ST 745

Analysis of Survival Data

Lecture Notes

(Modified from Dr. A. Tsiatis' Lecture Notes)

Daowen Zhang

Department of Statistics

North Carolina State University

©2005 by Anastasios Tsiatis and Daowen Zhang

Contents

1 Survival Analysis

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

- 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.

To study, we must introduce some notation and concepts for describing the distribution of "time to event" for a population of individuals. Let the random variable T denote time to the event of our interest. Of course, T is a positive random variable which has to be unambiguously defined; that is, we must be very specific about the **start** and **end** with the length of the time period in-between corresponding to T.

Some examples

- Survival time (in general): measured from **birth** to **death** for an individual. This is the survival time we need to investigate in a life expectancy study.
- Survival time of a treatment for a population with certain disease: measured from the time of **treatment initiation** until **death**.
- Survival time due to heart disease: (the event is death from heart disease): measured from **birth** (or other time point such as treatment initiation for heart disease patients) to death caused by heart disease. (This may be a bit tricky if individuals die from other causes. This is competing risk problem. That is, other risks are competing with heart disease to produce an event – death.)

The time of interest may be time to something "good" happening. For example, we may be interested in how long it takes to eradicate an infection after treatment with antibiotics.

Describing the Distribution of Time to An 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 our data. One way is to estimate the underlying true distribution. 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 the random variable T can be described in a number of equivalent ways. There is of course the usual (cumulative) distribution function

$$F(t) = P[T \leq t], \quad t \geq 0, \tag{1.1}$$

which is **right** continuous, *i.e.*, $\lim_{u \to t^+} F(u) = F(t)$. When T is a survival time, $F(t)$ 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

$$f(t) = \frac{dF(t)}{dt}, F(t) = \int_0^t f(u)du. \tag{1.2}$$

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

$$S(t) = P[T \geq t] = 1 - F(t^-), \tag{1.3}$$

where $F(t^-) = \lim_{u \to t^-} F(u)$. When T is a survival time, $S(t)$ is the probability that a randomly selected individual will **survive** to time t or beyond. (So $S(t)$ has the name of **survival function**.)

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. 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. 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.

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

$$S(t) = \int_t^\infty f(u)du, \quad f(t) = -\frac{dS(t)}{dt}. \tag{1.4}$$

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:

$$\mathrm{E}(T) = \int_0^\infty S(t)dt. \tag{1.5}$$

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):

$$mrl(t_0) = \mathrm{E}[T - t_0 | T \geq t_0], \tag{1.6}$$

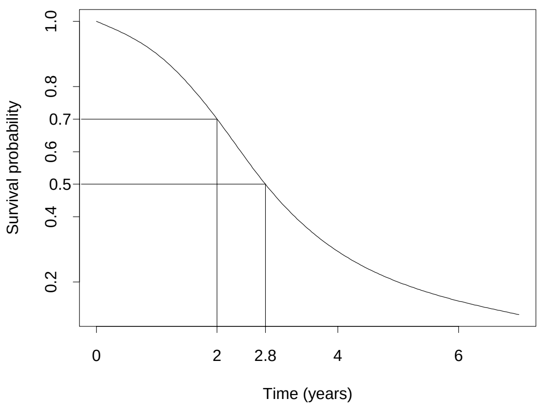
i.e., $mrl(t_0)$ = average remaining survival time **given** the population has survived beyond t_0. It can be shown that

$$mrl(t_0) = \frac{\int_{t_0}^\infty S(t)dt}{S(t_0)}. \tag{1.7}$$

For example, in the hypothetical population shown in Figure 1.1, 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).

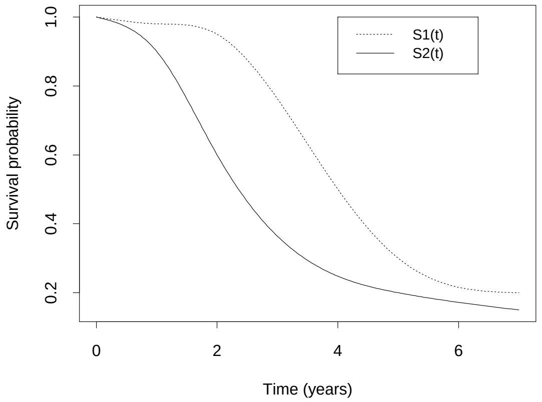
We say that the survival distribution for group 1 is stochastically larger than the survival distribution for group 2 if $S_1(t) \geq S_2(t)$, for all $t \geq 0$, where $S_i(t)$ is the survival function for group i. If T_i is the corresponding survival time for groups i, we also say that T_1 is stochastically (not deterministically) larger than T_2. Note that T_1 being stochastically larger than T_2 does NOT necessarily imply that $T_1 \geq T_2$. The situation is illustrated in Figure 1.2.

Figure 1.1: *The survival function for a hypothetical population*



Note: At any time point a greater proportion of group 1 will survive as compared to group 2.

Figure 1.2: *Illustration that T_1 is stochastically larger than T_2*



Hazard Rate

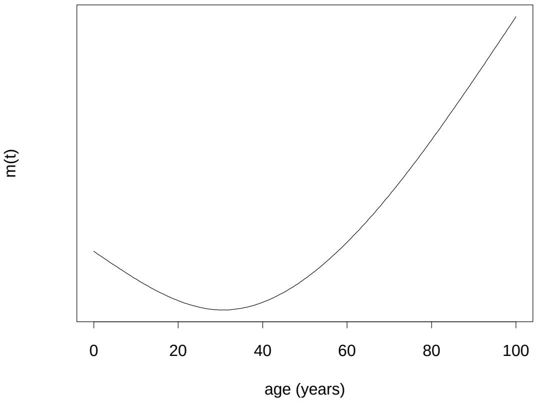
The hazard rate is a useful way of describing the distribution of "time to event" because it has a natural interpretation that relates to the aging of a population. This terminology is very popular in biomedical community. We motivate the definition of hazard rate by first defining **mortality rate**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

$$m(t) = P[t \leq T < t+1 | T \geq t]. \tag{1.8}$$

In a human population, the mortality rate has the typical pattern shown in Figure 1.3.

Figure 1.3: *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

$$\lambda(t) = \lim_{h \to 0} \frac{P[t \leq T < t+h | T \geq t]}{h}. \tag{1.9}$$

Therefore, if h is very small, we have

$$P[t \leq T < t+h | T \geq t] \approx \lambda(t) h. \tag{1.10}$$

The definition of the hazard function implies that

$$\lambda(t) \quad = \quad \frac{\lim_{h \to 0} \frac{P[t \leq T < t+h]}{h}}{P[T \geq t]} = \frac{f(t)}{S(t)} \tag{1.11}$$

$$= \quad -\frac{S'(t)}{S(t)} = -\frac{d\log\{S(t)\}}{dt}. \tag{1.12}$$

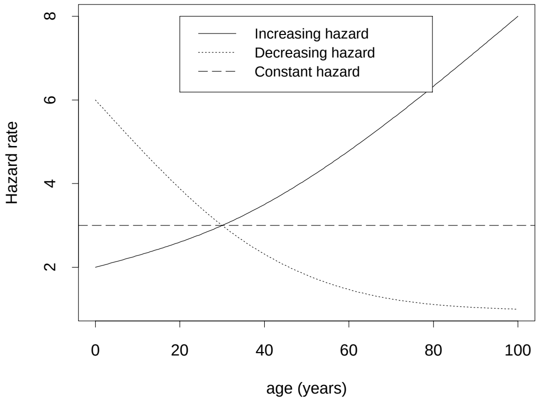
From this, we can integrate both sides to get

$$\Lambda(t) = \int_0^t \lambda(u)du = -\log\{S(t)\}, \tag{1.13}$$

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

$$S(t) = \mathrm{e}^{-\Lambda(t)} = \mathrm{e}^{-\int_0^t \lambda(u)du}. \tag{1.14}$$

Figure 1.4: *Three hazard patterns*



Note:

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

$$S(t) = \mathrm{e}^{-\int_0^t \lambda(u)du} \quad \text{and} \quad \lambda(t) = -\frac{d\log\{S(t)\}}{dt}. \tag{1.15}$$

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.

Common Parametric Models:

Distribution	$\lambda(t)$	$S(t)$	density $f(t)$	$\mathrm{E}(T)$
Exponential	$\lambda(>0)$	$e^{-\lambda t}$	$\lambda e^{-\lambda t}$	$\frac{1}{\lambda}$
Weibull	$\alpha\lambda t^{\alpha-1}(\alpha,\lambda>0)$	$e^{-\lambda t^{\alpha}}$	$\alpha\lambda t^{\alpha-1}e^{-\lambda t^{\alpha}}$	$\frac{\Gamma(1+1/\alpha)}{\lambda^{1/\alpha}}$
Gamma	$\frac{f(t)}{S(t)}$	$1-I(\lambda t,\beta)$	$\frac{\lambda^{\beta}t^{\beta-1}e^{-\lambda t}}{\Gamma(\beta)}$	$\frac{\beta}{\lambda}$

$I(t,\beta)=\int_0^t \frac{u^{\beta-1}e^{-u}}{\Gamma(\beta)}du.$

See page 38 of Klein and Moeschberger and Chapter 5 of the lecture notes for more distributions.

Exponential distribution: $\lambda(t)=\lambda$, $S(t)=e^{-\lambda t}$ and $f(t)=\lambda e^{-\lambda t}$. So **mean survival time**

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

Letting $S(t_{0.5})=e^{-\lambda t_{0.5}}=0.5$, then **median survival time** is $t_{0.5}=\frac{\log 2}{\lambda}$.

The **mean residual life time** after t_0 is

$$mrl(t_0)=\frac{\int_{t_0}^\infty S(t)dt}{S(t_0)}=\frac{\int_{t_0}^\infty \mathrm{e}^{-\lambda t}}{\mathrm{e}^{-\lambda t_0}}=\frac{1}{\lambda}=\mathrm{E}(T).$$

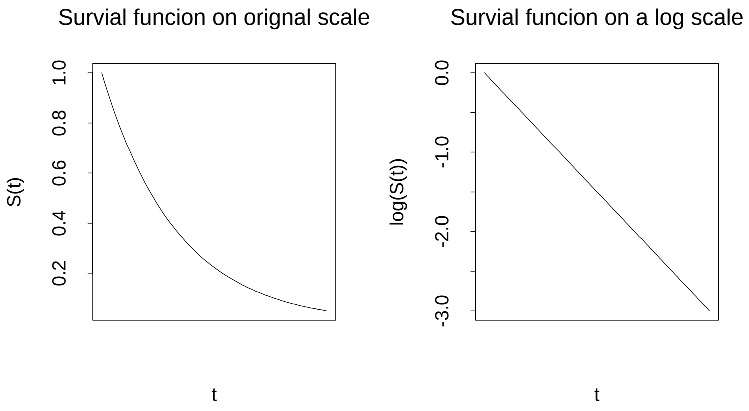
Sometimes it is useful to plot the survival distribution on a log scale. By so doing, we can identify the hazard rate as minus of the derivative of this function. In particular on a log scale the exponential distribution is a straight line. This is because $S(t)=e^{-\lambda t}$ for the exponential distribution, so

$$log[S(t)]=-\lambda t.$$

The above equation gives us a way to check if the underlying true distribution of the survival time is exponential or not given a data set. Suppose we can have an estimate $\hat{S}(t)$ of $S(t)$ without assuming any distribution of the survival time (the Kaplan-Meier estimate to be discussed in Chapter 2 is such an estimate). Then we can plot $log[\hat{S}(t)]$ vs t to see if it is approximately a straight line. A (approximate) straight line indicates that the exponential distribution may be a reasonable choice for the data.

Another alternative is to assume the exponential distribution for the data and get the estimate of $S(t) = e^{-\lambda t}$ (we only need to estimate λ; this kind of estimation will be discussed in Chapter 3). Denote this estimate by $\hat{S}_1(t)$ and Kaplan-Meier estimate by $\hat{S}_{KM}(t)$. If the exponential distribution assumption is correct, both estimates will be good estimates of the same survival function $S(t) = e^{-\lambda t}$. Therefore, $\hat{S}_1(t)$ and $\hat{S}_{KM}(t)$ should be close to each other and hence the plot $\hat{S}_1(t)$ vs $\hat{S}_{KM}(t)$ should be approximately a straight line. A non-straight line indicates that the exponential distributional assumption is not appropriate.

Figure 1.5: *The survival function of an exponential distribution on two scales*



Weibull distribution: $\lambda(t) = \alpha\lambda t^{\alpha-1}, S(t) = e^{-\lambda t^{\alpha}}$. Note this model allows:

- Constant hazard: $\alpha = 1$
- increasing hazard: $\alpha > 1$
- decreasing hazard: $\alpha < 1$.

and has the hazard patterns shown in Figure 1.4.

$$\mu = \mathrm{E}(T) = \int_0^\infty S(t)dt = \int_0^\infty \mathrm{e}^{-\lambda t^{\alpha}} dt = \frac{\Gamma(1+1/\alpha)}{\lambda^{1/\alpha}}.$$

The **mean survival time** $t_{0.5}$: $\mathrm{e}^{-\lambda t^{\alpha}} = 0.5 \Longrightarrow$

$$t_{0.5} = \left[\frac{\log 2}{\lambda}\right]^{1/\alpha}.$$

Since $\log S(t) = -\lambda t^{\alpha}$, so

$$\log\{-\log S(t)\} = \log\lambda + \alpha \log t.$$

A straight line in the plot of $\log\{-\log S(t)\}$ vs. $\log t$ indicates a Weibull model. We can use the above equation to check if the Weibull model is a reasonable choice for the survival time given a data set. Alternatively, we can assume a Weibull model for the survival time and use the data to estimate $S(t)$ and plot this estimate against the Kaplan-Meier estimate as we proposed for the exponential distribution. A (approximate) straight line indicates the Weibull model is a reasonable choice for the data.

Question: How do we check a Gamma model?