Analýza prežívania

Zlyhávanie a cenzúrovanie, funkcia vierohodnosti, základné charakteristiky a ich odhady, intervaly a pásy spoľahlivosti

Stanislav Katina[1]

[1]Ústav matematiky a statistiky
Přírodovědecká fakulta
Masarykova univerzita

ZS 2013

evropský sociální fond v ČR
EVROPSKÁ UNIE
MINISTERSTVO ŠKOLSTVÍ, MLÁDEŽE A TĚLOVÝCHOVY
OP Vzdělávání pro konkurenceschopnost
INVESTICE DO ROZVOJE VZDĚLÁVÁNÍ

Poďakovanie

Tento učební text vznikl za přispění Evropského sociálního fondu a státního rozpočtu ČR prostřednictvím Operačního programu Vzdělávání pro konkurenceschopnost v rámci projektu Univerzitní výuka matematiky v měnícím se světě

(CZ.1.07/2.2.00/15.0203)

Aké otázky v analýze prežívania riešime?

Príklady z praxe

- Odhadujeme a interpretujeme funkciu prežívania a riziko
- Porovnávame funkcie prežívania a riziká
- Modelujeme vzťah medzi vysvetľujúcimi premennými a časom prežívania

Otázky v analýze prežívania v aplikáciách

Príklady z praxe

Prežívanie pacientov po infarkte myokardu (IM) v rámci sekundárnej prevencie závažných kardiovaskulárnych problémov u pacientov s polymorfizmom glykoproteínu IV (GP VI 13254C/T) v membráne krvných doštičiek. [**Thrombosis Research** 125, 2: 61–4, 2009]

105 pacientov sledovaných v priemere 19($\pm$10.8) mesiacov

Otázky v analýze prežívania v aplikáciách
Príklady z praxe

Zlyhania: *smrť*, *ďalší IM*, *ďalšia selektívna koronarografia* (SKG: percutaneous coronary intervention (*PCI*, coronary angioplasty), coronary artery bypass graft (*CABG*)), *ďalša cievna mozgová príhoda* (CMP; stroke), *ďalšia hospitalizácia* (re-intervencia); sledované kombinácie: smrť/IM/re-intervencia a smrť/IM/re-intervencia/CMP [**MACE; Major Adverse Cardiac Events**, hlavné nepriaznivé srdcové udalosti]

Adjustujúce (rizikové) premenné:

1. *pohlavie* (žena=0, muž=1)
2. *hypertenzia* (nie=0, áno=1)
3. *hyperlipidémia* (nie=0, áno=1)
4. *fajčenie* (nefajčiar=0, fajčiar a bývalý fajčiar=1)
5. *diabetes* (nie=0, áno=1)
6. *srdcové zlyhanie* (**NYHA**; **New York Heart Association**; Classes: I = 0; II, III, IV = 1)

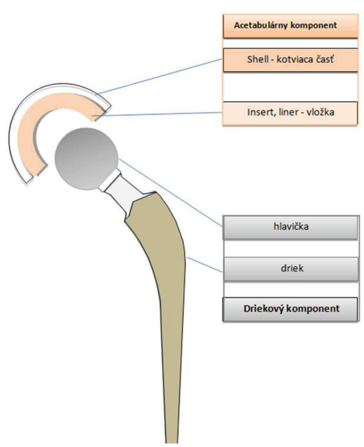
Otázky v analýze prežívania v aplikáciách
Príklady z praxe

Analýza prežívania implantátov bedra a kolena na Slovensku v rokoch 2003–2011. [**Acta Chir. Orthop. Traum. Čech.** 80: 1–85, 2013]

49 668 operácií (primárnych operácií a revízií) zo všetkých slovenských ortopedických kliník za roky 2003–2011:

- 38 485 THA (Total Hip Arthroplasty)
- 11 183 TKA (Total Knee Arthroplasty)

Otázky v analýze prežívania v aplikáciách
Príklady z praxe



Otázky v analýze prežívania v aplikáciách
Príklady z praxe

Zlyhania: zlyhanie komponentu implantátu

Adjustujúce (rizikové, prognostické) premenné:

1. *typ komponentu* (acetabulárny=0, femorálny=1)
2. *fixácia komponentu* (necementovaný=0, cementovaný=0)
3. *pohlavie* (žena=0, muž=1)
4. *cementovacia technika* (necementovaný=0, generácia cementu I = 1, generácia cementu II = 2, generácia cementu III = 3)
5. *diagnóza pri primárnej operácii* (primárna coxartróza = 1, dysplastická coxartróza = 2, poúrazová coxartróza = 3, aseptická nekróza hlavy = 4, M.Perthes = 5, reumatoidná artritída = 6, zlomenina krčku = 7)
6. *dôvod revízie* (spolu 18 dôvodov)
7. *revidované časti* (spolu 19 častí) a pod.

Prežívanie pacientov s chronickou myeloidnou leukémiou (CML). [**Neoplasma**, 92, 5: 381–7, 2005]

589 pacientov s CML, z ktorých 78 absolvovalo transplantáciu krvotvorných kmeňových buniek kostnej drene (allogeneic transplantation; *transplantácia od HLA-identického súrodenca alebo nepríbuzného darcu*; HLA znamená human leukocyte antigen) a zároveň majú odobrané vzorky periférnej krvi a kostnej drene pred a po transplantácii na Katedre genetiky Národného onkologického ústavu v Bratislave v rokoch 1990 až 2002

Zlyhania: úmrtie pacienta

Adjustujúce (prognostické, rizikové) premenné:

1. *vek* pacienta v čase transplantácie (skupina 1: $<$20 rokov, skupina 2: [20,40), skupina 3: $\geq$40)
2. *fáza CML* (spolu dve fázy; prvá chronická fáza = 1, ďalšie chronické fázy = 2)
3. *pohlavie darcu a príjemcu* (m–m, m–ž, ž–m, ž–ž)
4. *čas od diagnózy po transplantáciu* ($<$ 1 rok, $\geq$ 1 rok)

Example

Akútna myelogénna leukémia (acute myelogenous leukemia, **AML**). Po absolvovaní chemoterapie a zmiernení príznakov, boli pacienti náhodne rozdelení do dvoch skupín. Prvá skupina (skupina A) dostala udržujúcu chemoterapiu a druhá (kontrolná; skupina B) nie. Cieľom bolo zistiť, či udržujúca chemoterapia predlžuje čas do remisie (opätovného zhoršenia stavu).

Sk	čas po kompletnú remisiu (v týždňoch)	*n*	udalostí	cenzúr
A	$9, 13, 13+, 18, 23, 28+, 31, 34, 45+, 48, 161+$	11	7	4
B	$5, 5, 8, 8, 12, 16+, 23, 27, 30, 33, 43, 45$	12	11	1

(číslo = čas do zlyhania, číslo a plus $(+)$ = čas do cenzúry)

1. **problém 1**: po odstránení cenzúrovaných pozorovaní
2. **problém 2**: po ošetrení cenzúrovaných pozorovaní, ktoré zoberieme do úvahy akoby boli udalosťami (zlyhaniami)
3. **problém 3**: berúc do úvahy cenzúrované pozorovania

	problém 1		**problém 2**		**problém 3**	
	A	B	A	B	A	B
$\bar{x}$	25.1	21.7	38.5	21.3	52.6	22.7
$\tilde{x}$	23.0	23.0	28.0	19.5	31.0	23.0

(čísla sú v týždňoch)

Example

Cystická fibróza (CF) je autozomálna genetická choroba spôsobená mutáciou génu pre CFTR (cystic fibrosis transmembrane conductance regulator). Postihuje prevažne pľúca, ale aj pankreas, pečeň a črevo. V celoslovenskej databáze pacientov CF rozlišujeme pacientov s **jasnou klinickou formou** (**typická forma**, 259 živých, 112 zomrelých) a pacientov s **atypickou formu** (188 živých). Spolu teda 559 pacientov, 447 živých a 112 zomrelých. Aký je priemerný vek (prežívania) a medián (prežívania) v rokoch?

skupina/počty	typická forma CF	atypická forma CF	spolu
živí	259	188	447
zomrelí	112	0	112
spolu	371	188	559

Klasický prístup vs. analýza prežívania
Príklady

	problém 1	**problém 3**
	typická CF	typická CF
$\overline{x}$	9.22	45.05
$\widetilde{x}$	4.90	52.26

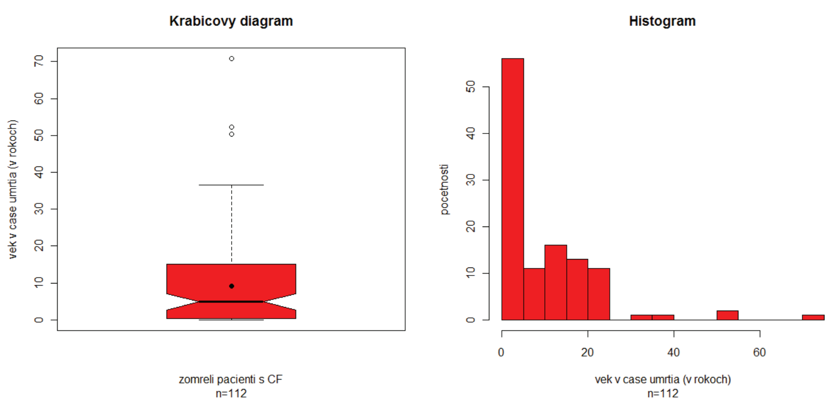
(čísla sú v rokoch)

- Rozdiel medzi priemerným vekom prežívania pacientov s typickou formou CF a priemerným vekom zomrelých je 35.83 roka (podobne pre medián je tento rozdiel 47.36 roka)
- 95% IS pre strednú hodnotu je $(7.02, 11.41)$ a pre medián $(2.72, 7.08)$

Klasický prístup vs. analýza prežívania
Príklady

- priemerný vek prežívania pre všetkých pacientov bez rozdielu typu CF je 53.94 ± 2.10 rokov, kde 95% IS je rovný $(49.82, 58.06)$
- medián prežívania pre všetkých pacientov bez rozdielu typu CF je 70.82 roka; 95% **IS pre medián zatiaľ nie je možné vypočítať**
- Priemerný vek prežívania pre pacientov s typickou formou CF je 45.05 ± 2.47 rokov, kde 95% IS je $(40.21, 49.89)$
- Median prežívania je 52.26 roka, dolna hranica 95% IS pre medián je 36.43 roka; **Hornú hranicu IS pre medián zatiaľ nie je možné vypočítať**

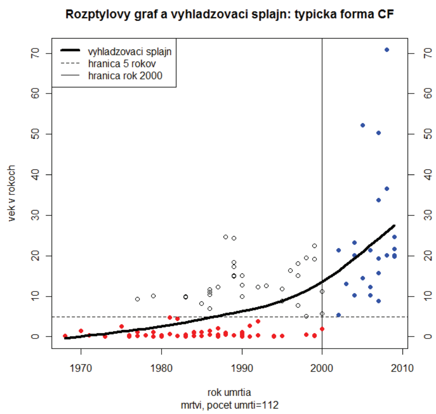
Klasický prístup vs. analýza prežívania
Príklady



Klasický prístup vs. analýza prežívania
Príklady

Tabuľka: Početnosti zomrelých v päťročných vekových intervaloch ($n = 112$). Označenia: vekové intervaly (zdola je interval otvorený a zhora uzavretý, okrem prvého, ktorý je aj zdola uzavretý): $I_1 = \langle 0, 5 \rangle$, $I_2 = (5, 10\rangle$, $I_3 = (10, 15\rangle$, $I_4 = (15, 20\rangle$, $I_5 = (20, 25\rangle$, $I_6 = 25, \max(vek)\rangle$.

vekové intervaly	I_1	I_2	I_3	I_4	I_5	I_6
početnosti	56	11	16	13	11	5
percentá	50%	9.82%	14.29%	11.61%	9.82%	4.46%



Ďalšie možné otázky
Zlyhanie: smrť
Adjustujúce (prognostické) premenné: *antropologické ukazovatele*, *funkčné charakteristiky pľúc* a pod.

Udalosť
Úvodné definície

Udalosť: *ukončenie pozorovania z dôvodu zlyhania alebo smrti pacienta* – do konca sledovaného obdobia

Príklady udalostí:

- **overall survival** – smrť z akéhokoľvek dôvodu
- **progression-free survival** – prvé znaky progresie choroby alebo smrť
- **disease-free survival** – prvé znovuobjavenie sa choroby alebo smrť
- **event-free survival** – prvé znovuobjavenie sa choroby, objavenie sa inej špecifikovanej choroby alebo smrť
- **disease-specific survival** (**cause-specific survival**) – smrť ako dôsledok špecifikovanej choroby
- **relapse-free survival** (**recurrence-free survival**) – prvé znaky recidívy (opakovania sa) chodoby
- **time-to-progression** – prvé znaky progresie choroby

Cenúrovanie
Úvodné definície

Cenzúra: *ukončenie pozorovania z dôvodu iného ako je zlyhanie alebo smrť pacienta* – do konca sledovaného obdobia dôjde k úmrtiu len niektorých pacientov, zatiaľ čo u ostatných k úmrtiu do konca sledovaného obdobia buď nedôjde alebo sa títo pacienti z pozorovania stratia

Príklady cenúr:

- **ukončenie štúdie (termination of the study)**: pacient prežije časový interval experimentu
- **konkurenčné riziko (competing risk)**: pacient zomrie z iného dôvodu, ako v dôsledku sledovanej choroby
- **prerušenie/vysadenie liečby (drop-out)**: pacient preruší liečbu a odíde z kliniky predčasne, napr. z dôvodu zlých vedľajších účinkov liečby, pacient sa sám rozhodne nepokračovať v liečbe
- **strata z ďalšieho sledovania (loss to follow-up)**: pacient sa rozhodne presťahovať a nemáme o ňom už žiadne informácie

Cenúrovanie
Cenzúrovanie I. typu

Základné princípy:

1. predpoklad – všetkých n jedincov vstupuje do experimentu súčasne
2. príčina cenzúrovania – plánované ukončenie experimentu
3. ide o **cenzúrovanie časom** – zvolíme pevné číslo t_c, ktoré nazveme *fixovaný cenzurujúci čas*
4. $T^{(1)} < T^{(2)} < \ldots < T^{(d)}$, kde $T^{(d)} < t_c < T^{(d+1)}$
5. **náhodná veličina** – počet skutočne pozorovaných zlyhaní $d \in \{0, 1, \ldots, n\}$
6. pozorujeme $X_1, X_2, \ldots, X_n$, kde

$$X_i = \min(T_i, t_c) = \begin{cases} T_i, T_i \leq t_c, & \text{pre necenzúrované } X_i \\ t_c, T_i > t_c, & \text{pre cenzúrované } X_i \end{cases}$$

7. skutočnému pozorovaniu potom zodpovedá náhodný vektor (X_i, δ_i), kde

$$\delta_i = \begin{cases} 1, T_i \leq t_c, & \text{pre necenzúrované } X_i \\ 0, T_i > t_c, & \text{pre cenzúrované } X_i \end{cases}$$

Cenúrovanie
Cenzúrovanie II. typu

Základné princípy:

1. predpoklad – všetkých n jedincov vstupuje do experimentu súčasne
2. príčina cenzúrovania – plánované ukončenie experimentu
3. ide o **cenzúrovanie zlyhaním** – zvolíme si pevné číslo d, ktoré nazveme *fixovaný počet zlyhaní*; ukončenie teda nastáva po vopred zvolenom počte d zlyhaní, kde $d = [np] + 1, p \in (0, 1)$
4. $X_1 = T^{(1)}, X_2 = T^{(2)}, \ldots, X_d = T^{(d)}, X_{d+1} = T^{(d)}, \ldots X_n = T^{(d)}$
5. **náhodná veličina** – čas trvania experimentu
6. pozorujeme $X_1, X_2, \ldots, X_n$, kde

$$X_i = \min(T_i, T^{(d)}) = \begin{cases} T_i, T_i \leq T^{(d)}, & \text{pre necenzúrované } X_i \\ T^{(d)}, T_i > T^{(d)}, & \text{pre cenzúrované } X_i \end{cases}$$

7. skutočnému pozorovaniu potom zodpovedá náhodný vektor (X_i, δ_i), kde

$$\delta_i = \begin{cases} 1, T_i \leq T^{(d)}, & \text{pre necenzúrované } X_i \\ 0, T_i > T^{(d)}, & \text{pre cenzúrované } X_i \end{cases}$$

Cenúrovanie
Progresívne (zrýchlené) cenzúrovanie I. typu

Základné princípy:

1. predpoklad – všetkých n jedincov vstupuje do experimentu súčasne
2. príčina cenzúrovania – plánované ukončenie experimentu
3. ide o **cenzúrovanie zlyhaním** – zvolíme čísla $t_{ci}, i = 1, 2, \ldots, k$, ktoré nazveme *fixované cenzurujúce časy*, v čase t_{ci} vyradíme m_i subjektov
4. $t_{c1} < t_{c2} < \ldots < t_{ck}$
5. v čase t_{c1} vyradíme m_1 subjektov, v čase t_{c2} vyradíme m_2 subjektov, $\ldots$, v čase t_{ck} vyradíme m_k subjektov
6. po k-tom kroku máme vyradených $m_1 + m_2 + \ldots + m_k$ subjektov
7. **náhodná veličina** – počet skutočne pozorovaných zlyhaní $d \in \{0, 1, \ldots, n\}$

Cenúrovanie
Progresívne (zrýchlené) cenzúrovanie II. typu

Základné princípy:

1. predpoklad – všetkých n jedincov vstupuje do experimentu súčasne
2. príčina cenzúrovania – plánované ukončenie experimentu
3. ide o **cenzúrovanie časom** – zvolíme čísla d_i, ktoré nazveme *fixované počty zlyhaní*; vyradenie teda nastáva po vopred zvolenom počte d_i zlyhaní, kde $d_i = [np_i] + 1, p_i \in (0, 1)$
4. po d_1 zlyhaniach vyradíme m_1 subjektov, po d_2 zlyhaniach vyradíme m_2 subjektov, $\ldots$, po d_k zlyhaniach vyradíme m_k subjektov
5. po k-tom kroku máme vyradených $m_1 + m_2 + \ldots + m_k$ subjektov
6. **náhodná veličina** – čas trvania experimentu

Základné princípy:

1. predpoklad – n jedincov nevstupuje do experimentu súčasne
2. **čas do zlyhania** $T_1, T_2, \ldots, T_n$ sú nezávislé, rovnako rozdelené náhodné premenné, kde náhodná veličina T_i $(i = 1, \ldots, n)$ má hustotu $f(t)$ a distribučnú funkciu $F(t)$
3. **čas do cenzúrovania** $C_1, C_2, \ldots, C_n$ sú nezávislé, rovnako rozdelené náhodné premenné, kde náhodná veličina C_i $(i = 1, \ldots, n)$ má hustotu $g(t)$ a distribučnú funkciu $G(t)$
4. pozorujeme $X_1, X_2, \ldots, X_n$, kde
$$X_i = \min(T_i, C_i) = \begin{cases} T_i, T_i \leq C_i, & \text{pre necenzúrované } X_i \\ C_i, T_i > C_i, & \text{pre cenzúrované } X_i \end{cases}$$
5. skutočnému pozorovaniu potom zodpovedá náhodný vektor (X, δ), kde $X_i = \min(T_i, C_i)$ a
$$\delta_i = \begin{cases} 1, T_i \leq C_i, & \text{pre necenzúrované } X \\ 0, T_i > C_i, & \text{pre cenzúrované } X \end{cases}$$
6. **náhodná veličina** – čas trvania experimentu a čas do cenzúry (ak $C_i = c$, ide o **ľubovoľné cenzúrovanie**)

Základné princípy:

Majme n subjektov. Označme $T_i, i = 1, 2, \ldots, n$, **nepozorovateľné časy zlyhania**. Skutočnému pozorovaniu potom zopovedá náhodný vektor (C_i, δ_i), kde C_i sú časy cenzúr a $\delta_i = I(T_i \leq C_i)$, t.j.

$$\delta_i = \begin{cases} 1, T_i \leq C_i, & \text{pre necenzúrované } X_i \\ 0, T_i > C_i, & \text{pre cenzúrované } X_i \end{cases}$$

Example (nádor pľúc, animálny model)

Laboratórne myši sú injektované látkou, ktorá spôsobuje nádor. Keďže tento druh nádoru nie je smrteľný, je potrebné myš najprv zabiť, aby sme zistili, či bol nádor indukovaný, t.j. po časovom úseku náhodnej dĺžky C je myš zabitá, aby sme zistili, či sa nádor vyvinul alebo nie. **Endpoint záujmu** je čas T do objavenia sa nádoru.

Základné princípy:

Majme opäť n subjektov. Označme $T_i, i = 1, 2, \ldots, n$, **nepozorovateľné časy zlyhania.** Vieme len, že T_i nastalo buď vnútri nejakého náhodného časového intervalu, pred jeho ľavou hranicou alebo po jeho pravej hranici. Označme C_{1i} a C_{2i} časy dvoch vyšetrení a indikačné funkcie definujeme nasledovne $\delta_{1i} = I(T_i \leq C_{1i})$, $\delta_{2i} = I(C_{1i} < T_i \leq C_{2i})$ a $\delta_{3i} = I(T_i > C_{2i})$, t.j.

$$\delta_{1i} = \begin{cases} 1, T_i \leq C_{1i}, & \text{pre necenzúrované } X_i \\ 0, T_i > C_{1i}, & \text{pre cenzúrované } X_i \end{cases},$$

$$\delta_{2i} = \begin{cases} 1, C_{1i} < T_i \leq C_{2i}, & \text{pre necenzúrované } X_i \\ 0, T_i > C_{2i}, & \text{pre cenzúrované } X_i \end{cases}$$

a nakoniec $\delta_{3i} = 0$.

Example (nádor pľúc, pacienti)

Pacienti navštevovali kliniku opakovane každých 4 až 6 mesiacov, kde pozorovania sú buď intervaly $(C_{1i}, C_{2i}\rangle$ ak sa retrakcia prsníka vyskytla medzi poslednými dvoma návštevami alebo (C_{2i}, ∞), ak sa do C_{2i} retrakcia nevyskytla.

Základné princípy:

Máme nasledovné tri možnosti:

1. udalosť mohla nastať niekedy pred prvým vyšetrením C_{1i}, kde $\delta_{1i} = 1$ a $\delta_{2i} = \delta_{3i} = 0$,
2. udalosť mohla nastať niekedy medzi prvým a druhým vyšetrením, t.j. v intervale $(C_{1i}, C_{2i}\rangle$, kde $\delta_{1i} = 0$, $\delta_{2i} = 1$ a $\delta_{3i} = 0$,
3. udalosť sa do druhého vyšetrenia nevyskytla, t.j. mohla nastať niekedy po C_{2i} (ale nevieme kedy), kde $\delta_{1i} = 0$, $\delta_{2i} = 0$ a $\delta_{3i} = 0$.

Nech $X_{1i} = C_{1i}$ a $X_{2i} = C_{2i}$. Skutočnému pozorovaniu potom zopovedá náhodný vektor
$$(X_{1i}, X_{2i}, \delta_{1i}, \delta_{2i}).$$

Všimnime si, že δ_{3i} nie je potrebné použiť, pretože nemáme ďalšie vyšetrenie po C_{2i}. Keby sme mali C_{3i} alebo aj ďalšie (po ňom nasledujúce) vyšetrenia, hovorili by sme **zovšeobecnenom intervalovom cenzúrovaní**.

Cenúrovanie
Funkcia vierohodnosti – pravé typy cenzúrovania

1. cenzúrovanie I. typu

$$L = \prod_{i=1}^{n} f(x_i)^{\delta_i} \times S_f(t_c)^{1-\delta_i}$$

2. cenzúrovanie II. typu

$$L = \frac{n!}{(n-d)!} f(t_{(1)}) f(t_{(2)}) \ldots f(t_{(d)}) \times S_f(t_{(d)})^{n-d}$$

3. náhodné cenzúrovanie

$$L = \prod_{i=1}^{n} f(x_i)^{\delta_i} S_f(x_i)^{1-\delta_i} = \prod_{i=1}^{n} \lambda(x_i)^{\delta_i} S_f(x_i)$$

Cenúrovanie
Funkcia vierohodnosti – intervalové cenzúrovanie

1. intervalové cenzúrovanie I. typu

$$L = \prod_{i=1}^{n} [S_f(x_i)]^{1-\delta_i} [F(x_i)]^{\delta_i}$$

2. intervalové cenzúrovanie II. typu

$$L = \prod_{i=1}^{n} [F(x_{1i})]^{\delta_{1i}} [F(x_{2i}) - F(x_{1i})]^{\delta_{2i}} [S_f(x_{2i})]^{\delta_{3i}},$$

kde $\delta_{3i} = 1 - \delta_{1i} - \delta_{2i}$

Označenia
Časy do zlyhania

Definition

Majme neusporiadané časy $t_1, t_2, \ldots, t_n$. Zoradené časy zapíšeme ako $t_{(1)}, t_{(2)}, \ldots, t_{(n)}$. Pokiaľ predpokladáme, že $t_1, t_2, \ldots, t_n$ sú už zoradené, t.j. $t_1 < t_2 < \ldots < t_n$, označenia v ďalšom texte sa týmto preznačením zjednodušia. Potom $t_n = t_{\max}$. Ak $t_{\max} < c_{\max}$, potom bez straty na všeobecnosti bude $t_n = c_{\max}$ (pozri aj výpočet strednej hodnoty času prežívania, kde je potrebné situáciu $t_{\max} < c_{\max}$ zohľadniť). V časoch cenzúr c sú hodnoty $S(c)$ a $\Lambda(c)$ – ako aj ostatných charakteristík – identické ich hodnotám v najbližšom čase zlyhania t, ktorý predchádza c. Preto, bez straty na všeobecnosti, uvažujeme n zoradených časov, v ktorých sa charakteristiky prežívania počítajú. Tieto časy označujeme $t_1, t_2, \ldots, t_n$. Ak máme v časoch t_i zhody, t.j. $t_1 \leq t_2 \leq \ldots \leq t_n$, potom počet rôznych časov bude $l \leq n$ a $t_l = t_{\max}$.

Riziko
Príklady

Example (zadanie z prednášky)

Závislosť hodnoty rizika na jednotkách času.

$\lambda(t) = \lim_{\Delta t \to 0^+} \frac{\Pr(t \leq T < t+\Delta t | T \geq t)}{\Delta t}$

$\Pr(t \leq T < t + \Delta t | T \geq t) = \frac{1}{4}$

Ak $\Delta t = \frac{1}{3}$ dňa, potom $\lambda(t) = \frac{\frac{1}{4}}{\frac{1}{3}} = 0.75$ na deň

Ak $\Delta t = \frac{1}{21}$ týždňa, potom $\lambda(t) = \frac{\frac{1}{4}}{\frac{1}{21}} = 5.25$ na týždeň

Definície
Príklady

Example (zadanie z prednášky)

Majme náhodný vektor (X_i, δ_i), definovaný nasledovne (pre nejaku fiktívnu i-tu štatistickú jednotku, t.j. subjekt)

1. $(X_i, \delta_i) = (3, 0)$, t.j. v čase $X_i = 3$ je cenzúra, $N_i(t) = N_i(3) = 0$, $Y_i(3) = Y_i(3) = 1$ $\rightarrow (N_i(3), Y_i(3)) = (0, 1)$
2. $(X_i, \delta_i) = (4, 1)$, t.j. v čase $X_i = 4$ je udalosť (zlyhanie), $N_i(4) = 1$, $Y_i(4) = 1$, t.j. $(N_i(4), Y_i(4)) = (1, 1)$
3. Ak máme viac udalostí: $(N_i(0.5), Y_i(0.5)) = (1, 1)$, $(N_i(2), Y_i(2)) = (2, 1)$

Zoznam zadaní príkladov
Príklady

Example (domáca úloha)

Nech nezáporná náhodná veličina T je charakterizovaná funkciou prežívania $S(T)$. Nech je k-ty moment, $\mathbb{E}(T^k)$, konečný, $\mathbb{E}(T^k) < \infty, k \in \mathbb{N}$. (a) Ukážte, že platí $\mathbb{E}(T) = \sum_{t \in \mathbb{N}_0} \Pr(T > t) = \sum_{t \in \mathbb{N}_0} 1 - F(T) = \sum_{t \in \mathbb{N}_0} S(T)$. Použite pri tom definíciu strednej hodnoty $\mathbb{E}(T) = \sum_{t \in \mathbb{N}_0} t \Pr(t)$ a pomocné tvrdenie $\underbrace{1 + 1 + \ldots + 1}_{t\text{-krát}} = \sum_{\xi=1}^{t} 1 = \sum_{\xi=0}^{t-1} 1 = \sum_{\xi < t, \xi \in \mathbb{N}_0} 1$.
(b) Ukážte, že platí $\mathbb{E}(T) = \int_0^\infty S(t)\, dt$. Použite pri tom definíciu strednej hodnoty $\mathbb{E}(T) = \int_0^\infty t f(t) dt$, aplikujte vlastnosti súm z DÚ 1A ako aj $\int_0^\infty S(t)\, dt = \int_0^\infty (\int_0^t 1 dx) S(t)\, dt$. Výpočet Vám uľahčí metóda per-partes.
(c) Pomocou metódy per-partes ukážte, že $\mathbb{E}(T^k) = k \int_0^\infty t^{k-1} S(t) dt$.

Zoznam zadaní príkladov
Príklady

Example (zadanie z cvičenia)

AML (pokrač.) Vypočítajte empirickú funkciu prežívania pre skupinu A.

Skupina	čas po kompletnú remisiu (v týždňoch)	n	udalostí	cenzúr
skupina A	$9, 13, 13+, 18, 23, 28+, 31, 34, 45+, 48, 161+$	11	7	4

Example (zadanie z cvičenia)

AML (pokrač.) Naprogramujte v R algoritmus na výpočet empirickej funkcie prežívania a aplikujte ho na skupinu A.

Example (zadanie z cvičenia)

AML (pokrač.) Naprogramujte v R algoritmus na výpočet empirickej funkcie prežívania len pre zlyhania (cenzúry nemeberime do úvahy) a aplikujte ho na skupinu A.

Zoznam zadaní príkladov
Príklady

Example (zadanie z prednásky)

Odvoďte maximálne vierohodný odhad funkcie prežívania $\widehat{S}_{KM}(t)$.

Example (zadanie z cvičenia)

AML (pokrač.) Naprogramujte v R algoritmus na výpočet KM odhadu funkcie prežívania a aplikujte ho na skupinu A.

Example (zadanie z cvičenia)

AML (pokrač.). Výpočtom a graficky porovnajte empirickú funkciu prežívania $S_n(t)$ s KM odhadom funkcie prežívania $\widehat{S}_{KM}(t)$ pre skupinu A.

Example (zadanie z cvičenie)

AML (pokrač.). Výpočtom a graficky porovnajte empirickú funkciu prežívania $S_n(t)$ len pre časy zlyhania (cenzúry nemeberime do úvahy) s KM odhadom funkcie prežívania $\widehat{S}_{KM}(t)$ pre skupinu A.

Example (domáca úloha)

Použitím funkcie vierohodnosti $L = \prod_{i=1}^{I} \lambda_i^{d_i}(1-\lambda_i)^{n_i-d_i}$ odvoďte maximálne vierohodný odhad $\widehat{Var}\left[\widehat{\lambda}_i\right]$, $i = 1, 2, \ldots, I$ a $\widehat{Var}\left[\widehat{\Lambda}(t)\right]$.

Example (zadanie z cvičenie)

AML (pokrač.). Vypočítajte rozptyl KM odhadu funkcie prežívania v čase 13 (pre skupinu A). Využite Greenwoodovu formulu.

Example (zadanie z cvičenie)

AML (pokrač.). Vypočítajte odhad rizika $\widehat{\lambda}$, odhad rizika v intervale $t_i \leq t < t_{i+1}$, odhad kumulatívneho rizika $\widehat{\Lambda}_{KM}$ a $\widehat{\Lambda}_{NA}$ spolu s ich rozptylmi $\widehat{Var}[\widehat{\Lambda}_{KM}]$ a $\widehat{Var}[\widehat{\Lambda}_{NA}]$ v čase 26 (pre skupinu A).

Example (zadanie z cvičenie)

AML (pokrač.). Nakreslite a porovnajte odhady kriviek prežívania $\widehat{S}_{KM}(t)$, $\widehat{S}_B(t)$ a $\widehat{S}_{FHmodB}(t)$ pre skupinu A.

Example (zadanie z cvičenie)

AML (pokrač.). Vypočítajte KM odhad funkcie prežívania $\widehat{S}_{KM}(t)$ a 95% IS pre $S(t)$ vo všetkých bodoch t v 1) plain škále, 2) log-škále a 3) log-log škále (pre skupinu A).

Example (domáca úloha)

Ak v náhodnom výbere nie sú cenzúry, skóre testovacia štatistika Z_S má za platnosti H_0 štandardizované normálne rozdelenie $Z_S = \frac{\widehat{S}(t)-S(t)}{\sqrt{\frac{S(t)(1-S(t))}{n}}} \sim N(0,1)$, kde platí $\{S(t) : |z| \leq z_{\alpha/2}\}$. Potom riešením kvadratickej rovnice $\{S(t) : (1+\frac{z_{\alpha/2}^2}{n})S^2(t) - (2\widehat{S}(t)+\frac{z_{\alpha/2}^2}{n})S(t) + \widehat{S}(t) \leq 0\}$ bude $100 \times (1-\alpha)$% IS pre $S(t)$. Odvoďte tento interval a upravte ho do podoby: vzorec pre stred IS $\pm z_{\alpha/2}\sqrt{\text{vzorec}}$.

Example (zadanie z cvičenie)

AML (pokrač.) Naprogramujte v R algoritmus na výpočet obsahu pod $\widehat{S}_{KM}(t)$ krivkou. Aplikujte ho na skupinu A. Porovnajte s aritmetickým priemerom časov do zlyhania a aritmetickým priemerom časov do zlyhania a časov cenzúr.

Example (zadanie z cvičenie)

AML (pokrač.). Vypočítajte priemerný čas prežívania $\widehat{\mu}$ a jeho rozptyl $\widehat{Var}(\widehat{\mu})$, medián času prežívania $\widetilde{\mu}$ a jeho rozptyl $\widehat{Var}[\widetilde{\mu}]$ (pre skupinu A). Porovnajte s necenzurovaným mediánom.

Example (zadanie z cvičenie)

Naprogramujte v R funkciu na výpočet kvantilov času prežívania t_p a ich $100 \times (1-\alpha)$% intervalov spoľahlivosti.

Example (zadanie z cvičenie)

(a) Naprogramujte v R funkcie na výpočet nasledovných odhadov funkcií prežívania $\widehat{S}_{KM}(t)$, $\widehat{S}_{KMmod}(t)$, $\widehat{S}_B(t)$ a $\widehat{S}_{FHmodB}(t)$, kde

1. $\widehat{S}_{KMmod}(t) = \prod_{i:t_i \leq t}(1 - \frac{1}{n_i})$ [pre nerozsekané a aj rozsekané zhody]
2. $\widehat{Var}\left[\widehat{S}_{KM}(t)\right]$, dolnú (DH) a hornú (HH) hranicu 95% *IS* v log-škále,
3. $\widehat{S}_B(t) = \exp\left(-\widehat{\Lambda}_{NA}(t)\right)$ [pre nerozsekané zhody]
4. $\widehat{S}_{FHmodB}(t) = \exp\left(-\widehat{\Lambda}_{FHmodB}(t)\right)$ [pre rozsekané zhody]

(b) Vypočítajte tieto odhady pre dáta (pozri tabuľku)

Example (zadanie z cvičenie; pokrač.)

(c) Naprogramujte v R výpočet počtu cenzúr v čase t_i, ak poznáte d_i a n_i (pozri tabuľku).
(d) Naprogramujte v R funkciu na výpočet odhadu rozptylu funkcie prežívania $\widehat{Var}\left[\widehat{S}_{KM}(t)\right]$, dolnú a hornú hranicu 95% IS pre $S(t)$ v log-škále.
(e) Vypočítajte tento odhad a IS pre $S(t)$ pre dáta (pozri tabuľku).

t	d_i	n_i
4.5	1	70
11.5	2	68
16.0	1	65
20.7	2	55
20.8	1	53
31.0	1	47
34.5	1	45
46.0	1	34
61.0	1	25
87.5	5	15

Example (zadanie z cvičenie)

AML (pokrač.) Naprogramujte v R algoritmus na výpočet $100 \times (1-\alpha)$% pásov spoľahlivosti pre funkciu prežívania v škále $S(t)$ – (a) Nairov a (b) Hall-Walnerov. Aplikujte na skupinu A. Výsledok porovnajte s IS pre $S(t)$ v škále $S(t)$. Na obrázku zobrazte IS bodovo (zobrazenie, ktoré je prednastavené v R je nesprávne).

Example (domáca úloha)

AML (pokrač.) Naprogramujte v R algoritmus na výpočet $100 \times (1-\alpha)$% pásov spoľahlivosti pre funkciu prežívania v log-log škále – (a) Nairov a (b) Hall-Walnerov. Aplikujte na skupinu A. Výsledok porovnajte s IS pre $S(t)$ v log-log škále. Na obrázku zobrazte IS bodovo (zobrazenie, ktoré je prednastavené v R je nesprávne).

Example (domáca úloha)

AML (pokrač.) Naprogramujte v R algoritmus na výpočet $100 \times (1-\alpha)$% pásov spoľahlivosti pre kumulatívne riziko v log-log škále – (a) Nairov a (b) Hall-Walnerov. Aplikujte na skupinu A. Výsledok porovnajte s IS pre kumulatívne riziko v log-log škále. Na obrázku zobrazte IS bodovo.

Example (domáca úloha)

AML (pokrač.) Naprogramujte v R algoritmus na výpočet odhadu strednej hodnoty zostatkového života a jej rozptylu a aplikujte ho na skupinu A v čase $t = 30$ týždňov.

Empirická funkcia prežívania
Príklady

Example

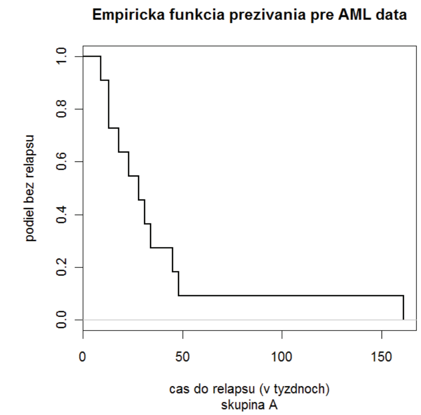
AML (pokrač.) Vypočítajte empirickú funkciu prežívania pre skupinu A.

	čas po kompletnú remisiu (v týždňoch)	n	udalostí	cenzúr
A	$9, 13, 13+, 18, 23, 28+, 31, 34, 45+, 48, 161+$	11	7	4

$$S_n(t) = \frac{\#\text{pozorovaní} > t}{n} = \frac{\#\{t_i > t\}}{n} = \frac{\sum_{i=1}^{n} I(t_i > t)}{n}$$

t	0	9	13	18	23	28	31	34	45	48	161
$S_n(t)$	$\frac{11}{11}$	$\frac{10}{11}$	$\frac{8}{11}$	$\frac{7}{11}$	$\frac{6}{11}$	$\frac{5}{11}$	$\frac{4}{11}$	$\frac{3}{11}$	$\frac{2}{11}$	$\frac{1}{11}$	$\frac{0}{11}$

Empirická funkcia prežívania
Príklady



Empirická funkcia prežívania a KM odhad
Príklady

Example

AML (pokrač.). Porovnajte empirickú funkciu prežívania $S_n(t)$ s KM odhadom funkcie prežívania $\widehat{S}_{KM}(t)$ pre skupinu A.

	čas po kompletnú remisiu (v týždňoch)	n	udalostí	cenzúr
A	$9, 13, 13+, 18, 23, 28+, 31, 34, 45+, 48, 161+$	11	7	4

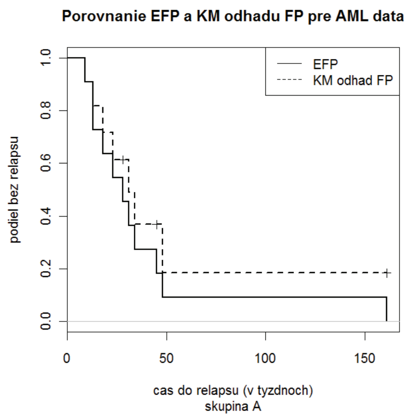
$$\widehat{S}_{KM}(t) = \prod_{i:t_i \leq t} \left[1 - \widehat{\lambda}_i\right], \text{ kde } \widehat{\lambda}_i = \frac{d_i}{n_i}$$

Empirická funkcia prežívania a KM odhad
Príklady

$$\widehat{S}_{KM}(0) = 1$$
$$\widehat{S}_{KM}(9) = \widehat{S}_{KM}(0) \times \tfrac{11-1}{11}$$
$$\widehat{S}_{KM}(13) = \widehat{S}_{KM}(9) \times \tfrac{10-1}{10}$$
$$\widehat{S}_{KM}(13+) = \widehat{S}_{KM}(13) \times \tfrac{9-0}{9} = \widehat{S}_{KM}(13)$$
$$\widehat{S}_{KM}(18) = \widehat{S}_{KM}(13) \times \tfrac{8-1}{8}$$
$$\widehat{S}_{KM}(23) = \widehat{S}_{KM}(18) \times \tfrac{7-1}{7}$$
$$\widehat{S}_{KM}(28+) = \widehat{S}_{KM}(23) \times \tfrac{6-0}{6} = \widehat{S}_{KM}(23)$$
$$\widehat{S}_{KM}(31) = \widehat{S}_{KM}(23) \times \tfrac{5-1}{5}$$
$$\widehat{S}_{KM}(34) = \widehat{S}_{KM}(31) \times \tfrac{4-1}{4}$$
$$\widehat{S}_{KM}(45+) = \widehat{S}_{KM}(34) \times \tfrac{3-0}{3} = \widehat{S}_{KM}(34)$$
$$\widehat{S}_{KM}(48) = \widehat{S}_{KM}(34) \times \tfrac{2-1}{2}$$
$$\widehat{S}_{KM}(161+) = \widehat{S}_{KM}(48) \times \tfrac{1-0}{1} = \widehat{S}_{KM}(48)$$

Empirická funkcia prežívania a KM odhad
Príklady

t	0	9	13	13+	18	23	28+	31	34	45+	48	161+
$S_n(t)$	$\frac{11}{11}$	$\frac{10}{11}$	$\frac{8}{11}$	-	$\frac{7}{11}$	$\frac{6}{11}$	$\frac{5}{11}$	$\frac{4}{11}$	$\frac{3}{11}$	$\frac{2}{11}$	$\frac{1}{11}$	$\frac{0}{11}$
$\widehat{S}_{KM}(t)$	1	0.91	0.82	0.82	0.72	0.61	0.61	0.49	0.37	0.37	0.18	0.18



Odhady
Príklady

Example

AML (pokrač.). Vypočítajte rozptyl KM odhadu funkcie prežívania v čase 13 (pre skupinu A). Využite Greenwoodovu formulu.

$$\widehat{Var_G[\widehat{S}_{KM}(t)]} = \widehat{S}^2_{KM}(t)\,\widehat{Var\left[\ln \widehat{S}_{KM}(t)\right]} = \widehat{S}^2_{KM}(t)\sum_{i:t_i\le t}\frac{d_i}{n_i(n_i-d_i)}$$

$$\widehat{Var_G[\widehat{S}_{KM}(t)]} = \widehat{S}^2_{KM}(t)\sum_{i:t_i\le t}\frac{d_i}{n_i(n_i-d_i)}$$

$$\widehat{Var_G[\widehat{S}_{KM}(13)]} = 0.82^2\left(\frac{1}{11(11-1)}+\frac{1}{10(10-1)}\right) = 0.0136$$

$$\widehat{SE_G[\widehat{S}_{KM}(13)]} = 0.1166$$

Odhady
Príklady

Example

AML (pokrač.). Vypočítajte riziko $\widehat{\lambda}$ a kumulatívne riziko $\widehat{\Lambda}_{KM}$ a $\widehat{\Lambda}_{NA}$ spolu s ich rozptylmi $\widehat{Var[\widehat{\Lambda}_{KM}]}$ a $\widehat{Var[\widehat{\Lambda}_{NA}]}$ v čase 26 (pre skupinu A) [viď R kód v prílohe, Príklad 2].

$\widehat{\lambda}(t_i) = \frac{d_i}{n_i}$

Odhad rizika v intervale $t_i \le t < t_{i+1}$ je rovný $\widehat{\lambda}(t) = \frac{d_i}{n_i(t_{i+1}-t_i)}$; hovoríme mu aj KM typ odhadu; odhad rizika zlyhania na jednotku času v intervale $\langle t_i, t_{i+1})$

$$\widehat{\Lambda}_{KM}(t) = -\ln(\widehat{S}_{KM}(t)) = -\ln\left(\prod_{i:t_i\le t}\frac{n_i-d_i}{n_i}\right)$$

$$\widehat{Var[\widehat{\Lambda}_{KM}(t)]} = \sum_{i:t_i\le t}\frac{d_i}{n_i(n_i-d_i)}$$

$$\widehat{\Lambda}_{NA}(t) = \sum_{i:t_i\le t}\frac{d_i}{n_i},\ \widehat{Var[\widehat{\Lambda}_{NA}(t)]} = \sum_{i:t_i\le t}\frac{d_i}{n_i^2}$$

Odhady
Príklady

$$\widehat{\lambda}(23) = \frac{1}{7} = 0.143$$

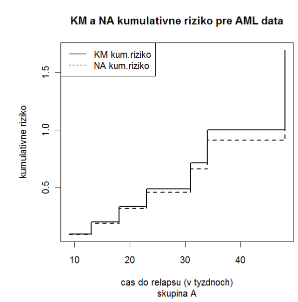
$$\widehat{\lambda}(26) = \widehat{\lambda}(23) = \frac{1}{7(31-23)} = 0.018$$

$$\widehat{\Lambda}_{KM}(26) = -\ln(\widehat{S}_{KM}(26)) = -\ln(0.61) = 0.49,$$

$$\widehat{\Lambda}_{NA}(26) = \frac{1}{11}+\frac{1}{10}+\frac{1}{8}+\frac{1}{7} = 0.4588$$

$$\widehat{Var[\widehat{\Lambda}_{KM}(t)]} = \frac{1}{11(11-1)}+\frac{1}{10(10-1)}+\frac{1}{8(8-1)}+\frac{1}{7(7-1)} = 0.0619$$

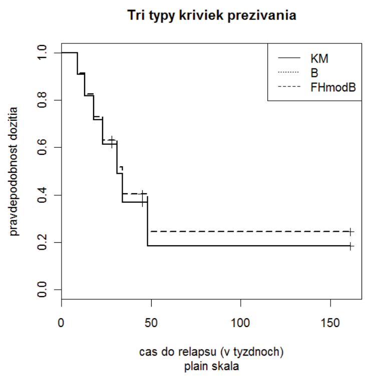
$$\widehat{Var[\widehat{\Lambda}_{NA}(t)]} = \frac{1}{11^2}+\frac{1}{10^2}+\frac{1}{8^2}+\frac{1}{7^2} = 0.0543$$



Odhady
Príklady

Example

AML (pokrač.). Nakreslite a porovnajte odhady kriviek prežívania $\widehat{S}_{KM}(t)$, $\widehat{S}_{B}(t)$ a $\widehat{S}_{FHmodB}(t)$ [viď R kód v prílohe, Príklad 1].



Medián a priemerný čas prežívania
Príklady

Example

AML (pokrač.). Vypočítajte priemerný čas prežívania $\widehat{\mu}$ a jeho rozptyl $\widehat{Var(\widehat{\mu})}$, medián času prežívania $\widetilde{\mu}$ a jeho rozptyl $\widehat{Var(\widetilde{\mu})}$ (pre skupinu A). Porovnajte s necenzurovaným mediánom.

$\widehat{\mu} = \sum_{i=1}^{I}(t_i - t_{i-1})\widehat{S}(t_{i-1}) = \sum_{i=0}^{I}\Delta t_i \widehat{S}(t_i)$, kde $\Delta t_i = t_{i+1} - t_i$, $I \le n$ je počet rôznych zlyhaní, $t_0 = 0$, $\widehat{S}(t_0) = 1$ a $\widehat{S}(t_{i-1})$ je výška funkcie v bode t_{i-1}.

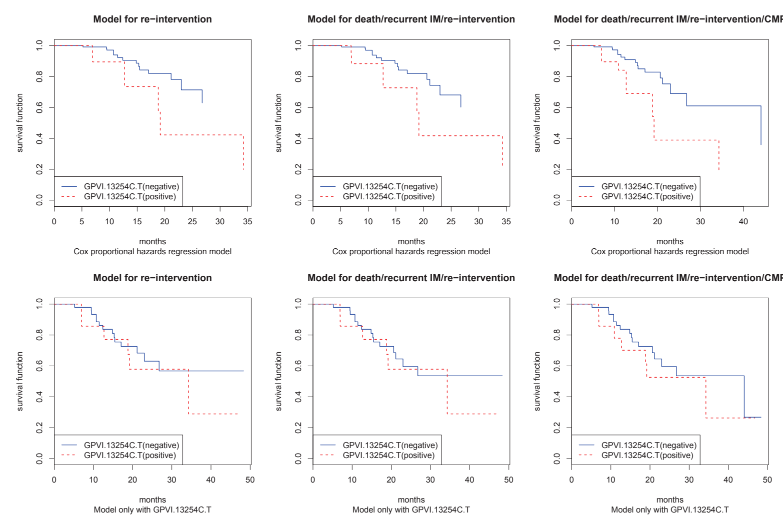
$$\widehat{Var}\left[\widehat{\mu}\right] = \sum_{i=1}^{I}\left[\sum_{t_i \le t_j \le t_n}\widehat{S}\left(t_j\right)\right]^2 \frac{d_i}{n_i(n_i-d_i)}$$

$$\widetilde{\mu} = \widehat{t}_{0.5},\ \widehat{Var}[\widetilde{\mu}] = \frac{Var_G[\widehat{S}(\widehat{t}_{0.5})]}{[\widehat{f}(\widehat{t}_{0.5})]^2}$$

Medián a priemerný čas prežívania
Príklady

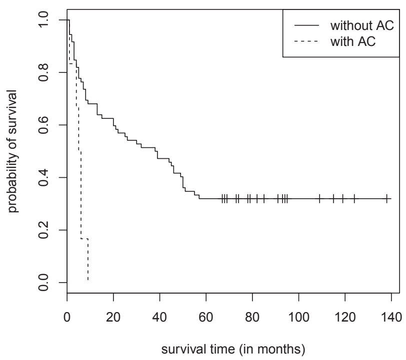
$t_{(n)} = 161$
$\widehat{\mu} = 52.6$ týždňa
$\widehat{Var}\left[\widehat{\mu}\right] = 19.8^2$
95% IS $= (13.792, 91.408)$ týždňa
$\widehat{t}_{0.5} = 31$ týždňov
$\widehat{u}_{0.5} = \max\{t_i : \widehat{S}(t_i) \ge 0.55\} = 23$
$\widehat{l}_{0.5} = \min\{t_i : \widehat{S}(t_i) \le 0.45\} = 34$
$\widehat{f}(31) = \frac{\widehat{S}(\widehat{u}_{0.5}) - \widehat{S}(\widehat{l}_{0.5})}{\widehat{l}_{0.5} - \widehat{u}_{0.5}} = \frac{\widehat{S}(23) - \widehat{S}(34)}{34 - 23} = \frac{0.6136364 - 0.3681818}{11} = 0.022$
$\widehat{Var}[31] = (\frac{0.16419327}{0.02231405})^2 = 54.144$
$\sqrt{\widehat{Var}[31]} = 7.358$
95% IS $= (16.578, 45.422)$ týždňa

Prežívanie pacientov po infarkte myokardu
Príklad: MACE (rôzne kombinácie; s a bez adjustácie) [funkcia prežívania]



Chronická myeloidná leukémia
Príklad: zmeny po transplantácii (typické a netypické) [funkcia prežívania]

- typické zmeny: $\widehat{\mu} = 58.08$ (± 6.70) mesiaca ($\widetilde{\mu} = 38.00$) a pravdepodobnosť úmrtia $49/72$
- netypické zmeny: $\widehat{\mu} = 5.17$ (± 0.98) mesiaca ($\widetilde{\mu} = 5.50$) pravdepodobnosť úmrtia 6/6



Odhady
Príklady

Example

Vypočítajte odhady nasledovné odhady funkcií prežívania $\widehat{S}_{KM}(t)$, $\widehat{S}_{KMmod}(t)$, $\widehat{S}_{B}(t)$ a $\widehat{S}_{FHmodB}(t)$ pre dáta (pozri tabuľku).

t	d_i	n_i
4.5	1	70
11.5	2	68
16.0	1	65
20.7	2	55
20.8	1	53
31.0	1	47
34.5	1	45
46.0	1	34
61.0	1	25
87.5	5	15

Odhady
Príklady

Pre 2 zhody v čase 12 platí:

- $\widehat{S}_{KM}(12) = (69/70)(66/68) = 0.9567$
- $\widehat{S}_{KMmod}(12) = (69/70)(67/68)(66/67) = (69/70)(66/68) = 0.9567$

Druhý prípad predstavuje úpravu $\widehat{S}_{KM}(t)$ pri zlome zhôd rozdelením času 11.5 na 11.48 a 11.52

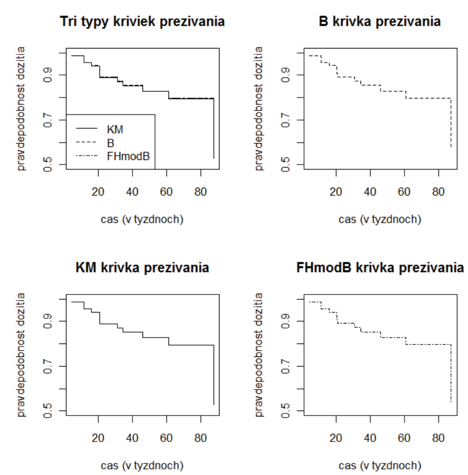
- $\widehat{S}_{B}(12) = \exp\left[-(1/70 + 2/68)\right] = 0.9572$
- $\widehat{S}_{FHmodB}(12) = \exp\left[-(1/70 + 1/68 + 1/67)\right] = 0.9570$

Pre 5 zhôd v čase 88 platí:

- $\widehat{S}_{KM}(88) = 0.5294$
- $\widehat{S}_{FHmodB}(88) = 0.5395$
- $\widehat{S}_{B}(88) = 0.5709$

$\widehat{S}_{FHmodB}(t)$ dáva vo všeobecnosti odhad bližšie ku $\widehat{S}_{KM}(t)$ a je menší ako $\widehat{S}_{B}(t)$ pri zhodách

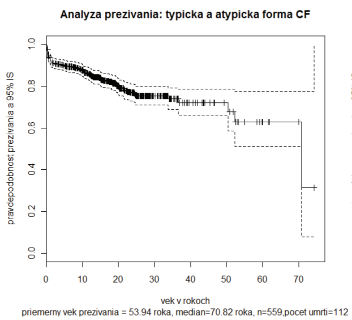
Odhady
Príklady

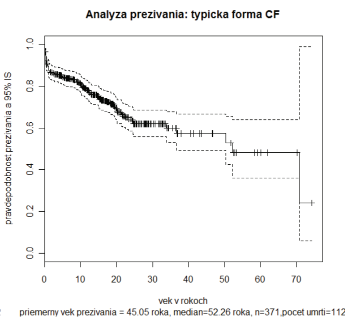


Odhady
Príklady

Example

CF (pokrač.) Nakreslite KM odhad funkcie prežívania $\widehat{S}_{KM}(t)$ a 95% IS pre $S(t)$ vo všetkých bodoch t pre pacientov s CF (všetci pacienti a typická forma).

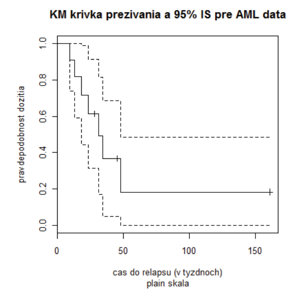


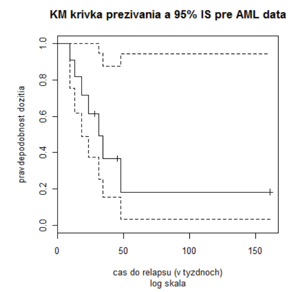


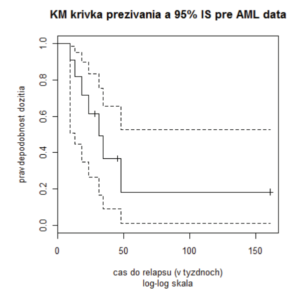
Odhady
Príklady

Example

AML (pokrač.). Vypočítajte KM odhad funkcie prežívania $\widehat{S}_{KM}(t)$ a 95% IS pre $S(t)$ vo všetkých bodoch t v 1) plain škále, 2) log-škále a 3) log-log škále (pre skupinu A).







Prehľad vzorcov
Odhady kumulatívneho rizika

- *Nelson-Aalenov (NA) odhad kumulatívneho rizika*

$$\widehat{\Lambda}_{NA}(t) = \int_0^t \frac{d\overline{N}(s)}{\overline{Y}(s)} ds \approx \sum_{i:t_i \le t} \frac{\Delta \overline{N}(t_i)}{\overline{Y}(t_i)}$$

- *Flemingom a Harringtonom (FH) modifikovaný NA odhad kumulatívneho rizika*

$$\begin{aligned}\widehat{\Lambda}_{FHmodNA}(t) &= \int_0^t \left[\sum_{j=0}^{\Delta\overline{N}(s)-1} \frac{1}{\overline{Y}(s) - j} \right] ds \\ &\approx \sum_{i:t_i \le t} \left[\sum_{j=0}^{\Delta\overline{N}(t_i)-1} \frac{1}{\overline{Y}(t_i) - j} \right]\end{aligned}$$

Prehľad vzorcov
Odhady kumulatívneho rizika

- *Nelson-Aalenov (NA) odhad kumulatívneho rizika*

$$\widehat{\Lambda}_{NA}(t) = \sum_{i:t_i \le t} \widehat{\lambda}(t_i) = \sum_{i:t_i \le t} \frac{d_i}{n_i},$$

- *Flemingom a Harringtonom (FH) modifikovaný NA odhad kumulatívneho rizika*

$$\widehat{\Lambda}_{FHmodNA}(t) = \sum_{i:t_i \le t} \left[\sum_{j=0}^{d_i - 1} \frac{1}{n_i - j} \right]$$

- *Kamplan-Meierov odhad funkcie prežívania*

$$\widehat{S}_{KM}(t)=\prod_{i:t_i\leq t}\left[1-\Delta\widehat{\Lambda}(t_i)\right],\Delta\widehat{\Lambda}(t_i)=\frac{\Delta\overline{N}(t_i)}{\overline{Y}(t_i)}$$

- *Breslowov odhad funkcie prežívania*

$$\widehat{S}_{B}(t)=\exp\left(-\widehat{\Lambda}(t)\right)=\prod_{i:t_i\leq t}e^{-\Delta\widehat{\Lambda}(t_i)},\Delta\widehat{\Lambda}(t_i)=\frac{\Delta\overline{N}(t_i)}{\overline{Y}(t_i)}$$

- *Flemingom a Harringtonom modifikovaný Breslowov odhad funkcie prežívania*

$$\widehat{S}_{FHmodB}(t)=\exp\left(-\widehat{\Lambda}_{FHmodNA}(t)\right)=\prod_{i:t_i\leq t}e^{-\Delta\widehat{\Lambda}_{FHmodNA}(t_i)}$$

- *Kamplan-Meierov odhad funkcie prežívania*

$$\widehat{S}_{KM}(t)=\prod_{i:t_i\leq t}\left[1-\frac{d_i}{n_i}\right]=\widehat{S}_{KMmod}=\prod_{i:t_i\leq t}\left[1-\sum_{j=0}^{d_i-1}\frac{1}{n_i-j}\right]$$

- *Breslowov odhad funkcie prežívania*

$$\widehat{S}_{B}(t)=\exp\left(-\widehat{\Lambda}_{NA}(t)\right)=\prod_{i:t_i\leq t}e^{-\frac{d_i}{n_i}}=e^{-\sum_{i:t_i\leq t}\frac{d_i}{n_i}}$$

- *Flemingom a Harringtonom modifikovaný Breslowov odhad funkcie prežívania*

$$\widehat{S}_{FHmodB}(t)=\exp\left(-\widehat{\Lambda}_{FHmodNA}(t)\right)=e^{-\sum_{i:t_i\leq t}\left[\sum_{j=0}^{d_i-1}\frac{1}{n_i-j}\right]}$$

- *Greenwoodov odhad rozptylu kumulatívneho rizika*

$$\widehat{Var_G\left[\widehat{\Lambda}(t)\right]}=\widehat{Var_G\left[-\ln\widehat{S}_{KM}(t)\right]}=\int_0^t\frac{d\overline{N}(s)}{\overline{Y}(s)\left[\overline{Y}(s)-d\overline{N}(s)\right]}ds$$

- *NA odhad rozptylu kumulatívneho rizika*

$$\widehat{Var\left[\widehat{\Lambda}_{NA}(t)\right]}=\int_0^t\frac{d\overline{N}(s)}{\overline{Y}^2(s)}ds$$

- *Flemingom a Harringtonom modifikovaný NA odhad rozptylu kumulatívneho rizika*

$$\widehat{Var\left[\widehat{\Lambda}_{FHmodNA}(t)\right]}=\int_0^t\left[\sum_{j=0}^{\Delta\overline{N}(s)-1}\frac{1}{\left[\overline{Y}(s)-j\right]^2}\right]ds$$

- *Greenwoodov odhad rozptylu kumulatívneho rizika*

$$\widehat{Var_G\left[\widehat{\Lambda}(t)\right]}=\sum_{i:t_i\leq t}\frac{\Delta\overline{N}(t_i)}{\overline{Y}(t_i)\left[\overline{Y}(t_i)-\Delta\overline{N}(t_i)\right]}$$

- *NA odhad rozptylu kumulatívneho rizika*

$$\widehat{Var\left[\widehat{\Lambda}_{NA}(t)\right]}=\sum_{i:t_i\leq t}\frac{\Delta\overline{N}(t_i)}{\overline{Y}^2(t_i)}$$

- *Flemingom a Harringtonom modifikovaný NA odhad rozptylu kumulatívneho rizika*

$$\widehat{Var\left[\widehat{\Lambda}_{FHmodNA}(t)\right]}=\sum_{i:t_i\leq t}\left[\sum_{j=0}^{\Delta\overline{N}(t_i)-1}\frac{1}{\left[\overline{Y}(t_i)-j\right]^2}\right]$$

Prehľad vzorcov
Odhady rozptylu kumulatívneho rizika

- *Greenwoodov odhad rozptylu kumulatívneho rizika*

$$\widehat{Var_G\left[\widehat{\Lambda}(t)\right]} = \sum_{i:t_i \le t} \frac{\Delta \overline{N}(t_i)}{\overline{Y}(t_i)\left[\overline{Y}(t_i) - \Delta \overline{N}(t_i)\right]} = \sum_{i:t_i \le t} \frac{d_i}{n_i(n_i - d_i)}$$

- *NA odhad rozptylu kumulatívneho rizika*

$$Var\,\widehat{\left[\widehat{\Lambda}_{NA}(t)\right]} = \sum_{i:t_i \le t} \frac{d_i}{n_i^2}$$

- *Flemingom a Harringtonom modifikovaný NA odhad rozptylu kumulatívneho rizika*

$$Var\,\widehat{\left[\widehat{\Lambda}_{FHmodNA}(t)\right]} = \sum_{i:t_i \le t} \left[\sum_{j=0}^{d_i - 1} \frac{1}{(n_i - j)^2}\right]$$

Prehľad vzorcov
$(1-\alpha)100\%$ IS $S(t)$ v t

- *škála $S(t)$* (**survival (plane) scale**)

$$\widehat{S}(t) \pm u_{\alpha/2}\sqrt{Var\,\widehat{\left[\widehat{S}(t)\right]}},\ \text{kde } Var\left[\widehat{S}(t)\right] = \widehat{S}^2(t)\,Var\left[\widehat{\Lambda}(t)\right]$$

- *škála kumulatívneho rizika* (**log-survival scale**)

$$\ln \widehat{S}(t) \pm u_{\alpha/2}\sqrt{Var\,\widehat{\left[\ln \widehat{S}(t)\right]}},\ \widehat{S}(t)\exp\left(\pm u_{\alpha/2}\sqrt{Var\,\widehat{\left[\widehat{\Lambda}(t)\right]}}\right)$$

$$\widehat{\Lambda}(t)\exp\left(\pm u_{\alpha/2}\sqrt{Var\,\widehat{\left[\ln \widehat{\Lambda}(t)\right]}}\right)$$

Prehľad vzorcov
$(1-\alpha)100\%$ IS $S(t)$ v t

$$\text{kde } Var\,\widehat{\left[\ln \widehat{\Lambda}(t)\right]} \approx Var\,\widehat{\left[\widehat{\Lambda}(t)\right]} / \left[\widehat{\Lambda}(t)\right]^2$$

- *log-log škála* (**log-log (survival) scale**; škála log-kumulatívneho rizika)

$$\ln(-\ln \widehat{S}(t)) \pm u_{\alpha/2}\sqrt{\widehat{Var[W]}},$$

$$\text{kde } W = \ln(-\ln \widehat{S}(t))),\ \widehat{Var[W]} = Var\,\widehat{\left[-\ln \widehat{S}(t)\right]} / (\ln \widehat{S}(t))^2,$$

$$(\widehat{S}(t))^{\exp(\pm u_{\alpha/2}\sqrt{\widehat{Var[W]}})},$$

a

$$\exp\left[-\widehat{\Lambda}(t)\exp\left(\pm u_{\alpha/2}\sqrt{Var\,\widehat{\left[\ln \widehat{\Lambda}(t)\right]}}\right)\right]$$

Prehľad vzorcov
Odhad strednej hodnoty do času prežívania a jej rozptylu

(Urezaný) odhad strednej hodnoty času do zlyhania $E[T]$

$$\widehat{\mu} = \int_0^{t_{\max}} \widehat{S}(t)\,dt,\ \widehat{\mu} = \sum_{i=0}^{I} \Delta t_i \widehat{S}(t_i),\ \Delta t_i = t_{i+1} - t_i,\ I \le n,$$

kde $\widehat{S}(t)$ je *KM* odhad funkcie prežívania, $t_{\max}$ je maximum pozorovaných časov, I počet rôznych zlyhaní

Odhad rozptylu $Var[\widehat{\mu}]$

$$\widehat{Var[\widehat{\mu}]} = \int_0^T \left[\int_t^T \widehat{S}(u)du\right]^2 \frac{d\overline{N}(t)}{\overline{Y}(t)\left[\overline{Y}(t) - d\overline{N}(t)\right]} dt$$

$$\widehat{Var[\widehat{\mu}]} = \sum_{i:t_i \le t_n} \left[\sum_{t_i \le t_j \le t_n} \widehat{S}(t_j)\right]^2 \frac{\Delta \overline{N}(t_i)}{\overline{Y}(t_i)\left[\overline{Y}(t_i) - \Delta \overline{N}(t_i)\right]}$$

Prehľad vzorcov
Odhad mediánu do času prežívania a jeho IS

Medián času prežívania je 50-ty percentil $t_{0.5}$
Medián funkcie prežívania je $S(t_{0.5}) = 0.5$
Výberový medián – prvý čas, v ktorom $\widehat{S}(t) \leq 0.5$, t.j. $\widetilde{\mu} = \widehat{S}^{-1}(0.5)$
Niekedy je potrebné použiť lineárnu interpoláciu pre $\widehat{t}_{0.5}$ v podobe

$$\widetilde{\mu}_{int} = t_i + (t_{i+1} - t_i)\frac{\widehat{S}(t_i) - 0.5}{\widehat{S}(t_i) - \widehat{S}(t_{i+1})}$$

Horná a dolná hranica IS pre medián – definovaná na základe IS pre $S(t)$ v danom čase, t.j.

- horná hranica IS pre medián je prvý čas, v ktorom je horná hranica IS pre $S(t)$ väčšia alebo rovná 0.5
- dolná hranica IS pre medián je prvý čas, v ktorom je dolná hranica IS pre $S(t)$ menšia alebo rovná 0.5

To korešponduje s narysovaním horizontálnej úsečky na grafe krivky prežívania, tj. pretnutím tejto úsečky s krivkou prežívania, dolnou a hornou hranicou IS pre $S(t)$ v danom čase

Prehľad vzorcov
Odhad kvantilov do času prežívania a ich rozptylov

Nech t_p je *p-ty kvantil* rozdelenia T ($100 \times$ *p-ty percentil*), teda $F(t_p) = \Pr(T < t_p) = p, t_p = F^{-1}(p)$. Potom

$$S(t_p) = \Pr(T \geq t_p) = 1 - p, t_p \leq S^{-1}(1-p)$$

Keďže KM krivka prežívania je schodovitá funkcia, inverzia $\widehat{S}^{-1}(t_p)$ nie je jednoznačne definovaná; *odhad kvantilu* bude potom

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

Aplikovaním delta metódy na $\widehat{Var_G(\widehat{S}(\widehat{t}_p))}$ dostaneme

$$\widehat{Var[\widehat{t}_p]} = \frac{\widehat{Var_G[\widehat{S}(\widehat{t}_p)]}}{[\widehat{f}(\widehat{t}_p)]^2}, \widehat{f}(\widehat{t}_p) = \frac{\widehat{S}(\widehat{u}_p) - \widehat{S}(\widehat{l}_p)}{\widehat{l}_p - \widehat{u}_p},$$

kde $\widehat{u}_p = \max\{t_i : \widehat{S}(t_i) \geq 1 - p + \epsilon\}$ a $\widehat{l}_p = \min\{t_i : \widehat{S}(t_i) \leq 1 - p - \epsilon\}$ pre $i = 1, 2, \ldots, I \leq n$, I je počet rozdielnych časov zlyhania, ϵ je veľmi malé číslo; vo všeobecnosti je $\epsilon = 0.05$ akceptovateľné, ale musí byť veľké, ak $|\widehat{l}_p - \widehat{u}_p| \approx 0$

Prehľad vzorcov
Odhady strednej hodnoty zostatkového života a jej rozptyl

$$\widehat{\text{mrl}}(t) = E(\widehat{T - t|T} > t) = \frac{\int_t^{t_{\max}} \widehat{S}(u)du}{\widehat{S}(t)}$$

$$\widehat{\text{mrl}(t)} = \frac{(t_{i+1} - t)\widehat{S}(t_i) + \sum_{j \geq i+1}(t_{j+1} - t_j)\widehat{S}(t_j)}{\widehat{S}(t)}, t_i \leq t < t_{i+1}$$

$$\begin{aligned} Var[\widehat{\widehat{\text{mrl}}(t)}] &= \frac{1}{\widehat{S}^2(t)}\left(\sum_{i:t \leq t_i \leq t_{\max}}\left(\int_{t_i}^{t_{\max}} \widehat{S}(u)du\right)^2 \frac{d_i}{n_i(n_i - d_i)}\right. \\ &+ \left.\left(\int_t^{t_{\max}} \widehat{S}(u)du\right)^2 \sum_{i:t_i \leq t} \frac{d_i}{n_i(n_i - d_i)}\right). \end{aligned}$$

$$\begin{aligned} Var[\widehat{\widehat{\text{mrl}}(t)}] &= \frac{1}{\widehat{S}^2(t)}\left(\sum_{i:t \leq t_i \leq t_{\max}}\left[\sum_{j:t_i \leq t_j \leq t_{\max}} \widehat{S}(t_j)\right]^2 \frac{d_i}{n_i(n_i - d_i)}\right. \\ &+ \left.\left(\sum_{i:t \leq t_i \leq t_{\max}} \widehat{S}(t_i)\right)^2 \sum_{i:t_i \leq t} \frac{d_i}{n_i(n_i - d_i)}\right) \end{aligned}$$

R
R príkazy [R je voľne dostupné na http://cran.r-project.org/]

Voľby argumentov fcie `Survfit` v knižnici `library(survival)`:

```
Surfit(surv(time,status)~1, type="...",error="...",conf.type
="...")
```

1. $\widehat{S}_{KM}(t)$: `type="kaplan-meier"` (prednastavené)
2. $\widehat{S}_{B}(t)$: `type="fleming-harrington"`
3. $\widehat{S}_{FHmodB}(t)$: `type="fh2"`
4. $Var_G\left[\widehat{\widehat{\Lambda}_{KM}(t)}\right]$: `error = "greenwood"` (prednastavené)
5. $Var\left[\widehat{\widehat{\Lambda}_{NA}(t)}\right] = \widehat{\sigma}_T^2$: `error = "tsiatis"`
6. **žiadny**: `conf.type ="none"`
7. **survival (plane) scale**: `conf.type ="plain"`
8. **log-survival scale**: `conf.type ="log"` (prednastavené)
9. **log-log (survival) scale**: `conf.type="log-log"`
10. koeficient spoľahlivosti `conf.int=0.95` (prednastavené)

Označme `surv.obj <- survfit(Surv(cas,status)~1)`. Priemerný vek prežívania a jeho smerodajná odchýlka (medián a jeho smerodajná odchýlka je súčasťou výstupu) sa vypočíta ako
1) `print(surv.obj,print.rmean=TRUE)` alebo
2) `print(surv.obj, rmean="individual")`

Na rozlíšenie typu cenzúrovania je dôležitý počet argumentov funkcie `Surv()`. Ak sú dva, t.j. `Surv(cas,status)`, ide o pravý typ cenzúrovania. Ak sú tri, t.j. `Surv(cas,cas1, status)`, potom ide o intervalové cenzúrovanie. Pomocným argumentom je `type="..."`, kde rozlišujeme `type="right"` (pravý typ), `type="interval"` (intervalový typ cenzúrovania I. typu; kde interval $(-\infty, t_i\rangle$ označujeme $(\mathrm{NA}, t_i\rangle$), `type="interval2"` (intervalový typ cenzúrovania II. typu; kde interval je typu $(t_{1i}, t_{2i}\rangle$ alebo interval (t_i, ∞), ktorý označujeme (t_i, NA)). Dolnou hranicou intervalu môže byť aj 0 a hornou hranicou $t_{\max}$.

Example

Intervalový typ cenzúrovania pre dáta `heart` – intervaly sa nachádzajú v stĺpcoch `heart$start` a `heart$stop`, status (udalosť) je v stĺpci `heart$event`. Význam premenných pozri v `help(heart)`.

Príklad 1 (poznámky sú uvedené za znakom #):

```
library(survival)
attach(aml)
names(aml)
aml.A <- aml[x=="Maintained",1]
status.A <- aml[x=="Maintained",2]
KM.aml.A.KM <- survfit(Surv(aml.A,status.A)~1,conf.type =
"plain",type="kaplan-meier") # $\widehat{S}_{KM}(t)$
KM.aml.A.B <- survfit(Surv(aml.A,status.A)~1,conf.type =
"plain",type="fleming-harrington") # $\widehat{S}_{B}(t)$
KM.aml.A.FHmodB <- survfit(Surv(aml.A,status.A)~1,conf.type =
"plain",type="fh2") # $\widehat{S}_{FHmodB}(t)$
# obrazok (lwd: hrubka ciary, lty: typ ciary)
plot(KM.aml.A.KM,xlab="cas do relapsu (v tyzdnoch)",
ylab="pravdepodobnost dozitia",conf.int=FALSE,lwd=2)
lines(KM.aml.A.KM,lty=1,lwd=2)
lines(KM.aml.A.B,lty=3,lwd=2)
lines(KM.aml.A.FHmodB,lty=2,lwd=2)
legend("topright",c("KM","B","FHmodB"),lty=c(1,3,2))
title(main="Tri typy kriviek prezivania",sub="plain skala")
```


Príklad 2 (poznámky sú uvedené za znakom #):

```
cas <- summary(KM.aml.A)$time # casy zlyhania $t_i$
n.i <- summary(KM.aml.A)$n.risk # pocet jedincov v riziku $n_i$
d.i <- summary(KM.aml.A)$n.event # pocet zlyhani $d_i$
KM <- summary(KM.aml.A)$surv # $\widehat{S}_{KM}(t)$
SE.KM <- summary(KM.aml.A)$std.err # $\widehat{SE(\widehat{S}_{KM}(t))}$
lambda.KM <- d.i/n.i
DIFF <- diff(cas, lag = 1) # dlzka intervalu napravo od $t_i$
DIFF[length(DIFF) + 1] <- NA # NA su chybajuce hodnoty
lambda.INT <- lambda.KM/DIFF
Lambda.KM <- -log(KM) # $\widehat{\Lambda}_{KM}(t)$
Lambda.NA <- cumsum(lambda.KM) # $\widehat{\Lambda}_{NA}(t)$
sumand <- d.i/(n.i*n.i)
se.Lambda.KM <- SE.KM/lambda.KM # $\widehat{SE(\widehat{\Lambda}_{KM}(t))}$
se.Lambda.NA <- sqrt(cumsum(sumand)) # $\widehat{SE(\widehat{\Lambda}_{NA}(t))}$
# round(cislo,kolko.des.miest)
# data.frame: datovy ramec
RIZ <- round(data.frame(cas, n.i, d.i, lambda.KM,lambda.INT,
Lambda.KM, se.Lambda.KM, Lambda.NA,se.Lambda.NA),4)
```


Príklad 2 (pokrač.):

```
# obrazok
# schodovita funkcia: type="s", prazdny obrazok: type="n"
# $ znamena indexaciu stlpca z datoveho ramca v podobe
# RIZ$nazov.stlpca
plot(RIZ$cas,RIZ$Lambda.KM,xlab="cas do relapsu (v
tyzdnoch)",ylab="kumulativne riziko",type="n")
lines(RIZ$cas,RIZ$Lambda.KM,lty=1,lwd=2,type="s")
lines(RIZ$cas,RIZ$Lambda.NA,lty=2,lwd=2,type="s")
abline(h=0,col="gray")
title(main="KM a NA kumulativne riziko pre AML
data",sub="skupina A")
legend("topleft",c("KM kum.riziko","NA kum.riziko"),lty=c(1,2))
```

Viac sa o R dozviete
1) v mojich skriptách na http://www.iam.fmph.uniba.sk/skripta/katina/, kde "podtržník _"vo význame priradenia výpočtu nejakému názvu treba nahradiť "<-"(ide o v minulosti používanú syntax v komerčnej verzii R, programe S-PLUS)
2) v knižke **An Introduction to R** na http://cran.r-project.org/ v časti *Manuals*

Inštalácia: *R Binaries*→*windows*→*base*→Download R 3.0.1 for Windows

Vybrané kapitoly z analýzy prežívania

Testovanie hypotéz

Stanislav Katina[1]

[1]Ústav matematiky a statistiky
Přírodovědecká fakulta
Masarykova univerzita v Brne

ZS 2013

INVESTICE DO ROZVOJE VZDĚLÁVÁNÍ

Poďakovanie

Tento učební text vznikl za přispění Evropského sociálního fondu a státního rozpočtu ČR prostřednictvím Operačního programu Vzdělávání pro konkurenceschopnost v rámci projektu Univerzitní výuka matematiky v měnícím se světě

(CZ.1.07/2.2.00/15.0203)

Testy na porovnanie kriviek prežívania

Prehľad testov

Delenie testov na porovnanie kriviek prežívania

- **typ dát:** necenzúrované alebo cenzúrované
- **počet porovnávaných kriviek prežívania:** $k = 2$ alebo $k \geq 3$
- **typ pozorovaní:** nepárové alebo párové
- **typ alternatívy:** všeobecná alebo zoradená
- **crossing efekt** (prekrižovanie sa kriviek prežívania): neexistuje alebo existuje
- **časový efekt liečby**: neexistuje alebo existuje

Testy na porovnanie kriviek prežívania

Prehľad testov

Neparametrické testy porovnania kriviek prežívania pre ***necenzurované dáta***

- testy porovnania **dvoch** kriviek prežívania ($k = 2$)
  - Wilcoxonov test (W)
  - Mann-Whitney test (MW)
  - Siegel-Tukey test (ST)
- testy porovnania **viac** kriviek prežívania ($k \geq 3$)
  - Kruskal-Wallis test (KW)
  - Jonckheere test (J)
  - Cuzick test (C)
  - Le test (L)

Neparametrické testy porovnania kriviek prežívania pre ***cenzurované dáta***

- testy porovnania **dvoch** kriviek prežívania ($k = 2$)
  - Gehan-Wilcoxon test, zovšeobecný Wilcoxonov test (GB)
  - Cox-Mantel test, log-rank test (CM)
  - Tarone-Ware test (TW)
  - Peto-Peto test (PP)
- testy porovnania **viac** kriviek prežívania ($k \geq 3$)
  - Gehan-Breslow test, zovšeobecný Wilcoxonov test, zovšeobecný Kruskal-Wallis test (GB)
  - Cox-Mantel test, log-rank test (CM)
  - Mantel-Haenszel test, log-rank test (MH)
  - Peto-Peto test (PP)

Testované hypotézy

- **nulová hypotéza** $H_0 : S_1(t) = S_2(t) = S(t)$
- **alternatíva hypotéza** H_1 :
  - $S_1(t) \neq S_2(t)$
  - $S_1(t) \overset{st}{\prec} S_2(t)$
  - $S_1(t) \overset{st}{\succ} S_2(t)$

  pre $\forall t$

  $S(t)$ je funkcia prežívania

  $\overset{st}{\prec}$ a $\overset{st}{\succ}$ stochasticky menší, resp. stochasticky väčší

Predpoklady

- $X_1, ..., X_{n_1}$ je náhodný výber (NV) z nejakého spojitého rozdelenia
- $Y_1, ..., Y_{n_2}$ je NV z rovnakého spojitého rozdelenia a je oproti prvému rozdeleniu posunuté o nejakú konštantu δ
- veličiny $X_1, ..., X_{n_1}$ a $Y_1 - \delta, ..., Y_{n_2} - \delta$ majú rovnaké rozdelenie
- oba výbery sú nezávislé

Hypotézy

- $H_0 : \delta = 0$ ($S_1(t) = S_2(t), \forall t$)
- $H_1 : \delta \neq 0$ ($S_1(t) \neq S_2(t)$, pre aspoň jedno t)

Označenia

- n_j je počet pozorovaní v j-tom NV, $j = 1, 2$
- $n_1 + n_2 = n$
- nech $R_1, R_2, ..., R_{n_1}$ sú poradia prvého NV v rámci usporiadaného združeného NV

Wilcoxonova štatistika

$$W_X = S_W = \sum_{i=1}^{n_1} R_i$$

Stredná hodnota a rozptyl S_W :

$$E_0\left[S_W\right]=\frac{n_1\left(n+1\right)}{2}$$

$$Var_0\left[S_W\right]=\frac{n_1 n_2\left(n+1\right)}{12}$$

Wilcoxonov test

Ak $n_1, n_2 \geq 10$

$$Z_W=\frac{S_W-E_0\left[S_W\right]}{\sqrt{Var_0\left[S_W\right]}}\sim N\left(0,1\right)$$

Stredná hodnota a rozptyl S_W :

$$E_0\left[S_W\right]=\frac{n_1\left(n+1\right)}{2}$$

$$Var_0[S_W|\mathbf{t}]=\frac{n_1 n_2\left(n+1\right)}{12}\left[1-\frac{1}{n\left(n^2-1\right)}\sum_{j=1}^{L} t_j\left(t_j^2-1\right)\right]$$

Wilcoxonov test

Ak $n_1, n_2 \geq 10$

$$Z_{W,kor}=\frac{S_W-E_0\left[S_W\right]}{\sqrt{Var_0\left[S_W\right]-\frac{n_1 n_2\sum_j\left(t_j^3-t_j\right)}{12(n_1+n_2)(n_1+n_2-1)}}}\sim N\left(0,1\right)$$

Predpoklady

- $X_1, ..., X_{n_1}$ je NV z nejakého spojitého rozdelenia
- $Y_1, ..., Y_{n_2}$ je NV z rovnakého spojitého rozdelenia a je oproti prvému rozdeleniu posunuté o nejakú konštantu δ
- oba výbery sú nezávislé
- nech $\left(X_i, Y_j\right)$ sú možné páry pozorovaní, pre ktoré môže nastať buď $X_i < Y_j$ alebo $X_i > Y_j$

Hypotézy

- $H_0 : \delta = 0$ $(S_1(t) = S_2(t), \forall t)$
- $H_1 : \delta > 0$ $(S_1(t) \overset{st}{\prec} S_2(t)$, pre aspoň jedno $t)$

Označenia

- n_j je počet pozorovaní v j-tom NV, $j = 1, 2$
- $n_1 + n_2 = n$

Mann-Whitney štatistika

$$S_{MW}=\#\left(x_i, y_j\right),\ \text{kde } x_i > y_j$$

Testy na porovnanie dvoch kriviek prežívania
Mann-Whitney test

Stredná hodnota a rozptyl S_{MW} :

$$E_0\left[S_{MW}\right] = \frac{n_1 n_2}{2}$$

$$Var_0\left[S_{MW}\right] = Var_0\left[S_W\right] = \frac{n_1 n_2\left(n_1+n_2+1\right)}{12}$$

Mann-Whitney test

Ak $n_1, n_2 \geq 10$

$$Z_{MW} = \frac{S_{MW} - E_0\left[S_{MW}\right]}{\sqrt{Var_0\left[S_{MW}\right]}} \sim N(0,1)$$

Testy na porovnanie dvoch kriviek prežívania
Mann-Whitney vs Wilcoxonov test

U_X vyjadruje počet dvojíc X_i, Y_j, kde platí $X_i < Y_j$

$$U_X = n_1 n_2 + \frac{n_1\left(n_1+1\right)}{2} - W_X,$$

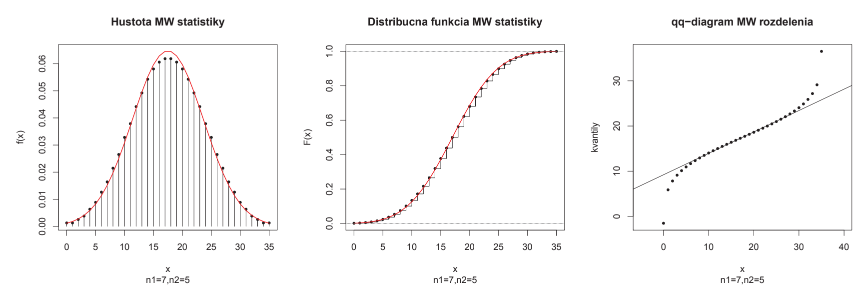
U_Y vyjadruje počet dvojíc X_i, Y_j, kde platí $X_i > Y_j$

$$U_Y = n_1 n_2 + \frac{n_2\left(n_2+1\right)}{2} - W_Y$$

Asymptotická normalita rozdelenia S_{MW}

```
n1 <- 7; n2 <- 5; n <- n1+n2
SUMmax <- sum(6:12); SUMmin <- sum(1:5)
Smw.max <- n1*n2-SUMmin+n2*(n2+1)/2
Smw.min <- n1*n2-SUMmax+n1*(n1+1)/2
x <- Smw.min:Smw.max
fx <- dwilcox(x, n1, n2); Fx <- pwilcox(x, n1, n2)
E.Smw <- n1*n2/2; Var.Smw <- n1*n2*(n+1)/12
fx.norm <- dnorm(x,E.Smw,sqrt(Var.Smw))
Fx.norm <- cumsum(dnorm(x,E.Smw,sqrt(Var.Smw)))/sum(dnorm(x,E.Smw,sqrt(Var.Smw)))
par(mfcol=c(1,3))
plot(x, fx,type="h", col="black", sub="n1=7,n2=5", xlab="x",
        ylab="f(x)",main= "Hustota MW statistiky",ylim=range(c(fx.norm,fx)))
points(x, fx,pch=16,cex=0.7)
lines(x,fx.norm,col="red",lwd=5)
plot(x, Fx,type="s", col="black", sub="n1=7,n2=5", xlab="x",
        ylab="F(x)",main= "Distribucna funkcia MW statistiky")
points(x, Fx,pch=16,cex=0.7)
lines(x,Fx.norm,col="red",lwd=5)
abline(h=0:1, col="gray20",lty=2)
p.mw <- seq(0.001,0.999,length=length(x))
q.mw <- qnorm(p.mw,E.Smw,sqrt(Var.Smw))
plot(x,q.mw,main = "qq-diagram MW rozdelenia", ylab="kvantily", sub="n1=7,n2=5",
        pch=16,cex=0.7,asp=1)
x1 <- quantile(x,0.25); x2 <- quantile(x,0.75)
y1 <- quantile(q.mw,0.25); y2 <- quantile(q.mw,0.75)
b1 <- (y2 - y1)/(x2 - x1); b0 <- y1 - b1 * x1
abline(a=b0,b=b1,lwd=2)
```


Testy na porovnanie dvoch kriviek prežívania
Mann-Whitney vs Wilcoxonov test



Obr.: Rozdelenie Mann-Whitneyho štatistiky S_{MW} $(n_1 = 7, n_2 = 5)$

Predpoklady

- efekt nového typu liečby spôsobí rast alebo klesanie krivky prežívania oproti pôvodnému typu liečby
- liečba nespôsobuje zmenu priemernej odpovede, ale výsledná odpoveď
- $X_1, ..., X_{n_1}$ je náhodný výber (NV) z nejakého spojitého rozdelenia
- $Y_1, ..., Y_{n_2}$ je náhodný výber (NV) z nejakého spojitého rozdelenia
- oba výbery sú nezávislé

Hypotézy

- $H_0 : Var(S_1(t)) = Var(S_2(t)), \forall t$
- $H_1 : Var(S_1(t)) \neq Var(S_2(t))$, pre aspoň jedno t

Podstata Siegel-Tukey testu je nasledovná:

- poradie R_1 priradíme najmenšiemu pozorovaniu
- poradie R_2 priradíme najväčšiemu pozorovaniu
- poradie R_3 priradíme druhému najmenšiemu pozorovaniu
- poradie R_4 priradíme druhému najväčšiemu pozorovaniu atď.

Siegel-Tukey štatistika

$$S_{ST} = \sum_{i=1}^{n_1} R_i,$$

kde daná suma prislúcha súčtu poradí pre prvý NV

Stredná hodnota a rozptyl S_{ST} (resp. S_{ST}^{alt}):

$$E_0\left[S_{ST}\right] = \frac{n_1\left(n_1 + n_2 + 1\right)}{2}$$

$$Var_0\left[S_{ST}\right] = \frac{n_1 n_2\left(n_1 + n_2 + 1\right)}{12}$$

Siegel-Tukey test

Ak $n_1, n_2 \geq 10$

$$Z_{ST} = \frac{S_{ST} - E_0\left[S_{ST}\right]}{\sqrt{Var_0\left[S_{ST}\right]}} \sim N\left(0, 1\right)$$

Podstata Leveneho alternatívy ST testu (Levene testu) je nasledovná:

- odchýlky $D_X = |X - \mu_X|$ a $D_Y = |Y - \mu_Y|$
- odchýlky $\{d_i = |x_i - \overline{x}|\}_{i=1}^{n_1}$ a $\{d_j = |y_j - \overline{y}|\}_{j=1}^{n_2}$
- $d_{(1)} < d_{(2)} < ... < d_{(n)}, n = n_1 + n_2$

Levene štatistika

$$S_L = S_{ST}^{alt} = \sum_{i=1}^{n_1} R_{diff,(i)},$$

kde $R_{diff,(i)}$ predstavujú poradia odchýlok od priemeru pre prvý NV

Stredná hodnota a rozptyl S_{ST}^{alt}:

$$E_0\left[S_{ST}\right] = E_0\left[S_{ST}^{alt}\right] = \frac{n_1\left(n_1+n_2+1\right)}{2}$$

$$Var_0\left[S_{ST}\right] = Var_0\left[S_{ST}^{alt}\right] = \frac{n_1 n_2\left(n_1+n_2+1\right)}{12}$$

Levene test

Ak $n_1, n_2 \geq 10$

$$Z_L = Z_{SL}^{alt} = \frac{S_{ST}^{alt} - E_0\left[S_{ST}^{alt}\right]}{\sqrt{Var_0^{alt}\left[S_{ST}\right]}} \sim N\left(0,1\right)$$

Example

Nech $t_{ij}, i = 1, ..., n_j, j = 1, 2$ sú časy do zlyhania (úmrtia) od diagnostiky nádoru pľúc v mesiacoch, kde $j = 1$ predstavuje I. typ terapie a $j = 2$ zasa II. typ terapie, nech $\psi_{ij} = 0$ ak $j = 1$ a $\psi_{ij} = 1$ ak $j = 2$ (pozri tab.). Otestujete $H_0 : S_1\left(t\right) = S_2\left(t\right)$, alternatíva $H_1 : S_1\left(t\right) \neq S_2\left(t\right)$. Použite aj $S_{MW}, S_{ST}, S_{ST}^{alt}$. Vždy presne naformulujte H_1.

t_{ij}	52	240	19	53	15	43	340	133	111	231	378	49
ψ_{ij}	0	1	1	0	0	1	1	0	0	1	0	0

t_{ij}	52	240	19	53	15	43	340	133	111	231	378	49
ψ_{ij}	0	1	1	0	0	1	1	0	0	1	0	0

$t_{(i)}$	15	19	43	49	52	53	111	133	231	240	340	378
ψ_{ij}	0	1	1	0	0	0	0	0	1	1	1	0
$R_i^{(1)}$	1	-	-	4	5	6	7	8	-	-	-	12
$R_i^{(2)}$	-	2	3	-	-	-	-	-	9	10	11	-

Wilcoxon test

$W_{X_2} = S_W = \sum_{i=1}^{5} R_i^{(2)} = 35$

$E_0\left[S_W\right] = \frac{5(7+5+1)}{2}$

$Var_0\left[S_W\right] = \frac{7\times5(7+5+1)}{12}$

$Z_W = (35 - 32.5)/6.157651 = 0.405999$, p-hodnota=0.6847

Mann-Whitney test

$S_{MW} = n_1 n_2 - \sum R_i^{(2)} + n_1(n_1+1)/2 = 20$

$E_0[S_{MW}] = n_1 n_2/2 = 17.5$,

$Var_0[S_{MW}] = n_1 n_2(n+1)/12 = 37.91667$

$Z_{ST} = (20 - 17.5)/37.91667 = 0.405999$, p-hodnota=0.6847

Siegel-Tukey test

t_{ij}	52	240	19	53	15	43	340	133	111	231	378	49
ψ_{ij}	0	1	1	0	0	1	1	0	0	1	0	0
$R_i^{(1)}$	9	-	-	11	1	-	-	10	12	-	2	7
$R_i^{(2)}$	-	6	3	-	-	5	4	-	-	8	-	-

$S_{ST} = \sum_{i=1}^{5} R_i^{(2)} = 26$

$E_0\left[S_{ST}\right] = \frac{5(7+5+1)}{2}$, $Var_0\left[S_W\right] = \frac{7\times5(7+5+1)}{12}$

$Z_{ST} = (26 - 32.5)/6.157651 = -3.166792$, p-hodnota=0.2912

Testy na porovnanie kriviek prežívania
Prehľad testov

Testované hypotézy

- **nulovná hypotéza** $H_0 : S_i(t) = S_j(t) = S(t)$
- **alternatíva hypotéza** H_1 :
  - $S_i(t) \neq S_j(t)$ pre aspoň jedno i, j (KW test)
  - $S_i(t) \overset{st}{\prec} S_j(t)$ (J,C,L testy)
  - $S_i(t) \overset{st}{\succ} S_j(t)$ (J,C,L testy)

pre $\forall t$, $i < j$, $i, j = 1, 2, ..., k$
k je počet porovnávaných kriviek prežívania
$S(t)$ je funkcia prežívania
$\overset{st}{\prec}$ a $\overset{st}{\succ}$ stochasticky menší, resp. stochasticky väčší

Testy na porovnanie viacerých kriviek prežívania
Kruskall-Wallis test

Označenia

- n_i je počet pozorovaní v i-tom NV, $i = 1, 2, ..., k$
- $n = \sum_{i=1}^{k} n_i$
- $R_i = \sum_{j=1}^{n_i} R_{ij}$, $R = \sum_{i=1}^{k} R_i$, $\overline{R}_i = R_i / n_i$, $\overline{R} = R/n = (n+1)/2$

Kruskall-Wallisova testovacia štatistika

$$\begin{aligned} S_{KW} &= \frac{12}{n(n+1)} \sum_{i=1}^{k} n_i \left(\frac{R_i}{n_i} - \frac{n+1}{2} \right)^2 \\ &= \frac{12}{n(n+1)} \sum_{i=1}^{k} \frac{R_i^2}{n_i} - 3(n+1) \\ &= \frac{1}{Var[R]} \left(\sum_{i=1}^{k} \frac{R_i^2}{n_i} - \frac{n(n+1)^2}{4} \right) \sim \chi^2_{k-1} \end{aligned}$$

Testy na porovnanie viacerých kriviek prežívania
Kruskall-Wallisov test

Rozptyl S_{KW} :

$$Var[R] = \frac{1}{n-1} \sum_{i=1}^{k} \sum_{j=1}^{n_i} \left(R_{ij} - \frac{n+1}{2} \right)^2$$

$$\begin{aligned} Var[R|\mathbf{t}] &= \frac{1}{n-1} \sum_{i=1}^{k} \sum_{j=1}^{n_i} \left(R_{ij|\mathbf{t}} - \frac{n+1}{2} \right)^2 \\ &= \frac{n(n+1)}{12} \left[1 - \frac{1}{n(n^2-1)} \sum_{j=1}^{L} t_j \left(t_j^2 - 1 \right) \right] \end{aligned}$$

Kruskall-Wallisov test

$$\widetilde{S}_{KW} = \frac{1}{Var[R|\mathbf{t}]} \sum_{i=1}^{k} n_i \left(\frac{R_i}{n_i} - \frac{n+1}{2} \right)^2 \sim \chi^2_{k-1}$$

Testy na porovnanie viacerých kriviek prežívania
Jonckheere test

Označenia

- $i < j$, teda $i = 1, 2, ..., k-1; j = 1+i, ..., k$, ďalej $\alpha_i = 1, 2, ..., n_i$, $\alpha_j = 1, 2, ..., n_j$
- nech S_{MW}^{ij} je Mann-Whitney štatistika porovnávajúca i-ty a j-ty výber

$$S_{MW}^{ij} = \# \left(x_{i\alpha_i}, x_{j\alpha_j} \right), \text{ kde } x_{i\alpha_i} < x_{j\alpha_j}$$

Jonckheere štatistika

$$S_J = \sum_{i<j} S_{MW}^{ij} = \sum_{i=1}^{k-1} \sum_{j=1+i}^{k} S_{MW}^{ij}$$

Stredná hodnota a rozptyl S_J :

$$E_0\left[S_J\right]=\frac{n^2-\sum n_i^2}{4}$$

$$Var_0\left[S_J\right]=\frac{n^2\left(2n+3\right)-\sum n_i^2\left(2n_i+3\right)}{72}$$

Jonckheere test

$$Z_J=\frac{S_J-E_0\left[S_J\right]}{\sqrt{Var_0\left[S_J\right]}}\sim N\left(0,1\right)$$

Alternatívna podoba Jonckheere štatistiky

$$\widetilde{S}_J=2\sum_{i=1}^{k-1}\sum_{j=1+i}^{k}S_{MW}^{ij}-\sum_{i=1}^{k-1}\sum_{j=1+i}^{k}n_in_j,$$

kde $\sum_{i=1}^{k-1}\sum_{j=1+i}^{k}n_in_j=\max_{\forall S}S_J$

Stredná hodnota a rozptyl $\widetilde{S}_J$:

$$E[\widetilde{S}_J]=0,\ Var_0\left[\widetilde{S}_J\right]=\frac{n^2\left(2n+3\right)-\sum n_i^2\left(2n_i+3\right)}{18}$$

Alternatívna podoba Jonckheere testu

$$\widetilde{Z}_J=\frac{\widetilde{S}_J-E_0\left[\widetilde{S}_J\right]}{\sqrt{Var_0\left[\widetilde{S}_J\right]}}\sim N\left(0,1\right)$$

Označenia

- nech $(X_1,Y_1)^T,\ldots,(X_n,Y_n)^T$ NV z dvojrozmerného rozdelenia
- realizácie $(x_i,y_i)^T, i=1,2,\ldots,n$
- dvojicu jednotiek s indexami i a j, (x_i,y_i) a (x_j,y_j), nazveme
  - *konkordantná* (usporiadaná) pokiaľ platí $x_i<x_j\wedge y_i<y_j$ alebo $x_i>x_j\wedge y_i>y_j$
  - *diskordantná* (neusporiadaná) pokiaľ platí $x_i<x_j\wedge y_i>y_j$ alebo $x_i<x_j\wedge y_i>y_j$
  - ak platí $x_i=x_j$ alebo $y_i=y_j$, nejde ani o konkordantný ani o diskordantný vzťah
- $C+D\leq n\left(n-1\right)$, C je počet konkordantných dvojíc, D počet diskonkordantných dvojíc medzi $\left(x_1,y_1\right),\ldots,\left(x_n,y_n\right)$

Kendallov korelačný koeficient

$$\tau=\tfrac{1}{n(n-1)}\sum_{i=1}^{n}\sum_{j=1}^{n}sign\left(x_i-x_j\right)sign\left(y_i-y_j\right),$$

čo je identické s

$$\widetilde{\tau}=\tfrac{2}{n(n-1)}\sum_{i=1}^{n-1}\sum_{j=i+1}^{n}sign\left(x_i-x_j\right)sign\left(y_i-y_j\right),$$

kde

$$sign(x_i-x_j)=\begin{cases}1 & \text{ak } x_i>x_j\\ -1 & \text{ak } x_i<x_j\\ 0 & \text{ak } x_i=x_j\end{cases}$$

$$sign(y_i-y_j)=\begin{cases}1 & \text{ak } y_i>y_j\\ -1 & \text{ak } y_i<y_j\\ 0 & \text{ak } y_i=y_j\end{cases}$$

Poradovú alternatíva Kendallovho korelačného koeficientu

- nech $R_1, \ldots, R_n$ sú poradia veličín $x_1, \ldots, x_n$
- nech $Q_1, \ldots, Q_n$ sú poradia veličín $y_1, \ldots, y_n$

Potom platí

$$\tau = \frac{1}{n(n-1)} \sum_{i=1}^{n} \sum_{j=1}^{n} sign\left(R_i - R_j\right) sign\left(Q_i - Q_j\right),$$

čo je identické s

$$\widetilde{\tau} = \frac{2}{n(n-1)} \sum_{i=1}^{n-1} \sum_{j=i+1}^{n} sign\left(R_i - R_j\right) sign\left(Q_i - Q_j\right)$$

Platí

$$\widetilde{\tau} \in \langle -1, 1 \rangle, E[\widetilde{\tau}] = 0, Var[\widetilde{\tau}] = \frac{2(2n+5)}{9n(n-1)}$$

a

$$Z_{\widetilde{\tau}} = \frac{\widetilde{\tau}}{\sqrt{Var[\widetilde{\tau}]}} \sim N(0,1)$$

Vzťah Kendallovho a Pearsonovho korelačného koeficientu
Ak $(X, Y) \sim N_2(\mu, \Sigma)$ a $\rho_{X,Y}$, potom $\tau = \frac{2}{\pi} \arcsin(\rho_{X,Y})$, kde arcsin nadobúda hodnoty z $\langle -\frac{\pi}{2}, \frac{\pi}{2} \rangle$

Vzťah Kendallovho korelačného koeficientu a Jonckheere štatistiky

$$\tau = \frac{\widetilde{S}_J}{\sum_{i=1}^{k-1} \sum_{j=1+i}^{k} n_i n_j}, \tau \in \langle -1, 1 \rangle,$$

kde τ nazývame *zovšeobecnený Kendallov korelačný koeficient*

Example

Majme pacientov s akútnou myeloidnou leukémiou (AML) a v rámci nich skupinu AG-pozitívnych (výskyt určitých špecifických indikátorov choroby v kostnej dreni). Pre chorobu je charakteristické, že s počtom bielych krviniek (white blood cells counts, WBC) vzrastná závažnosť choroby. Nech $t_i, i = 1, 2, \ldots, 17$ sú časy do zlyhania v týždňoch prislúchajúce zoradeným WBC (pozri tab.). Vypočítajte τ, $\widetilde{\tau}$ a R_S. Otestujte nezávislosť medzi počtom bielych krviniek a časmi do zlyhania.

WBC	t_i	C_{ij}	D_{ij}
750	156	0	16
2300	65	5	9
2600	134	1	13
4300	100	3	10
5400	39	5	7
6000	16	7	4
7000	143	0	10
9400	56	3	6
10000	121	0	8
10500	108	0	7
17000	4	4	2
32000	26	1	4
35000	22	1	3
52000	5	1	2
100000	1	1	0
100000	1	1	0
100000	65	0	0

$C_{ij} = \#(\uparrow t_i, \uparrow WBC_i)$ pod i
$C = \sum C_{ij} = 33$
$D_{ij} = \#(\downarrow t_i, \uparrow WBC_i)$ pod i
$D = \sum D_{ij} = 101$
$\frac{n(n-1)}{2} = \frac{17 \times 16}{2}$
$\widetilde{\tau} = \frac{C-D}{\frac{n(n-1)}{2}} = -0.5$
$Var[\widetilde{\tau}] = \frac{2(2 \times 17+5)}{9 \times 17(17-1)} = 0.032$
$Z_{\widetilde{\tau}} = -0.5/\sqrt{0.032} = -2.801$
p-hodnota=0.002

Označenia

- nech R_{ij} je poradie ij-teho pozorovania v združenom NV
- nech z_{ij} je *skóre* prislúchajúce NV, do ktorého ij-te pozorovanie patrí
- $n = \sum_{j=1}^{k} n_j$

Cuzickova štatistika

$$S_C = \sum_{i=1}^{n_i} \sum_{j=1}^{k} z_{ij} R_{ij},$$

Stredná hodnota

$$E_0\left[S_C\right] = \left(\sum_{i=1}^{n} i\right) E\left[Z\right] = \frac{1}{2} n\left(n+1\right) E\left[Z\right],$$

kde $E\left[Z\right] = \sum_{j=1}^{k} z_j p_j$, k je počet skupín, $z_{ij} = z_j = j$, $p_j = n_j/n$

Rozptyl

$$Var\left[S_C\right] = \left[\frac{n^2\left(n+1\right)}{12}\right] Var\left[Z\right],$$

kde $Var\left[Z\right] = \sum_{j=1}^{k} z_j^2 p_j - \left(E\left[Z\right]\right)^2$

Cuzickov test

$$Z_C = \frac{S_C - E_0\left[S_C\right]}{\sqrt{Var_0\left[S_C\right]}} \sim N\left(0,1\right)$$

Označenia

- nech $\left(X_1, Y_1\right)^T, ..., \left(X_n, Y_n\right)^T$ je výber z dvojrozmerného rozdelenia
- nech $R_1, ..., R_n$ sú poradia veličín $X_1, ..., X_n$
- nech $Q_1, ..., Q_n$ sú poradia veličín $Y_1, ..., Y_n$

Spearmanova štatistika

$$S_N = \sum_{i=1}^{n} R_i Q_i$$

Stredná hodnota

$$E\left[S_N\right] = n\left(\frac{n+1}{2}\right)^2$$

Rozptyl

$$Var\left[S_N\right] = \frac{1}{n-1}\left(\frac{n(n^2-1)}{12}\right)^2$$

Spearmanov test

$$\sqrt{n-1} R_S \sim N(0,1),$$

kde $R_S = \frac{1}{\sqrt{(n-1) Var[S_N]}}\left(S_N - E\left[S_N\right]\right)$, $E\left[R_S\right] = 0$, $Var\left[R_S\right] = \frac{1}{n-1}$

Testy na porovnanie viacerých kriviek prežívania
Zovšeobecnený Spearmanov korelačný koeficient

Vzťah Spearmanovho a Pearsonovho korelačného koeficientu
Ak $(X_1, Y_1)^T, ..., (X_n, Y_n)^T$ je výber z dvojrozmerného normálneho rozdelenia s korelačným koeficientom $\rho_{X,Y}$, potom platí

$$\rho_{X,Y} = 2\sin\left(\frac{\pi}{6}\right) R_S$$

Vzťah Spearmanovho korelačného koeficientu a Cuzickovej štatistiky
Cuzickova štatistika S_C je rovná Spearmanovej štatistike S_N, kde jedna premenná predstavuje zoradenú (ordinálnu) premennú a druhá spojitú premennú

Testy na porovnanie viacerých kriviek prežívania
Le test

Označenia

- n_i je rozsah i-teho NV
- $L_i = \sum_{j<i} n_j = \#$ pozorovaní vo všetkých výberoch naľavo od i-teho výberu, $L_1 = 0$
- $M_i = \sum_{j>i} n_j = \#$ pozorovaní vo všetkých výberoch napravo od i-teho výberu, $M_k = 0$
- $L_i - M_i \in \langle -n, n \rangle$
- $\overline{R}_i$ je priemerné poradie pre i-ty výber

Le štatistika

$$S_L = \sum_{i=1}^{k} n_i \left(L_i - M_i\right) \overline{R}_i$$

Testy na porovnanie viacerých kriviek prežívania
Le test

Stredná hodnota

$$E\left[S_L\right] = 0$$

Rozptyl

$$Var\left[S_L\right] = \frac{n\left(n+1\right)}{12} \sum_{i=1}^{k} n_i \left(L_i - M_i\right)^2$$

Le test

$$Z_L = \frac{S_L - E_0\left[S_L\right]}{\sqrt{Var_0\left[S_L\right]}} \sim N\left(0, 1\right)$$

Testy na porovnanie viacerých kriviek prežívania
Zovšeobecnenia

Test odklonu od trendu

$$S_{KW} - \frac{S_L^2}{Var\left[S_L\right]} \sim \chi^2_{k-2}$$

Všeobecný tvar testovacej štatistiky

$$S_A = \sum_{i=1}^{k} n_i s_i r_i,$$

kde n_i sú rozsahy jednotlivých NV, s_i skóre prislúchajúce jednotlivým NV a r_i sú priemerné charakeristiky polohy prislúchajúce jednotlivým NV
Potom

$$\frac{\left(S_A - E_0[S_A]\right)^2}{Var_0\left[S_A\right]} \sim \chi^2_1$$

Example (pokrač.)

Nech $t_{ij}, i = 1, ..., n_j, j = 1, 2$ sú časy do zlyhania (úmrtia) od diagnostiky nádoru pľúc v mesiacoch, kde $j = 1$ predstavuje I. typ terapie a $j = 2$ zasa II. typ terapie, nech $\psi_{ij} = 0$ ak $j = 1$ a $\psi_{ij} = 1$ ak $j = 2$ (pozri tab.). Otestujte $H_0 : S_1(t) = S_2(t)$, alternatíva $H_1 : S_1(t) \neq S_2(t)$. Použite aj S_{KW}, S_J, S_C. Vždy presne naformulujte H_1.

t_{ij}	52	240	19	53	15	43	340	133	111	231	378	49
ψ_{ij}	0	1	1	0	0	1	1	0	0	1	0	0

$S_{KW} = S_W$
$S_J = S_{MW}$
$S_{MW} = S_C$

Example (pokrač.)

Majme pacientov s akútnou myeloidnou leukémiou (AML) a v rámci nich skupinu AG-pozitívnych (výskyt určitých špecifických indikátorov choroby v kostnej dreni). Pre chorobu je charakteristické, že s počtom bielych krviniek (white blood cells counts, WBC) vzrastná závažnosť choroby. Nech $t_i, i = 1, 2, ..., 17$ sú časy do zlyhania v týždňoch (pozri tab.). Otestujte $H_0 : S_1(t) = S_2(t) = S_3(t)$, alternatíva a) $H_1 : S_i(t) \neq S_j(t), i < j; i, j = 1, 2, 3$, b) $S_1(t) \overset{st}{\succ} S_2(t) \overset{st}{\succ} S_3(t)$, c) $S_1(t) \overset{st}{\prec} S_2(t) \overset{st}{\prec} S_3(t)$.

Rozdeľme AG-pozitívnych pacientov do troch skupín nasledovne

- Skupina 1: WBC ≥ 100000, $n_1 = 3$, (1, 1, 65)
- Skupina 2: WBC $\in \langle 10000, 100000)$, $n_2 = 6$, (108, 121, 4, 26, 22, 5),
- Skupina 3: WBC < 10000, $n_3 = 8$, (65, 156, 100, 134, 16, 39, 143, 56)

sk.1	1	1						
sk.2			4	5		22	26	
sk.3					16			39
poradie	1.5	1.5	3	4	5	6	7	8

sk.1		65						
sk.2				108	121			
sk.3	56	65	100			134	143	156
poradie	9	10.5	12	13	14	15	16	17

$\overline{R}_1 = 4.50$, $\overline{R}_2 = 7.8\overline{33}$, $\overline{R}_3 = 11.5625$

$S_{KW} = \frac{12}{n(n+1)} \sum_{i=1}^{k} \frac{R_i^2}{n_i} - 3(n+1) = \frac{12}{17 \times 18} \left[3 \times 4.5^2 + 6 \times 7.8\overline{33}^2 + 8 \times 11.5625^2\right] - 3 \times 18 = 4.762662$, p-hodnota$= 0.0924$

$S_L = \sum_{i=1}^{3} n_i (L_i - M_i) \overline{R}_i =$
$3 \times (0-14) \times 4.5 + 6 \times (3-8) \times 7.8\overline{33} + 8 \times (9-0) \times 11.5625 = 408.5$

$Var[S_L] = \frac{n(n+1)}{12} \sum_{i=1}^{k} n_i (L_i - M_i)^2 = \frac{17(17+1)}{12} \left[3 \times (0-14)^2 + 6 \times (3-8)^2 + 8 \times (9-0)^2\right] = 187.9973^2$

$Z_L = 408.5/187.9973 = 2.172903$, p-hodnota=0.0298

$S_{KW} - (Z_L)^2 = 4.762662 - 2.172903^2 = 0.0412$, p-hodnota=0.8392

Formulácie sčítacím procesom

- (X_i, δ_i) nahradíme $(N_i(t), Y_i(t))$, kde $N_i(t)$ je počet pozorovaných udalostí v intervale $\langle 0, t\rangle$ v jednotke i,

$$Y_i(t) = \begin{cases} 1 & \text{jednotka } i \text{ je v riziku v čase } t \\ 0 & \text{inak} \end{cases}$$

- $Y_i(t) = I(\{T_i \geq t\})$ a $N_i(t) = I(\{T_i \leq t,\ \delta(x) = 1\})$
- agregovaný proces $\overline{Y}(t) = \sum_i Y_i(t)$, $\overline{N}(t) = \sum_i N_i(t)$, $d\overline{N}(t) = \Delta\overline{N}(t) = \overline{N}(t) - \overline{N}(t^-)$, kde $\overline{N}(t)$ je suma udalostí do času t vrátane, $\overline{Y}(t)$ je počet jednotiek v riziku v čase t

Testovacia štatistika na porovnanie dvoch kriviek prežívania

$$T(W,t) = \int_0^t W(s)\frac{\overline{Y}_1(s)\overline{Y}_2(s)}{\overline{Y}_1(s)+\overline{Y}_2(s)}\left(\frac{d\overline{N}_1(s)}{\overline{Y}_1(s)} - \frac{d\overline{N}_2(s)}{\overline{Y}_2(s)}\right)ds,$$

kde $W(s)$ je funkcia váh a $s \geq 0$

- **stredná hodnota** $E_0[T(W,t)] = 0$
- **rozptyl**

$$\begin{aligned} Var_0[T(W,t)] &= \sum_{i=1}^{2}\int_0^t \frac{W^2(s)}{\overline{Y}_i(s)}\left(\frac{\overline{Y}_1(s)\overline{Y}_2(s)}{\overline{Y}_1(s)+\overline{Y}_2(s)}\right)^2 \\ &\times \left(1 - \frac{\Delta\overline{N}_1(s)+\Delta\overline{N}_2(s)-1}{\overline{Y}_1(s)+\overline{Y}_1(s)-1}\right) \\ &\times \left(\frac{\Delta\overline{N}_1(s)+\Delta\overline{N}_2(s)}{\overline{Y}_1(s)+\overline{Y}_1(s)}\right)ds \end{aligned}$$

Potom asymptoticky

$$\frac{T(W,t) - E_0[T(W,t)]}{\sqrt{Var_0[T(W,t)]}} \sim N(0,1)$$

Tarone a Ware (1977) trieda váh

1. konštantný rozdiel v **logitovej** škále $f(p) = \ln\frac{p}{1-p}$, potom váhy $W(t) = 1$
2. konštantný rozdiel v **aritmetickej** škále: $f(p) = p$, potom váhy $W(t) = (\frac{1/\overline{Y}(t)}{1-1/\overline{Y}(t)})^{-1} = \overline{Y}^2(t)/(\overline{Y}(t)-1) \approx \overline{Y}(t)$
3. konštantný rozdiel v **arcsin** škále: $f(p) = \arcsin\sqrt{p}$, potom sú váhy rovné $W(t) = \frac{\overline{Y}(t)}{\sqrt{\overline{Y}(t)-1}} \approx \sqrt{\overline{Y}(t)}$ kde $p_t = 1/\overline{Y}(t)$ a $\overline{Y}(t)$ počet osôb v riziku v združenom výbere v čase t

Vo všeobecnosti môžeme váhy zapísať ako $W(t) = g(\overline{Y}(t)/n)$

Harrington a Fleming (1982) trieda váh $W(t) = \widehat{S}^{\rho}(t), \rho \geq 0$

1. $\rho = 0$, a teda $W(t) = \widehat{S}^{0}(t) = 1$ (**Cox-Mantel test** alebo **log-rank test**; Cox, 1972; Mantel, 1966)
2. $\rho = 1$, a teda $W(t) = \widehat{S}^{1}(t) = \widehat{S}(t)$ (**Gehan-Wilcoxon test** alebo **Peto-Peto-Wilcoxon test**, Gehan, 1965; Peto a Peto, 1972)

Formálna formulácia
Testovacia štatistika na porovnanie dvoch kriviek prežívania

$$T(W,t) = \sum_{j=0}^{t} W_j \left(d_{1j} - d_j \frac{n_{1j}}{n_k} \right),$$

potom

- **stredná hodnota** $E_0[T(W,t)] = 0$
- **rozptyl**

$$Var_0[T(W,t)] = \sum_{j=0}^{t} W_j^2 \frac{n_{1j}}{n_j} \left(1 - \frac{n_{1j}}{n_j}\right) \frac{d_j(n_j - d_j)}{n_j - 1}$$

Označenia

- majme dva náhodné výbery s rozsahmi n_1 a n_2
- prvý výber má r_1 cenzúr a $n_1 - r_1$ zlyhaní
- druhý výber r_2 cenzúr a $n_2 - r_2$ zlyhaní
- realizácie

$$\underbrace{x_{j1}^{(c)}, x_{j2}^{(c)}, \ldots, x_{jr_j}^{(c)}}_{\text{cenzúry}} \underbrace{x_{jr_j+1}, x_{jr_j+2}, \ldots, x_{jn_j}}_{\text{zlyhania}}, j = 1, 2$$

Testovacia štatistika $S_W^{(c)} = \sum_{i,j} U_{ij}$, kde

$$U_{ij} = \begin{cases} -1, & \text{ak } x_{1i} < x_{2j} \vee x_{1i} \leq x_{2j}^{(c)} \\ 0, & \text{ak } x_{1i} = x_{2j} \vee x_{1i}^{(c)} = x_{2j}^{(c)} \vee x_{1i}^{(c)} < x_{2j} \vee x_{1i} > x_{2j}^{(c)} \\ +1, & \text{ak } x_{1i} > x_{2j} \vee x_{1i}^{(c)} \geq x_{2j} \end{cases},$$

kde sumujeme cez všetkých $n_1 n_2$ porovnávaní

Ak nemáme prítomné cenzúry, $S_W^{(c)}$ sa redukuje na **Wilcoxonovu štatistiku** S_W, **Mann-Whitneho štatistiku** S_{MW} a **Kendallovo** τ nasledovne

$$S_W^{(c)} = n_2(n_1 + n_2 + 1) - 2S_W = 2S_{MW} - n_1 n_2 = \tau,$$

kde S_W je suma poradí druhého výberu v združenom výbere, $S_{MW} = \#(x_{1i}, x_{2j})$, $x_{1i} < x_{2j}$ a posledná rovnosť platí aj v prítomnosti zhôd
$S_W^{(c)}$ je možné zjednodušit ak všetky cenzúry nastali v čase t_n (Halperin, 1960), teda

$$S_W^{(c)} = 2S^{(z)} + r_1 r_2 - n_1 n_2,$$

kde $S^{(z)} = S_{MW}^{(z)} - r_1(n_2 - r_2)$, $S_{MW}^{(z)}$ je Mann-Whitney štatistika počítaná len na základe $(n_1 - r_1) + (n_2 - r_2)$ zlyhaní

Testy na porovnanie dvoch kriviek prežívania
Zovšeobecnený Wilcoxon test

Za platnosti H_0 platí

- **stredná hodnota** $E_0[S_W^{(c)}] = 0$
- **rozptyl**

$$Var_0[S_W^{(c)}] = \frac{n_1 n_2}{n(n-1)} \left\{ \sum_{j=1}^{L} d_j D_{j-1}(D_{j-1}+1) + \sum_{j=1}^{L} l_j D_j(D_j+1) + \sum_{j=1}^{L} \left(d_j(n - D_j - L_{j-1})(n - 3D_{j-1} - d_i - L_{j-1} - 1) \right) \right\},$$

kde $n_1 + n_2 = n, D_i = \sum_{j=1}^{i} d_j, D_0 = 0, L_i = \sum_{j=1}^{i} l_j, L_0 = 0$, d_j je počet necenzúrovaných pozorovaní (zlyhaní) s poradím j (R_j) v združenom výbere zlyhaní, l_j je počet cenzúr z intervalu (R_j, R_{j+1})

Testy na porovnanie dvoch kriviek prežívania
Zovšeobecnený Wilcoxon test

Ak **nemáme žiadne zhody**, $d_1 = \ldots = d_L = 1$, ako aj žiadne cenzúry, $l_1 = \ldots = l_L = 0$ ($L = n$), potom

$$Var_0[S_W^{(c)}] = \frac{1}{3} n_1 n_2 (n+1)$$

Ak **nemáme zhody a všetky cenzúry sa objavili po** $L = (n_1 - r_1) + (n_2 - r_2)$ **zlyhaniach**, teda $d_1 = \ldots = d_L = 1, l_1 = \ldots = l_{L-1} = 0, l_L = r_1 + r_2$.

$$Var_0[S_W^{(c)}] = \frac{n_1 n_2 (n - r_1 - r_2)}{n(n-1)} \left(n(r_1 + r_2) + \frac{1}{3}\left((n - r_1 - r_2)^2 - 1 \right) \right)$$

Ak **máme len zlyhania, ale vyskytujú sa zhody**, dostaneme (Hemelrijk, 1952)

$$Var_0[S_W^{(c)}] = Var_0[S_{MW}|\mathbf{t}] = Var_0\left[S_{MW}\right] - \frac{n_1 n_2 \sum_j t_j \left(t_j^2 - 1\right)}{12\left(n_1 + n_2\right)\left(n_1 + n_2 - 1\right)}$$

Testy na porovnanie dvoch kriviek prežívania
Zovšeobecnený Wilcoxon test vs. kontingenčné tabuľky

Majme špeciálny prípad, kde

- $d_1 = d, l_1 = l, d_j = l_j = 0$ pre $i \neq 0$
- pozorovania usporiadajme do **kontingenčnej tabuľky (KT)** 2×2 s fixovanými marginálnymi početnosťami ako výsledok náhodného výberu z **hypergeometrického rozdelenia**
- následne sa testovacia štatistika $S_W^{(c)}$ redukuje na **rozdiel súčinov diagonálnych a mimodiagonálnych prvkov**, čo je v úzkom vzťahu ku **pomeru šancí**
- **rozptyl** $Var_0[S_W^{(c)}] = \frac{l d n_1 n_2}{n_1 + n_2 - 1}$

Testy na porovnanie dvoch a viac kriviek prežívania
Kontingenčné tabuľky

KT 2×2

	y_1	y_2	$\sum$
x_1	a	b	$n_{1\cdot}$
x_2	c	d	$n_{2\cdot}$
$\sum$	$n_{\cdot 1}$	$n_{\cdot 2}$	1

početnosti $n_{j\cdot}, n_{\cdot j}, j = 1, 2$ sa nazývajú marginálne početnosti a sú v tomto prípade fixované

χ^2 **test nezávislosti** (alebo **homogenity**) pre KT 2×2

$$\chi^2 = \left(\frac{a - E_0[A]}{\sqrt{Var_0[A]}} \right)^2 \sim \chi^2_{df}, df = 1,$$

kde a je početnosť v prvej bunke KT

Kombinovanie L jednoduchých KT (Gart, 1970; Cox, 1972) v L časoch zlyhania do **mnohorozmernej (L-rozmernej) KT**

- pre **dvojvýberový prípad** je KT $(2 \times 2) \times L$
- pre k**-rozmerný prípad** je KT $(2 \times k) \times L$

Použitý χ^2 test porovnania nezávislých kriviek prežívania bude potom formálne identický s **Birch-Armitage štatistikou asociácie týchto KT** (Mantel, 1966; Birch, 1965; Armitage, 1966)

χ^2 test pre kombináciu L KT $2 \times k$ (Mantel a Haenszel, 1959)

$$\chi^2 = \left(\frac{\sum_{i=1}^{L} (a_i - E_0[A_i])}{\sqrt{\sum_{i=1}^{L} Var_0[A_i]}} \right)^2 \sim \chi^2_{df}, df = k - 1$$

Pre každé $t_{(i)}$, $1 \leq i \leq I$, môžeme dáta zapísať do KT 2×2

výber/status	zlyhanie v $t_{(i)}$	naživе v $t_{(i)}$	spolu v $t_{(i)}$
1	d_{1i}	a_{1i}	n_{1i}
2	d_{2i}	a_{2i}	n_{2i}
spolu	d_i	a_i	n_i

- $n_{1i} = \#$ subjektov v prvom NV, ktorí boli v riziku tesne pred časom $t_{(i)}$, $n_{2i} = \#$ subjektov v druhom NV, ktorí boli v riziku tesne pred časom $t_{(i)}$, $n_i = n_{1i} + n_{2i}$
- $d_{1i} = \#$ zlyhaní z prvého NV, $d_{2i} = \#$ zlyhaní z druhého NV, $d_i = d_{1i} + d_{2i}$
- $a_i = n_i - d_i = a_{1i} + a_{2i} = \#$ subjektov, ktorí ostali nažive v čase $t_{(i)}$
- $\#$ zlyhaní do času $t_{(i)}$ vrátane $d = \sum_{j:t_j \leq t_{(i)}} d_j$

Testované hypotézy

$$H_0 : \lambda_1(t) = \lambda_2(t), H_1 : \lambda_1(t) = \theta\lambda_2(t),$$

$$H_0 : S_1(t) = S_2(t), H_1 : S_1(t) = [S_2(t)]^\theta$$

kde

- $\lambda(t)$ je riziko v čase t
- θ neznáma konštanta proporcionality rizík

Ak $\theta < 1$, liečba 1 je efektívnejšia ako liečba 2, naopak v prípade $\theta > 1$

- marginálne početnosti v tab., n_{1i}, n_{2i} a d_i sú náhodné premenné závislé iba na minulosti pred časom $t_{(i)}$
- Mantel a Haenzel (1959): **rozdelenie pozorovaní (realizácií) v bunkách KT podmienené pozorovanými marginálnymi početnosťami** $(d_i, a_i, n_{1i}, n_{2i})$ za platnosti H_0
- to implikuje **rozdelenie iba jednej bunky**, d_{1i}, pretože ostatné početnosti sú ľahko odvoditeľné od marginálnych
- za platnosti nulovej hypotézy, H_0, rozdelenie d_{1i} je *hypergeometrické*, teda

$$\Pr(d_{1i}) = \frac{\binom{n_{1i}}{d_{1i}}\binom{n_{2i}}{d_i - d_{1i}}}{\binom{n_i}{d_i}}$$

- v tejto forme H_0 o rovnosti kriviek prežívania implikuje **nezávislosť výberu a statusu (nažive alebo zlyhanie)**

Za platnosti H_0

- **očakávaná (stredná) hodnota** $E_0[d_{1i}] = n_{1i}\frac{d_i}{n_i}$
- **rozptyl**

$$Var_0[d_{1i}] = \left[n_{1i}\frac{d_i}{n_i}\left(1-\frac{d_i}{n_i}\right)\right]\left(\frac{n_i-n_{1i}}{n_i-1}\right) = \frac{n_{1i}n_{2i}a_id_i}{n_i^2(n_i-1)}$$

Informáciu o KT v čase $t_{(i)}$ nám dá nasledovný vzťah

$$\chi_i^2 = \frac{[d_{1i}-E_0[d_{1i}]]^2}{Var_0[d_{1i}]} \sim \chi^2_{df}, df = 1$$

Pre všetky tabuľky ($i = 1, 2, ..., I$) píšeme

$$U = \sum_{i=1}^{I}(d_{1i} - E_0[d_{1i}]),$$

$$E_0[U] = 0, Var_0[U] = \sum_{i=1}^{I} Var_0[d_{1i}] = \sum_{i=1}^{I}\frac{n_{1i}n_{2i}a_id_i}{n_i^2(n_i-1)}$$

Ak máme fixované d_i, n_{1i}, n_{2i}, potom platí (Mantel a Haenzel, 1959)

$$Q_{MH} = \frac{\left[\sum_{i=1}^{I}(d_{1i}-E_0[d_{1i}])\right]^2}{\sum_{i=1}^{I} Var_0[d_{1i}]} = \frac{U^2}{Var_0[U]} \sim \chi^2_{df}, df = 1,$$

$$Q = \frac{(U-E_0[U])^2}{Var_0[U]} \sim \chi^2_1, Z = \frac{U-E_0[U]}{\sqrt{Var_0[U]}} \sim N(0,1)$$

Example

Majme dáta z klinickej štúdie zhrnuté v nasledovnej tabuľke.

t_i	n_i	d_i	n_{1i}	d_{1i}	$E_0[d_{1i}]$	$d_{1i} - E_0[d_{1i}]$	$Var_0[d_{1i}]$
3	10	1	5	1	0.50	0.50	0.2500
5	9	1	4	1	0.44	0.56	0.2469
7	8	1	3	1	0.38	0.62	0.2344
12	6	1	1	0	0.17	−0.17	0.1389
18	5	1	1	1	0.20	0.80	0.1600
19	4	1	0	0	0.00	0	0
20	3	1	0	0	0.00	0	0
suma				4	1.69	2.31	1.0302

$\chi^2 = 2.31^2/1.0302 = 5.179674$
p-hodnota=0.0229
$z = \sqrt{2.31^2/1.0302} = 2.275890$
p-hodnota=$2 \times 0.0114 = 0.0229$

Majme váhy w_i asociované s KT v čase t_i, potom

$$U = \sum_{i=1}^{I} w_i(d_{1i} - E_0[d_{1i}]),$$

$$E_0[U] = 0, Var_0[U] = \sum_{i=1}^{I} w_i^2 Var_0[d_{1i}] = \sum_{i=1}^{I} w_i^2\frac{n_{1i}n_{2i}a_id_i}{n_i^2(n_i-1)}$$

Ak máme fixované d_i, n_{1i}, n_{2i}, potom platí (Mantel a Haenzel, 1959)

$$Q = \frac{(U-E_0[U])^2}{Var_0[U]} \sim \chi^2_1, Z = \frac{U-E_0[U]}{\sqrt{Var_0[U]}} \sim N(0,1)$$

Podľa výberu váh w_i rozoznávame nasledovné typy testov:

- ak $w_i = n_i$, $Q = Q_G$, ide o **Gehan-Wilcoxon test** (**zovšeobecnený Wilcoxonov test**; Gehan, 1965), ktorý môžeme zredukovať na Wilcoxonovu štatistiku pri absencii cenzúr
- ak $w_i = 1$, $Q = Q_{MH}$, ide o **Cox-Mantel test** (**log-rank test**; Mantel a Haenzel, 1959)
- ak $w_i = \sqrt{n_i}$, $Q = Q_{TW}$, ide o **Tarone-Ware test** (Tarone a Ware, 1977)
- ak $w_i = \widehat{S}_{pooled}(t_i) = \prod_{j:t_j \leq t_i} \frac{n_j - d_j + 1}{n_j + 1}$, $Q = Q_P$, ide o **Peto-Peto test** (Peto a Peto, 1972; Prentice, 1978).

Za platnosti alternatívnej hypotézy $H_1 : \lambda_1(t_i) \neq \lambda_2(t_i)$, pre $\forall i$, očakávaná hodnota rozdielu

$$E[d_{1i} - E_0[d_{1i}]] = \tau_i \left(\frac{\lambda_1(t_i)}{\lambda_2(t_i)} - 1 \right), \tau_i = \frac{n_{1i} n_{2i}}{n_i \left(n_{1i} \frac{\lambda_1(t_i)}{\lambda_2(t_i)} + n_{2i} \right)},$$

a suma bude

$$NC_{parameter} = \sum_i \tau_i \left(\frac{\lambda_1(t_i)}{\lambda_2(t_i)} - 1 \right),$$

U čo nie je očakávaná hodnota sumy $\sum_i [d_{1i} - E_0[d_{1i}]]$, no Schoenfeld (1981) ukázal, že ide asymptoticky o *parameter necentrality* **normálneho rozdelenia**

Testované hypotézy

$$H_0 : \lambda_1(t) = \lambda_2(t) = \ldots = \lambda_k(t)$$

$$H_1 : \exists \text{ aspoň jedno } i < j, \lambda_i(t) \neq \lambda_j(t)$$

Pre každé $t_{(i)}, 1 \leq i \leq I$, môžeme dáta zapísať do KT $2 \times k$

status/výber	1	2	...	j	...	k	spolu
zlyhanie v $t_{(i)}$	d_{1i}	d_{2i}	...	d_{ji}	...	d_{ki}	d_i
nažive v case $t_{(i)}$	a_{1i}	a_{2i}	...	a_{ji}	...	a_{ki}	a_i
v riziku pred časom $t_{(i)}$	n_{1i}	n_{2i}	...	n_{ji}	...	n_{ki}	n_i

$a_i = \sum_j a_{ji} = n_i - d_i$, $a_{ji} = n_{ji} - d_{ji}$
$n_i = \sum_j n_{ji}$
$d_i = \sum_j d_{ji}$

Za platnosti nulovej hypotézy a fixovaných marginálnych početnostiach sa dá ukázať, že počet zlyhaní v k výberoch má **hypergeometrické rozdelenie s dimenziou** $k - 1$

$$U_j = \sum_{i=1}^{I} [d_{ji} - E_0[d_{ji}]], j = 1, 2, \ldots, k,$$

stredné hodnoty v časoch $t_{(i)}$

$$E_0[d_{ji}] = n_{ji} \frac{d_i}{n_i},$$

kovariancie v časoch $t_{(i)}$

$$Cov(d_{li}, d_{si}) = \begin{cases} \frac{n_{li}(n_i - n_{li}) d_i a_i}{n_i^2 (n_i - 1)} & \text{pre } l = s \\ -\frac{n_{li} n_{si} d_i a_i}{n_i^2 (n_i - 1)} & \text{pre } l \neq s \end{cases},$$

$j = 1, 2, \ldots, k - 1$

Testy na porovnanie viacerých kriviek prežívania
Prehľad testov

U_j substituujeme $(k-1)$-stĺpcovým vektorom $\mathbf{U}$ s elementami $U_j, j = 1, 2, ..., k-1$
$\mathbf{V}\left(t_{(i)}\right)$ je $(k-1) \times (k-1)$ **kovariančná matica** s komponentami v časoch $t_{(i)}$ daných nasledovne

$$\left(\mathbf{V}\left(t_{(i)}\right)\right)_{ls} = Cov\left(d_{li}, d_{si}\right)$$

$$\mathbf{V} = \sum_{i=1}^{I} \mathbf{V}\left(t_{(i)}\right)$$

Testovacia štatistika

$$Q_{overall} = \mathbf{U}^T \mathbf{V}^{-1} \mathbf{U} \sim \chi^2_{k-1}$$

Ak $k = 2$, štatistika $Q_{overall} = Q_{MH}$

Testy na porovnanie viacerých kriviek prežívania
Prehľad testov

V prípade pridania váh, bude platiť

$$\mathbf{U}_w = \sum_{i=1}^{I} w_i \mathbf{U}_i,$$

kde w_i sú váhy asociované s pozorovaním v čase $t_{(i)}$

$$\mathbf{U}_i = \begin{pmatrix} d_{1i} - E_0\left[d_{1i}\right] \\ \vdots \\ d_{ji} - E_0\left[d_{ji}\right] \\ \vdots \\ d_{ki} - E_0\left[d_{ki}\right] \end{pmatrix}$$

Podobne ako vyššie bude platiť

$$\mathbf{U}_w^T \mathbf{V}_w^{-1} \mathbf{U}_w \sim \chi^2_{k-1},$$

kde $\mathbf{V}_w = \mathbf{w}^T \mathbf{V} \mathbf{w}$

Testy na porovnanie viacerých kriviek prežívania
Prehľad testov

Voľba váh je nasledovná:

- ak $w_i = n_i$, ide o **zovšeobecnený Wilcoxonov test (zovšeobecnený Kruskal-Wallis test, Gehan-Breslow test)**,
- ak $w_i = 1$, ide o **Cox-Mantel test (log-rank test),**
- ak $w_i = \hat{S}\left(t_i^-\right)^\rho$, $\rho = 0$ ide o **Mantel-Haenszelov test (log-rank test)**, ak $\rho = 1$, ide o **Peto-Peto-Wilcoxon test**.

Testy na porovnanie viacerých kriviek prežívania
Test trendu

Test nulovej hypotézy oproti stochasticky usporiadanej alternatíve je *testom trendu*, kde testujeme zoradený vzťah medzi k funkciami prežívania definovanými v zmysle vektora váh $\mathbf{w}_* = \left(w_1^*, w_2^*, ..., w_j^*, ..., w_k^*\right)^T$, potom

$$H_0 : \lambda_1(t) = \lambda_2(t) = ... = \lambda_k(t)$$

a H_1 :

$$\begin{aligned} \lambda_1(t) &= w_1^* \lambda_k(t), \\ \lambda_2(t) &= w_2^* \lambda_k(t), \\ &\cdots \\ \lambda_j(t) &= w_j^* \lambda_k(t), \\ \lambda_{k-1}(t) &= w_{k-1}^* \lambda_k(t), \end{aligned}$$

kde bez straty na všeobecnosti môžeme predpokladať, že $w_k = 1$

Testovacia štatistika

$$Q_{trend} = \frac{\left[U_{trend}\right]^2}{\mathbf{w}_*^T \mathbf{V}_* \mathbf{w}_*} \sim \chi_1^2,$$

kde

$$U_{trend} = \sum_{j=1}^{k} U_{trend}^{(j)} = \sum_{j=1}^{k} w_j^* U_j$$
$$= \sum_{j=1}^{k} w_j^* \sum_{i=1}^{l} \left[d_{ji} - E_0\left[d_{ji}\right]\right],$$

a $\mathbf{V}_*$ počítame tak ako $\mathbf{V}$, ale s tým rozdielom, že ide o maticu $k \times k$

Testovaciu štatistiku Q_{trend} budeme potom písať nasledovne

$$Q_{trend} = \frac{\left(\sum_{j=1}^{k} w_j^* \sum_{i=1}^{l}\left[d_{ji} - E_0\left[d_{ji}\right]\right]\right)^2}{\sum_{i=1}^{l} \frac{n_i - d_i}{n_i - 1}\left(\sum_{j=1}^{k}\left(w_j^*\right)^2 E_0\left[d_{ji}\right] - \frac{1}{d_i}\left[\sum_{j=1}^{k}\left(w_j^*\right) E_0\left[d_{ji}\right]\right]^2\right)}$$

a ak sú váhy rozdelené lineárne, $w_j^* = j$, potom hovoríme o *teste lineárneho trendu*
Platí

$$Q_{residual} = Q_{overall} - Q_{trend} \sim \chi_{k-2}^2,$$

čo vedie ku záveru, že ak je tento test štatisticky signifikantný, potom váhy použité v Q_{trend} treba nahradiť inými, ktoré by lepšie modelovali trend v dátach

Example

Majme experiment (Thomas et all, 1977), kde sme mali 3 rôzne koncentrácie látky ($konc_1 = 2.0$, $konc_2 = 1.5$ a $konc_3 = 0$) a hľadali sme jej účinok na pacientov, u ktorých sme sledovali objavenie sa nádoru (pozri Tab.). Zaujíma nás testovanie A) $H_0 : \lambda_1(t) = \lambda_2(t) = \lambda_3(t)$ oproti $H_1 : \exists$ aspoň jedno $i < j, \lambda_i(t) \neq \lambda_j(t)$ B) $H_0 : \lambda_1(t) = \lambda_2(t) = \lambda_3(t)$ oproti $\lambda_1(t) = w_1^* \lambda_3(t), \lambda_2(t) = w_2^* \lambda_3(t)$, kde $w_j^* = konc_j, j = 1, 2$ (test trendu).

Pozn.: časy do zlyhania alebo cenzúry; $^+$ znamená cenzúra, $n_0 = \#$ pozorovaní v t_0, $konc_j$ je koncentrácia látky v skupine j,

$konc_j$	n_0										
2.0	10	41^+	41^+	47	47^+	47^+	58	58	58	100^+	117
1.5	10	43^+	44^+	45^+	67	68^+	136	136	150	150	150
0	9	73^+	74^+	75^+	76	76	76^+	99	166	246^+	

d_{ji}	47	58	67	76	99	117	136	150	166
konc. 1	1	3	0	0	0	1	0	0	0
konc. 2	0	0	1	0	0	0	2	3	0
konc. 3	0	0	0	2	1	0	0	0	1

a_{ji}	47	58	67	76	99	117	136	150	166
konc. 1	7	2	2	2	2	0	0	0	0
konc. 2	7	7	6	5	5	5	3	0	0
konc. 3	9	9	9	4	2	2	2	2	1

n_{ji}	47	58	67	76	99	117	136	150	166
konc. 1	8	5	2	2	2	1	0	0	0
konc. 2	7	7	7	5	5	5	5	3	0
konc. 3	9	9	9	6	3	2	2	2	2

Testy na porovnanie viacerých kriviek prežívania
Príklad

Prvý čas, kedy sa nádor objavil je $t_{(1)} = 47$ a KT v $t_{(1)}$ je nasledovná

dávka	udalosť v $t_{(1)}$	bez udalosti v $t_{(1)}$	v riziku pred $t_{(1)}$
2.0	1	7	8
1.5	0	7	7
0.0	0	9	9
spolu	1	23	24

Čiastkové výpočty v $t_{(1)}$

dávka	$E_0\,[d_{j1}]$	U_{j1}
2.0	0.33333	$1 - 0.33333 = 0.66667$
1.5	0.29167	$0 - 0.29167 = -0.29167$
0.0	0.37500	

Testy na porovnanie viacerých kriviek prežívania
Príklad

$$\mathbf{V}\left(t_{(1)}\right) = \begin{bmatrix} \frac{8\times(24-8)\times1\times(24-1)}{24^2(24-1)} = 0.222 & -\frac{8\times7\times1\times23}{24^2\times23} = -0.097 \\ -0.097 & \frac{7\times(24-7)\times1\times23}{24^2 23} = 0,207 \end{bmatrix}$$

$$\mathbf{U} = \begin{pmatrix} 3.209 \\ -0.803 \end{pmatrix}$$

$$\mathbf{V} = \begin{pmatrix} 1.319 & -0.641 \\ -0.641 & 2.663 \end{pmatrix}, \mathbf{V}^{-1} = \begin{pmatrix} 0.859 & 0.207 \\ 0.207 & 0.425 \end{pmatrix}$$

$Q_{overall} = \mathbf{U}^T\mathbf{V}^{-1}\mathbf{U} = 8.050$, p-hodnota=0.018

Ak $w^* = (0.0, 1.5, 2.0)^T$, potom

$U^2_{trend} = 36.186, \mathbf{w}_*^T\mathbf{V}_*\mathbf{w}_* = 5.991,$

$Q_{trend} = 6.04$, p-hodnota=0.014,

a naviac $Q_{residual} = Q_{overall} - Q_{trend} = 2.01$, p-hodnota=0.156

R príkazy [R je voľne dostupné na http://cran.r-project.org/]

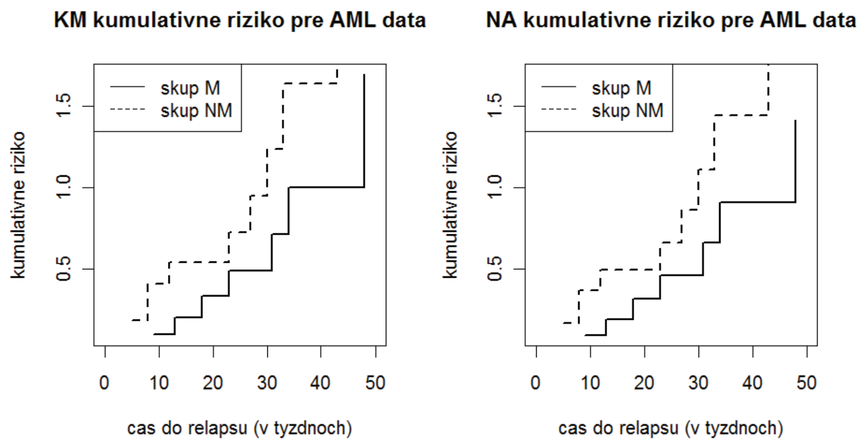
Príklad 1 (poznámky sú uvedené za znakom #):

```
library(survival)
attach(aml)
names(aml)
## odhady pre obe skupiny zvlast
# median, priemerny cas prezivania, rozsahy, pocty zlyhani
kmfit.aml <- survfit(Surv(cas,status)~skupina, data=aml)
print(kmfit.aml,show.rmean=TRUE)
## odhady v jednotlivych casoch
# fcia prezivania,IS, pocet v riziku, pocet udalosti v case
## testovanie hypotez
# default rho = 0 [log-rank]
test.aml<-survdiff(Surv(cas,status)~skupina, data=aml)
# Peto test rho = 1
test.aml<-survdiff(Surv(cas,status)~skupina, data=aml,rho = 1)
```


R príkazy [R je voľne dostupné na http://cran.r-project.org/]

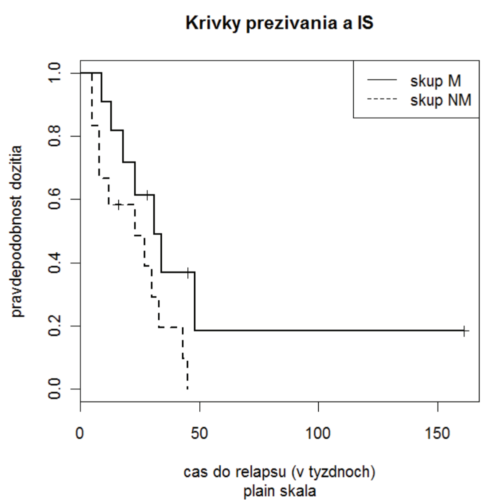
Príklad 1 (pokrač.):

```
skup <- as.numeric(aml$skupina)
cas1 <- kmfit.aml.1$time; cas2 <- kmfit.aml.2$time;
CAS <- c(cas1,cas2)
## kumulativne riziko (KM a NA)
# (pocitane rovnako ako v priklade 2, handout 1)
LAM.KM <- c(Lambda.KM.1,Lambda.KM.2)
par(mfcol=c(1,2)) # nastavenie poctu podokien
## obrazok (limitacia osi-argument xlim,ylim, kde beriem do
## uvahy rozsah zdruzeneho vektora kum.rizik)
## KM odhad kum.rizika (podobne aj NA odhad)
plot(CAS,LAM.KM,xlab="cas do relapsu (v tyzdnoch)",
ylab="kumulativne riziko", type="n", ylim=range(LAM.KM),
xlim=c(0,50))
lines(cas1,Lambda.KM.1,lty=1,lwd=2,type="s")
lines(cas2,Lambda.KM.2,lty=2,lwd=2,type="s")
abline(h=0,col="gray")
title(main="KM kumulativne riziko pre AML data")
legend("topleft",c("skup M","skup NM"),lty=c(1,2))
```



Príklad 1 (pokrač.):

```
## odhady funkcie prezivania pre kazdu skupinu zvlast
kmfit.aml.1<-survfit(Surv(cas[skup == 1],status[skup == 1])~1)
kmfit.aml.2<-survfit(Surv(cas[skup == 2],status[skup == 2])~1)
## obrazok (hrubka ciary lwd, conf.int kresli IS)
plot(kmfit.aml.1,xlab="cas do relapsu (v tyzdnoch)",
ylab="pravdepodobnost dozitia",conf.int=FALSE,lwd=2)
lines(kmfit.aml.2,lty=2,lwd=2)
legend("topright",c("skup M","skup NM"),lty=c(1,2))
title(main="Krivky prezivania a IS",sub="plain skala")
```

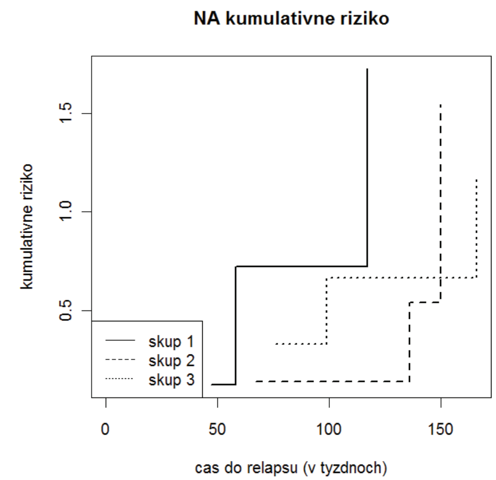



Príklad 2:

```
## odhady funkcie prezivania pre kazdu skupinu zvlast
## data z prednasky
konc1 <- c(41,41,47,47,47,58,58,58,100,117)
status1 <- c(0,0,1,0,0,1,1,1,0,1)
konc2 <- c(43,44,45,67,68,136,136,150,150,150)
status2 <- c(0,0,0,1,0,1,1,1,1,1)
konc3 <- c(73,74,75,76,76,76,99,166,246)
status3 <- c(0,0,0,1,1,0,1,1,0)
data.all <- c(konc1,konc2,konc3)
status.all <- c(status1,status2,status3)
skup.all <- c(rep(1,10),rep(2,10),rep(3,9))
## testovanie hypotez
# default rho = 0 [log-rank]
survdiff(formula = Surv(data.all, status.all)~skup.all)
# Peto test rho = 1
survdiff(formula = Surv(data.all, status.all)~skup.all,rho=1)
```

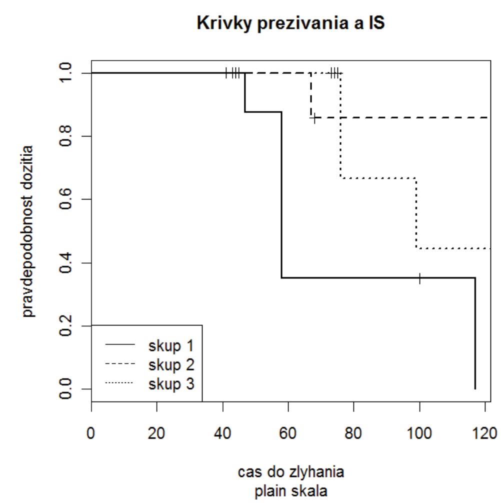
Príklad 2 (pokrač.):

```
## NA kumulativne riziko
# (pocitane rovnako ako v priklade 2, handout 1)
LAM <- Lambda.NA.1
plot(cas,Lambda.NA,xlab="cas do relapsu (v
tyzdnoch)",ylab="kumulativne riziko", type="n", ylim=range(LAM),
xlim=c(0,166))
lines(cas,Lambda.NA.1,lty=1,lwd=2,type="s")
lines(cas,Lambda.NA.2,lty=2,lwd=2,type="s")
lines(cas,Lambda.NA.3,lty=3,lwd=2,type="s")
abline(h=0,col="gray")
title(main="NA kumulativne riziko")
legend("bottomleft",c("skup 1","skup 2","skup 3"),lty=c(1,2,3))
```

Príklad 2 (pokrač.):

```
## odhady funkcie prezivania pre kazdu skupinu zvlast
kmfit.1 <- survfit(Surv(konc1,status1)~1)
kmfit.2 <- survfit(Surv(konc2,status2)~1)
kmfit.3 <- survfit(Surv(konc3,status3)~1)
## obrazok (hrubka ciary lwd, conf.int kresli IS)
plot(kmfit.1,xlab="cas do zlyhania",ylab="pravdepodobnost
dozitia",conf.int=FALSE,lwd=2)
lines(kmfit.2,lty=2,lwd=2)
lines(kmfit.3,lty=3,lwd=2)
legend("bottomleft",c("skup 1","skup 2","skup 3"),lty=c(1,2,3))
title(main="Krivky prezivania a IS",sub="plain skala")
```

Vybrané kapitoly z analýzy prežívania

Regresné modely

Stanislav Katina[1]

[1]Ústav matematiky a statistiky
Přírodovědecká fakulta
Masarykova univerzita v Brne

ZS 2013

INVESTICE DO ROZVOJE VZDĚLÁVÁNÍ

Poďakovanie

Tento učební text vznikl za přispění Evropského sociálního fondu a státního rozpočtu ČR prostřednictvím Operačního programu Vzdělávání pro konkurenceschopnost v rámci projektu Univerzitní výuka matematiky v měnícím se světě

(CZ.1.07/2.2.00/15.0203)

Parametrické regresné modely

Prehľad rozdelení pravdepodobnosti

1. **Exponenciálne rozdelenie**
2. **Weibullovo rozdelenie**
3. **Rozdelenie extrémnych (minimálnych) hodnôt**
4. **Log-normálne rozdelenie**
5. **Log-logistické rozdelenie**
6. **Gama rozdelenie**

Parametrické regresné modely

Exponenciálne rozdelenie

Náhodná premenná T má ***exponenciálne rozdelenie*** $Exp\left(\lambda\right)$ práve vtedy, ak jej **hustota** má tvar

$$f\left(t,\lambda\right)=\lambda e^{-\lambda t},\text{ kde } \lambda>0, t\geq 0$$

a naviac

- **funkcia rizika** $\lambda\left(t,\lambda\right)=\lambda$
- **distribučná funkcia** $F\left(t,\lambda\right)=\int_0^t\lambda e^{-\lambda t}=1-e^{-\lambda t}$
- **funkcia prežívania** $S\left(t,\lambda\right)=e^{-\lambda t}$
- **stredná hodnota** $E\left[T\right]=\frac{1}{\lambda}$
- **rozptyl** $Var\left[T\right]=\frac{1}{\lambda^2}$
- p**-ty kvantil** $t_p=-\frac{1}{\lambda}\log(1-p)$

Parametrické regresné modely
Exponenciálne rozdelenie

- jednoduchosť modelu – **riziko je konštantné**, teda
  - riziko nezávisí na t a
  - pravdepodobnosť zlyhania v intervale $[y, y + \delta y]$ je nezávislá na tom, aký dlhý čas jedinec doteraz prežil
- **riziko bez pamäte** – obmedzenie v praxi, nakoľko riziko zvyčajne s časom narastá
- **podmienka konštantnosti rizika** – odhad kumulatívneho rizika zobraziť voči t – graf bude potom predstavovať priamku
  - **kumulatívne riziko** $\Lambda(t) = -\log S(t)$
  - pre exponenciálne rozdelenie $\log(\Lambda(t)) = \log(-\log S(t)) = \log(\lambda) + \log(t)$, kde $\log(t) = -\log(\lambda) + \log(-\log S(t))$
  - graf $\log(t)$ voči $\log(-\log S(t))$ je priamka so **sklonom** rovným jednej a **interceptom** $-\log(\lambda)$
  - graf času t voči riziku λ je **horizontálna** priamka
- **medián času prežívania** je daný riešením rovnice $F(t, \lambda) = \frac{1}{2}$, a teda $t_{0.5} = \frac{1}{\lambda} \log 2$
- exponenciálny model je **špeciálnym prípadom Weibullovho a Gama rozdelenia s parametrom tvaru rovným jednej**

Parametrické regresné modely
Exponenciálne rozdelenie

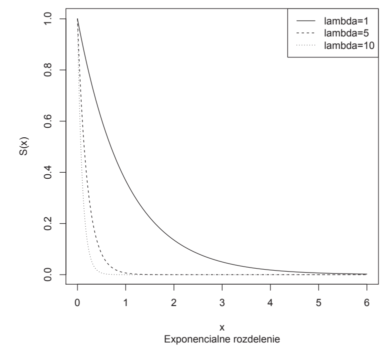


Figure: Funkcie prežívania (Exponenciálne rozdelenie)

Parametrické regresné modely
Weibullovo rozdelenie

Náhodná premenná T má ***Weibullovo rozdelenie*** $W(\lambda, \theta)$ práve vtedy, ak jej **hustota** má tvar

$$f(t, \lambda, \theta) = \lambda\theta (\lambda t)^{\theta-1} e^{-(\lambda t)^{\theta}}, \text{kde } \lambda > 0, \theta > 0, t \geq 0$$

a naviac

- **funkcia rizika** $\lambda(t, \lambda, \theta) = \lambda\theta(\lambda t)^{\theta-1}$
- **distribučná funkcia** $F(t, \lambda, \theta) = 1 - e^{-(\lambda t)^{\theta}}$
- **funkcia prežívania** $S(t, \lambda, \theta) = e^{-(\lambda t)^{\theta}}$
- **stredná hodnota** $E[T] = \int_0^{\infty} \theta(\lambda t)^{\theta} e^{-(\lambda t)^{\theta}} = \lambda^{-1}\Gamma\left(1 + \frac{1}{p}\right)$
- **rozptyl** $Var[T] = \lambda^{-2}\left[\Gamma\left(1 + \frac{2}{\theta}\right) - \Gamma^2\left(1 + \frac{1}{\theta}\right)\right]$
- **p-ty kvantil** $t_p = \frac{1}{\lambda}(-\log(1-p))^{\frac{1}{\theta}}$

Parametrické regresné modely
Weibullovo rozdelenie

- $\Gamma(x)$ predstavuje **Gama funkciu** a je definovaná ako $\int_0^{\infty} u^{x-1} e^{-u} du, x > 0$
- parameter θ sa nazýva **parameter tvaru**
- **funkcia rizika** je
  - **rastúca**, keď je $\theta > 1$
  - **klesajúca**, keď $\theta < 1$
  - **konštantná**, keď $\theta = 1$
- parameter λ sa nazýva **škála**, nakoľko jeho hodnoty menia iba škálu na horizontálnej časovej t osi a nie tvar funkcie
- Weibullov model je veľmi flexibilný a ukázalo sa, že je v praxi často aplikovateľný
- často očakávame **rastúce riziko** s časom, ako napr. modelovanie času prežívania pacientov s leukémiou, ktorí neodpovedajú na liečbu, kde je udalosťou smrť
- avšak môžeme mať aj opačnú situáciu, teda **klesajúce riziko**, ak sa pacienti napr. zotavujú po chirurgickom zákroku

Parametrické regresné modely
Weibullovo rozdelenie

- pre Weibullovo rozdelenie
  $\log(\Lambda(t)) = \log(-\log S(t)) = \theta(\log(\lambda) + \log(t))$, kde
  $\log(t) = -\log(\lambda) + \sigma \log(-\log S(t))$, kde $\sigma = \frac{1}{\theta}$
- graf $\log(t)$ voči $\log(-\log S(t))$ je priamka so **sklonom** rovným $\sigma = \frac{1}{\theta}$ a **interceptom** $-\log(\lambda)$
- **medián času prežívania** je daný riešením rovnice $F(t, \lambda, \theta) = \frac{1}{2}$, a teda $t_{0.5} = \frac{1}{\lambda}(-\log 2)^{\frac{1}{\theta}}$
- Weibullovo rozdelenie je úzko previazané s **rozdelením extrémnych hodnôt** [extreme value distribution] – **logaritmická transformácia** Weibullovej náhodnej premennej nám dáva náhodnú premennú majúcu rozdelenie extrémnych hodnôt

Parametrické regresné modely
Weibullovo rozdelenie

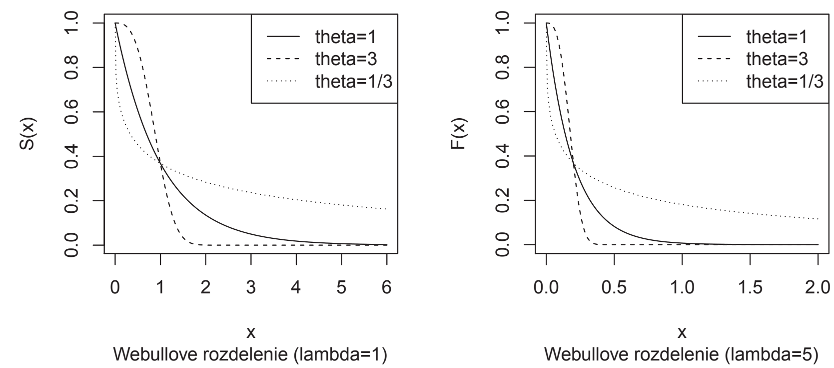


Figure: Funkcie prežívania (Weibullove rozdelenie)

Parametrické regresné modely
Weibullovo rozdelenie

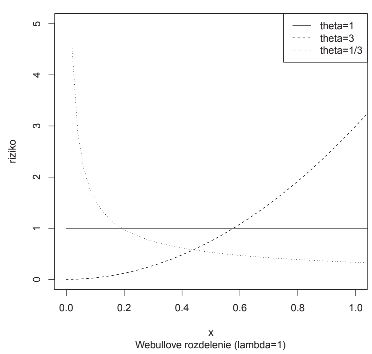


Figure: Funkcie rizika (Weibullove rozdelenie)

Parametrické regresné modely
Rozdelenie extrémnych (minimálnych) hodnôt

Nech $\mu \in \mathbb{R}$ a $\sigma > 0$ sú parametre polohy a škály – ***štandardizované rozdelenie extrémnych hodnôt*** má $\mu = 0$ a $\sigma = 1$; potom

- **hustota** $f(t, \mu, \sigma) = \frac{1}{\sigma}\exp(\frac{t-\mu}{\sigma} - \exp\frac{t-\mu}{\sigma})$
- **distribučná funkcia** $F(t, \mu, \sigma) = 1 - \exp(-\exp\frac{t-\mu}{\sigma})$
- **funkcia prežívania** $S(t, \mu, \sigma) = \exp(-\exp\frac{t-\mu}{\sigma})$
- **stredná hodnota** $E[T] = \mu - \gamma\sigma$
- **rozptyl** $Var[T] = \frac{\pi^2}{6}\sigma^2$
- p**-ty kvantil** $t_p = \mu + \sigma\log(-\log(1-p))$

kde $\gamma \doteq 0.5772$ je Eulerovo číslo, parameter polohy μ je 0.632-tý kvantil a $t \in \mathbb{R}$

- potom ak T je Weibullova náhodná premenná s paramatrami θ a λ, $Y = \log(T)$ má rozdelenie extrémnych hodnôt s parametrami $\mu = -\log\lambda$ a $\sigma = \frac{1}{\theta}$
- naviac $Y = \mu + \sigma Z$, kde Z má štandardizované rozdelenie extrémnych hodnôt
- parametre μ a σ nemenia tvar rozdelenia, iba polohu a škálu

Parametrické regresné modely
Log-normálne rozdelenie

Ak je čas prežívania T ***log-normálne** rozdelený*, potom $Y = \log(T)$ je normálne rozdelené s parametrami μ a σ, teda $Y \sim N(\mu, \sigma^2)$. Nech $Y = \mu + \sigma Z$, potom $Z \sim N(0,1)$. Nech $\theta, \lambda > 0$, potom

- **hustota** $f(t,\theta,\lambda) = \frac{1}{\sqrt{2\pi}} \theta t^{-1} \exp(\frac{-\theta^2 (\log(\lambda t))^2}{2})$
- **distribučná funkcia** $F(t,\theta,\lambda) = \Phi(\theta \log(\lambda t))$
- **funkcia prežívania** $S(t,\theta,\lambda) = 1 - \Phi(\theta \log(\lambda t))$
- **funkcia rizika** $\lambda(t,\theta,\lambda) = \frac{f(t,\theta,\lambda)}{S(t,\theta,\lambda)}$
- **stredná hodnota** $E[T] = \exp(\mu + \frac{\sigma^2}{2})$
- **rozptyl** $Var[T] = (\exp(\sigma^2) - 1)(\exp(2\mu + \sigma^2))$

kde $\mu = -\log\lambda$ a $\sigma = \frac{1}{\theta}$

- **funkcia rizika** sa rovná nule v čase $t = 0$, **rastie** do maxima, potom **klesá** smerom k nule, keď t dosahuje veľké hodnoty – riziko klesá pre veľké hodnoty t, čo sa môže zdať ako nepoužiteľná vlastnosť tohoto rozdelenia, napr. pri modelovaní prežívania tuberkulóznych pacientov, kde ich zlyhávanie najprv stúpa a potom neskôr klesá, tento model vhodný je
- dobrou aproximáciou log-normálneho rozdelenia je **log-logistické rozdelenie**

Parametrické regresné modely
Log-logistické rozdelenie

Ak je čas prežívania T ***log-logisticky** rozdelený*, potom $Y = \log(T)$ je logisticky rozdelené s parametrom **polohy** μ a **škály** σ. Nech $Y = \mu + \sigma Z$, potom Z má štandardizované logistické rozdelenie s hustotou

$$f(z) = \frac{\exp z}{(1 + \exp z)^2}, z \in \mathbb{R},$$

ktorá je symetrická, jej chvosty sú o niečo ťažšie ako chvosty normálneho rozdelenia, šikmosť a špicatosť sú rovné 1.2, $E[Z] = 0$ a $Var[Z] = \frac{\pi^2}{3}$. Potom pre T a $\lambda, \theta > 0$ platí

- **hustota** $f(t,\lambda,\theta) = \lambda\theta(\lambda t)^{\theta-1}(1 + (\lambda t)^\theta)^{-2}$
- **distribučná funkcia** $F(t,\lambda,\theta) = 1 - \frac{1}{1+(\lambda t)^\theta}$
- **funkcia prežívania** $S(t,\lambda,\theta) = \frac{1}{1+(\lambda t)^\theta}$
- **funkcia rizika** $\lambda(t,\lambda,\theta) = \frac{\lambda\theta(\lambda t)^{\theta-1}}{1+(\lambda t)^\theta}$
- p**-ty kvantil** $t_p = \lambda^{-1}(\frac{p}{1-p})^{\frac{1}{\theta}}$

kde $\mu = -\log\lambda$ a $\sigma = \frac{1}{\theta}$

Parametrické regresné modely
Log-logistické rozdelenie

- tento model sa stal populárnym, podobne ako Weibullov model, aj pre jednoduché vyjadrenie funkcie prežívania a funkcie rizika
- aproximácia cenzurovaných dát pomocou tohoto modelu sa v praxi ukázala lepšia ako aproximácia log-normálny rozdelením (okrem extrémov v chvostoch rozdelenia)
- **riziková funkcia je podobná Weibullovej**, líšia sa len menovateľom $1 + (\lambda t)^\theta$.
- ak $\theta < 1 (\sigma > 1)$ je riziková funkcia monotónne **klesajúca** z ∞ a
- ak $\theta = 1 (\sigma = 1)$, je riziková funkcia monotónne **klesajúca** z λ
- Ak $\theta > 1 (\sigma < 1)$ je riziková funkcia monotónne **rastúca** z nuly do maxima v $t = \frac{(\theta-1)^{\frac{1}{\theta}}}{\lambda}$ a **klesajúca** potom k nule
- dá sa ľahko ukázať, že **šanca prežitia za časom** t sa dá zapísať ako

$$\frac{S(t)}{1 - S(t)} = (\lambda t)^{-\theta}$$

  a potom $\log t$ je lineárnou funkciou logaritmu šance prežívania za časom t, teda $\log t = \mu + \sigma(-\log\frac{S(t)}{1-S(t)})$, kde $\mu = -\log\lambda$ a $\sigma = \frac{1}{\theta}$
- graf $\log t$ oproti $-\log\frac{S(t)}{1-S(t)}$ je priamka so **sklonom** σ a **interceptom** μ

Parametrické regresné modely
Log-logistické rozdelenie

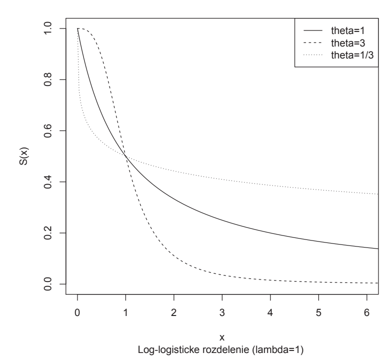


Figure: Funkcie prežívania (Log-logistické rozdelenie)

Parametrické regresné modely
Log-logistické rozdelenie

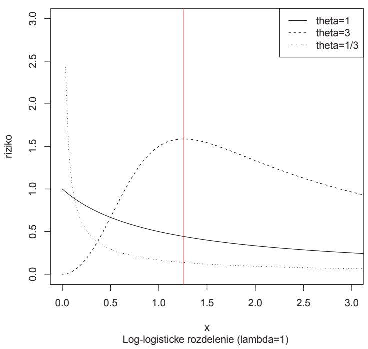


Figure: Funkcie rizika (Log-logistické rozdelenie)

Parametrické regresné modely
Gama rozdelenie

Náhodná premenná T má **Gama rozdelenie** $Gama(\lambda, k)$ práve vtedy, ak jej **hustota** má tvar

$$f(t) = \frac{\lambda^k}{\Gamma(k)} t^{k-1} e^{-\lambda t}, \text{ kde } t \geq 0, \lambda > 0, k > 0.$$

Potom ak neúplná Gama funkcia $I_k(s) = \left(\int_0^s t^{k-1} e^{-t} dt\right) / \Gamma(k)$, tak

- **funkcia rizika** $\lambda(t, \lambda, k) = \frac{\lambda(\lambda t)^{k-1} e^{-\lambda t} \Gamma^{-1}(k)}{1 - I_k(\lambda t)}$
- **distribučná funkcia** $F(t, \lambda, k) = I_k(\lambda t)$
- **funkcia prežívania** $S(t, \lambda, k) = 1 - I_k(\lambda t)$
- **stredná hodnota** $E[T] = \frac{k}{\lambda}$
- **rozptyl** $Var[T] = \frac{k}{\lambda^2}$

Parametrické regresné modely
Gama rozdelenie

- ak $k > 1$ funkciu rizika monotónne **rastie** z 0
- ak $k < 1$ funkciu rizika monotónne **klesá** z ∞
- ak $k = 1$, potom sa Gama rozdelenie redukuje na **exponenciálne rozdelenie** a riziková funkcia je **konštantná**, t.j. rovná λ
- **model** pre $Y = \log T$ môžeme písať v tvare $Y = \mu + Z$, kde Z má hustotu $\frac{\exp(kz - \exp(z))}{\Gamma(k)}$
- náhodná premenná Y sa nazýva **log-Gama náhodná premenná** s parametrami k a $\mu = -\log \lambda$
- Z má negatívne zošikmené rozdelenie s klesajúcou šikmosťou ak k rastie
- ak $k = 1$ potom ide o **exponenciálny model** a Z má štandardizované rozdelenie extrémnej hodnoty
- s výnimkou $k = 1$ **nie je Gama rozdelenie členom rodiny polohovo-škálových rozdelení, ale iba členom rodiny polohových rozdelení**

Parametrické regresné modely
Rozdelenia pravdepodobnosti – zhrnutie

Okrem Gama rozdelenia majú všetky spomenuté rozdelenia času prežívania T vlastnosť, že **rozdelenie** $\log T$ **je členom rodiny polohovo-škálových rozdelení**. Spoločné črty spomenutých rozdelení sú nasledovné

1. rozdelenie T má dva parametre škálu λ a tvar θ
2. rozdelenie $\log T$ má dva parametre škálu $\mu = -\log \lambda$ a tvar $\sigma = 1/\theta$
3. pre všetky rozdelenia platí, že $Y = \log T = \mu + \sigma Z$, kde Z má štandardizované rozdelenie s $\mu = 0 (\lambda = 1)$ a $\sigma = 1 (\theta = 1)$
4. tieto modely sú **log-lineárnymi modelmi**

T	$Y = \log T$
Weibullove rozdelenie	rozdelenie extrémnych hodnôt
log-normálne rozdelenie	normálne rozdelenie
log-logistické rozdelenie	logistické rozdelenie

Parametrické regresné modely
Konštrukcia qq-diagramu

- nech $\widehat{S}_{KM}(t)$ je **Kaplan-Meierov odhad funkcie prežívania** v čase t
- nech $t_i, i = 1, 2, \ldots r \leq n$ sú **zoradené necenzurované časy do zlyhania**
- pre každý necenzurovaný výberový kvantil $y_i = \log t_i$ je odhadnutá **pravdepodobnosť zlyhania** $\widehat{p}_i^{(KM)} = 1 - \widehat{S}_{KM}(t_i)$
- **parametrický štandardizovaný kvantil** získame použitím $\widehat{p}_i = F_{0,1}(z_i) = \Pr(Z \leq z_i)$, kde $F_{0,1}(z_i)$ je **distribučná funkcia štandardizovaného parametrického modelu** ($\mu = 0, \sigma = 1$)
- tieto štandardizované kvantily porovnávame s kvantilmi neparametrického odhadu $\widehat{S}_{KM}(t)$ – ak je navrhovaný model adekvátny, potom graf (z_i, y_i) leží na priamke so **sklonom** σ a **interceptom** μ

t_p kvantil	$y_p = \log t_p$ kvantil	štandardizovaný kvantil z_p
Weibull r.	roz.extr.h.	$\log(-\log \widehat{S}(t_p)) = \log \Lambda(t_p) = \log(-\log(1-p))$
log-norm.r.	norm.r.	$\Phi^{-1}(p)$
log-logis.roz.	logis.r.	$= -\log \frac{\widehat{S}(t_p)}{1-\widehat{S}(t_p)} = -\log(\frac{1-p}{p})$

Regresné modely pre jednovýberový prípad
Odhady a testy pre exponenciálny model pre necenzurované dáta

- **funkcia vierohodnosti**

$$L(\lambda|\mathbf{t}) = \prod_{i=1}^{n} \lambda \exp(-\lambda t_i) = \lambda^n \exp(-\lambda \sum_{i=1}^{n} t_i)$$

- **logaritmus funkcie vierohodnosti**

$$l(\lambda|\mathbf{t}) = n \log \lambda - \lambda \sum_{i=1}^{n} t_i$$

- **maximálne vierohodné odhady (MLE)**

$$\frac{\partial l(\lambda|\mathbf{t})}{\partial \lambda} = \frac{n}{\lambda} - \sum_{i=1}^{n} t_i = 0, \text{ potom}$$

$$\widehat{\lambda} = \frac{n}{\sum_{i=1}^{n} t_i} = \frac{1}{\overline{T}}, \widehat{E[T]} = \frac{1}{\widehat{\lambda}} = \overline{T}, \widehat{Var[T]} = \frac{1}{\widehat{\lambda}^2} = \overline{T}^2$$

Regresné modely pre jednovýberový prípad
Odhady a testy pre exponenciálny model pre necenzurované dáta

- **exaktné rozdelenie** – ak $T_i \sim Exp(\lambda)$, potom $\sum_{i=1}^{n} T_i \sim Gama(\lambda, k = n)$, potom

$$2\lambda \sum_{i=1}^{n} T_i = 2n \frac{\lambda}{\widehat{\lambda}} \sim \chi^2_{2n}$$

potom $(1-\alpha) \times 100\%$ **interval spoľahlivosti (IS) pre** λ

$$\left\{ \lambda : \lambda \in (\chi^2_{2n}(1-\alpha/2)\frac{\widehat{\lambda}}{2n}, \chi^2_{2n}(\alpha/2)\frac{\widehat{\lambda}}{2n}) \right\},$$

$(1-\alpha) \times 100\%$ **IS pre** $1/\lambda$ (IS pre strednú hodnotu $E[T]$)

$$\left\{ 1/\lambda : 1/\lambda \in (\frac{2n\overline{T}}{\chi^2_{2n}(\alpha/2)}, \frac{2n\overline{T}}{\chi^2_{2n}(1-\alpha/2)}) \right\}$$

Regresné modely pre jednovýberový prípad
Odhady a testy pre exponenciálny model pre necenzurované dáta

- p**-ty kvantil** t_p, kde $p = F(t_p|\lambda) = 1 - \exp(-\lambda t_p)$, potom $t_p = -\log(1-p)/\lambda$; nech $\widehat{t}_p$ je MLE t_p, potom $\widehat{t}_p = -\log(1-p)/\widehat{\lambda} = -\overline{T}\log(1-p)$ a MLE mediánu je $\widehat{t}_{0.5} = -\overline{T}\log(\frac{1}{2}) = \overline{T}\log 2$
- **test pomerom vierohodnosti**

$$LR(\lambda_0|\mathbf{t}) = -2\log\frac{L(\lambda_0|\mathbf{t})}{L(\lambda|\mathbf{t})} \sim \chi^2_1$$

kde

$$l(\lambda_0|\mathbf{t}) = n\log\lambda_0 - \lambda_0 n\overline{T}$$

a

$$l(\lambda|\mathbf{t}) = n\log\frac{1}{\overline{T}} - \frac{1}{\overline{T}}n\overline{T}$$

Example (exponenciálny model)

(pokrač.) Majme AML dáta. Ak predpokladáme, že cenzúry sú zlyhania (problém B, kap.1), potom vypočítajte MLE pre strednú hodnotu $1/\lambda$, medián $\widehat{t}_{0.5}$ ako aj 95%IS pre λ a $1/\lambda$. Otestujte $H_0 : E[T] = 30$ oproti $H_1 : E[T] \neq 30$ na $\alpha = 0.05$

Skupina	Čas po kompletný relaps (v týždňoch)	n	udalostí	cenzúr
skupina A	$9, 13, 13+, 18, 23, 28+, 31, 34, 45+, 48, 161+$	11	7	4
skupina B	$5, 5, 8, 8, 12, 16+, 23, 27, 30, 33, 43, 45$	12	11	1

- $u \in \mathbb{U}$ – **necenzurované pozorovania**
- $c \in \mathbb{C}$ – **cenzurované pozorovania**
- n_u **počet necenzurovaných pozorovaní**
- **vektor pozorovaní $\mathbf{x} = (x_1, \ldots, x_n)^T$** a
- **vektor indikácií zlyhania a cenzúry** $\delta = (\delta_1, \ldots, \delta_n)^T$ (zodpovedajúci **x**; 1 je zlyhanie a 0 je cenzúra)
- **Fisherova informačná matica** $I(\theta) = -E\left[\frac{\partial^2}{\partial\theta_j\partial\theta_k} l(\theta|\mathbf{x})\right] = nI_1(\theta)$
- **Fisherova informačná matica nejakého jedného pozorovania** x_1 $I_1(\theta) = -E\left[\frac{\partial^2}{\partial\theta_j\partial\theta_k} f(x_1|\theta)\right]$

- Potom platí $\widehat{\theta} \sim N_d(\theta, I^{-1}(\theta))$, kde
  - d je dimenzia **x**
  - $I(\theta)$ je matica $d \times d$
  - i-ty diagonálny element $I^{-1}(\theta)$ je asymptotickým rozptylom i-teho elementu θ
  - mimodiagonálne elementy sú asymptotickými kovarianciami korešpondujúcich elementov θ
- ak je θ skalár, potom $Var(\theta) = \frac{1}{I(\theta)}$, kde $I(\theta) = -E[\frac{\partial^2}{\partial\theta^2} l(\theta|\mathbf{x})]$
- **pre cenzurované dáta** ide o funkciu cenzurovaného rozdelenia G ako aj rozdelenia času prežívania F
- preto je potrebné aproximovať $I(\theta)$ **pozorovanou informačnou maticou** $I^*(\theta)$ v bode $\widehat{\theta}$, kde $I^*(\theta) = -E\left[\frac{\partial^2}{\partial\theta_j\partial\theta_k} l(\theta|\mathbf{x})\right]$
- v jednorozmernom prípade $I^*(\theta) = -E\left[\frac{\partial^2}{\partial\theta^2} l(\theta|\mathbf{x})\right]$, kde rozptyl $Var(\theta) = (I^*(\theta))^{-1}$

- **funkcia vierohodnosti** – vo všeobecnosti pre náhodne cenzurované dáta platí

$$L(\theta|\mathbf{x}) = \prod_{i=1}^{n} f(x_i|\theta)^{\delta_i} S_f(x_i|\theta)^{1-\delta_i}$$

potom

$$\begin{aligned} L(\lambda|\mathbf{x}) &= \prod_{i:u\in\mathbb{U}} \lambda \exp(-\lambda x_i) \prod_{i:c\in\mathbb{C}} \exp(-\lambda x_i) \\ &= \lambda^{n_u} \exp(-\lambda \sum_{i:u\in\mathbb{U}} x_i) \exp(-\lambda \sum_{i:c\in\mathbb{C}} x_i) \\ &= \lambda^{n_u} \exp(-\lambda \sum_{i=1}^{n} x_i) \end{aligned}$$

- **logaritmus funkcie vierohodnosti**

$$l(\theta|\mathbf{x}) = \sum_{i:u\in\mathbb{U}} \log f(x_i|\theta) + \sum_{i:c\in\mathbb{C}} \log S_f(x_i|\theta),$$

$$l(\lambda|\mathbf{x}) = n_u \log\lambda - \lambda\sum_{i=1}^{n} x_i$$

- **maximálne vierohodné odhady (MLE)** – prvá a druhá derivácia $l(\lambda|\mathbf{x})$

$$\frac{\partial l(\lambda|\mathbf{x})}{\partial\lambda} = \frac{n_u}{\lambda} - \sum_{i=1}^{n} x_i, \frac{\partial^2 l(\lambda|\mathbf{x})}{\partial\lambda^2} = -\frac{n_u}{\lambda^2} = -I^*(\lambda),$$

potom

$$\widehat{\lambda} = \frac{n_u}{\sum_{i=1}^{n} x_i}, \widehat{Var\left[\widehat{\lambda}\right]} = \frac{\widehat{\lambda}^2}{E[n_u]}, \text{ kde } E[n_u] = n\Pr(T \le C)$$

- **asymtpotické rozdelenie** λ

$$\frac{\widehat{\lambda}-\lambda}{\sqrt{\widehat{Var(\widehat{\lambda})}}} \sim N(0,1), \text{ kde } \widehat{Var(\widehat{\lambda})} \approx \frac{\widehat{\lambda}^2}{n_u} = \frac{1}{I^*(\lambda)}$$

kde $E[n_u]$ substituujeme za n_u, lebo rozdelenie $G(\cdot)$ nepoznáme. Treba si uvedomiť, že rozptyl je závislý na λ; potom $(1-\alpha)\times 100\%$ **IS pre** λ

$$\left\{\lambda : \lambda \in (\widehat{\lambda} - u_{\alpha/2}SE(\widehat{\lambda}), \widehat{\lambda} + u_{\alpha/2}SE(\widehat{\lambda}))\right\},$$

$(1-\alpha)\times 100\%$ **IS pre** $1/\lambda$

$$\left\{1/\lambda : 1/\lambda \in (1/\widehat{\lambda} - u_{\alpha/2}SE(1/\widehat{\lambda}), 1/\widehat{\lambda} + u_{\alpha/2}SE(1/\widehat{\lambda}))\right\},$$

kde $Var(1/\lambda)$ dostaneme pomocou *delta metódy*, kde $g(\lambda) = 1/\lambda$, $g'(\lambda) = -1/\lambda^2$ a $Var(1/\lambda) = \frac{1}{\lambda^2 E[n_u]} \approx \frac{1}{\lambda^2 n_u}$

- ***p*-ty kvantil** t_p, $\widehat{t}_p = -\log(1-p)/\widehat{\lambda} = -\frac{\sum_{i=1}^{n} x_i}{n_u}\log(1-p)$ a MLE mediánu je $\widehat{t}_{0.5} = -\frac{\sum_{i=1}^{n} x_i}{n_u}\log(\frac{1}{2})$; **rozptyl *p*-teho kvantilu**

$$\widehat{Var(\widehat{t}_p)} = \log^2(1-p)Var(1/\widehat{\lambda}) \approx \log^2(1-p)\frac{1}{\widehat{\lambda}^2 n_u}$$

$(1-\alpha)\times 100\%$ **IS pre** $t_{0.5}$

$$\left\{t_{0.5} : t_{0.5} \in (\widehat{t}_{0.5} - u_{\alpha/2}SE(\widehat{t}_{0.5}), \widehat{t}_{0.5} + u_{\alpha/2}SE(\widehat{t}_{0.5}))\right\}$$

- pomocou **delta metódy** môžeme vypočítať IS, ktorého hranice sú menej vychýlené nájdením **transformácií**, ktoré eliminujú závislosť rozptylu nejakého parametra na parametri samotnom. Napr. $g(\lambda) = \log\lambda$, $g'(\lambda) = 1/\lambda$ a potom $Var(\log\lambda) = \lambda^{-2}\frac{\lambda^2}{E[n_u]} \approx \frac{1}{n_u}$

Potom $\log(\widehat{\lambda}) \sim N(\log\lambda, 1/n_u)$, $(1-\alpha)\times 100\%$ **IS pre** $\log\lambda$

$$\left\{\log\lambda : \log\lambda \in (\log\widehat{\lambda} - u_{\alpha/2}\frac{1}{\sqrt{n_u}}, \log\widehat{\lambda} + u_{\alpha/2}\frac{1}{\sqrt{n_u}})\right\}$$

Analogicky dostaneme

$$\log\frac{1}{\widehat{\lambda}} \sim N(\log\frac{1}{\lambda}, \frac{1}{n_u}), \log\widehat{t}_{0.5} \sim N(\log t_{0.5}, \frac{1}{n_u}),$$

potom $(1-\alpha)\times 100\%$ **IS pre** $\log(1/\lambda)$

$$\left\{\log(1/\lambda) : \log(1/\lambda) \in (\log(1/\widehat{\lambda}) - u_{\alpha/2}\frac{1}{\sqrt{n_u}}, \log(1/\widehat{\lambda}) + u_{\alpha/2}\frac{1}{\sqrt{n_u}})\right\}$$

a $(1-\alpha)\times 100\%$ **IS pre** $\log t_{0.5}$

$$\left\{\log t_{0.5} : \log t_{0.5} \in (\log\widehat{t}_{0.5} - u_{\alpha/2}\frac{1}{\sqrt{n_u}}, \log\widehat{t}_{0.5} + u_{\alpha/2}\frac{1}{\sqrt{n_u}})\right\}$$

Spätnou transformáciou hraníc IS pre log(paramater) dostaneme hranice pre IS parametra ako exp(koncové body)

- **ML odhad funkcie prežívania** $\widehat{S}(t) = \exp(-\widehat{\lambda} t)$, ktorej rozdelenie môžeme dostať pomocou delta metódy

Alternatívne zoberieme **log-log transformáciu**, ktorá zvyčajne urýchli konvergenciu k normalite (kvôli nezávislosti rozptylu na neznámom parametri λ). Vieme, že $\log(\widehat{\lambda}) \sim N(\log\lambda, 1/n_u)$, potom

$$\log(-\log\widehat{S}(t)) = \log(\widehat{\lambda}) + \log t$$

a pre rozptyl platí

$$\widehat{Var}(\log(-\log\widehat{S}(t))) = \widehat{Var}(\log(\widehat{\lambda})) \approx \frac{1}{n_u}$$

Z delta metódy pre každé fixované t platí

$$\log(-\log\widehat{S}(t)) \sim N\left(\log(-\log S(t)), \frac{1}{n_u}\right), \text{ kde } \log(-\log S(t)) = \log(\lambda t)$$

$(1-\alpha)\times 100\%$ **IS pre** $S(t)$

$$\left\{S(t) : S(t) \in \left(\exp\left(\log\widehat{S}(t)\exp\left(\frac{u_{\alpha/2}}{\sqrt{n_u}}\right)\right), \exp\left(\log\widehat{S}(t)\exp\left(\frac{-u_{\alpha/2}}{\sqrt{n_u}}\right)\right)\right)\right\}$$

Regresné modely pre jednovýberový prípad
Odhady a testy pre exponenciálny model pre cenzúrované dáta

- **test pomerom vierohodnosti**

$$LR(\lambda_0) = -2\log\frac{L(\lambda_0)}{L(\lambda)} \sim \chi^2_1,$$

kde

$$l(\lambda_0|\mathbf{x}) = n_u\log\lambda_0 - \lambda_0\sum_{i=1}^{n} x_i$$

a

$$l(\lambda|\mathbf{x}) = n_u\log\frac{n_u}{\sum_{i=1}^{n} x_i} - n_u$$

Stanislav Katina Vybrané kapitoly z analýzy prežívania

Regresné modely pre jednovýberový prípad
Odhady a testy pre exponenciálny model pre cenzúrované dáta

Example (exponenciálne rozdelenie)

(pokrač.) Majme AML dáta. Vypočítajte MLE pre strednú hodnotu $1/\lambda$, medián $\widehat{t}_{0.5}$ ako aj 95%IS pre λ a $1/\lambda$. Otestujte $H_0 : E[T] = 30$ oproti $H_1 : E[T] \neq 30$ na $\alpha = 0.05$.

Skupina	Čas po kompletný relaps (v týždňoch)	n	udalostí	cenzúr
skupina A	$9, 13, 13+, 18, 23, 28+, 31, 34, 45+, 48, 161+$	11	7	4
skupina B	$5, 5, 8, 8, 12, 16+, 23, 27, 30, 33, 43, 45$	12	11	1

Example (exponenciálne rozdelenie)

(pokrač.) Použitím funkcie `survreg` pre cenzurované dáta zopakujte výpočty urobené ručne, teda vypočítajte bodové odhady strednej hodnoty a mediánu ako aj ich 95%IS a zostrojte qq-diagram. Nakreslite funkcie prežívania pre exponenciálne, Weibullovo a log-logistické rozdelenie a porovnajte ich s $\widehat{S}_{KM}(t)$.

Stanislav Katina Vybrané kapitoly z analýzy prežívania

Regresné modely pre jednovýberový prípad
Odhady a testy pre exponenciálny model pre cenzúrované dáta

Pozn.: **Exponenciálne rozdelenie** času t je Weibullovo rozdelenie s parametrom tvaru $\theta = 1$ alebo $\log t$ má rozdelenie extrémnej hodnoty so škálou $\sigma = 1$. Funkcia `survreg` fituje $\log t$ a výstupom je $\widehat{\mu} = -\log\widehat{\lambda}$ (parameter polohy rozdelenia extrémnej hodnoty). Pre **Weibullove rozdelenie** $\widehat{\lambda} = \exp(-\widehat{\mu})$ a odhad $1/\widehat{\lambda} = \exp(\widehat{\mu})$ a naviac $\widehat{\sigma} = 1/\widehat{\theta}$. Funkcia `summary(fit)`, kde `fit` je vypočítaný pomocou funkcie `survreg`, nám poskytuje práve $\widehat{\mu}$ a $\widehat{\sigma}$. Vo funkcii `survreg` treba špecifikovať rozdelenie `dist="weibull"`. Funkcia `predict` je doplnkom ku funkcii `survreg` a jej výstupom sú odhady kvantilov a ich štandardných chýb. Jeden z argumentov funkcie `predict` je `type`. Nastavte `type="uquantile"`, kedy ide o log-transformáciu. Default počíta odhady rozptylu a IS pomocou delta metódy. Ak chceme odhadnúť lineárny prediktor (`lp`), potom použijeme funkciu `predict(fit, type="lp", newdata = list(skupina = 1))`.

Pri výpočte $\widehat{S}(t)$ pre **log-logistické rozdelenie** treba mať na zreteli, že parametre z log-logistického modelu ($\widehat{\mu}$ a $\widehat{\sigma}$, funkcia `survreg`) sa aplikujú na $Y = \log T$. Vo funkcii `survreg` treba špecifikovať rozdelenie `dist="loglogistic"`. Pre **lognormálne rozdelenie** `dist="lognormal"`. Označenia distribučných funkcií v `R` sú nasledovné

	Weibullove r.	logistické r.($Y = \log T$)	normálne r.($Y = \log T$)
$F(t)$	pweibull$(q, \theta, \lambda^{-1})$	plogis(q, μ, σ)	pnorm(q, μ, σ)
t_p	qweibull$(p, \theta, \lambda^{-1})$	qlogis(p, μ, σ)	qnorm(p, μ, σ)

Stanislav Katina Vybrané kapitoly z analýzy prežívania

Regresné modely pre jednovýberový prípad
Odhady a testy pre exponenciálny model pre cenzúrované dáta

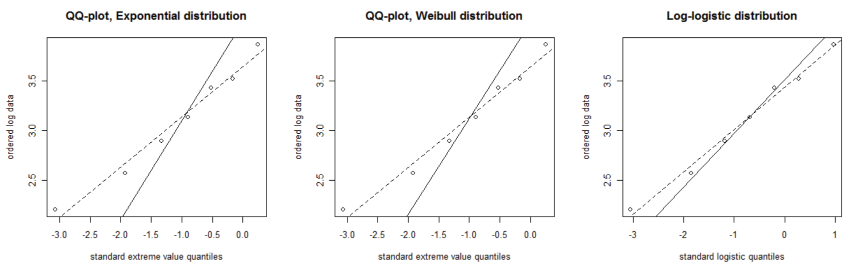


Figure: Porovnanie qq-diagramov

Regresné modely pre jednovýberový prípad
Odhady a testy pre exponenciálny model pre cenzúrované dáta

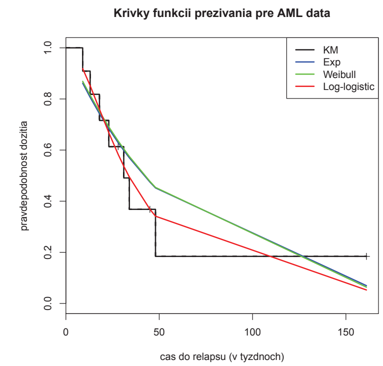


Figure: Porovnanie funkcií prežívania

Regresný model pre dvojvýberový prípad
Úvod

- zamerajme sa na **porovnanie dvoch škálových parametrov** λ – v **log-transformácii** pôjde o porovnanie dvoch parametrov polohy $\mu = -\log\lambda$
- použijeme `AML dáta`, kde **naparametrický log-rank test** zamietol H_0 o rovnosti dvoh kriviek prežívania (p-hodnota=0.03265)
- najprv si položíme otázku, **či nejaký model** (log-rozdelenie, ktoré patrí do rodiny rozdelení polohy a škály) **fituje dáta adekvátne**
- **saturovaný (plný) log-lineárny model** môžeme písať ako

$$\begin{aligned} Y = \log T &= \widetilde{\mu} + \epsilon \\ &= \beta_0^* + \beta^*\text{skupina} + \epsilon \\ &= \begin{cases} \beta_0^* + \beta^*\text{skupina} + \epsilon, & \text{kde skupina=1} \\ \beta_0^* + \epsilon, & \text{kde skupina=0} \end{cases} \end{aligned}$$

Regresný model pre dvojvýberový prípad
Úvod

- parameter $\widetilde{\mu} = \beta_0^* + \beta^*\text{skupina}$ – **lineárny prediktor**, ktorý nadobúda dve hodnoty
  - $\mu_1 = \beta_0^* + \beta^*$
  - $\mu_2 = \beta_0^*$
- vieme, že $\widetilde{\mu} = -\log\widetilde{\lambda}$, kde $\widetilde{\lambda}$ znamená **parameter škály** rozdelenia premennej T
- potom $\widetilde{\lambda} = \exp(-\beta_0^* - \beta^*\text{skupina})$ nadobúda dve hodnoty
  - $\lambda_1 = \exp(-\beta_0^* - \beta^*)$
  - $\lambda_2 = \exp(-\beta_0^*)$
- **nulová hypotéza**

  $H_0 : \lambda_1 = \lambda_2$ vtedy a len vtedy, keď $\mu_1 = \mu_2$ vtedy a len vtedy, keď $\beta^* = 0$
- treba si uvedomiť, že **parameter škály** σ v log-transformovanom modeli je rovný (**parameter polohy**)$^{-1}$, kde ide o parameter polohy θ originálneho (netransformovaného) modelu, teda $\sigma = 1/\theta$

Regresný model pre dvojvýberový prípad
Úvod

H_0 testujeme za nasledovných dvoch predpokladov

P1 Predpokladáme **rovnaký parameter tvaru** θ, teda predpokladáme, že $\sigma_1 = \sigma_2$ (**parametre škály**) sa rovnajú. Potom sa chyba $\epsilon = \sigma Z$, kde náhodná premenná Z pochádza z rozdelenia extrémnych hodnôt, štandardizovaného logistického rozdelenia alebo štandardizovaného normálneho rozdelenia.

P2 Predpokladáme **rôzne parametre tvaru** θ, teda $\sigma_1 \neq \sigma_2$

Na testovanie H_0 použijeme tri nasledovné modely

M1 Dáta pochádzajú z **rovnakého rozdelenia**, kde **nulový model** bude mať tvar $Y = \log T = \beta_0^* + \sigma Z$, kde Z pochádza z rozdelenia extrémnych hodnôt.

M2=P1 **Model s rôznymi parametrami polohy** μ **a rovnakými parametrami škály** σ bude mať tvar $Y = \log T = \beta_0^* + \beta^*\text{skupina} + \sigma Z$

M3=P2 **Model s rôznymi parametrami polohy** μ **a škály** σ bude mať tvar $Y = \log T = \beta_0^* + \beta^*\text{skupina} + \epsilon$

Regresný model pre dvojvýberový prípad
Úvod

model	parametre	slovné vyjadrenie
M1	β_0^*, σ	rovnaká poloha a škála
M2	$\beta_0^*, \beta^*, \sigma \equiv \mu_1, \mu_2, \sigma$	rôzna poloha a rovnaká škála
M3	$\beta_0^*, \beta^*, \sigma_1, \sigma_2 \equiv \mu_1, \mu_2, \sigma_1, \sigma_2$	rôzna poloha a škála

Regresný model pre dvojvýberový prípad
Exponenciálny a Weibullov regresný model pre dve skupiny

Nech $\widetilde{\mu} = -\log\widetilde{\lambda}$ a $\sigma = 1/\theta$. Potom

$$\lambda(t|\text{skupina}) = \theta\widetilde{\lambda}^\theta t^{\theta-1} = \theta\lambda^\theta t^{\theta-1}\exp(\beta\text{skupina}) = \lambda_0(t)\exp(\beta\text{skupina}),$$

kde $\lambda = \exp(-\beta_0^*)$ a $\beta = -\beta^*/\sigma$, $\lambda_0(t)$ je **baseline riziko** (teda ak skupina=0 alebo $\beta = 0$). Teda $\lambda_0(t)$ je **riziková funkcia pre Weibullovo rozdelenie so škálou** λ **nezávislá na ďalších premenných**.
Pomer rizík skupina=1 a skupiny=0

$$\text{HR} = \frac{\lambda(t|1)}{\lambda(t|0)} = \frac{\exp\beta}{\exp 0} = \exp\beta,$$

teda môžeme povedať, že Weibullov model spĺňa **podmienku proporcionality rizík**. Ak $\theta = 1$ ide o exponenciálny regresný model.

Všeobecné zápisy regresných modelov
Exponenciálny regresný model

Nech **funkcia rizika** $\lambda(t,\lambda) = \lambda(t) = \lambda$ je konštantná pri rôznych hodnotách t a $E[T] = 1/\lambda$. Modelujme riziko λ ako funkciu vektora premenných $\mathbf{x}$, potom **funkcia rizika** bude mať tvar

$$\lambda(t|\mathbf{x}) = \lambda_0(t)\exp(\mathbf{x}^T\beta) = \lambda\exp(\sum_{i=1}^k \beta_i x_i),$$

odkiaľ je zreteľné, že premenné ovplyvňujú riziko **multiplikatívne**. V log-lineárnom modeli

$$\log(\lambda(t|\mathbf{x})) = \log(\lambda) + \mathbf{x}^T\beta = \log(\lambda) + \sum_{i=1}^k \beta_i x_i$$

premenné ovplyvňujú logaritmus rizika **aditívne** a $\sum_{i=1}^k \beta_i x_i$ sa nazýva *lineárny prediktor log-rizika*.

Funkcia prežívania

$$S(t|\mathbf{x}) = \exp(-\lambda(t|\mathbf{x})t) = \exp(-\lambda t \exp(\mathbf{x}^T\beta))$$

Hustota

$$f(t|\mathbf{x}) = \lambda(t|\mathbf{x})S(t|\mathbf{x}) = \lambda \exp(\mathbf{x}^T\beta)\exp(-\lambda t \exp(\mathbf{x}^T\beta))$$

Ak T je **rozdelené exponenciálne**, potom $Y = \log T$ má **rozdelenie extrémnych hodnôt s parametrom škály** $\sigma = 1$. Potom

$$\widetilde{\mu} = -\log(\lambda(t|\mathbf{x})) = -\lambda\exp(\mathbf{x}^T\beta) = -\log\lambda - \mathbf{x}^T\beta, \sigma = 1$$

a platí

$$Y = \log T = \widetilde{\mu} + \sigma Z = \beta_0^* + \mathbf{x}^T\beta^* + Z,$$

kde $\beta_0^* = -\log\lambda$, $\beta^* = -\beta$ a $Z \sim f(z) = \exp(z - e^z), z \in \mathbb{R}$ je **rozdelenie extrémnych hodnôt**.
Zhrnutie: $\lambda(t|\mathbf{x}) = \lambda\exp(\mathbf{x}^T\beta)$ je log-lineárny model zlyhania a je transformovaný na lineárny model $Y = \log T$.

Zovšeobecnime teraz funkciu rizika $\lambda(t) = \theta\lambda^\theta t^{\theta-1}$ tak, že budeme modelovať $\lambda(t)$ ako funkciu vektora premenných $\mathbf{x}$. Potom **funkcia rizika** bude mať tvar

$$\begin{aligned}\lambda(t|\mathbf{x}) &= \lambda_0(t)\exp(\mathbf{x}^T\beta)\\ &= \theta\lambda^\theta t^{\theta-1}\exp(\mathbf{x}^T\beta)\\ &= \theta\left(\lambda(\exp(\mathbf{x}^T\beta))^{\frac{1}{\theta}}\right)^\theta t^{\theta-1} = \theta\widetilde{\lambda}^\theta t^{\theta-1},\end{aligned}$$

kde $\widetilde{\lambda} = \lambda(\exp(\mathbf{x}^T\beta))^{\frac{1}{\theta}}$. Potom **log-lineárny model** bude mať tvar

$$\begin{aligned}\log(\lambda(t|\mathbf{x})) &= \log(\theta) + \theta\log(\widetilde{\lambda}) + (\theta-1)\log(t)\\ &= \log(\theta) + \theta\log(\lambda) + \sum_{i=1}^k \beta_i x_i + (\theta-1)\log(t)\end{aligned}$$

Ak $T \sim W(\lambda, \theta)$, potom $Y = \log T = \widetilde{\mu} + \sigma Z$, podmienené dátami $\mathbf{x}$,

$$\widetilde{\mu} = -\log(\widetilde{\lambda}) = \log(\lambda(\exp(\mathbf{x}^T\beta))^{\frac{1}{\theta}}) = -\log(\lambda) - \frac{1}{\theta}\sum_{i=1}^k \beta_i x_i, \sigma = \frac{1}{\theta}$$

a Z má **štandardizované rozdelenie extrémnych hodnôt**. Preto

$$Y = \widetilde{\mu} + \sigma Z = \beta_0^* + \mathbf{x}^T\beta^* + \sigma Z$$

kde $\beta_0^* = -\log\lambda$, $\beta^* = -\sigma\beta$.

Funkcia prežívania

$$S(t|\mathbf{x}) = \exp\left(-(\widetilde{\lambda}t)^\theta\right).$$

Vieme, že $\Lambda(t|\mathbf{x}) = -\log(S(t|\mathbf{x}))$. Potom **logaritmus funkcie kumulatívneho rizika**

$$\begin{aligned}\log(\Lambda(t|\mathbf{x})) &= \theta\log(\widetilde{\lambda}) + \theta\log(t)\\ &= \theta\log(\lambda) + \theta\log(t) + \sum_{i=1}^k \beta_i x_i\\ &= \log(\Lambda_0(t)) + \sum_{i=1}^k \beta_i x_i,\end{aligned}$$

kde $\Lambda_0(t) = -\log(S_0(t)) = (\lambda t)^\theta$ je **baseline funkcie kumulatívneho rizika**.

Všeobecné zápisy regresných modelov
Weibullov regresný model

Logaritmus funkcie kumulatívneho rizika je lineárny v $\log t$ a v regresných koeficientoch β. Potom vo fixovaných x graf $\Lambda(t|x)$ oproti t v log-log škále je priamka so **sklonom** θ a **interceptom** $\mathbf{x}^T\beta + \theta\log\lambda$. Dá sa ľahko ukázať, že

$$\Lambda(t|\mathbf{x}) = \Lambda_0(t)\exp(\mathbf{x}^T\beta) = (\lambda t)^\theta \exp(\mathbf{x}^T\beta).$$

Weibullov regresný model je jediný log-lineárny model, ktorý spĺňa podmienku **proporcionality rizika**. Vieme, že $\lambda(t|\mathbf{x}) = \lambda_0(t)\exp(\mathbf{x}^T\beta)$, kde $\lambda_0(t) = \theta\lambda^\theta t^{\theta-1}$ ako aj to, že $\Lambda_0(t) = (\lambda t)^\theta$. Ďalej vieme, že $\log(\Lambda(t|\mathbf{x})) = \theta\log(\lambda) + \theta\log(t) + \mathbf{x}^T\beta$. **Vzťah medzi koeficientami log-lineárnom modeli a koeficientami v rizikovej funkcii** je nasledovný

$$\beta = -\sigma^{-1}\beta^* \text{ a } \lambda = \exp(-\beta_0^*).$$

Pomer rizík sa dá napísať nasledovne

$$\mathrm{HR}(t|\mathbf{x}_1,\mathbf{x}_2) = \frac{\lambda(t|\mathbf{x}_2)}{\lambda(t|\mathbf{x}_1)} = \left(\exp((\mathbf{x}_2^T - \mathbf{x}_1^T)\beta^*)\right)^{1/\sigma}.$$

Všeobecné zápisy regresných modelov
Log-logistický regresný model

Majme model

$$T = \exp Y = \exp(\beta_0^* + \mathbf{x}^T\beta^* + \sigma Z) = \exp(\mathbf{x}^T\beta^*)T^*,$$

kde $T^* = \exp Z^*$, $Z^* = \beta_0^* + \sigma Z$. Teda **x** má **multiplikatívny efekt** na čas T. **Riziková funkcia** času T (pre dané **x**) sa dá zapísať ako funkcia $\lambda_0^*(\cdot)$ v podobe

$$\lambda(t|\mathbf{x}) = \lambda_0^*\left(\exp(-\mathbf{x}^T\beta^*)t\right)\exp(-\mathbf{x}^T\beta^*),$$

čo vyplýva z transformačnej vety.
Pozn.: $f(t) = f^*(g^{-1}(t))\left|\frac{\partial g^{-1}(t)}{\partial t}\right|$, kde $t = g(t^*)$.
Nech $Y = \log T$, potom má **log-lineárny model** tvar

$$Y = \beta_0^* + \mathbf{x}^T\beta^* + \sigma Z,$$

kde $Z \sim$ **štandardizované logistické rozdelenie**.

Všeobecné zápisy regresných modelov
Log-logistický regresný model

Nech $Z^* = \sigma Z$. Potom môžeme písať

$$Y = \beta_0^* + \mathbf{x}^T\beta^* + Z^*,$$

kde $Z^* \sim$ **logistické rozdelenie so strednou hodnotou** β_0^* **a škálou** σ. Potom **baseline riziková funkcia** log-logistického času $T^* = \exp Z^*$ je

$$\lambda_0^*(t^*) = \frac{\lambda\theta(\lambda t^*)^{\theta-1}}{1+(\lambda t^*)^\theta}$$

kde $\beta_0^* = -\log\lambda$ a $\sigma = 1/\theta$. Treba si uvedomiť, že $\lambda_0^*(t^*)$ je nezávislý na β^*. Potom pre **rizikovú funkciu** T píšeme

$$\lambda(t|\mathbf{x}) = \frac{\theta\widetilde{\lambda}^\theta t^{\theta-1}}{1+\widetilde{\lambda}^\theta t^\theta},$$

kde $\widetilde{\lambda} = \lambda\exp(-\mathbf{x}^T\beta^*)$. A naviac pre dané **x**, $T \sim$ **log-logisticky rozdelené s parametrami** $\widetilde{\lambda}$ **a** θ.

Všeobecné zápisy regresných modelov
Log-logistický regresný model

Teda

$$Y = \log T = \widetilde{\mu} + \sigma Z,$$

kde $\widetilde{\mu} = -\log(\widetilde{\lambda}) = -\log\lambda + \mathbf{x}^T\beta^* = \beta_0^* + \mathbf{x}^T\beta^*$ a $Z \sim$ **štandardizované logistické rozdelenie**. Log-logistický model je log-lineárnym modelom, ale nie je modelom proporcionálneho rizika ako napr. Weibullov regresný model.
Log-logistický regresný model má jednu výhodu oproti ostatným modelom, pretože ide o **model propocionálnych šancí** a **proporcionálneho času**. Log-logistická funkcia prežívania má tvar $S(t|\mathbf{x}) = S_0^*(\exp(-\mathbf{x}^T\beta^*)t) = S_0^*(t^*)$, kde $= S_0^*(\cdot)$ je **baseline funkcia prežívania**.
Pozor ! **x** mení škálu horizontálnej (t) osi. Ak $\mathbf{x}^T\beta^*$ rastie, potom čas do zlyhania klesá a naopak. Preto sa takýto log-lineárny model nazýva **spomaľovací (zrýchlovací) model času zlyhania (accelerated (decelerated) failure time model)**. Do tejto skupiny modelov patrí aj exponenciálny, Weibullov a aj log-normálny regresný model.

Pre **funkciu prežívania** platí

$$S(t|\mathbf{x}) = S_0^*(\exp(-\mathbf{x}^T\beta^*)t) = \frac{1}{1+\left(\exp(y-\beta_0^*-\mathbf{x}^T\beta^*)\right)^{\theta}},$$

kde $y = \log t$, $\beta_0^* = -\log\lambda$ a $\theta = 1/\sigma$.
Šanca prežitia za časom t je daná ako

$$\frac{S(t|\mathbf{x})}{1-S(t|\mathbf{x})} = \left(\exp(y-\beta_0^*-\mathbf{x}^T\beta^*)\right)^{-\theta},$$

kde treba zdôrazniť, že $-\log$(šanca) je lineárnou funkciou $\log t$ ako aj premenných $x_j, j = 1, 2, \ldots k$. **Pomer šancí za časom** t počítaný medzi x_1 a x_2 sa dá vyjadriť nasledovne

$$\mathrm{OR}(t|\mathbf{x}_2, \mathbf{x}_1) = \left(\exp(\mathbf{x}_2-\mathbf{x}_1)^T\beta^*\right)^{\theta}$$

naviac **časový pomer** v t počítaný medzi x_1 a x_2

$$\mathrm{TR}(t|\mathbf{x}_2, \mathbf{x}_1) = (\mathrm{OR}(\mathbf{x}_2, \mathbf{x}_1))^{1/\theta}$$

- pomer šancí je často používaný na meranie efektu premenných x_j a je nezávislý na čase t, čo nazývame **vlastnosť proporcionality šancí**.

Example

Ak OR $= 2$, potom šanca prežitia za časom t jedincov s x_2 je $2\times$ tak veľká ako u jedincov s x_1, čo platí pre všetky t.

- rovnako aj TR je nezávislý na čase, čo nazývame **vlastnosť proporcionality časov**.

Example

Podobne ak TR $= 2$, tak čas prežívania jedincov s x_2 je $2\times$ väčší ako u jedincov s x_1.

- $\mathrm{OR} = \mathrm{TR}^{\theta}$, kde pomer OR je kontrolovaný θ, tvarovým parametrom log-logistického rozdelenia

Example

Ak $\theta = 1$, tak OR $=$ TR. Ak $\theta = 2$, tak TR $= 2$ a OR $= 2^2 = 4$.

- na jednotku nárastu jednej premennej, fixovaním ostatných premenných v x, OR $\to +\infty \vee 0$ ak $\theta \to \infty$ (v závislosti od znamienka korešpondujúcej premennej pri β^*)
- na zistenie $\widehat{\mathrm{HR}}$ a $\widehat{\mathrm{OR}}$ – odhady vypočítané funkciou `survreg`
- ľubovoľný $p \times 100\%$ **percentil** log-logistického modelu času t

$$t_p(\mathbf{x}) = \left(\frac{p}{1-p}\right)^{\sigma}\exp(\beta_0^*+\mathbf{x}^T\beta^*)$$

- log-logistický regresný model je jediným akcelerovaným modelom zlyhania, ktorý spĺňa podmienky **propocionálity šancí** a **proporcionálity času**

Example

(pokrač.) Použitím funkcie `survreg` pre cenzurované dáta odhadnite parametre Weibullovho, log-logistického a log-normálneho modelu a samotné modely porovnajte.

Všeobecné zápisy regresných modelov
Príklad v R

- Na základe výsledkov LR testu medzi modelmi M1 a M2 ako aj M2 a M3 môžeme konštatovať, že model M2, ktorý predpokladá rovnakú škálu je adek-vátny.
- Výsledky z modelov M1, M2 a M3 zhrnieme v nasledovnej tabuľke

rozdelenie	$L(\widehat{\beta}_0^*, \widehat{\beta}^*)$	p-hodnota	$\widehat{\beta}_0^*$	$\widehat{\beta}^*$	p-hodnota (skupina)
Weibullovo	-80.5	0.021	3.180	0.929	0.0151
log-logistické	-79.4	0.120	2.899	0.604	0.1240
log-normálne	-78.9	0.062	2.854	0.724	0.0568

- Vo Weibullovom modeli je efekt skupiny signifikantný (prvá skupina zostáva v remisii dlhšie, t.j. prežívanie je v druhej skupine kratšie) s odhadmi $\widehat{\mu}_1 = 4.109$ a $\widehat{\mu}_2 = 4.109 - 0.929 = 3.18$ rozdelenia extrémnych hodnôt.
- Log-normálny model má maximálnu vierohodnosť najväčšiu a Weibullov model najmenšiu. Avšak LR test fitu modelu je signifikantný len pre Weibullov model, kde p-hodnota=0.0151.

Všeobecné zápisy regresných modelov
Príklad v R

- Mediány s ich 95%IS zhŕňa pre všetky modely nasledovná tabuľka

model	med 1	95%IS pre med 1	med 2	95%IS pre med 2
Weibullov	45.566	(24.901,83.381)	17.990	(10.966,29.514)
Log-logistický	33.215	(18.902,58.366)	18.147	(10.747,30.641)
Log-normálny	35.833	(20.510,62.605)	17.364	(10.556,28.561)

- Odhad lineárneho prediktora $\widehat{\widehat{\mu}} = \widehat{\beta}_0^* + \widehat{\beta}^*$skupina, $\widehat{\widehat{\mu}} = \log(\widehat{\widehat{\lambda}})$, $\widehat{\widehat{\lambda}} = \exp(-\widehat{\widehat{\mu}})$ = $\exp(-\widehat{\beta}_0^* - \widehat{\beta}^*$skupina), $\widehat{\theta} = 1/\widehat{\sigma}$ pre všetky tri modely (Weibullov, log-logistický a log-normálny; zľava doprava) zhŕňa nasledovná tabuľka

skupina	$\widehat{\widehat{\lambda}}$	$\widehat{\sigma}$	$\widehat{\widehat{\lambda}}$	$\widehat{\sigma}$	$\widehat{\widehat{\lambda}}$	$\widehat{\sigma}$
0	0.042	1.264	0.055	1.950	0.058	1.160
1	0.016	1.264	0.030	1.950	0.028	1.160

- Z qq-diagramu je zrejmé, že log-logistický a log-normálny model fitujú skupinu `Maintained` lepšie ako Weibullov regresný model. Avšak fit skupiny `Nonmain- tained` sa použitím týchto modelov nezlepšil.

Všeobecné zápisy regresných modelov
Príklad v R

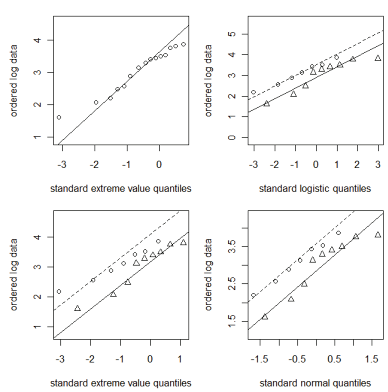


Figure: Porovnanie kvantilových diagramov (Weibullove modely M1 a M2, log-logistický model, log-normálny model)

Všeobecné zápisy regresných modelov
Príklad v R

- Neparametrický prístup (KM odhad) dáva lepší náhľad na AML dáta

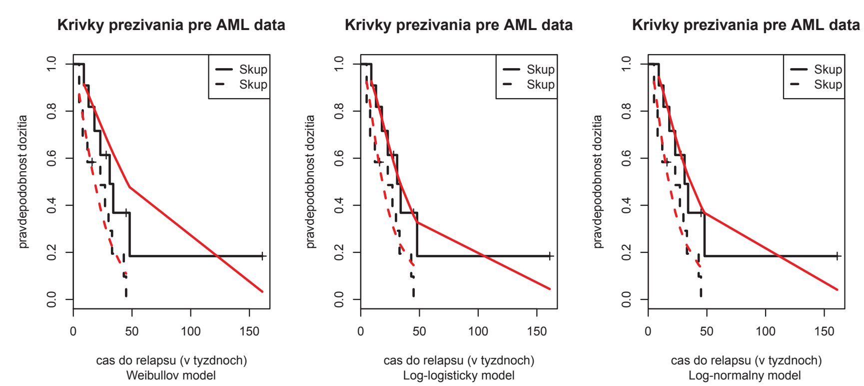


Figure: Porovnanie funkcií prežívania

- Majme Weibullov model. Nech $\widehat{\beta} = -\widehat{\beta}^*/\widehat{\sigma} = -0.929/0.791 = -1.1745$, potom odhad pomeru rizík $\widehat{\mathrm{HR}} = \widehat{\lambda}(t|1)/\widehat{\lambda}(t|0) = \exp(\widehat{\beta})/\exp(0) = \exp(\widehat{\beta}) = 0.31$. Skupina `Maitained` má 31% riziko z rizika kontrolnej skupiny `Nonmaitained`. Alebo kontrolná skupina má $1/0.31 = 3.23\times$ väčšie riziko ako skupina `Maitained`. HR je teda miera efektu vplyvu skupiny na čas prežívania (čas do relapsu alebo recidívy).
- Majme opäť Weibullov model. Ak vypočítame pomer odhadnutých pravdepodobností prežívania, napr. v čase $t = 31$ týždňov ($\widehat{\widehat{\lambda}} = \exp(-\widehat{\widehat{\mu}})$), dostaneme $\widehat{S}(31|1)/\widehat{S}(31|0) = 0.6531634/0.2517891 = 2.594$, čo znamená, že skupina `Maitained` má $2.594\times$ väčšiu pravdepodobnosť zostať v remisii (dočasné zlepšenie stavu) najmenej 31 týždňov ako skupina `Nonmaitained`.

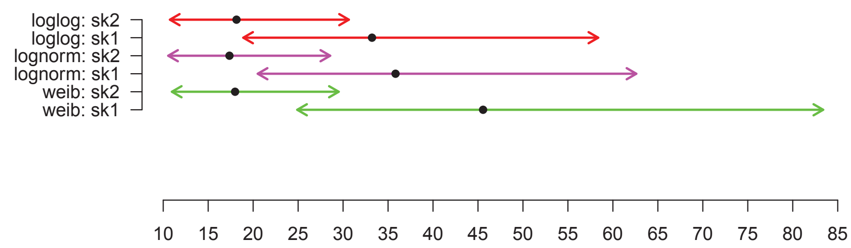


Figure: Porovnanie intervalov spoľahlivosti pre medián (Weibullov, log-logistický a log-normálny model)