Analýza přežití

Motivace: Analýza přežití je obor statistiky zabývající se popisem a analýzou dat, která korespondují době od vstupní události (tzv. čas počátku) do výskytu sledované události (tzv. koncový bod). Za vstupní událost můžeme pokládat například
- narození,
- počátek léčby,
- začátek nemoci,
- vstup jedince do studie,
- svatbu,
- zavedení nového přístroje do výroby

a jiné.
Koncovou událostí může být
- úmrtí jedince,
- návrat příznaků nemoci,
- uzdravení pacienta,
- rozvod,
- porucha přístroje

a další.
Dobu mezi těmito dvěma událostmi označujeme jako dobu přežití.
Analýza přežití, jak vyplývá z předchozího, má velmi široké uplatnění, třeba
- ve zdravotnictví,
- v průmyslu,
- v zemědělství,
- v demografii

apod.
Pro jednoduchost budeme za čas počátku považovat vstup jedince do nějaké studie či experimentu a za koncový bod smrt jedince.

Specifika dat v analýze přežití

Data analýza přežití nejsou vhodná ke zpracování standardními statistickými metodami používanými v analýze dat. Hlavním důvodem je fakt, že doby přežití jsou často cenzorovány. Doba přežití jedince je cenzorována, jestliže sledovaná koncová událost není u tohoto jedince během pozorování uskutečněna. To nastane například v případě, že

- s pozorovaným jedincem ztratíme kontakt, přestěhuje se nebo přestane docházet na pravidelné prohlídky nutné ke studii (nevíme, zda je na konci studie živ či mrtev)
- data jsou zpracovávána v době, kdy se sledovaná událost u jedince ještě nevyskytla
- pozorovaný jedinec zemřel na jinou nemoc

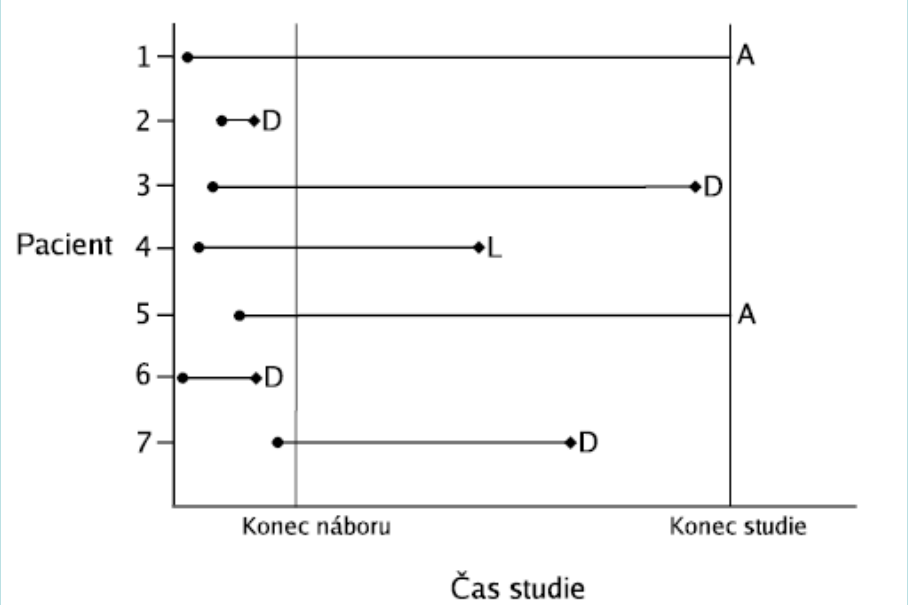
a podobně.

V každé z těchto situací jedinec, který vstoupil do studie v čase t_0, zemřel v čase $t_0 + t$, avšak čas t je neznámý. Víme pouze, že jedinec byl živ v čase $t_0 + c$, kde čas c se nazývá cenzorovaná doba přežití. V tomto případě, kdy se cenzorované události staly napravo od posledního známého času přežití, mluvíme o cenzorování zprava - skutečná doba přežití je vyšší než doba pozorování.

V případě, že doba přežití jedince je menší než sledovaná, jedná se o další typ cenzorování, a to cenzorování zleva. Tímto druhem cenzorování je případ, kdy jedinec zemře dříve než oficiálně započne studie, například během výběru jedinců vhodných ke sledování.

Posledním typem cenzorování je intervalové cenzorování, které odpovídá případu, že jedince je možno sledovat jen v určitých okamžicích (například jednou za měsíc), ne tedy bez přerušení po celou dobu trvání studie. V takovém případě vždy dostaneme jen informaci, že smrt nastala v časovém rozmezí určeném okamžikem posledního pozorování a současností.

Ilustrace různých druhů cenzorování



Na obrázku je znázorněna doba přežití u 7 pacientů. Období sledování je započato koncem náboru jedinců a ukončeno koncem studie. Písmeno D označuje smrt, L ztrátu kontaktu s jedincem a A znamená, že pacient je stále naživu. Na začátku studie zjistíme, že pacienti 2 a 6 již zemřeli, avšak nevíme přesně čas úmrtí. Jediné, co víme, je, že tito jedinci zemřeli někdy před začátkem studie, a tudíž se jedná o cenzorování zleva. V případě pacientů 1, 4 a 5 se jedná naopak o cenzorování zprava. Jelikož s pacientem 4 jsme během studie ztratili kontakt, neznáme jeho stav na konci studie. Pacienti 1 a 5 jsou na konci studie stále naživu, a tedy sledovaná událost, v našem případě smrt, se u nich během pozorování nevyskytla. Jedná se tedy také o cenzorovaný čas přežití zprava. Nakonec u jedinců 3 a 7 se sledovaná událost vyskytla během doby pozorování, a tudíž se jedná o necenzorované časy přežití, jelikož přesně víme dobu úmrtí.

Upozornění: Nadále budeme předpokládat, že data jsou cenzorovaná zprava, jelikož v praxi se jedná o nejčastěji se vyskytující typ cenzorování.

Funkce přežití

Nechť spojitá nezáporná náhodná veličina T udává čas, který uplyne od počátku sledování jedince do jeho smrti. Rozložení pravděpodobností této náhodné veličiny je popsáno hustotou pravděpodobnosti $\varphi(t)$.
Distribuční funkce $\Phi(t) = P(T \le t)$ je s hustotou spjata vztahem

$$\forall t \ge 0 : \Phi(t) = \int_0^t \varphi(u)du \, .$$

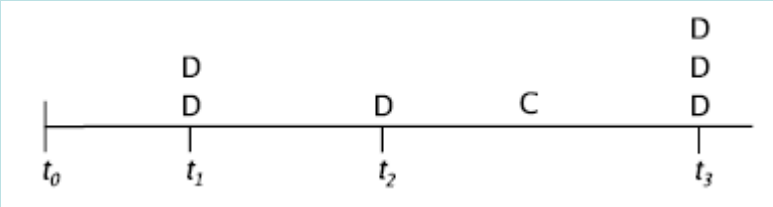
Zavedeme funkci přežití $\Psi(t) = P(T > t)$.
Hodnota funkce přežití v bodě t je pravděpodobnost, že doba přežití sledovaného jedince je větší než t. Funkci přežití lze pomocí hustoty vyjádřit vztahem

$$\forall t \ge 0 : \Psi(t) = \int_t^\infty \varphi(u)du$$

a pomocí distribuční funkce vztahem
$\forall t \ge 0 : \Psi(t) = 1 - \Phi(t)$.

Kaplanův - Meierův odhad funkce přežití

Kaplanův - Meierův odhad funkce přežití je metoda, která poskytuje odhad funkce přežití v každém okamžiku, ve kterém došlo k alespoň jedné sledované události. Opět budeme pro jednoduchost za tuto událost považovat smrt.
K určení Kaplanova - Meierova odhadu funkce přežití z cenzorovaných dat se nejprve rozdělí doba pozorování do souboru časových intervalů. Každý z těchto intervalů je zkonstruován tak, aby v každém z nich bylo obsaženo alespoň jedno úmrtí, přičemž čas smrti je vzat jako počátek jednotlivých intervalů. Například předpokládejme, že t_1, t_2, t_3 jsou tři zaznamenané časy přežití uspořádané dle velikosti tak, že $t_1 < t_2 < t_3$, a c je cenzorovaný čas přežití, který spadá mezi časy t_2, t_3. Zkonstruované intervaly tedy začínají v časech t_1, t_2, t_3. Každý z intervalů obsahuje jeden čas úmrtí, ačkoliv zde může být více než jeden jedinec, který zemřel v některém z jednotlivých časů t_1, t_2, t_3. Stojí za povšimnutí, že žádný interval nezačíná v cenzorovaném čase c. Situace je ilustrována na následujícím obrázku, kde D reprezentuje smrt a C cenzorovaný čas přežití. Vidíme, že dva jedinci umřeli v čase t_1, jeden v čase t_2 a tři zemřeli v čase t_3.



Čas počátku, například studie, je označen jako t_0. Zde je také počátek prvního období, které končí před t_1, získáme tedy interval $\langle t_0, t_1)$. Tento interval neobsahuje žádné úmrtí. První zkonstruovaný interval $\langle t_1, t_2)$ obsahuje první čas úmrtí v čase t_1. Druhý interval $\langle t_2, t_3)$ obsahuje čas smrti v čase t_2 a cenzorovaný čas přežití c. Poslední - třetí interval - započne v čase t_3 a obsahuje nejvyšší čas přežití, čas t_3.

Označení používaná v K – M odhadu funkce přežití

n … počet sledovaných jedinců

$t_1, t_2, \ldots, t_n$ … časy přežití sledovaných jedinců

(Některá z těchto pozorování mohou být zprava cenzorovaná, takže se zde může vyskytnout několik jedinců se stejnou dobou přežití.)

r … počet časů úmrtí mezi n sledovanými jedinci ($r \leq n$)

$t_{(1)} < t_{(2)} < \ldots < t_{(r)}$ … r uspořádaných časů úmrtí

n_j … počet jedinců, kteří jsou živí před časem t_j, j = 1, 2, … , r

d_j … počet jedinců, kteří zemřou v čase t_j, j = 1, 2, … , r

$\frac{n_j - d_j}{n_j}$ … odhad pravděpodobnosti přežití pro interval $\langle t_j, t_{j+1})$

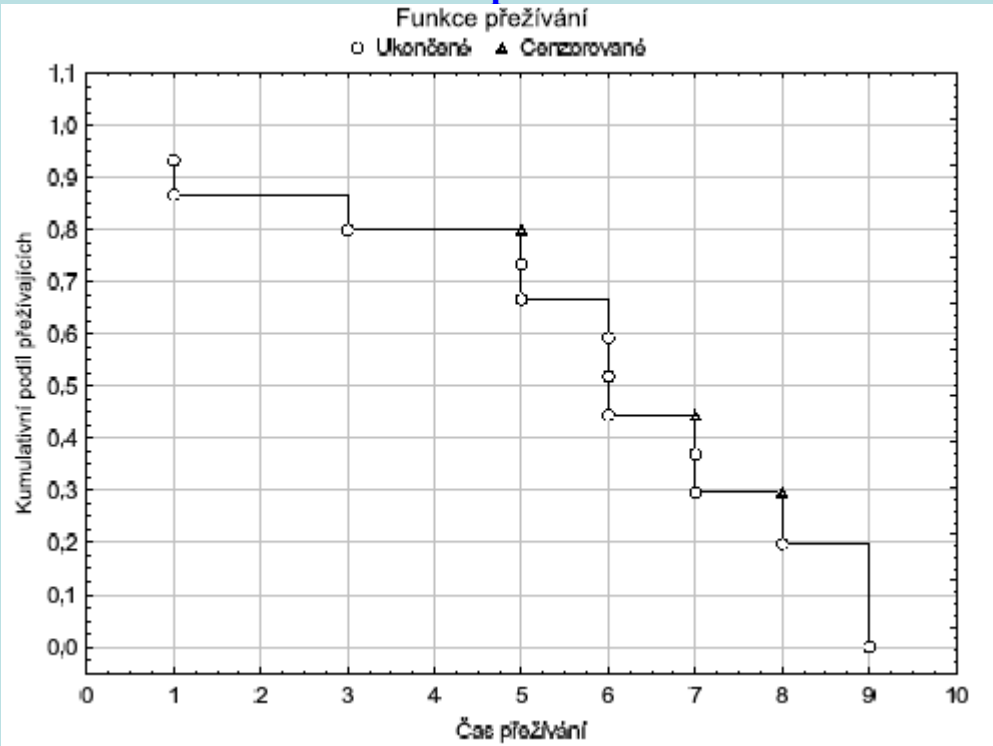
Předpokládáme, že smrti jedinců nastávají ve stejném okamžiku nezávisle na sobě.

Kaplanův -Meierův odhad funkce přežití je dán vztahem

$$\hat{\Psi}(t) = \prod_{j=1}^{k} \left(\frac{n_j - d_j}{n_j} \right) = \prod_{j=1}^{k} \left(1 - \frac{d_j}{n_j} \right), \text{ kde } t_{(k)} \leq t < t_{(k+1)}, \; k = 1, 2, \ldots, r$$

Graf funkce přežití odhadnuté K - M odhadem má schodovitý průběh s tím, že odhadnuté pravděpodobnosti přežití jsou konstantní mezi každými dvěma sousedními časy smrtí a v jednotlivých časech úmrtí funkce klesá.

Ukázka K – M odhadu funkce přežití



Máme 15 jedinců a 12 časů úmrtí. 3 jedinci mají cenzorované časy přežití. V čase 1 zemřeli 2 jedinci, v čase 3 jeden jedinec, v čase 5 dva jedinci a jeden je cenzorován atd.

Testování hypotézy o rozdílu mezi dvěma a více skupinami

V analýze přežití se zajímáme o to, zda existují rozdíly mezi různými skupinami, např.

- mezi pacienty s různými druhy léčby,
- mezi muži a ženami trpícími stejnou chorobou,
- mezi pacienty s různými stádii téže choroby.

Nejjednodušší způsob je vykreslení odhadů funkce přežití do jednoho grafu. Výsledný graf již může být dostatečně informativní.

Existují ovšem i statistické testy, které mohou být použity ke zjištění rozdílnosti mezi skupinami.

Nulová hypotéza: funkce přežití jsou ve všech sledovaných skupinách stejné.

Alternativní hypotéza: existuje aspoň jedna dvojice rozdílných funkcí přežití.

Případ dvou skupin

Pro porovnání dvou skupin jedinců se často používá **log-rank test** a **Gehanův – Wilcoxonův test**.

Oba testy jsou založeny na porovnání pozorovaného a očekávaného počtu úmrtí v 1. skupině.

Testová statistika obou testů se v případě platnosti nulové hypotézy asymptoticky řídí rozložením $\chi^2(1)$.

Nulovou hypotézu tedy zamítáme na asymptotické hladině významnosti α, když testová statistika $\geq \chi^2_{1-\alpha}(1)$.

Srovnání Wilcoxonova a log-rank testu

Při testování nulové hypotézy, že neexistuje rozdíl mezi funkcemi přežití dvou skupin jedinců, je dobré vědět, který z uvedených dvou testů je vhodnější použít.

Jednoduchou pomůckou, jak to zjistit, je vykreslit si obě funkce přežití do jednoho grafu a podle jejich průběhu vybrat vhodnější test. Pokud se tyto dvě funkce přežití kříží, je vhodnější k testování nulové hypotézy zvolit Gehanův - Wilcoxonův test. Pokud je tomu však naopak a funkce se nekříží, je lépe zvolit log-rank test.

Případ tří a více skupin

Předpokládáme, že máme $p \geq 3$ skupin jedinců

Nulová hypotéza: funkce přežití jsou ve všech p sledovaných skupinách stejné.

Alternativní hypotéza: existuje aspoň jedna dvojice rozdílných funkcí přežití.

χ^2-kvadrát test je opět založen na porovnání pozorovaného a očekávaného počtu úmrtí v p-1 skupinách.

V případě platnosti nulové hypotézy se testová statistika asymptoticky řídí rozložením $\chi^2(p-1)$. Nulovou hypotézu tedy zamítáme na asymptotické hladině významnosti α, když testová statistika $\geq \chi^2_{1-\alpha}(p-1)$.

V případě zamítnutí nulové hypotézy chceme odhalit, které dvojice skupin se liší. Při porovnávání dvojic skupin však musíme použít Bonferroniho korekci. Původní hladinu významnosti podělíme počtem prováděných testů (tj. $\binom{p}{2}$). Rozdíl mezi dvěma skupinami pak považujeme za významný, když p-hodnota dvouvýběrového testu poklesne pod $\frac{\alpha}{\binom{p}{2}}$.

Riziková funkce a kumulativní riziková funkce

Riziková funkce h(t) je definována jako intenzita pravděpodobnosti úmrtí, tj. pravděpodobnost, že jedinec nepřežije krátký časový interval Δt za předpokladu, že přežil čas t:

$$h(t)\Delta t = P(t < T \le t + \Delta t / T > t)$$

$$h(t) = \frac{P(t < T \le t + \Delta t)}{\Delta t P(T > t)} = \frac{\Phi(t + \Delta t) - \Phi(t)}{\Delta t \Psi(t)} = \frac{1}{\Psi(t)} \frac{\Phi(t + \Delta t) - \Phi(t)}{\Delta t}$$

Limitním přechodem $\Delta t \to 0$ dostaneme

$$h(t) = \frac{1}{\Psi(t)} \frac{d}{dt} \Phi(t) = \frac{\frac{d}{dt}(1 - \Psi(t))}{\Psi(t)} = -\frac{\frac{d}{dt}\Psi(t)}{\Psi(t)} = -\frac{d}{dt} \ln \Psi(t)$$ nebo $$h(t) = \frac{\varphi(t)}{\Psi(t)}$$

Riziková funkce má často „vanovitý“ průběh: na počátku klesá, pak je takřka konstantní a ke konci opět roste – projeví se stárnutí.

Kumulativní riziková funkce:

$$H(t) = \int_0^t h(u) du$$

Vztah mezi kumulativní rizikovou funkcí a funkcí přežití:

$$H(t) = -\ln \Psi(t)$$

Jádrový odhad rizikové funkce

Na stránce http://www.math.muni.cz/veda-a-vyzkum/vyvijeny-software/274-matlab-toolbox.html je umístěn balík kerns i nápověda k němu. Pomocí funkce kshazard.m lze pro různá jádra a různé šířky vyhlazovacího okna získat odhady rizikové funkce.

Ukázka odhadu rizikové funkce s Epanečnikovým jádrem: