	1	17	0
16	partial	23	1	35	0
17	partial	5	1	6	1
18	complete	11	1	13	1
19	complete	4	1	9	0
20	complete	1	1	6	0
21	complete	8	1	10	0

Note: Duration time t_i is recorded in weeks and δ_i is censoring indicator. *Source* : Gehan (1965); Freireich *et al.* (1963).

2. Describing the Distribution of Time-to-Event

In routine data analysis, we may first present some summary statistics such as mean, standard error for the mean, etc. In analyzing survival data, however, because of possible censoring, the summary statistics may not have the desired statistical properties, such as unbiasedness. For example, the sample mean is no longer an unbiased estimator of the population mean (of survival time). So we need to use other methods to present survival data. One way is to estimate the underlying distribution of survival time. When this distribution is estimated (either parametrically or nonparametrically), we then can estimate other quantities of interest such as mean, median, etc. of the survival time.

The distribution of survival time T can be described in a number of equivalent ways. There is of course the usual (cumulative) distribution function

$$F(t) = P[T \leq t], \quad t \geq 0, \tag{1}$$

which is **right** continuous, *i.e.*, $\lim_{u \to t^+} F(u) = F(t)$. $F(t)$ is the probability that a randomly selected subject from the population will die **before** time t.

If T is a continuous random variable, then it has a density function $f(t)$, which is related to $F(t)$ through following equations

$$f(t) = \frac{dF(t)}{dt}, \; F(t) = \int_0^t f(u)du. \tag{2}$$

In biomedical applications, it is often common to use the **survival function**

$$S(t) = P[T \geq t] = 1 - F(t^-), \tag{3}$$

where $F(t^-) = \lim_{u \to t^-} F(u)$. $S(t)$ is the probability that a randomly selected individual will **survive** to time t or beyond.

Note: Some authors use the following definition of a survival function

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

This definition will be identical to the above one if T is a continuous random variable, which is the case we will focus on in this course.

The survival function $S(t)$ is a non-increasing function over time taking on the value 1 at $t = 0$, *i.e.,* $S(0) = 1$. For a proper random variable T, $S(\infty) = 0$, which means that everyone will eventually experience the event if there was no censoring. However, we will also allow the possibility that $S(\infty) > 0$. This corresponds to a situation where there is a positive probability of not "dying" or not experiencing the event of interest. For example, if the event of interest is the time from response until disease relapse and the disease has a cure for some proportion of individuals in the population, then we have $S(\infty) > 0$, where $S(\infty)$ corresponds to the proportion of cured individuals or long-term survivors (Ref. Maller and Zhou, 1996).

Obviously if T is a continuous r.v., we have

$$S(t) = \int_t^\infty f(u)du, \qquad f(t) = -\frac{dS(t)}{dt}. \tag{4}$$

That is, there is a one-to-one correspondence between $f(t)$ and $S(t)$.

Mean Survival Time: $\mu = \mathrm{E}(T)$. Due to censoring, sample mean of observed survival times is no longer an unbiased estimate of $\mu = \mathrm{E}(T)$. If we can estimate $S(t)$ well, then we can estimate $\mu = \mathrm{E}(T)$ using the following fact:

$$\mathrm{E}(T) = \int_0^\infty S(t)dt. \tag{5}$$

Median Survival Time: Median survival time m is defined as the quantity m satisfying $S(m) = 0.5$. Sometimes denoted by $t_{0.5}$. If S(t) is not strictly decreasing, m is the smallest one such that $S(m) \leq 0.5$.

***p*th quantile of Survival Time** (100pth percentile): t_p such that $S(t_p) = 1-p$ ($0 < p < 1$). If S(t) is not strictly decreasing, t_p is the smallest one such that $S(t_p) \leq 1 - p$.

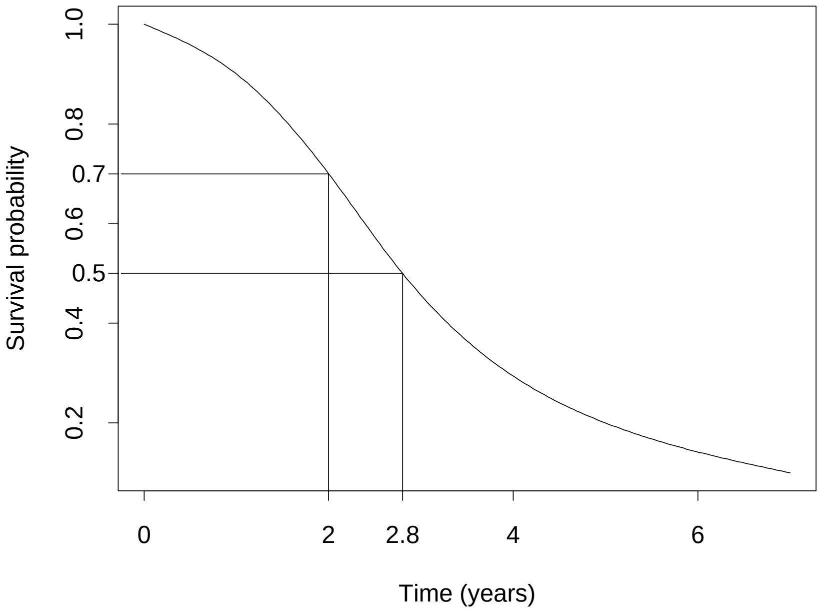
Mean Residual Life Time(mrl):

$$mrl(t_0) = \mathrm{E}[T - t_0 | T \geq t_0], \tag{6}$$

i.e., $mrl(t_0)$ = average remaining survival time **given** the population has survived beyond t_0. It can be shown that

$$mrl(t_0) = \frac{\int_{t_0}^{\infty} S(t)dt}{S(t_0)}. \tag{7}$$

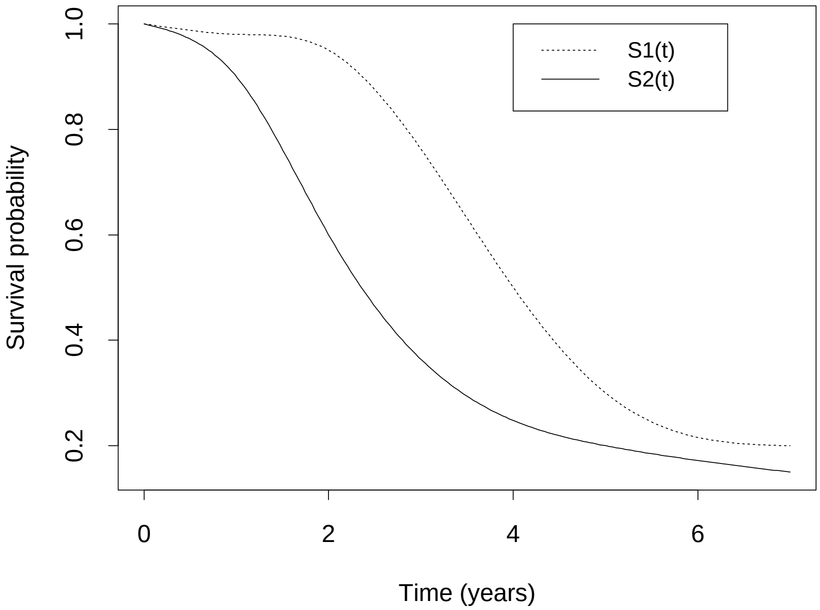
Figure 1: *The survival function for a hypothetical population*



For example, in the hypothetical population shown in Figure 1, we have a population where 70% of the individuals will survive 2 years (i.e., $t_{0.3} = 2$) and the median survival time is 2.8 years (*i.e.*, 50% of the population will survive at least 2.8 years).

We say that the survival distribution for group 1 is stochastically larger than the survival distribution for group 2 if $S_1(t) \geq S_2(t)$, for all $t \geq 0$, where $S_i(t)$ is the survival function for group i. If T_i is the corresponding survival time of a subject from group i, we also say that T_1 is stochastically (not deterministically) larger than T_2. Note that T_1 being stochastically larger than T_2 does NOT necessarily imply that $T_1 \geq T_2$. It implies that at any time point a greater proportion of group 1 will survive as compared to group 2 (see Figure 2 below).

Figure 2: *Illustration that T_1 is stochastically larger than T_2*



Hazard Rate

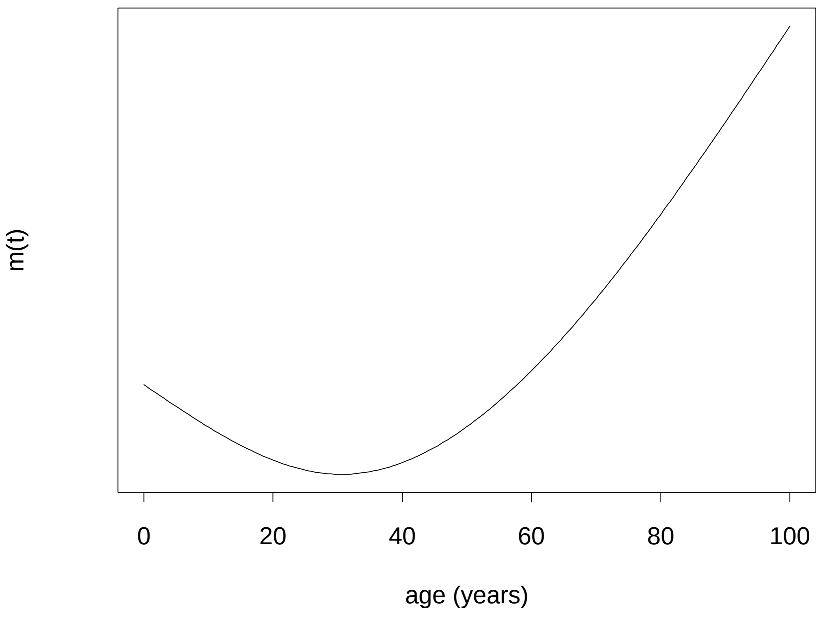
The hazard rate is a useful way of describing the distribution of "time-to-event" because it has a natural interpretation that relates to the aging of a population. This terminology is very popular in biomedical community. We motivate the definition of hazard rate by first defining **mortality rate** which is a discrete version of the hazard rate.

The **mortality rate** at time t, where t is generally taken to be an integer in terms of some unit of time (*e.g.*, years, months, days, etc), is the proportion of the population who fail (die) between times t and $t + 1$ **among** individuals **alive** at time t, *i.e.*,

$$m(t) = P[t \leq T < t + 1 | T \geq t]. \tag{8}$$

In a human population, the mortality rate has the typical pattern shown in Figure 3.

Figure 3: *A typical mortality pattern for human*



The hazard rate $\lambda(t)$ is the limit of a mortality rate if the interval of time is taken to be small (rather than one unit). The hazard rate is the instantaneous rate of failure (experiencing the event) at time t given that an individual is alive at time t.

Specifically, hazard rate $\lambda(t)$ is defined by the following equation

$$\lambda(t) = \lim_{h \to 0} \frac{P[t \le T < t+h | T \ge t]}{h}. \tag{9}$$

Therefore, if h is very small, we have

$$P[t \le T < t+h | T \ge t] \approx \lambda(t) h. \tag{10}$$

The definition of the hazard function implies that

$$\begin{aligned} \lambda(t) &= \frac{\lim_{h \to 0} \frac{P[t \le T < t+h]}{h}}{P[T \ge t]} = \frac{f(t)}{S(t)} && (11) \\ &= -\frac{S'(t)}{S(t)} = -\frac{d\log\{S(t)\}}{dt}. && (12) \end{aligned}$$

From this, we can integrate both sides to get

$$\Lambda(t) = \int_0^t \lambda(u)du = -\log\{S(t)\}, \tag{13}$$

where $\Lambda(t)$ is referred to as the cumulative hazard function. Here we used the fact that $S(0) = 1$. Hence,

$$S(t) = \exp\{-\Lambda(t)\} = \exp\{-\int_0^t \lambda(u)du\}. \tag{14}$$

Note:

1. There is a one-to-one relationship between hazard rate $\lambda(t), t \geq 0$ and survival function $S(t)$, namely,

$$S(t) = \exp\{-\int_0^t \lambda(u)du\} \quad \text{and} \quad \lambda(t) = -\frac{d\log\{S(t)\}}{dt}. \tag{15}$$

2. The hazard rate is NOT a probability, it is a probability rate. Therefore it is possible that a hazard rate can exceed one in the same fashion as a density function $f(t)$ may exceed one.