2009 Quality & Productivity Research Conference

June 3-5, New York

Intensity Models for Randomly Censored Data

Ivanka Horová, Zdeněk Pospíšil and Jiří Zelinka

Masaryk University

Czech Republic

Gregor Reisch: Margarita Philosophica (1508)

CONTENTS

- Survival function and intensity
- Kernel estimate of intensity
- Deterministic model of intensity
- Application
- References

EXAMPLE

Number of patients – 236

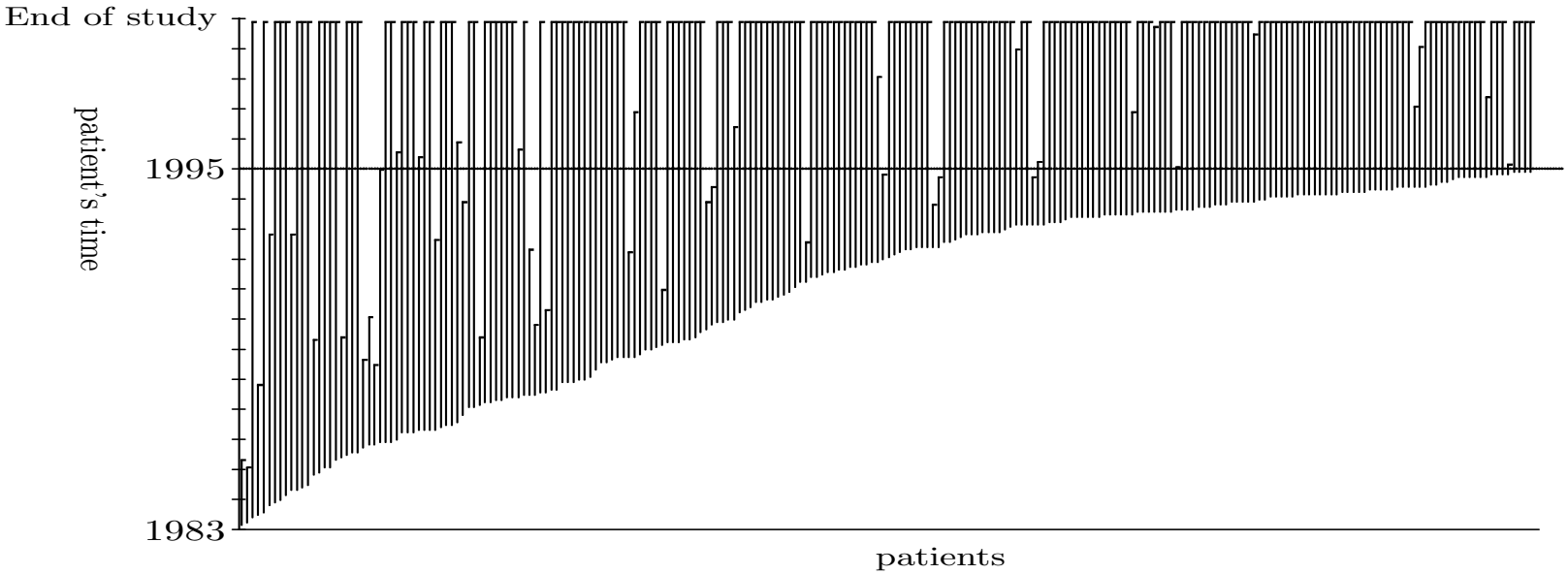
Diagnosis – breast carcinoma of the I and II clinical stage

Treatment – surgical treatment and radiotherapy

The beginning of the study – 1983

The end of the study – 2000, the last patients entered in 1995

Number of death of cancer – 47



SURVIVAL FUNCTION AND INTENSITY

- T – the survival time or lifetime
- F – the cumulative distribution function of T:
  $F(x) = P(T < x)$
- f – the density of F: $F(x) = \int_0^x f(t)\mathrm{dt}$
- $\bar{F}$ – the survival function: $\bar{F}(x) = 1 - F(x)$, i.e. $\bar{F}(x)$ is the probability that an individual survives for a time greater or equal to x
- λ – the intensity (or the hazard function)

$$\lambda(x) = -\frac{d}{dx}\log\bar{F}(x) = \frac{f(x)}{\bar{F}(x)},\ \bar{F}(x) = e^{-\int_0^x \lambda(t)dt},\ \bar{F}(x) > 0$$

i.e. λ is the probability that an individual dies at time x conditional on he or she having survived to the time x.

DIFFERENT POINT OF VIEW

- $N(x)$ – **the size of the cohort at time** x
- $N_0 = N(0)$ – **the initial size of the cohort**
- $\mu(x)$ – **the time dependent death rate or rate of change**
- $N'(x) = -\mu(x)\, N(x),\ N(0) = N_0$**,**

$$N(x) = N_0 \exp\left(-\int_0^x \mu(t)dt\right)$$

- $\bar{F}$ – **the survival function:** $\bar{F}(x) = \frac{N(x)}{N_0} = \exp\left(-\int_0^x \mu(t)dt\right)$
- **the relation between the death rate and intensity:**

$$\lambda(x) = \mu(x) = -\frac{N'(x)}{N(x)}$$

RANDOM CENSORSHIP MODEL

- $T_1, \ldots, T_n$ – **i.i.d. lifetimes**
  F – **the distribution function**
- $C_1, \ldots, C_n$ – **i.i.d. censoring times**
  G – **the distribution function**
- $X_1, \ldots, X_n$ – **i.i.d. observation times:** $X_i = \min\limits_{i=1,\ldots,n} (C_i, T_i)$
  L – **the distribution function:** $\bar{L}(x) = \bar{F}(x)\,\bar{G}(x)$
- **Random censorship model**
  (X_i, δ_i), $i = 1, \ldots, n$, $\delta_i = I_{\{X_i = T_i\}}$ – **the indicator of censoring status**

ESTIMATE OF SURVIVAL FUNCTION

- $\hat{\bar{F}}(x) = \prod_{X_{(j)}<x} \left(\frac{n-j}{n-j+1}\right)^{\delta_{(j)}}$, **(Kaplan-Meier 1958)**
- $X_{(j)}$ **– the** j**th order statistics of** $X_1, \ldots, X_n$
- $\delta_{(j)}$ **– the corresponding indicator of the censoring status.**

MODIFIED EMPIRICAL SURVIVAL FUNCTION OF OBSERVATION TIMES:

- $\bar{L}_n = 1 - L_n$

$$L_n(x) = \frac{1}{n+1} \sum_{i=1}^{n} I_{\{X_i \le x\}}.$$

KERNEL ESTIMATES OF THE INTENSITY

Assumptions and notation

- $\lambda \in C^{k_0}[0,\tau]$, $k_0 \geq 2$, $L(\tau) < 1$
- ν, k **nonnegative integers,** $0 \leq \nu < k \leq k_0$
- $K \in S_{\nu,k}$

  (i) K **– real valued function on** $\mathbb{R}$

  (ii) support$(K) = [-1,1]$

  (iii) $K \in \mathrm{Lip}[-1,1]$

  (iv) $\int_{-1}^{1} x^j K(x)\mathrm{d}x = \begin{cases} 0, & 0 \leq j < k,\ j \neq \nu \\ (-1)^\nu \nu!, & j = \nu \\ \beta_k \neq 0, & j = k. \end{cases}$

- $h = h(n)$ – **a bandwidth or a smoothing parameter**
  $\{h(n)\}$ – **a sequence of non-random positive numbers:**

$$\lim_{n\to\infty} h(n) = 0, \quad \lim_{n\to\infty} nh^{2\nu+1}(n) = \infty$$

- **The kernel estimate of the ν-th derivative of the intensity λ at the point $x \in [0, \tau]$**

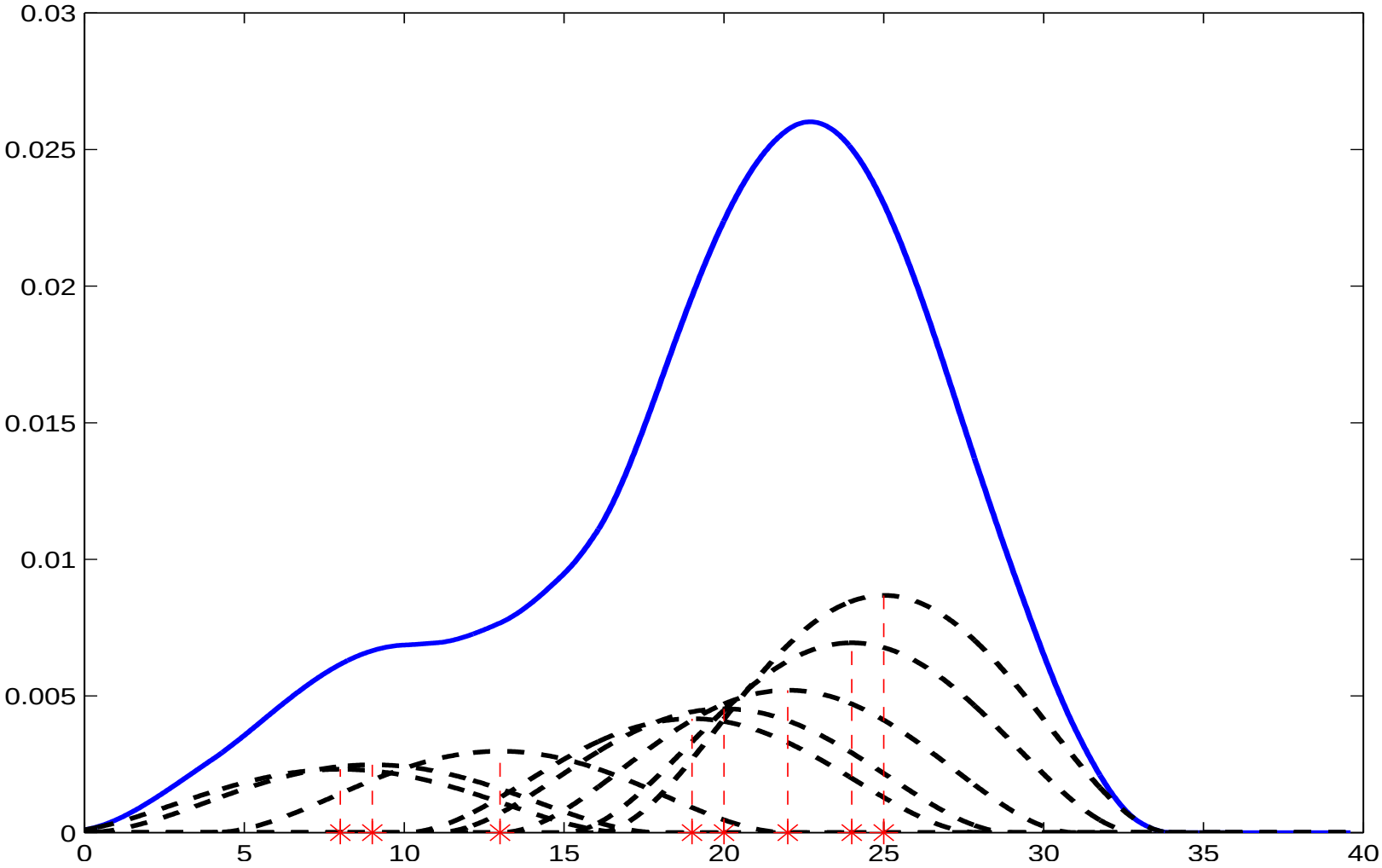
$$\hat{\lambda}^{(\nu)}_{h,K}(x) = \frac{1}{h^{\nu+1}} \sum_{i=1}^{n} K\left(\frac{x - X_{(i)}}{h}\right) \frac{\delta_{(i)}}{n - i + 1}$$

- **The quality of the estimate**
  The mean integrated square error

$$MISE\left(\hat{\lambda}^{(\nu)}_{h,K}\right) = \int_0^{\tau} E\left(\hat{\lambda}^{(\nu)}_{h,K}(x) - \lambda^{(\nu)}(x)\right)^2 \mathrm{d}x,$$

E **denotes the expectation of the random variable**

Construction of kernel estimate of intensity



particular parts of estimate of the hazard function, kernel estimate of the hazard fucntion, $*$ – non-censored observations

- **Notation**

$$V(K) = \int_{-1}^{1} K^2(x)\mathrm{d}x, \quad \beta_k = \int_{-1}^{1} x^k K(x)\mathrm{d}x,$$

$$\Lambda = \int_{0}^{\tau} \frac{\lambda(x)}{\bar{L}(x)}\mathrm{d}x, \quad D_k = \int_{0}^{\tau} \left(\frac{\lambda^{(k)}(x)}{k!}\right)^2 \mathrm{d}x.$$

- $MISE\left(\hat{\lambda}_{h,K}^{(\nu)}\right) = \beta_k^2 D_k\, h^{2(k-\nu)} + \frac{V(K)\Lambda}{nh^{2\nu+1}} + o\left(h^{2(k-\nu)} + \frac{1}{nh^{2\nu+1}}\right)$
- **The leading term**

$$\overline{MISE}\left(\hat{\lambda}_{h,K}^{(\nu)}\right) = \underbrace{\beta_k^2 D_k\, h^{2(k-\nu)}}_{\text{leading bias}^2} + \underbrace{\frac{V(K)\Lambda}{nh^{2\nu+1}}}_{\text{leading variance}}$$

- **The asymptotically optimal bandwidth**

$$h_{opt}^{2k+1} = \frac{V(K)\Lambda(2\nu+1)}{2n(k-\nu)D_k\beta_k^2}$$

- **Relation between $h_{opt,2,k}$ and $h_{opt,0,k}$, k an even integer:**

$$h_{opt,2,k}^{2k+1} = \left(\frac{5k}{k-2}\right)\left(\frac{\gamma_{2,k}}{\gamma_{0,k}}\right)^{2k+1} h_{opt,0,k}^{2k+1}$$

$$\gamma_{0,k}^{2k+1} = \frac{V(K)}{\beta_k^2},\ K \in S_{0,k}, \quad \gamma_{2,k}^{2k+1} = \frac{V(K)}{\beta_k^2},\ K \in S_{2,k}$$

- **The asymptotic $(1-\alpha)$ confidence intervals**

$$\hat{\lambda}_{h,K}(x) \pm \left\{\frac{\hat{\lambda}_{h,K}(x)V(K)}{(1-L_n(x))nh}\right\}^{1/2} \Phi^{-1}(1-\alpha/2)$$

Φ – standard normal cumulative distribution function

CHOICE OF PARAMETERS

- **Choice of the kernel**

(ν, k)	**kernel**	$\gamma_{\nu,k}$
(0,2)	$K(x) = \frac{3}{4}(1 - x^2)$	**1.7188**
(0,4)	$K(x) = \frac{15}{32}(1 - x^2)(3 - 7x^2)$	**2.0165**
(2,4)	$K(x) = \frac{105}{16}(1 - x^2)(5x^2 - 1)$	**1.3925**

- **Choice of the bandwidth**

  Available methods: cross-validation, plug-in, maximal likelihood, iterative

- $\hat{h}_{opt,\nu,k}$ **an estimate of** $h_{opt,\nu,k}$

POINTS OF THE MOST RAPID CHANGE

- θ – the point of the most rapid change, i.e.

$$|\lambda'(\theta)| > |\lambda'(x)|, \quad x \neq \theta, \quad 0 \leq x \leq \tau$$

- $\hat{\theta}$ – the estimate of θ
  – the root of the second derivative

$$\hat{\lambda}^{(2)}_{h,K}(\hat{\theta}) = 0, \quad \hat{\theta} \xrightarrow{p} \theta$$

- The asymptotic $(1-\alpha)$ confidence intervals for $\hat{\theta}$

$$\hat{\theta} \pm \left[\frac{\hat{\lambda}_{h,K}(\hat{\theta}) V(K)}{(1 - L_n(\hat{\theta}))\hat{\lambda}^{(3)}_{h,K}(\hat{\theta}) n \hat{h}^5} \right]^{1/2} \Phi^{-1}(1-\alpha/2), \; K \in S_{2,k}$$

DETERMINISTIC MODEL

- $y = y(x)$ – **time dependent size of the cancer cells population**
- $\lambda = \lambda(x)$ – **proportional to the rate of proliferation for cancer cells**
- $\lambda(x) = \rho y'(x)$, ρ – **the positive rate of proportionality**

$$\left.\begin{aligned} y' &= -\,a\,y\log\frac{b}{y} \\ y(0) &= y_0 \end{aligned}\right\}$$ **the Gompertzian model of cancer cells growth**

y_0 **the initial size of cancer cells population**

b **the maximal possible size of cancer cells population,** $b \gg y_0$

a **the maximal possible rate of increase of the tumor**

- $\lambda(x) = \lambda(x; a, t^*, \lambda^*) = \lambda^* \exp\left(1 - a(x - t^*) - \mathrm{e}^{-a(x-t^*)}\right)$
- $\lambda^* = \rho\frac{ab}{2} = \max \lambda(x)$ **– the maximal hazard or maximal intensity**
- $t^* = \frac{1}{a}\log\left(\log\frac{b}{y_0}\right) = \arg\max \lambda(x)$ **– the time of achieving of the maximal hazard**

COMPOSED INTENSITY FUNCTION

- $N(x) = \sum_{i=1}^{l} \alpha_i N_i(x)$, $N_i(0) = \alpha_i N_0$, N_i **– the subcohort,** $\alpha_i > 0$, $\sum_{i=1}^{l} \alpha_i = 1$
- $N_i(x) = \alpha_i N_0 \exp\left(-\int_0^x \lambda_i(t)dt\right)$, $\lambda_i(t) = \lambda(t; a_i, t_i^*, \lambda_i^*)$, $i = 1, \ldots, l$
- λ_c **– the composed intensity for the complete cohort**

$$\lambda_c(x) = -\frac{N'(x)}{N(x)} = \frac{\sum_{i=1}^{l} \alpha_i \lambda_i(x) e^{-\int_0^x \lambda_i(t)dt}}{\sum_{i=1}^{l} \alpha_i e^{-\int_0^x \lambda_i(t)dt}}$$

- $\lambda_c(x) = \lambda_c(x; \alpha_1, a_1, t_1^*, \lambda_1^*, \alpha_2, a_2, t_2^*, \lambda_2^*, \ldots, \alpha_l, a_l, t_l^*, \lambda_l^*)$

ESTIMATE OF PARAMETERS

- $\hat{\lambda}_j = \hat{\lambda}_{h,K}(x_j)$, $j = 1, \ldots, s$ **– the kernel estimate of the intensity at the given points** $x_j,\ j = 1, \ldots, s$**.**

- **The least square method:** $\alpha_i > 0$, $\sum_{i=1}^{l} \alpha_i = 1$

$$(\hat{\alpha}_1, \ldots, \hat{\alpha}_l, \hat{a}_1, \ldots, \hat{a}_l, \hat{\lambda}_1^*, \ldots, \hat{\lambda}_l^*, \hat{t}_1^*, \ldots, \hat{t}_l^*) =$$

$$= \arg\min \left\{ \sum_{j=1}^{s} \left[\hat{\lambda}_j - \lambda_c(x_j; \alpha_1, a_1, t_1^*, \lambda_1^*, \ldots, \alpha_l, a_l, t_l^*, \lambda_l^*) \right]^2 \right\}$$

$$\hat{\lambda}_c(x) = \lambda_c(x; \hat{\alpha}_1, \hat{a}_1, \hat{t}_1^*, \hat{\lambda}_1^*, \ldots, \hat{\alpha}_l, \hat{a}_l, \hat{t}_l^*, \hat{\lambda}_l^*)$$

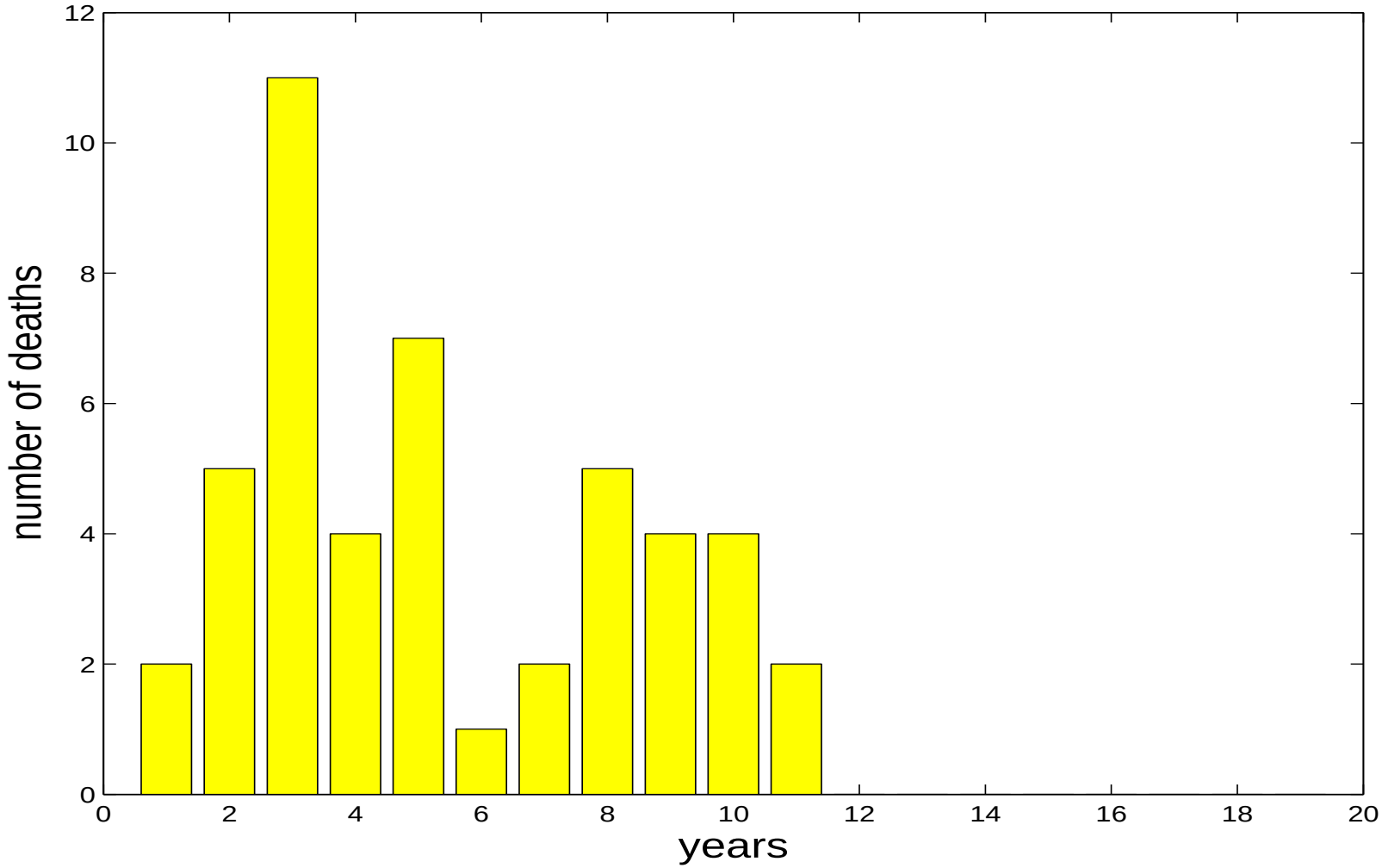
- $\hat{\lambda}_c$ **– the estimate of** λ_c

APPLICATION

- **Breast carcinoma data – BRB (number of patients 236)**
- **Cervical carcinoma data – CCB (number of patients 51)**
- **Acute lymfoblastic leukemia data – ALL (number of patients 38)**

BREAST CARCINOMA

Number of deaths in particular years

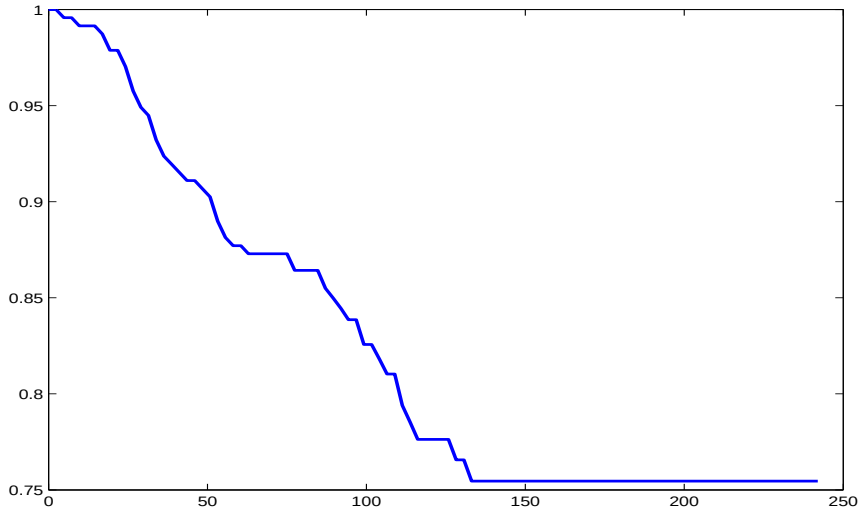


BRB

n	τ	n_d	p_d
236	**220**	**47**	**19,9**

n **number of patients** τ **max. observation time**

n_d **number of deaths** p_d **percentage of death**

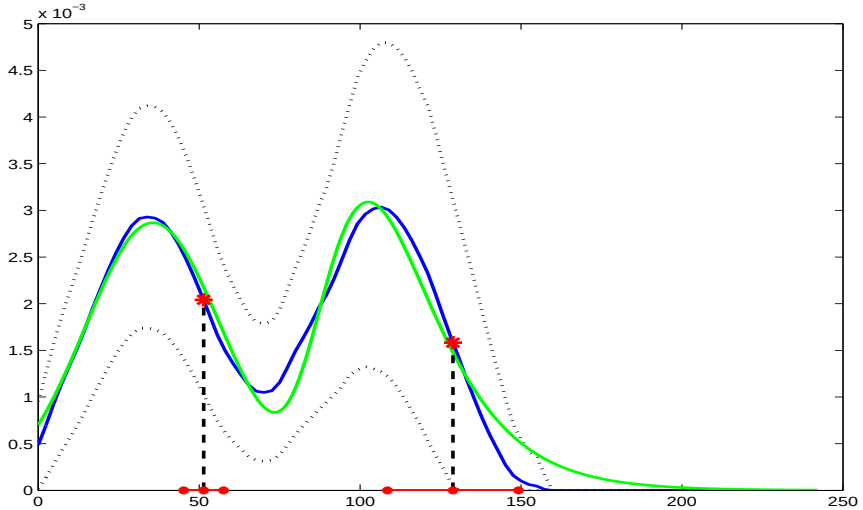


Kaplan-Meier estimate of the survival function

The points of the most rapid change

$$\hat{\theta}_1 = 51.39$$

$$\hat{\theta}_2 = 128.87$$

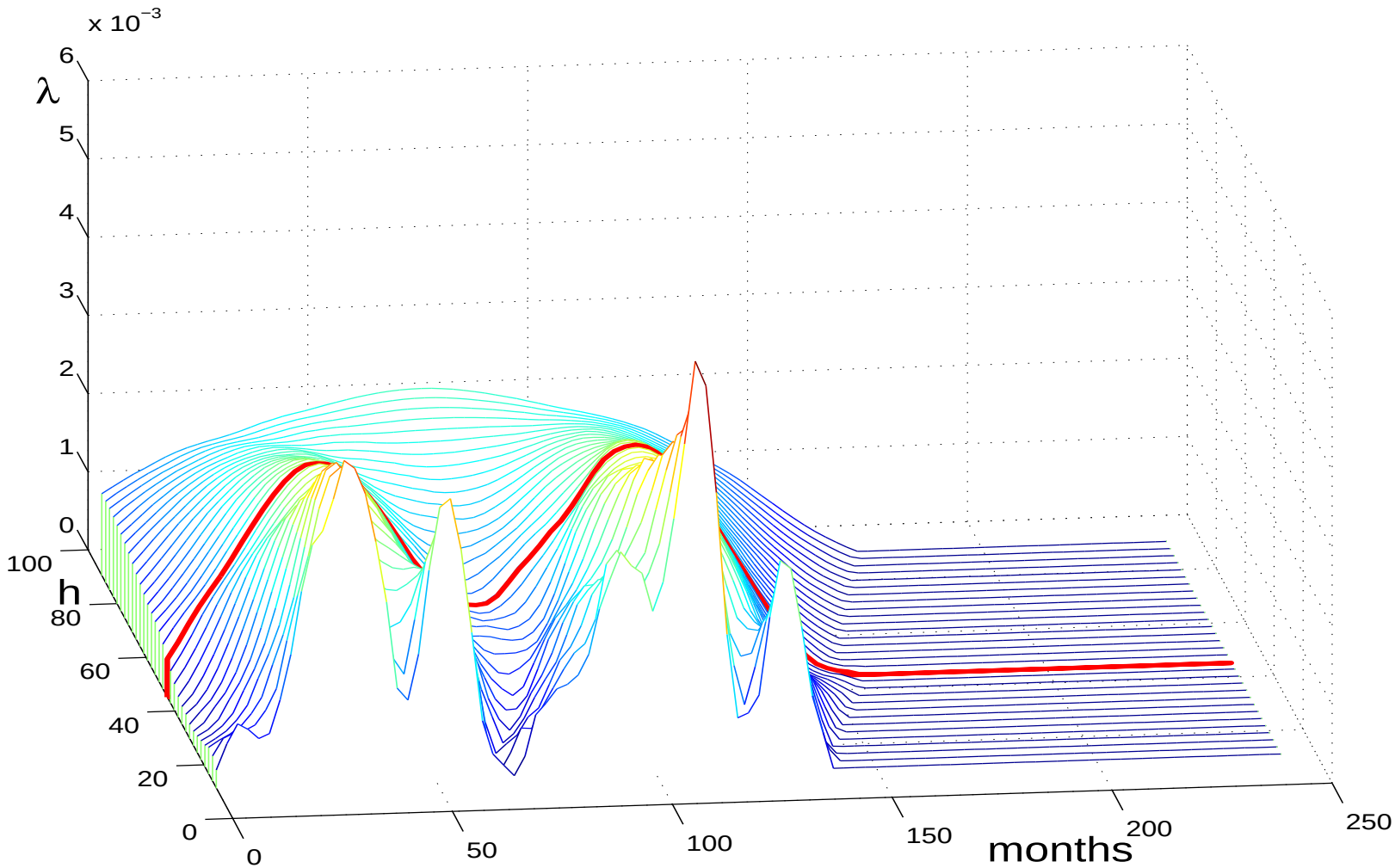


$$\hat{h}_{0,opt} = 45.205$$

kernel estimate, deterministic estimate,

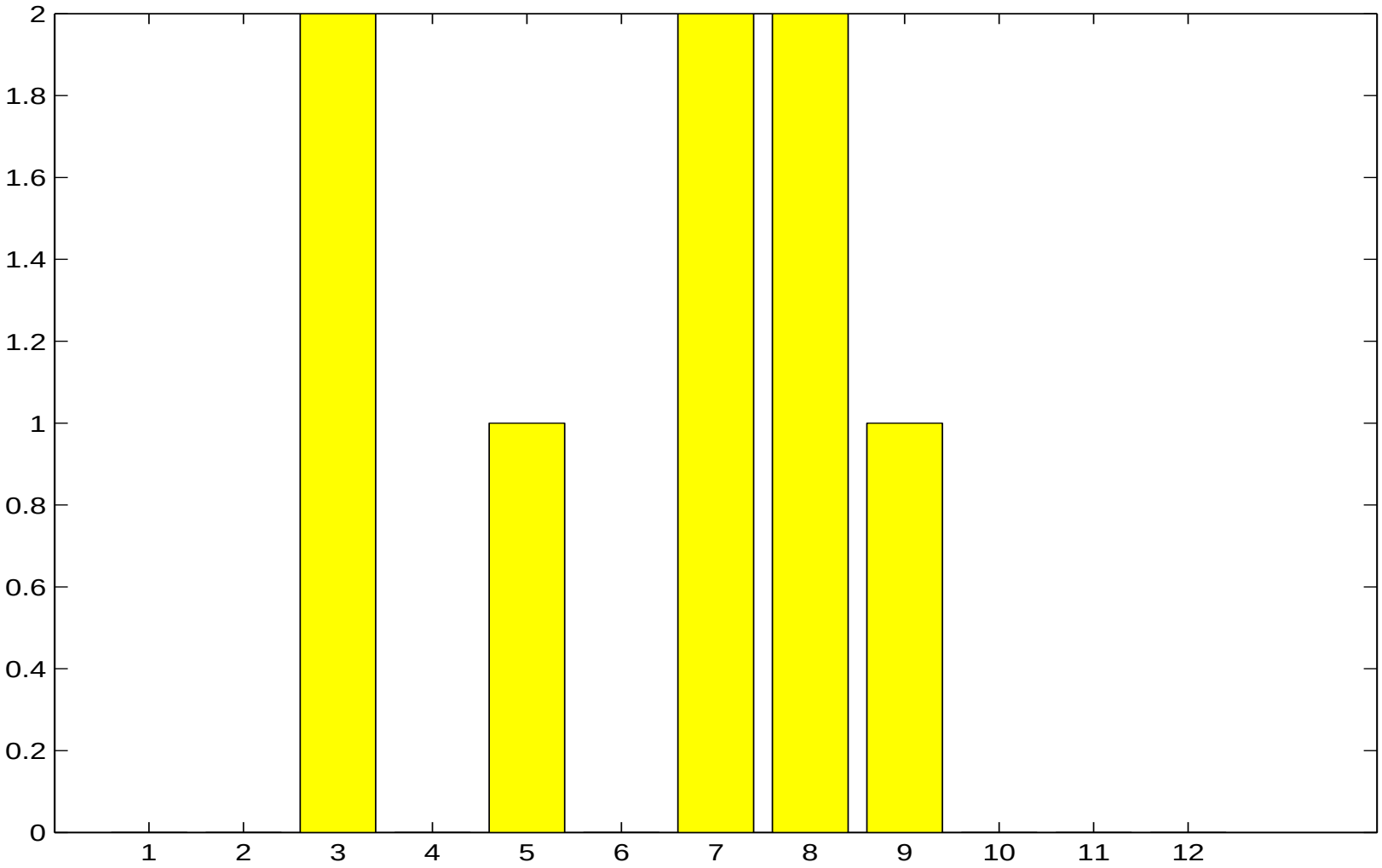
points of the most rapid change

Influence of the bandwidth to the shape of intensity



CERVICAL CARCINOMA

Number of deaths in particular quarters

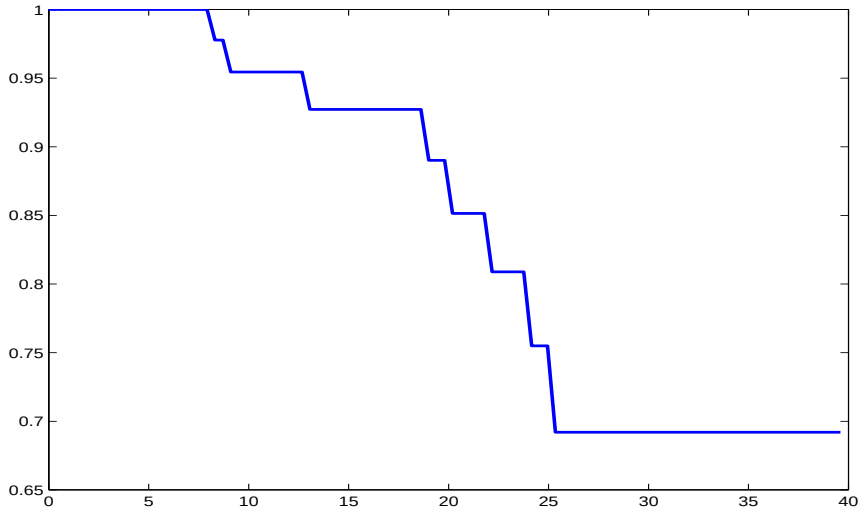


CCB

n	τ	n_d	p_d
51	**36**	**8**	**15,7**

n **number of patients** τ **max. observation time**

n_d **number of deaths** p_d **percentage of death**

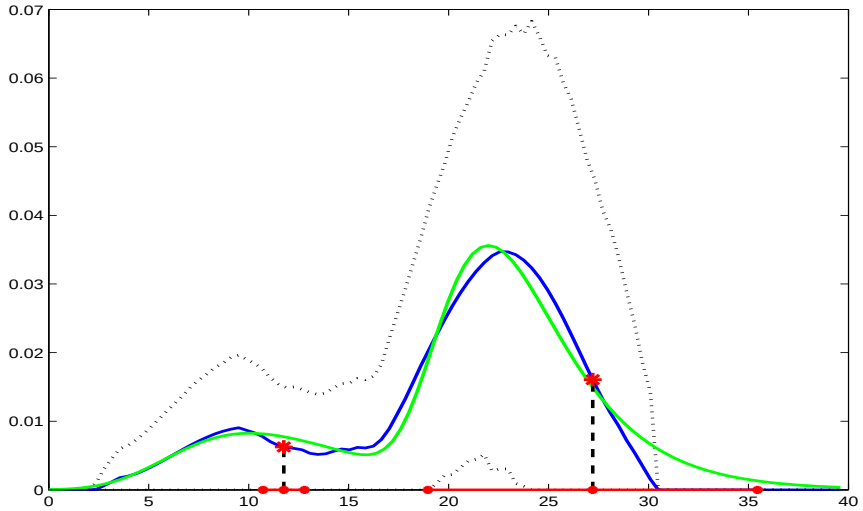


Kaplan-Meier estimate of the survival function

The points of the most rapid change

$$\hat{\theta}_1 = 11.76$$

$$\hat{\theta}_2 = 27.21$$



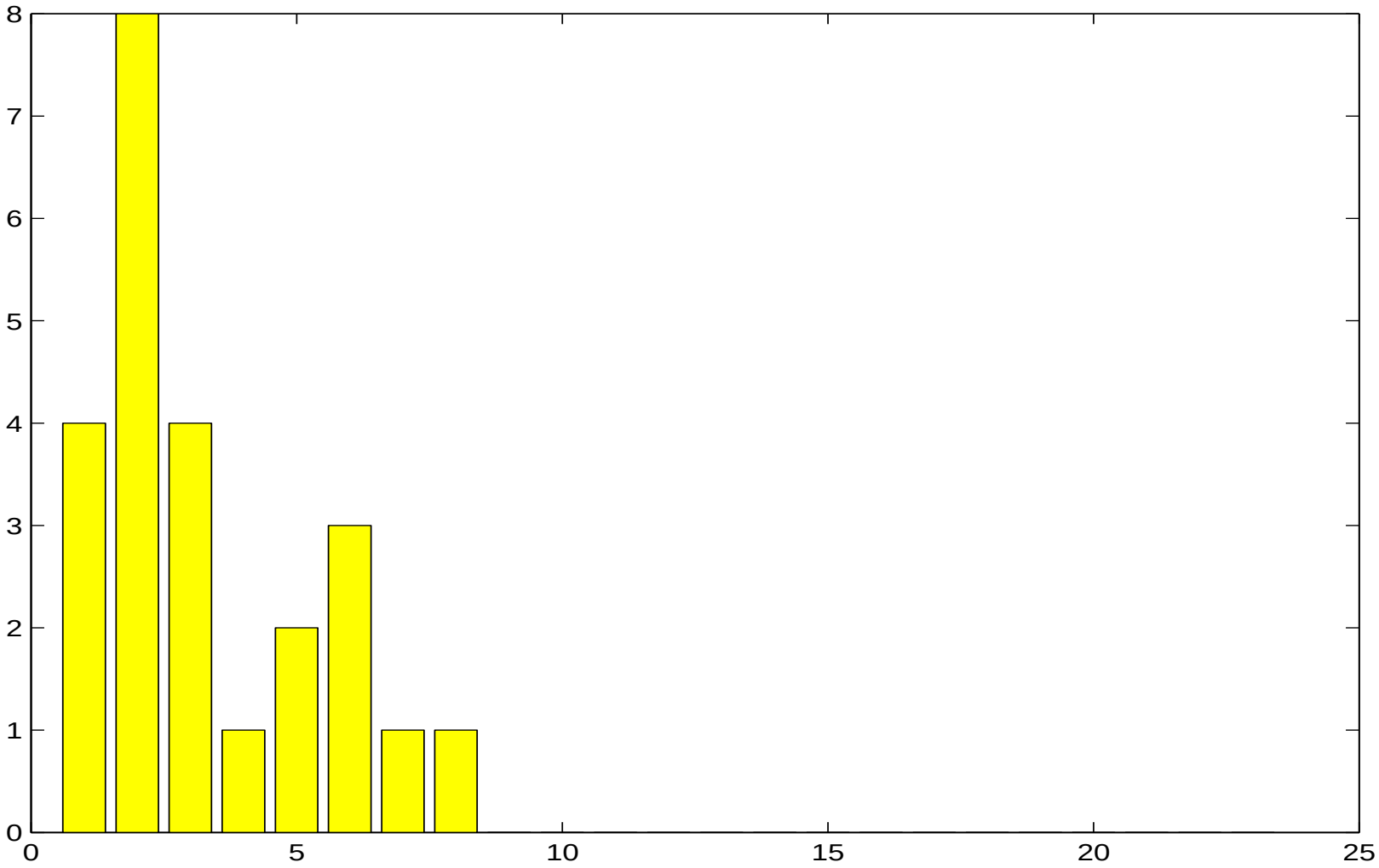
$$\hat{h}_{0,opt} = 9.4341$$

kernel estimate, deterministic estimate,

points of the most rapid change

ACUTE LEUKEMIA

Number of deaths in particular quarters

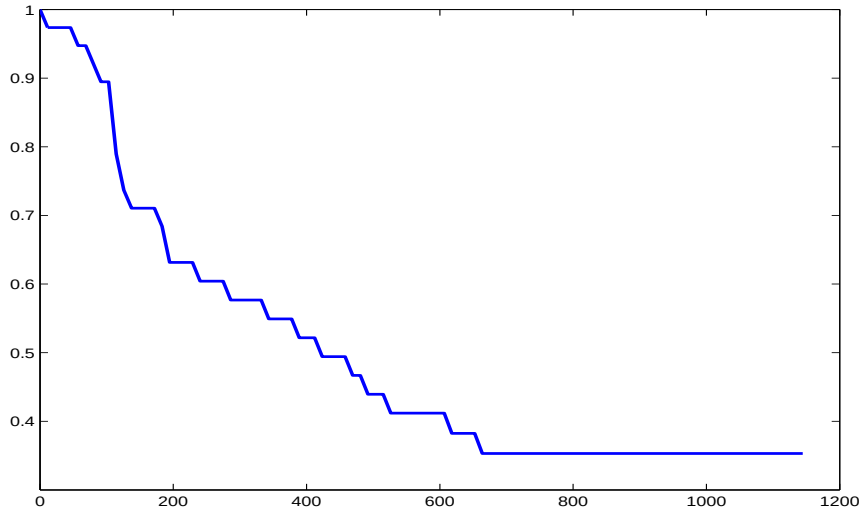


ALL

n	τ	n_d	p_d
38	**2081**	**24**	**60.5**

n **number of patients** τ **max. observation time**

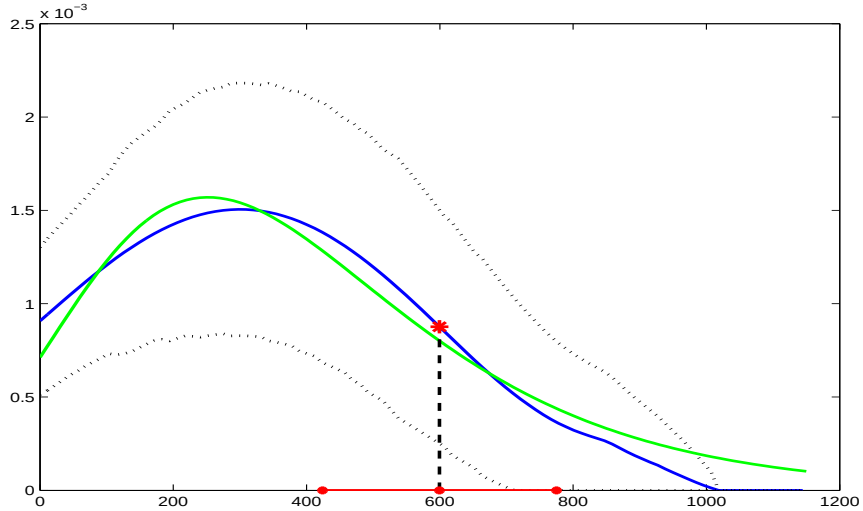
n_d **number of deaths** p_d **percentage of death**



Kaplan-Meier estimate of the survival function

The points of the most rapid change

$$\hat{\theta}_1 = 599.52$$



$$\hat{h}_{0,opt} = 735.855$$

kernel estimate, deterministic estimate,

points of the most rapid change

CONCLUSION

Cancer patients survival can be characterized by parameters:

- a **the maximal possible rate of increase of the tumor – characterizes the disease**
- λ^* **the maximal risk,** $\lambda^* = \lambda(a, b, \rho)$ **– aggressiveness of the disease**
- t^* **time of the maximal risk,** $t^* = t(a, b, y_0)$ **– effectiveness of the treatment**

EQUATABILITY INDEX

Equatability index provides an information about rarity or commonness of particular subcohorts total set of objects.

Simpson's equatability index

$$E = \frac{1}{l \sum_{i=1}^{l} \alpha_i^2}, \quad 0 < E \leq 1$$

Breast carcinoma, $E = 0.656$

i	1	2
α_i	0.138	0.862
a_i	0.0181	0.0579
λ_i^*	0.0865	0.00302
t_i^*	94.3	103

Cervical carcinoma, $E = 0.733$

i	1	2
α_i	0.198	0.802
a_i	0.198	0.326
λ_i^*	0.0503	0.0389
t_i^*	11.0	22.1

Acute leukemia

a	0.126
λ^*	0.0478
t^*	8.28

REFERENCES

[1] COLLETT D. (2003). *Modelling Survival Data in Medical Research*. Chapman & Hall/CRC: Boca Raton-London-New York-Washington,D.C.

[2] DOLEŽELOVÁ H., ŠLAMPA P., ONDROVÁ B., GOMBOŠOVÁ J., SOVADINOVÁ Š., NOVOTNÝ T., BOLČÁK K., RŮŽIČKOVÁ J., HYNKOVÁ L and FORBELSKÁ M. (2008). The impact of PET in radiotherapy treatment planning and in the prediction on patients with cervix carcinoma – result of pilot study. *Neoplasma*, **55**(5), 437–441.

[3] HOROVÁ I., ZELINKA J and BUDÍKOVÁ M. (2006). Estimates of Hazard Functions for Carcinoma Data Sets. *Environmetrics*, **17**, 239–255.

[4] HOROVÁ I. and ZELINKA J. (2006). Kernel Estimates of Hazard Functions for Biomedical Data Sets. In *Applied Biostatistics: Case studies and Interdisciplinary Methods*, Springer.

[5] HOROVÁ I., POSPÍŠIL Z. and ZELINKA J. (2008). Semiparametric Estimation of Hazard Function for Cancer Patients, *Sankhya*, **69**, 494–513.

REFERENCES

[6] HOROVÁ I., POSPÍŠIL Z. and ZELINKA J. (2009). Hazard function for cancer patients and cancer cell dynamics, *Journal of Theoretical Biology*, **258**(3), 437–443.

[7] KOZUSKO F and BAJZER Ž. (2003). Combining Gompertzian growth and cell population dynamics. *Mathematical Biosciences*, **185**, 153–167.

[8] MÜLLER H.G. and WANG J.L. (1990). Nonparametric Analysis of Changes in Hazard Rates for Censored Survival Data: An alternative Change-Point Models. *Biometrika*, **77**(2), 305–314.

[9] RAMLAU-HANSEN H. (1983). Counting Processes Intensities by Means of Kernel Functions. *The Annals of Statistics*, **11**(2), 453–466.

[10] SOUMAROVÁ R., HOROVÁ H., RŮŽIČKOVÁ J., ČOUPEK P., ŠLAMPA P., ŠENEKLOVÁ Z., PETRÁKOVÁ K., BUDÍKOVÁ M. and HOROVÁ I. (2002). Local and Distant Failure in Patients with Stage I and II Carcinoma of the Breast Treated with Breast-Conserving Surgery and Radiation Therapy (in Czech, English summary). *Radiační onkologie*, **2**(1), 17–24.

REFERENCES

[11] TANNER M.A. and WONG W.H. (1983). The Estimation of the Hazard Function from Randomly Censored Data by the Kernel Method. *The Annals of Statistics*, **11**(3), 989–993.

[12] UZUNOGULLARI U. and WANG J.L. (1992). A comparision of Hazard Rate Estimators for Left Truncated and Right Censored Data. *Biometrika*, **79**(2), 297–310.