18.650: Fundamentals of Statistics November 27, 2023

Lecture 27 — Survival analysis

Prof. Philippe Rigollet *Scribe: Anya Katsevich*

1 Overview

Survival analysis is about estimating the probability distribution of time until failure, or more generally, time until some event of interest happens.

As a motivating example, consider the duration of Zoom subscriptions started in April 2020. 500 participants are observed over the 40 months since then. We want to understand the statistics of subscription duration, i.e. time till cancellation. Below is what our data might look like for 8 Zoom users.

# months since subs.	cancelled
5	1
12	1
7	1
35	1
40	0 (censored)
6	1
40	0 (censored)
40	0 (censored)

The left column lists the number of months until cancellation for 8 users. However, 3 of the users are at 40 months, because they haven't yet cancelled! Therefore, the information about the time till cancellations is considered "censored" for these three users — we don't actually know how much longer they'll use the subscription before cancelling, only that they've used it for 40 months so far.

Here are some other more common but less fun examples of this kind of data.

- rats given cancer (how long until they die?)
- patients treated with a drug (how long until they recover?)
- engines running in cold temperatures (how long until failure?)

The main question is how to make the most of all *the data we have, including the partial observations from the censored data.*

2 Basic concepts

Definition 2.1: Failure time & survival function

Let $T \geq 0$ be a random variable denoting failure time, or time until the event of interest. Let $F(t)$ be the cdf of T. Then the *survival function* $S(t)$ is

$$S(t) = \mathbb{P}(T > t) = 1 - F(t).$$

In the Zoom example, T is the time until cancellation.

Goal: estimate $S(t)$.

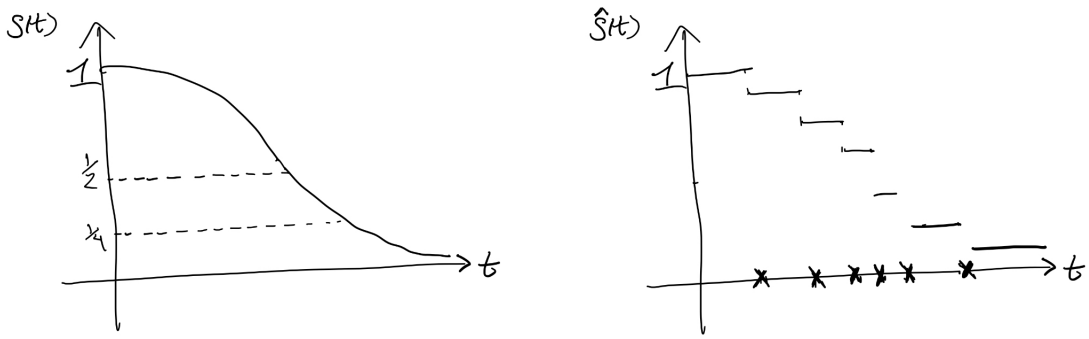

Figure 1: The lefthand curve depicts a survival function $S(t)$. We might be interested in the first time the survival probability dips below $1/2$ or $1/4$, for example. The righthand curve depicts what an estimator $\hat{S}(t)$ of $S(t)$ looks like: a step function, whose value changes at the observed times (the x's on the t axis).

Suppose we observe i.i.d. draws $T_1, \ldots, T_n \overset{\text{i.i.d.}}{\sim} T$. The plug-in estimator of S is just

$$\hat{S}(t) = \frac{1}{n} \sum_{i=1}^{n} \mathbb{1}(T_i > t). \tag{1}$$

This is the same as one minus the empirical cdf. The estimator $\hat{S}$ has the following properties:

- unbiased
- consistent
- asymptotically normal: $\sqrt{n}(\hat{S}_n(t) - S(t)) \rightsquigarrow \mathcal{N}(0, S(t)(1 - S(t)))$.

Using the estimator $\hat{S}$ would be the right approach if we did actually get to observe $T_1, \ldots, T_n$. But we do not get to observe these random variables. For example, in the Zoom study, we only get to observe

$$\tilde{T}_i = \min(T_i, 40) < T_i, \quad i = 1, \ldots, n$$

More generally, the censoring times could be different. For example, consider a study of recovery times in COVID-positive participants. The participants are asked every day whether or not they have tested negative. A participant could be negligent and stop reporting their test status after a certain number of days. The drop-out time would then be specific to that participant.

Definition 2.2: Censoring time and censored random variables

We let C_i denote the censoring time of sample i, so that our observations are given by

$$\tilde{T}_i = \min(T_i, C_i).$$

If observation i is never censored, set C_i to be infinity.

In the COVID study, the uncensored observations correspond to the recovery times of the participants who finally tested negative and reported it to the study.

2.1 Simple estimators which account for censoring

The naive alternative to $\hat{S}(t)$ from (1) is to construct an estimator by throwing out all censored data — in the COVID example, this means throwing out the data from all participants who dropped out at any point.

$$\hat{S}^{\text{naive}}(t) = \frac{\sum_{i=1}^n \mathbb{1}(\tilde{T}_i > t, C_i = \infty)}{\sum_{i=1}^n \mathbb{1}(C_i = \infty)} = \frac{\#\text{COVID-negative after} > t \text{ days \& never dropped out}}{\#\text{never dropped out}}.$$

A better option is to include censored data in the estimator, as long as the censoring time occurs *after* t.

$$\hat{S}^{\text{better}}(t) = \frac{\sum_{i=1}^n \mathbb{1}(\tilde{T}_i > t)}{\sum_{i=1}^n \mathbb{1}(C_i > t)} = \frac{\#\text{COVID-negative after} > t \text{ days}}{\#\text{dropped out later than } t \text{ (or never dropped)}}. \tag{2}$$

$\hat{S}^{\text{better}}$ is better because it uses more of the data. In the next section, we'll see an even more advanced estimator that takes advantage of more of the data.

3 The Kaplan-Meier estimator

Assume for simplicity $t = 0, 1, 2, \ldots$ (discrete time). Write

$$\begin{aligned} S(t) &= \mathbb{P}(T > t) = \mathbb{P}(T > t | T > t-1)\mathbb{P}(T > t-1) \\ &= (1 - \mathbb{P}(T \leq t | T > t-1))\mathbb{P}(T > t-1) \\ &= \underbrace{(1 - \mathbb{P}(T = t | T > t-1))}_{q(t)} \underbrace{\mathbb{P}(T > t-1)}_{S(t-1)} \end{aligned} \tag{3}$$

We can recursively apply the relationship $S(t) = q(t)S(t-1)$ to get

$$S(t) = q(t)q(t-1)\ldots q(1)q(0) = \prod_{s=0}^{t} q(s)$$

where $q(0) = 1 - \mathbb{P}(T = 0 | T > -1) = 1 - \mathbb{P}(T = 0)$.
The function $q(s)$ is given by 1 minus the so-called hazard rate:

Definition 3.1: Hazard rate

$$h(s) = \mathbb{P}(T = s | T > s-1) = \frac{\mathbb{P}(T = s)}{\mathbb{P}(T \geq s)}$$

is the hazard rate.

We will now estimate $h(s)$ by $\hat{h}(s)$, then set $\hat{q}(s) = 1 - \hat{h}(s)$, and then finally define the Kaplan-Meier estimator to be

$$\hat{S}(t) = \hat{q}(t)\hat{q}(t-1)\ldots\hat{q}(0).$$

To estimate $h(s)$, we take

$$\hat{h}(s) = \frac{\sum_{i=1}^{n} \mathbb{1}(\tilde{T}_i = s, C_i > s)}{\sum_{i=1}^{n} \mathbb{1}(\tilde{T}_i \geq s)}, \tag{4}$$

and then

$$\hat{q}(s) = 1 - \hat{h}(s).$$

Definition 3.2: Kaplan-Meier estimator

The Kaplan-Meier estimator is

$$\hat{S}(t) = \hat{q}(t)\hat{q}(t-1)\dots\hat{q}(0),$$

where $\hat{q}(s) = 1 - \hat{h}(s)$ and $\hat{h}(s)$ is given in (4). Explicitly,

$$\hat{S}(t) = \prod_{s=0}^{t} \left(1 - \frac{\#\{i : \tilde{T}_i = s, C_i > s\}}{\#\{i : \tilde{T}_i \geq s\}}\right).$$

The advantage of the Kaplan-Meier estimator is that in estimating the survival time past t, *it successfully incorporates observations that were censored before time t!* This is thanks to the product structure. In contrast, the better but still naive estimator $\hat{S}^{\text{better}}$ from (2) only uses observation that were censored *after time t*.

3.1 Variations.

We often want to know how the survival function $S(t)$ depends on certain features $x \in \mathbb{R}^k$. For example, to estimate the probability a person will live past age t, it makes sense to take their health metrics into account. So the function now becomes $S(t, x)$: what is probability of survival past t based on your specific x?

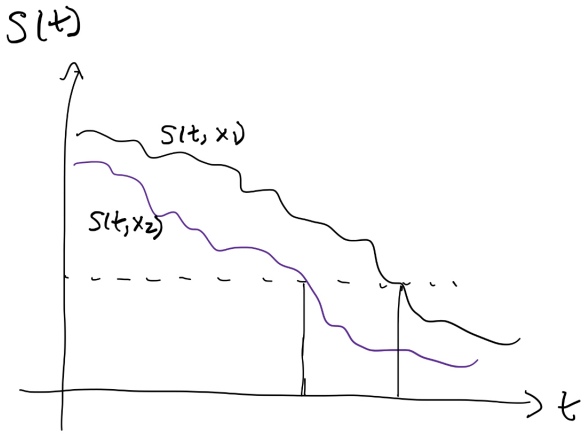

Figure 2: Two survival curves for two different feature vectors x_1 and x_2

The Cox proportional hazard regression model for the function h is

$$h(t, x) = h_0(t) \exp(\beta^T x).$$

It decouples the effect of x and t.

There is also a continuous time version of the Kaplan-Meier estimator:

$$\text{discrete time: } h(t) = \frac{\mathbb{P}(T = t)}{\mathbb{P}(T > t - 1)}$$

$$\text{cts. time: } h(t) = \frac{\mathbb{P}(t \leq T \leq t + dt)}{\mathbb{P}(T \geq t)} = \frac{f(t)}{S(t)} = \frac{-S'(t)}{S(t)} = -\frac{d}{dt}\log(S(t)).$$