STAT3655 Survival Analysis

Yu Gu, PhD
Assistant Professor

Department of Statistics & Actuarial Science
The University of Hong Kong

Table of Contents

Table of Contents

Table of Contents

Stochastic Processes

- A stochastic process is a collection of random variables

$$X = \{X(t) : \ t \in \Gamma\}$$

  indexed by a set Γ.
  - $X(t)$ is a random variable for each t
  - Γ is regarded as time, either $\{0, 1, 2, \cdots\}$ (discrete-time process) or $[0, \infty)$ (continuous-time process)

- The realization of $X(t)$ (seen as a function as t) is called the sample path.

- A stochastic process is called increasing or right-continuous if its sample paths have the corresponding property.

Counting Process

- Counting process is a continuous-time stochastic process $\{N(t) : \ t \geq 0\}$ with $N(0) = 0$ whose sample paths are right-continuous, piecewise constant and have jumps of size 1 only.
- In survival analysis, $N(t)$ records the number of failures observed within the time interval $[0, t]$.
- Without censoring, $N(t) = I(T \leq t)$.
- In the presence of censoring, $N(t) = I(Y \leq t, \delta = 1) = \delta I(Y \leq t)$.

Gaussian Process

A continuous-time stachastic process $\{X(t) : \ t \in \Gamma\}$ is Gaussian if and only if for every finite set of indices $t_1, \ldots, t_m \in \Gamma$, $(X(t_1), \ldots, X(t_m))$ is a multivariate Gaussian random variable.

Martingale

- A right-continuous stochastic process $\{X(t) : \ t \in \Gamma\}$ is a martingale with respect to the filtration $\mathcal{F}_t$ if

$$E\{X(t+s) \mid \mathcal{F}_t\} = X(t), \forall t \geq 0, s \geq 0$$

  - $\mathcal{F}_t$ is generated by the stochastic process up to t: $\mathcal{F}_t = \sigma\{X(s) : 0 \leq s \leq t\}$.
  - $\mathcal{F}_t$ represents the available data at time t or the past history up to t.
  - $s \leq t \Rightarrow \mathcal{F}_s \subset \mathcal{F}_t$.

- Properties of a martingale:
  (i) $E\{X(t+s)\} = E\{X(t)\}, \quad \forall t \geq 0, s \geq 0.$
  (ii) $\text{Cov}\{X(t+s), X(t)\} = \text{Var}\{X(t)\}, \quad \forall t \geq 0, s \geq 0.$

- In survival analysis, $\mathcal{F}_t$ is the information of failure and censoring over the interval $[0, t]$.

Table of Contents

Parametric vs Nonparametric Estimation

	Parametric	**Nonparametric**
Survival function	$S(t;\theta)$	arbitrary $S(t)$
Parameter	θ (finite-dimensional)	function (infinite-dimensional)
Estimation	MLE on θ	nonparametric methods such as NPMLE and kernel smoothing
Computation	easy	more challenging
Flexibility	restrictive	flexible

Empirical Survival Function

- A useful way of portraying survival data is to compute and plot the empirical survival function (ESF), which is a nonparametric estimator of the survival function of the failure time T.

- When there is no censoring, ESF is simply defined as

$$\widehat{S}(t) = \frac{\#\text{ failures after time } t}{\text{total number of subjects}}, \quad t \geq 0.$$

- This is just a step function which takes jumps at all distinct failure times observed in the data.

- If there are d failures recorded at time t among all n subjects, then $\widehat{S}(t)$ drops by d/n at t.

Empirical Survival Function (Cont.)

- It can be easily observed that $\widehat{S}(t)$ is a binomial proportion of subjects still alive at time t:

$$\widehat{S}(t) = n^{-1} \sum_{i=1}^{n} I(T_i > t),$$

with mean $S(t)$ and variance $S(t)\{1 - S(t)\}/n$.

- For example, consider $n = 35$ patients with colon cancer, 8 of them died during the first year of follow-up. The estimator of 1-year survival probability is

$$\widehat{S}(1) = (35 - 8)/35 = 0.771,$$

with an estimated standard error of

$$\sqrt{0.771 \times (1 - 0.771)/35} = 0.071.$$

When Censoring Exists

- Suppose that 10 patients died within 2 years of follow-up, but 2 patients were censored at 13 and 14 months, respectively. How can we estimate the 2-year survival probability?

- Excluding the two censored patients from the analysis, we have $\widehat{S}(2) = (33 - 10)/33 = 0.697$. This will underestimate the true survival probability since we ignore the fact that each of these two patients were at risk of death between one and two years but did not die while under observation.

- If we use $\widehat{S}(2) = (35 - 10)/35 = 0.714$, it will overestimate the true survival probability since we are assuming that both two censored patients survived beyond the 2-year follow-up.

- Two common methods for the estimation of $S(t)$ in the presence of censoring are the Kaplan-Meier method and the life-table method.

Kaplan-Meier Estimator

- A heuristic derivation of the KM estimator has been discussed in Chapter 1. Here, we study its rigorous derivation based on MLE.

- Notation:
  - $t_1 < t_2 < \cdots < t_K$: distinct failure times observed in the data
  - d_k: number of failures occurring at time t_k $(k = 1, \ldots, K)$
  - c_{kl} $(l = 1, \ldots, m_k)$: all censoring times observed within $[t_k, t_{k+1})$

- The likelihood function can be written as

$$
\begin{aligned}
L &= \prod_{k=1}^{K} \left\{ \Pr(T = t_k)^{d_k} \times \prod_{l=1}^{m_k} \Pr(T > c_{kl}) \right\} \\
&= \prod_{k=1}^{K} \left[\left\{ S(t_k^-) - S(t_k) \right\}^{d_k} \times \prod_{l=1}^{m_k} S(c_{kl}) \right]
\end{aligned}
$$

Kaplan-Meier Estimator (Cont.)

- We adopt the nonparametric maximum likelihood estimation (NPMLE) method and treat $S(t)$ as a step function with jumps only at $t_1 < t_2 < \cdots < t_K$.

- In addition, we consider the reparameterization

$$1 - \lambda_k = \frac{S(t_k)}{S(t_{k-1})} \quad \Rightarrow \quad S(t_k) = \prod_{l=1}^{k}(1 - \lambda_l).$$

- Under this set-up, the likelihood becomes

$$\begin{aligned} L &= \prod_{k=1}^{K}\left[\left\{S(t_{k-1}) - S(t_k)\right\}^{d_k} \times S(t_k)^{m_k}\right] \\ &= \prod_{k=1}^{K}\left[\left\{\lambda_k \prod_{l=1}^{k-1}(1-\lambda_l)\right\}^{d_k} \times \left\{\prod_{l=1}^{k}(1-\lambda_l)\right\}^{m_k}\right] \end{aligned}$$

Kaplan-Meier Estimator (Cont.)

- Maximizing the likelihood function with respect to λ_k yields

$$\widehat{\lambda}_k = \frac{d_k}{r_k} \quad \Rightarrow \quad \widehat{S}(t) = \prod_{k:t_k \leq t} (1 - \frac{d_k}{r_k}) \quad \text{(KM estimator)},$$

  where $r_k = \sum_{l=k}^{K} (d_l + m_l)$ is the number of subjects at risk at time t_k.

- A slightly problematic point of the KM estimator $\widehat{S}(t)$ is that it never reduces to zero. Thus, $\widehat{S}(t)$ is usually taken to be undefined for $t > \tau$, where τ is the study end time.

Variance of the KM Estimator

- The asymptotic variance of $\widehat{S}(t)$ can be estimated by the inverse of the information matrix

$$\widehat{\text{Var}}(\widehat{\lambda}_k) = \left\{ -\frac{\partial^2 \ell}{\partial \lambda_k^2}\bigg|_{\lambda_k = \widehat{\lambda}_k} \right\}^{-1} = \frac{d_k(r_k - d_k)}{r_k^3}.$$

- Since $\log \widehat{S}(t) = \sum_{k:t_k \le t} \log(1 - \widehat{\lambda}_k)$, applying the Delta method yields

$$\begin{aligned} \widehat{\text{Var}}\left\{\log \widehat{S}(t)\right\} &= \sum_{k:t_k \le t} \frac{1}{(1-\widehat{\lambda}_k)^2} \widehat{\text{Var}}(\widehat{\lambda}_k) \\ &= \sum_{k:t_k \le t} \frac{d_k}{r_k(r_k - d_k)}. \end{aligned} \tag{1}$$

- Exponentiating $\log \widehat{S}(t)$ and applying the Delta method again, we obtain

$$\widehat{\text{Var}}\left\{\widehat{S}(t)\right\} = \widehat{S}(t)^2 \sum_{k:t_k \le t} \frac{d_k}{r_k(r_k - d_k)} \quad \text{(Greenwood's Formula)} \tag{2}$$

Limiting Distribution of the KM Estimator

Theorem (Limiting distribution of $\widehat{S}(t)$)

Under mild conditions on the censoring mechanism, the process $G(t) = \sqrt{n}\{\widehat{S}(t) - S(t)\}$ *converges weakly to a mean-zero Gaussian process whose covariance function can be consistently estimated by*

$$\widehat{Cov}\{G(s), G(t)\} = n\widehat{S}(s)\widehat{S}(t) \sum_{k:t_k \leq \min(s,t)} \frac{d_k}{r_k(r_k - d_k)}.$$

Confidence Interval of $S(t)$

- By the asymptotic normality of $\widehat{S}(t)$, a 95% CI for $S(t)$ is given by

$$\widehat{S}(t) \pm 1.96\sqrt{\widehat{\mathrm{Var}}\left\{\widehat{S}(t)\right\}}.$$

- To avoid impossible values outside the range $[0,1]$, we apply the log-log transformation $\widehat{B}(t) = \log\{-\log \widehat{S}(t)\}$. By (1) and the Delta method,

$$\widehat{\mathrm{Var}}\left\{\widehat{B}(t)\right\} = \frac{1}{\left\{\log \widehat{S}(t)\right\}^2}\widehat{\mathrm{Var}}\left\{\log \widehat{S}(t)\right\} = \frac{\sum_{k:t_k \le t} \frac{d_k}{r_k(r_k - d_k)}}{\left\{\sum_{k:t_k \le t} \log\left(\frac{r_k - d_k}{r_k}\right)\right\}^2}.$$

- Therefore, a 95% CI for $B(t)$ is given by $\widehat{B}(t) \pm 1.96\sqrt{\widehat{\mathrm{Var}}\{\widehat{B}(t)\}}$. Exponentiation twice yields a 95% CI for $S(t)$ by

$$\widehat{S}(t)^{\exp\left[\pm 1.96\sqrt{\widehat{\mathrm{Var}}\{\widehat{B}(t)\}}\right]}$$

Nelson-Aalen Estimator

- Recall that in Chapter 1, we have also studied the NA estimator for $\Lambda(t)$:

$$\widehat{\Lambda}(t) = \sum_{k:t_k \le t} \frac{d_k}{r_k} \quad \text{(NA estimator)}$$

- By the martingale theory for counting processes, we can establish the limiting distribution of $\widehat{\Lambda}(t)$.

Theorem (Limiting distribution of $\widehat{\Lambda}(t)$)

Under mild conditions on the censoring mechanism, the process $M(t) = \sqrt{n}\{\widehat{\Lambda}(t) - \Lambda(t)\}$ *converges weakly to a mean-zero Gaussian martingale whose variance function can be consistently estimated by*

$$\widehat{Var}\{M(t)\} = \sum_{k:t_k \le t} \frac{n d_k}{r_k^2} \quad \Rightarrow \quad \widehat{Var}\left\{\widehat{\Lambda}(t)\right\} = \sum_{k:t_k \le t} \frac{d_k}{r_k^2}$$

Compare NA and KM Estimators

- Another estimator for $\Lambda(t)$ is based on the KM estimator:

$$\widetilde{\Lambda}(t) = -\log \widehat{S}(t) = -\sum_{k:t_k \le t} \log(1 - \frac{d_k}{r_k}),$$

  with variance estimator [by (1)]

$$\widehat{\text{Var}}\left\{\widetilde{\Lambda}(t)\right\} = \sum_{k:t_k \le t} \frac{d_k}{r_k(r_k - d_k)}$$

- The variance of $\widehat{\Lambda}(t)$ is slightly smaller than that of $\widetilde{\Lambda}(t)$. Thus, the NA estimator has better finite-sample performance (more efficient) and is more commonly used.

Life Table

- Life tables are often used in actuarial applications, where the survival data are grouped into successive time intervals $I_1, I_2, \ldots, I_J$.

- The life table presents the number of failures and censored survival times falling within each interval.

- Notation ($j = 1, \ldots, J$):
  - d_j: number of observed failure times within the interval I_j
  - m_j: number of censored failure times within the interval I_j
  - $r_j = \sum_{k \geq j}(d_k + m_k)$: number of subjects at risk at the start of I_j

- The standard life-table estimator of the conditional probability of failure in I_j given survival to enter I_j is

$$\widehat{q}_j = \begin{cases} \frac{d_j}{r_j - m_j/2} & \text{if } r_j > 0, \\ 1 & \text{otherwise.} \end{cases}$$

Life Table (Cont.)

- The $m_j/2$ term in the denominator is used to adjust for the fact that not all of the r_j subjects are at risk for the whole of I_j.

- The corresponding life-table estimator of the survival function at the end of I_j is

$$\widehat{S}_j = \prod_{k=1}^{j} (1 - \widehat{q}_k).$$

  The variance estimator of $\widehat{S}_j$ is given by Greenwood's formula (2), with r_j replaced by $r_j - m_j/2$.

- The life-table method is designed primarily for situations where actual failure and censoring times are unavailable but only numbers of failures and censored subjects are known for each time interval.

Table of Contents

Estimation of Quantiles

- Quantiles of the failure time T are most conveniently estimated by the graphical method based on the plot of the estimated survival function such as the KM estimator.

- To obtain graphical estimates of the pth quantile Q_p, say the median $Q_{0.5}$, we search for the first time point at which the estimated survival probability attains 0.5.

- Mathematically, the estimator of the pth quantile is given by

$$\widehat{Q}_p = \min\left\{t : \widehat{S}(t) \leq 1 - p\right\}$$

- $\widehat{S}(t)$ refers to the KM estimator hereafter.

Variance of the Quantile Estimator

- By the large-sample theory for $\widehat{S}(t)$ and the Delta method, $\widehat{Q}_p$ is asymptotically normal, with mean Q_p and variance estimator

$$\widehat{\text{Var}}(\widehat{Q}_p) = \frac{\widehat{\text{Var}}\left\{\widehat{S}(\widehat{Q}_p)\right\}}{\left\{\hat{f}(\widehat{Q}_p)\right\}^2}$$

- The numerator can be computed using the Greenwood's formula. The denominator is commonly estimated by

$$\hat{f}(\widehat{Q}_p) = \frac{\widehat{S}(\hat{l}_p) - \widehat{S}(\hat{u}_p)}{\hat{u}_p - \hat{l}_p}.$$

- The values $\hat{l}_p$ and $\hat{u}_p$ satisfy $\hat{l}_p < \widehat{Q}_p < \hat{u}_p$ and are most often chosen by

$$\hat{l}_p = \max\left\{t : \widehat{S}(t) \geq 1 - p + c\right\} \text{ and } \hat{u}_p = \min\left\{t : \widehat{S}(t) \leq 1 - p - c\right\},$$

where c is a positive constant and is usually taken as 0.05.

Confidence Interval of Quantiles

- We may construct the $(1-\alpha)\times 100\%$ CI for the quantile Q_p by

$$\widehat{Q}_p \pm z_{1-\alpha/2}\sqrt{\widehat{\mathrm{Var}}(\widehat{Q}_p)}$$

- However, the above CI requires to estimate the density function at Q_p and depends on the assumption that $\widehat{Q}_p$ is normal. The sensitivity of the CI to the choice of the density estimator and the normal assumption has not been studied.

- An alternative $(1-\alpha)\times 100\%$ CI for Q_p consists of all the values of t such that

$$\left|\frac{\widehat{S}(t)-(1-p)}{\sqrt{\widehat{\mathrm{Var}}\left\{\widehat{S}(t)\right\}}}\right| \leq z_{1-\alpha/2}, \tag{3}$$

  which can be obtained easily by the graphical method described previously.

Estimation of the Mean

- In the presence of censoring, nonparametric estimation of the mean of the failure time T has been extremely difficult and inaccurate.
- We can only estimate the mean of T within the interval $[0, \tau]$, where τ is usually the largest observed failure time t_K.
- This estimate may not make much sense in practice and we would not discuss much on it.
- Instead, we usually estimate the restricted mean survival time (RMST), defined as

$$\text{RMST} = E\{\min(T, \tau)\} = \int_0^\tau S(t)dt.$$

Table of Contents

Multi-Group Comparison

- Suppose that there are G groups of subjects, whose failure times are iid within each group with survival functions $S_1, S_2, \ldots, S_G$, respectively.
- We want to test whether the G survival functions are all the same, i.e.,

$$H_0 : S_1(t) = S_2(t) = \cdots = S_G(t) \quad \text{for all } t > 0$$

- In Chapter 1, we have considered the simplest case with $G = 2$ and studied the log-rank test. The basic idea is similar here.

$2 \times G$ Contingency Table

- Let $t_1 < t_2 < \cdots < t_K$ be the distinct failure times from all G groups.
- At each observed failure time t_k $(k = 1, \ldots, K)$, we create a $2 \times G$ contingency table.

Group	1	2	$\cdots$	G	Total
Failures	d_{1k}	d_{2k}	$\cdots$	d_{Gk}	d_k
Non-failures	$r_{1k} - d_{1k}$	$r_{2k} - d_{2k}$	$\cdots$	$r_{Gk} - d_{Gk}$	$r_k - d_k$
At risk	r_{1k}	r_{2k}	$\cdots$	r_{Gk}	r_k

- Under H_0, the conditional distribution of $(d_{1k}, \ldots, d_{Gk})$ given the marginals is the multivariate hypergeometric distribution, with conditional mean and covariance

$$e_{gk} = \frac{d_k r_{gk}}{r_k} \quad \text{and} \quad v_{ghk} = \frac{d_k(r_k - d_k) r_{gk} \{r_k I(g = h) - r_{hk}\}}{r_k^2 (r_k - 1)},$$

for $g, h = 1, \ldots, G$.

Weighted Log-Rank Test

- For $g = 1, \ldots, G$, define the statistic Z_g for Group g to be a weighted sum of $(O - E)$ over all K failure times:

$$Z_g = \sum_{k=1}^{K} w(t_k)(d_{gk} - e_{gk}),$$

  where $w(t)$ is a prespecified bounded nonnegative weight function.

- It can be shown that the statistic

$$Z = (Z_1, Z_2, \ldots, Z_{G-1})^{\mathsf{T}} \xrightarrow{d} N_{G-1}(0, \Sigma),$$

  with the (g, h)th element of Σ given by $\sum_{k=1}^{K} w(t_k)^2 v_{ghk}$.

- The weighted log-rank statistic is thus

$$Q^2 = Z^{\mathsf{T}} \Sigma^{-1} Z \xrightarrow{d} \chi^2_{G-1} \quad \text{under } H_0.$$

Remarks

- When $G = 2$ and $w(t) \equiv 1$, Q reduces to the two-sample log-rank statistic discussed in Chapter 1.
- Some commonly used weight functions are listed below.

Test	$w(t_k)$
Log-rank	1
Gehan	r_k
Tarone-Ware	$\sqrt{r_k}$
Prentice-Wilcoxon	$\widehat{S}(t_k^-)$
Harrington-Fleming	$\widehat{S}^\rho(t_k^-)$ for $\rho \geq 0$

- In general, the log-rank test and the Gehan test are most commonly used partly because they are available in most statistical software.

Table of Contents

Kernel Smoothed Estimator

- Based on the NA estimator for $\Lambda(t)$, a crude estimator of the hazard function $\lambda(t)$ is given by

$$\Delta\widehat{\Lambda}(t) = \begin{cases} \frac{d_k}{r_k} & \text{if } t = t_k \text{ for some } k \in \{1, \ldots, K\} \\ 0 & \text{otherwise} \end{cases}$$

- However, this crude estimator does not make much sense since it implies that a subject cannot fail at any other times than $t_1, \ldots, t_K$.

- We can use the kernel smoothing method to obtain a smooth estimator of $\lambda(t)$, which is a weighted average of the crude estimator over the distinct failure times close to t:

$$\hat{\lambda}(t) = \sum_{k=1}^{K} w_k(t) \frac{d_k}{r_k},$$

where $w_k(t)$ are kernel weights depending on $|t - t_k|$.

Kernel Smoothed Estimator (Cont.)

- Given a bandwidth h and a kernel function $K(\cdot)$ defined on $[-1, 1]$, the kernel weights are chosen as

$$w_k(t) = K_h(t - t_k) = h^{-1} K\left(\frac{t - t_k}{h}\right),$$

- This yields the kernel smoothed estimator for $\lambda(t)$:

$$\hat{\lambda}(t) = \sum_{k=1}^{K} K_h(t - t_k) \frac{d_k}{r_k} \quad \text{(Ramlau-Hansen estimator)}$$

- Since $K \equiv 0$ outside the interval $[-1, 1]$, only those distinct failure times within $[t - h, t + h]$ will contribute to $\hat{\lambda}(t)$.

Variance of $\hat{\lambda}(t)$

- The variance of $\hat{\lambda}(t)$ can be estimated by

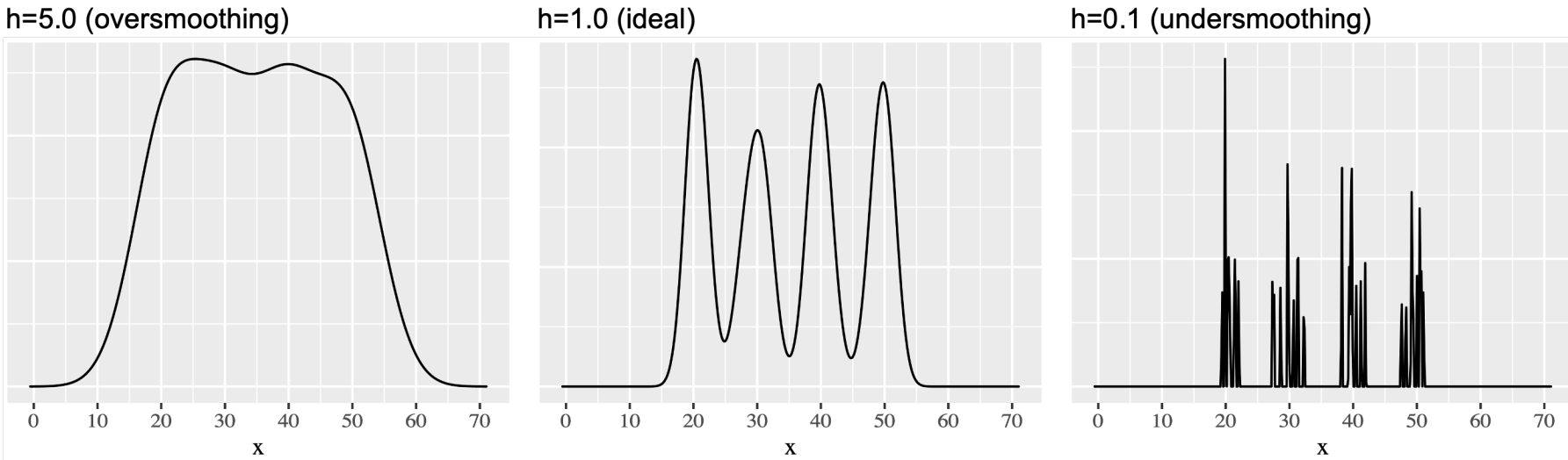
$$\begin{aligned}
\widehat{\text{Var}}\left\{\hat{\lambda}(t)\right\} &= \sum_{k=1}^{K} K_h^2(t-t_k)\widehat{\text{Var}}\left(\frac{d_k}{r_k}\right) \\
&= \sum_{k=1}^{K} K_h^2(t-t_k)\widehat{\text{Var}}\left\{\hat{\Lambda}(t_k)-\hat{\Lambda}(t_{k-1})\right\} \\
&= \sum_{k=1}^{K} K_h^2(t-t_k)\left[\widehat{\text{Var}}\left\{\hat{\Lambda}(t_k)\right\}-\widehat{\text{Var}}\left\{\hat{\Lambda}(t_{k-1})\right\}\right] \\
&= \sum_{k=1}^{K} K_h^2(t-t_k)\frac{d_k}{r_k^2}
\end{aligned}$$

- A $(1-\alpha)\times 100\%$ CI for $\lambda(t)$ can be obtained by the log transformation

$$\hat{\lambda}(t)\exp\left[\pm z_{1-\alpha/2}\sqrt{\widehat{\text{Var}}\{\hat{\lambda}(t)\}}/\hat{\lambda}(t)\right]$$

Bandwidth

- The bandwidth h determines the smoothness of $\hat{\lambda}(t)$. The bigger h, the smoother $\hat{\lambda}(t)$.



- Muller & Wang (1994) [1] suggested setting $h = c(t_K - t_1)D^{-1/5}$, where c is a tuning parameter, $D = \sum_{k=1}^{K} d_k$ is the total number of observed failures.

[1] *Muller H. G., Wang J. L. (1994). Hazard rate estimation under random censoring with varying kernels and bandwidths. Biometrics, 61–76.*

Kernel Function

- The kernel function $K(\cdot)$ must satisfy the following conditions:
  - (i) $\int_{-1}^{1} K(u)du = 1$ (weights sum up to 1)
  - (ii) $\int_{-1}^{1} u^{2\ell-1}K(u)du = 0$ for $\ell = 1, 2, \cdots$ (symmetric weights)
  - (iii) $\int_{-1}^{1} |u^p|K(u)du < \infty$ (finite moments)

- To some extent, kernel function can be viewed as a density function defined over $[-1, 1]$.

- Some common choices of $K(\cdot)$ are
  - Uniform kernel: $K(u) = \frac{1}{2}$ for $u \in [-1, 1]$
  - Epanechnikov kernel: $K(u) = 0.75(1 - u^2)$ for $u \in [-1, 1]$
  - Biweight kernel: $K(u) = \frac{15}{16}(1 - u^2)^2$ for $u \in [-1, 1]$

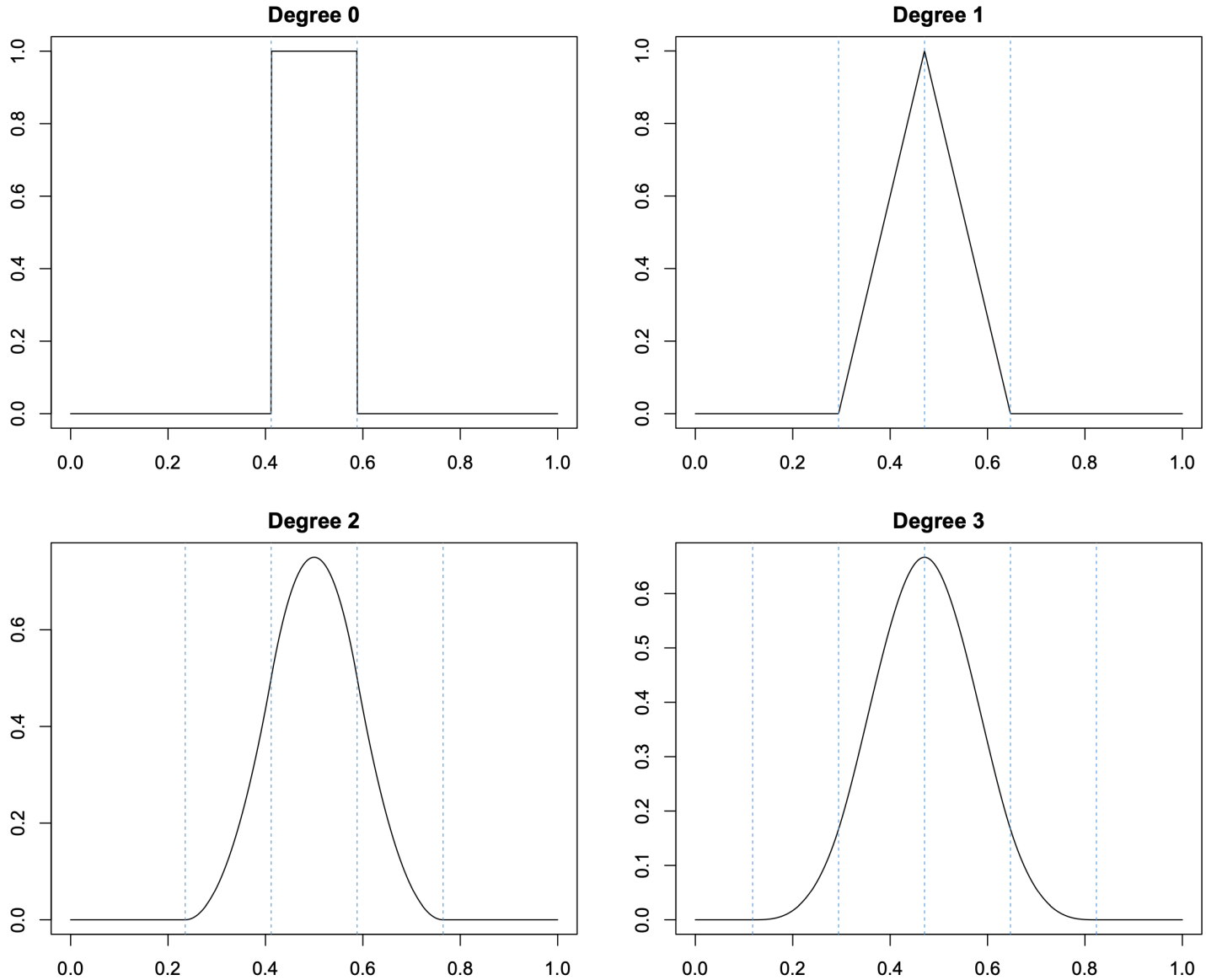
Alternative Method: B-Splines

- Alternatively, we can estimate the hazard function $\lambda(t)$ using B-splines.
- Tuning parameters in B-splines:
  - d: degree of spline basis functions (order $d+1$)
  - $0 < x_1 < x_2 < \cdots < x_G < \tau$: G internal knots that partition the study period $[0, \tau]$
- The hazard function is specified as a linear combination of $(G+d+1)$ spline basis functions:

$$\lambda(t) = \sum_{l=1}^{G+d+1} e^{\alpha_l} B_l(t)$$

  - e^{α_l}: unknown spline parameters. Exponentiation ensures that $\lambda(t)$ is nonnegative.
  - $B_l(t)$: spline basis functions, derived from the Cox-de Boor recursion formula [2]

[2] *De Boor, C. (2001). A Practical Guide to Splines. Springer-Verlag New York.*



B-splines of degrees 0 through 3. The knot points are marked by dashed blue vertical lines.

B-Splines (Cont.)

- Under B-splines, nonparametric estimation of $\lambda(t)$ reduces to parameteric estimation of the spline parameters $\alpha = (\alpha_1, \ldots, \alpha_{G+d+1})$, which can be done easily through maximum likelihood estimation.

- Specifically, the likelihood function is

$$L(\alpha) = \prod_{i=1}^{n} \left\{ \sum_{l=1}^{G+d+1} e^{\alpha_l} B_l(Y_i) \right\}^{\Delta_i} \exp\left\{ - \int_0^{Y_i} \sum_{l=1}^{G+d+1} e^{\alpha_l} B_l(t) dt \right\}$$

$$= \prod_{i=1}^{n} \left\{ \sum_{l=1}^{G+d+1} e^{\alpha_l} B_l(Y_i) \right\}^{\Delta_i} \exp\left\{ - \sum_{l=1}^{G+d+1} e^{\alpha_l} IB_l(Y_i) \right\},$$

  where $IB_l(t) = \int_0^t B_l(u) du$ is the lth integrated spline basis function.

- The MLE for α can be computed using the Newton-Raphson algorithm.

Penalized Maximum Likelihood Estimation

- In practice, it may occur that $\alpha_l \to -\infty$ for some l, indicating that the corresponding basis functions are not needed (since $e^{\alpha_l} \to 0$).

- To avoid numerical difficulties caused by sparsity of e^{α}, we can add a penalty term to the log-likelihood function, i.e.,

$$\hat{\alpha} = \arg\max_{\alpha} \ell(\alpha) - p_\eta(\alpha),$$

  where $p_\eta(\alpha)$ is some penalty function for α and $\eta > 0$ is a tuning parameter.

- A proper penalty function should encourage small values of e^{α_l}.
  - $p_\eta(\alpha) = \eta \sum_{l=1}^{G+d+1} (10 + \alpha_l)_+^3$ when $d = 3$ (Rosenberg, 1995) [3].
  - $p_\eta(\alpha) = \eta \sum_{l=1}^{G+d+1} e^{r\alpha_l}$ (r=1: Lasso penalty; r=2: Ridge penalty)

[3] *Rosenberg, P. S. (1995). Hazard function estimation using B-splines. Biometrics, 874–887.*

Selection of Tuning Parameters

- Theoretically, the internal knots should be equally distributed. In practice, however, we often place the knots at equal quantiles of the distinct failure times $t_1, \ldots, t_K$.

- The degree d is often taken as 1, 2, or 3, which corresponds to linear, quadratic, or cubic splines (most commonly used), respectively.

- The optimal values for d and G can be determined by the Akaike information criterion (AIC):

$$(d, G)^{(\text{opt})} = \arg\min_{(d,G)} \text{AIC} = \arg\min_{(d,G)} 2(G + d + 1) - 2\hat{\ell},$$

where $\hat{\ell}$ is the maximum of the (penalized) log-likelihood function.

Class Test

- Time: March 21, 12:30–1:20 pm
- Location: TBD
- Contents to be covered: Chapters 1–3
- Format: closed-book
- You should bring a calculator.