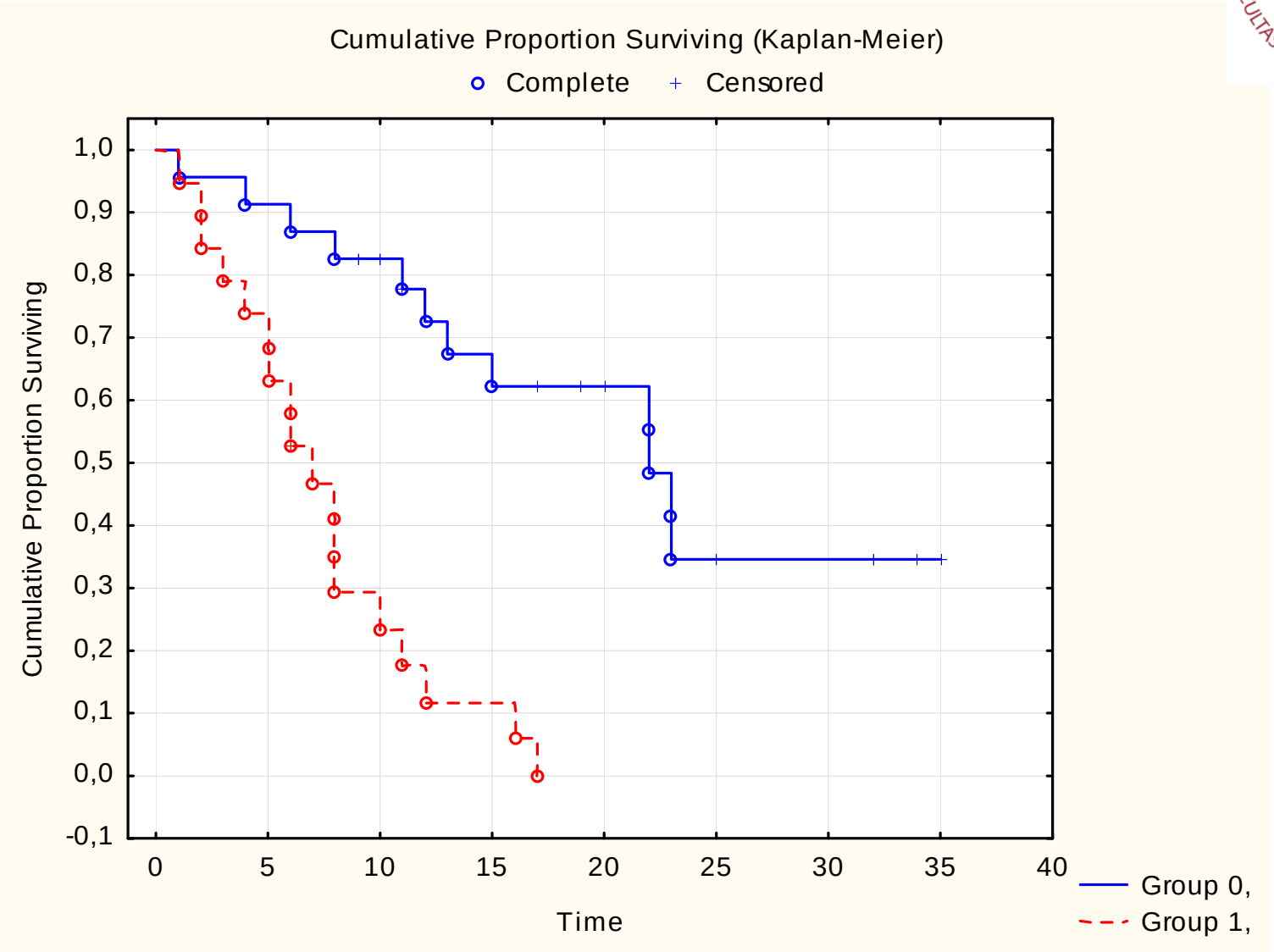
Analýza přežití



Petr Škarka
348736

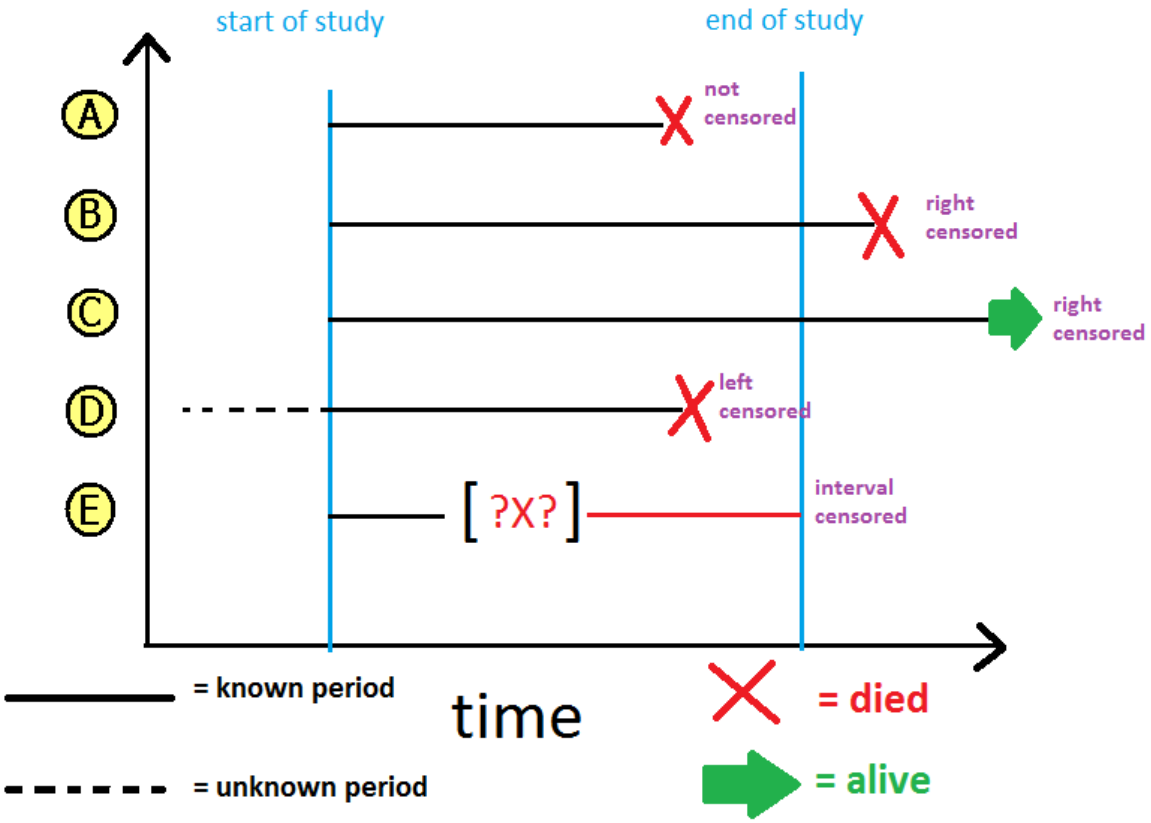
Analýza přežití (Survival analysis; SA)

- Soubor statistických procedur analyzujících chování NV **doba přežití**
- Doba přežití udává *čas od doby počátku sledování do události*
- Jednoduše zkoumáme dobu do nastání nějaké události
    - Událostí je obecně jakékoliv selhání (smrt, default,…)
    - Čas měřen v hodinách, dnech, měsících, kvartálech,…

- Aplikace:
    - Demografie
        - Odchod od rodičů,recidiva vězně po propuštění, rozchod páru či rozvod,…
    - (Bio)medicína
        - Doba do propuknuti nákazy virem HIV u narkomanů, remise u onkolog. pacientů,…
    - Technika
        - Porucha přístroje,…
    - Marketing
        - Míra návratnosti po zavedení výrobku na trh, požadavek na reklamaci po jeho prodeji
    - ***Finance***
        - Kdy klient přestane splácet úvěr
        - Jaká je doba do zaplacení dlužného pojistného
        - Modelování úmrtnosti pro účely finančního rozhodování
        - Atd.

Proč analýza přežití a ne jiné metody?

- Využití neúplné informace, kdy nenastala sledovaná událost – ***cenzorovaná pozorování***
  - Sledování času do selhání objektu (po uzavření úvěru, podepsání pojistné smlouvy,…)
    - Doba trvání studie < doba do selhání… neznáme skutečnou dobu přežití
    - Objekt přestal vykazovat činnost… známe datum poslední kontaktu
    - Vyloučení objektu z nějakých důvodů nesouvisejících se studií… stále cenná informace
  - Více dat pro modelování doby přežití a následné (finanční) rozhodování
- Ideální by byl dataset bez cenzorovaných dat → nereálný předpoklad



- Studie = pojistná doba, doba trvání úvěru,…

- 5 sledovaných subjektů:
  - A – *selhání* (zde smrt)
  - B – přežil dobu studie (a zanedlouho zemřel (*cenzorování*))
  - C – přežil dobu studie (*cenzorování*)
  - D – *cenzorování* – vyňat se studie (např. subjekt přejel vlak nebo přestal reagovat, což nesouvisí s naší studií)
  - E – *cenzorování* – nevíme, kdy přesně zemřel, ale víme, že zemřel

Cenzorování a jiné důvody pro analýzu přežití

- Cenzorování
  - ***Zprava*** – víme pouze to, že skutečná doba do události je větší než pozorovaná doba (konec studie)
  - *Zleva* – víme jen, že doba přežití není větší než pozorovaná doba (HIV a narkomani)
  - *Intervalové* – událost nastala v intervalu (viz subjekt E)

  - ***Neinformativní*** – důvody pro cenzorování nesouvisí s rizikem výskytu události;subjekty cenzorované v čase t_i jsou stále v daném časem součástí „přežívajících“ subjektů
  - *Informativní* – souvislost s rizikem nastání události; vychyluje odhady estimátorů

- „Usekávání“ (truncation)

- Další důvody upřednostnění metod SA:
  - Vysvětlující proměnná je čas
  - Data přežívání nemívají symetrické rozdělení
    - Zešikmené rozložení v histogramu – posunuté těžiště
    - U rizikové funkce (definice později) často proložení přes Weibull, log-normal,…

Označení a terminologie

doba přežití* T** — Spojitá NV udávající čas, jež uplyne od počátku sledování po selhání, příp. cenzorování. Realizaci T značíme ***t a nabývá nezáporných hodnot. Rozložení prstí popsáno hustotou pravděpodobností *f(t)* a distribuční funkcí *F(t)*.

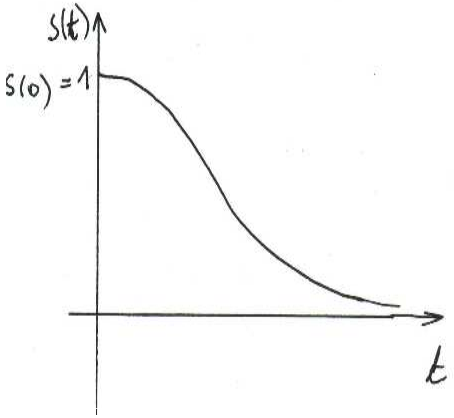
***indikace selhání* δ** — Alternativní NV, kde 1 = selhání a 0 = cenzorování, tj. přežití doby studie nebo vyloučení ze sledovaného souboru (vlak, ztracený kontakt,…).

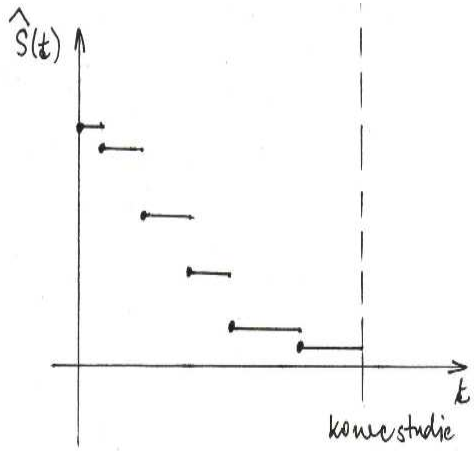
***funkce přežití* S(t)** — **S(t) = P(T > t)**, tedy vyjadřuje pravděpodobnost přežití okamžiku t.
Teoretická funkce přežití je spojitá a nerostoucí.
Většinou uvažujeme S(0) = 1 a S(t) $\longrightarrow$ 0 pro t $\longrightarrow$ inf.
Opačný jev distribuční funkce S(t) = P(T > t) = 1 – P(T ≤ t) = 1 – F(t).

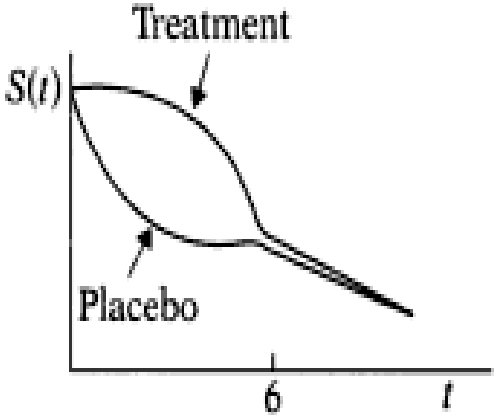
***odhad* Ŝ(t)** — Data nesbíráme nekonečně dlouho, pak v čase i může platit $S(t_i) \geq 0$.

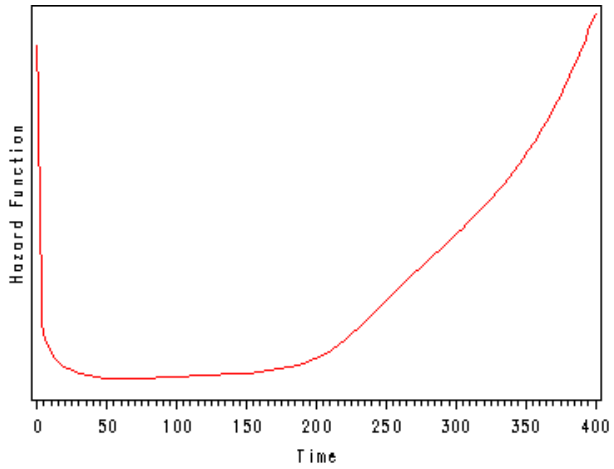
***riziková funkce* h(t)** — Platí, že $$h(t) = \lim_{dt \longrightarrow 0} \frac{P(t \leq T < t + dt / T \geq t)}{dt}$$

Nevyjadřuje pravděpodobnost selhání v čase t, ale *podmíněnou míru rizika selhání v čase t* při dožití se času t („tendence selhat po přežití okamžiku t“).
Nezáporná (P) a hodnoty závisí na jednotkách času, v nichž měříme.









•Jak funkce přežití, tak riziková funkce popisují chování NV T. Interpretace se ale liší. S(t) jako pravděpodobnost přežití okamžiku *t* (P neselhání), h(t) naopak podmíněným selháním.

•Jednu lze odvodit z druhé; po vyjádření: $h(t) = -\frac{1}{S(t)} \cdot \frac{\partial S(t)}{\partial t}$ a $S(t) = \exp\left\{-\int_0^t h(u)du\right\}$

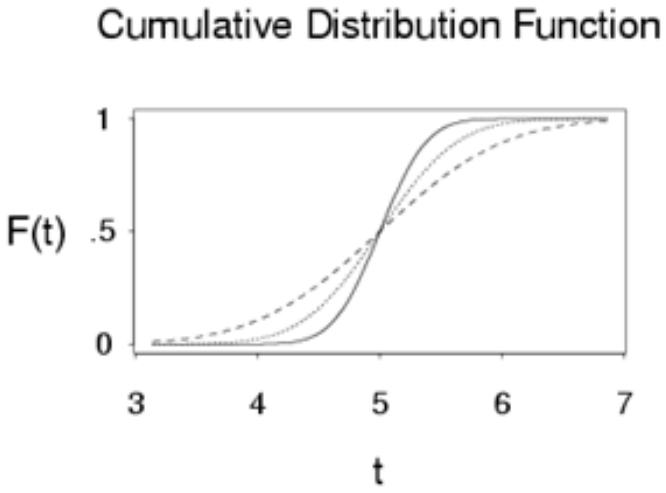
•Znalost průběhů je vhodná k posouzení pravděpodobnosti přežití t (S(t)), popř. podmíněného rizika selhání v t (h(t)).

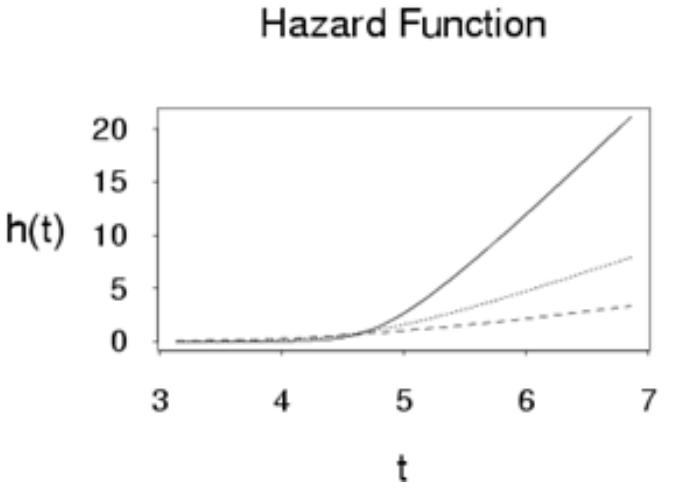
•***Cíle SA***:

I. Odhad a interpretace funkce přežití a rizikové funkce
II. Srovnání funkce přežití a rizikové funkce ve více skupinách
III. Modelování vztahu mezi časem přežití a vysvětlujícími proměnnými

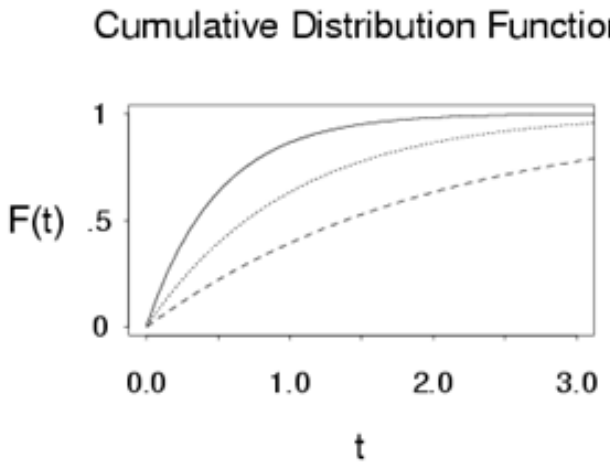
Ukázka distribučních a rizikových funkcí pro různá rozdělení*

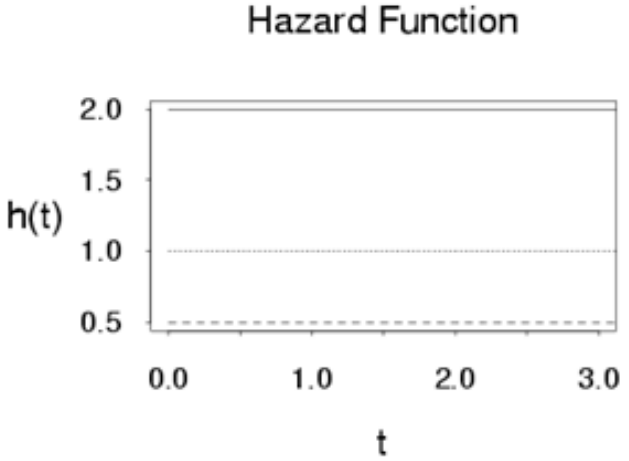
Normální rozdělení



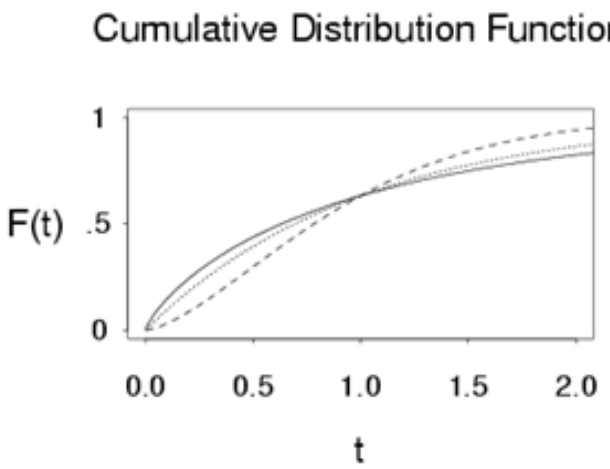


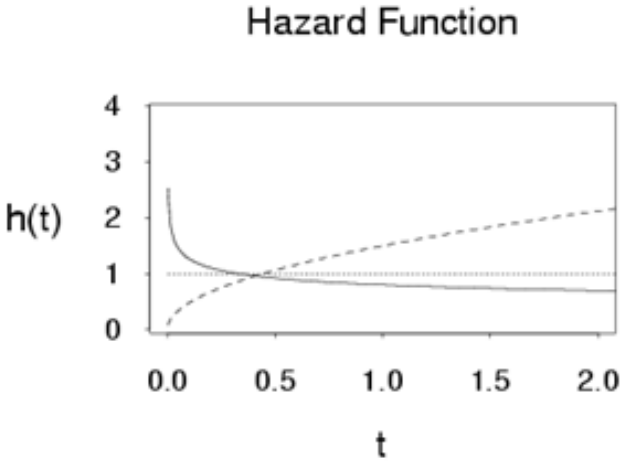
Exponenciální rozdělení





Weibullovo rozdělení





*a různé parametry

Kaplan-Meier estimátor

$t_{(j)}$	n_j	m_j	q_j	$\hat{S}(t_{(j)})$
0	21	0	0	21/21=1,00
1	21	2	0	19/21= 0,90
2	19	2	0	17/21= 0,81
3	17	1	0	16/21=0,76
4	16	2	0	14/21=0,67
5	14	2	0	12/21=0,57
8	12	4	0	8/21=0,38
11	8	2	0	6/21=0,29
12	6	2	0	4/21=0,19
15	4	1	0	3/21=0,14
17	3	1	0	2/21=0,10
22	2	1	0	1/21=0,05
23	1	1	0	0/21=0,00

- S(t) není obvykle známá → odhad z dat
- Bez cenzorovaných údajů (*klasický přístup*):

P(T > t_i) = (počet objektů s T > t_i / počet všech objektů souboru)

- Smysl SA je však právě ve využití cenzorovaných dat
- Nejčastěji užívaný odhad S(t) je ***Kaplan-Meierův***
- Uvažujme okamžiky, kdy došlo k selhání: $t_0 < t_1 < t_2 < \ldots < t_k$.
- Lze odvodit: $S(t_j) = S(t_{j-1}) \cdot P(T > t_j / T \geq t_j)$ pro j = 1,…,k
- Odvození:

$$S(t_j) = P(T > t_j) = P(T \geq t_j \wedge T > t_j) = P(T > t_{j-1} \wedge T > t_j) =$$
$$= P(T > t_{j-1}) \cdot P(T > t_j / T > t_{j-1}) = S(t_{j-1}) \cdot P(T > t_j / T \geq t_j)$$

- Vztah je rekurentní, pak:

$$S(t_j) = S(t_0) \cdot P(T > t_1 / T \geq t_1) \cdot P(T > t_2 / T \geq t_2) \cdot \ldots$$
$$\ldots \cdot P(T > t_j / T \geq t_j) = 1 \cdot \prod_{i=1}^{j} P(T > t_i / T \geq t_i)$$

Kaplan-Meier estimátor

$t_{(j)}$	m_j	q_j	n_j
$t_{(0)} = 0$	$m_0 = 0$	q_0	n_0
$t_{(1)}$	m_1	q_1	n_1
$t_{(2)}$	m_2	q_2	n_2
$\vdots$			
$t_{(k)}$	m_k	q_k	n_k

- K-M odhaduje $P(T > t_i / T \geq t_i)$ podílem $\frac{n_i - m_i}{n_i}$, kde:
  - n_i = počet objektů, které jsou v čase t_i v riziku (tj. platí pro ně $T \geq t_i$)
  - m_i = počet objektů, které selhaly přesně v čase t_i
- Čím větší rozsah výběru n a zároveň čím více selhání, tím přesnější výsledky
  - U SA je především důležitý počet selhání
  - Mnoho cenzorovaných dat → moc se toho nedozvíme
- Pro j = 1,…,k lze pak odhad pomocí K-M napsat jako $\hat{S}(t_j) = \prod_{i=1}^{j} \frac{n_i - m_i}{n_i}$
  - Pro $t \in [\, t_i ; t_{i+1})$ klademe $\hat{S}(t) = \hat{S}(t_i)$
- Cenzorovaná pozorování nezpůsobují skok v odhadu – navyšují rozsah výběru ve smyslu počtu objektů v riziku (n_i)
- (při cenzorování zleva trochu komplikovanější)

- K-M odhad fce přežití je neparametrický ML odhad

Kaplan-Meier estimátor

- Jsou-li data necenzorovaná, pak K-M vede ke stejným hodnotám jako klasický přístup
  - Necenzorovaná data: $n_j = n_{j-1} - m_{j-1}$

$$\hat{S}(t_j) = \prod_{i=1}^{j} \frac{n_i - m_i}{n_i} = \frac{n_1 - m_1}{n_1} \cdot \frac{n_2 - m_2}{n_1 - m_1} \cdot \ldots \cdot \frac{n_j - m_j}{n_{j-1} - m_{j-1}} = \frac{n_j - m_j}{n_1} = \frac{n_i - m_i}{n}$$

- Na obrázku vpravo výpočet Ŝ(t)
- Sloupec q_j obsahuje cenzorovaná pozorování
- Lze si všimnout, že v t_j nejsou součástí m_j , tzn. stále je využíváme jako součást n_j.

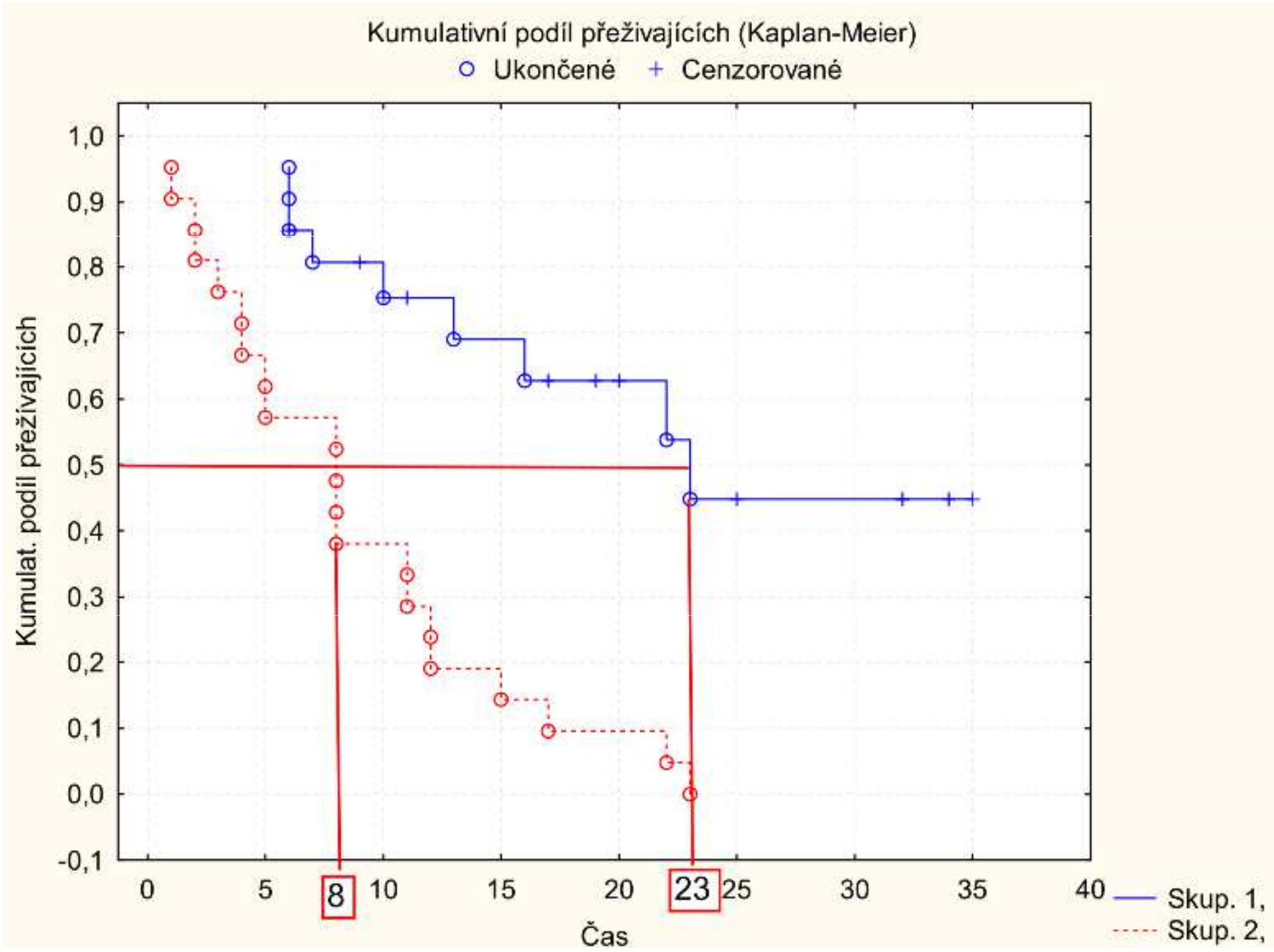
$t_{(j)}$	n_j	m_j	q_j	$\hat{S}(t_{(j)})$
0	21	0	0	1
6	21	3	1	$1 \cdot \frac{18}{21} = 0,8571$
7	17	1	1	$0,8571 \cdot \frac{16}{17} = 0,8067$
10	15	1	2	$0,8067 \cdot \frac{14}{15} = 0,7529$
13	12	1	0	$0,7529 \cdot \frac{11}{12} = 0,6902$
16	11	1	3	$0,6902 \cdot \frac{10}{11} = 0,6275$
22	7	1	0	$0,6275 \cdot \frac{6}{7} = 0,5378$
23	6	1	5	$0,5378 \cdot \frac{5}{6} = 0,4482$

- Ŝ(t_j) je statistika → var(Ŝ(t_j))?
  - Odhad pomocí ***Greenwoodovy formule***

$$\text{var}(\hat{S}(t_j)) = \hat{S}^2(t) \sum_{i=1}^{j} \frac{m_i}{n_i \cdot (n_i - m_i)}$$

  - Pro fixní t ($t \in [\, t_i ; t_{i+1} \,)$) má statistika přibližně N rozložení Ŝ(t) ~ N(S(t), var (S(t))
  - Pak lze odvodni 100(1-α)%-ní asymptotický IS pro skutečnou hodnotu S(t)
  - IS je symetrický kolem hodnoty Ŝ(t)

Grafický výstup (STATISTICA)



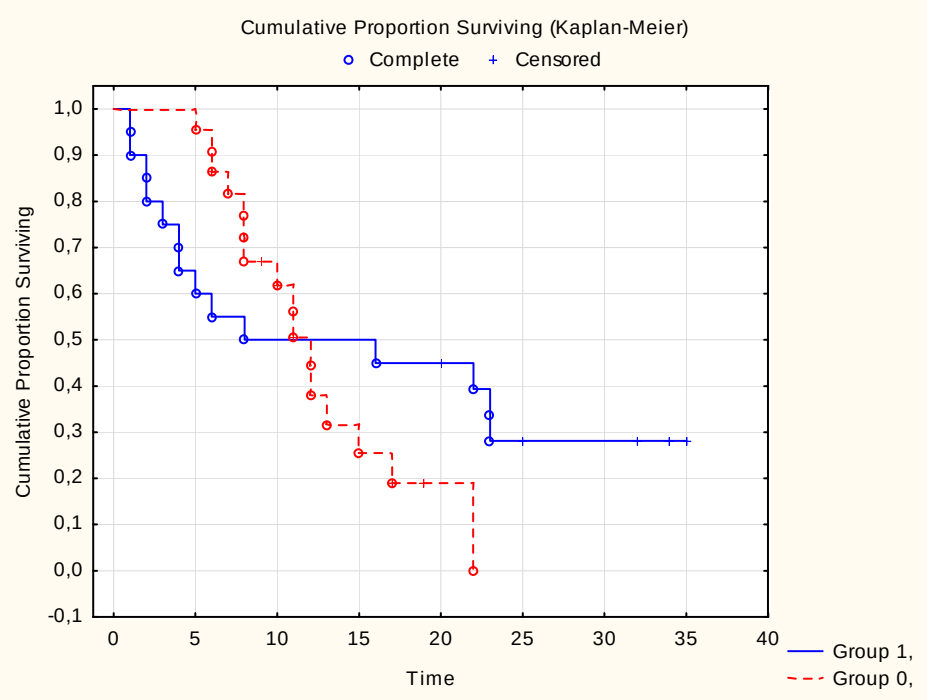
Logrank test

- Obecně test pro posouzení rozdílu mezi dvěma skupinami za určitých předpokladů (zde OK)
- Postaveno na χ^2 statistice
  - Rozdíl mezi pozorovanými (O) a očekávanými (E) četnostmi

- Funkční hodnoty funkce přežití mohou být pro různé skupiny rozdílné*
  - Je tato rozdílnost statisticky významná?
  - H_0: neexistuje rozdíl mezi objekty ze skupin × H_1: doba přežití se mezi skupinami liší
  - Nezamítnutí H_0
    - Rozdíl být nemusí,
    - Ale může – problém křížení funkcí

- Formálně: $LR = \frac{(O_2 - E_2)^2}{\mathrm{var}(O_2 - E_2)}$, kde LR ≈ χ^2 (1)

(více viz stanford.edu/~kcobb/hrp262/lecture2.ppt)



*různé odlišnosti mezi skupinami – různá síla logrank testu → vážené logrank testy

Coxův* model proporciálního rizika

- Regresní model v SA; na závislou proměnnou doba přežití působí několik regresorů
- Alternativa k parametrickému regresnímu modelu AFT (accelerated failure time)

- Nejčastěji ve tvaru vycházejícím z rizikové funkce h(t), lze jej přepsat i do tvaru s S(t)

- Coxův model: $h(t, X) = h_0(t) \cdot e^{\sum_{i=1}^{p} \beta_i X_i} = h_0(t) \cdot \exp\{X' \beta\}$ pro t ≥ 0

  - $\boldsymbol{X} = (X_1, \ldots, X_p)'$ je vektor prediktorů
  - $\boldsymbol{\beta} = (\beta_1, \ldots, \beta_p)'$ je vektor neznámých parametrů, které odhadujeme
  - $h_0(t)$ je *základní riziko* (baseline hazard function). Neznáme ji, plní roli interceptu. Pro $(X_1, \ldots, X_p) = (0, \ldots, 0)$ je $h(t, \boldsymbol{X}) = h_0(t)$.
  - $e^{\sum_{i=1}^{p} \beta_i X_i}$ výraz nezávislý na čase!

*Sir ***David Roxbee Cox*** (1924 –), britský statistik
- *Nonparametric estimation from incomplete observations*, #2 WoS (#1 K-M)

Coxův model

- Coxův model je *semiparametrický*
  - Základní riziko $h_0(t)$ není specifikovaná funkce – o jejím rozložení nic nepředpokládáme
  - Parametrická regresní složka

- Proč je Coxův model populární?
  - exp{$\boldsymbol{X}'\boldsymbol{\beta}$} vždy kladné, pak $0 \leq h(t,\boldsymbol{X}) < \inf$
  - Dobře aproximuje parametrické modely
  - ($\beta_1, \ldots, \beta_p$) a poměry rizik (viz dále) lze odhadnout i při nespecifikovaném riziku $h_0(t)$ **!!!**
  - Lze s ním získat odhady zákl. rizika $h_0(t)$, rizikové funkce $h(t,\boldsymbol{X})$ a funkce přežití $S(t,\boldsymbol{X})$

- Interpretace:
  - Př.) datový soubor určitého pojistného kmene (selhání je přestání placení pojistného), u jehož objektů sledujeme 5 regresorů (pohlaví, věk, vzdělání, plat, rodinný stav). Zajímá nás, jak se liší riziko selhání v čase t liší u *SŠ* vzdělaných *žen* ve věku *50-55 let* s platem z intervalu *[15k;20k)* lišící se v rodinném stavu (*0* = svobodná, *1* = vdaná). Poměr rizik lze pak zapsat jako:

$$HR = \frac{h_1(t,x^*)}{h_2(t,x)} = \frac{h_0(t)\cdot\exp\left\{\beta_1\cdot x_1+\beta_2\cdot x_2+\beta_3\cdot x_3+\beta_4\cdot x_4+\beta_5\cdot x_5^*\right\}}{h_0(t)\cdot\exp\left\{\beta_1\cdot x_1+\beta_2\cdot x_2+\beta_3\cdot x_3+\beta_4\cdot x_4+\beta_5\cdot x_5\right\}} = e^{\beta_5\cdot(x_5^*-x_5)}$$

  - $e^{\beta_5\cdot(x_5^*-x_5)}$ říká, kolikrát vzroste riziko selhání u vdané (h_1), viz $h_1(t,x) = e^{\beta_5\cdot(x_5^*-x_5)}\cdot h_2(t,x)$

Coxův model

- Číslo $e^{\beta_i \cdot (x_i^* - x_i)}$ nazýváme relativní riziko vzhledem k prediktoru x_i.
  - Nezmíněné prediktory jsou fixní

- Obecně lze podíl rizikových funkcí, *hazard ratio* (HR), zapsat jako

$$HR = \frac{h(t, x^*)}{h(t, x)} = \exp\left\{\sum_{i=1}^{p} \beta_i (x_i^* - x_i)\right\}$$

- Výše uvedený vztah platí pro libovolné rozdíly v hodnotách prediktorů (o počtu 1 až p)

- Odhady $\hat{\beta}_i$ a $e^{\hat{\beta}_i}$ získáváme pomocí parciální věrohodnostní funkce
  - Hustota doby přežití v Coxově modelu není specifikovaná parametrické funkce → úpravy
- Intervaly spolehlivosti pro dané odhady statistik pak pomocí Waldovy statistiky

- Lze kromě parametrických $\beta_1, \ldots, \beta_p$ odhadnout i neparametrickou složku $h_0(t)$?
  - Ano, Gehan-Breslow test
  - Pak lze kreslit odhady funkce přežití pro specificky nastavené hodnoty regresorů
    - Např. jak vypadá funkce přežití (defaultu) pro objekty „muž – [20;25) let – ZŠ“.

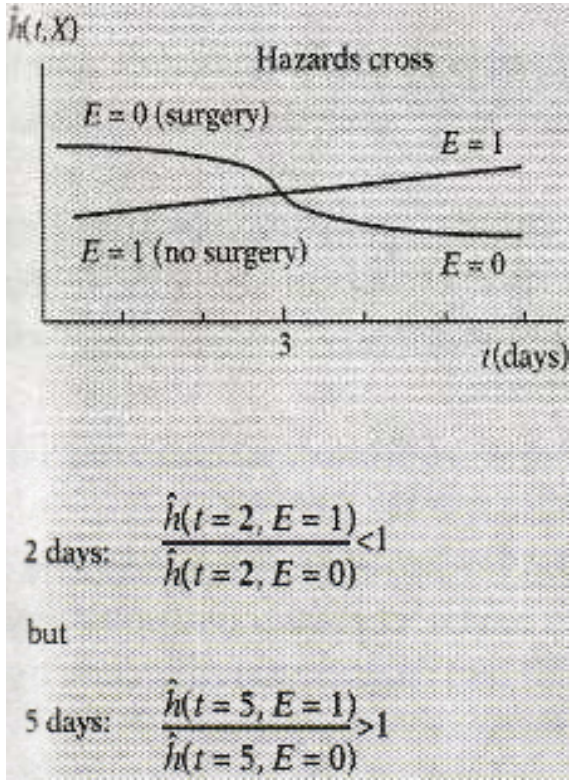
Coxův model - předpoklady

•Žádné na specifikaci základního rizika $h_0(t)$, ALE za to ***předpoklad proporciálního rizika*** – relativní riziko HR se vzhledem k jednotlivým regresorům v čase *t* nemění. Pak pro dva objekty, které se liší hodnotou regresorů platí $HR = \frac{h(t, x^*)}{h(t, x)} = konst.$

•Na obrázku vpravo porušení proporcionality (překřížení rizik. funkcí) ⟶ Coxův model není vhodný

•Jak ověříme proporcionalitu?

- •*Graficky* – na ose Y logaritmy kumulativních rizikových funkcí (lze získat přes Kaplan-Meiera) pro každý regresor a na ose X čas nebo log času ⟶ musí být rovnoběžné
- •*Rezidua* – vhodná jsou Schoenfeldova rezidua pro každý z regresorů



Aplikace ve financích

- Metody credit scoringu = ***jestli*** dojde k defaultu × SA = ***kdy*** k defaultu dojde
  - Využití ⟶ kvantifikace profitability klienta
  - Výstupem je odhad funkce v čase
- Navíc na rozdíl od credit scoringu SA zohledňuje cenzorovaná data (v CS *zřejmě* zmenšení tréningové matice vypuštěním těchto dat, čímž ale přicházíme o data, se kterými se v realitě také setkáváme)

Survival analysis in credit scoring (2009)

- Dataset k bankovním úvěrům
- Skupiny proměnných
  - Popis dlužníků, typ úvěru, definice defaultu (ihned/90 dní/180 dní po splatnosti)
- Úpravy dat (outliers, kategorizace proměnných, korelující kovariáty,…)

Number of defaults for different definitions

10.6%	Client definition
5.5%	90 days past due
4.3%	180 days past due

The list of given variables

sex	age	marital status
education	employment status (since)	employer
housing status	repayment type	credit card holder
kind of employment	type of private phone	type of phone at work
No. of people in the household	monthly income	other income
limit of loan	distribution channel of loan	type of loan
date of the first drawing	maturity of loan	

Survival analysis in credit scoring (2009) – average client

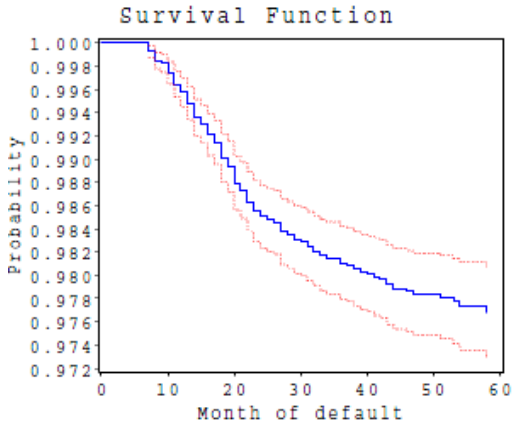
- Průměrný klient datasetu + (M 58 %, Ž 42 %):

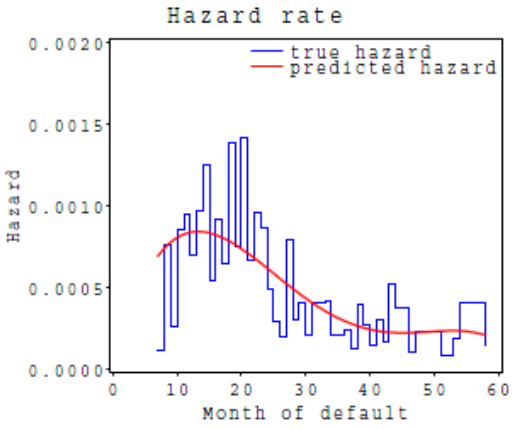
variable	category	average client
education	less than full secondary education	0.3919
	full secondary education (with school leaving exam)	0.3691
	higher then full secondary education	0.2390
employment stability	employed < 1 year	0.0851
	employed > 1 year and < 5 years or student or other	0.3661
	employed > 5 years	0.5488
repayment type	payment on account from same bank	0.9407
	payment in cash or other	0.0593

- Ukázka odhadů parametru β tří statisticky nejvýznamnějších proměnných:

variable	category	$\hat{\beta}$
education	less than full secondary education	1.18102
	full secondary education (with school leaving exam)	0.50992
	higher then full secondary education	0
employment stability	employed < 1 year	1.09880
	employed > 1 year and < 5 years or student or other	0.82072
	employed > 5 years	0
repayment type	payment on account from same bank	-1.3370
	payment in cash or other	0

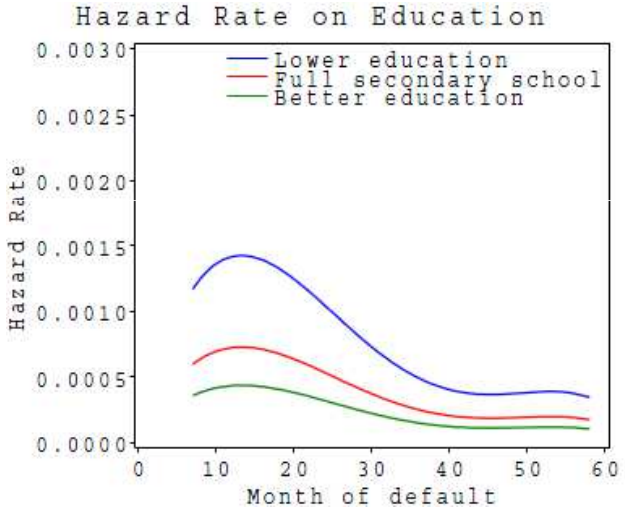
- Funkce přežití s IS (α = 5 %) a riziková funkce průměrného klienta:

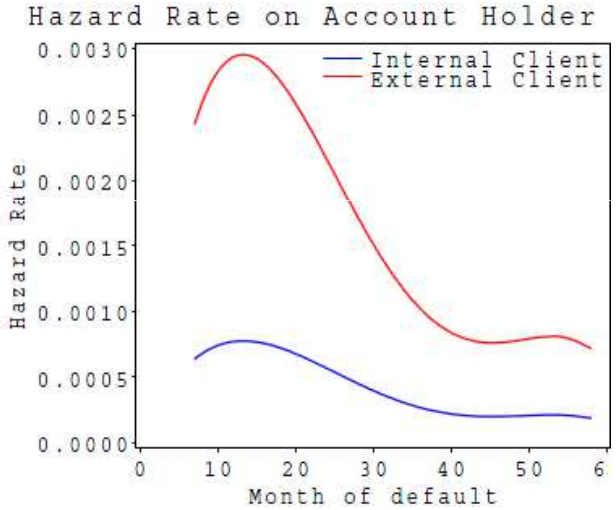




Survival analysis in credit scoring (2009)

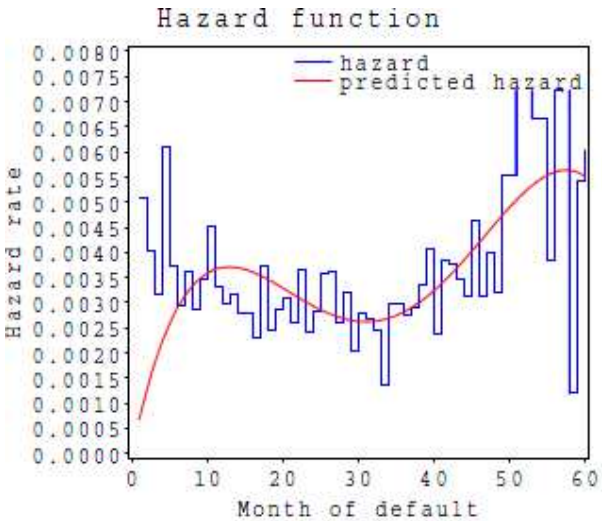
- Efekt vzdělání (nalevo) a splátek úvěru z BÚ u úvěrové či externí instituce (napravo) na *h(t)*:



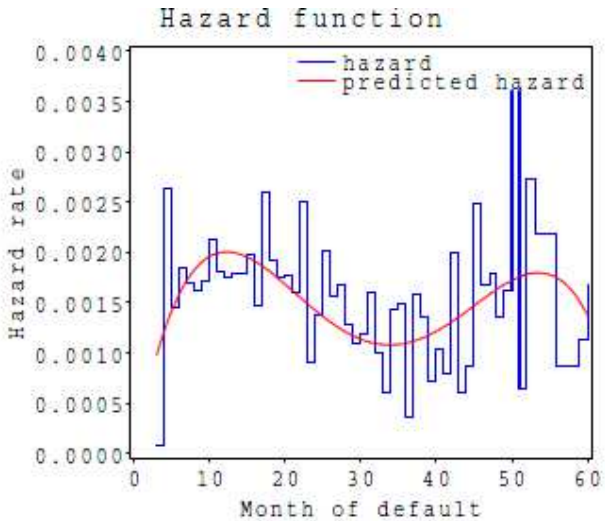


- Ploty rizikové funkce dle definic defaultu:

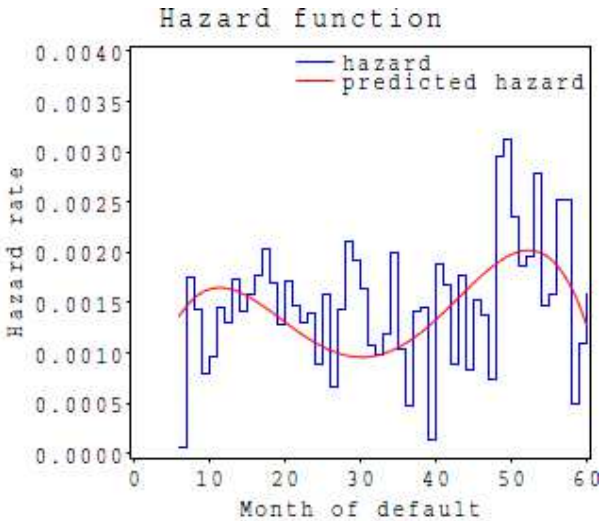
Ihned



90 dnů po splatnosti



180 dnů po splatnosti



Zdroje

•Mgr. Maria Králová, PhD. – materiály k předmětu *Statistika 3* (BPM_STA3), ESF MU.
•Kristin Sainani, PhD. – materiály k předmětu *Longitudinal Data Analysis*, dostupné z: <http://www.stanford.edu/~kcobb/courses/hrp262/index.html>

•Mgr. Michaela Stehlíková – diplomová práce *Analýza přežití a populační dynamika*, PřF MU. Dostupné z: <http://is.muni.cz/th/175404/prif_m/Diplomova_prace.pdf>
•Mgr. Martin Hrba – prezentace *Odhady aktuárských předpokladů pomocí analýzy přežití*, dostupné z: <http://www.actuaria.cz/upload/SAV%2011-05-2012.pdf>

•Pazdera, Rychnovský a Zahradník - projekt *Survival analysis in credit scoring*, MFF UK. Dostupné z: <http://artax.karlin.mff.cuni.cz/~rychm5am/Project.pdf>