Survival Analysis: Weeks 1-2

Lu Tian and Richard Olshen

Stanford University

Survival Time/Failure Time/Event Time

- What is the survival outcomes? the time to a clinical event of interest: terminal and non-terminal events.
  1. the time from diagnosis of cancer to death
  2. the time between administration of a vaccine and infection date
  3. the time from the initiation of a treatment to the time of the disease progression.

Survival time T

- Let T be a **nonnegative** random variable denoting the time to the event of interest (survival time/event time/failure time).
- The distribution of T could be discrete, continuous or a mixture of both. We will focus on the continuous distribution.

Survival time T

- The distribution of a random variable $T \geq 0$ can be characterized by its probability density function (pdf) and cumulative distribution function (CDF). However, in survival analysis, we often focus on

  1. Survival function: $S(t) = \text{pr}(T > t)$. If T is time to death, then $S(t)$ is the probability that a subject survives beyond time t.

  2. Hazard function:

$$h(t) = \lim_{\epsilon \downarrow 0} \frac{\text{pr}(T \in (t, t+\epsilon] | T \geq t)}{\epsilon}.$$

  3. Cumulative hazard function: $H(t) = \int_0^t h(u)du$.

Relationships between survival and hazard functions

- Hazard function: $h(t) = f(t)/S(t)$.
- Cumulative hazard function

$$H(t) = \int_0^t h(u)du = \int_0^t \frac{f(u)}{S(u)}du = \int_0^t \frac{-dS(u)}{S(u)}du = -\log\{S(t)\}$$

- $f(t) = h(t)S(t) = h(t)\exp\{-H(t)\}$
- $S(t) = \exp\{-H(t)\}$.

Additional properties of hazard functions

- If $H(t)$ is the cumulative hazard function of T, then $H(T) \sim \text{EXP}(1)$, the unit exponential distribution. (Equivalent to the statement that $F(T) \sim U(0,1)$, where $F(\cdot)$ is the CDF of the random variable T.)
- If T_1 and T_2 are two independent survival times with hazard functions $h_1(t)$ and $h_2(t)$, respectively, then $T = \min(T_1, T_2)$ has a hazard function $h_T(t) = h_1(t) + h_2(t)$. (This statement can be generalized to the case with more than two survival times)

Hazard functions

- The hazard function $h(t)$ is NOT the probability that the event (such as death) occurs at time t or before time t. (The latter is often called "risk" in epidemiology.)
- $h(t)\Delta$ is approximately the conditional probability that the event occurs within the interval $(t, t+\Delta]$ given that the event has not occurred before time t for small $\Delta > 0$.
- If the hazard function $h(t)$ increases $X\%$ at $[0, \tau]$, the probability of failure before τ in general does not increase $X\%$.

Why hazard

- Interpretability: in general, it could be fairly straightforward to understand how the hazard (qualitatively) changes with time, e.g., think about the hazard (of death) for a person since his/her birth.
- Advantage in analyzing censored data.

Exponential distribution

- In survival analysis the exponential distribution is the " simplest" parametric distribution for survival time.
- Denote the exponential distribution by EXP(λ) :
  - $f(t) = \lambda e^{-\lambda t}$
  - $F(t) = 1 - e^{-\lambda t}$
  - $h(t) = \lambda$; **constant hazard**
  - $H(t) = \lambda t$

Exponential distribution

- $E(T) = \lambda^{-1}$
  The higher the hazard, the shorter the expected survival time.
- $\text{Var}(T) = \lambda^{-2}$.
- Memoryless property: $\text{pr}(T > t) = \text{pr}(T > t + s | T > s), t, s > 0$.
- $c_0 \times \text{EXP}(\lambda) \sim \text{EXP}(\lambda / c_0)$ for $c_0 > 0$.
- The log-transformed exponential distribution is the so called extreme value distribution.

Gamma distribution

- Gamma distribution is a generalization of the simple exponential distribution.
- Be careful about the parametrization $G(\alpha, \lambda), \alpha, \lambda > 0$:
  1. The density function

$$f(t) = \frac{\lambda^{\alpha} t^{\alpha-1} e^{-\lambda t}}{\Gamma(\alpha)} \propto t^{\alpha-1} e^{-\lambda t},$$

  where

$$\Gamma(\alpha) = \int_0^{\infty} t^{\alpha-1} e^{-t} dt$$

  is the Gamma function. For integer α, $\Gamma(\alpha) = (\alpha - 1)!$.

  2. There is no close formulae for survival or hazard function.

Gamma distribution

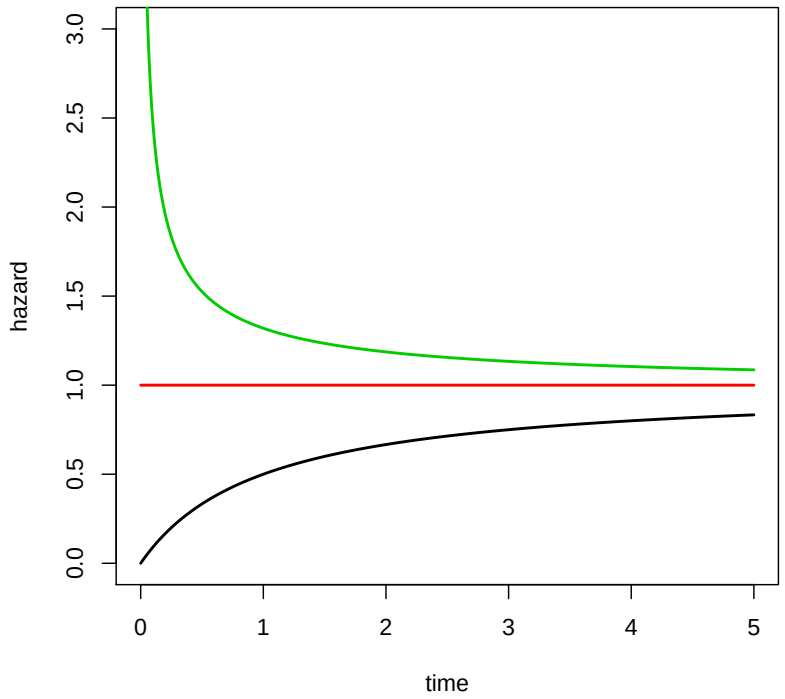
- $E(T) = \alpha\lambda^{-1}$.
- $\mathrm{Var}(T) = \alpha\lambda^{-2}$.
- If $T_i \sim G(\alpha_i, \lambda)$, $i = 1, \cdots, K$ and $T_i, i = 1, \cdots, K$ are independent, then

$$\sum_{i=1}^{K} T_i \sim G(\sum_{i=1}^{K} \alpha_i, \lambda).$$

- $G(1, \lambda) \sim \mathrm{EXP}(\lambda)$. (A generalization of the exponential distribution)

Gamma distribution

Figure 1: increasing hazard $\alpha > 1$; constant hazard $\alpha = 1$; decreasing hazard $0 < \alpha < 1$



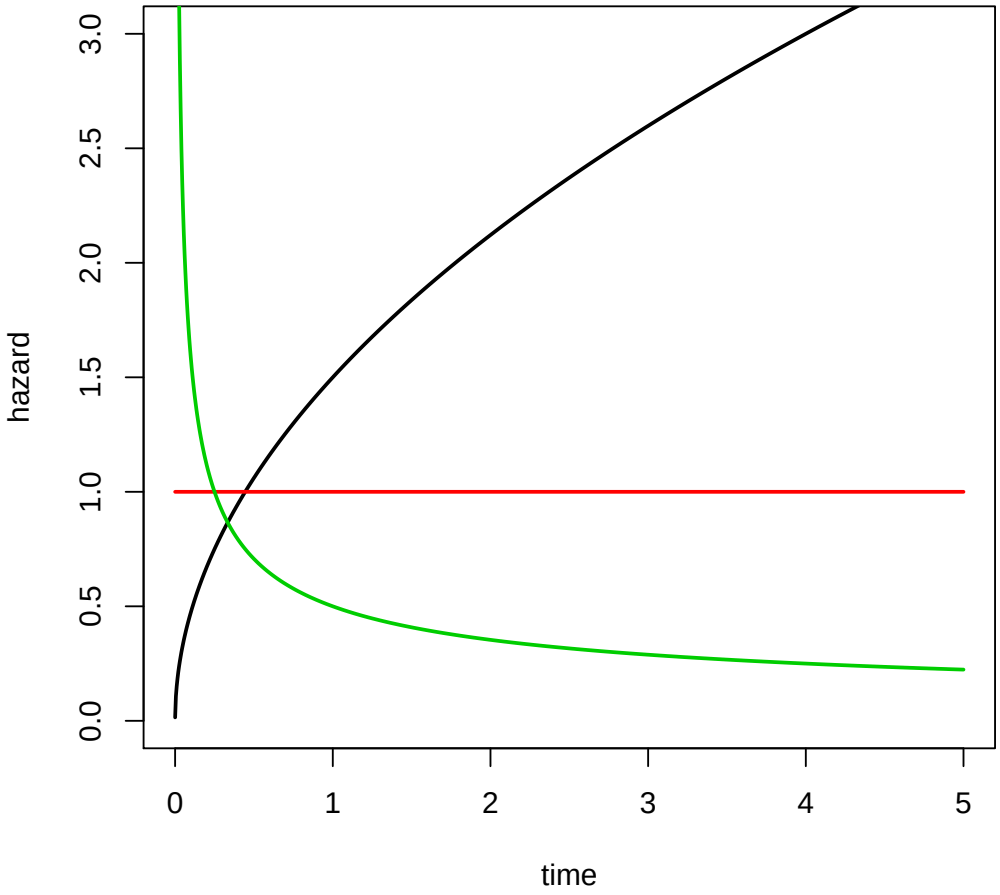
Weibull distribution

- Weibull distribution is also a generalization of the simple exponential distribution.
- Be careful about the parametrization $W(p, \lambda), \lambda > 0$(scale parameter) and $p > 0$(shape parameter):
  1. $S(t) = e^{-(\lambda t)^p}$
  2. $f(t) = p\lambda(\lambda t)^{p-1} e^{-(\lambda t)^p} \propto t^{p-1} e^{-(\lambda t)^p}$.
  3. $h(t) = p\lambda(\lambda t)^{p-1} \propto t^{p-1}$
  4. $H(t) = (\lambda t)^p$.

Weibull distribution

- $E(T) = \lambda^{-1}\Gamma(1 + 1/p)$.
- $\mathrm{Var}(T) = \lambda^{-2}\left[\Gamma(1 + 2/p) - \{\Gamma(1 + 1/p)\}^2\right]$
- $W(1, \lambda) \sim \mathrm{EXP}(\lambda)$.
- $W(p, \lambda) \sim \{\mathrm{EXP}(\lambda^p)\}^{1/p}$

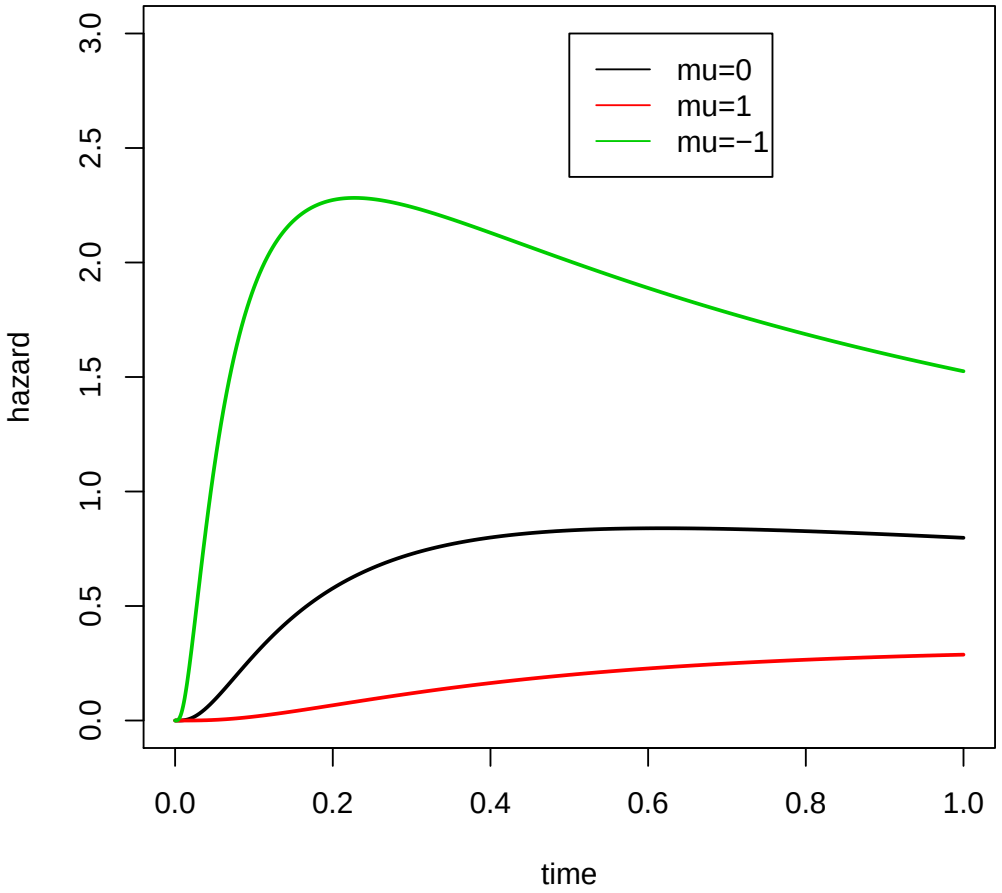
Hazard function of the Weibull distribution



Log-normal distribution

- The log-normal distribution is another commonly used parametric distribution for characterizing the survival time.
- $LN(\mu, \sigma^2) \sim \exp\{N(\mu, \sigma^2)\}$
- $E(T) = e^{\mu+\sigma^2/2}$
- $\text{Var}(T) = e^{2\mu+\sigma^2}(e^{\sigma^2} - 1)$

The hazard function of the log-normal distribution



Generalized gamma distribution

- The generalized gamma distribution becomes popular due to its flexibility.
- Again be careful about its parametrization $GG(\alpha, p, \lambda)$:
  - $f(t) = p\lambda(\lambda t)^{\alpha-1}e^{-(\lambda t)^p}/\Gamma(\alpha/p) \propto t^{\alpha-1}e^{-(\lambda t)^p}$
  - $S(t) = 1 - \gamma\{\alpha/p, (\lambda t)^p\}/\Gamma(\alpha/p)$, where

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

    is the incomplete gamma function.

Generalized gamma distribution

- For $k = 1, 2, \cdots$

$$E(T^k) = \frac{\Gamma\{(\alpha + k)/p\}}{\lambda^k \Gamma(\alpha/p)}$$

- If $p = 1$, $GG(\alpha, 1, \lambda) \sim G(\alpha, \lambda)$
- if $\alpha = p$, $GG(p, p, \lambda) \sim W(p, \lambda)$
- if $\alpha = p = 1$, $GG(1, 1, \lambda) \sim \text{EXP}(\lambda)$
- The generalized gamma distribution can be used to test the adequacy of commonly used Gamma, Weibull and Exponential distributions, since they are all nested within the generalized gamma distribution family.

Homogeneous Poisson Process

- $N(t) =$# events occurring in $(0, t)$
- T_1 denotes the time to the first event;
  T_2 denotes the time from the first to the second event
  T_3 denotes the time from the second to the third event et al.
- If the gap times $T_1, T_2, \cdots$ are i.i.d EXP(λ), then

$$N(t+s) - N(t) \sim Poisson(\lambda s).$$

  The process $N(t)$ is called the homogeneous Poisson process.

- The interpretation of the intensity function (similar to hazard function)

$$\lim_{\epsilon \downarrow 0} \frac{\text{pr}\{N(t+\epsilon) - N(t) > 0\}}{dt} = \lambda$$

Censoring

- A common feature of survival data is the presence of censoring.
- There are different types of censoring. Suppose that $T_1, T_2, \cdots, T_n$ are i.i.d survival times.
  1. Type I censoring: observe only

$$(U_i, \delta_i) = \{\min(T_i, c), I(T_i \leq c)\}, i = 1, \cdots, n,$$

     i.e,, we only have the survival information up to a fixed time c.

  2. Type II censoring: observe only

$$T_{(1,n)}, T_{(2,n)}, \cdots, T_{(r,n)}$$

     where $T_{(i,n)}$ is the ith smallest survival time, i.e, we only observe the first r smallest survival times.

3. Random censoring (**The most common type of censoring**): $C_1, C_2, \cdots, C_n$ are potential censoring times for n subjects, observe only

$$(U_i, \delta_i) = \{\min(T_i, C_i), I(T_i \leq C_i)\}, i = 1, \cdots, n.$$

   We often treat the censoring time C_i as i.i.d. random variables in statistical inferences.

4. Interval censoring: observe only $(L_i, U_i), i = 1, \cdots, n$ such that $T_i \in [L_i, U_i)$.

Non-informative Censoring

- If T_i and C_i are independent, then censoring is non-informative.
- Examples of informative and non-informative censoring.
  1. administrative censoring
  2. random drop off
  3. competing risk

Non-informative Censoring

- Noninformative censoring condition:

$$h(t) = \lim_{\epsilon \downarrow 0} \frac{\text{pr}(T \in [t, t+\epsilon] | T \geq t, C \geq t)}{\epsilon}$$

- It is slightly weaker than the independence between T and C.
- Consequences of informative censoring:
  - There are more than one distribution for (T, C) with different marginal distribution of T correspond to the same distribution of $(U, \delta) = \{\min(T, C), I(T < C)\}$.
  - Based on the distribution (U, δ) alone, it is impossible to determine the distribution of T.

Likelihood Construction

- In the presence of right censoring, we only observe $(U_i, \delta_i), i = 1, \cdots, n$.
- The likelihood construction must be with respect to the bivariate random variable $(U_i, \delta_i), i = 1, \cdots .n$.
  1. If $(U_i, \delta_i) = (u_i, 1)$, then $T_i = u_i, C_i > u_i$
  2. If $(U_i, \delta_i) = (u_i, 0)$, then $T_i \geq u_i, C_i = u_i$.

Likelihood Construction

- Assuming $C_i, 1 \leq i \leq n$ are i.i.d random variables with a CDF $G(\cdot)$.

$$L_i(F,G) = \begin{cases} f(u_i)(1-G(u_i)), \text{ if } \delta_i = 1 \\ S(u_i)g(u_i), \text{ if } \delta_i = 0 \end{cases}$$

$$\Rightarrow L(F,G) = \prod_{i=1}^{n} L_i(F) = \prod_{i=1}^{n} \left[\{f(u_i)(1-G(u_i))\}^{\delta_i} \{S(u_i)g(u_i)\}^{1-\delta_i}\right].$$

$$= \prod_{i=1}^{n} L_i(F,G) = \left\{\prod_{i=1}^{n} f(u_i)^{\delta_i} S(u_i)^{1-\delta_i}\right\} \left\{\prod_{i=1}^{n} g(u_i)^{1-\delta_i} (1-G(u_i))^{\delta_i}\right\}.$$

Likelihood Construction

- We have used the noninformative censoring assumption in the likelihood construction.
- $L(F, G) = L(F) \times L(G)$ and therefore the likelihood-based inference for F can be made based on

$$L(F) = \prod_{i=1}^{n} \left\{ f(u_i)^{\delta_i} S(u_i)^{1-\delta_i} \right\} = \prod_{i=1}^{n} h(u_i)^{\delta_i} S(u_i)$$

only.

Parametric inference: One-Sample Problem

- Suppose that $T_1, \cdots, T_n$ are i.i.d. EXP(λ) and subject to noninformative right censoring.

$$L(\lambda) = \prod_{i=1}^{n} \lambda^{\delta_i} e^{-\lambda u_i} = \lambda^r e^{-\lambda W},$$

where

1. $R = \sum_{i=1}^{n} \delta_i =$ #failures
2. $W = \sum_{i=1}^{n} u_i =$ total follow up time

- The score function: $\partial l(\lambda)/\partial\lambda = R/\lambda - W.$
- The observed information: $-\partial^2 l(\lambda)/\partial\lambda^2 = R/\lambda^2$

Parametric inference: One-Sample Problem

- $\hat{\lambda} = R/W$

  In epidemiology, the incidence rate is often estimated by the ratio of total events and total exposure time, which is the MLE for the constant hazard under the the exponential distribution.

- The information: $\hat{I}(\lambda) = R/\lambda^2$ and
  $I(\lambda) = E\{\hat{I}(\lambda)\} = n\text{pr}(C_i > T_i)/\lambda^2 = np/\lambda^2$.

- It follows from the property of MLE

$$\frac{\hat{\lambda} - \lambda}{\sqrt{\lambda^2/np}} \to N(0,1)$$

  in distribution as $n \to \infty$.

- $\hat{\lambda}$ approximately follows $N(\lambda, R/W^2)$ for large n.

Parametric inference: One-Sample Problem

- With $\delta-$method

$$\log(\hat{\lambda}) \sim N(\log(\lambda), R^{-1}),$$

where the variance R^{-1} is free of the unknown parameter λ.

- Hypothesis testing for $H_0 : \lambda = \lambda_0$

$$Z = \sqrt{R}\{\log(\hat{\lambda}) - \log(\lambda_0)\} \sim N(0, 1)$$

under H_0. We will reject the null hypothesis if $|Z|$ is too big.

- Confidence interval:

$$\text{pr}\left[\log(\hat{\lambda}) - 1.96\sqrt{1/R} < \log(\lambda) < \log(\hat{\lambda}) + 1.96\sqrt{1/R}\right] \approx 0.95,$$

which suggests that the .95 confidence interval for λ is

$$\left(\hat{\lambda}e^{-1.96/\sqrt{R}}, \hat{\lambda}e^{1.96/\sqrt{R}}\right).$$

PBC Example

Mayo Clinic Primary Biliary Cirrhosis Data

This data is from the Mayo Clinic trial in primary biliary cirrhosis (PBC) of the liver conducted between 1974 and 1984. A total of 424 PBC patients, referred to Mayo Clinic during that ten-year interval, met eligibility criteria for the randomized placebo controlled trial of the drug D-penicillamine. The first 312 cases in the data set participated in the randomized trial and contain largely complete data. The additional 106 cases did not participate in the clinical trial, but consented to have basic measurements recorded and to be followed for survival.

Parametric inference: One Sample Problem

- In the PBC example, $R = 186$, $W = 801633$ days, $\hat{\lambda}$ is 0.0232% and the 95% CI for λ is $[0.0201, 0.0268]$%.
- The value of λ depends on the unit of T!

Parametric inference: One-Sample Problem

```
library(survival)
head(pbc)
u=pbc$time
delta=1*(pbc$status>0)
r=sum(delta)
W=sum(u)
lambdahat=r/W
c(lambdahat,
  lambdahat*exp(-1.96/sqrt(r)),
  lambdahat*exp(1.96/sqrt(r)))
```

Parametric inference: Two-Sample Problem

- Compare the hazards λ_A and λ_B for two groups (e.g., treatment and placebo arms in a clinical trial).
- Test the null hypothesis $H_0 : \lambda_A = \lambda_B$.
  1. Under the exponential model, summarizing the data as (R_A, W_A) and (R_B, W_B).
  2. The test statistics

$$Z = \frac{\log(\hat{\lambda}_A) - \log(\hat{\lambda}_B)}{\sqrt{R_A^{-1} + R_B^{-1}}} \sim N(0,1) \quad \text{under} \quad H_0.$$

  We will reject H_0, if $|Z| > 1.96$.

- In the PBC example, $Z = 0.29$, which corresponds to a p value of 0.77.

Parametric inference: Two-Sample Problem

```
library(survival)
head(pbc)
u=pbc$time[1:312]
delta=1*(pbc$status>0)[1:312]
trt=pbc$trt[1:312]

r1=sum(delta[trt==1])
W1=sum(u[trt==1])
lambda1=r1/W1
r2=sum(delta[trt==2])
W2=sum(u[trt==2])
lambda2=r2/W2
z=(log(lambda1)-log(lambda2))/sqrt(1/r1+1/r2)
c(z, 1-pchisq(z^2, 1))
```

Parametric inference: Regression Problem

- $\mathbf{z}$ is a $p \times 1$ vector of covariates measured for each subject and we are interested in assessing the association between $\mathbf{z}$ and T
- Observed data: $(U_i, \delta_i, \mathbf{z}_i), i = 1, \cdots, n$
- Noninformative censoring: $T_i \perp C_i | \mathbf{Z}_i = \mathbf{z}_i$ (different from $T_i \perp C_i$)
- What is the appropriate statistical model to link T_i with $\mathbf{z}_i$?

Parametric inference: Regression Problem

- If we observe $(T_i, \mathbf{z}_i), i = 1, \cdots n$, then the linear regression model can be used:

$$\log(T_i) = \beta' \mathbf{z}_i + \epsilon_i,$$

  which is called the accelerated failure time (AFT) model in survival analysis, where ϵ_i follows a parametric distribution such as the extreme value distribution.

- Model the association between the hazard function and covariates $\mathbf{z}_i$
  - $T_i \sim \text{EXP}(\lambda_i)$
  - $\log(\lambda_i) = \log(\lambda_0) + \beta' \mathbf{z}_i$
- The likelihood function can be used to derive the MLE and make the corresponding statistical inferences.

Parametric Inference

- In practice, the exponential distribution is rarely used due to its over-simplicity: one parameter λ characterizes the entire distribution.
- Alternatives such as Weibull, Gamma, Generalized Gamma distribution and log-normal distribution are more popular, but they also put specific constraints on the hazard function.
- An intermediate model from parametric to nonparametric model is the "piecewise exponential" distribution.

Parametric Inference

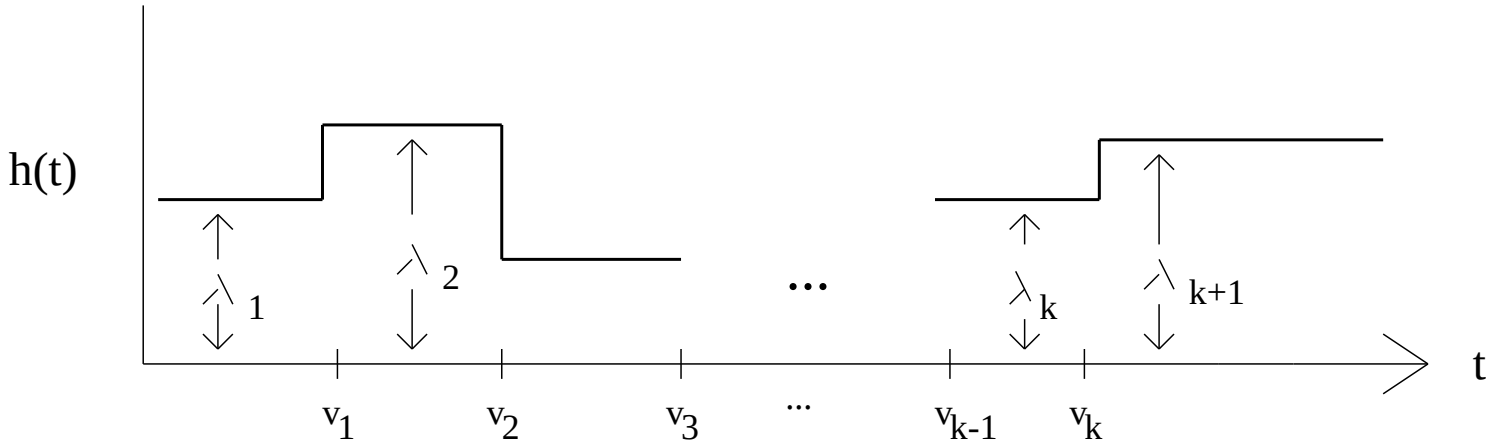
- $T_1, \cdots, T_n$ i.i.d random variables
- Suppose that the hazard function of T is in the form of

$$h(t) = \lambda_j \quad \text{for} \quad v_{j-1} \leq t < v_j,$$

where $0 = v_0 < v_1 < \cdots < v_k < v_{k+1} = \infty$ are given cut-off values.

Parametric Inference

Figure 2: The hazard function of piece-wise exponential



Parametric Inference

- $L(\lambda_1, \cdots, \lambda_{k+1}) = \cdots$
- Let R_j and W_j be the total number of events and follow-up time within the interval $[v_j, v_{j+1})$, respectively. $\hat{\lambda}_j = R_j/W_j$
- $\log(\hat{\lambda}_1), \cdots, \log(\hat{\lambda}_{k+1})$ are approximately independently distributed with $\log(\hat{\lambda}_j) \sim N(\log(\lambda_j), R_j^{-1})$ for large n.
- The statistical inference for the hazard function follows.