

Survival Analysis / Modèles de Durée

Chapitre 1 : Survival Data

Gilles de Truchis

Master 2 ESA

Plan du chapitre

- 1 Introduction
- 2 Principles of Survival Analysis
- 3 Parametric estimation principles
- 4 Nonparametric estimation
- 5 Nonparametric comparisons

Plan

- 1 Introduction
- 2 Principles of Survival Analysis
- 3 Parametric estimation principles
- 4 Nonparametric estimation
- 5 Nonparametric comparisons

Survival Analysis

- Study of survival times of a particular phenomenon...

... and the factor that influence them

- Data with survival outcomes are numerous

- ⇒ Clinical trials

- ⇒ Biomedical studies

- ⇒ Industrial settings (failure of a device)

- ⇒ Labor market

- ⇒ **Credit default**

- Statistical analysis of survival data requires

- ⇒ Estimation of survival distribution

- ⇒ Comparisons of various survival distributions

- ⇒ Elucidations of the factors that influence survival times (regressions)

Survival Data

- The variable of interest has key characteristics
 - ⇒ Non-negative discrete (or continuous) random variable
 - ⇒ Represents the time from a well-defined origin to a well-defined event
 - ⇒ Often subject to censoring : the starting or ending event is not observed
- Example of right censoring
 - Let T^* be a random variable representing the time to failure
 - Let U be a random variable representing the time to censoring event
 - The recorded event will be $T = \min(T^*, U)$ and we can define
$$\delta = I(T^* < U)$$
a censoring indicator taking value 1 or 0
 - ⇒ $\delta = 1$ if T is an observed failure time and $\delta = 0$ if T is a censored time

Note 1 Left censoring are possible albeit less frequent

Note 2 Interval censoring are also possible : the failure time has occurred within an unobserved time interval

Censoring classification

- There are 3 types of censoring times :

Type I Pre-specified censored times

- e.g. In a study with a pre-specified ending time, if an individual has not experienced the event of interest before the end, it is censored at that time

Type II Pre-specified fraction of failure

- e.g. If the study runs until a pre-specified fraction of failure is reached (e.g. 25 %), individuals or objects that have not failed (75%) are censored

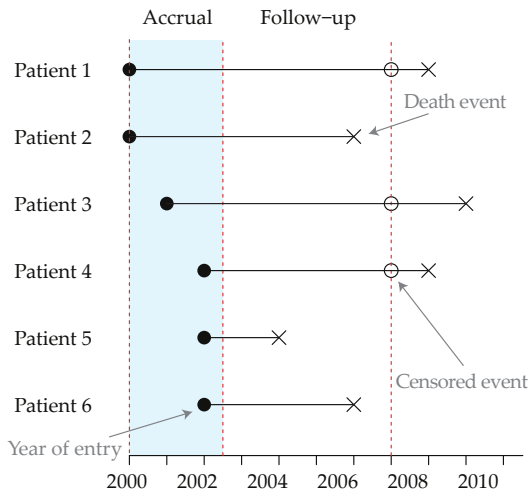
Random Censoring that occurs randomly and independently of the study

- e.g. In a biomedical study, patient dropout that are unrelated to the disease process (e.g. death unrelated to the disease under investigation)

Note The random nature of this type of censoring is crucial to avoid bias

Type I censored data

- In biomedical studies, administrative censoring is of type I
- ⇒ It occurs when patients are still alive at the end of the follow-up period



Patient time structure

- Survival database are generally structured as follows

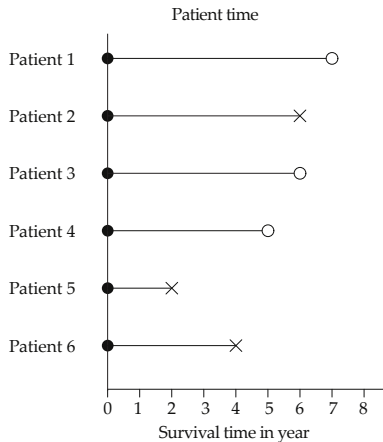
⇒ For each individual, the survival time and δ (“Status”) are reported

Table – Survival data example

Patient	Survtime	Status
1	7	0
2	6	1
3	6	0
4	5	0
5	2	1
6	4	1

Patient time representation

- The patient time graphical representation is as follows



Database example (1)

■ Additional informations can include additional outcomes

- individual characteristics
- competing risks factors

⇒ Below, $\delta \in \{0, 1, 2\}$ where 2 to indicate death from other causes

Table – Survival prospects of prostate cancer patients with high-risk disease

Patient	grade	stage	ageGroup	survTime	status
88	poor	T2	75-79	33	0
89	mode	T2	75-79	6	0
90	mode	T1c	75-79	15	2
91	mode	T2	70-74	6	2
92	mode	T1ab	80+	93	1
93	poor	T2	80+	60	2
94	mode	T2	80+	1	0
95	mode	T1ab	75-79	34	0

Database example (2)

- Comparisons survival data is also of crucial interest

e.g. triple-medication v.s. nicotine patch therapy alone

Note 1 δ is set to 0 for individuals who remained non-smokers for 6 months

Note 2 Below, the variable *ttr* is time until return to smoking

⇒ The objective is to compare the two treatment therapies by identifying the factors related to this outcome

Table – Comparison of medical therapies to aid smokers to quit

	ttr	relapse	grp	age	gender	morphotype	employment
1	182	0	patchOnly	36	Male	white	ft
2	14	1	patchOnly	41	Male	white	other
3	5	1	combination	25	Female	white	other
4	16	1	combination	54	Male	white	ft
5	0	1	combination	45	Male	white	other
6	182	0	combination	43	Male	hispanic	ft

Plan

- 1 Introduction
- 2 Principles of Survival Analysis
- 3 Parametric estimation principles
- 4 Nonparametric estimation
- 5 Nonparametric comparisons

Hazard and Survival Functions

- Survival Analysis relies on the survival distribution that is specified by
 - either the Survival Function (SF)
 - or the Hazard Function (HF)

- The SF is defined as the probability of surviving up to a point t

$$S(t) = \mathbb{P}(T > t), \quad 0 < t < \infty$$

$\Rightarrow S(t)$ is right continuous, equals 1 at time 0 and decreases over time

Note In some cases, $S(t)$ can also remain constant and never reach 0

- The HF is defined as the instantaneous failure rate

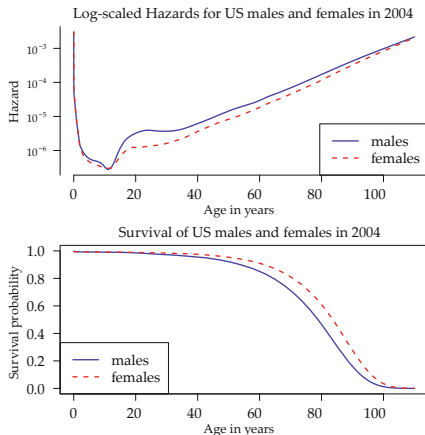
$$h(t) = \lim_{\Delta \rightarrow 0} \frac{\mathbb{P}(t < T < t + \Delta | T > t)}{\Delta}$$

$\Rightarrow h(t)$ is the probability of failing in the next interval of time Δ , given that the subject has survived up to time t , divided by that interval

Hazard and Survival Functions representation

Data The daily hazard rates of men and women by age from 1940 to 2004

- The initial days and weeks of life are particularly dangerous
- The hazard increases during the teen years, then levels off
- It starts a steady increase in midlife



Other representations of the Survival Distribution

- The complement of the SF is just the so-called CDF

$$F(t) = \mathbb{P}(T \leq t), \quad 0 < t < \infty$$

⇒ known as cumulative risk function in the survival analysis

- The PDF is also an obvious alternative representation

$$f(t) = -\frac{d}{dt}S(t) = \frac{d}{dt}F(t)$$

⇒ it is the rate of change of $F(t)$ or minus the rate of change of $S(t)$

- $f(t)$ is also related to $h(t)$ by

$$h(t) = \frac{f(t)}{S(t)}$$

⇒ the hazard at time t is the probability that an event occurs in the neighborhood of t divided by the probability that the subject is alive at t

The Survival Function as function of the Hazard Function

- The area under the HF up to time t is the cumulative HF

$$H(t) = \int_0^t h(u) du$$

- Then, one can define the survival function in terms of the CHF

$$S(t) = \exp \left(- \int_0^t h(u) du \right) = \exp(-HF)$$

Mean and Median Survival time

- The expected value of the survival time is simply

$$\mathbb{E}(T) = \int_0^{\infty} t f(t) dt = \mu$$

- An alternative equivalent measurement is

$$\mu = \int_0^{\infty} S(t) dt$$

Note 1 it is defined ($\mu < \infty$) only if $S(\infty) = 0$: all subjects eventually fail

\Rightarrow this might not be the case if, e.g., the survival outcome is time to cancer recurrence and a fraction c of subjects are completely cured

- The Median survival time is the time τ such that $S(\tau) = 1/2$

Note 2 If $S(t)$ is a step function, it is not continuous at $1/2$ and the Median is the smallest t such that $S(t) \leq 1/2$

Note 3 If $S(t)$ never drop below $c = 1/2$ during the observation period, the Median is undefined

Introduction to parametric Survival Distributions

- In view of modeling the survival process, we need to specify a distribution
- The simplest survival distribution is the exponential one

$$f(t) = \lambda e^{-\lambda t},$$

- The definitions of S13 allows to compute the SF

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

and alternative representations of S15 give

$$h(t) = \lambda$$

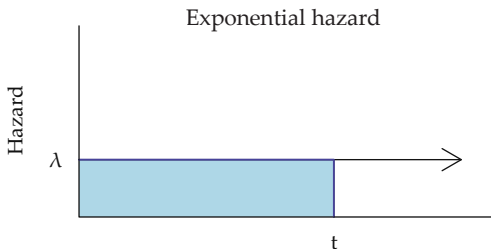
⇒ This SD has constant hazard function $h(t) = \lambda$

The Exponential Survival Distribution

- The cumulative hazard function is hence

$$H(t) = \int_0^t h(u)du = \int_0^t \lambda du = \lambda t$$

and is represented by the shaded area below



- The mean survival time is simply

$$\mathbb{E}(T) = \int_0^\infty S(t)dt = \int_0^\infty e^{-\lambda t} dt = 1/\lambda$$

and the median survival time is obtained for $e^{-\lambda\tau} = 0.5$, i.e. $\tau = \log(2)/\lambda$

The Weibull Survival Distribution

- The constant hazard is a strong assumption in many practical cases
- ⇒ a first generalization is obtained by considering

$$h(t) = \alpha \lambda^\alpha t^{\alpha-1}$$

the hazard function derived from the Weibull distribution

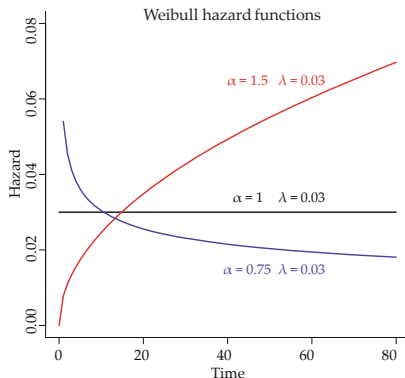
Note For $\alpha = 1$ it comes down to the exponential distribution

- From $h(t)$ one can easily derive $H(t) = (\lambda t)^\alpha$ and hence

$$S(t) = e^{-(\lambda t)^\alpha}$$

The Weibull Hazard Function

- For several parameter choices the behavior of $h(t)$ is represented below



- The mean survival time formula is not obvious

$$\mathbb{E}(T) = \int_0^{\infty} S(t)dt = \frac{\Gamma(1 + 1/\alpha)}{\lambda}$$

and the median survival time is given by $\tau = \log(2)^{1/\alpha} / \lambda$

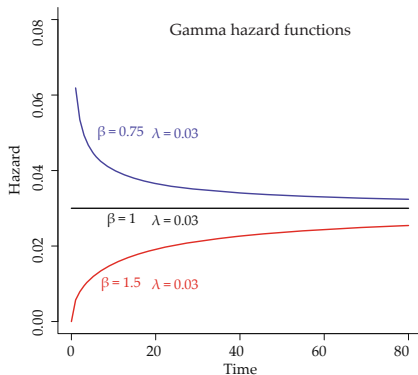
Note The Gamma function generalizes the factorial function to real numbers

The Gamma Hazard Function

- Another choice for survival modeling is the Gamma distribution

$$f(t) = \frac{\lambda^\beta t^{\beta-1} \exp(-\lambda t)}{\Gamma(\beta)}$$

which comes down to the exponential one for $\beta = 1$ as $\Gamma(1) = 1$



Note No closed form exist for the HF and SF \Rightarrow numerical computations

Numerical approximation to the Hazard and Survival Functions

- In some cases (see e.g. S14), the distribution is much more complicated
- An alternative way is numerical computation :
 - 1 Take people dead at birth, after 1 day, week, month, year, 2 years, ...
 - 2 Take the data in difference to obtained rectangles
 - 3 Compute the cumulated sum of data in each rectangle to get $\hat{H}(t)$
 - 4 The SF is simply given by $\hat{S}(t) = \exp(-\hat{H}(t))$
- One can use $\hat{S}(t)$ to compute the mean that is

73.80

for the male and

78.90

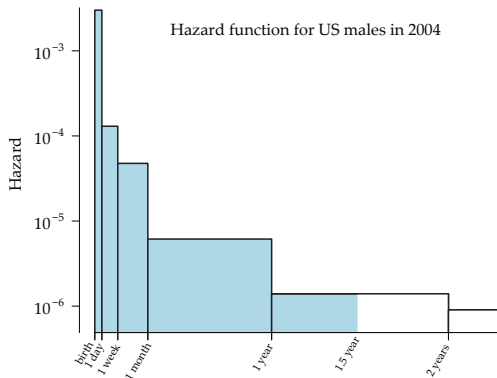
for the women when considering the US lifetime data of S14

Example of CHF approximation

- Step 1 to 3 allow to approximate the integral of $H(t)$

e.g. The male lifetime CHF up to 1.5 years is given by the blue area

⇒ Applying this method beyond 2 years leads to the blue CHF curve in S14



Plan

- 1 Introduction
- 2 Principles of Survival Analysis
- 3 Parametric estimation principles
- 4 Nonparametric estimation
- 5 Nonparametric comparisons

Unknown distribution parameters

- In general, we have poor knowledge upon

$$S(t)$$

the underlying Survival Distribution

- We only have realizations

$$t_1, t_2, \dots, t_n$$

of random variables for which a distributional assumption is done

e.g. Under exponential distribution hypothesis, the parameter

$$\lambda$$

is unobserved and we would like to estimate it

⇒ A natural candidate is the Maximum Likelihood estimator

MLE principle for Survival data

- As in time series analysis, the likelihood function take the general form

$$L(\theta; t_1, t_2, \dots, t_n) = f(t_1, \theta) \cdot f(t_2, \theta) \dots f(t_n, \theta) = \prod_{i=1}^n f(t_i, \theta)$$

with $\theta = \lambda$ in the exponential distribution case

Note However, particular attention has to be paid to censored data

e.g. For right-censored data we use δ and the Survival Function

$$S(t_i, \theta)^{1-\delta_i}$$

to indicate that observation i is known only to exceed t_i as

$$S(t_i, \theta) = \mathbb{P}(T_i > t_i)$$

⇒ The likelihood is hence transformed to

$$L(\theta; t_1, t_2, \dots, t_n) = \prod_{i=1}^n f(t_i, \theta)^{\delta_i} S(t_i, \theta)^{1-\delta_i} = \prod_{i=1}^n h(t_i, \theta)^{\delta_i} S(t_i, \theta)$$

Note For left-censored data we use δ and $1 - S(t_i, \theta) = \mathbb{P}(T_i \leq t_i) = F(t_i, \theta)$

MLE principle for exponential distribution

- In the particular case of the exponential distribution,

$$L(\theta; t_1, t_2, \dots, t_n) = \prod_{i=1}^n \left(\lambda e^{-t_i/\mu} \right)^{\delta_i} \left(e^{-\lambda t_i} \right)^{1-\delta_i} = \lambda^d e^{-\lambda V}$$

where $d = \delta_1 + \dots + \delta_n$ is the total number of failure and

$$V = t_1 + \dots + t_n$$

is the total amount of time of patients

- The MLE is given by the value of λ that maximizes $L(\lambda; t_1, t_2, \dots, t_n)$
- As log-transformation simplifies the likelihood function we prefer

$$\ell(\lambda) = \log L(\theta; t_1, t_2, \dots, t_n) = d \log \lambda - \lambda V$$

- Under regularity conditions, the MLE is asymptotically Gaussian

Solution of exponential-based MLE

- The first derivative (score function) give

$$\ell'(\lambda) = \frac{d}{\lambda} - V$$

and hence the maximum likelihood estimate is $\hat{\lambda} = d/V$

- The second derivative (Hessian function) is

$$\ell''(\lambda) = -\frac{d}{\lambda^2} = -I(\lambda)$$

where $I(\lambda) > 0$ is the Fisher information

- As $\ell''(\lambda) < 0$ the solution is a maximum and inversing $I(\lambda)$ we obtain

$$\mathbb{V}(\hat{\lambda}) = \sigma_{\lambda}^2 \approx I^{-1}(\lambda) = \lambda^2/d$$

- In practice we will use

$$\hat{\sigma}_{\lambda}^2 \approx I^{-1}(\lambda) = \hat{\lambda}^2/d = d/V^2$$

Note For most of distributions, no explicit solutions exist \Rightarrow numerical resolution

Exercise

- Consider the data of Table 1
- Plot the log-likelihood and compute the MLE of λ and $\mathbb{V}(\hat{\lambda})$

Exercise

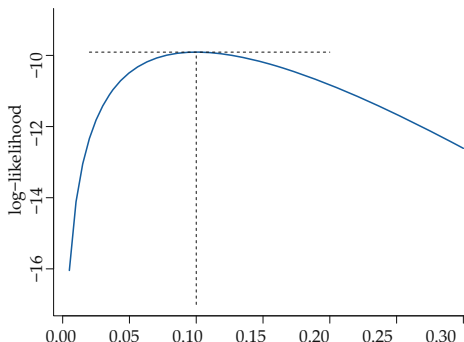
- Consider the data of Table 1
- Plot the log-likelihood and compute the MLE of λ and $\mathbb{V}(\hat{\lambda})$
- Simple observation of the data gives $d = 3$ and

$$V = 7 + 6 + 6 + 5 + 2 + 4 = 30$$

⇒ The log-likelihood function is

$$\ell(\lambda) = 3 \log \lambda - 30\lambda$$

and hence we obtain $\hat{\lambda} = 3/30 = 0.1$ with $\hat{\sigma}_{\lambda}^2 \approx 3/(30^2) = 0.0033$



Plan

- 1 Introduction
- 2 Principles of Survival Analysis
- 3 Parametric estimation principles
- 4 Nonparametric estimation
- 5 Nonparametric comparisons

The Kaplan-Meier estimator (KPE)

- In practice, the distribution/survival/hazard function is hard to choose
- ⇒ The parametric approach is likely to be misspecified
- Nonparametric estimation procedures offer more flexibility
- ⇒ The most widely used of these procedures is the Kaplan-Meier estimator

$$\hat{S}(t) = \prod_{t_i \leq t} (1 - \hat{q}_i) = \prod_{t_i \leq t} \left(1 - \frac{d_i}{n_i}\right)$$

where d_i is the number of failure at time t_i and n_i the number of individuals at risk at that time

- ⇒ $\hat{S}(t)$ is the product over failure times of the conditional probabilities of surviving to the next failure time

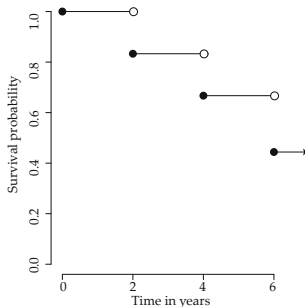
Application of the KPE

- By using the data of the Table 1, one can easily obtain

Table – Kaplan-Meier estimator

t_i	n_i	d_i	q_i	$1 - q_i$	\hat{S}_i
2	6	1	0.167	0.833	0.833
4	5	1	0.200	0.800	0.667
6	3	1	0.333	0.667	0.444

- One can use \hat{S}_i to reconstruct graphically the Survival Function



Interpretation of \hat{S}_t

- \hat{S}_t is a non-increasing right continuous step function
 - t_i is the failure time
 - n_i is the number of individuals at risk at time t_i
 - d_i is the number of individuals who fail at time t_i
 - $q_i = d_i/n_i$ is the failure probability
 - $1 - q_i$ is the conditional survival probability
 - S_i is the Survival Function at time t_i

- The right-continuity is illustrated by open and closed circles

e.g. $S(4) = 0.667$ while $S(3.99) = 0.833$

Note The median is obtained for

$$t_i = \hat{\tau} = 6,$$

that is the smallest time such that $S(t) \leq 1/2$ ($\hat{S}(\tau) = 0.444$)

KPE and inference

- The variance of the KPE can be approximated by

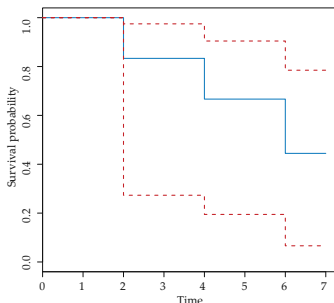
$$\mathbb{V}(\hat{S}_t) \approx \hat{S}_t^2 \sum_{t_i \leq t} \frac{d_i}{n_i(n_i - d_i)}$$

- Unfortunately, CIs derived from $\mathbb{V}(\hat{S}_t)$ may extend above 1 or below 0

Note Remind that $S(t) \in [0, 1]$

⇒ One often overcome this issue by using a log-log transformation of $\hat{S}(t)$

$$\mathbb{V}(\log(-\log \hat{S}_t)) \approx \frac{1}{(\log \hat{S}_t)^2} \sum_{t_i \leq t} \frac{d_i}{n_i(n_i - d_i)}$$



Nelson-Altschuler estimate of the SF

- An alternative estimator is the one of Nelson-Altschuler based on $H(t)$

$$\hat{S}_t = e^{-\hat{H}(t)}, \quad \hat{H}(t) = \sum_{t_i \leq t} \frac{d_i}{n_i}$$

Table – Nelson-Altschuler estimator

t_i	n_i	d_i	q_i	\hat{H}_i	\hat{S}_i
2	6	1	0.167	0.167	0.846
4	5	1	0.200	0.367	0.693
6	3	1	0.333	0.700	0.497

- Confidence intervals can be obtained in a similar way to KPE

Median and inference

- As stated previously, the median is

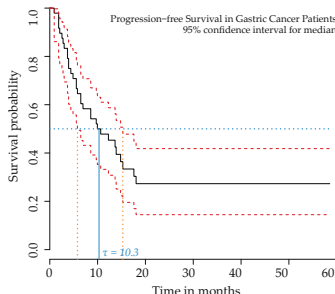
$$\hat{\tau} = \inf\{t : \hat{S}(t) \leq 1/2\}$$

- For a risk level α confidence intervals are given by

$$-z_{\alpha/2} \leq \frac{g(\hat{S}(t)) - g(1/2)}{\mathbb{V}(L(\hat{S}(t)))^{1/2}} \leq z_{\alpha/2}$$

with $g(x) = \log(-\log(x))$ and $z_{\alpha/2}$ a Standard Normal quantile

e.g. Consider the data of Table 2 and the KPE : $\hat{\tau} = 10.3$



Kernel smoothing and Hazard Function estimation

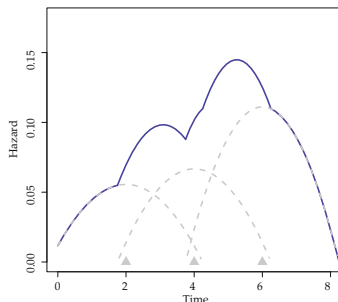
- The Nelson-Altschuler estimate of $h(t)$ can be rough and quite instable
- A kernel function can be used to smooth $\hat{h}(t)$

$$\hat{h}(t) = \frac{1}{b} \sum_{i=1}^D \mathcal{K}\left(\frac{t - t_i}{b}\right) \frac{d_i}{n_i}$$

where $t_1 < \dots < t_D$ are ordered failure times and b a tuning parameter

Note Many kernel function exist but the Epanechnikov kernel is very common

$$\mathcal{K}(x) = 3/4(1 - x^2), \quad -1 \leq x \leq 1$$



Corrected Kernel smoothing and Hazard Function estimation

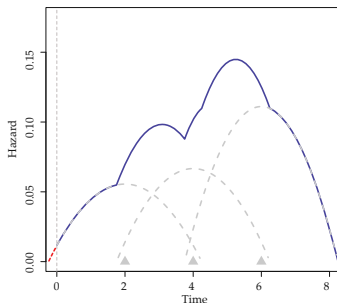
- Without corrections $\mathcal{K}(x)$ is likely to be $\neq 0$ at time $t < 0$

\Rightarrow The first kernel below is centered at $t = 2$ and $b = 2.5$ meaning that

$$t - b = -0.5 \quad t + b = 4.5$$

and hence, the actual area under the first kernel is too small

\Rightarrow The modified Epanechnikov kernel is recommended



- Another approach consists in setting a time-varying b :

\Rightarrow wider $b(t)$ is used than for time regions densely populated with events

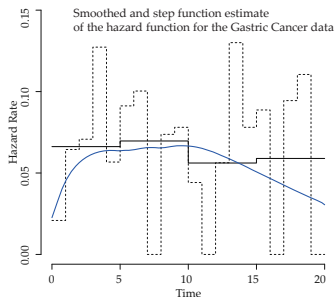
Example Kernel smoothing and Hazard Function estimation

- Consider again the data of Table 2

⇒ Choose the modified Epanechnikov kernel with $b = 20$

Note Selection of b can be critical :

- if b is too small, the estimate may gyrate widely
- if b is too wide, the hazard function may be too smooth to observe real variations in the hazard function

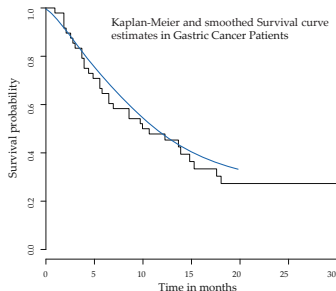


Example Kernel smoothing and Survival Function estimation

- One can use \hat{h} to obtain a smooth estimate of $S(t)$

$$\hat{S}(t) = \exp \left(- \int_{u=0}^t \hat{h}(u) du \right)$$

- In practice the integral is approximated by the rectangles method



Plan

- 1 Introduction
- 2 Principles of Survival Analysis
- 3 Parametric estimation principles
- 4 Nonparametric estimation
- 5 Nonparametric comparisons

Comparing Two Groups of Survival Times

- Comparison of distributional features is of crucial interest

e.g. In medical trials you need to compare treatment and control groups

$$H_0 : S_1(t) = S_0(t)$$

- Let $S_1(t)$ be the SF of the treatment group

⇒ Two alternative hypotheses can be specified (one-sided or two-sided)

$$H_1 : S_1(t) > S_0(t) \text{ or } H_1 : S_1(t) \neq S_0(t)$$

⇒ Unfortunately, Survival data imply several serious issues

- Survival distributions can be similar for some t and differ for others
- Survival distributions can cross

Lehman alternatives

- One solution is to consider Lehman-type alternatives defined as

$$H_1 : S_1(t) = (S_0(t))^\psi$$

where $\psi \neq 1$ unless under

$$H_0 : S_1(t) = (S_0(t))^1$$

⇒ The one-sided alternative is now

$$H_1 : \psi < 1$$

and imposes that $S_1(t)$ is uniformly higher than $S_0(t)$

- These hypotheses can be formulated in terms of proportional hazards

$$h_1(t) = \psi h_0(t)$$

The 2-by-2 Table representation

- In the spirit of the rank tests à la Mann-Whitney H_0 can be tested against Lehman alternatives

Note Complications arise from the presence of censoring

⇒ To solve this issue consider a two-by-two table representation of the data

Table – 2-by-2 Table representation

	Control	Treatment	Sums
Failure	d_{0i}	d_{1i}	d_i
Non-failure	$n_{0i} - d_{0i}$	$n_{1i} - d_{1i}$	$n_i - d_i$
At risk	n_{0i}	n_{1i}	n_i

- Numbers at risk for the control and treatment groups are n_{0i} and n_{1i}
- Numbers of failure for the control and treatment groups are d_{0i} and d_{1i}
- This representation is adopted for any distinct ordered failure time t_i

Hypergeometric distribution

- If one holds d_i , n_{0i} and n_{1i} fixed (and hence n_i too) we can derive

$$\mathbb{P}(d_{0i}|n_{0i}, n_{1i}, d_i) = \binom{n_{0i}}{d_{0i}} \binom{n_{1i}}{d_{1i}} \binom{n_i}{d_i}^{-1}$$

the hypergeometric distribution of d_{0i} where

$$\binom{n_i}{d_i} = \frac{n_i!}{d_i!(n_i - d_i)!}$$

represents the number of combinations of n items taken d at time t_i

- The 2 first moments of that distribution are

$$\mathbb{E}(d_{0i}) = \frac{n_{0i}d_i}{n_i} = \mu_{0i}$$

and

$$\mathbb{V}(d_{0i}) = \frac{n_{0i}n_{1i}d_i(n_i - d_i)}{n_i^2(n_i - 1)} = \sigma_{0i}^2$$

The log-rank test statistics

- Based on the 2-by-2 representation and $\mathbb{E}(d_{0i})$ one can define

$$U_0 = \sum_{i=1} (d_{0i} - \mathbb{E}(d_{0i}))$$

a simple linear test statistic and its variance

$$\mathbb{V}(U_0) = \sum_{i=1} \mathbb{V}(d_{0i})$$

- One can show that $U_0/\sqrt{\mathbb{V}(U_0)} \sim \mathcal{N}(0, 1)$ or equivalently

$$\frac{U_0^2}{\mathbb{V}(U_0)} \sim \chi_1^2$$

- This test statistic is known as the log-rank test of group comparison

Note 1 This test is also known as the Mantel-Haenzel test

Note 2 A comparison of k groups is possible and modify the distribution to

$$\chi_{k-1}^2$$

but is slightly different from the stratified tests discussed in S55

Exercise : application of the log-rank test

- Consider the following survival data
- C and T stand for Control and Treatment groups respectively

Table – Survival data

Patient	Survtime	Censor	Group
1	6	1	C
2	7	0	C
3	10	1	T
4	15	1	C
5	19	0	T
6	25	1	T

- When required, construct the 2-by-2 tables
- Compute the log-rank test and interpret the result

Exercise : computation

- Failures appear at $t = 6, 10, 15, 25$ and result in four 2-by-2 tables

Table – 2-by-2 tables for $t = 6, 10, 15, 25$

	t = 6			t = 10			t = 15			t = 25		
	C	T	Σ	C	T	Σ	C	T	Σ	C	T	Σ
Failure	1	0	1	0	1	1	1	0	1	0	1	1
Non-failure	2	3	5	1	2	3	0	2	2	0	0	0
At risk	3	3	6	1	3	4	1	2	3	0	1	1

Table – Intermediate calculus to compute the log-rank test statistic

t_i	n_i	d_i	n_{0i}	d_{0i}	n_{1i}	d_{1i}	μ_{0i}	σ_{0i}^2
6	6	1	3	1	3	0	0.500	0.2500
10	4	1	1	0	3	1	0.250	0.1875
15	3	1	1	1	2	0	0.333	0.2222
25	1	1	0	0	1	1	0.000	0.0000
Σ				2		2	1.083	0.6597

Exercise : interpretation

- From Tables in previous slide we easily obtain

$$U_0 = \sum_i d_{0i} - \sum_i \mu_{0i} = O_0 - E_0 = 2 - 1.083 = 0.917$$

$$\text{and } \mathbb{V}(U_0) = \sum_i \sigma_{0i}^2 = V_0 = 0.6597$$

⇒ The log-rank test statistic is

$$\frac{U_0^2}{\mathbb{V}(U_0)} \approx 1.26$$

which we compare to a χ_1^2 distribution

⇒ The corresponding p -value is

$$p = 0.259$$

meaning that we cannot reject H_0 and hence the group difference is not statistically significant

Note When applying the test to d_{1i} , the result is identical as it also sums to 2

The generalized log-rank test statistics

- An important generalization of the log-rank test is

$$U_0(w) = \sum_{i=1} w_i (d_{0i} - \mathbb{E}(d_{0i}))$$

with the corresponding variance $\mathbb{V}(U_0) = \sum_{i=1} w_i^2 \mathbb{V}(d_{0i})$

- This leads to the so called Fleming-Harrington $G(\rho)$ test

$$G(\rho) = \frac{U_0(w)^2}{\mathbb{V}(U_0(w))}$$

- The most common way of setting weights is à la Gehan-Wilcoxon

$$w_i = \mathcal{F}(\hat{S}(t_i))^\rho, \quad \mathcal{F}(\cdot) \text{ being a certain function}$$

Note 1 When $\rho = 1$ we get the Prentice modification : places higher weight on earlier survival times

Note 2 When $w_i = \sqrt{n_i}$ we get the Tarone-Ware modification : intermediate weight compared to $\rho = 0$ and $\rho > 0$

Note 3 When $w_i = \hat{S}(t_i)^p (1 - \hat{S}(t_i))^q$ we get the Harrington-Fleming(p, q) test : more flexible

Example : Prentice modification of Gehan-Wilcoxon test

- Let consider pancreatic cancer data from a clinical trial (41 patients)
- We are interested in the progression-free survival (PFS)

⇒ the time from assignment in the trial to disease progression or death

Table – Locally Advanced Pancreatic Cancer or Metastatic Pancreatic Cancer

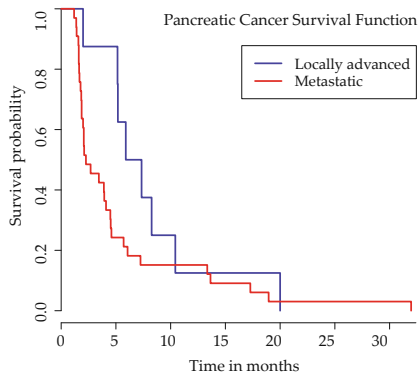
	stage	onstudy	progression	death
1	MPC	16/12/2005	02/02/2006	19/10/2006
2	MPC	06/01/2006	26/02/2006	19/04/2006
3	LAPC	03/02/2006	02/08/2006	19/01/2006
4	MPC	30/03/2006	“NA”	11/05/2006
5	LAPC	27/04/2006	11/03/2007	29/05/2007
6	MPC	07/05/2006	25/06/2006	11/10/2006
⋮	⋮	⋮	⋮	⋮

Note 1 “NA” means that the patient died with no recorded progression and the PFS is time to death

Note 2 For all other patients, the PFS is time to the date of progression

Example : Prentice modification of Gehan-Wilcoxon test

- The graphical analysis of SF reveals :
 - the LAPC group shows an early survival advantage over the MPC
 - but the survival curves converge after about 10 months



Example : Prentice modification of Gehan-Wilcoxon test

- When computing the Gehan-Wilcoxon test for

$$\rho = 0$$

i.e. the log-rank test and

$$\rho = 1$$

i.e. the Prentice modification, we obtain

Table – Fleming-Harrington $G(\rho)$ for $\rho = 0$ and $\rho = 1$, with $k = \{0, 1\}$

$\rho = 0$	N	O_k	E_k	$(O_k - E_k)^2/V_k$
LAPC	8	8	1.49	2.25
MPC	33	33	0.64	2.25
We cannot reject H_0 (no difference) as				$p\text{-value} = 0.134$
$\rho = 1$	N	O_k	E_k	$(O_k - E_k)^2/V_k$
LAPC	8	2.34	2.13	4.71
MPC	33	18.76	0.82	4.71
We reject H_0 as				$p\text{-value} = 0.0299$

- The two tests produce conflicting results as they are optimized for different alternatives

\Rightarrow For $\rho = 1$, the test places higher weight on earlier survival times

Stratified tests

- To compare two groups while adjusting for another covariate, one can
 - 1 include the other covariate as regression terms for the hazard function (see next Chapter)
 - 2 construct a stratified log-rank test if the covariate we are adjusting for is categorical

⇒ denote h_{0j} the population hazard of level $j = 1, 2, \dots, G$, with G small

- For the G categories of the covariate we can test

$$H_0 : h_{0j}(t) = h_{1j}(t), \quad j = 1, 2, \dots, G$$

- Accordingly, the stratified version of the log-rank test statistic is

$$X^2 = \frac{\left(\sum_{g=1}^G U_{0g} \right)^2}{\sum_{g=1}^G V_{0g}} \sim \chi_1^2$$

Example 1 of stratified test

- Consider the dataset of Table 3 (time to return smoking)
- We first compare the 2 treatment groups by means of the log-rank test

$\rho = 0$	N	O_k	E_k	$(O_k - E_k)^2/V_k$
Combination	61	37	49.9	8.03
Patch only	64	52	39.1	8.03
We reject H_0 (no difference) as				$p\text{-value} = 0.00461$

- If now we are interested by the influence of the age we may define

$$g = 1 : 21 - 49 \parallel g = 2 : 50 \text{ or more}$$

a categorical variable that divides the subjects in 2 groups

- The resulting stratified log-rank test is close to the unadjusted test

\Rightarrow the stratification based on the age seems unnecessary

$\rho = 0$	N	O_k	E_k	$(O_k - E_k)^2/V_k$
Combination	61	37	49.1	7.03
Patch only	64	52	39.9	7.03
We reject H_0 (no difference) as				$p\text{-value} = 0.008$

Example 2 of stratified test

- Consider simulated data representing an artificial clinical trial
- This trial compares a standard therapy (control) and an experimental one (treatment)
- The survival times are simulated as exponentially distributed and produces no censoring
- A confounding genotype factor is also simulated with only 2 levels

$g = 1$: wild type genotype || $g = 2$: mutant genotype

with $g = 2$ leading to poorer prognosis as the hazard rate is

$$\lambda = 0.03 \text{ per day}$$

for a mutant patient in the control group whilst the effect of treatment leads to

$$\lambda = 0.0165$$

- For wild type patients $\lambda = 0.006$ whilst the effect of treatment leads to

$$\lambda = 0.0033$$

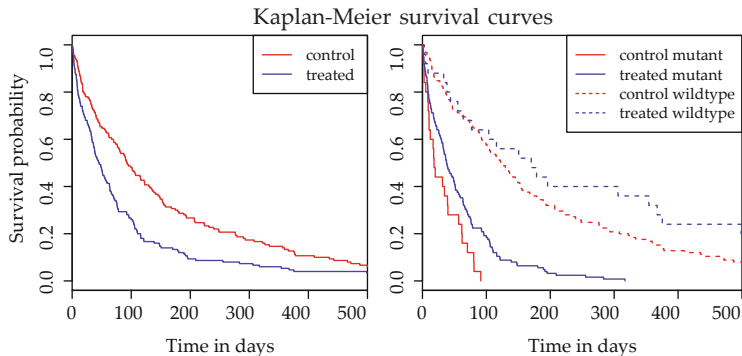
Example 2 of stratified test

- The Kaplan-Meier survival curves are computed both naively and accounting for the gene confounder

Note 1 The naive estimate concludes against the experimental therapy

Note 2 When accounting for the gene confounder the results are at the opposite

⇒ within each genotype, the treatment is actually superior to the control



Example 2 of stratified test

- The stratified log-rank test is now used to confirm the graphical analysis

Unadjusted	N	O_k	E_k	$(O_k - E_k)^2/V_k$
Control	150	150	183	15.9
Treatment	150	150	117	15.9
We reject H_0 (no difference) as				$p\text{-value} = 0.00006$

Note 1 The unadjusted test shows that the treatment reduces survival

Stratified	N	O_k	E_k	$(O_k - E_k)^2/V_k$
Control	150	150	133	7.57
Treatment	150	150	167	7.57
We reject H_0 (no difference) as				$p\text{-value} = 0.00595$

Note 2 The stratified test confirms that the treatment improves survival compared to the control

Note 3 Patients carrying the wild type form of the gene have better survival than do patients carrying the mutation

Note 4 There are more mutation-carrying patients in the treatment group than in the control group, whereas the reverse is true for wild type patients