Survival Analysis

Davi Sales Barreira

April 11, 2022

Contents

List of Definitions

List of Theorems

List of Examples

1 Introduction

Survival Analysis is about studying the time until an event occurs. Usually, one will be interested in predicting such time, for example, how long will it take for case to be given a sentence in the judiciary system.

An important concept in Survival Analysis is the idea of Censoring. This refers to data in which the information is unknown for some reason. For example, suppose that for some reason a case is dropped by the judiciary before the sentence. Thus, the output for such case was "censored" since we can't observe when it would've occur.

In this scenario, there are three types of censoring, the right censoring, which we described in the example, where the occurrence of the event is unseen. The left censoring occurs when the starting time is unknown. Consider for example that we know that a case entered the justice system in January, but we don't know which day. Lastly, interval censoring is, for example, the case in which we can only know the monthly information, so we know the case entered in January, and it was sentenced in December, but we don't know the day.

Let's denote T_i the random variable for the survival time of event i, and C_i the right censoring time. Thus, we defined another random variable Δ_i such that $\Delta_i(\omega) = 1$ if $T_i(\omega) \geq C_i(\omega)$ and 0 otherwise.

The survival function of T is $S(t) = P(T > t) = 1 - P(T \leq t)$.

Definition 1.1 (Hazard). Given a r.v. T and a survival function S, suppose that T is continuous with pdf $p(t)$. Thus, $h(t) := p(t \mid T \geq t)$. Note that this is similar, but not equal, to the conditional distribution $p(t \mid T \geq t_0)$, since our t_0 is also varying.

The hazard can also be expressed as:

$$h(t) = \frac{p(t)}{S(t)}.$$

The hazard is not a probability distribution, but, in a sense, it's the rate that at a time t one will fail. At first, one might think that such rate would be $p(t)$ itself, but note, for example, that if $T \sim \text{Exp}()$, the peak is right in the beginning, but the chance of actually failing in the beginning is zero. Thus, $p(t)$ is not the rate of failure experience at such time, but it's more like the rate at which your risk of failure grows.

Hence, in an exponential distribution the risk of failure grows faster in earlier times, but the actual rate of failure is given by the hazard function.

References