Hazard Models

Rahul Telang

Timing

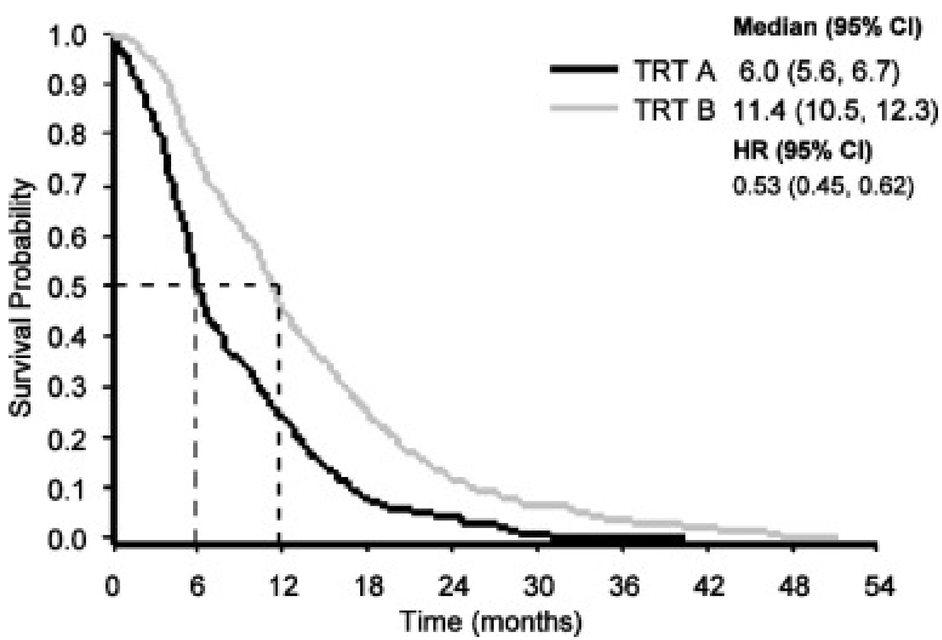
- We encounter data where the prediction is about when an event will occur.
  - When will this strike end?
  - When will customer adopt?
  - When will customer visit my web site next time?
  - How long a patient survive after going though a surgery?
- How does one answer these questions?

How to think about “when”

- We are interested in knowing when an event takes place.
- We want to formally model the “timing” decision.
- This is commonly referred to as hazard/survival models

Hazard in Medicine

- In most medical research, the following is a standard measure to access the effectiveness of treatment



Some basics

- To be able to formally understand hazard, we need to understand basic statistical terminology

Probability Density Function (pdf)

- The distribution of data is described by a “probability density function” (PDF)
- PDF is the relative probability, or likelihood, to observe data.
- If *p(x)* is a density function for some attribute of a population, then for it to be a probability density function, it must be.
  - $\int_a^b p(x)dx$= (fraction or probability of population for which a ≤ x ≤b)

- The probability will be higher in an interval where data is more likely to be found
- Since

$$\int_{-\alpha}^{\alpha} p(x)dx = 1$$

- The probability of (x) occurring between this range (-ꝏ,ꝏ) is 1.

Cumulative Distribution Function (cdf)

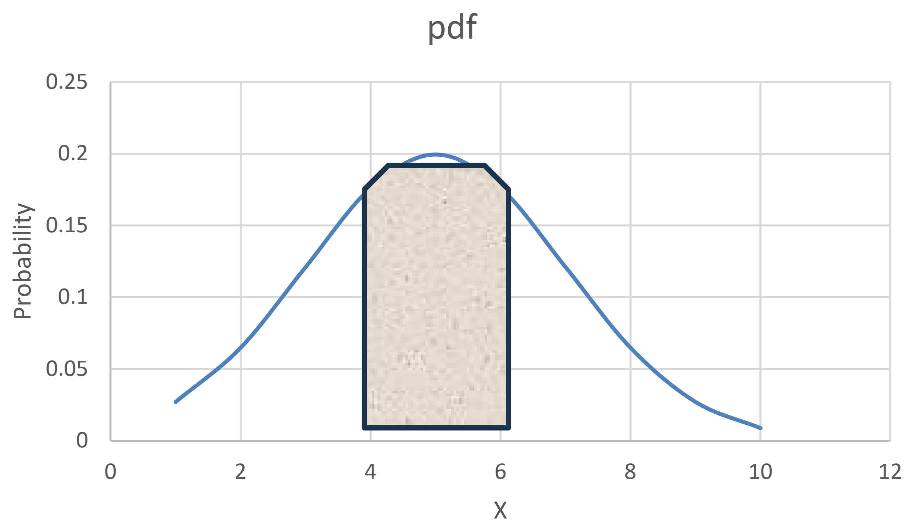
- Suppose f(x) is the density function of a quantity
- Then cumulative distribution function (cdf) for the quantity is defined as

$$F(x) = \int_{-\infty}^{X} f(x)dx$$

- Then F(x) calculates the proportion of items which have value less than x.
- Probability of observing value < x

PDF example

- Normal distribution with mean 5 and standard deviation 2. PDF is charting the probability of finding different value of data.
- The probability of finding data in range [4,6] is
  - $P = \int_0^6 f(x)dx - \int_0^4 f(x)dx = F(6) - F(4)$



Relating PDF and CDF

- Because F(x) $= \int_{-\alpha}^{x} f(t)dt$

- Then $\frac{d}{dx}F(x) = F'(x) = f(x)$

- Thus, pdf is a derivative (rate of change) of cdf.

Implied failure rates

The event we model can be thought of as "failure". Failure rate is a very well-developed concept in engineering).

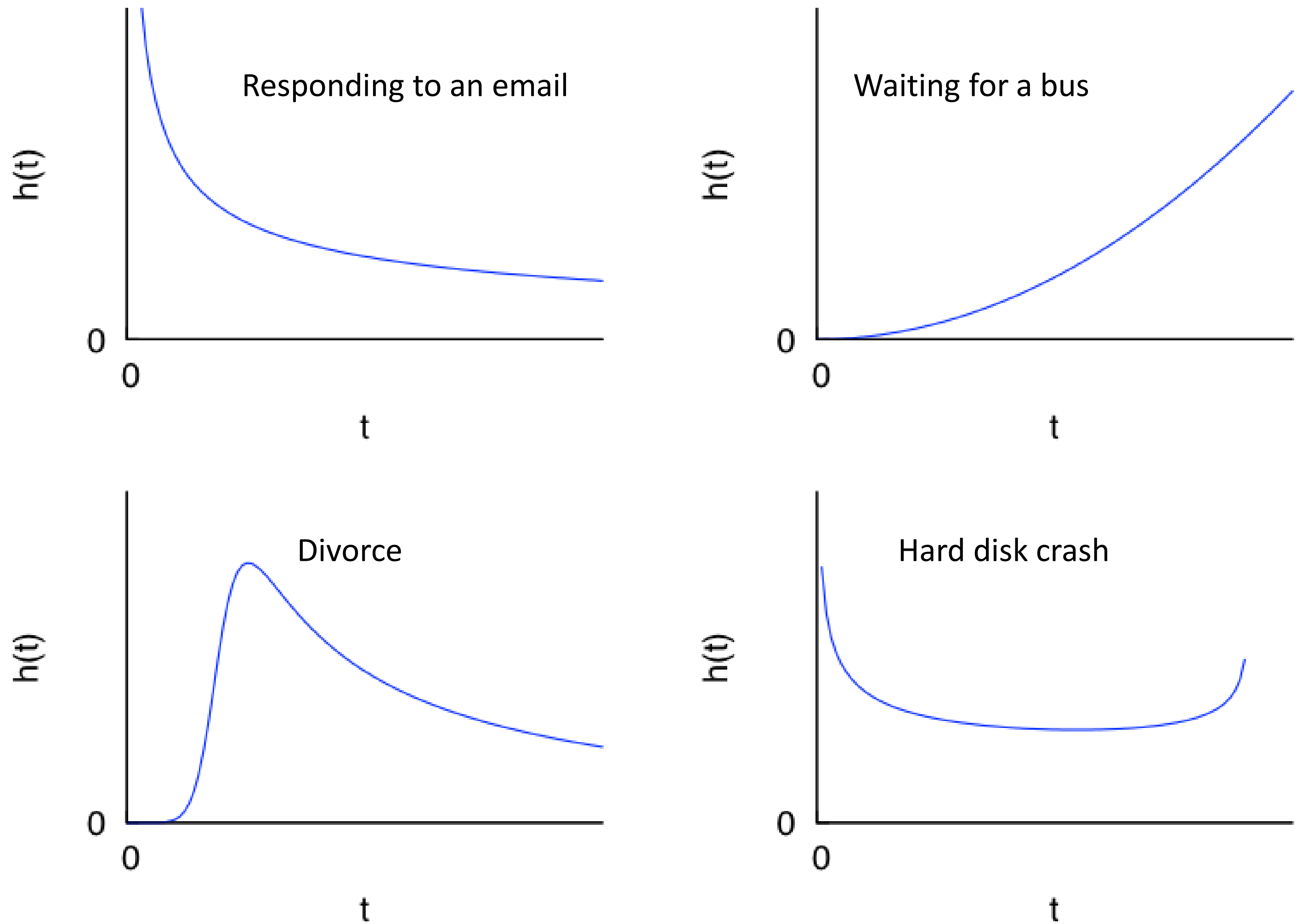
Failure can be though of as "death" in medical literature and

The models that describe such a behavior are called "Hazard" models.

Hazard processes

- What is the probability that the event will happen *right now*, given that it has *not yet happened?*
  - This is called the *hazard rate*.
- They are also referred to as "duration dependent" models. As these models imply that an action at a given time depends on the fact that the action has not occurred till that time
- Hazard rate is "conditional on surviving till time t"
- $hazard\ rate = \frac{probability\ of\ an\ event\ occurting\ at\ t}{event\ has\ not\ occurred\ until\ t}$

Some shapes of hazard rate functions



Definitions

- Suppose T is a non-negative random variable representing the time until some event of interest. For example, T might denote:
  - the time from diagnosis of a disease until death
  - the time between administration of a vaccine and development of an infection,
  - the time from the start of subscription of the service to the end (churn)
- We assume that T is continuous. The probability density function (pdf) and cumulative distribution function (cdf) are denoted these by f (·) and F (·), respectively:

pdf : $f(t)$
cdf : $F(t) = P(T \leq t)$

The hazard rate

The hazard rate function h(t) is defined by

$$h(t) = \lim_{\Delta \to 0} \frac{P(t < T \leq t + \Delta t | T > t)}{\Delta t}$$
$$= \frac{f(t)}{1 - F(t)}$$

The terminology “hazard rate” is derived from industrial engineering, and represents the instantaneous rate of “failure” at time t.

The denominator is called survival function. It is the probability that a certain object of interest will survive beyond a certain time t. So, it can be also interpreted as “Survival”

S(t) = P(T >t), or

= 1-F(t)

Calculate the hazard rate

- At time =0, there are 1000 items, and we provide time-line on when they fail. What is the empirical hazard?

time	failures	N = observations at risk	hazard
0	0	1000	0.00
1	350	1000	0.35
2	200	650	0.31
3	90	450	0.20
4	85	360	0.24
5	75	275	0.27
6	70	200	0.35
7	70	130	0.54
8	60	60	1.00

Starting Point for Hazard Analysis

- Exponential function is a commonly used function.

$$h(t) = \frac{f(t)}{1 - F(t)} = \frac{\lambda e^{-\lambda t}}{e^{-\lambda t}} = \lambda$$

- Notice that exponential implies a “constant” hazard rate.
  - This is why exponential is called a “memory-less” distribution because hazard is not a function of time.
- S(T) = 1-F(T) is the survivor function.

Hazard Rate and Distribution Function

- Remember that Hazard functions have one-on-one mapping with the distribution function. If you are defining one, you are automatically defining the other.

$$F(t) = 1 - \exp\left(-\int_0^t h(u)\,du\right)$$

  - Sometimes it is much easier to define Hazard function first.
  - By defining hazard rate, one has timing process to work with.

Distributions

- Commonly used distribution for survival analysis?

  Exponential pdf

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

  Exponential cdf

  $$F(t) = 1 - e^{-\lambda t}$$

  Hazard

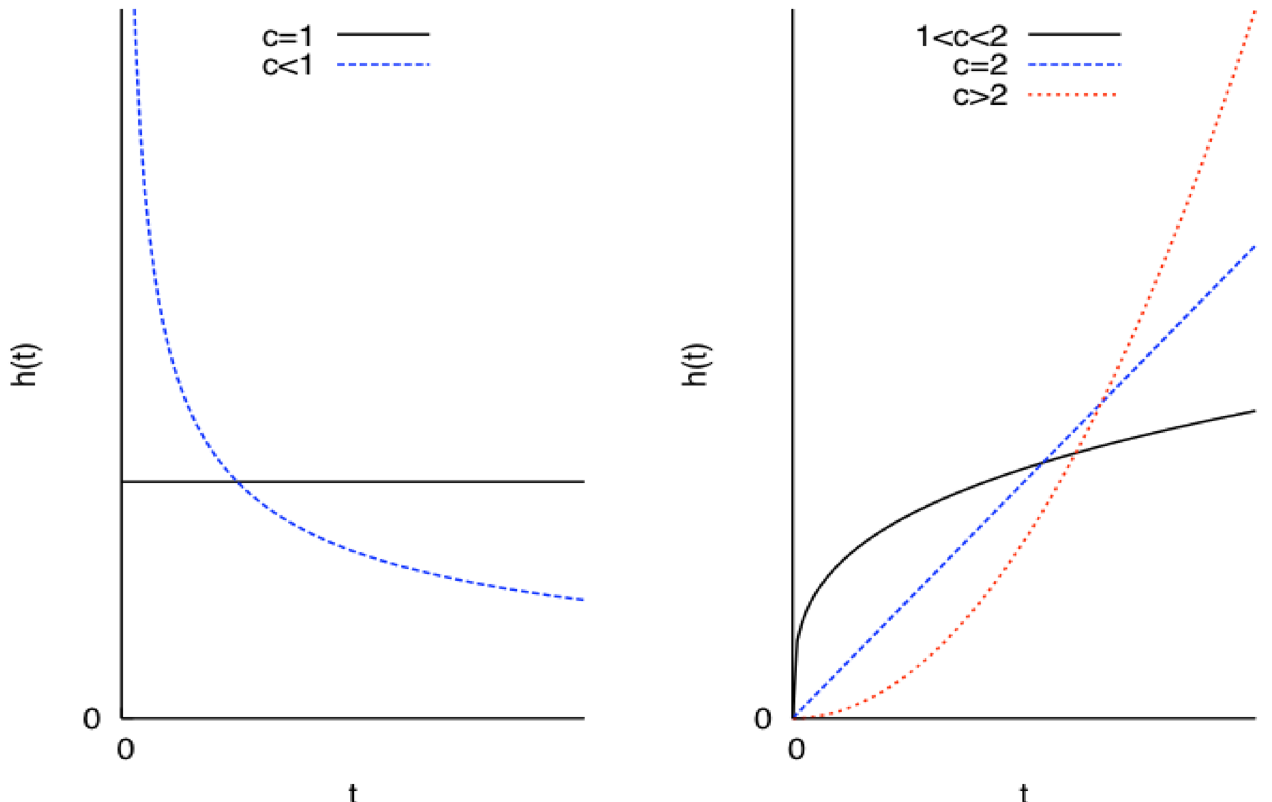
  $$h(t) = \frac{f(t)}{1 - F(t)} = \frac{\lambda e^{-\lambda t}}{e^{-\lambda t}} = \lambda$$

  Constant hazard. That is the hazard rate is not a function of time.

Weibull Distribution

- A flexible distribution that can represent increasing or decreasing hazard rates readily. This allows for λ to change with time
  - $F(t) = 1 - e^{-\lambda t^c}$
  - $f(t) = \lambda c t^{c-1} e^{-\lambda t^c}$
  - $h(t) = \frac{f(t)}{1-F(t)} = \lambda c t^{c-1}$
- The hazard is function of time. That means depending on c, the hazard can be increasing or decreasing. For c=1, this boils down to exponential

Weibull Hazard



- Decreasing hazard rate (negative duration dependence) when c<1
- Increasing hazard rate (positive duration dependence) when c>1

Distributions

- We have other commonly used hazard functions like log normal, gamma etc.
- We will use a very commonly used “Cox proportional hazard” as a tool for analyzing survival data. The interesting part of the model is that “hazard” in Cox is not specified unlike Exponential or Weibull.
- We will see why it is popularly used.

Summary

- Survival models are a powerful tool to model the “when” decision.
- They are widely used in variety of situation to fit the data.