Notation and Quantities of Interest

Survival analysis is concerned with estimating the time until an event (of interest) occurs. For a population of individuals, $i = 1, \dots, n$ we are interested in T_i which is defined to be *the event time*. Most of our interest is going to be centered on characterizing the distribution of T_i. The exact method of doing this is going to depend on whether we are interested in treating T_i as though it is a **continuous** or a **discrete** random variable. The nature of survival data is such that, often, the event of interest is not observed in the study. Suppose that, for each individual, there is an observation window which we denote $(0, C_i]$. Here C_i can either be a fixed quantity, known in advance, or it can be a random quantity. If the event time, T_i, occurs outside of the window of observation, $(0, C_i]$, then we say that the event is censored.

If the event occurs after the observation window has ended ($T_i > C_i$) then we say that the event is *right censored*. This might occur, for instance, if the event of interest is death and we only observe patients for 10 years. It might also occur, for instance, if the event of interest is death due to cancer, and a patient dies of a heart attack during our observation window. If the event occurs before the observation window has begun ($T_i < 0$) then we say that the event is *left censored*. Note that $t = 0$ represents a well-defined origin (for instance, the start of the study), and so left censoring just refers to an individual who had their event occur *before* the start of the study.

In any event, we define $X_i = \min(T_i, C_i)$ and state that for each individual we observe X_i. We also define $\delta_i = I(T_i \leq C_i)$ to represent an observation indicator.

Discrete Time to Event Data

Suppose that, if $T_i \in [k, k+1)$ we assign it the value of k, for integer k. In this case we have discretized the time space, and we are dealing with *discrete time-to-event* data. We can think about characterizing the distribution of T_i, and generally use one of four quantities to do so.

1. **The (discrete-time) Hazard Function:** This is defined as
$$h(k) = P(T = k | T \geq k).$$
This represents the probability that the event will occur at the instant k, supposing we know that it has lasted until at least k.

2. **The Survivor Function:** This is defined as
$$S(k) = P(T > k) = 1 - P(T \leq k) = \prod_{s=1}^{k} P(T > s | T \geq s) = \prod_{s=1}^{k} (1 - h(s)).$$
The survivor function tells you the (unconditional) probability that the event happens *after* a specific time.

3. **The Mass Function:** This is defined as
$$P_T(k) = P(T = K),$$
which is the probability mass function that you are used to!

4. **The Distribution Function:** This is defined as

$$F_T(k) = P(T \leq k) = 1 - S(k).$$

This is the standard CDF that you are used to.

Each of these quantities will **uniquely** define the survival distribution, and working from one we can get the others. This should be clear based on the defining relationships. We already know that $F_T(k) = \sum_{t=0}^{k} P_T(t)$, and conversely that $P_T(t) = F_T(t) - F_T(t-1)$. To move between the survival function and the CDF, we simply take the complement $F_T(k) = 1 - S(k)$. The Hazard function gives the survival function, through the defining quantity, which also means that $h(k) = 1 - \frac{S(k)}{S(k-1)}$.

We will most often focus on the **hazard function** and the **survival function**, but you should feel comfortable with all four of these quantities.

Continuous Time to Event Data

While the times were discretized in the discrete case, we can also think of treating T_i (or C_i) as continuous valued. In this sense it would no longer make sense to consider the mass function (since $P(T_i = k) = 0$ for all k). Instead, we want to build continuous analogues for all of the aforementioned quantities.

1. **The (continuous-time) Hazard Function:** This is defined as

$$h(k) = \lim_{\Delta k \to 0} \frac{P(k \leq T \leq k + \Delta k | T \geq k)}{\Delta k}.$$

This represents the probability that the event will occur at the instant k, supposing we know that it has lasted until at least k, just as it did in the discrete-time case. The limit is necessary to ensure it is well-defined.

2. **The Survivor Function:** This is defined as

$$S(k) = P(T > k) = 1 - F_T(k).$$

The survivor function tells you the (unconditional) probability that the event happens *after* a specific time.

3. **The Density Function:** This is defined as

$$f_T(k) = \frac{\partial}{\partial k} F(k) = \lim_{\Delta k \to 0} \frac{P(k \leq T < k + \Delta k)}{\Delta k},$$

which is the probability density function that you are used to!

4. **The Distribution Function:** This is defined as

$$F_T(k) = P(T \leq k) = 1 - S(k).$$

This is the standard CDF that you are used to.

Each of these quantities will **uniquely** define the survival distribution, and working from one we can get the others.

To make this clear, first note that starting from the CDF (which we know defines a distribution uniquely) we get $S(t) = 1 - F_T(t)$ and $f_T(t) = F_T'(t)$. Similarly, we can integrate the density function to obtain the CDF, and take the complement to get the survival function, or work from the survival function, taking the complement to get the CDF. All that remains then is connecting these quantities to the hazard function.

Note that since the hazard function is defined based on a conditional probability, we can expand this to be

$$\begin{aligned} h(t) &= \lim_{\Delta t \to 0} \frac{P(t \leq T \leq t + \Delta t | T \geq t)}{\Delta t} \\ &= \lim_{\Delta t \to 0} \frac{P(t \leq T \leq t + \Delta t, T \geq t)}{P(T \geq t)\Delta t} \\ &= \lim_{\Delta t \to 0} \frac{P(t \leq T \leq t + \Delta t)}{\Delta t} \times \frac{1}{S_T(t)} \\ &= \frac{f_T(t)}{S_T(t)}. \end{aligned}$$

If we note that $S_T'(t) = -F_T'(t) = -f_T(t)$, then we see that

$$-\frac{\partial}{\partial t} \log S(t) = \frac{-S'(t)}{S(t)} = h(t).$$

This relationship can be inverted, as well, giving

$$S(t) = \exp\left(-\int_0^t h(s)ds\right).$$

Of course, these connections allow us to connect to the CDF and density as well, following through the same processes as above.

The Cumulative Hazard Function

Note, the term given by $\int_0^t h(s)ds$ is frequently of directly interest. As a result, we will call this the **cumulative hazard function** and denote it

$$H(t) = \int_0^t h(s)ds.$$

As a result, we can say that $S(t) = \exp(-H(t))$ more concisely.