Survival analysis notes

Rob Trangucci

January 6, 2025

Contents

Chapter 1

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. 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 $P_\theta(X_1 > x, C_1 > c)$. Then the joint density for the observed infection times is as follows:

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

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.

1.1 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:

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

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:

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

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

1.2 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 $\lim_{n \to \infty} \bar{T} \leq \mathbb{E}[X_i]$ with probability 1.

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

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

□

1.3 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 $S_{X_i}(t) = P(X_i > t)$, which is also known as the *survival function*. Recall this fact about non-negative random variables $X_i \geq 0$ w.p. 1:

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

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

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

1.3.1 Properties of the survival function

Given that the survival function is defined as $S_{X_i}(t) = 1 - F_{X_i}(t)$ (also known as the complementary CDF) the survival function inherits its properties from the CDF. The survival function:

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

An example of a discrete survival function is shown in Figure 1.1.

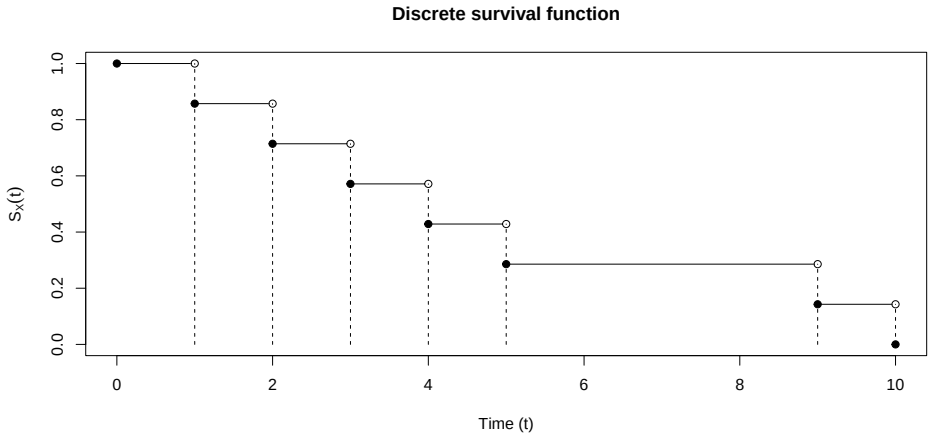

Figure 1.1: Example plot of a survival function for a discrete survival time, bounded between $[0, 10]$

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).