Survival analysis notes

Rob Trangucci

January 8, 2025

Contents

Chapter 1

Notation

Notation	Description
C_i	Random variable representing the time to censoring
$T_i = \min(X_i, C_i)$	Observable event time
$\delta_i = \mathbb{1}(T_i = X_i)$	Indicator variable equal to one if event time is a failure time
$P_\theta(X \leq t)$	Distribution function of X indexed by parameters θ
$S_X(t; \theta)$	Survival function for random variable X evaluated at t, parameters θ
$f_X(t; \theta)$	Density function for random variable X evaluated at t, parameters θ
$\lambda_X(t; \theta)$	Hazard function for random variable X evaluated at t, parameters θ

Table 1.1: List of notation used throughout the notes

Chapter 2

Introduction

This introduction is based in part on Klein, Moeschberger, et al. 2003, and in part on Aalen et al. 2008 plus Fleming and Harrington 2005.

Survival analysis is the modeling and analysis of time-to-event data; this means we will be studying how to model **nonnegative** random variables (time will always be measured in such a way so that the observations are nonnegative). Think about a clinical trial for a new COVID vaccine and how you might model the length of time between study entry and infection in each arm of the trial. Let X_i be the time from trial entry to infection for the i-th participant. These sorts of trials are typically run until a prespecified number of people have become infected. Let n be the total number of participants in the trial and let r be the prespecified number of infections. Let T_i be the observed infection time for the i-th participant. This means that for r participants, $T_i = X_i$, but for $n - r$ participants we know only that the time-to-infection is larger than the observed time. Let C_i denote the time from study entry for participant i to study end. Then $T_i = \min(X_i, C_i)$, and let $\delta_i = \mathbb{1}(T_i = X_i)$. The density of T_i is related to the joint probability for X_i and C_i, which is indexed by a possibly infinite dimensional parameter θ: $P_\theta(X_i > t, C_i > c)$. When $\delta_i = 1$, and $T_i = X_i$, the likelihood of the observation is

$$\left(-\frac{\partial}{\partial u} P_\theta(X_i > u, C_i > t)\right)\Big|_{u=t},$$

while the likelihood for $\delta_i = 0$ is

$$\left(-\frac{\partial}{\partial u} P_\theta(X_i > t, C_i > u)\right)\Big|_{u=t},$$

Then $T_i = C_i$ for the other $n - r$ participants. Under the null hypothesis that the vaccine has no effect, the population distribution function for all n participants for X_i, C_i is $P_\theta(X_1 > x, C_1 > c)$ (i.e. the distribution for survival times in the treatment group and the placebo

group is the same). Then the joint density for the observed infection times is as follows:

$$f_{T_1,\ldots,T_n}(t_1,\ldots,t_n;\theta) = n!\prod_{i=1}^{r}\left(-\frac{\partial}{\partial u}P_\theta(X_1 > u, C_1 > t_{(i)})\right)\Big|_{u=t_{(i)}}\prod_{i=r+1}^{n}\left(\left(-\frac{\partial}{\partial u}P_\theta(X_1 > t_{(i)}, C_1 > u)\right)\Big|_{u=t_{(i)}}\right),$$

where $t_{(i)}$ is the i-th order statistic of the set $\{t_1,\ldots,t_n\}$. Note that this is different from most other data analysis where missing observations are not expected to occur with much frequency. On the contrary, in survival analysis, missingness, both *truncation* and *censoring* are expected to occur with nearly every dataset, so much of our time will be spent ensuring our methods work when data arise with these peculiarities.

2.1 Independent censoring

Now suppose that $X_1 \perp\!\!\!\perp C_1$, and that θ partitions into η and ϕ, such that

$$P_\theta(X_1 > x, C_1 > c) = P_\eta(X_1 > x)P_\phi(C_1 > c).$$

Then we can rewrite the joint observational density for T_i as:

$$\begin{aligned} f_{T_1,\ldots,T_n}(t_1,\ldots,t_n;\theta) = n!&\left(\prod_{i=1}^{r} f_{X_1}(t_{(i)};\eta)\right)\prod_{i=r+1}^{n} P_\eta(X_1 > t_{(i)}) \\ &\times\left(\prod_{i=1}^{r} P_\phi(C_1 > t_{(i)})\right)\prod_{i=r+1}^{n} f_C(t_{(i)};\phi). \end{aligned}$$

If we are only interested about inference about η, the parameters that govern the distribution of the true time-to-infection random variables, we can ignore the the distribution for the censoring random variables C_1, and maximize the likelihood because, in η:

$$f_{T_1,\ldots,T_n}(t_1,\ldots,t_n;\eta) \propto \left(\prod_{i=1}^{r} f_{X_1}(t_{(i)};\eta)\right)\prod_{i=r+1}^{n} P_\eta(X_1 > t_{(i)})$$

We will talk in more detail about censoring in the coming lectures.

2.2 Mean time to failure

Aalen et al. 2008 notes that we cannot even compute a simple mean in this situation, so something like a t-test will be useless. As an aside, let's try to compute a mean from the data above. Let $\bar{T} = \frac{1}{n}\sum_{i=1}^{n} T_i$. We can show that $\lim_{n\to\infty}\bar{T} \le \mathbb{E}[X_i]$ with probability 1.

Proof. Let $T_i = X_i\mathbb{1}(X_i \le C_i) + C_i\mathbb{1}(X_i > C_i)$. Then by the SLLN $\bar{T} \overset{\text{a.s.}}{\to} \mathbb{E}[T_i]$.

$$\begin{aligned} \mathbb{E}[T_i] &= \mathbb{E}[X_i\mathbb{1}(X_i \le C_i)] + \mathbb{E}[C_i\mathbb{1}(X_i > C_i)] \\ &\le \mathbb{E}[X_i\mathbb{1}(X_i \le C_i)] + \mathbb{E}[X_i\mathbb{1}(X_i > C_i)] = \mathbb{E}[X_i] \end{aligned}$$

□

2.3 Survival function

How can we compute the mean time to infection then? One way to estimate the mean time to infection is to first estimate the function $S_{X_i}(t;\theta) = P_\theta(X_i > t)$, which is also known as the *survival function*. Recall this fact about non-negative random variables $X_i \geq 0$ w.p. 1:

$$\mathbb{E}[X_i] = \int_0^\infty P_\theta(X_i > t)dt$$

This follows from an application of Fubini's theorem applied to the integral:

$$\begin{aligned}
\mathbb{E}[X_i] &= \int_0^\infty u dP_{X_i}(u;\theta) \\
&= \int_0^\infty \int_0^\infty \mathbb{1}(0 \leq t \leq u)\, dt\, dP_{X_i}(u;\theta) \\
&= \int_0^\infty \int_0^\infty \mathbb{1}(0 \leq t \leq u)\, dP_{X_i}(u;\theta)dt \\
&= \int_0^\infty P_\theta(X_i > t)dt
\end{aligned}$$

2.3.1 Properties of the survival function

Let $F_{X_i}(t;\theta) = P_\theta(X_i \leq t)$. Then because the survival function is defined as $S_{X_i}(t;\theta) = 1 - F_{X_i}(t;\theta)$ (also known as the complementary CDF) the survival function inherits its properties from the CDF. The survival function:

1. $S_{X_i}(t;\theta)$ is a nonincreasing function
2. $S_{X_i}(0;\theta) = 1$
3. $\lim_{t\to\infty} S_{X_i}(t;\theta) = 0$
4. Has lefthand limits:
$$\lim_{s \nearrow t} S_{X_i}(s;\theta) = S_{X_i}(t-;\theta).$$
5. Is right continuous:
$$\lim_{s \searrow t} S_{X_i}(s;\theta) = S_{X_i}(t;\theta).$$

An example of a discrete survival function is shown in Figure 2.1.

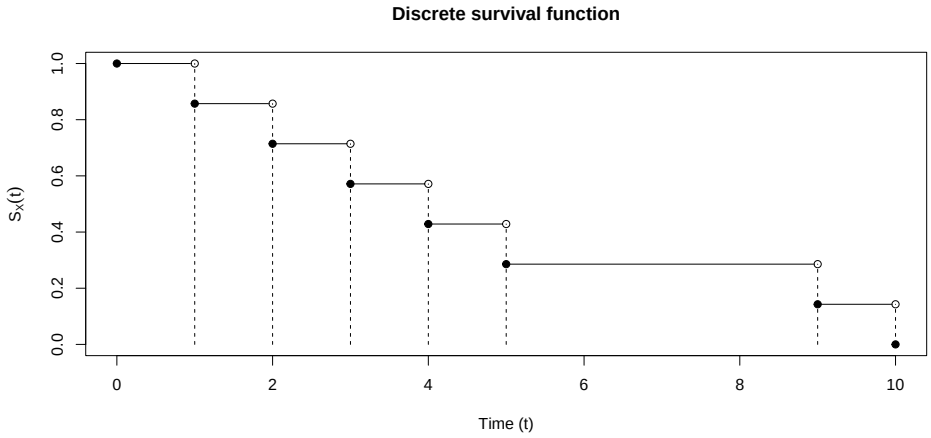

Figure 2.1: Example plot of a survival function for a discrete survival time, bounded between $[0, 10]$

2.4 Hazard function

Another way to characterize the random variable X_i is the *hazard function*, which is typically denoted as $\lambda(t)$ or $h(t)$ and is defined as

$$\begin{aligned}\lambda_{X_i}(t) &= \lim_{\Delta t \searrow 0} \frac{1}{\Delta t} \mathbb{P}_\theta \left(t \leq X_i < t + \Delta t \mid X_i \geq t\right) \\ &= \lim_{\Delta t \searrow 0} \frac{1}{\Delta t} \frac{\mathbb{P}_\theta \left(t \leq X_i < t + \Delta t\right)}{\mathbb{P}_\theta \left(X_i \geq t\right)}\end{aligned}$$

First, note that we can define $\mathbb{P}_\theta \left(X_i \geq t\right)$ in terms of the survival function as:

$$\mathbb{P}_\theta \left(X_i \geq t\right) = \lim_{s \nearrow t} S_{X_i}(s; \theta).$$

Using the notation introduced in Section 2.3.1, we can write this as

$$\mathbb{P}_\theta \left(X_i \geq t\right) = S_{X_i}(t-; \theta).$$

Of course, when X_i is absolutely continuous,$S_{X_i}(t-; \theta) = S_{X_i}(t; \theta)$, but when X_i is discrete, or mixed discrete and continuous, as noted above, it is not true in general that the survival function is left-continuous.

A few things to note about $\lambda_{X_i}(t; \theta)$: when X_i is an absolutely continuous random variable, which occurs when we're considering survival in continuous time, we can write this in terms of the probability density function $f_{X_i}(t; \theta)$ and the cumulative distribution function $F_{X_i}(t; \theta)$:

$$\begin{aligned}\lambda_{X_i}(t) &= \lim_{\Delta t \searrow 0} \frac{1}{\Delta t} \frac{\mathbb{P}_\theta \left(t \leq X_i < t + \Delta t\right)}{\mathbb{P}_\theta \left(X_i \geq t\right)} \\ &= \lim_{\Delta t \searrow 0} \frac{F_{X_i}(t + \Delta t; \theta) - F_{X_i}(t; \theta)}{\Delta t} \times \frac{1}{1 - F_{X_i}(t; \theta)} \\ &= \frac{f_{X_i}(t; \theta)}{1 - F_{X_i}(t; \theta)}.\end{aligned}$$

Let's examine how the survival function and the hazard function fit together.

$$\lambda_{X_i}(t) = \frac{f_{X_i}(t; \theta)}{S_{X_i}(t-; \theta)}.$$

Note that we can write the hazard function in terms of the survival function instead of the density, when X_i is absolutely continuous:

$$\begin{aligned}\lambda_{X_i}(t) &= \lim_{\Delta t \searrow 0} \frac{1}{\Delta t} \frac{\mathbb{P}_\theta \left(t \leq X_i < t + \Delta t\right)}{\mathbb{P}_\theta \left(X_i \geq t\right)} \\ &= \lim_{\Delta t \searrow 0} \frac{S_{X_i}(t; \theta) - S_{X_i}(t + \Delta t; \theta)}{\Delta t} \times \frac{1}{S_{X_i}(t; \theta)} \\ &= -\frac{d}{dt} S_{X_i}(t; \theta) / S_{X_i}(t; \theta).\end{aligned}$$

This implies that

$$\lambda_{X_i}(t) = -\frac{d}{dt}\log S_{X_i}(t;\theta).$$

If we integrate both sides, we get another important identity in survival analysis:

$$\int_0^u \frac{d}{dt}\log S_{X_i}(t;\theta)dt = -\int_0^u \lambda_{X_i}(t)dt \tag{2.1}$$

$$\log S_{X_i}(u;\theta) - \log S_{X_i}(0;\theta) = -\int_0^u \lambda_{X_i}(t)dt \quad \text{note } S_{X_i}(0;\theta) = 1 \tag{2.2}$$

$$S_{X_i}(u;\theta) = \exp\left(-\int_0^u \lambda_{X_i}(t)dt\right) \tag{2.3}$$

2.4.1 Properties of the hazard function

The relationship $S_{X_i}(u;\theta) = \exp\left(-\int_0^u \lambda_{X_i}(t)dt\right)$ and the properties of the survival function reveal the following facts about the hazard function and highlight its differences with a probability density.

1. $\lim_{t\to\infty} S_{X_i}(t;\theta) = 0$ implies that $\lim_{t\to\infty}\int_0^t \lambda_X(u)du = \infty$
2. Given that $S_{X_i}(t;\theta)$ is a nonincreasing function, $\lambda_X(t) \geq 0$ for all t.

So unlike a probability density function, $\lambda_X(t)$ isn't integrable over the support of the random variable.

2.5 Density function for survival time

Given that we have $S_{X_i}(t;\theta)$ and $\lambda(t) = \frac{f_{X_i}(t;\theta)}{S_{X_i}(t-;\theta)}$, we can recover the density, $f_{X_i}(t;\theta)$ easily:

$$f_{X_i}(t;\theta) = \lambda_{X_i}(t)S_{X_i}(t;\theta)$$

2.6 Cumulative hazard function

One final important quantity that describes a survival distribution is that of *cumulative hazard*, which we'll denote as $\Lambda_{X_i}(t)$, though it is also denoted as $H(t)$ in Klein, Moeschberger, et al. 2003. This is defined as you might expect:

$$\Lambda_{X_i}(t) = \int_0^t \lambda_{X_i}(u)du.$$

It has the important property that for any absolutely continuous failure time X_i with a given cumulative hazard function, the random variable $Y_i = \Lambda_{X_i}(X_i)$ is exponentially distributed

with rate 1. The derivation is straightforward. Remember that $P(X_i > t) = \exp\left(-\Lambda_{X_i}(t)\right)$

$$\begin{aligned} P(\Lambda_{X_i}(X_i) > t) &= P(X_i > \Lambda_{X_i}^{-1}(t)) \\ &= \exp\left(-\Lambda_{X_i}(\Lambda_{X_i}^{-1}(t))\right) \\ &= \exp\left(-t\right) \end{aligned}$$

2.7 Discrete survival time

We've been working with continuous survival times until now. If X_i is a discrete random variable with support on $\{t_1, t_2, \dots\}$, we lose some of the tidyness of the previous derivations. We can define the distribution of X_i in terms of the survival function, $P_\theta(X_i > t)$. First let $p_j = P_\theta(X_i = t_j)$, so

$$S_{X_i}(t;\, \theta) = P_\theta(X_i > t) = \sum_{j|t_j > t} p_j$$

We can also define the hazard function for a discrete random variable:

$$\lambda_{X_i}(t_j) = \frac{p_j}{S_{X_i}(t_{j-1};\, \theta)} = \frac{p_j}{p_j + p_{j+1} + \dots}$$

Note that $p_j = S_{X_i}(t_{j-1};\, \theta) - S_{X_i}(t_j;\, \theta)$, then

$$\lambda_{X_i}(t_j) = 1 - \frac{S_{X_i}(t_j;\, \theta)}{S_{X_i}(t_{j-1};\, \theta)}.$$

If we let $t_0 = 0$ then $S_{X_i}(t_0;\, \theta) = 1$. This allows us to write the survival function in a sort of telescoping product:

$$\begin{aligned} P_\theta(X_i > t_j) &= P_\theta(X_i > t_0)\frac{P_\theta(X_i > t_1)}{P_\theta(X_i > t_0)}\frac{P_\theta(X_i > t_2)}{P_\theta(X_i > t_1)} \cdots \frac{P_\theta(X_i > t_j)}{P_\theta(X_i > t_{j-1})} \\ &= 1\frac{S_{X_i}(t_1;\, \theta)}{S_{X_i}(t_0;\, \theta)}\frac{S_{X_i}(t_2;\, \theta)}{S_{X_i}(t_1;\, \theta)} \cdots \frac{S_{X_i}(t_j;\, \theta)}{S_{X_i}(t_{j-1};\, \theta)} \end{aligned}$$

This yields another way to write $S_{X_i}(t;\, \theta)$:

$$S_{X_i}(t;\, \theta) = \prod_{j|t_j \le t} (1 - \lambda_{X_i}(t_j)). \tag{2.4}$$

It turns out that we can write the survival function for continuous random variables in the same way.

2.7.1 Connection between discrete and continuous survival functions

Recall the definition of the hazard function:

$$\lambda_{X_i}(t) = \lim_{\Delta t \searrow 0} \frac{1}{\Delta t} \mathbb{P}_\theta \left(t \leq X < t + \Delta t \mid X \geq t \right)$$

Note that $\lambda_{X_i}(t)\,\Delta t$ is approximately $\mathbb{P}_\theta \left(t \leq X < t + \Delta t \mid X \geq t \right)$. Let $\mathcal{T}$ be a partition of $(0, \infty)$ with partition size Δt, $t_0 = 0$:

$$\mathcal{T} = \bigcup_{j=0}^{\infty} [t_j, t_j + \Delta t).$$

Then we can use Equation (2.4) to represent the survival function:

$$S_{X_i}(t;\, \theta) = \prod_{j | t_j + \Delta t \leq t} (1 - \lambda_{X_i}(t_j) \Delta t). \tag{2.5}$$

We can show that as the partition of the time domain gets finer and finer, we will recover $S_{X_i}(t;\, \theta) = \exp(-\int_0^t \lambda_{X_i}(u) du)$

$$S_{X_i}(t;\, \theta) = \prod_{j \in \mathcal{T} | t_j + \Delta t \leq t} (1 - \lambda_{X_i}(t_j) \Delta t) \tag{2.6}$$

$$\log S_{X_i}(t;\, \theta) = \sum_{j \in \mathcal{T} | t_j + \Delta t \leq t} \log(1 - \lambda_{X_i}(t_j) \Delta t) \tag{2.7}$$

We use the Taylor expansion of $\log(1 - \lambda_{X_i}(t_j)\Delta t)$ for small $\lambda_{X_i}(t_j)\Delta t$, assuming that $\lambda_{X_i}(t)$ is sufficiently well-behaved for all t.

$$\log(1 - \lambda_{X_i}(t_j) \Delta t) \approxeq -\lambda_{X_i}(t_j) \Delta t.$$

Then

$$\log S_{X_i}(t;\, \theta) \approxeq \sum_{j \in \mathcal{T} | t_j + \Delta t \leq t} -\lambda_{X_i}(t_j) \Delta t \tag{2.8}$$

As

$$\lim_{\Delta t \searrow 0} \sum_{j \in \mathcal{T} | t_j + \Delta t \leq t} -\lambda_{X_i}(t_j) \Delta t = -\int_0^t \lambda_{X_i}(u) du.$$

So, $S_{X_i}(t;\, \theta) = \exp(-\int_0^t \lambda_{X_i}(u) du)$, or

$$S_{X_i}(t;\, \theta) = \exp(-\lambda_{X_i}(t)) \tag{2.9}$$

Bibliography

[1] John P Klein, Melvin L Moeschberger, et al. *Survival analysis: techniques for censored and truncated data*. Vol. 1230. Springer, 2003.

[2] Odd Aalen, Ornulf Borgan, and Hakon Gjessing. *Survival and event history analysis: a process point of view*. Springer Science & Business Media, 2008.

[3] Thomas R Fleming and David P Harrington. "Counting Processes and Survival Analysis". In: *Wiley Series in Probability and Statistics* (2005).