Hazard functions

Patrick Breheny

August 27

Introduction

- Let T be a nonnegative random variable representing the time to an event
- The distribution of T can be specified in a variety of ways, three of which are special to survival analysis in the sense that they come up often in the field, but are encountered only rarely outside of it:
  - The survival function, $S(t)$
  - The hazard function, $\lambda(t)$
  - The cumulative hazard function, $\Lambda(t)$
- We will begin by discussing the case where T follows a continuous distribution, and come back to the discrete and general cases toward the end of lecture

Survival function

- The *survival function* of T, denoted $S(t)$, is defined as

  $$S(t) = \mathbb{P}(T > t)$$

  for $t > 0$
- Note that $S(t)$ is related to the distribution function $F(t)$ and the density function $f(t)$ in the following ways:

  $$\begin{aligned} S(t) &= 1 - F(t) \\ f(t) &= -S'(t) \\ S(t) &= \int_t^\infty f(s)ds \end{aligned}$$
- Note that $S(t)$ uniquely defines a distribution

Hazard function

- The *hazard function*, $\lambda(t)$, is the instantaneous rate of failure at time t, given that an individual has survived until at least time t:

$$\begin{aligned}\lambda(t) &= \lim_{h \to 0^+} \frac{\mathbb{P}(t \le T < t + h | T \ge t)}{h} \\ &= f(t)/S(t)\end{aligned}$$

- In addition, note that

$$\lambda(t) = -\frac{d}{dt} \log S(t)$$

- Note that hazard functions are nonnegative and, like $S(t)$, uniquely define a distribution (under the assumption that $f(t)$ is continuous)

Cumulative hazard function

- Finally, the *cumulative hazard function* is simply the accumulated hazard up until time t:

$$\Lambda(t) = \int_0^t \lambda(s)ds$$

- Note that

$$\begin{aligned} \Lambda(t) &= -\log S(t) \\ S(t) &= \exp\{-\Lambda(t)\} \\ f(t) &= \lambda(t)\exp\{-\Lambda(t)\} \end{aligned}$$

and once again, $\Lambda(t)$ uniquely defines a distribution

Mean residual life

- It is worth mentioning that there is an interesting connection between the mean and the survival function
- Specifically, for any nonnegative continuous variable, the *expected residual life*, $r(t) = \mathbb{E}(T - t|T \geq t)$, is equal to

$$r(t) = \frac{\int_t^\infty S(u)du}{S(t)}$$

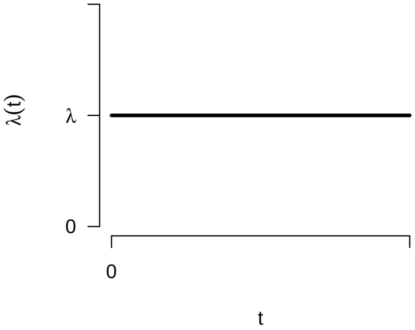
- In particular,

$$\mathbb{E}(T) = \int_0^\infty S(u)du,$$

which you are asked to show as a homework exercise

Constant hazard

- Let's consider the simplest meaningful survival distribution, that of the constant hazard rate:



- Obviously, this is overly simplistic for many situations, but is still a very convenient special case to work with

Cumulative hazard, survival, and density functions

- For the special case of constant hazard, the cumulative hazard is

  $$\Lambda(t) = \lambda t$$

  the survival function is therefore

  $$S(t) = e^{-\lambda t}$$

  and the density function is

  $$f(t) = \lambda e^{-\lambda t}$$

- Thus, by assuming a constant hazard, we arrive at the *exponential distribution*

Discrete distributions

- Let's consider the corresponding quantities and relationships for the analogous case where T has a discrete distribution: i.e., $T = t_j$ with probability $f(t_j)$ for a set of values $t_1 < t_2 < \cdots$
- In principle, survival data should always be continuous because time is a continuous quantity
- For a variety of reasons, however, the discrete case comes up often in survival analysis:
    - Nonparametric approaches typically condition on the observed failure times, resulting in discrete distributions
    - Data is not always recorded at precise times, but only at the level of day/month/year

Survival function in the discrete case

- The survival function is simply

$$S(t) = \sum_{t_j > t} f(t_j)$$

- Note that some authors define $S(t)$ as $\mathbb{P}(T \geq t)$, while others define it as $\mathbb{P}(T > t)$; we will use the latter definition to be consistent with the book

Hazards in the discrete case

- The hazard function is defined as in the continuous case:

$$\begin{aligned}\lambda_j &= \mathbb{P}(T = t_j | T \geq t_j) \\ &= f(t_j)/S(t_j^-),\end{aligned}$$

where

$$S(t_j^-) = \lim_{t \nearrow t_j} S(t)$$

- As in the continuous case, there is a relationship between S and λ:

$$S(t) = \prod_{t_j \leq t} \{1 - \lambda_j\}$$

Relationship between discrete and continuous cases

- On the surface, this relationship seems different than what we had for the continuous case
- However, consider discretizing a continuous hazard by dividing its range into intervals of equal length (i.e., failure at time t_j refers to failure in the jth interval)
- **Homework:** Show that, for a distribution with constant hazard $\lambda(t) = \lambda$, taking the limit of $S(t) = \prod_{t_j \leq t}\{1 - \lambda_j\}$ as the length of the intervals goes to zero yields the exponential distribution survival function
- This can be shown not just for constant hazards but for any continuous hazard function, although the proof is considerably longer in the general case

Mass function and hazard in the discrete case

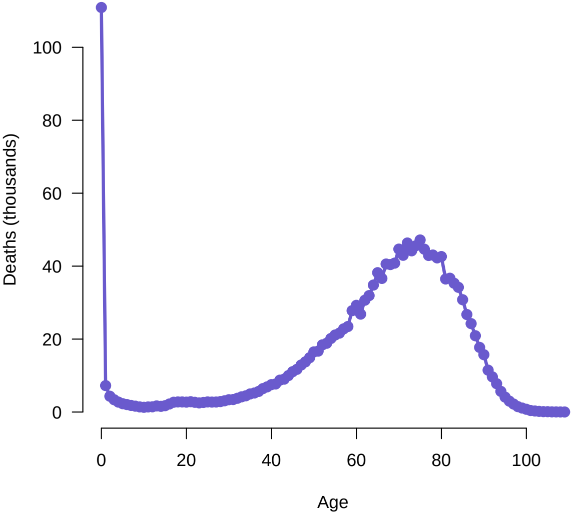
Finally, we have the following relationships between the probability mass function and hazard in the discrete case:

$$f(t_k) = \lambda_k \prod_{j<k} \{1 - \lambda_j\}$$

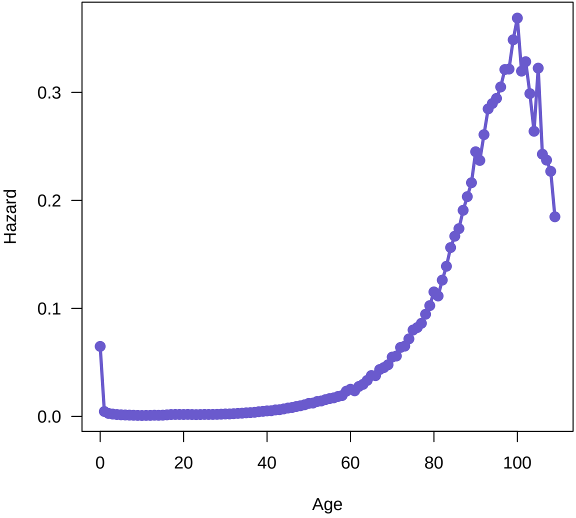
U.S. mortality data

- In the previous lecture, we discussed human life expectancy and the subtle relationship between distributions and hazards
- Let's look at some actual data from the Human Mortality Database at what the real distribution and hazard functions look like
- The data come from death counts by year for the United States; it is worth mentioning that I have made no effort to adjust for the fact that the size of the U.S. population was not constant over this time, so these results are clearly biased, but still serve to illustrate the general idea

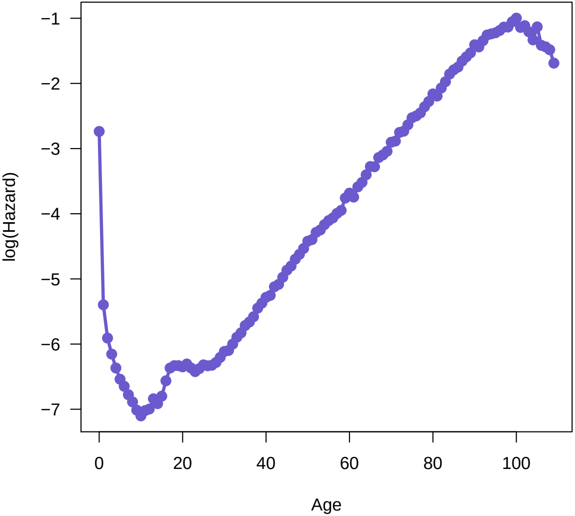
Distribution: Age at death for people who died in 1960



Hazard function (based on 1960 deaths)



Log(Hazard function) (based on 1960 deaths)



General case: Introduction

- Finally, let us consider the general case, without assuming that T is either continuous or discrete
- For the most part, everything we will do in this class falls into either the continuous or discrete cases, but seeing the general results are useful for a few reasons:
  - General results can help one to see things from a broader, more universal perspective
  - We need general results to deal with mixed distributions that have both continuous and discrete components
  - It provides a good exposure to some extensions of calculus that may be new to you

Differential increments

- Since F and Λ are not necessarily differentiable, expressions such as $f(t) = \frac{d}{dt}F(t)$ are not valid
- Instead, we must work in terms of *differential increments*:

$$\begin{aligned} dF(t) &= \mathbb{P}\{T \in [t, t+dt)\} \\ &= f_j 1\{t = t_j\} + f(t)dt, \end{aligned}$$

  where $f_j = \mathbb{P}\{T = t_j\}$ and $f(t)$ is the density of the continuous component of T
- Similarly, the differential increment of $\Lambda(t)$ is

$$d\Lambda(t) = \lambda_j 1\{t = t_j\} + \lambda(t)dt$$

Stieltjes integrals

- F and Λ can then be reconstructed from their differential increments using an extension of integration called *Stieltjes integration*
- Stieltjes integrals are written in terms of the differential increments as:

$$\begin{aligned}\Lambda(t) &= \int_0^t d\Lambda \\ &= \lim_{n\to\infty} \sum_{i=1}^{n} \Lambda(t_i) - \Lambda(t_{i-1}),\end{aligned}$$

  where $0 = t_0 < t_1 < \cdots < t_n = t$
- Obviously, we're glossing over some technical ideas here (does this limit exist? Does it depend on the partition we choose?, etc.), but hopefully the basic idea makes sense

Product integrals

- Finally, we can relate hazard functions and survival functions, but we need something called a *product integral* (basically, a product integral is to products what the integral is to sums):

$$
\begin{aligned}
S(t) &= \prod_0^t \{1 - d\Lambda\} \\
&\equiv \lim_{n\to\infty} \prod_{k=1}^n \{1 - [\Lambda(t_k) - \Lambda(t_{k-1})]\} \\
&= \exp\left\{-\int_0^t \lambda(u)du\right\} \prod_{t_j<t} (1-\lambda_j)
\end{aligned}
$$

- Next week, we'll start talking about inference – how to infer things about $\lambda(t)$ and $S(t)$ from data – particularly in the presence of censoring