Survival Analysis: Time To Event Modelling

Sudipta Das

Assistant Professor,
Department of Computer Science,
Ramakrishna Mission Vivekananda Educational & Research Institute

Outline I

1. Basic Quantities
   - Introduction
   - Survival Function
   - Hazard Function
   - The Mean Residual Life Function
   - Common Parametric Models for Survival Data

Basic Quantities I

- In this topic we consider the basic parameters used in modeling survival data. We shall define these quantities and show how they are interrelated.
- In addition, we shall also discuss some common parametric models, here.

Basic Quantities II

- Survival time, X (say) are characterized by *four* functions
    - The *survival function*,
        - the probability of an individual surviving to time x
    - The *hazard rate/function*, (sometimes termed risk function),
        - the chance an individual of age x experiences the event in the next instant in time
    - The *probability density* (or probability mass) *function*
        - the unconditional probability of the event's occurring at time x
    - The *mean residual life at time x*
        - the mean time to the event of interest, given the event has not occurred at x

- Note that
  - If we know any one of these four functions, then the other three can be uniquely determined.
  - In practice, these four functions, along with another useful quantity, *the cumulative hazard function*, are used to illustrate different aspects of the distribution of X.

Survival Function I

- The basic quantity employed to describe time-to-event phenomena is the survival function,
  - the probability of an individual surviving beyond time x (experiencing the event after time x).
- Survival Function is defined as

$$
\begin{aligned}
S_X(x) &= Pr(X > x) \\
&= 1 - F_X(x)
\end{aligned}
$$

- Survival functions for continuous life-time

$$S(x) = \int_x^\infty f(u)du$$

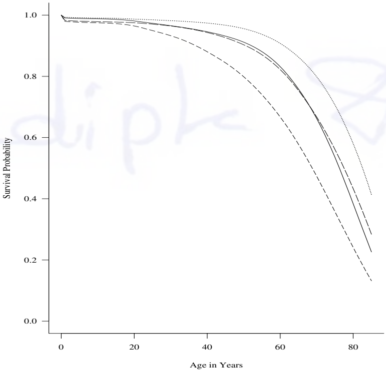
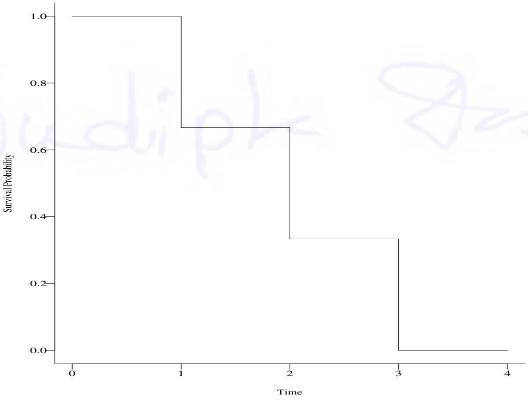


Figure 2.2 *Survival Functions for all cause mortality for the US population in 1989. White males (———); white females (······); black males (------); black females (— — —).*

Survival Function III

- Survival functions for discrete life-time

$$S(x) = \sum_{u>x} p(u)$$



Survival function for a discrete random lifetime

Hazard Function I

- A basic quantity, fundamental in survival analysis, is the hazard function/rate.
- This function is also known as
  - the conditional failure rate in reliability,
  - the force of mortality in demography,
  - the intensity function in stochastic processes,
  - the age-specific failure rate in epidemiology,
  - the inverse of the Mill's ratio in economics.

Hazard Function II

- The hazard rate is defined by

$$h(x) = \lim_{\Delta x \to 0} \frac{P(x \leq X < x + \Delta x | X \geq x)}{\Delta x}$$

  - One can see that $h(x)\Delta x$ may be viewed as the “approximate" probability of an individual of age x experiencing the event in the next instant.
  - This function is particularly useful in determining the appropriate failure distributions utilizing qualitative information about the mechanism of failure and for describing the way in which the chance of experiencing the event changes with time.

Hazard Function III

- There are many general shapes for the hazard rate.
- The only restriction on $h(x)$ is that it be nonnegative, i.e., $h(x) \geq 0$.

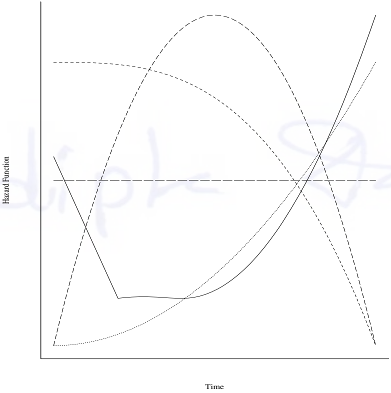


Figure 2.4 *Shapes of hazard functions. Constant hazard (——————); increasing hazard (------); decreasing hazard (- - - - - -); bathtub shaped (————); humpshaped (———).*

Hazard Function IV

- Models with increasing hazard rates may arise when there is natural aging or wear.
- Decreasing hazard functions are much less common but find occasional use in certain types of electronic devices or in patients experiencing certain types of transplants.

Hazard Function V

- Most often, a bathtub-shaped hazard is appropriate in populations followed from birth.
  - In this type of hazard function where, during an early period, deaths result, primarily, from infant diseases, after which the death rate stabilizes, followed by an increasing hazard rate due to the natural aging process.
  - Similarly, some manufactured equipment may experience early failure due to faulty parts, followed by a constant hazard rate which, in the later stages of equipment life, increases.
- Finally, if the hazard rate is increasing early and eventually begins declining, then, the hazard is termed hump-shaped.
  - This type of hazard rate is often used in modeling survival after successful surgery where there is an initial increase in risk due to infection, hemorrhaging, or other complications just after the procedure, followed by a steady decline in risk as the patient recovers.

Continuous Hazard Function I

- Hazard functions for continuous life-time

$$
\begin{aligned}
h(x) &= \lim_{\Delta x \to 0} \frac{P(x \leq X < x + \Delta x | X \geq x)}{\Delta x} \\
&= \lim_{\Delta x \to 0} \frac{P(x \leq X < x + \Delta x)}{P(X \geq x)\Delta x} \\
&= \lim_{\Delta x \to 0} \frac{f(x)\Delta x}{P(X \geq x)\Delta x} \\
&= \frac{f(x)}{S(x)} = -\frac{d}{dx} \ln[S(x)]
\end{aligned}
$$

- Thus, the cumulative hazard function $H(x)$, defined by

$$H(x) = \int_0^x h(u)du = -\ln[S(x)]$$

- Hence,

$$\begin{aligned} S(x) &= e^{-H(x)} \\ &= e^{-\int_0^x h(u)du} \end{aligned}$$

Discrete Hazard Function I

- Hazard functions for discrete life-time

$$
\begin{aligned}
h(x_j) &= Pr(X = x_j | X \geq x_j), \text{ for } j = 1, 2, \ldots \\
&= \frac{P(X = x_j)}{P(X > x_{j-1})} \\
&= \frac{p(x_j)}{S(x_{j-1})} \\
&= \frac{S(x_{j-1}) - S(x_j)}{S(x_{j-1})} = 1 - \frac{S(x_j)}{S(x_{j-1})}
\end{aligned}
$$

- Thus,

$$\frac{S(x_j)}{S(x_{j-1})} = 1 - h(x_j)$$

Discrete Hazard Function II

- Since, $S(x_0) = 1$

$$S(x) = \prod_{x_j \le x} \frac{S(x_j)}{S(x_{j-1})}$$
$$= \prod_{x_j \le x} \left[1 - h(x_j)\right]$$

- For discrete lifetimes, the cumulative hazard function by

$$H(x) = \sum_{x_j \le x} h(x_j)$$

Discrete Hazard Function III

- Sometimes, it is also defined as

$$H(x) = -\ln\left[\prod_{x_j \le x}[1 - h(x_j)]\right]$$

- Note that
  - the relationship for continuous lifetimes $S(x) = \exp[-H(x)]$ will be preserved for discrete lifetimes, for the later definition.
  - If the $h(x_j)$ are small, then the former will be an approximation of the later.

Mixed Hazard Function I

- If X is a positive random variable with a hazard rate $h(t)$, which is a sum of
  - a continuous function $h_c(t)$ and
  - a discrete function which has mass $h_d(x_j)$ at times $0 \leq x_1 < x_2 < \cdots$,
- then the survival function is related to the hazard rate by the so called “product integral" of $[1 - h(t)]dt$ defined as follows

$$S(x) = \prod_{x_j \leq x} [1 - h_d(x_j)] \exp\left[-\int_0^x h_c(t)dt\right].$$

The Mean Residual Life Function I

- For individuals of age x, the mean residual life at time x measures their expected remaining lifetime.
- It is defined as

$$mrl(x) = E(X - x | X > x).$$

The Mean Residual Life Function II

- For a continuous random variable,

$$
\begin{aligned}
mrl(x) &= \frac{\int_x^{\infty}(t-x)f(t)dt}{S(x)} \\
&= \frac{\int_x^{\infty} S(t)dt}{S(x)}
\end{aligned}
$$

The Mean Residual Life Function III

- Note that
  - Mean of X

$$mrl(0) = E[X] = \mu = \int_0^{\infty} S(t)dt$$

- From the definition of $mrl(x)$

$$\frac{d}{dx} mrl(x)S(x) = -S(x)$$

$$\Rightarrow \quad \left(1 + \frac{d}{dx} mrl(x)\right) S(X) = mrl(x)f(x)$$

$$\Rightarrow \quad h(x) = \frac{1 + \frac{d}{dx} mrl(x)}{mrl(x)}$$

The Mean Residual Life Function IV

- For a discrete random variable,

$$mrl(x) = \frac{(x_{i+1} - x)S(x_i) + \sum_{j \geq i+1}(x_{j+1} - x_j)S(x_j)}{S(x)}, \text{ for } x_i \leq x \leq x_{i+1}$$

Gamma Distribution I

$X \sim Gamma(\beta, \lambda)$

- Probability density function:

$$f(x, \beta, \lambda) = \frac{\lambda(\lambda x)^{\beta-1}}{\Gamma(\beta)} e^{-\lambda x}$$

  - $\beta (> 0)$: the shape parameter
  - $\lambda (> 0)$: the rate parameter
    - $\frac{1}{\lambda}$: the scale parameter

- Survival Function:

$$S(x) = 1 - \frac{1}{\Gamma(\beta)} \gamma(\beta, \lambda x)$$

  - $\gamma(\beta, \lambda x) = \int_0^{\lambda x} t^{\beta-1} e^{-t} dt$

Gamma Distribution II

- Hazard Function:

$$h(x) = \frac{\lambda(\lambda x)^{\beta-1} e^{-\lambda x}}{\Gamma(\beta) - \gamma(\beta, \lambda x)}$$

  - Monotonic
  - Converging
- **FIGURE 2A**
- Expectation:

$$E[X] = \frac{\beta}{\lambda}$$

Weibull Distribution I

$X \sim \textit{Weibull}(\beta, \lambda)$

- Probability density function:

$$f(x, \beta, \lambda) = \beta\lambda(\lambda x)^{\beta-1} e^{-(\lambda x)^{\beta}}$$

  - $\beta(> 0)$: the shape parameter
  - $\frac{1}{\lambda}(> 0)$: the scale parameter

- Survival Function:

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

Weibull Distribution II

- Hazard Function:

$$h(x) = \beta\lambda(\lambda x)^{\beta-1}$$

  - Monotonic
  - Not converging
- **FIGURE 2B**
- Expectation:

$$E[X] = \frac{\Gamma(1+\frac{1}{\beta})}{\lambda}$$

Log-normal Distribution I

- Let Z be a standard normal random variable, and let μ and $\sigma > 0$ be two real numbers.
  - $f_Z(z) = \phi(z) = \frac{1}{\sqrt{2\pi}} e^{-\frac{1}{2}z^2}$
- Then, the random variable

$$X = e^{\mu + \sigma Z}$$

follows the log-normal distribution with parameters μ and σ.

Log-normal Distribution II

- Probability density function:

$$
\begin{aligned}
f_X(x, \mu, \sigma) &= \left|\frac{dz}{dx}\right| f_Z\left(\frac{\ln x - \mu}{\sigma}\right) \\
&= \frac{1}{x\sigma}\phi\left(\frac{\ln x - \mu}{\sigma}\right) \\
&= \frac{1}{x\sigma}\left[\frac{1}{\sqrt{2\pi}} e^{-\frac{1}{2}\left(\frac{\ln x - \mu}{\sigma}\right)^2}\right],
\end{aligned}
$$

  - $-\infty < \mu < \infty$: the location parameter
  - $\sigma > 0$: the shape parameter

- Survival Function:

$$S(x) = 1 - \Phi\left(\frac{\ln x - \mu}{\sigma}\right) = \Phi\left(\frac{-\ln x + \mu}{\sigma}\right),$$

  - where $\Phi(x) = \int_{-\infty}^{x} \phi(z)dz = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{z} e^{-\frac{1}{2}z^2} dz.$

- Hazard Function:

$$h(x) = \frac{\frac{1}{x\sigma}\phi\left(\frac{\ln x - \mu}{\sigma}\right)}{\Phi\left(\frac{-\ln x + \mu}{\sigma}\right)}$$

  - Non-monotonic

- **FIGURE 2C**
- Expectation:

$$E[X] = e^{\mu + 0.5\sigma^2}$$

Log-logistic distribution I

- Let Z be a standard logistic random variable, and let μ and $\sigma > 0$ be two real numbers.
    - $f_Z(z) = \dfrac{e^z}{(1+e^z)^2}$
- Then, the random variable

$$X = e^{\mu+\sigma Z}$$

follows the log-logistic distribution with parameters μ and σ.

Log-logistic distribution II

- Probability density function:

$$
\begin{aligned}
f_X(x,\beta,\lambda) &= \left|\frac{dz}{dx}\right| f_Z\left(\frac{\ln x-\mu}{\sigma}\right)\\
&= \frac{1}{\sigma x}\frac{e^{\frac{\log x-\mu}{\sigma}}}{\left(1+e^{\frac{\log x-\mu}{\sigma}}\right)^2}\\
&= \frac{1}{\sigma x}\frac{x^{\frac{1}{\sigma}}e^{-\frac{\mu}{\sigma}}}{\left(1+x^{\frac{1}{\sigma}}e^{-\frac{\mu}{\sigma}}\right)^2} = \frac{\beta\lambda(\lambda x)^{\beta-1}}{[1+(\lambda x)^{\beta}]^2}
\end{aligned}
$$

- $\beta=\dfrac{1}{\sigma}(>0)$: the shape parameter
- $\lambda=e^{-\mu}(>0)$: the scale parameter

Log-logistic distribution III

- Survival Function:

$$S(x) = \frac{1}{1 + (\lambda x)^{\beta}}$$

- Hazard Function:

$$h(x) = \frac{\beta\lambda(\lambda x)^{\beta-1}}{1 + (\lambda x)^{\beta}}$$

    - Non-Monotonic

- **FIGURE 2D**

Extreme value distribution I

- Let $X \sim Weibull(\beta, \lambda)$, with $\beta > 0$ and $\lambda > 0$.
  - $f_X(x) = \beta\lambda(\lambda x)^{\beta-1} e^{-(\lambda x)^\beta}$
- Then the random variable

$$Y = \log X$$

  follows an extreme value distribution with parameters $\sigma > 0$ and $\mu(-\infty < \mu < \infty)$, i.e.
- $Y \sim EV(\mu, \sigma)$.
  - σ: the scale parameter
  - μ: the location parameter

Extreme value distribution II

- Probability density function:

$$
\begin{aligned}
f_Y(y,\sigma,\mu) &= \left|\frac{dx}{dy}\right| \times f_X(e^y) \\
&= e^y \times \beta\lambda(\lambda e^y)^{\beta-1} e^{-(\lambda e^y)^\beta} \\
&= \beta\lambda^\beta \exp\left[y\beta - \lambda^\beta e^{y\beta}\right] \\
&= \beta \exp\left[y\beta + \beta\log\lambda - e^{y\beta+\beta\log\lambda}\right] \\
&= \frac{1}{\sigma}\exp\left[\left(\frac{y-\mu}{\sigma}\right) - \exp\left(\frac{y-\mu}{\sigma}\right)\right]
\end{aligned}
$$

  - $\sigma = \beta^{-1}$: the scale parameter
  - $\mu = -\log\lambda$: the location parameter

Extreme value distribution III

- Survival Function:

$$S(y) = \exp\left[-\exp\left(\frac{y-\mu}{\sigma}\right)\right]$$

Extreme value distribution IV

- If $Y \sim EV(\mu, \sigma)$, then

$$Z = \frac{Y - \mu}{\sigma} \sim EV(0, 1).$$

- Probability density function:

$$f_Z(z) = e^{z - e^z}$$

- Survival Function:

$$S(z) = e^{-e^z}$$

Gompertz distribution I

$X \sim Gompertz(\alpha, \theta)$

- Probability density function:

$$f(x, \alpha, \theta) = \theta e^{\alpha x} e^{\frac{\theta}{\alpha}(1 - e^{\alpha x})}$$

  - $\alpha (> 0)$: the scale parameter
  - $\theta (> 0)$: the shape parameter

- Survival Function:

$$S(x) = e^{\frac{\theta}{\alpha}(1 - e^{\alpha x})}$$

Gompertz distribution II

- Hazard Function:

$$h(x) = \theta e^{\alpha x}$$

    - Monotonic
    - Not converging
- **FIGURE 2E**

Some other distributions I

- Other distributions which have received some attention in the literature are the
  - the exponential power distribution,
  - the inverse Gaussian distribution and
  - the Pareto distribution.
- The generalized gamma distribution introduces an additional parameter allowing additional flexibility in selecting a hazard function.

Some other distributions II

- The density function of generalized gamma distribution

$$f(x) = \frac{\alpha\lambda^{\beta}x^{\alpha\beta-1}e^{-\lambda x^{\alpha}}}{\Gamma(\beta)}$$

- The survival function of generalized gamma distribution

$$S(x) = 1 - \gamma(\lambda x^{\alpha}, \beta)$$

- Note that, this distribution reduces
  - to the exponential when $\alpha = \beta = 1$,
  - to the Weibull when $\beta = 1$,
  - to the gamma when $\alpha = 1$ and
  - approaches the log normal as $\beta \to \infty$.