Survival Analysis
APTS 2016/17 Preliminary material

Ingrid Van Keilegom

KU Leuven
(ingrid.vankeilegom@kuleuven.be)

August 2017

1. Introduction
2. Common functions in survival analysis
3. Parametric survival distributions
4. Exercises
5. References

1. Introduction
2. Common functions in survival analysis
3. Parametric survival distributions
4. Exercises
5. References

What is ‘Survival analysis’ ?

- Survival analysis (or duration analysis) is an area of statistics that models and studies the time until an event of interest takes place.
- In practice, for some subjects the event of interest cannot be observed for various reasons, e.g.
  - the event is not yet observed at the end of the study
  - another event takes place before the event of interest
  - ...
- In survival analysis the aim is
  - to model ‘time-to-event data’ in an appropriate way
  - to do correct inference taking these special features of the data into account.

Examples

- Medicine :
  - time to death for patients having a certain disease
  - time to getting cured from a certain disease
  - time to relapse of a certain disease
- Agriculture :
  - time until a farm experiences its first case of a certain disease
- Sociology (‘duration analysis’) :
  - time to find a new job after a period of unemployment
  - time until re-arrest after release from prison
- Engineering (‘reliability analysis’) :
  - time to the failure of a machine

1 Introduction

2 Common functions in survival analysis

3 Parametric survival distributions

4 Exercises

5 References

- Let T be a non-negative continuous random variable, representing the time until the event of interest.
- Denote

$F(t) = P(T \leq t)$	distribution function
$f(t)$	probability density function

- For survival data, we consider rather

$S(t)$	survival function
$H(t)$	cumulative hazard function
$h(t)$	hazard function
$mrl(t)$	mean residual life function

- Knowing one of these functions suffices to determine the other functions.

Survival function

$$S(t) = P(T > t) = 1 - F(t)$$

- Probability that a randomly selected individual will survive beyond time t
- Decreasing function, taking values in $[0, 1]$
- Equals 1 at $t = 0$ and 0 at $t = \infty$

Cumulative Hazard Function

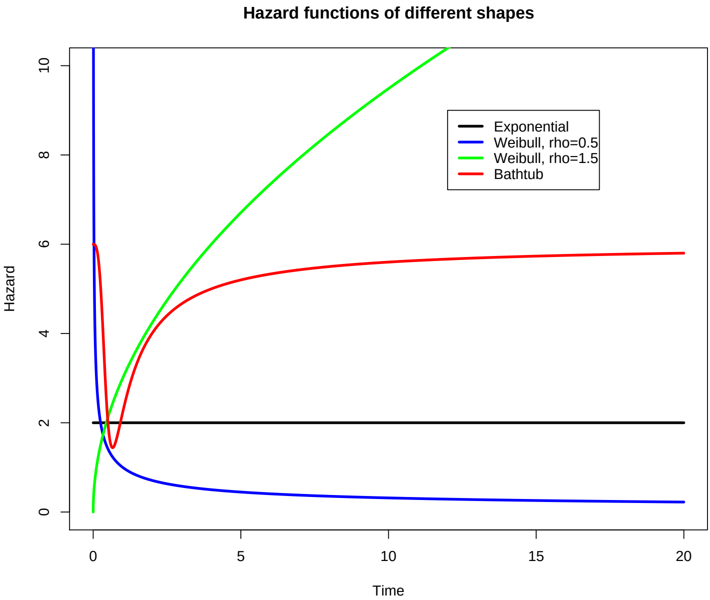
$$H(t) = -\log S(t)$$

- Increasing function, taking values in $[0, +\infty]$
- $S(t) = exp(-H(t))$

Hazard Function (or Hazard Rate)

$$
\begin{aligned}
h(t) &= \lim_{\Delta t \to 0} \frac{P(t \leq T < t + \Delta t \mid T \geq t)}{\Delta t} \\
&= \frac{1}{P(T \geq t)} \lim_{\Delta t \to 0} \frac{P(t \leq T < t + \Delta t)}{\Delta t} \\
&= \frac{f(t)}{S(t)} = \frac{-d}{dt} \log S(t) = \frac{d}{dt} H(t)
\end{aligned}
$$

- $h(t)$ measures the instantaneous risk of dying right after time t given the individual is alive at time t
- Positive function (not necessarily increasing or decreasing)
- The hazard function $h(t)$ can have many different shapes and is therefore a useful tool to summarize survival data



Hazard functions of different shapes
Exponential
Weibull, rho=0.5
Weibull, rho=1.5
Bathtub
Hazard
Time
0
2
4
6
8
10
0
5
10
15
20

Mean Residual Life Function

- The mrl function measures the expected remaining lifetime for an individual of age t. As a function of t, we have

$$mrl(t) = \frac{\int_t^\infty S(s)ds}{S(t)}$$

- This result is obtained from

$$mrl(t) = \mathrm{E}(T - t \mid T > t) = \frac{\int_t^\infty (s-t)f(s)ds}{S(t)}$$

- Mean life time:

$$\mathrm{E}(T) = \mathrm{mrl}(0) = \int_0^\infty sf(s)ds = \int_0^\infty S(s)ds$$

1 Introduction

2 Common functions in survival analysis

3 Parametric survival distributions

4 Exercises

5 References

Exponential distribution

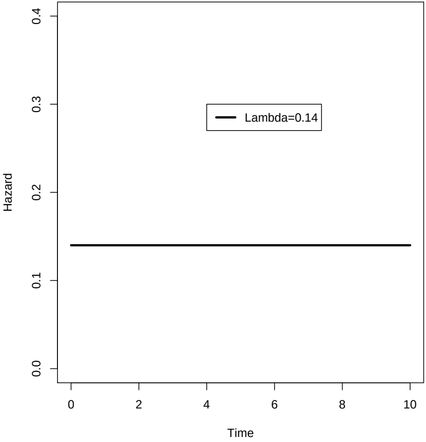
- Characterized by one parameter $\lambda > 0$:

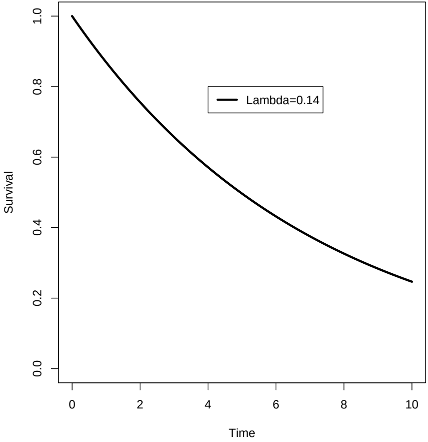
$$
\begin{aligned}
S_0(t) &= \exp(-\lambda t) \\
f_0(t) &= \lambda \exp(-\lambda t) \\
h_0(t) &= \lambda
\end{aligned}
$$

  $\rightarrow$ leads to a constant hazard function

- Empirical check: plot of the log of the survival estimate versus time

Hazard and survival function for the exponential distribution





Weibull distribution

- Characterized by a scale parameter $\lambda > 0$ and a shape parameter $\rho > 0$:

$$\begin{aligned} S_0(t) &= \exp(-\lambda t^{\rho}) \\ f_0(t) &= \rho\lambda t^{\rho-1}\exp(-\lambda t^{\rho}) \\ h_0(t) &= \rho\lambda t^{\rho-1} \end{aligned}$$

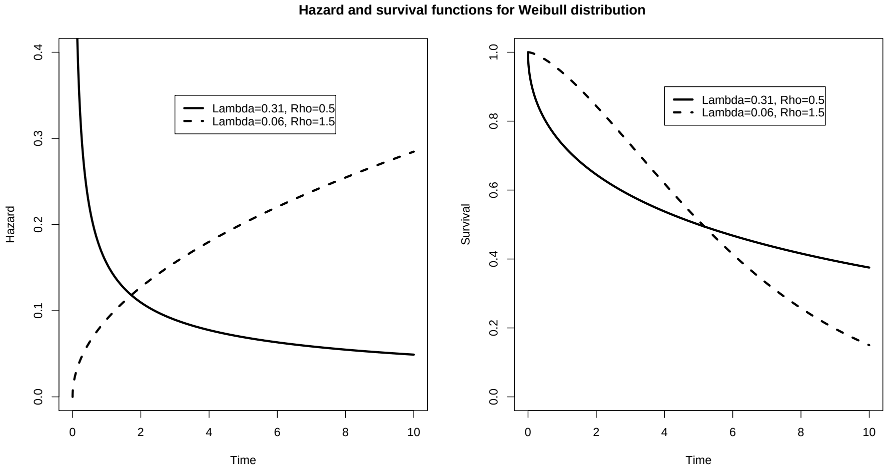
  $\rightarrow$ hazard decreases monotonically with time if $\rho < 1$

  $\rightarrow$ hazard increases monotonically with time if $\rho > 1$

  $\rightarrow$ hazard is constant over time if $\rho = 1$ (exponential case)

- Empirical check: plot log cumulative hazard versus log time

Hazard and survival function for the Weibull distribution



Gompertz distribution

- Characterized by two parameters $\lambda > 0$ and $\gamma > 0$:

$$\begin{aligned} S_0(t) &= \exp\left[-\lambda\gamma^{-1}\left(\exp(\gamma t) - 1\right)\right] \\ f_0(t) &= \lambda \exp(\gamma t) \exp\left[-\lambda\gamma^{-1}\left(\exp(\gamma t) - 1\right)\right] \\ h_0(t) &= \lambda \exp(\gamma t) \end{aligned}$$

  $\rightarrow$ hazard increases from λ at time 0 to ∞ at time ∞

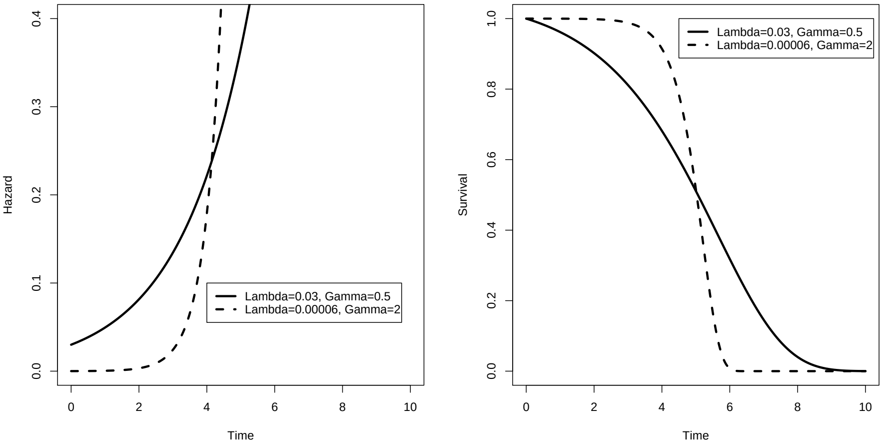
  $\rightarrow$ $\gamma = 0$ corresponds to the exponential case

- Gompertz distribution can also be presented with $\gamma \in R$

  $\rightarrow$ for $\gamma < 0$ the hazard is decreasing and the cumulative hazard is not going to ∞ when $t \rightarrow \infty$

  $\rightarrow$ part of the population will never experience the event

Hazard and survival function for the Gompertz distribution



Log-logistic distribution

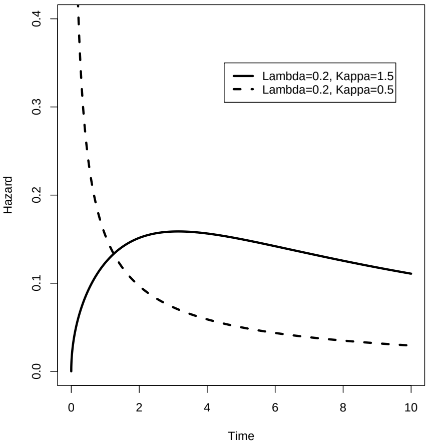
- A random variable T has a log-logistic distribution if $logT$ has a logistic distribution
- Characterized by two parameters λ and $\kappa > 0$:

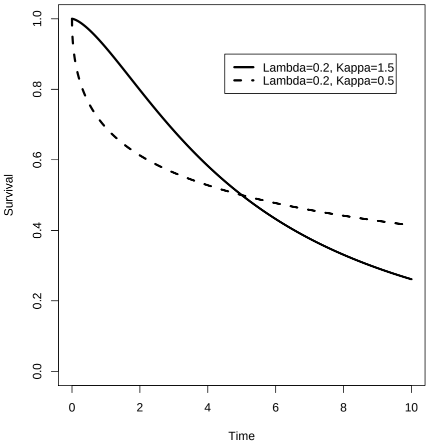
$$\begin{aligned} S_0(t) &= \frac{1}{1+(t\lambda)^{\kappa}} \\ f_0(t) &= \frac{\kappa t^{\kappa-1}\lambda^{\kappa}}{[1+(t\lambda)^{\kappa}]^2} \\ h_0(t) &= \frac{\kappa t^{\kappa-1}\lambda^{\kappa}}{1+(t\lambda)^{\kappa}} \end{aligned}$$

- The median event time is only a function of the parameter λ :

$$Med(T) = \exp(1/\lambda)$$

Hazard and survival function for the log-logistic distribution





Log-normal distribution

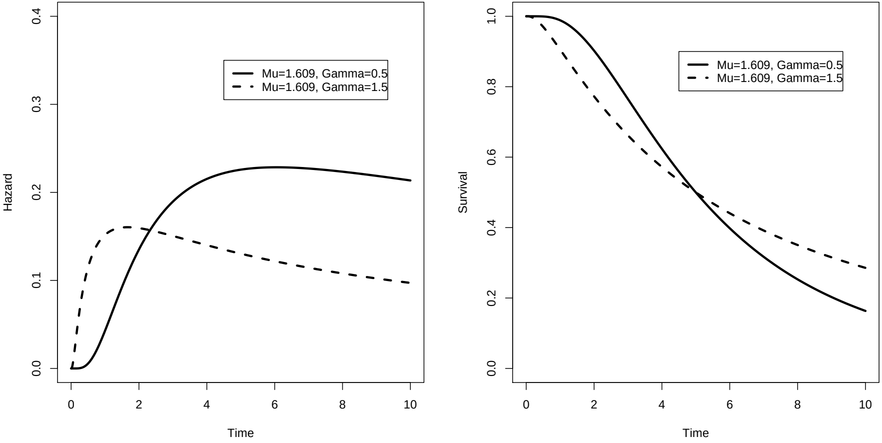
- Resembles the log-logistic distribution but is mathematically less tractable
- A random variable T has a log-normal distribution if *logT* has a normal distribution
- Characterized by two parameters μ and $\gamma > 0$:

$$
\begin{aligned}
S_0(t) &= 1 - F_N\left(\frac{\log(t) - \mu}{\sqrt{\gamma}}\right) \\
f_0(t) &= \frac{1}{t\sqrt{2\pi\gamma}} \exp\left[-\frac{1}{2\gamma}\left(\log(t) - \mu\right)^2\right]
\end{aligned}
$$

- The median event time is only a function of the parameter μ :

$$Med(T) = \exp(\mu)$$

Hazard and survival function for the log-normal distribution



1 Introduction

2 Common functions in survival analysis

3 Parametric survival distributions

4 Exercises

5 References

1. Find a few more practical situations where time-to-event data are of interest, and try to imagine why the event of interest can sometimes not be observed in these situations.

2. Show that the four common functions in survival analysis (survival function, cumulative hazard function, hazard function and mean residual life function) all determine the law of the random variable of interest in a unique way.

1 Introduction

2 Common functions in survival analysis

3 Parametric survival distributions

4 Exercises

5 References

Some textbooks on survival analysis :

- Cox, D.R. et Oakes, D. (1984). *Analysis of survival data*, Chapman and Hall, New York.
- Fleming, T.R. et Harrington, D.P. (1981). *Counting processes and survival analysis*, Wiley, New York.
- Hougaard, P. (2000). *Analysis of multivariate survival data*. Springer, New York.
- Kalbfleisch, J.D. et Prentice, R.L. (1980). *The statistical analysis of failure time data*, Wiley, New York.
- Klein, J.P. and Moeschberger, M.L. (1997). *Survival analysis, techniques for censored and truncated data*, Springer, New York.
- Kleinbaum, D.G. et Klein, M. (2005). *Survival analysis, a self-learning text*, Springer, New York.