Analýza prežívania

Neparametrické modely

Stanislav Katina[1]

[1]Ústav matematiky a statistiky, Masarykova univerzita
Honorary Research Fellow, The University of Glasgow

23. mája 2018

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. nech $X_1, X_2, \ldots, X_n$, 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 $(X_i, \delta_i)^T$, 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 $T^{(d)}$
6. nech $X_1, X_2, \ldots, X_n$, kde
$$X_i = \min(T_i, T^{(d)}) = \begin{cases} T_i, T_i \le T^{(d)}, & \text{pre necenzúrované } X_i \\ T^{(d)}, T_i > T^{(d)}, & \text{pre cenzúrované } X_i \end{cases}$$
7. skutočnému pozorovaniu potom zodpovedá náhodný vektor $(X_i, \delta_i)^T$, kde
$$\delta_i = \begin{cases} 1, T_i \le T^{(d)}, & \text{pre necenzúrované } X_i \\ 0, T_i > T^{(d)}, & \text{pre cenzúrované } X_i \end{cases}$$

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 časom** – 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, ..., 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 zlyhaním** – 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, ..., 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 $T^{(d_k)}$

Cenúrovanie
Náhodné a ľubovoľné cenzúrovanie

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. nech $X_1, X_2, \ldots, X_n$ je náhodný výber časov, kde
$$X_i = \min(T_i, C_i) = \begin{cases} T_i, T_i \le C_i, & \text{pre necenzúrované } X_i \\ C_i, T_i > C_i, & \text{pre cenzúrované } X_i \end{cases}$$
5. nech $(X_i, \delta_i)^T$ je náhodný vektor, kde $X_i = \min(T_i, C_i)$ a
$$\delta_i = \begin{cases} 1, T_i \le C_i, & \text{pre necenzúrované } X \\ 0, T_i > C_i, & \text{pre cenzúrované } X \end{cases}$$
6. **náhodná veličina** – čas trvania experimentu a čas do cenzúry (ak $C_i = c$, ide o **ľubovoľné cenzúrovanie**)

Základné princípy:

Majme n subjektov. Nech časy do zlyhania $T_i, i = 1, 2, \ldots, n$, sú náhodné premenné, ktorých realizácie **nedokážeme pozorovať**. Nech $(X_i, \delta_i)^T$ je náhodný vektor, kde $X_i = C_i$ 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. Kedže 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. Nech časy do zlyhania $T_i, i = 1, 2, \ldots, n$ sú náhodné premenné, ktorých realizácie **nedokážeme pozorovať**. Vieme len, že T_i nastalo buď vnútri nejakého náhodného časového intervalu, pred jeho ľavou hranicou alebo po jeho pravej hranici. Označme C_{i1} a C_{2i} časy dvoch vyšetrení a indikačné funkcie definujeme nasledovne $\delta_{i1} = I(T_i \leq C_{i1})$, $\delta_{i2} = I(C_{i1} < T_i \leq C_{i2})$ a $\delta_{i3} = I(T_i > C_{i2})$, t.j.

$$\delta_{i1} = \begin{cases} 1, T_i \leq C_{i1}, & \text{pre necenzúrované } X_i \\ 0, T_i > C_{i1}, & \text{pre cenzúrované } X_i \end{cases},$$

$$\delta_{i2} = \begin{cases} 1, C_{i1} < T_i \leq C_{i2}, & \text{pre necenzúrované } X_i \\ 0, T_i > C_{i2}, & \text{pre cenzúrované } X_i \end{cases}$$

a nakoniec $\delta_{i3} = 0$.

Example (nádor pľúc, pacienti)

Pacienti navštevovali kliniku opakovane každých 4 až 6 mesiacov, kde pozorovania sú buď intervaly $(C_{i1}, C_{i2}\rangle$ ak sa retrakcia prsníka vyskytla medzi poslednými dvoma návštevami alebo (C_{i2}, ∞), ak sa do C_{i2} retrakcia nevyskytla.

Základné princípy:

Máme nasledovné tri možnosti:

1. udalosť mohla nastať niekedy pred prvým vyšetrením C_{i1}, kde $\delta_{i1} = 1$ a $\delta_{i2} = \delta_{i3} = 0$,
2. udalosť mohla nastať niekedy medzi prvým a druhým vyšetrením, t.j. v intervale $(C_{i1}, C_{i2}\rangle$, kde $\delta_{i1} = 0$, $\delta_{i2} = 1$ a $\delta_{i3} = 0$,
3. udalosť sa do druhého vyšetrenia nevyskytla, t.j. mohla nastať niekedy po C_{i2} (ale nevieme kedy), kde $\delta_{i1} = 0$, $\delta_{i2} = 0$ a $\delta_{i3} = 0$.

Nech $X_{i1} = C_{i1}$ a $X_{2i} = C_{2i}$. Potom dostaneme náhodný vektor

$$(X_{i1}, X_{2i}, \delta_{i1}, \delta_{i2})^T.$$

Všimnime si, že δ_{i3} nie je potrebné použiť, pretože nemáme ďalšie vyšetrenie po C_{i2}. Keby sme mali C_{i3} alebo aj ďalšie (po ňom nasledujúce) vyšetrenia, hovorili by sme **zovšeobecnenom intervalovom cenzúrovaní**.

Nech T je nezáporná náhodná premenná ($T \geq 0$) reprezentujúca čas úmrtia (zlyhania) indivídua z homogénnej populácie. Rozdelenie pravdepodobnosti T môže byť charakterizované rôznym spôsobom. V analýze prežívania sa najčastejšie používajú:

1. **distribučná funkcia** (*distribution function*) $F(t)$
2. **funkcia hustoty** (hustota; *probability density function*) $f(t)$
3. **funkcia prežívania** (*survivor function*, *reliability function*) $S(t)$
4. **riziková funkcia** (funkcia rizika, intenzita úmrtnosti, riziko; *hazard function*, *hazard rate*, *risk*, *mortality rate*) $\lambda(t)$, v poistných aplikáciách $\mu(t)$
5. **kumulatívna riziková funkcia** (funkcia kumulatívneho rizika, kumulatívne riziko; *cumulative hazard function*) $\Lambda(t)$

Ďalej sa budeme zaoberať charakteristikami:

1. *p*-**ty kvantil** t_p rozdelenia T, špeciálne **medián času prežívania** (*median survival time*, *median survival*) $t_{0.5}$
2. **medián funkcie prežívania** $S(t_{0.5})$
3. **stredná hodnota času prežívania** (*mean survival*) μ
4. **stredná hodnota zostatkového života** v čase t (*mean residual life*) mrl(t)
5. **medián zostatkového života** v čase t (*median remaining lifetime*, *median residual life*) mrlt(t)

Distribučná funkcia

$$F(t) = \Pr(T \leq t) = \int_0^t f(x)\,dx = 1 - S(t)$$

Funkcia hustoty

$$\begin{aligned} f(t) &= \frac{\partial}{\partial t} F(t) = F(t + \Delta t) - F(t) \\ &= \frac{\partial}{\partial t}(1 - S(t)) = S(t) - S(t + \Delta t) \end{aligned}$$

Funkcia prežívania

$$S(t) = 1 - F(t) = \Pr(T > t) = \int_t^\infty f(x)dx$$

Riziková funkcia

$$\lambda(t) = \frac{f(t)}{S(t)} = -\frac{\frac{\partial}{\partial t} S(t)}{S(t)} = -\frac{\partial}{\partial t} \ln S(t)$$

Kumulatívna riziková funkcia

$$\Lambda(t) = \int_0^t \lambda(x)\,dx = -\ln S(t) - (-\ln S(0)) = -\ln S(t)$$

Funkcia prežívania vyjadrená pomocou rizika a kumulatívneho rizika

$$S(t) = e^{-\int_0^t \lambda(s)ds} = e^{-\Lambda(t)}$$

Stredná hodnota času prežívania

$$\mu = \int_0^\infty S(t)\,dt \text{ a často aj } \mu = \int_0^{t_{\max}} S(t)\,dt$$

Stredná hodnota zostatkového života v čase t

$$\text{mrl}(t) = E[T - t | T > t] = \frac{\int_t^\infty (u - t) f(u) du}{S(t)} = \frac{\int_t^\infty S(u) du}{S(t)}$$

Funkcia prežívania vyjadrená pomocou mrl(t)

$$S(t)=\frac{\mathrm{mrl}(0)}{\mathrm{mrl}(t)}\exp\left(-\int_0^t\frac{du}{\mathrm{mrl}(u)}\right)$$

Funkcia rizika vyjadrená pomocou mrl(t)

$$\lambda(t)=\left(\frac{\partial}{\partial t}\mathrm{mrl}(t)+1\right)\frac{1}{\mathrm{mrl}(t)}$$

Funkcia hustoty vyjadrená pomocou mrl(t)

$$f(t)=\left(\frac{\partial}{\partial t}\mathrm{mrl}(t)+1\right)\frac{\mathrm{mrl}(0)}{(\mathrm{mrl}(t))^2}\exp\left(-\int_0^t\frac{du}{\mathrm{mrl}(u)}\right)$$

Funkcia prežívania

$$S(t)=\prod_{i:t_i\leq t}\left(1-\lambda(t_i)\right),$$

Funkcia hustoty

$$f(t)=\lambda(t_i)\left[\prod_{j:t_j\leq t_{i-1}}\left(1-\lambda(t_j)\right)\right],\text{ kde } t_{i-1}<t_i\leq t$$

Kumulatívne riziko

$$\Lambda(t)=-\sum_{i:t_i\leq t}\ln\left(1-\lambda(t_i)\right)$$

Ak $\lambda(t_i)$ sú malé, potom $\ln(1-\lambda(t_i))\approx-\lambda(t_i)$ a naviac

$$\Lambda(t)=\sum_{i:t_i\leq t}\lambda(t_i)$$

Stredná hodnota času prežívania

$$\mu=\sum_{i=1}^{l}(t_i-t_{i-1})S(t_{i-1})=\sum_{i=0}^{l-1}t_{i+1}(S(t_i)-S(t_{i+1})),$$

kde $t_0=0$ a $l\leq n$ je počet rôznych zlyhaní; ak $t_{\max}<c_{\max}$, kde $c_{\max}$ je maximálnym časom do cenzúry, potom strednú hodnotu počítame na intervale $\langle 0,c_{\max}\rangle$

Stredná hodnota zostatkového života v čase t

$$\mathrm{mrl}(t)=\frac{(t_{i+1}-t)S(t_i)+\sum_{j:t_j\geq t_{i+1}}(t_{j+1}-t_j)S(t_j)}{S(t)},$$

kde $t_i\leq t<t_{i+1}$

Funkcia vierohodnosti pre jednotlivé typy cenzúrovania

- cenzúrovanie I. typu

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

- cenzúrovanie II. typu

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

- náhodné cenzúrovanie

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

Funkcia vierohodnosti
Pravé typy cenzúrovania

- intervalové cenzúrovanie I. typu

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

- intervalové cenzúrovanie II. typu

$$L = \prod_{i=1}^{n} \left[F(x_{i1})\right]^{\delta_{i1}} \left[F(x_{i2}) - F(x_{i1})\right]^{\delta_{i2}} \left[S_f(x_{i2})\right]^{\delta_{i3}},$$

kde $\delta_{i3} = 1 - \delta_{i1} - \delta_{i2}$

Funkcia vierohodnosti
Pravé typy cenzúrovania – funkcia vierohodnosti

Náhodné cenzúrovanie

$$\begin{aligned} L &= \prod_{i=1}^{n} f(x_i)^{\delta_i} S_f(x_i)^{1-\delta_i} = \prod_{i=1}^{l} [f(t_i)]^{d_i} \prod_{i=1}^{n_c} \prod_{j=1}^{n_i - n_{i+1} - d_i} S_f(c_{ij}) \\ &= \prod_{i=1}^{l} [f(t_i)]^{d_i} \prod_{j=1}^{l} \left[S(t_j)\right]^{n_j - n_{j+1} - d_j} \\ &= \prod_{i=1}^{l} [S(t_{i-1})\lambda(t_i)]^{d_i} \prod_{j=1}^{l} \left[S(t_j)\right]^{n_j - n_{j+1} - d_j} = \dots \\ &= \prod_{i=1}^{l} [\lambda(t_i)]^{d_i} [1 - \lambda(t_i)]^{n_i - d_i} \end{aligned}$$

Prehľad vzorcov
Charakteristiky definované sčítacím procesom

Vo formuláciách sčítacím procesom $(X_i, \delta_i)^T$ nahradíme $(N_i(t), Y_i(t))$, kde $N_i(t)$ je počet pozorovaných udalostí v intervale $\langle 0, t\rangle$ v jednotke i,

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

Táto formulácia obsahuje dáta s pravým typom cenzúr ako špeciálny prípad, teda $Y_i(t) = I(\{T_i > t\})$ a $N_i(t) = I(\{T_i \le t,\ \delta = 1\})$. Všimnime si, že $N(t)$ je zprava spojitá a $Y(t)$ zľava spojitá. $Y(t)$ je príkladom *predikovateľného procesu*, ktorého hodnoty v čase t sú známe nekonečne krátko pred t, v čase t^-, ak nie skôr. *Sčítací proces* je stochastický proces začínajúci v čase 0, ktorého trajektória je zprava spojitá funkcia so skokmi veľkosti 1. Pre $N(t)$ potom platí $\{N(t) : t \ge 0\}, N(0) = 0$.

Prehľad vzorcov
Charakteristiky definované sčítacím procesom

Odhad kumulatívneho rizika je definovaný na základe agregovaného procesu $\overline{Y}(t) = \sum_i Y_i(t), \overline{N}(t) = \sum_i N_i(t)$, $d\overline{N}(t) = \Delta\overline{N}(t) = \overline{N}(t) - \overline{N}(t^-)$, kde $\overline{N}(t)$ je suma udalostí do času t vrátane, $\overline{Y}(t)$ je počet jednotiek v riziku v čase t (formálne ide o počet jednotiek v riziku v časovom intervale $(t - \epsilon, t\rangle$ pre malé ϵ).

Example

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

Riešenie

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

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*

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

Prehľad vzorcov
Odhady funkcie prežívania

- *Kaplan-Meierov odhad funkcie prežívania*

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

- *Breslowov odhad funkcie prežívania*

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

- *Flemingom a Harringtonom modifikovaný Breslowov odhad funkcie prežívania*

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

Prehľad vzorcov
Odhady funkcie prežívania

- *Kaplan-Meierov odhad funkcie prežívania*

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

- *Breslowov odhad funkcie prežívania*

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

- *Flemingom a Harringtonom modifikovaný Breslowov odhad funkcie prežívania*

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

Prehľad vzorcov
Odhady rozptylu kumulatívneho rizika

- *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*

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

Prehľad vzorcov
Odhady rozptylu kumulatívneho rizika

- *Greenwoodov odhad rozptylu kumulatívneho rizika*

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

- *NA odhad rozptylu kumulatívneho rizika*

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

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

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

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*

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

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

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

Prehľad vzorcov
Odhad strednej hodnoty a jeho rozptyl

Odhad strednej hodnoty času do zlyhania $E[T]$ (**priemerný čas do zlyhania**) v spojitom prípade je definovaný ako

$$\widehat{\mu} = \int_0^{t_{\max}} \widehat{S}(t)\, dt,$$

v diskrétnom prípade

$$\widehat{\mu} = \sum_{i=0}^{t_{\max}-1} (t_{i+1} - t_i)\widehat{S}(t_i),$$

kde $t_0 = 0$ a $l \le n$ je počet rôznych zlyhaní a $t_l = t_{\max}$.

Odhad rozptylu priemerného času do zlyhania je definovaný ako

$$\widehat{Var}\left[\widehat{\mu}\right] = \int_0^{t_{\max}} \left[\int_t^{t_{\max}} \widehat{S}(u)\, du\right]^2 \widehat{\sigma}^2(t) dt$$

a v diskrétnom prípade

$$\widehat{Var}\,[\widehat{\mu}] = \sum_{i:t_i \le t_{\max}-1} \left[\sum_{j:t_i \le t_j \le t_{\max}-1} (t_{j+1} - t_j)\widehat{S}(t_j) \right]^2 \widehat{\sigma}^2(t_i).$$

Za $\widehat{S}\,(t)$ dosadíme $\widehat{S}_{\mathrm{KM}}\,(t)$, $\widehat{S}_{\mathrm{B}}\,(t)$ alebo $\widehat{S}_{\mathrm{FHmodB}}\,(t)$. Podobne ako pri rozptyle funkcie prežívania dostaneme päť nasledovných rozptylov

$$\widehat{Var_{\mathrm{G}}}\,[\widehat{\mu}_{\mathrm{KM}}] = \widehat{Var_{\mathrm{G}}}\,[\widehat{\mu}_{\mathrm{KMmod}}],\ \widehat{Var_{\mathrm{NA}}}\,[\widehat{\mu}_{\mathrm{B}}],\ \widehat{Var_{\mathrm{AJ}}}\,[\widehat{\mu}_{\mathrm{KM}}],$$

$$\widehat{Var_{\mathrm{NA}}}\,[\widehat{\mu}_{\mathrm{FHmodB}}] \text{ a } \widehat{Var_{\mathrm{G}}}\,[\widehat{\mu}_{\mathrm{FHmodB}}].$$

Odhad strednej hodnoty zostatkového života (**priemerný zostatkový život**) v čase t je definovaný v spojitom prípade ako

$$\widehat{\mathrm{mrl}}(t) = \frac{\int_t^{t_{\max}} \widehat{S}(u)du}{\widehat{S}(t)}$$

a v diskrétnom prípade ako

$$\widehat{\mathrm{mrl}}(t) = \frac{1}{\widehat{S}(t)} \left((t_{i+1} - t)\widehat{S}(t_i) + \sum_{j:t_{i+1} \le t_j \le t_{\max}-1} (t_{j+1} - t_j)\widehat{S}(t_j) \right),$$

kde $t_i \le t < t_{i+1}$. Za $\widehat{S}(\cdot)$ dosadíme $\widehat{S}_{\mathrm{KM}}(\cdot)$, $\widehat{S}_{\mathrm{B}}(\cdot)$ alebo $\widehat{S}_{\mathrm{FHmodB}}(\cdot)$.

Odhad rozptyl priemerného zostatkového života

$$\begin{aligned}\widehat{Var}[\widehat{\mathrm{mrl}}(t)] &= \frac{1}{\widehat{S}^2(t)} \left(\int_t^{t_{\max}} \left[\int_u^{t_{\max}} \widehat{S}(x)dx \right]^2 \widehat{\sigma}^2(u)\, du \right. \\ &\quad \left. + \left[\int_t^{t_{\max}} \widehat{S}(u)du \right]^2 \int_0^t \widehat{\sigma}^2(u)du \right),\end{aligned}$$

pre spojitý prípad, kde $u \in \langle t, t_{\max} \rangle$, a pre diskrétny prípad

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

Za $\widehat{S}\,(t)$ dosadíme $\widehat{S}_{\mathrm{KM}}\,(t)$, $\widehat{S}_{\mathrm{B}}\,(t)$ alebo $\widehat{S}_{\mathrm{FHmodB}}\,(t)$. Potom dostaneme $\widehat{\mathrm{mrl}}_{\mathrm{KM}}\,(t)$, $\widehat{\mathrm{mrl}}_{\mathrm{B}}\,(t)$ a $\widehat{\mathrm{mrl}}_{\mathrm{FHmodB}}\,(t)$ päť nasledovných rozptylov

$$\widehat{Var_{\mathrm{G}}}\left[\widehat{\mathrm{mrl}}_{\mathrm{KM}}\right] = \widehat{Var_{\mathrm{G}}}\left[\widehat{\mathrm{mrl}}_{\mathrm{KMmod}}\right],\ \widehat{Var_{\mathrm{NA}}}\left[\widehat{\mathrm{mrl}}_{\mathrm{B}}\right],\ \widehat{Var_{\mathrm{AJ}}}\left[\widehat{\mathrm{mrl}}_{\mathrm{KM}}\right],$$

$$\widehat{Var_{\mathrm{NA}}}\left[\widehat{\mathrm{mrl}}_{\mathrm{FHmodB}}\right] \text{ a } \widehat{Var_{\mathrm{G}}}\left[\widehat{\mathrm{mrl}}_{\mathrm{FHmodB}}\right].$$

1. **Waldov princíp**, **skóre princíp** a **vierohodnostný princíp** (všetky tri vychádzajúce z funkcie vierohodnosti),
2. **princíp transformácie funkcie vierohodnosti v** $S(t)$ **pomocou nejakej funkcie** $g(S(t))$ aplikovaný na Waldov princíp (hranice IS vypočítané pomocou $g(S(t))$ sa spätne transformujú do škály $S(t)$), kde podľa $g(S(t))$ rozoznávame nasledovné škály
3. princíp **úpravy hraníc IS pomocou efektívneho rozsahu súboru v čase** t ($n^*(t)$ alebo $n^{**}(t)$) z dôvodu výskytu cenzúr v dátach (resp. rozdielu medzi rozptylom $S(t)$ v čase t vypočítaným pre cenzúrované dáta a rozptylom za predpokladu, že cenzúry v dátach nie sú)
4. **princíp korekcie hraníc IS z dôvodu zlých štatistických vlastností** (konzervatívny alebo liberálny IS, t.j. predpokladáme, že nominálny koeficient spoľahlivosti $1-\alpha$ je iný ako teoretický)

- $g(S(t)) = S(t)$ – **škála funkcie prežívania**
- $g(S(t)) = \ln S(t)$ – **škála logaritmu funkcie prežívania** (**log škála funkcie prežívania**; **škála kumulatívneho rizika**) – spätná transformácia pre funkciu prežívania $\exp(\ln S(t))$,
- $g(S(t)) = \ln(-\ln S(t)) = \ln \Lambda(t)$ – **log-log škála funkcie prežívania** (**log škála kumulatívneho rizika**; **škála logaritmu kumulatívneho rizika**) – spätná transformácia $\exp(\exp(\ln(-\ln S(t))))$,
- $g(S(t)) = \arcsin\left((S(t))^{1/2}\right)$ – **arcus sínus škála funkcie prežívania** – spätná transformácia $\left(\sin\left(\arcsin\left((S(t))^{1/2}\right)\right)\right)^2$
- $g(\Lambda(t)) = \arcsin\left(\exp\left(-\frac{\Lambda(t)}{2}\right)\right) = \arcsin\left(\exp\left(\frac{\ln S(t)}{2}\right)\right)$ – **arcus sínus škála kumulatívneho rizika** – spätná transformácia $-2\ln\left(\sin\left(\arcsin\left(\exp\left(-\frac{\Lambda(t)}{2}\right)\right)\right)\right)$


```
surv.obj <- survfit(Surv(cas,status)~1,
type="...",error="...",conf.type ="..."))
```

Argumenty:

- **odhady funkcie prežívania** vypočítame ako
  1. $\widehat{S}_{\text{KM}}(t)$: `type="kaplan-meier"` (default)
  2. $\widehat{S}_{\text{B}}(t)$: `type="fleming-harrington"`
  3. $\widehat{S}_{\text{FHmodB}}(t)$: `type="fh2"`
- **odhady rozptylu** ako
  1. $\widehat{Var_G\left[\widehat{S}_{\text{KM}}(t)\right]}$: `error="greenwood"` (default)
  2. $\widehat{Var_G\left[\widehat{S}_{\text{B}}(t)\right]}$: `error="tsiatis"`
- **typy intervalov spoľahlivosti** (IS) ako
  1. žiadny: `conf.type="none"`
  2. škála funkcie prežívania: `conf.type="plain"`
  3. škála kumulatívneho rizika: `conf.type ="log"` (default)
  4. škála log. kumulatívneho rizika: `conf.type="log-log"`

Ďalšími argumentami sú:

- koeficient spoľahlivosti `conf.int=0.95` (default)
- úprava spodnej hranice IS, kde argument
  1. `conf.lower="usual"` (nemodifikovaná dolná hranica IS)
  2. `conf.lower="peto"` (používa Petov efektívny rozsah súboru $n^{**}(t)$)
  3. `conf.lower= "modified"` (používa Dorey-Korn modifikáciu)

Priemerný vek prežívania a jeho smerodajná odchýlka ako aj **medián a jeho smerodajná odchýlka** sa vypočítajú 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

1. `type="right"` (pravý typ), `type="interval"` (intervalový typ cenzúrovania I. typu, kde interval $(-\infty, t_i\rangle$ označujeme (NA, $t_i\rangle$)
2. `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 $t_{\max}$.

Výstupmi objektu `surv.obj` a `summary(surv.obj)` sú nasledovné:

1. rozsah – n
2. počet jedincov v riziku v jednotlivých časoch – `n.risk`
3. počet udalostí (zlyhaní) v jednotlivých časoch – `n.event`
4. počet cenzúr v jednotlivých časoch – `n.censor`
5. odhad funkcie prežívania v jednotlivých časoch – `surv`
6. odhad odmocniny z rozptylu funkcie prežívania v jednotlivých časoch – `std.err`
7. dolná hranica IS pre funkciu prežívania v jednotlivých časoch – `lower`
8. horná hranica IS pre funkciu prežívania v jednotlivých časoch – `upper`
9. časy, v ktorých nastalo
   - zlyhanie alebo cenzúra – `time` (pre `surv.obj`)
   - zlyhanie – `time` (pre `summary(surv.obj)`)

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, zovšeobecnený Wilcoxonov test (GW)
  - 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, zovšeobecnený Wilcoxonov test, zovšeobecnený Kruskal-Wallis test (GB)
  - Cox-Mantel test, log-rank test (CM)
  - Mantel-Haenszel test, log-rank test (MH)
  - Peto-Peto test (PP)

Testované hypotézy

- **nulovná hypotéza** $H_0 : S_1(t) = S_2(t) = S(t)$
- **alternatíva hypotéza** H_1 :
  - $S_1(t) \neq S_2(t)$, pre $\forall t$
  - $S_1 \overset{st}{\prec} S_2$, čo je ekvivalentné s $S_1(t) < S_2(t)$, pre $\forall t$
  - $S_1 \overset{st}{\succ} S_2$, čo je ekvivalentné s $S_1(t) > S_2(t)$, pre $\forall t$

$S(t)$ je funkcia prežívania

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

Predpoklady

- $X_1, \ldots, X_{n_1}$ je náhodný výber (NV) z nejakého spojitého rozdelenia
- $Y_1, \ldots, Y_{n_2}$ je NV z rovnakého spojitého rozdelenia a je oproti prvému rozdeleniu posunuté o nejakú konštantu δ
- veličiny $X_1, \ldots, X_{n_1}$ a $Y_1 - \delta, \ldots, Y_{n_2} - \delta$ 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 $X_1, X_2, \ldots, X_{n_1}$ prvého NV v rámci usporiadaného združeného NV
- ich realizácie $r_1, r_2, \ldots, r_{n_1}$ sú poradia realizácií $x_1, x_2, \ldots, x_{n_1}$

Wilcoxonova štatistika

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

Realizáciu S_W ozn. $w_X = s_W = \sum_{i=1}^{n_1} r_i$.

Stredná hodnota a rozptyl S_W :

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

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

Wilcoxonova testovacia štatistika

Ak $n_1, n_2 \geq 10$

$$Z_W = \frac{S_W - E_0[S_W]}{\sqrt{\widehat{Var_0}[S_W]}} \overset{\mathcal{D}}{\sim} N(0,1)$$

Realizáciu Z_W ozn. z_W.

Stredná hodnota a rozptyl S_W:

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

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

Wilcoxonova testovacia štatistika

Ak $n_1, n_2 \geq 10$

$$\widetilde{Z}_W = \frac{S_W - E_0[S_W]}{\sqrt{\widehat{Var_0}[S_W] - \frac{n_1 n_2 \sum_j \left(t_j^3 - t_j\right)}{12(n_1+n_2)(n_1+n_2-1)}}} \overset{\mathcal{D}}{\sim} N(0,1)$$

Realizáciu $\widetilde{Z}_W$ ozn. $\widetilde{z}_W$.

Predpoklady

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

Hypotézy

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

Označenia

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

Mann-Whitneyho štatistika

$$S_{MW} = \sum_{i=1}^{n_1}\sum_{j=1}^{n_2} I(X_i > Y_j) = \#\left(X_i, Y_j\right), \text{ kde } X_i > Y_j$$

Realizáciu S_{MW} ozn. $s_{MW} = \sum_{i=1}^{n_1}\sum_{j=1}^{n_2} I(x_i > y_j)$.

Stredná hodnota a rozptyl S_{MW} :

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

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

Mann-Whitneyho testovacia štatistika

Ak $n_1, n_2 \geq 10$

$$Z_{MW} = \frac{S_{MW} - E_0[S_{MW}]}{\sqrt{\widehat{Var_0}[S_{MW}]}} \overset{\mathcal{D}}{\sim} N(0,1)$$

Realizáciu Z_{MW} ozn. z_{MW}.

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

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

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

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

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

Wilcoxonov a Mann-Whitneyho test

Example (nádor pľúc; cvič.)

Nech $t_{ij}, i = 1, \ldots, n_j, j = 1, 2$ 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 (pozri tabuľku). Otestujte $H_0 : S_1(t) = S_2(t)$ oproti alternatíve $H_1 : S_1(t) \neq S_2(t)$ pomocou S_W a S_{MW} nesledovne (1) $S_W = W_Y$ a $S_W = W_X$, (2) $S_{MW} = U_Y$ a $S_{MW} = U_X$. Vždy presne naformulujte H_1. Okomentujte výsledky.

t_{ij}	52	240	19	53	15	43	340	133	111	231	378	49
skup	1	2	2	1	1	2	2	1	1	2	1	1

Wilcoxonov vs Mann-Whitneyho test

Example (Wilcoxonov vs Mann-Whitneyho test; DÚ)

Nech U_X vyjadruje počet dvojíc X_i, Y_j, kde platí $X_i < Y_j$, podobne U_Y vyjadruje počet dvojíc X_i, Y_j, kde platí $X_i > Y_j$. Dokážte, že

$$U_X = n_1 n_2 + \frac{n_1 (n_1 + 1)}{2} - W_X, U_Y = n_1 n_2 + \frac{n_2 (n_2 + 1)}{2} - W_Y.$$

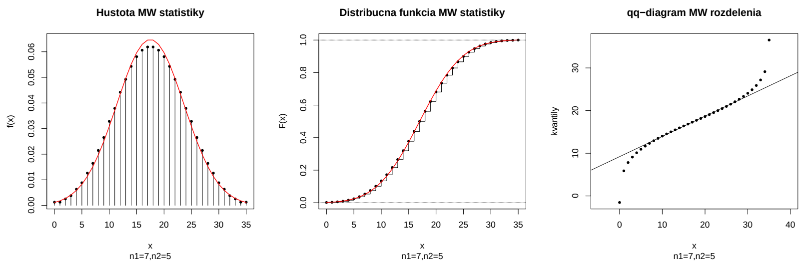
Pozn.: Ekvivalentne sa dá ukázať, že $W_X = U_Y + \frac{n_1(n_1+1)}{2}$ (podobne pre W_Y a U_X) a dosadiť do vzorcov pre U_X a U_Y.

Asymptotická normalita rozdelenia S_{MW}

Example (asymptotická normalita S_{MW})

Pre $n_1 = 7$ a $n_2 = 5$ porovnajte v R asymptotické rozdelenie S_{MW} s jej exaktným rozdelením. Na výpočet asymptotickej hustoty použite funkciu `dnorm()` a na výpočet asymptotickej distribučnej funkcie použite funkcie `dnorm()` a `cumsum()`. Na výpočet exaktnej hustoty použite funkciu `dwilcox()` a na výpočet exaktnej distribučnej funkcie použite funkciu `pwilcox()`. Teoretické a exaktné rozdelenie superponujte v podobe (1) hustoty, (2) distribučnej funkcie a (3) qq-diagramu s qq-priamkou (na x-ovej osi bude sekvencia x od teoreticky možného $\min(S_{MW})$ po teoreticky možné $\max(S_{MW})$ a na y-ovej osi teoretické kvantily y vypočítané pomocou funkcie `qnorm()`; qq-priamka bude prechádzať bodmi $(\widetilde{x}_{0.25}, \widetilde{y}_{0.25})$ a $(\widetilde{x}_{0.75}, \widetilde{y}_{0.75})$.

Testy na porovnanie dvoch kriviek prežívania
Asymptotická normalita rozdelenia S_{MW}



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

Example (asymptotická normalita S_{MW})

Porovnajte v R asymptotické rozdelenie S_{MW} s jej exaktným rozdelením pre (1) $n_1 = 5$ a $n_2 = 50$, (2) $n_1 = 50$ a $n_2 = 50$, (3) $n_1 = 50$ a $n_2 = 100$ a (4) $n_1 = 100$ a $n_2 = 100$.

Testy na porovnanie dvoch kriviek prežívania
Siegel-Tukey a Levene test

Predpoklady

- liečba nespôsobuje zmenu priemernej odpovede, ale výsledná odpoveď má menší rozptyl okolo strednej hodnoty
- $X_1, ..., X_{n_1}$ je NV z nejakého spojitého rozdelenia
- $Y_1, ..., Y_{n_2}$ je 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

Testy na porovnanie dvoch kriviek prežívania
Siegel-Tukey test

Podstata Siegel-Tukey testu je nasledovná

Algoritmus 1:

- 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. Realizáciu S_{ST} ozn. $s_{ST} = \sum_{i=1}^{n_1} r_i$.

Testy na porovnanie dvoch kriviek prežívania
Siegel-Tukey test

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

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

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

Siegel-Tukey testovacia štatistika

Ak $n_1, n_2 \geq 10$

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

Realizáciu $Z_{ST} = Z_{ST}^{alt}$ ozn. $z_{ST} = z_{ST}^{alt}$.

Testy na porovnanie dvoch kriviek prežívania
Levene test

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

- odchýlky $D_X = \left|X - \overline{X}\right|$ a $D_Y = \left|Y - \overline{Y}\right|$
- $D_{(1)} < D_{(2)} < \ldots < D_{(n)}, n = n_1 + n_2$
- realizácie odchýlok $\{d_i = |x_i - \overline{x}|\}_{i=1}^{n_1}$ a $\{d_j = |y_j - \overline{y}|\}_{j=1}^{n_2}$
- $d_{(1)} < d_{(2)} < \ldots < d_{(n)}$

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. Realizáciu S_{ST}^{alt} ozn. s_{ST}^{alt}.

Testy na porovnanie dvoch kriviek prežívania
Levene test

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

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

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

Levene testovacia štatistika

Ak $n_1, n_2 \geq 10$

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

Realizáciu $Z_L = Z_{ST}^{alt}$ ozn. $z_L = z_{ST}^{alt}$.

Testy na porovnanie dvoch kriviek prežívania
Siegel-Tukey test a Levene test

Example (nádor pľúc pokrač., cvič.)

Nech $t_{ij}, i = 1, \ldots, n_j, j = 1, 2$ 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 (pozri tabuľku). Otestujte v R $H_0 : Var[S_1(t)] = Var[S_2(t)]$ oproti alternatíve $H_1 : Var[S_1(t)] \neq Var[S_2(t)]$ pomocou S_{ST} a S_{ST}^{alt}. Okomentujte výsledky.

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

Siegel-Tukey test

$s_{ST} = \sum_{i=1}^{5} r_i^{(2)} = 26$

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

$Z_{ST} = (26 - 32.5)/6.157651 = -3.167$ a p-hodnota=0.291

Testy na porovnanie dvoch kriviek prežívania
Siegel-Tukey test a Levene test

Example (Siegel-Tukey test a Levene test)

Naprogramujte v R test $H_0 : Var[S_1(t)] = Var[S_2(t)]$ oproti alternatíve $H_1 : Var[S_1(t)] \neq Var[S_2(t)]$ pomocou S_{ST} a S_{ST}^{alt} použitím algoritmu 2. Okomentujte výsledky.

Algoritmus 2:

- poradie R_1 priradíme najmenšiemu pozorovaniu
- poradia R_2 a R_3 priradíme dvom najväčším pozorovaniam
- poradia R_4 a R_5 priradíme druhému a tretiemu najmenšiemu pozorovaniu
- atď.

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)
  - $S_i \overset{st}{\prec} S_j$ (čo je ekv. s $S_i(t) < S_j(t)$, pre $\forall t$; J,C,L testy)
  - $S_i \overset{st}{\succ} S_j$ (čo je ekv. s $S_i(t) > S_j(t)$, pre $\forall t$; J,C,L testy)

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

Označenia

- $X_{j1}, \ldots, X_{jn_j}$ je j-ty NV, $i = 1, 2, \ldots, n_j; j = 1, 2, \ldots, k$
- R_{ji} poradia X_{ji} v združenom výbere

Ďalšie označenia

- $n = \sum_{j=1}^{k} n_j$, n_j je počet pozorovaní v j-tom NV
- $S_j = \sum_{i=1}^{n_j} R_{ji}$, $S = \sum_{j=1}^{k} R_j$, $\overline{S}_j = S_j / n_j$,
  $\overline{S} = S/n = (n+1)/2$

Kruskall-Wallisova testovacia štatistika

$$
\begin{aligned}
U_{KW} &= \frac{12}{n(n+1)} \sum_{j=1}^{k} n_j \left(\frac{S_j}{n_j} - \frac{n+1}{2} \right)^2 \\
&= \frac{12}{n(n+1)} \sum_{j=1}^{k} \frac{S_j^2}{n_j} - 3(n+1) \\
&= \frac{1}{\widehat{Var}[R]} \left(\sum_{j=1}^{k} \frac{S_j^2}{n_j} - \frac{n(n+1)^2}{4} \right) \overset{\mathcal{D}}{\approx} \chi^2_{k-1}
\end{aligned}
$$

Realizáciou U_{KW} je $u_{KW} = \frac{12}{n(n+1)} \sum_{j=1}^{k} n_j \left(\frac{s_j}{n_j} - \frac{n+1}{2} \right)^2$.

Rozptyl poradí R :

$$
\widehat{Var}[R] = \frac{1}{n-1} \sum_{j=1}^{k} \sum_{i=1}^{n_j} \left(S_{ji} - \frac{n+1}{2} \right)^2
$$

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

Kruskall-Wallisova testovacia štatistika

$$
\widetilde{U}_{KW} = \frac{1}{\widehat{Var}[R|\mathbf{t}]} \sum_{j=1}^{k} n_j \left(\frac{S_j}{n_j} - \frac{n+1}{2} \right)^2 \overset{\mathcal{D}}{\approx} \chi^2_{k-1}
$$

Realizáciou $\widetilde{U}_{KW}$ je $\widetilde{u}_{KW}$.

Označenia

- $i < j$, teda $i = 1, 2, \ldots, k-1; j = 1+i, \ldots, k$, ďalej
  $\alpha_i = 1, 2, \ldots, n_i$, $\alpha_j = 1, 2, \ldots, n_j$
- 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

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

Realizáciami štatistík S_{MW}^{ij} a S_J sú s_{MW}^{ij} a s_J.

Stredná hodnota a rozptyl S_J :

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

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

Jonckheereho testovacia štatistika

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

Realizáciu Z_J ozn. z_J.

Alternatívna podoba Jonckheereho štatistiky

$$S_J^{alt} = 2\sum_{i=1}^{k-1}\sum_{j=1+i}^{k} S_{MW}^{ij} - \sum_{i=1}^{k-1}\sum_{j=1+i}^{k} n_i n_j,$$

kde $\sum_{i=1}^{k-1}\sum_{j=1+i}^{k} n_i n_j = \max_{\forall S_J} S_J$
Stredná hodnota a rozptyl S_J^{alt} :

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

Alternatívna podoba Jonckheereho testovacej štatistiky

$$Z_J^{alt} = \frac{S_J^{alt} - E_0\left[S_J^{alt}\right]}{\sqrt{\widehat{Var_0\left[S_J^{alt}\right]}}} \overset{\mathcal{D}}{\sim} N\left(0,1\right)$$

Realizáciami S_J^{alt} a Z_J^{alt} sú s_J^{alt} a z_J^{alt}.

Označenia

- nech $\left(X_1, Y_1\right)^T, \ldots, \left(X_n, Y_n\right)^T$ NV z dvojrozmerného rozdelenia
- dvojicu s indexami i a j, $\left(X_i, Y_i\right)$ a $\left(X_j, Y_j\right)$, 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 $Y_i = Y_j$ 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
- $\widetilde{\tau} = \frac{C-D}{\frac{n(n-1)}{2}}$

Kendallov korelačný koeficient

$$\tau = \tfrac{1}{n(n-1)}\textstyle\sum_{i=1}^{n}\sum_{j=1}^{n} sign\left(X_i - X_j\right) sign\left(Y_i - Y_j\right),$$

čo je identické s

$$\widetilde{\tau} = \tfrac{2}{n(n-1)}\textstyle\sum_{i=1}^{n-1}\sum_{j=i+1}^{n} sign\left(X_i - X_j\right) sign\left(Y_i - Y_j\right),$$

kde

$$sign(X_i - X_j) = \begin{cases} 1, & \text{ak } X_i > X_j \\ -1, & \text{ak } X_i < X_j \\ 0, & \text{ak } X_i = X_j \end{cases}$$

$$sign(Y_i - Y_j) = \begin{cases} 1, & \text{ak } Y_i > Y_j \\ -1, & \text{ak } Y_i < Y_j \\ 0, & \text{ak } Y_i = Y_j \end{cases}$$

Alternatíva Kendallovho korelačného koeficientu

- nech $R_1, \ldots, R_n$ sú poradia $X_1, \ldots, X_n$
- nech $Q_1, \ldots, Q_n$ sú poradia $Y_1, \ldots, Y_n$

Potom platí

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

to 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

$$\widetilde{\tau} \in \langle -1, 1 \rangle, E_0[\widetilde{\tau}] = 0, \widehat{Var_0[\widetilde{\tau}]} = \frac{2(2n+5)}{9n(n-1)}$$

a

$$Z_{\widetilde{\tau}} = \frac{\widetilde{\tau} - E_0[\widetilde{\tau}]}{\sqrt{\widehat{Var_0[\widetilde{\tau}]}}} \overset{\mathcal{D}}{\sim} N(0,1)$$

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 $\arcsin(\cdot)$ nadobúda hodnoty z $\langle -\frac{\pi}{2}, \frac{\pi}{2} \rangle$

Vzťah Kendallovho korelačného koeficientu a Jonckheere štatistiky

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

kde τ nazývame **zovšeobecnený Kendallov korelačný koeficient**

Example (WBC; predn.)

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 pomocou počtu konkordantných a diskonkordantných párov. Otestujte nezávislosť medzi počtom bielych krviniek a časmi do zlyhania na hladine významnosti $\alpha = 0.05$.

WBC	t_i	c_{ji}	d_{ji}
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

$c_{ji} = \#(\uparrow t_i, \uparrow WBC_i)$ pod i
$c = \sum c_{ji} = 33$
$d_{ji} = \#(\downarrow t_i, \uparrow WBC_i)$ pod i
$d = \sum d_{ji} = 101$
$\frac{n(n-1)}{2} = \frac{17 \times 16}{2}$
$\widehat{\tau} = \frac{c-d}{\frac{n(n-1)}{2}} = -0.5$
$\widehat{Var[\widehat{\tau}]} = \frac{2(2 \times 17 + 5)}{9 \times 17(17-1)} = 0.032$
$z_{\widehat{\tau}} = -0.5/\sqrt{0.032} = -2.801$
p-hodnota=0.005

Označenia

- nech R_{ji} je poradie X_{ji} v združenom NV
- nech s_{ji} je *skóre* prislúchajúce NV, do ktorého X_{ji} patrí
- $n = \sum_{j=1}^{k} n_j$

Cuzickova štatistika

$$S_C = \sum_{j=1}^{k} \sum_{i=1}^{n_j} s_{ji} R_{ji},$$

Realizácia S_C je $s_C = \sum_{j=1}^{k} \sum_{i=1}^{n_j} s_{ji} r_{ji}$.

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

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

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

Cuzickova testovacia štatistika

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

Realizácia Z_C je z_C.

Ak máme v pozorovaniach zhody, potom **Cuzickova testovacia štatistika**

$$\widetilde{Z}_C = \frac{S_C - E_0\left[S_C\right]}{\sqrt{\widehat{Var_0\left[S_C\right]} - \frac{1}{n(n^2-1)} \sum_j t_j \left(t_j^2 - 1\right)}} \overset{\mathcal{D}}{\sim} N\left(0, 1\right)$$

$Var_0\left[S_C|\mathbf{t}\right] = Var_0\left[S_{KW}|\mathbf{t}\right]$
Realizácia $\widetilde{Z}_C$ je $\widetilde{z}_C$.

Označenia

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

Spearmanova štatistika

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

Realizáciu S_N ozn. $s_N = \sum_{i=1}^{n} r_i q_i$.

Stredná hodnota

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

Rozptyl

$$\widehat{Var_0}\left[S_N\right] = \frac{1}{n-1}\left(\frac{n(n^2-1)}{12}\right)^2$$

Spearmanova testovacia štatistika

$$Z_S = \frac{S_N - E_0\left[S_N\right]}{\sqrt{\widehat{Var_0}\left[S_N\right]}} \overset{\mathcal{D}}{\sim} N(0,1), \text{čo je ekv. } \sqrt{n-1}R_S \overset{\mathcal{D}}{\sim} N(0,1),$$

kde $R_S = \frac{1}{\sqrt{(n-1)\widehat{Var_0}[S_N]}}\left(S_N - E_0\left[S_N\right]\right)$, $E_0\left[R_S\right] = 0$,
$Var_0\left[R_S\right] = \frac{1}{n-1}$
Realizácie R_S a Z_S ozn. r_S a z_S.

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

Example (pokrač. WBC)

Vypočítajte Spearmanov korelačný koeficient r_S. Otestujte nezávislosť medzi počtom bielych krviniek a časmi do zlyhania pomocou Z_S na hladine významnosti $\alpha = 0.05$.

Označenia

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

Le štatistika

$$S_L = \sum_{j=1}^{k} n_j\left(L_j - M_j\right)\overline{R}_j$$

Realizáciu S_L ozn. $s_L = \sum_{j=1}^{k} n_j\left(l_j - m_j\right)\overline{r}_j$.

Stredná hodnota

$$E_0\left[S_L\right] = 0$$

Rozptyl

$$\widehat{Var_0}\left[S_L\right] = \frac{n(n+1)}{12}\sum_{j=1}^{k} n_j\left(L_j - M_j\right)^2$$

Le testovacia štatistika

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

Realizáciu Z_L ozn. z_L.

Testovacia štatistika odklonu od trendu

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

Všeobecný tvar štatistiky

$$S_A = \sum_{j=1}^{k} n_j s_j R_j,$$

kde n_j sú rozsahy jednotlivých NV, s_j skóre prislúchajúce jednotlivým NV a R_j sú priemerné charakteristiky polohy prislúchajúce jednotlivým NV
Potom

$$Z_A = \frac{\left(S_A - E_0[S_A]\right)^2}{\widehat{Var_0}\left[S_A\right]} \overset{\mathcal{D}}{\sim} \chi^2_1$$

Realizáciu Z_A ozn. z_A.

Example (nádor pľúc pokrač.)

Nech $t_{ij}, i = 1, \ldots, n_j, j = 1, 2$ 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 (pozri tabuľku). Otestujte $H_0 : S_1(t) = S_2(t)$, alternatíva $H_1 : S_1(t) \neq S_2(t)$. Použite (1) S_{KW}, (2) S_J, (3) S_C a (4) S_L. Vždy presne naformulujte H_1. Prečo nemôžeme testovať odkolon od trendu? Aký je vzťah medzi týmito testovacími štatistikami a testovacími štatistikami S_W a S_{MW} pre dva výbery?

t_{ij}	52	240	19	53	15	43	340	133	111	231	378	49
skup	1	2	2	1	1	2	2	1	1	2	1	1

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

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

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

$S_L = \sum_{j=1}^{3} n_j \left(L_j - M_j\right) \overline{R}_j =$
$3 \times (0-14) \times 4.5 + 6 \times (3-8) \times 7.8\overline{33} + 8 \times (9-0) \times 11.5625 = 408.5$

$\widehat{Var}\left[S_L\right] = \frac{n(n+1)}{12} \sum_{j=1}^{k} n_j \left(L_j - M_j\right)^2 =$
$\frac{17(17+1)}{12} \left[3 \times (0-14)^2 + 6 \times (3-8)^2\right.$
$\left. + 8 \times (9-0)^2\right] = 187.9973^2$

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

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

Tarone a Ware (1977) trieda váh (Cochran, 1954; Mantel a Haenszel, 1959; Armitage, 1966)

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 $w(t) = (1/\overline{Y}(t) \times (1 - 1/\overline{Y}(t)))^{-1} = \overline{Y}^2(t)/(\overline{Y}(t) - 1) \approx \overline{Y}(t)$
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

$$T(w,t) = \sum_{j=0}^{L} w_j \left(d_{1j} - d_j \frac{n_{1j}}{n_k} \right), L = l \leq n$$

potom

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

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

KT 2×2 (označenia typické v epidemiológii – vľavo, označenia typické v analýze prežívania – vpravo)

	y_1	y_2	$\sum$		y_1	y_2	$\sum$
x_1	a	b	$n_{1\cdot}$	x_1	d_1	d_2	d
x_2	c	d	$n_{2\cdot}$	x_2	a_1	a_2	a
$\sum$	$n_{\cdot 1}$	$n_{\cdot 2}$	n	$\sum$	n_1	n_2	n

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

$$\chi^2 = \left(\frac{d_1 - E_0[D]}{\sqrt{\widehat{Var_0[D]}}} \right)^2 \overset{\mathcal{D}}{\sim} \chi^2_{df}, df = 1,$$

kde d_1 je početnosť v prvej bunke KT.

Testy na porovnanie dvoch a viac kriviek prežívania
Kontingenčné tabuľky

$x \backslash y$	y_1	y_2	$\sum$	$x \backslash y$	y_1	y_2	$\sum$
x_1	p_{11}	p_{12}	$p_{1\cdot}$	x_1	n_{11}	n_{12}	$n_{1\cdot}$
x_2	p_{21}	p_{22}	$p_{2\cdot}$	x_2	n_{21}	n_{22}	$n_{2\cdot}$
$\sum$	$p_{\cdot 1}$	$p_{\cdot 2}$	1	$\sum$	$n_{\cdot 1}$	$n_{\cdot 2}$	n

- predpoklad **binomického/multinomického rozdelenia** – fixované riadkové marginálne početnosti
- predpoklad **Poissonovho rozdelenia** – žiadne marginálne početnosti fixované
- predpoklad **hypergeometrického rozdelenia** – fixované všetky marginálne početnosti

Testované efekty

- **rozdiel pravdepodobností** $p_{11} - p_{12}$
- **pomer rizík** RR $= \frac{n_{11}/n_{1\cdot}}{n_{21}/n_{2\cdot}} = \frac{p_{11}}{p_{21}}$
- **pomer šancí** OR $= \frac{p_{11}/p_{12}}{p_{21}/p_{22}} = \frac{n_{11}/n_{12}}{n_{21}/n_{22}}$

Testy na porovnanie dvoch a viac kriviek prežívania
Kontingenčné tabuľky

Kombinovanie L ($L = I$) 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)

$$\chi^2 = \left(\frac{\sum_{i=1}^{L} (d_{1i} - E_0[D_i])}{\sqrt{\sum_{i=1}^{L} \widehat{Var_0[D_i]}}} \right)^2 \overset{\mathcal{D}}{\sim} \chi^2_{df}, df = k - 1.$$

Testy na porovnanie dvoch kriviek prežívania
Prehľad testov

Pre každé $t_i, 1 \leq i \leq I$, môžeme dáta zapísať do KT 2×2

výber/status	1	2	spolu v t_i
zlyhanie v t_i	d_{1i}	d_{2i}	d_i
nažive v t_i	a_{1i}	a_{2i}	a_i
spolu v t_i	n_{1i}	n_{2i}	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 $d = \sum_{j: t_j \leq t_i} d_j$

Testy na porovnanie dvoch kriviek prežívania
Prehľad testov

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

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

- 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**

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

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

$$\chi_i^2 = \frac{[d_{1i} - E_0[d_{1i}]]^2}{\widehat{Var_0}[d_{1i}]} \overset{\mathcal{D}}{\sim} \chi_{df}^2, df = 1.$$

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

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

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

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

$$Q = \frac{\left[\sum_{i=1}^{I}(d_{1i} - E_0[d_{1i}])\right]^2}{\sum_{i=1}^{I}\widehat{Var_0}[d_{1i}]} = \frac{U^2}{\widehat{Var_0}[U]} \overset{\mathcal{D}}{\sim} \chi_{df}^2, df = 1,$$

$$Q = \frac{(U - E_0[U])^2}{\widehat{Var_0}[U]} \sim \chi_1^2, Z_Q = \frac{U - E_0[U]}{\sqrt{\widehat{Var_0}[U]}} \overset{\mathcal{D}}{\sim} N(0,1).$$

Majme váhy w_i asociované s KT v čase t_i, potom

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

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

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

$$Q = \frac{(U - E_0[U])^2}{\widehat{Var_0}[U]} \overset{\mathcal{D}}{\sim} \chi_1^2, Z_Q = \frac{U - E_0[U]}{\sqrt{\widehat{Var_0}[U]}} \overset{\mathcal{D}}{\sim} N(0,1).$$

Podľa výberu váh w_i rozoznávame nasledovné typy testov:

- ak $w_i = n_i$, $Q = Q_{GW}$, ide o **Gehan-Wilcoxon test** (**zovšeobecnený Wilcoxonov test**; Gehan, 1965), ktorý môžeme zredukovať na Wilcoxonovu štatistiku pri absencii cenzúr [TW trieda (2)]
- ak $w_i = 1$, $Q = Q_{CM}$, ide o **Cox-Mantel test** (**log-rank test**; Mantel a Haenzel, 1959) [TW trieda (1) a HF trieda (1)]
- ak $w_i = \sqrt{n_i}$, $Q = Q_{TW}$, ide o **Tarone-Ware test** (Tarone a Ware, 1977) [TW trieda (3)]
- ak $w_i = \widehat{S}_{pooled}(t_i) = \prod_{j:t_j \le t_i} \frac{n_j - d_j + 1}{n_j + 1}$, $Q = Q_{PP}$, ide o **Peto-Peto test** (Peto a Peto, 1972; Prentice, 1978) [HF trieda (2)]


```
surv.test <- survdiff(Surv(cas,status)~x,rho=0)
```

Argumenty:

1. **typ testu**
   - rho=0 (Q_{MH})
   - rho=1 (Q_{PP})

Výstupy objektu `surv.test` sú nasledovné:

1. n – **počet pozorovaní** n v každej skupine
2. obs – **počet udalostí** v každej skupine (d_1 a d_2)
3. exp – **počet očakávaných udalostí** v každej skupine $E_0[d_1]$ a $E_0[d_2]$
4. var – **rozptyl** alebo **kovariančná matica** $Var_0[U]$

Veličina U podelená jej rozptylom $Var_0[U]$ nám dáva podľa práce Peto (1976)

$$\ln \widehat{\theta}_P = \frac{U}{Var_0[U]}$$

ako maximálne vierohodný odhad $\log \theta$. Táto štatistika má rozptyl rovný

$$Var_0\left[\ln \widehat{\theta}_P\right] = (Var_0[U])^{-1},$$

preto obojstranný $100(1-\alpha)\,\%$ interval spoľahlivosti pre $\log \theta$ (na základe asymptotickej normality) bude rovný

$$\left\{\ln \theta_P : \ln \widehat{\theta}_P \pm z_{\alpha/2}\sqrt{Var_0[U]}\right\},$$

potom

$$\left\{\theta_P : \widehat{\theta}_P \exp\left(\pm z_{\alpha/2}\sqrt{Var_0[U]}\right)\right\}.$$

Podľa Mantel a Haenzela (1959), odhadneme θ nasledovným spôsobom

$$\widehat{\theta}_{MH} = \frac{\sum_{i=1}^{I} \frac{d_{1i}(n_{2i}-d_{2i})}{n_i}}{\sum_{i=1}^{I} \frac{d_{2i}(n_{1i}-d_{1i})}{n_i}} = \frac{\sum_{i=1}^{I} R_i}{\sum_{i=1}^{I} S_i} = \frac{R_+}{S_+},$$

kedy tento odhad môžeme písať ako vážený priemer odhadu pomeru šancí zlyhania ($\widehat{OR}_i$) pre každú kontingenčnú tabuľku, teda

$$\widehat{\theta}_{MH} = \frac{\sum_{i=1}^{I} w_i \widehat{OR}_i}{\sum_{i=1}^{I} w_i},$$

kde

$$\widehat{OR}_i = \frac{d_{1i}(n_{2i} - d_{2i})}{d_{2i}(n_{1i} - d_{1i})},$$

$$w_i = \left(\frac{1}{n_{1i}} + \frac{1}{n_{2i}}\right)^{-1} (1 - \widehat{p}_{1i})\,\widehat{p}_{2i},$$

$\widehat{p}_{1i} = \frac{d_{1i}}{n_{1i}}, \widehat{p}_{2i} = \frac{d_{2i}}{n_{2i}}$ sú podmienené pravdepodobnosti zlyhania.

Prvá časť $\left(\frac{1}{n_{1i}}+\frac{1}{n_{2i}}\right)^{-1}$ ukazuje, že väčšie váhy zodpovedajú väčším rozsahom n_{1i} a/alebo n_{2i}. Sato (1990) odvodil $100\left(1-\alpha\right)\%$ interval spoľahlivosti pre θ ako riešenie kvadratickej rovnice

$$\frac{\left(R_+-\theta S_+\right)^2}{\theta W_+}=z_{\alpha/2}^2,$$

kde

$$W_+=\sum_{i=1}^{I}w_i=\sum_{i=1}^{I}\left[\begin{array}{c} d_{1i}\left(n_{2i}-d_{2i}\right)\left(n_{1i}-d_{1i}+d_{2i}+1\right)+\\ d_{2i}\left(n_{1i}-d_{1i}\right)\left(n_{2i}-d_{2i}+d_{1i}+1\right)\end{array}\right]/n_i^2.$$

Riešením vyššie uvedenej rovnice dostaneme

$$\frac{2R_+S_++z_{\alpha/2}^2W_+\pm\sqrt{\left(4R_+S_++z_{\alpha/2}^2W_+\right)z_{\alpha/2}^2W_+}}{2S_+^2}.$$

Ak nemáme zhody v čase t_i, potom $d_i=1$ a $d_{1i}-E_0\left[d_{1i}\right]=1-\frac{n_{1i}}{n_i}$, ak je zlyhanie pozorované v prvej skupine, alebo $-\frac{n_{1i}}{n_i}$, ak je zlyhanie pozorované v druhej skupine a korešpondujú príspevkom do R_+, resp. S_+. Ak rozptyl $Var_0\left[d_{1i}\right]=\frac{n_{1i}n_{2i}}{n_i^2}$, ľahko sa dá vidieť, že $Var_0\left[d_{1i}\right]$ korešponduje s w_i a výpočet IS pre θ má členy, ktoré sa vyskytujú aj v Q_{MH}. Simulačné Monte-Carlo štúdie ukázali, že pravdepodobnosť pokrytia tohoto približného IS je veľmi podobná očakávanej pravdepodobnosti pokrytia.

Alternatívou ku $\widehat{\theta}_{MH}$ je nasledovný odhad (Anderson a Bernstein, 1985)

$$\widehat{\theta}_{MH}^*=\frac{\sum_{i=1}^{I}\frac{d_{1i}n_{2i}}{n_i}}{\sum_{i=1}^{I}\frac{d_{2i}n_{1i}}{n_i}},$$

kde ide o vážený priemer odhadovaného pomeru zlyhaní v dvoch skupinách s váhami

$$w_i=\frac{n_{1i}n_{2i}}{n_i}.$$

Ak máme konštatnú proporcionalitu rizika cez všetky časy (θ sa nemení časom), $\widehat{\theta}_{MH}$ a aj $\widehat{\theta}_{MH}^*$ odhadujú túto konštantu. Ak máme nekonštantnú proporcionalitu rizika θ_i, $\widehat{\theta}_{MH}$ a aj $\widehat{\theta}_{MH}^*$ dávajú vážený priemer θ_i.
Ak máme konštatnú proporcionalitu rizika cez všetky časy (θ sa nemení časom), $\widehat{\theta}_{MH}$ a aj $\widehat{\theta}_{MH}^*$ odhadujú túto konštantu. Ak máme nekonštantnú proporcionalitu rizika θ_i, $\widehat{\theta}_{MH}$ a aj $\widehat{\theta}_{MH}^*$ dávajú vážený priemer θ_i.

Example (dvojvýberové testy)

Majme dáta z klinickej štúdie zhrnuté v nasledovnej tabuľke (pozri tabuľku). (a) Vytvorte kontingenčné tabuľky v každom čase zlyhania $t_i, i=1,2,\ldots,7$ použitím celkového počtu subjektov v riziku n_i v čase t_i, celkového počtu zlyhaní d_i v čase t_i, celkového počtu subjektov prvej skupiny v riziku n_{1i} v čase t_i a celkového počtu zlyhaní d_{1i} subjektov prvej skupiny v čase t_i. (b) Vypočítajte stredné hodnoty $E_0[d_{1i}]$, rozdiely empirických a očakávaných početností $d_{1i}-E_0[d_{1i}]$, ako aj rozptyly $Var_0[d_{1i}]$. (c) Otestujte $H_0:\lambda_1\left(t\right)=\lambda_2\left(t\right)$ oproti $H_1:\lambda_1\left(t\right)=\theta\lambda_2\left(t\right)$ pomocou testovacích štatistík Q_{GW}, Q_{CM}, Q_{TW} a Q_{PP}. (d) Nakreslite Kaplan-Meierove odhady funkcie prežívania pre obe skupiny do jedného obrázka. (e) Vypočítajte (1) $\widehat{\theta}_P$, $Var[\widehat{\theta}_{MH}]$ a 95%IS pre θ_P, (2) $\widehat{\theta}_{MH}$ a 95%IS pre θ_{MH} a (3) $\widehat{\theta}_{MH}^*$.

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

$Q = 2.31^2/1.0302 = 5.179674$
p-hodnota=0.02285261
$z_Q = \sqrt{2.31^2/1.0302} = 2.275890$
p-hodnota=$2 \times 0.01142630 = 0.02285259$

Example (dvojvýberová situácia)

(1) Nakreslite (a) kumulatívne riziko $\widehat{\Lambda}_{KM,j}(t)$, (b) kumulatívne riziko $\widehat{\Lambda}_{NA,j}(t)$, (c) Kaplan-Meierove krivky prežívania $\widehat{S}_{KM,j}(t)$ a (d) Breslowove krivky prežívania $\widehat{S}_{B,j}(t), j = 1, 2$ (vždy po dvojiciach do jedného obrázka).
(2) Otestujte $H_0 : \lambda_1(t) = \lambda_2(t)$ oproti $H_1 : \lambda_1(t) = \theta\lambda_2(t)$ pomocou testovacích štatistík Q_{MH} a Q_{PP}. Použite funkcie `survdiff()` s argumentami `rho=0` (Q_{MH}) a `rho=1` (Q_{PP}).
(3) Otestujte $H_0 : \lambda_1(t) = \lambda_2(t)$ oproti $H_1 : \lambda_1(t) = \theta\lambda_2(t)$ pomocou testovacích štatistík Q_{GW}, Q_{CM}, Q_{TW} a Q_{PP}.
(4) Vypočítajte (a) $\widehat{\theta}_P$, $Var[\widehat{\theta}_P]$ a Waldov 95% IS pre θ_P, (b) $\widehat{\theta}_{MH}$ a Waldov 95% IS pre θ_{MH} a (c) $\widehat{\theta}^*_{MH}$.

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 v t_i
zlyhanie v t_i	d_{1i}	d_{2i}	...	d_{ji}	...	d_{ki}	d_i
naživе 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** $k-1$ so **strednou hodnotou** $E_0[d_{ji}] = n_{ji}\frac{d_i}{n_i}$ v čase t_i
Potom

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

kde $\mathbf{U}_i$ je vektor merajúci rozdiel medzi pozorovaným a očakávaným počtom zlyhaní v čase t_i a je definovaný ako

$$\mathbf{U}_i = \begin{pmatrix} U_{1i} \\ \vdots \\ U_{ji} \\ \vdots \\ U_{k-1,i} \end{pmatrix} = \begin{pmatrix} d_{1i} - E_0[d_{1i}] \\ \vdots \\ d_{ji} - E_0[d_{ji}] \\ \vdots \\ d_{k-1,i} - E_0[d_{k-1,i}] \end{pmatrix}.$$

Kovariančná matica V (t_i) s komponentami v časoch t_i je daná nasledovne

$$(\mathbf{V}(t_i))_{ls} = \widehat{Cov\,[d_{li}, d_{si}]} = \begin{cases} \frac{n_{li}(n_i - n_{li}) d_i a_i}{n_i^2 (n_i - 1)} & \text{pre } l = s \\ -\frac{n_{li} n_{si} d_i a_i}{n_i^2 (n_i - 1)} & \text{pre } l \neq s \end{cases},$$

kde $l, s = 1, 2, \ldots, k-1$. Potom, ak berieme do úvahy všetky časy zlyhania, dostaneme

$$\mathbf{V} = \sum_{i=1}^{I} \mathbf{V}(t_i).$$

Testovacia štatistika

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

Ak $k = 2$, štatistika $Q_{overall} = Q_{CM}$.

V prípade pridania váh w_i v čase t_i bude platiť

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

kde $\mathbf{U}_i^{(w)}$ je vektor merajúci vážený rozdiel medzi pozorovaným a očakávaným počtom zlyhaní v čase t_i a je definovaný ako

$$\mathbf{U}_i^{(w)} = \begin{pmatrix} U_{1i}^{(w)} \\ \vdots \\ U_{ji}^{(w)} \\ \vdots \\ U_{k-1,i}^{(w)} \end{pmatrix} = w_i \begin{pmatrix} U_{1i} \\ \vdots \\ U_{ji} \\ \vdots \\ U_{k-1,i} \end{pmatrix} = w_i \begin{pmatrix} d_{1i} - E_0[d_{1i}] \\ \vdots \\ d_{ji} - E_0[d_{ji}] \\ \vdots \\ d_{k-1,i} - E_0[d_{k-1,i}] \end{pmatrix}.$$

Pre **kovariančnú maticu** bude platiť

$$\mathbf{V}_w = \sum_{i=1}^{I} w_i \mathbf{V}(t_i).$$

Nakoniec bude **testovacia štatistika** rovná

$$\mathbf{U}_w^T \mathbf{V}_w^{-1} \mathbf{U}_w \overset{\mathcal{D}}{\sim} \chi^2_{k-1}.$$

Voľba váh je nasledovná:

- ak $w_i = n_i$, potom $Q = Q_{GB}$ a ide o **zovšeobecnený Wilcoxonov test (zovšeobecnený Kruskal-Wallis test, Gehan-Breslow test)** [TW trieda (2)]
- ak $w_i = 1$, potom $Q = Q_{CM}$ a ide o **Cox-Mantel test (log-rank test)** [TW trieda (1)]
- ak $w_i = \sqrt{n_i}$, potom $Q = Q_{TW}$ a ide o **Tarone-Ware test)** [TW trieda (3)]
- ak $w_i = \hat{S}(t_i^-)^\rho$, $\rho = 0$, potom $Q = Q_{MH}$ a ide o **Mantel-Haenszelov test (log-rank test)** [HF trieda (1)]
- ak $w_i = \hat{S}(t_i^-)^\rho$, $\rho = 1$, potom $Q = Q_{PP}$ a ide o **Peto-Peto-Wilcoxon test** [HF trieda (2)]

Test nulovej hypotézy oproti stochasticky usporiadanej alternatíve je **testom trendu**, kde testujeme zoradený vzťah medzi k funkciami prežívania definovanými v zmysle vektora váh $\boldsymbol{\theta} = \left(\theta_1, \theta_2, \ldots, \theta_j, \ldots, \theta_k\right)^T$, potom

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

a H_1 :

$$\begin{aligned} \lambda_1(t) &= \theta_1 \lambda_k(t), \\ \lambda_2(t) &= \theta_2 \lambda_k(t), \\ &\ldots \\ \lambda_j(t) &= \theta_j \lambda_k(t), \\ \lambda_{k-1}(t) &= \theta_{k-1} \lambda_k(t), \end{aligned}$$

kde bez straty na všeobecnosti môžeme predpokladať, že $\theta_k = 1$.

Testovacia štatistika pre trend bude daná nasledovným vzťahom

$$Q_{trend} = \frac{\left(\boldsymbol{\theta}^T \mathbf{U}_*\right)^2}{\boldsymbol{\theta}^T \mathbf{V}_* \boldsymbol{\theta}}$$

a po zložkách

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

kde $\mathbf{U}_*$ je $\mathbf{U}$ doplnené o k-ty element. Ďalej $\mathbf{V}_*$ počítame tak ako $\mathbf{V}$, ale s tým rozdielom, že ide o maticu $k \times k$. Ak sú váhy lineárne, napr. $\theta_j = j$, potom hovoríme o **teste lineárneho trendu**.
Platí

$$Q_{residual} = Q_{overall} - Q_{trend} \overset{\mathcal{D}}{\sim} \chi^2_{k-2}.$$

Example (trojvýberová situácia)

Majme experiment, 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 tabuľku). (1) Nakreslite (a) kumulatívne riziko $\widehat{\Lambda}_{KM,j}(t)$, (b) kumulatívne riziko $\widehat{\Lambda}_{NA,j}(t)$, (c) Kaplan-Meierove krivky prežívania $\widehat{S}_{KM,j}(t)$ a (d) Breslowove krivky prežívania $\widehat{S}_{B,j}(t), j = 1, 2, 3$ (vždy po trojiciach do jedného obrázka). (2) Otestujte (a) $H_0 : \lambda_1(t) = \lambda_2(t) = \lambda_3(t)$ oproti $H_1 : \exists$ aspoň jedno $i < j, \lambda_i(t) \neq \lambda_j(t)$ pomocou testovacích štatistík Q_{GW}, Q_{CM}, Q_{TW} a Q_{PP}; (b) $H_0 : \lambda_1(t) = \lambda_2(t) = \lambda_3(t)$ oproti $\lambda_1(t) = \theta_1 \lambda_3(t)$, $\lambda_2(t) = \theta_2 \lambda_3(t)$, kde $\theta_j = konc_j, j = 1, 2$ (test trendu) pomocou testovacej štatistiky Q_{trend}.

<u>Pozn.:</u> 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^+	