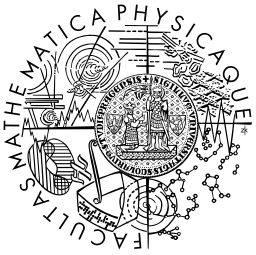


MATEMATICKO-FYZIKÁLNÍ
FAKULTA
Univerzita Karlova

BAKALÁŘSKÁ PRÁCE

Zuzana Teichmannová

Odhad rozdělení pravděpodobnosti při cenzorovaných datech

Katedra pravděpodobnosti a matematické statistiky

Vedoucí bakalářské práce:	doc. RNDr. Petr Lachout, CSc.
Studijní program:	Matematika
Studijní obor:	Obecná matematika

Praha 2017

Prohlašuji, že jsem tuto bakalářskou práci vypracoval(a) samostatně a výhradně s použitím citovaných pramenů, literatury a dalších odborných zdrojů.

Beru na vědomí, že se na moji práci vztahují práva a povinnosti vyplývající ze zákona č. 121/2000 Sb., autorského zákona v platném znění, zejména skutečnost, že Univerzita Karlova má právo na uzavření licenční smlouvy o užití této práce jako školního díla podle §60 odst. 1 autorského zákona.

V ..... dne ..... Podpis autora

Ráda bych poděkovala panu docentovi P. Lachoutovi za zajímavé téma a vedení práce. Rovnež bych chtěla poděkovat panu profesorovi J. Antochovi za rady a připomínky.

Název práce: Odhad rozdělení pravděpodobnosti při cenzorovaných datech

Autor: Zuzana Teichmannová

Katedra: Katedra pravděpodobnosti a matematické statistiky

Vedoucí bakalářské práce: doc. RNDr. Petr Lachout, CSc., Katedra pravděpodobnosti a matematické statistiky

Abstrakt: V této práci se zabýváme odhadem rozdělení pravděpodobnosti cenzorovaných dat. Cenzorovaná data jsou data, která nebylo možné dopozorovat celá, jelikož před pozorovanou událostí nastala událost jiná, která nám zabránila výsledek dopozorovat. Nejvíce se věnujeme Kaplan-Meierovu odhadu a nějakým jeho základním vlastnostem. Podíváme se také na Nelson-Aalenův odhad. V závěru dochází k porovnání těchto odhadů s naivním odhadem, ve kterém cenzorovaná data vynecháváme. Toto porovnání je ilustrováno na dvou numerických příkladech, kde je vidět zásádní rozdíl v přesnostech odhadů a vidíme, že není vhodné cenzorovaná data při odhadování rozdělení vynechávat.

Klíčová slova: rozdělení náhodné veličiny, distribuční funkce, odhad, cenzorovaná data

Title: Estimation of probability distribution for censored data

Author: Zuzana Teichmannová

Department: Department of Probability and Mathematical Statistics

Supervisor: doc. RNDr. Petr Lachout, CSc., Department of Probability and Mathematical Statistics

Abstract: In this thesis, we look into estimation of probability distribution for censored data. These data are not complete, because for some reason it was impossible to observe them all. We use the Kaplan-Meier estimator and study some of its properties. We also use the Nelson-Aalen estimator. In the end we make a comparison of these estimators with a naive estimator, which omits the censored data. The comparison is illustrated on two numerical examples where we can see the main differences in the accuracy of the estimators. We will see that it is better to include the censored data to our estimations.

Keywords: distribution of a random variable, distribution function, estimation, censored data

Obsah

Úvod

Při práci s reálnými daty se často stává, že se mezi získanými daty objeví data neúplná, taková jež nebylo možné dopozorovat celá, např. jelikož pozorování skončilo nebo nějaký z pozorovaných subjektů přestal být sledován. Taková data se pak nazývají cenzorovaná. Typickým příkladem jsou lékařské studie, které běží po nějakou přesně danou dobu (pak se kýžený výsledek nemusí objevit před koncem studie), a při kterých je snadné nějakého pacienta ztratit (pacient přestane docházet, odstěhuje se nebo zemře). V této práci se na fiktivním numerickém příkladě ukazuje, jak odhadnout rozdělení, pokud máme takováto data. Budou porovnány tři přístupy a to klasický odhad, při kterém se pracuje jen s necenzorovanými daty, Kaplan-Meierův odhad a Nelson-Aalenův odhad.

V první části této práce je seznámení s potřebnou teorií k cenzorovaným datům a jednotlivým odhadům, poté následuje provedení odhadů a jejich porovnání. Cílem je pochopit základy odhadu rozdělení při cenzorovaných datech a to především Kaplan-Meierův odhad a tento odhad porovnat s klasickým odhadem a Nelson-Aalenovým odhadem.

Data v této práci jsou fiktivní a náhodně vygenerovaná a budou představovat doby životností strojů, tedy studii, ve které pozorujeme deset tisíc strojů po dobu dvou set hodin a měříme, za jak dlouho stroje přestanou fungovat.

1. Klíčové pojmy

1.1 Analýza přežití

Analýza přežití se zabývá metodami, které zkoumají data měřící čas výskytu nějaké události. Výskyt takové události se nazývá selhání (často selhání značí smrt, proto analýza přežití). Nicméně selhání nemusí vždy být negativní jev. ůže se jednat o jakoukoliv událost, která nás zajímá (můžeme třeba sledovat, po kolika dnech od zasazení vyroste rostlina). Na čas do výskytu takové události se můžeme dívat jako na nezápornou náhodnou veličinu. Tuto veličinu budeme značit T a nazývat čas selhání.

1.2 Rozdělení času selhání

Nyní se budeme zabývat nezápornou náhodnou veličinou T a jejím rozdělení.

Náhodná veličina T může být diskrétní, spojitá i kombinovaná. Její rozdělení může být popsáno mnoha způsoby, nejčastěji však pomocí distribuční funkce, hustoty, funkce přežití a rizikové funkce.

Definice 1. *Distribuční funkce náhodné veličiny $t \geq 0$ je definována*

$$F(t) = \mathsf{P}(T \leq t), \qquad t \geq 0.$$

Definice 2. *Funkce přežití náhodné veličiny $T \geq 0$ je definována*

$$S(t) = \mathsf{P}(T > t), \qquad t \geq 0.$$

Lemma 1 (Vlastnosti funkce přežití)**.** *Pro funkci přežití $S(t)$ náhodné veličiny $T \geq 0$ platí následující tvrzení*

(i) *platí $S(t) = 1 - F(t)$ pro všechna $t \in [0, \infty)$;*

(ii) *$S(t)$ je nerostoucí, zprava spojitá funkce času $t \in [0, \infty)$;*

(iii) *$S(0) = 1$ a $\lim_{t\to\infty} S(t) = 0$.*

Důkaz.

(i) Máme $S(t) = P(T > t) = 1 - P(T \leq t) = 1 - F(t)$.

(ii) Plyne z (i), protože o distribuční funkci $F(t)$ víme, že je neklesající, zprava spojitá funkce.

(iii) Jelikož pro distribuční funkci nezáporné náhodné veličiny platí $F(0) = 0$ a $\lim_{t\to\infty} F(t) = 1$, z (i) snadno plyne $S(0) = 1 - 0 = 1$ a $\lim_{t\to\infty} S(t) = \lim_{t\to\infty} 1 - F(t) = 0$.

□

U cenzorovaných dat se více pracuje s funkcí přežití $S(t)$, protože $S(t)$ vyjadřuje pravděpodobnost, že v intervalu $[0, t)$ nedojde k selhání.

1.2.1 Spojitá náhodná veličina T

Definice 3. *Hustotu spojité náhodné veličiny $T \geq 0$ definujeme*

$$f(t) = -\frac{dS(t)}{dt}, \qquad t \geq 0.$$

Platí $f(t) = dF(t)/dt$, $f(t) \geq 0$, $\int_0^\infty f(t)dt = 1$, $F(t) = \int_0^t f(s)ds$ a $S(t) = \int_t^\infty f(s)ds$.

Definice 4. *Riziková funkce spojité náhodné veličiny $T \geq 0$ je definována*

$$\lambda(t) = \lim_{h \to 0^+} \frac{P(t \leq T < t + h \,|\, T \geq t)}{h}, \qquad t \geq 0.$$

Riziková funkce je nezáporná a udává riziko, že v čase t nastane selhání. Někdy se rizikové funkci také říká intenzita, jelikož ji můžeme interpretovat jako intezitu selhání během celého procesu.

Pro rizikovou funkci platí

$$\begin{aligned} \lambda(t) &= \frac{f(t)}{S(t)} \\ &= -\frac{d \log S(t)}{dt}. \end{aligned}$$

Integrací a použitím $S(0) = 1$ se dostane

$$\begin{aligned} S(t) &= \exp[-\int_0^t \lambda(s)ds] \\ &= \exp[-\Lambda(t)]. \end{aligned} \tag{1.1}$$

Definice 5. *Pro náhodnou veličinu $T \geq 0$ s rizikovou funkcí $\lambda(t)$ se funkce*

$$\Lambda(t) = \int_0^t \lambda(s)ds$$

nazývá kumulativní riziková funkce.

Derivací výrazu (1.1) získáme

$$f(t) = \lambda(t) \exp[-\Lambda(t)]. \tag{1.2}$$

Rovněž platí, že libovolná spojitá funkce $\lambda(t)$ splňující

$$\int_0^t \lambda(s)ds < \infty$$

pro nějaké $t \geq 0$ a

$$\int_0^\infty \lambda(s)ds = \infty,$$

je rizikovou funkcí nějaké spojité náhodné veličiny.

Definice 6. *Nechť T je nezáporná náhodná veličina. Pak střední zbytkovou dobou života budeme rozumět*

$$r(t) = E(T - t \,|\, T \geq t).$$

Všimněme si, že platí

$$r(t) = \frac{\int_t^\infty (s-t) f(s) ds}{S(t)}.$$

Zintegrujeme-li funkci $\int_t^\infty (s-t) f(s) ds$ per partes, získáme $[(s-t)(-S(s))]_t^\infty + \int_t^\infty S(s) ds$. Protože $\lim_{s\to\infty} S(s) = 0$, obdržíme

$$r(t) = \frac{\int_t^\infty S(s) ds}{S(t)}. \tag{1.3}$$

Dosadíme-li $t = 0$ do (1.3), vidíme, že

$$E(T) = r(0) = \int_0^\infty S(s) ds,$$

protože $r(0) = E(T \,|\, T \geq 0)$ a $S(0) = 1$.

1.2.2 Diskrétní náhodná veličina T

Je-li $T \geq 0$ diskrétní náhodná veličina, pak nabývá hodnot $0 \leq a_1 < a_2 < \dots$ s příslušnou pravděpodobnostní funkcí

$$f(a_i) = P(T = a_i), \qquad i = 1, 2, \dots,$$

kde funkce přežití má tvar

$$S(t) = \sum_{i | a_i > t} f(a_i), \qquad t \geq 0.$$

Definice 7. *Riziková funkce diskrétní náhodné veličiny $T \geq 0$ v čase a_i je definována jako podmíněná pravděpodobnost selhání v čase a_i za podmínky, že selhání nenastalo před časem a_i, tedy*

$$\lambda_i = P(T = a_i \,|\, T \geq a_i) = \frac{f(a_i)}{S(a_i^-)}, \qquad i = 1, 2, \dots,$$

kde $S(a^-) = \lim_{t\to a^-} S(t)$.

Obdobně jako v (1.1) a v (1.2) můžeme vyjádřit funkci přežití a pravděpodobnost pomocí rizikové funkce (viz Kalbfleisch a Prentice, 2002)

$$S(t) = \prod_{i | a_i \leq t} (1 - \lambda_i)$$

a

$$f(a_i) = \lambda_i \prod_{j=1}^{i-1} (1 - \lambda_j).$$

Stejně jako ve spojitém případě platí, že riziková funkce $(\lambda_i, i = 1,2,\dots)$ jednoznačně určuje rozdělení diskrétní náhodné veličiny T.

1.2.3 T s kombinovaným rozdělením

Obecně může být rozdělení náhodné veličiny $T \geq 0$ zároveň jak diskrétní, tak spojité. V takovém případě má riziková funkce spojitou komponentu $\lambda_c(t)$ a diskretní komponenty $\lambda_1, \lambda_2, \ldots$ v diskrétních časech $0 \leq a_1 < a_2 < \ldots$. Kalbfleisch a Prentice (2002) uvádí tvar celkové funkce přežití

$$S(t) = exp[-\int_0^t \lambda_c(u)du] \prod_{j|a_j \leq t} (1 - \lambda_j)$$

a tvar kumulativní rizikové funkce

$$\Lambda(t) = \int_0^t \lambda_c(u)du + \sum_{j|a_j \leq t} \lambda_j.$$

Uvedené formule jsou pouze aproximace těchto funkcí.

Kumulativní riziková je zprava spojitá, neklesající funkce. Pomocí $\Lambda(t)$ definujeme přírůstek

$$\begin{aligned}\Delta\Lambda(t) &= \Lambda(t^- + \Delta t) - \Lambda(t^-) \\ &= P[T \in [t, t + \Delta t) \,|\, T \geq t] \\ &= \begin{cases} \lambda_i, & t = a_i, \quad i = 1, 2, \ldots, \\ \lambda_c(t)dt, & \text{jinak.} \end{cases}\end{aligned}$$

Přírustek $\Delta\Lambda(t)$ specifikuje riziko selhání na nekonečně malém intervalu $[t, t+\Delta t)$.

Funkci přežití pro náhodnou veličinu s kombinovaným diskrétním i spojitím rozdělením můžeme psát jako

$$S(t) = \mathscr{P}_0^t[1 - \Delta\Lambda(u)].$$

$\mathscr{P}$ je součinový integrál definovaný způsobem

$$\mathscr{P}_0^t[1 - \Delta\Lambda(u)] = lim \prod_{k=1}^{r} \{1 - [\Lambda(u_k) - \Lambda(u_{k-1})]\},$$

kde $0 = u_0 < u_1 < \ldots < u_r = t$ a limita je pro $r \to \infty$ a $max(u_i - u_{i-1}) \to 0$.

Pro spojitý případ ($\lambda_i = 0$ pro všechna i) máme

$$S(t) = \mathscr{P}_0^t[1 - \Delta\Lambda(u)] = \mathscr{P}_0^t[1 - \lambda_c(u)du] = exp[-\int_0^t \lambda_c(u)du].$$

Pro diskrétní případ ($\lambda_c(t) = 0$ pro všechna t) máme

$$S(t) = \mathscr{P}_0^t[1 - \Delta\Lambda(u)] = \prod_{j|a_j \leq t} (1 - \lambda_i).$$

Reprezentaci funkce přežití pomocí součinu si můžeme představit na pokusu s hody mincí, ve kterém se pravděpodobnost rubu liší v průběhu času. Mince je

házena opakovaně a selhání je v tomto případě první padlý líc. Tudíž je pravděpodobnost přežití v čase t získána jako součin podmíněných pravděpodobností přežití $1 - \Delta\Lambda(u)$ na nekonečně malých intervalech až do času t.

Funkce $f(t)$, $S(t)$ a $F(t)$ jsou používány k obecnému popisu rozdělení náhodné veličiny. Riziková funkce $\lambda(t)$ je více specifická charakterizace užitečná při modelování dat z analýzy přežití.

1.3 Počáteční čas, cenzorování a krácení

1.3.1 Počáteční čas

Na začátku je důležité správně definovat, jaký čas je pro nás nulou. U klinických studií někdy můžeme čas měřit jako věk subjektů, pak je počáteční čas jejich datum narození, jindy může jít o čas výskytu nějaké události, například nástup do studie či diagnostikování nemoci.

1.3.2 Cenzorování

Data často obsahují jedince, u nichž nejsme schopni určit, kdy požadovaná událost nastala. Víme pouze, že nastala v nějakém intervalu. Tomu se říká cenzorování. Rozlišujeme cenzorování zleva, zprava a cenzorování intervalové.

Vysvětlíme si pravé a levé cenzorování na jednoduchých příkladech.

Cenzorování zprava

Můžeme rozlišit dva důvody cenzorování zprava. Prvním z nich je uplynutí doby pozorování, to znamená, pro t z nějakého intervalu $[0, M)$, kde M je čas ukončení našeho pozorování, nenastalo selhání. Potom je veličina $T \geq 0$ cenzorovaná a čas cenzorování je roven M (cenzorování typu I nebo také cenzorování časem). Druhým důvodem k cenzorování zprava je výskyt nějaké jiné události, kvůli které nejsme schopni subjekt dále pozorovat. Pak časem cenzorování je U, kde U je čas výskytu této jiné události (cenzorování typu II nebo také cenzorování poruchou).

Představme si lékařskou studii trvající jeden rok, ve které je nemocným pacientům podáván lék a „čas selhání“ je zde čas, kdy se pacient vylečí. Potom máme tři možnosti: pacient se během roku vyléčí, pacient je nemocný i ve chvíli ukončení studie po celém roce a pacient v nějaké části roku přestane být sledován (odstěhuje se, přestane docházet,...). U druhé a třetí možnost se jedná právě o cenzorování zprava, kdy nevíme, zda se pacient vyléčil dvacet minut po ukončení sledování, dvacet let nebo třeba nikdy.

Cenzorování zleva

Cenzorování zleva je analogické cenzorování zprava. Změna je v tom, že událost mohla nastat předtím, než jsme začali vůbec výskyt události sledovat. Tedy v tom, že naše sledování trvá na intervalu (M, ∞) a selhání nastalo pro nějaké

$t \in [0, M]$.

Příkladem je následující pokus: vědci pozorovali skupinu paviánů, která spí každou noc na stromě a sledovanou událostí byl moment, kdy více jak polovina paviánů po probuzení slezla dolů. Vědci nicméně začínali se sledováním každý den v určitý čas M. Někdy k „selhání“ došlo až poté a přesný čas byl známý, někdy ovšem byla větší část skupiny dole dřív než sledování začalo a vědci věděli pouze, že k „selhání“ došlo někdy před časem M.

Značení cenzorování

Na čas pozorování se opět můžeme dívat jako na nezápornou náhodnou veličinu, označme ji C. Rozdělení C lze také popsat distribuční funkcí a hustotou. Můžeme říci, že náhodná veličina C měří dobu pozorování.

Pokud je doba pozorování kratší než doba do selhání, pak není čas selhání napozorován. Pokud máme $T \geq 0$ a $C \geq 0$, můžou nastat tyto dva případy:

- $T \leq C$, T je napozorováno a C není,
- $T > C$, C je napozorováno a T není.

Předpokládáme, že veličiny T a C jsou nezávislé.

Definice 8. *Nechť T je nezáporná náhodná veličina popisující čas selhání a C je nezáporná náhodná veličina popisující dobu pozorování. Pak náhodnou veličinu $X = \min(T, C)$, $X \geq 0$, nazveme cenzorovaný čas selhání a $\delta = 1(T \leq C)$ nazveme indikátor selhání.*

Nezávislé cenzorování

Definice 9. *Řekneme, že cenzorování je nezávislé, pokud je v každém čase t riziková funkce stejná, jako kdyby k žádnému cenzorování nedocházelo. Tedy*

$$\lim_{h \to 0} \frac{P\{T \in [t, t+h) \,|\, T \geq t\}}{h} = \lim_{h \to 0} \frac{P\{T \in [t, t+h) \,|\, T \geq t, C \geq t\}}{h}, \qquad (1.4)$$

kde T je nezáporná náhodná veličina popisující čas selhání a C je nezáporná náhodná veličina popoisující dobu pozorování.

Pokud tedy máme n subjektů, T_i čas selhání, C_i čas cenzorování i-tého subjektu a předpokládáme, že T_i a C_i jsou nezávislé náhodné veličiny, tedy (T_i, C_i), $i = 1, 2, \ldots, n$ jsou nezávislé, pak je (1.4) ekvivalentní

$$\lim_{h \to 0} \frac{P\{T_i \in [t, t+h) \,|\, T_i \geq t\}}{h} = \lim_{h \to 0} \frac{P\{T_i \in [t, t+h) \,|\, T_i \geq t, C_i \geq t\}}{h}.$$

1.3.3 Krácení

Dalším jevem, který může nastat, je krácení. O krácení se jedná, pokud jsou do studie zahrnuty pouze takové subjekty, u nichž událost nastala v nějakém intervalu (K_D, K_H). Pokud $K_H = \infty$, říkáme, že se jedná o krácení zleva a pokud

$K_D = 0$, říkáme, že jde o krácení zprava. Jinými slovy, krácení zleva nastává, pokud jsou subjekty pozorovány pouze v případě, že k události došlo až po nějakém daném čase t. Obdobně pro krácení zprava, kde požadujeme výskyt události před daným časem t.

Typičtější je krácení zleva. Příkladem je studie sledující pacienty trpící cukrovkou, kteří museli brát pravidelně inzulín. Podle počtu napsaných předpisů bylo na začátku studie 1. července 1973 n takových pacientů. Tito pacienti byli sledováni až do 1. ledna 1982 za účelem odhadnout úmrtnost pacientů trpící cukrovkou. Předpokládejme, že chceme funkci úmrtnosti sledovat jako funkci trvání nemoci. Protože pacienti byli zahrnuti pouze tehdy, pokud byli 1. července 1973 naživu, příslušná distribuční funkce pro časy přežití $X_i, i = 1, 2, \ldots, n$ je podmíněná distribuční funkce X s $X > V$, kde V je doba trvání nemoci k 1.červenci 1973. Taková data jsou pak krácená zleva.

1.3.4 Příklad

Ukážeme si jednoduchý příklad týkající se cenzorování zprava. Představme si, že máme $n = 6$ nemocných pacientů, kterým podáme lék, a po dobu deseti dní sledujeme, za jak dlouho lék začne účinkovat. Tři nemocné pacienty, označme je A, B, C, jsme měli již na začátku studie. Pacient A se uzdravil po 8 dnech, pacient B třetí den ze studie odejde a pacient C se během deseti dní nevyléčí.

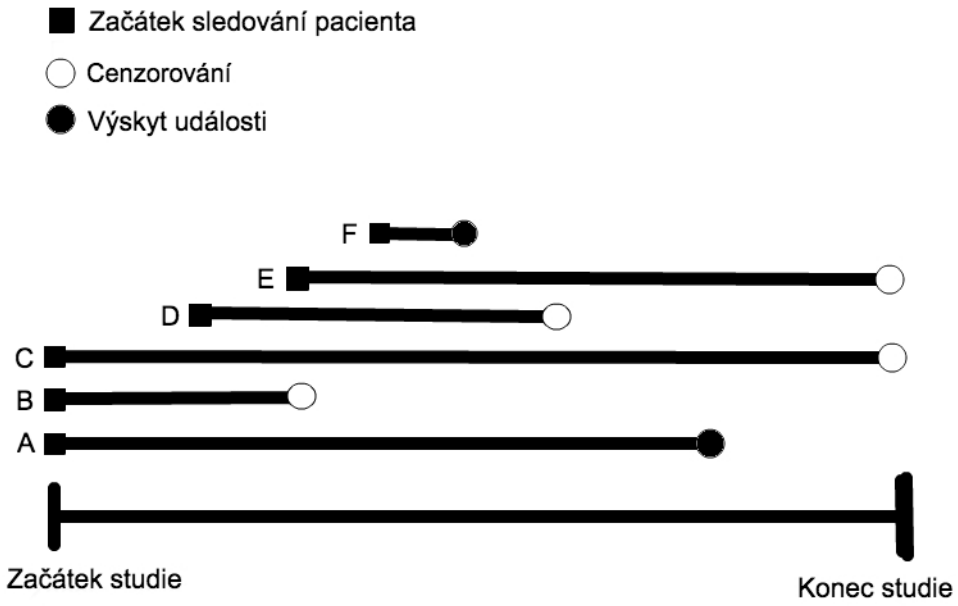
Druhý den přidáme pacienta D, ten však šestý den také ze studie odejde. Třetí a pátý den získáme dva nové pacienty, pacienta E a pacienta F. Pacient E se během celé studie nevyléčí, kdežto pacient F se vyléčí hned den po podání léku, tedy patý den naší studie.

Pacient	**A**	**B**	**C**	**D**	**E**	**F**
T	8	> 3	> 10	> 5	> 6	2
C	8	3	10	5	6	2
X	8	3	10	5	6	2
δ	1	0	0	0	0	1

Tabulka 1.1: Čas selhání (T), čas cenzorování (C), cenzorovaný čas selhání (X) a indikátor selhání (δ) pro data napozorovaná během studie.

Z tabulky 1.1 můžeme vyčíst informace o nasbíraných datech. Počáteční čas v tomto příkladu je podání léku, sledovaná událost je vyléčení pacienta. Máme dva výskyty události, dvě cenzorování časem a dvě cenzorování poruchou. Z dat jsme schopni určit všechny potřebné hodnoty: čas selhání (měřený ve dnech) známe pouze u pacientů A (8) a F (2), pro všechny známe dobu pozorování a minimem těchto dvou hodnot získáme cenzorovaný čas selhání. Rovněž si indikátorem označíme, zda došlo k události (A, F) nebo k cenzorování (B, C, D, E).

Pro lepší představu reprezentace dat se můžeme podívat na obrázek 1.1.



Obrázek 1.1: Grafické znázornění nasbíraných dat.

2. Odhadování

Máme tedy nasbíraná data obsahující data cenzorovaná i necenzorovaná a snažíme se odhadnout, z jakého rozdělení pochází. Nabízí se jednoduché řešení, a to pracovat pouze s těmi daty, která cenzorovaná nejsou. Potom máme náhodný výběr s nějakou distribuční funkcí, kterou můžeme odhadnout pomocí empirické distribuční funkce.

Nicméně pokud chceme zahrnout i data cenzorovaná, musíme s nimi pracovat jinak a používat jiné odhady. Konkrétně se podíváme na Kaplan-Meierův odhad, na Nelson-Aalenův odhad a na nějaké základní vlastnosti těchto odhadů.

2.1 Empirická distribuční funkce

Definice 10. *Nechť $X_1, \ldots, X_n$ je náhodný výběr z rozdělení s distribuční funkcí F_X. Potom*

$$\hat{F}_n(x) = \frac{1}{n} \sum_{i=1}^{n} \mathbb{1}_{(-\infty, x]}(X_i),$$

se nazývá empirická distribuční funkce.

Empirická distribuční funkce je jednoduchým odhadem distribuční funkce $F_X(x) = P(X \leq x)$. Její graf vizuálně reprezentuje napozorovaná data a umožňuje pozorovat obecné vlastnosti rozdělení, ze kterého napozorovaná data pochází. Empirická distribuční funkce v bodě x je poměr počtu naměřených hodnot menších nebo rovných x a všech naměřených hodnot. Pro pevné x se na $\hat{F}_n(x)$ můžeme dívat jako na relativní četnost jevu $[X_i \leq x]$ počítanou z n pozorování. Pravděpodobnost jevu $[X_i \leq x]$ je $P([X_i \leq x]) = F_X(x)$.

Věta 2 (Vlastnosti empirické distribuční funkce)**.** *Pro $x \in \mathbb{R}$ platí*

(i) $E\hat{F}_n(x) = F_X(x)$ *(nestrannost),* $var\hat{F}_n(x) = \frac{F_X(x)(1-F_X(x))}{n}$*;*

(ii) $\hat{F}_n(x) \xrightarrow{P} F_X(x)$ *pro* $n \to \infty$ *(konzistence);*

(iii) $\sqrt{n}(\hat{F}_n(x) - F_X(x)) \xrightarrow{D} N(0, F_X(x)(1 - F_X(x)))$*;*

(iv) $n\hat{F}_n(x) \sim Bi(n, F_X(x))$.

Důkaz. Protože se na empirickou distribuční funkci v každém bodě x můžeme dívat jako na relativní četnost, k důkazu výše uvedených vlastností nám stačí vlastnosti relativní četnosti a to, že pro $\overline{X}_n = \frac{1}{n}\sum_{i=1}^{n} X_i, X_i \sim Alt(p)$ platí

(i) $E\overline{X}_n = p$, $var\overline{X}_n = \frac{p(1-p)}{n}$;

(ii) $\overline{X}_n \xrightarrow{P} p$;

(iii) $\sqrt{n}(\overline{X}_n - p) \xrightarrow{D} N(0, p(1-p))$;

(iv) $n\overline{X}_n \sim Bi(n, p)$.

□

V analýze přežití se častěji než distribuční funkce odhaduje funkce přežití. Odhad funkce přežití získáme z empirické distribuční funkce snadno. Protože $S(x) = 1 - F(x)$, máme empirickou funkci přežití

$$\hat{S}_n(x) = 1 - \hat{F}_n(x) = 1 - \frac{1}{n}\sum_{i=1}^{n} \mathbb{1}_{(-\infty, x]}(X_i). \tag{2.1}$$

Empirická funkce přežití je nerostoucí schodovitá funkce se skoky o velikosti $\frac{1}{n}$ v každém napozorovaném čase selhání.

2.2 Kaplan-Meierův odhad

2.2.1 Odvození

Mějme tedy nezápornou náhodnou veličiny T popisující čas selhání. Dále mějme pozorování $0 \leq t_1 < t_2 < \ldots < t_k$ reprezentující časy selhání napozorované na homogenní populaci o velikosti $n = n_0$ s nějakou neznámou funkcí přežití S.

Napozorováné časy selhání definují $k+1$ intervalů, konkrétně intervaly $[0, t_1)$, $[t_1, t_2), \ldots, [t_k, \infty)$. Předpokládejme, že d_j selhání nastane v čase t_j a m_j pozorování je cenzorovaných v intervalu $[t_j, t_{j+1})$, $j = 0, 1, \ldots, k$, kde $t_0 = 0$ a $t_{k+1} = \infty$. Označme časy cenzorování v intervalu $[t_j, t_{j+1})$ jako $t_{j1}, t_{j2}, \ldots, t_{jm_j}$, $j = 0, 1, \ldots, k$. Dále položme $n_j = (m_j + d_j) + \cdots + (m_k + d_k)$. n_j značí počet subjektů těsně před časem t_j, u kterých ještě selhání nenastalo, neboli počet těch subjektů, kteří jsou před časem t_j stále „v riziku“.

Dle Kalbfleische a Prentice (2002) pravděpodobnost selhání v čase t_j můžeme psát

$$P(T = t_j) = S(t_j^-) - S(t_j)$$

a přírůstek cenzorovaného času přežití v čase t_{jl} k pravděpodobnosti je

$$P(T > t_{jl}) = S(t_{jl}).$$

Za předpokladu nezávislého cenzorování nám napozorovaný cenzorovaný čas t_{jl} říká pouze to, že nenapozorovaný čas selhání je větší než t_{jl}.

Pravděpodobnost dat $P(t_1 < t_2 < \ldots < t_k)$ je tvaru

$$L = \prod_{j=0}^{k} \{[S(t_j^-) - S(t_j)]^{d_j} \prod_{l=1}^{m_j} S(t_{jl})\}. \tag{2.2}$$

Funkci L můžeme brát jako funkci na prostoru všech funkcí přežití. Maximálně věrohodný odhad funkce přežití je funkce $\hat{S}$, která maximalizuje L. Zřejmě $\hat{S}(t)$ je nespojitá v napozorovaných bodech výskytu selhání t_j. V těchto bodech je L nenulová, v ostatních bodech je $L = 0$. Protože platí $t_{jl} \geq t_j$, je $S(t_{jl})$ maximalizováno pro $S(t_{jl}) = S(t_j)$, $j = 1, \ldots, k$, $l = 1, \ldots, m_j$. Hledaný maximálně

věrohodný odhad $\hat{S}(t)$ je tedy diskrétní funkce přežití s rizikovými komponenty $\hat{\lambda}_1, \ldots, \hat{\lambda}_k$ v bodech $t_1, \ldots, t_k$. Proto

$$\hat{S}(t_j) = \prod_{l=1}^{j}(1 - \hat{\lambda}_l)$$

a

$$\hat{S}(t_j^-) = \prod_{l=1}^{j-1}(1 - \hat{\lambda}_l),$$

kde $\hat{\lambda}_l$, $l = 1, \ldots, j$ jsou voleny tak, aby maximalizovaly funkci

$$L(\theta) = \prod_{j=1}^{k}[\lambda_j^{d_j} \prod_{l=1}^{j-1}(1-\lambda_l)^{d_j} \prod_{l=1}^{j}(1-\lambda_l)^{m_j}] = \prod_{j=1}^{k} \lambda_j^{d_j}(1-\lambda_j)^{n_j - d_j}, \;\; \theta = (\lambda_1, \ldots, \lambda_k),$$

kterou získáme dosazením $\hat{S}(t_j) = \prod_{l=1}^{j}(1-\hat{\lambda}_l)$ a $\hat{S}(t_j^-) = \prod_{l=1}^{j-1}(1-\hat{\lambda}_l)$ do vzorce (2.2). Logaritmická věrohodnostní funkce má potom tvar

$$\log L(\theta) = l(\theta) = \sum_{j=1}^{k} d_j \log \lambda_j + \sum_{j=1}^{k}(n_j - d_j)\log(1 - \lambda_j).$$

Podíváme se na i-tou parciální derivaci $l(\theta)$, kterou položíme rovnu 0.

$$\frac{\partial l(\theta)}{\partial \lambda_i} = \frac{d_i}{\lambda_i} + \frac{(-1)(n_i - d_i)}{1 - \lambda_i} = \frac{d_i - n_i\lambda_i}{\lambda_i(1 - \lambda_i)} = 0.$$

Platí, že $(d_i - n_i\lambda_i)/(\lambda_i(1 - \lambda_i))$ je rovna 0 právě tehdy, když $\hat{\lambda}_i = d_i/n_i$ pro $i = 1, \ldots, k$.
Dosazením do $\hat{S}(t_j) = \prod_{l=1}^{j}(1-\hat{\lambda}_l)$ získáme Kaplan-Meierův odhad funkce přežití.

Definice 11. *Definujeme Kaplan-Meierův odhad funkce přežití jako*

$$\hat{S}(t) = \prod_{j|t_j \le t} \frac{n_j - d_j}{n_j}, \tag{2.3}$$

kde d_j je počet selhání, která nastala v čase t_j, m_j je počet pozorování cenzorovaných v intervalu $[t_j, t_{j+1})$, $j = 0, 1, \ldots, k$, $t_0 = 0$, $t_{k+1} = \infty$ a $n_j = (d_j + m_j) + \cdots + (d_k + m_k)$.

V Kaplan-Meierově odhadu odhadnuté riziko nebo podmíněná pravděpodobnost selhání v čase t_j přesně koresponduje s napozorovaným poměrem (d_j/n_j) pro n_j subjektů, které jsou v čase t_j stále v riziku. Opět se na funkci přežití díváme sekvenčně a v každém čase selhání odhadujeme riziko selhání jako napozorované množství selhání. Všimněme si, že $\hat{S}(t)$ je vždy nenulová, pokud $m_k > 0$. V takovém případě je největší napozorovaná hodnota cenzorovaná a obvykle bereme $\hat{S}(t)$ jako nedefinovanou pro $t > t_{km_k}$.

2.2.2 Vlastnosti

Věta 3 (Asymptotické rozdělení Kaplan-Meierova odhadu)**.** *Nechť distribuční funkce F času selhání T a distribuční funkce G času pozorování C jsou spojité. Nechť $t > 0$ je takové, že $S(t) = 1 - F(t) > 0$ a nechť I je indikátor selhání. Potom*

$$\sqrt{n}(\hat{S}(t) - S(t)) \xrightarrow{D} N(0, S^2(t) \int_0^t [(1 - F(x))(1 - G(x))]^{-2} dP(X < x, I = 1)).$$

Tuto větu (viz Hurt 1984) necháme bez důkazu.

V praxi je třeba nahradit rozptyl užitý ve větě nějakým odhadem.

Věta 4 (Greenwoodova formule)**.** *Rozptyl Kaplan-Meierova odhadu funkce přežití $\hat{S}(t)$ je*

$$\widehat{V}_S(t) = \widehat{var}[\hat{S}(t)] = \hat{S}^2(t) \sum_{j|t_j \le t} \frac{d_j}{n_j(n_j - d_j)}. \tag{2.4}$$

Důkaz. Uvažujme asymptotické rozdělení $\hat{S}(t)$ v nějakém daném t. Rozptyl odhadu $\hat{S}(t)$ můžeme získat, budeme-li se dívat na vzorec

$$L(\lambda_1, \ldots, \lambda_k) = \prod_{j=1}^{k} [\lambda_j^{d_j} \prod_{l=1}^{j-1} (1 - \lambda_l)^{d_j} \prod_{l=1}^{j} (1 - \lambda_l)^{m_j}] = \prod_{j=1}^{k} \lambda_j^{d_j} (1 - \lambda_j)^{n_j - d_j}$$

jako na věrohodnostní funkci parametrů $\lambda_1, \ldots, \lambda_k$.

Heuristickým odvozením získáme rozptyl odhadu λ_j, $j = 1, \ldots, k$. Jelikož

$$\frac{\partial^2 \log L(\lambda_1, \ldots, \lambda_k)}{\partial \lambda_j^2} = \frac{-n_j \lambda_j^2 - d_j + 2 d_j \lambda_j}{\lambda_j^2 (1 - \lambda_j)^2},$$

po dosázení $\lambda_j = d_j / n_j$ obdržíme

$$K_j = \frac{n_j^3 (d_j - n_j)}{d_j (n_j - d_j)^2}.$$

Odhad rozptylu můžeme získat jako

$$\widehat{var}(\hat{\lambda_j}) = \frac{1}{-K_j} = \frac{d_j (n_j - d_j)}{n_j^3}.$$

Máme $\widehat{var}(\hat{\lambda_j}) = d_j(n_j - d_j)/n_j^3$ a protože

$$\hat{S}(t) = \prod_{j|t_j \le t} \frac{(n_j - d_j)}{n_j} = \prod_{j|t_j \le t} (1 - \frac{d_j}{n_j}) = \prod_{j|t_j \le t} (1 - \hat{\lambda}_j),$$

máme

$$log \hat{S}(t) = log \prod_{j|t_j \le t} (1 - \hat{\lambda}_j) = \sum_{j|t_j \le t} log(1 - \hat{\lambda}_j).$$

Potom pro asymptotický odhad rozptylu platí

$$\begin{aligned}\widehat{var}[log\hat{S}(t)] &= \widehat{var}[\sum_{j|t_j\le t} log(1-\hat{\lambda}_j)] = \sum_{j|t_j\le t} \widehat{var}[log(1-\hat{\lambda}_j)] \\ &= \sum_{j|t_j\le t} (1-\hat{\lambda}_j)^{-2}\widehat{var}(1-\hat{\lambda}_j) = \sum_{j|t_j\le t} (1-\frac{d_j}{n_j})^{-2}\frac{d_j(n_j-d_j)}{n_j^3} \\ &= \sum_{j|t_j\le t} \frac{d_j}{n_j(n_j-d_j)}.\end{aligned}$$

kde jsme v druhé rovnosti použili nezávislost a ve třetí rovnosti transformaci $g(x) = log(x)$.
Pokud nyní použijeme transformaci $h(x) = e^x, dh/dx(x) = e^x$ a protože $e^{\log \hat{F}(t)} = \hat{F}(t)$, získáme asymptotický rozptyl

$$\widehat{V}_S(t) = \widehat{var}[\hat{S}(t)] = \hat{S}^2(t) \sum_{j|t_j\le t} \frac{d_j}{n_j(n_j-d_j)}.$$

□

2.2.3 Konfidenční intervaly

Při odhadování distribuční funkce je často lepší a praktičtější používat konfidenční intervaly. Stejně tak se používají konfidenční intervaly při odhadu funkce přežití založené na Kaplan-Meierově odhadu.

Využijeme toho, že Kaplan-Meierův odhad má asymptoticky normální rozdělení. Potom 95% konfidenční interval pro odhad funkce přežití $S(t)$ je

$$(\hat{S}(t) - 1{,}96[\widehat{var}\hat{S}(t)]^{\frac{1}{2}}, \hat{S}(t) + 1{,}96[\widehat{var}\hat{S}(t)]^{\frac{1}{2}}).$$

Takovýto interval nicméně může obsahovat nesmyslné hodnoty, tedy hodnoty mimo interval [0,1]. Tomu se můžeme vyhnout, použijeme-li nějakou transformaci. Zkusme se tedy podívat na rozptyl veličiny

$$\hat{H}(t) = \log[-\log \hat{S}(t)].$$

Transformace $g(x) = \log(-\log x)$ zobrazuje interval $(0, 1)$ na interval $(-\infty, \infty)$. Protože známe rozptyl $\widehat{var}[\log \hat{S}(t)] = \widehat{var}[-\log \hat{S}(t)]$, můžeme rozptyl $\hat{H}$ spočítat pomocí transformace $h(x) = \log x$, pro kterou $dh/dx(x) = 1/x$, a tedy $dh/dx(-\log(\hat{S}(t))) = -1/\log \hat{S}(t)$. Získáváme

$$\widehat{var}\hat{H}(t) = \frac{\widehat{var}\hat{S}(t)}{[\log \hat{S}(t)]^2}.$$

Označme $v(t) = \widehat{var}\hat{H}(t)$, potom máme 95% konfidenční interval $\hat{H}(t) \pm 1{,}96v(t)$ pro $H(t) = \log[-\log S(t)]$. Aplikujeme-li $g^{-1}(x) = e^{-\exp x}$ na interval $\hat{H}(t) \pm 1{,}96v(t)$, získáme 95% konfidenční interval pro $S(t)$

$$([\hat{S}(t)]^{\exp[-1{,}96v(t)]}, [\hat{S}(t)]^{\exp[+1{,}96v(t)]}).$$

Při počítání konfidenčních intervalů se častěji používá právě tento postup.

2.3 Nelson-Aalenův odhad

Často se prvně místo funkce přežití uvažuje kumulativní riziková funkce $\Lambda(t)$. Protože $\Lambda(t) = \int_0^t \lambda_c(u)du + \sum_{j|t_j \leq t} \lambda_j$, můžeme snadno vytvořit odhad $\hat{\Lambda}(t) = \sum_{j|t_j \leq t} \hat{\lambda_j}$. Maximálně věrohodný odhad $\hat{\lambda_j} = d_i/n_i$ jsme získali v předchozí části.

Definice 12. *Definujme Neslon-Aalenův odhad kumulativní rizikové funkce jako*

$$\hat{\Lambda}(t) = \sum_{j|t_j \leq t} \frac{d_j}{n_j}, \tag{2.5}$$

kde d_j je počet selhání, která nastala v čase t_j, m_j je počet pozorování cenzorovaných v intervalu $[t_j, t_{j+1})$, $j = 0, 1, \ldots, k$, $t_0 = 0$, $t_{k+1} = \infty$ a $n_j = (d_j + m_j) + \cdots + (d_k + m_k)$.

Funkce $\hat{\Lambda}(t)$ je zprava spojitá schodovitá funkce se skoky o velikosti odhadů $\hat{\lambda}_j$.

Protože je $S(t) = \mathscr{P}_0^t[1 - \Delta\Lambda(u)]$, máme mezi Kaplan-Meierovým odhadem a Neslon-Aalenovým odhadem následující vztah:

$$\hat{S}(t) = \mathscr{P}_0^t[1 - \Delta\hat{\Lambda}(u)] = \prod_{j|t_j \leq t} (1 - \hat{\lambda}_j).$$

3. Numerické příklady

Skripty použité k následujcím příkladům stejně jako použitá data můžeme nalézt v příloze.

3.1 Příklad 1

Nyní se podíváme, jak tato teorie funguje na konkrétním numerickém příkladě. Podíváme se na reálný příklad, nicméně s daty uměle vygenerovanými. Pro taková data známe přesné rozdělení a můžeme tedy lépe porovnat, jak jednotlivé způsoby odhadování fungují.

Představme si, že pozorujeme životnost nějakých strojů. Životnost budeme měřit v hodinách. Strojů je pozorováno $n = 10000$ a to na různých místech a různými lidmi. Proto se může snadno stát, že se k nám nedostanou kompletní data, někdo nějaký stroj omylem vypne dřív nebo na něj zapomene, přestane ho sledovat či sdílet s námi informace. Zároveň chceme někdy s pokusem skončit, proto si sami nastavíme maximální dobu sledování a to $M = 200$ hodin. V našem příkladu se tedy objevuje cenzorování zprava a to obojího typu.

Data

Máme náhodný výběr $\mathbb{X} = (X_1, \delta_1), \ldots, (X_{10000}, \delta_{10000})$ z neznámého rozdělení. Předpokládáme, že toto rozdělení je spojité, přestože jsme schopni měřit jen v diskrétním čase, a to ještě s chybami vzniklými kvůli zaokrouhlování. Všechno je měřeno jen na jedno desetinné místo. Podíváme se na nějaké základní popisné statistiky.

Data	Minimum	Medián	Průměr	Maximum	n	m
$\mathbb{X}$	0,0	69,1	85,85	200,0	10000	1481

Tabulka 3.1: Základní popisné statistiky náhodného vektoru $\mathbb{X}$, n je počet složek vektoru a m je počet cenzorovaných složek vektoru.

Protože jsme si data vygenerovali sami, známe jejich přesné rozdělení. Vektor $\mathbb{X}$ je výběr z exponenciálního rozdělení $Exp(\frac{1}{100})$, tedy očekávaná doba životnosti stroje je 100 hodin. Distribuční funkce je

$$F(t) = 1 - e^{-\frac{1}{100}t}, \qquad t \geq 0,$$

proto funkce přežití má tvar

$$S(t) = e^{-\frac{1}{100}t}, \qquad t \geq 0.$$

Protože máme hustotu

$$f(t) = \frac{1}{100} e^{-\frac{1}{100}t}, \qquad t \geq 0,$$

lze snadno dopočítat riziková a kumultivní riziková funkce pro $t \geq 0$, tedy

$$\lambda(t) = \frac{\frac{1}{100}e^{-\frac{1}{100}t}}{e^{-\frac{1}{100}}} = \frac{1}{100},$$

$$\Lambda(t) = \int_0^t \frac{1}{100} ds = \frac{1}{100}t.$$

Snadno můžeme dopočítat i střední zbytkovou dobu života, která je v tomto případě konstantní. Pro $t \in [0, \infty)$ totiž máme

$$r(t) = \frac{\int_t^\infty e^{-\frac{1}{100}s} ds}{e^{-\frac{1}{100}t}} = 100.$$

3.1.1 Odhady

Empirická funkce přežití

Z vektoru $\mathbb{X}$ můžeme snadno vytvořit vektor $\mathbb{Y}$, který nebude obsahovat žádná cenzorovaná data. V tom případě se nám změní některé statistiky, jak můžeme vidět v následující tabulce 3.2.

Data	Minimum	Medián	Průměr	Maximum	n	m
$\mathbb{X}$	0,1	57,0	68,73	199,9	8514	0

Tabulka 3.2: Základní popisné statistiky náhodného vektoru $\mathbb{Y}$, n je počet složek vektoru a m je počet cenzorovaných složek vektoru.

Nicméně pro vektor $\mathbb{Y}$ můžeme velmi snadno spočítat empirickou distribuční funkci $\hat{F}$ a pomocí vzorce (2.1) získáme empirickou funkci přežití $\hat{S}$.

Podíváme se na obrázek 3.1, jak tento odhad vypadá.

Vidíme, že odhadujeme-li funkci přežití $S(t) = e^{-\frac{1}{100}t}$, ale nepoužíváme žádná cenzorovaná data, odhad pomocí empirické funkce se pro větší t více vzdaluje od reálné hodnoty.

Kaplan Meierův odhad funkce přežití

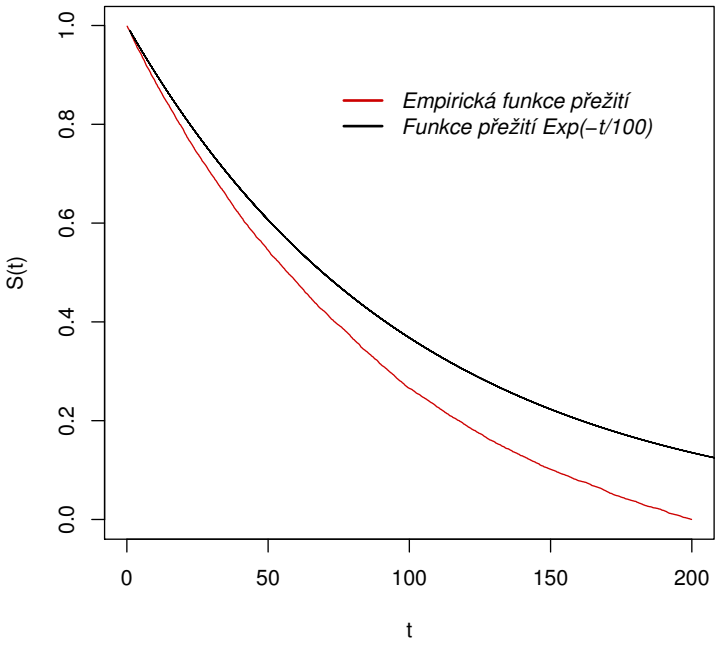
Zkusíme tedy funkci přežití odhadnout Kaplan-Meierovým odhadem, viz vzorec (2.3), který nám umožňuje zahrnout všechna naměřená data, a to i ta cenzorová. Podívejme se tedy na obrázek 3.2, jak takovýto odhad vypadá a jak konfidenční intervaly založené na Kaplan-Meierově odhadu pokrývají opravdovou funkci přežití.

Vidíme, že konfidenční intervaly pokrývají funkci přežití pro všechna $t \in$ [0,200], pro $t > 200$ není Kaplan-Meierův odhad funkce přežití definován, tedy nemáme ani konfidenční intervaly.

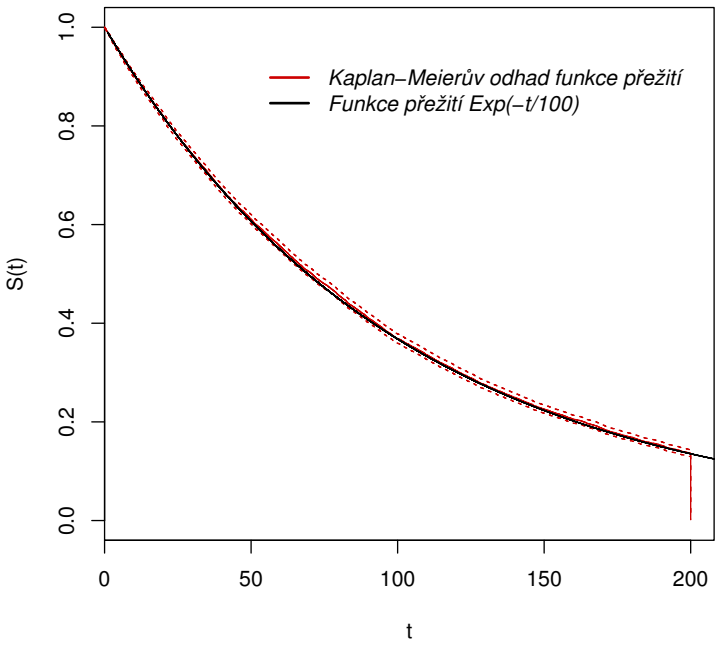
Nelson-Aalenův odhad

Nelsonův-Aalenův odhad neodhaduje funkci přežití, nýbrž kumulativní rizikovou funkci, v našem případě

$$\Lambda(t) = \int_0^t \frac{1}{100} ds = \frac{1}{100}t.$$

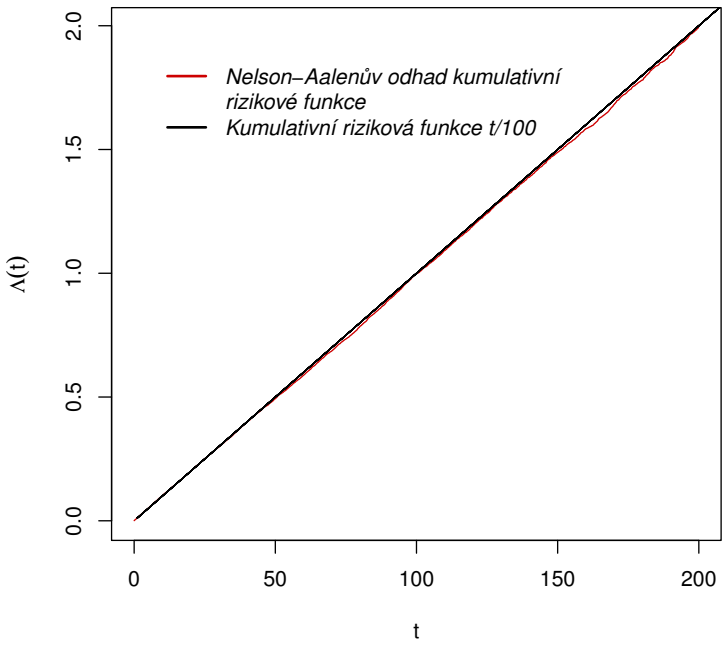


Obrázek 3.1: Odhad funkce přežití pomocí empirické funkce přežití.



Obrázek 3.2: Odhad funkce přežití pomocí Kaplan-Meierova odhadu a konfidenční interval.

Odhadovaná kumulativní riziková funkce je lineární a my ji budeme odhadovat pomocí vzorce (2.5). Na výsledný odhad se můžeme podívat na obrázku 3.3.



Obrázek 3.3: Odhad kumulativní rizikové funkce pomocí Nelson-Aalenova odhadu.

Z odhadu vidíme, že je kumulativní riziková funkce lineární, tedy riziková funkce bude konstantní. Protože riziková funkce je v tomto případě směrnicí kumulativní rizikové funkce, můžeme se ji pokusit odhadnout tak, že vezmeme průměr poměrů $\hat{\Lambda}(t)/t$.

Tento odhad nám dává

$$\hat{\lambda}(t) = 0{,}009942,$$

přičemž naše odhadovaná funkce je

$$\lambda(t) = \frac{1}{100} = 0{,}01.$$

Pokud by kumulativní riziková funkce nebyla lineární, nešlo by tak snadno odhadnout rizikovou funkci pomocí $\hat{\Lambda}$.

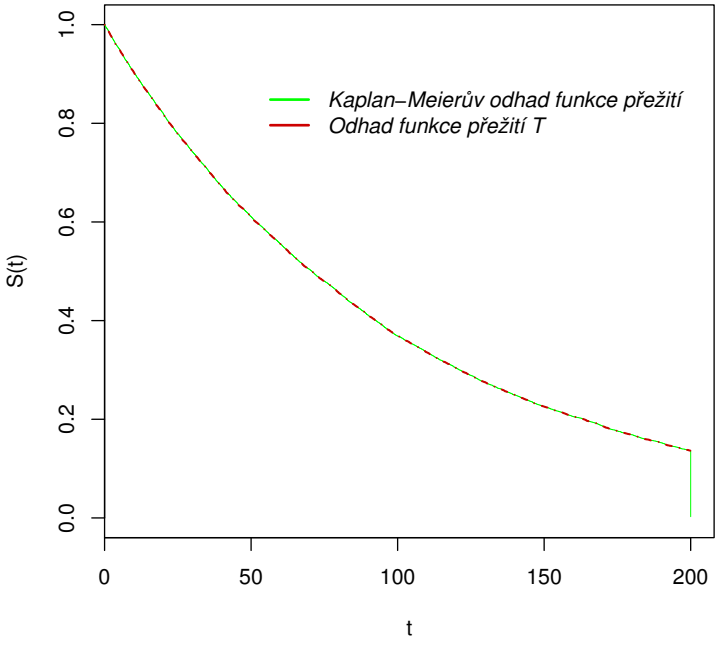
Jelikož z (1.1) víme, že $S(t) = \exp(-\Lambda(t))$, můžeme rovněž snadno odhadnout $S(t)$ následujícím způsobem

$$T(t) = \exp(-\hat{\Lambda}(t)).$$

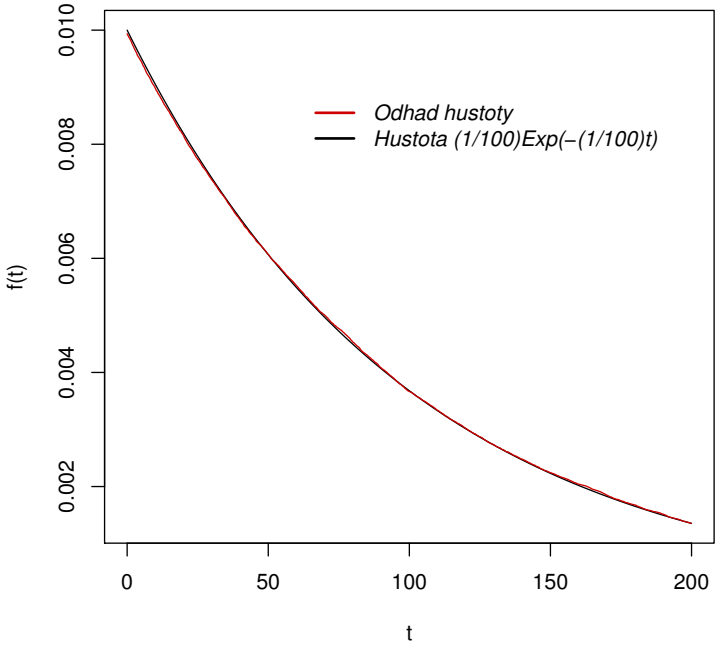
Tento odhad je identický Kaplan-Meierově odhadu, jak vidíme na obrázku 3.4

Dále víme $f(t) = \lambda(t)\exp[-\Lambda(t)]$, proto můžeme i velmi snadno odhadnout hustotu

$$f(t) = \frac{1}{100} e^{-\frac{1}{100}t}$$



Obrázek 3.4: Odhad funkce přežití pomocí Nelson-Aalenova odhadu a Kaplan-Meierův odhad funkce přežití.



Obrázek 3.5: Odhad hustoty.

pomocí

$$\hat{f}(t) = \hat{\lambda} e^{-\hat{\Lambda}(t)} = 0{,}009942 e^{\hat{\Lambda}(t)\cdot}$$

Na obrázku 3.5 vidíme, jak tento odhad vypadá pro náš příklad.

3.1.2 Porovnání odhadů

Je zřejmé, že odhad pomocí empirické distribuční funkce je vychýlený a není dobrým odhadem v analýze přežití. Kaplan-Meierův odhad a Nelson-Aalanův odhad nicméně fungovali na našem vzorku dat dobře, a protože jsme měli data z exponenciálního rozdělení, bylo snadné odvodit veškeré další funkce popisující rozdělení.

V praxi ovšem nebývají rozdělení takto jednoduchá a riziková funkce není obvykle konstantní. To ale nijak neovlivňuje kvalitu Kaplan-Meierova ani Nelson-Aalanova odhadu, tyto odhady se dají s dobrým výsledkem použít v jakémkoliv případě.

3.2 Příklad 2

Podívejme se nyní na příklad s daty z rozdělení, které je „složitější“ než exponenciální rozdělení, tedy takové, pro které riziková funkce není lineární.

Model

Volme

$$\lambda(t) = t^2, \qquad t \geq 0.$$

Z toho můžeme pro $t \geq 0$ odvodit kumulativní rizikovou funkci

$$\Lambda(t) = \int_0^t s^2 ds = \frac{1}{3}[s^3]_0^t = \frac{t^3}{3},$$

funkci přežití

$$S(t) = \exp(-\frac{t^3}{3})$$

a hustotu

$$f(t) = -\frac{dS(t)}{dt} = -\exp(-\frac{t^3}{3})(-t^2) = t^2 e^{-\frac{t^3}{3}}.$$

Následujícím způsobem ověříme, že se jedná o hustotu:

$$\int_0^\infty t^2 \exp(-\frac{t^3}{3}) dt = [-\exp(-\frac{t^3}{3})]_0^\infty = -(-1) = 1.$$

Budeme tedy generovat data z rozdělení s hustotou $t^2 \exp(-t^3/3)$.

Data

Opět máme náhodný výběr $\mathbb{X} = (X_1, \delta_1), \ldots, (X_{10000}, \delta_{10000})$ z neznámého rozdělení. Předpokládáme, že toto rozdělení je spojité, přestože jsme schopni měřit jen v diskrétním čase, a to ještě s chybami vzniklými kvůli zaokrouhlování. V tomto případě je všechno měřeno na tři desetinná místa.

I v tomto příkladě dochází k oběma typům cenzorování, maximální doba pozorování je nastavnena jako $M = 2$. Nyní se podíváme se na nějaké základní popisné statistiky.

Data	Minimum	Medián	Průměr	Maximum	n	m
$\mathbb{X}$	0,048	1,272	1,269	2,000	10000	917

Tabulka 3.3: Základní popisné statistiky náhodného vektoru $\mathbb{X}$, n je počet složek vektoru a m je počet cenzorovaných složek vektoru.

3.2.1 Odhady

Empirická funkce přežití

Z vektoru $\mathbb{X}$ snadno vytvoříme vektor $\mathbb{Y}$, který nebude obsahovat žádná cenzorovaná data. Některé statistky se opět změní, jak můžeme vidět v následující tabulce 3.4.

Data	Minimum	Medián	Průměr	Maximum	n	m
$\mathbb{X}$	0,048	1,229	1,214	199,9	9083	0

Tabulka 3.4: Základní popisné statistiky náhodného vektoru $\mathbb{Y}$, n je počet složek vektoru a m je počet cenzorovaných složek vektoru.

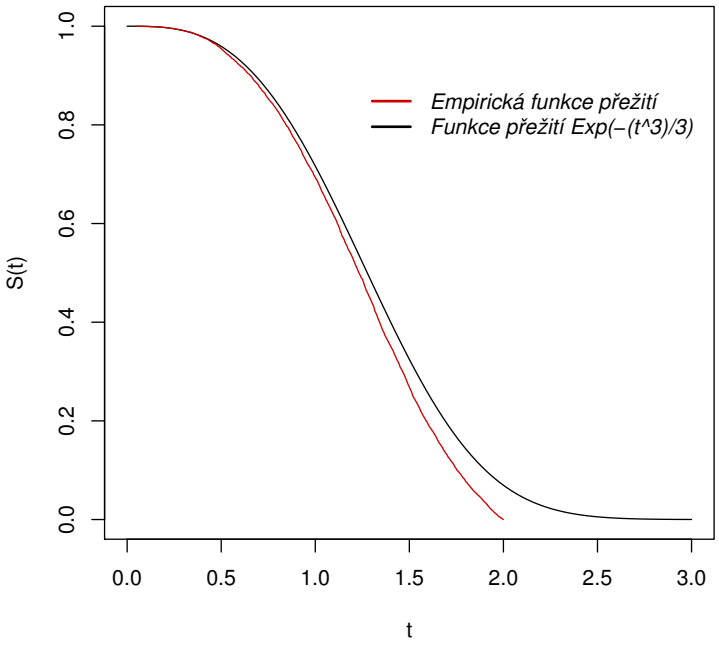
Pro vektor $\mathbb{Y}$ umíme spočítat empirickou distribuční funkci $\hat{F}$ (viz vzorec (2.1)) a získat empirickou funkci přežití $\hat{S}$. Na obrázku 3.6 můžeme vidět, jak tento odhad vypadá.

Vidíme, že odhadujeme-li funkci přežití $S(t) = e^{-\frac{1}{3}t^3}$, ale nepoužíváme žádná cenzorovaná data, odhad pomocí empirické funkce se rovněž pro větší t více vzdaluje od reálné hodnoty.

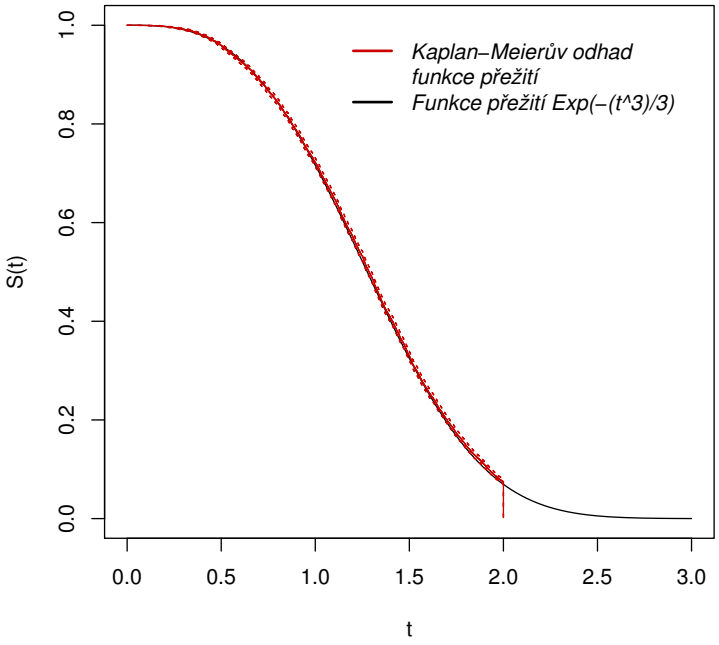
Kaplan-Meierův odhad

Nyní odhadneme funkci přežití Kaplan-Meierovým odhadem, viz vzorec (2.3), který nám umožňuje zahrnout všechna naměřená data, a to i ta cenzorová. Na obrázku 3.7 vidíme, jak takovýto odhad vypadá a jak konfidenční intervaly založené na Kaplan-Meierově odhadu pokrývají opravdovou funkci přežití.

Vidíme, že konfidenční intervaly pokrývají funkci přežití pro všechna $t \in [0,2]$, pro $t > 2$ není Kaplan-Meierův odhad funkce přežití definován, tedy nemáme ani konfidenční intervaly.



Obrázek 3.6: Odhad funkce přežití pomocí empirické funkce přežití.



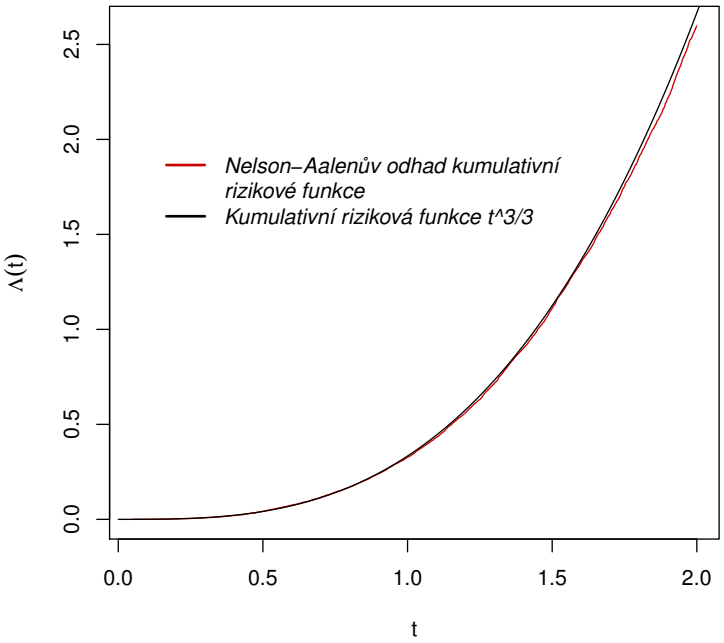
Obrázek 3.7: Odhad funkce přežití pomocí Kaplan-Meierova odhadu a konfidenční interval.

Nelson-Aalenův odhad

Poslední odhad je opět Nelsonův-Aalenův odhad, který neodhaduje funkci přežití, nýbrž kumulativní rizikovou funkci, v našem případě

$$\Lambda(t) = \frac{t^3}{3}.$$

Kumulativní rizikovou funkci budeme odhadovat pomocí vzorce (2.5). Výsledný odhad vidíme na obrázku 3.8.



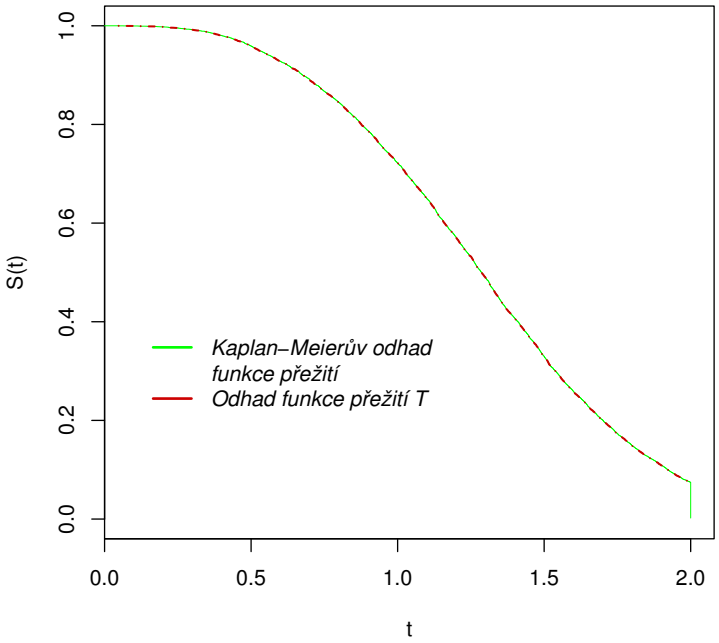
Obrázek 3.8: Odhad kumulativní rizikové funkce pomocí Nelson-Aalenova odhadu.

Jak víme z (1.1), platí $S(t) = \exp(-\Lambda(t))$. Použijeme-li Nelson-Aalenův odhad $\hat{\Lambda}(t)$, který jsme pozorovali před chvíli, můžeme snadno odhadnout $S(t)$. Následujícím odhad

$$T(t) = \exp(-\hat{\Lambda}(t))$$

je tedy odhad funkce přežití. Z obrázku 3.9 lze vidět, že tento odhad je opět identický Kaplan-Meierově odhadu.

Podíváme-li se zpět na obrázek 3.8, vidíme, že kumulativní riziková funkce není lineární. Není tedy možné udělat stejný „trik“ jako v předchozím příkladě a pokusit se odhadnout rizikovou funkci $\lambda(t)$ z odhadu kumulativní rizikové funkce. Nemůžeme tedy ani využít vzorce $\hat{f}(t) = \hat{\lambda}e^{-\hat{\Lambda}(t)}$ a pokusit se odhadnout hustotu pomocí Nelson-Aalenova odhadu.



Obrázek 3.9: Odhad funkce přežití pomocí Nelson-Aalenova odhadu a Kaplan-Meierův odhad funkce přežití.

3.2.2 Porovnání odhadů

I v druhém příkladu je jasně vidět, že vynechání cenzorovaných dat má za následek vychýlení odhadu, kdežto Kaplan-Meierův odhad odhaduje funkci přežití poměrně dobře, stejně jako Nelson-Aalenův odhad odhaduje dobře kumulativní rizikovou funkci.

Díky vyrobeným odhadům si můžeme udělat představu o rozdělení nasbíraných dat. Je sice pravda, že nad hranicí M (cenzorování časem) již nevíme, jak se funkce dále vyvýjí, nicméně pro $t \leq M$ odhady fungují správně.

Závěr

V první kapitole jsme se seznámili se základními pojmy potřebnými k popisu rozdělení veličiny. Toto rozdělení může být spojité, diskrétní i kombinované. Zadefinovali jsme si především funkci přežití $S(t)$, rizikovou funkci $\lambda(t)$ a kumulativní rizikovou funkci $\Lambda(t)$, které jsou v analýze přežití užívány více než obvyklá distribuční funkce a hustota. V druhé části první kapitoly jsme si vysvětlili základní principy cenzorování a potřebné pojmy spojené s cenzorováním.

V kapitole druhé se pak věnujeme jednotlivým odhadům, konkrétně odhadu pomocí empirické distribuční funkce, který s cenzorovanými daty nepracuje, dále Kaplan-Meierově odhadu a Nelson-Aalenově odhadu. Kaplan-Meierův odhad si odvodíme a rovněž si dokážeme Greenwoodovu formuli, která popisuje rozptyl Kaplan-Meierova odhadu. Díky tomu si dále můžeme odvodit konfidenční intervaly založené na tomto odhadu, které se v praxi užívají častěji. V závěru druhé kapitoly si ukážeme Nelson-Aalenův odhad a jeho vztah ke Kaplan-Meierově odhadu.

Kapitola třetí obsahuje dva příklady, a to s daty vygenerovanými z námi známého rozdělení. Odhady nicméně provadíme, jako kdybychom rozdělení neznali. Funkci přežití odhadujeme pomocí empirické funkce přežití, Kaplan-Meireova odhadu i pomocí Nelson-Aalenova odhadu. Rovněž odhadujeme kumulativní distribuční funkci, a to pomocí Nelson-Aalenova odhadu. V prvním příkladě jsou data z klasického exponenciálního rozdělení, v druhém pak z rozdělení, které má kvadratickou rizikovou funkci. V obou případech je jasně vidět, že odhad nepoužívající cenzorovaná data nefunguje dobře, a že odhady vhodné data obsahující data cenzorovaná odhadují reálné funkce poměrně přesně.

Seznam použité literatury

[1] Kalbfleisch, J.D.; Prentice, R.L.: *The Statistical Analysis of Failure Time Data.* Wiley, New York, 2002.

[2] Hurt, J.: *Teorie spolehlivosti.* Praha, SPN 1984. (Skripta.)

[3] Andersen, Per Kragh; Borgan, Ørnulf; Gill, Richard D.; Keiding, Niels: *Statistical Models Based on Counting Processes.* Springer-Verlag, Berlin 1993.

[4] Rich, Jason T. et al.: A practical guide to understanding Kaplan-Meier curves. Otolaryngol Head Neck Surg. 2010 September 143,3 (2010), 331–336.

[5] Kulich, M.: *Censored Data Analysis.* Praha, 2016. (Skripta.)