MATH 60604A
Statistical modelling
§ 7b - Likelihood for survival analysis

HEC Montréal
Department of Decision Sciences

Survival and hazard functions

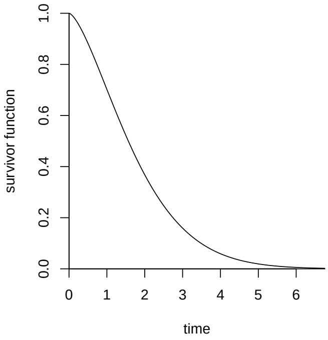
Let T denote the survival time

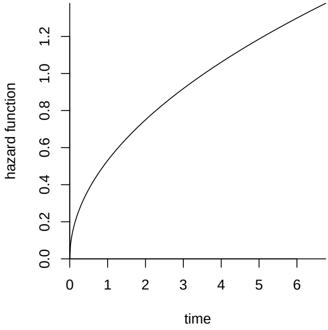
- The survival function, $S(t) = \mathsf{P}(T > t)$, completely caracterizes the law of T.
- Often, we are more interested in knowing what time periods are characterized by higher failure rates. The hazard of T is

$$
\begin{aligned}
h(t) &= \lim_{\delta \to 0} \frac{\mathsf{P}(t < T < t + \delta \mid T > t)}{\delta} \\
&= \lim_{\delta \to 0} \frac{1}{\delta} \frac{\mathsf{P}(t < T < t + \delta)}{\mathsf{P}(T > t)} \\
&= \frac{f(t)}{S(t)}
\end{aligned}
$$

We can think of the hazard rate as being the instantaneous probability of "dying" at time t, given survival to time t.

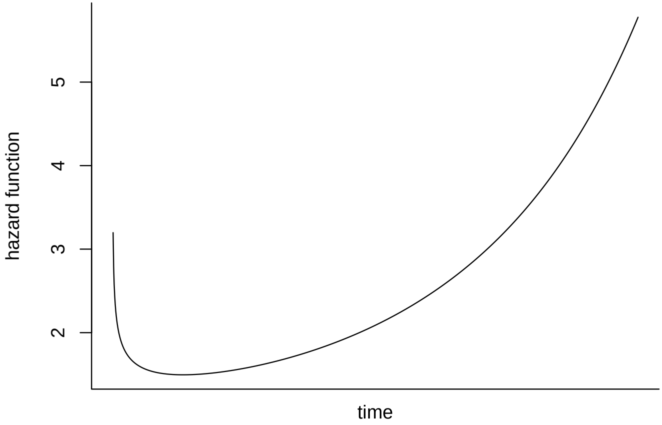
Survival function and hazard function





The survival function decreases monotically from $S(0) = 1$. The higher the hazard $h(t)$, the fastest the decrease of the survival function.

Bathtub shaped hazard



Typical hazard shape: the risk rate is high (e.g., childhood mortality, manufacturing defect) initially, then decreases and plateau. As time goes on, the hazard increases steadily.

Random censoring and likelihood

We observe $T_i = \min\{T_i^0, C_i\}$. If an observation is right-censored at time c, we know that $S(c) = \mathsf{P}\left(T_i^0 > c\right)$

- in other words, survival time exceeds c.

If we have censoring, the database includes an indicator variable δ_i where

$$T_i = \begin{cases} T_i^0, & \delta_i = 1 \text{ (observed failure time)} \\ C_i, & \delta_i = 0 \text{ (right-censored)} \end{cases}$$

Likelihood contribution

Let $S(t;\boldsymbol{\theta}) = \mathsf{P}\left(T_i^0 > t\right)$ denote the survival function of T_i^0. If T_i^0 is independent of C_i, the likelihood contribution of each observation is

$$L_i(\boldsymbol{\theta}) = \begin{cases} f(t_i;\boldsymbol{\theta}), & \delta_i = 1 \text{ (observed failure time)} \\ S(t_i;\boldsymbol{\theta}), & \delta_i = 0 \text{ (right-censored)} \end{cases}.$$

We can therefore write the log likelihood as

$$\ell(\boldsymbol{\theta}) \equiv \sum_{i:\delta_i=1} \ln f(t_i;\boldsymbol{\theta}) + \sum_{i:\delta_i=0} \ln S(t_i;\boldsymbol{\theta})$$

Inferential approaches

Many avenues are open for estimating the survival function (or hazard).

- parametric: choose a family of distributions (Weibull, log normal, Gompertz, exponential) for T.
  - \+ can easily incorporate explanatories
  - \+ continuous function, can be used to extrapolate
  - − subject to model misspecificiation
  - − not flexible: can fit poorly to the data.
- nonparametric: no distributional assumption
  - − no explanatory variable
  - \+ minimal hypotheses, theoretical guarantees for large sample size
  - \+ flexible
  - − yields discontinuous estimates
  - − cannot extrapolate beyond the largest observed time.

Parametric model for survival: exponential distribution

Consider $T_i \overset{\text{iid}}{\sim} \mathsf{E}(\lambda)$, i.e., exponential variables with expectation λ^{-1}.

- The survival function of T is $S(T) = \exp(-\lambda t)$ and
- the hazard $h(t) = \lambda$ is **constant**.

The log likelihood for a random sample of size n is

$$\ell(\lambda) = \sum_{i=1}^{n} \{\delta_i \ln \lambda - \lambda T_i\}.$$

The maximum likelihood estimator is $\widehat{\lambda} = \sum_{i=1}^{n} \delta_i / \sum_{i=1}^{n} T_i$.

- The estimated survival time is infinite if no one failed.
- The standard errors are obtained from the observed information matrix $j(\widehat{\lambda}) = \sum_{i=1}^{n} \delta_i / \widehat{\lambda}^2$; censored observations contribute no information.