MODERNÍ ANALÝZA BIOLOGICKÝCH DAT

ANALÝZA ČASU DO UDÁLOSTI A DALŠÍ METODY V PROSTŘEDÍ R

4

STANO PEKÁR
MAREK BRABEC

MASARYKOVA UNIVERZITA

MODERNÍ ANALÝZA BIOLOGICKÝCH DAT
ANALÝZA ČASU DO UDÁLOSTI A DALŠÍ METODY V PROSTŘEDÍ R

4. díl

STANO PEKÁR, MAREK BRABEC

MUNI
PRESS

Knihu recenzoval: doc. RNDr. PaedDr. Stanislav Katina, Ph.D.

MODERNÍ ANALÝZA BIOLOGICKÝCH DAT

ANALÝZA ČASU DO UDÁLOSTI A DALŠÍ METODY V PROSTŘEDÍ R

4. díl

STANO PEKÁR
MAREK BRABEC

Masarykova univerzita, Brno 2024

https://www.press.muni.cz/moderni-analyza-4

Pekár S. & Brabec M. 2024. Modern analysis of biological data. 4.
Time to event analysis and other methods in R. Masaryk University Press, Brno.

ISBN 978-80-280-0353-1 (vázáno)
ISBN 978-80-280-0366-1 (brožováno)
ISBN 978-80-280-0367-8 (online ; pdf)

OBSAH

PŘEDMLUVA

V prvních třech dílech MABD jsme představili především jednorozměrné regresní modely, čímž jsme pominuli celou velmi zajímavou a užitečnou skupinu statistických modelů. To chceme napravit v tomto díle, který bude na rozdíl od předchozích značně heterogenní. V tomto dílu se totiž zaměříme na několik tříd modelů. Nejen na vícerozměrné modely a metody, ale také na analýzu doby do události. Ta je velice běžná v mnoha různých biologických oborech, kde se pracuje s časem jako významnou proměnnou. Čas je důležitou proměnnou v medicíně, protože nás zajímá, jak dlouho budeme žít, jak dlouho budeme nemocní apod. Ale také v ekologii, především v demografii, měříme čas v různých kontextech. Zajímá nás doba, která uběhne před tím, než se objeví biologicky relevantní událost, jako například vylíhnutí vajíček, vykvetení rostliny nebo ulovení kořisti.

Struktura tohoto dílu je velice podobná předchozím dílům. I tento díl je postaven na řadě příkladů z reálných studií. Rozpracování je podobné jako v předchozích dílech. Data byla samozřejmě částečně pozměněna tak, abychom mohli ukázat různé situace. Typicky to znamenalo redukci datasetu jak co do počtu měření, tak do počtu vysvětlujících proměnných. Stejně tak text knihy je opět „dvojjazyčný". Dominuje samozřejmě čeština, ale používáme mnohé anglické termíny. Kromě užití anglických slov v prostředí R používáme anglické termíny i v teoretické části. Je to proto, že české statistické názvosloví není vždy zcela ustálené (zejména v oblasti novějších metod, modelů či technik) a bere si často za základ amatérsky „počeštěné" anglické termíny.

Na závěr bychom chtěli poděkovat kolegům, kteří laskavě souhlasili s použitím svých (byť upravených) dat k příkladům uvedeným v tomto dílu. Jsou to jmenovitě: P. Fila, A. Honěk, M. Horsák, J. Hubert, J. Křemenová, Z. Martinková, E. Řehulková a P. Samaš. Některým z nich, konkrétně P. Filovi, T. Bartoničkovi, M. Horsákovi, E. Řehulkové a P. Samašovi jsme vděčni také za připomínky k vybraným kapitolám. A také jsme vděčni doc. RNDr. PaedDr. Stanislavu Katinovi, Ph.D., za celou řadu připomínek, které text jednoznačně vylepšily. Jakékoliv připomínky k textu a obsahu knihy rádi uvítáme na e-mailových adresách: *pekar@sci.muni.cz* anebo *mbrabec@cs.cas.cz*.

Leden 2024

Stano Pekár
Marek Brabec

1 ÚVOD

V tomto v pořadí již čtvrtém díle představíme statistické metody, které sice nejsou žádnou novinkou, ale jejich použití nebylo v předchozích našich textech doposud dostatečně představeno. Jmenovitě se budeme věnovat třem třídám modelů: analýze času do události, regresním stromům a vybraným vícerozměrným metodám, které se používají v jiném kontextu, než je analýza společenstev druhů organismů.

Na první pohled se může obsah knihy zdát dost heterogenní. Ano, je určitě různorodější než všechny předešlé díly, které byly vždy zacílené na konkrétní metodu či třídu modelů. Metody, které zde uvedeme, jsou relativně specializovanější oproti třeba GLM, a proto jsme jich několik vtěsnali do jednoho dílu. Přestože třída vícerozměrných metod a modelů je velice široká a dalo by se o nich naplnit několik dílů, úmyslně jsme se nechtěli podrobně zabývat ordinačními metodami, pro které existuje dostupná a kvalitní specializovaná česká literatura (třeba Lepš & Šmilauer 2000).

Analýza doby do události, známá pod názvem analýza přežití, je v rámci biologických disciplín asi nejpoužívanější v lékařství. O výsledcích těchto metod se dočteme v podstatě denně. Třeba to, že ženy se dožívají v průměru o několik let déle než muži. Anebo v průběhu epidemie COVIDu jsme opakovaně slyšeli, za jak dlouho se můžeme nakazit. Ale také při návštěvě lékaře a při stanovení diagnózy se dozvíme, za jak dlouho se (typicky) můžeme vyléčit. Tyto metody mají však mnohem širší použití, protože sledovanou událostí může být cokoliv, což je zejména v biologických disciplínách opravdu pestré – narození, vyklíčení, páření, interakce molekul atd. Ne každý biolog si tuto metodologickou výhodu plně uvědomuje. A nejen takovým je tento text určen.

Klasifikační a regresní stromy jsou poněkud opomíjenou metodou, se kterou jste se doposud nemuseli setkat. Je to novější metoda používána spíše v explorační analýze. Ale lze ji úspěšně aplikovat také v koncové analýze, především v situacích, kdy klasifikujeme subjekty podle vybraných kritérií.

Doposud jsme se v regresních analýzách věnovali především datům s jednou závislou proměnnou. Ta vznikla třeba měřením nějaké vlastnosti jednoho druhu. Přitom typický biolog, a to nejen zoolog, botanik, mikrobiolog, ale také zemědělec nebo lékař, se na počátku studia potkává s celým společenstvem živočichů, rostlin, mikroorganismů nebo skupinou měřených vlastností (symptomů). Čili hned na začátku s velice komplexním – vícerozměrným problémem. Větší rozměr je dán nejen maticí různého počtu druhů jakožto závislé proměnné. Analýza takových dat je samozřejmě mnohem náročnější než analýza jednorozměrných dat, a proto se k ní dostáváme až teď. Vícerozměrné metody

jsou, zejména v českém prostředí, velice oblíbené především mezi ekology, kteří studují společenstva živočichů nebo rostlin.

1.1 Jak číst tuto knihu

Podobně jako předchozí díly, text knihy kombinuje nezbytnou (minimální) teorii s ukázkami vybraných statistických metod a popis jejich implementace v prostředí **R**. Každá metoda je nejprve představena stručně teoreticky a za ní následují příklady. Jelikož jsou v knize tři diametrálně odlišné skupiny metod, není nutné číst knihu od začátku, ale je možné se zaměřit na konkrétní skupinu.

Kapitoly 2–6 jsou zaměřeny na analýzu doby do události. Tři z těchto kapitol jsou „teoretické", které je nutné prostudovat a pochopit, aby byl čtenář schopen porozumět dalším kapitolám. Teorie v nich obsažená není (z pohledu typického biologa) tak úplně triviální a k dobrému pochopení jsou nutné znalosti všech třech předchozích dílů MABD. Dvě z těchto kapitol, 4 a 6, obsahují několik vzorových příkladů použití semiparametrického a parametrického modelu.

Kapitola 7 je o klasifikačních a regresních stromech. Do jedné kapitoly se tentokrát vešel základní popis principu metody i praktické ukázky. Samotná kapitola není nikterak dlouhá a věříme, že pro čtenáře bude vcelku snadno pochopitelná (intuitivní konstrukce ostatně stála na počátku celé metody). Použití této metody je velice obecné. Ve dvou kapitolách ukážeme provázanost s jinými metodami.

Kapitoly 8–11 obsahují vybrané vícerozměrné metody. Každá z metod je představena v separátní kapitole, která obsahuje teorii a pak alespoň jeden příklad.

V kap. 8 je nezbytná teorie zaměřená na vícerozměrný lineární model. Ten je nadstavbou obecného lineárního modelu, který jsme popsali již v MABD 1 (Pekár & Brabec 2020).

V kap. 9 se vracíme k PCA, kterou jsme v předchozích dílech použili v EDA, na úpravu matice vysvětlujících proměnných. Zde si tuto metodu představíme v poněkud jiném světle, jako nástroj koncové analýzy konkrétních dat.

Kap. 10 obsahuje diskriminační analýzu, která je velice užitečná pro řešení klasifikačních problémů.

Konečně v poslední kap. 11 ukazujeme, jak již známé metody LME, GLS či GEE, kterým jsme se podrobně věnovali v MABD 2 (Pekár & Brabec 2012), lze použít v docela jiných situacích, jež jsou velice běžné. Proto zde již neopakujeme teorii a čtenáře odkazujeme na předchozí díl.

Na konci textu uvádíme seznam použité a doporučené literatury. Obsahuje jak tuzemské, tak zahraniční tituly související s probíranými metodami. Vesměs jsou to publikace vztahující se buď přímo k R, anebo statistické texty s převahou těch, které se víceméně věnují aplikacím statistiky v biomedicínské oblasti.

Na úplný závěr jsou zařazeny rejstříky: a to jak obecných termínů funkcí, tak argumentů prostředí R. Index je záměrně dosti podrobný, aby umožnil snadné vyhledávání požadovaných termínů.

Datové soubory použité v této knize si můžete stáhnout z adresy:
https://www.press.muni.cz/moderni-analyza-4.
Všechna data jsou uložena v jednom excelovském souboru, přičemž každý datový soubor je na zvláštním listu označeném číslem kapitoly. Uživatel si je může importovat do R přes schránku (clipboard).

1.2 Konvence

Text knihy obsahuje několik typů písma. Využíváme je k odlišení základního textu knihy od příkazů (a jiných klíčových slov) jazyka R. Pokud uvádíme názvy příkazů a jejich argumentů, v textu používáme font Courier New tučný, o velikosti 10 bodů. Názvy objektů, které uživatelsky vytváříme v průběhu analýzy, jsou psány fontem Courier New obyčejný, o velikosti 10 bodů. Ostatní text je psán fontem Times New Roman, velikost 10. Názvy proměnných, matematické formulace jsou psány kurzivou, názvy úrovní faktorů ve strojopisných uvozovkách. Jména balíčků jsou podtržena.

K přepisu všeho, co se zobrazuje v oknech spuštěného prostředí R, používáme font Courier New, o velikosti 8 bodů. Pro lepší orientaci přitom rozlišujeme mezi uživatelem zadávanými příkazy, které píšeme tučně, a odpovědí programu v normálním stylu. Pro úsporu místa byly některé řádky odpovědí programu vynechány a nahrazeny tečkami.

Pracovní grafy, tj. ty, které jsou vytvořené na začátku analýzy, byly vytvořeny s použitím pokud možno co nejmenšího počtu příkazů a argumentů, proto jim často chybí popisky, legendy apod. Teprve finální grafy obsahují všechny detaily (za cenu delší syntaxe).

Přirozený logaritmus (se základem *e*) se ve statistice používá velmi často. Budeme ho tedy zapisovat jako log. Navzdory českému prostředí jako oddělovač desetinných míst používáme tečku místo čárky. Konečně, většina čísel je v textu zaokrouhlena na 4 cifry.

V analýze byla použita verze R 4.2.3 (R Core Team 2023).

2 ANALÝZA ČASU DO UDÁLOSTI

2.1 Úvod

V prvním dílu MABD (Pekár & Brabec 2020), konkrétně v kap. 7, jsme zmínili, že čas, přesněji čas do nějaké sledované události, lze analyzovat metodami analýzy přežití, a dále jsme se tomu tématu již nevěnovali. Teď nastal čas se na tyto metody podívat zblízka. Pod názvem „Analýza přežití" (z anglického Survival analysis, popřípadě Lifetime analysis) vystupují různé metody, parametrické, semiparametrické i neparametrické, které se používají k analýze času do nějaké sledované události (time to event, failure time, anebo time to failure). To jest doby, do které se vyskytne sledovaná událost. Tou nemusí být vždy exitus, jak by doslovné čtení sousloví „Analýza přežití" mohlo napovídat. Události, které sledujeme, mohou v biologii být velmi různého druhu, např. čas do klíčení, doba do napadení kořisti predátorem, čas do první reprodukce, doba do (prvního) poklesu cholesterolu od podání určitého léčiva, doba do rozpoznání partnera od počátku interakce, čas do vyléčení, a konečně i doba do úmrtí. Ty poslední jmenované události daly celé oblasti jméno. Nelze se tomu divit, protože tyto metody jsou velice často používány především v lékařství a demografii k tomu, aby zodpověděly otázky typu, jaké je riziko onemocnění, jak zkoumané léčivo mění rozdělení dob přežití, nebo jaká je pravděpodobnost uzdravení se z dané infekční choroby atd.

V závislosti na konkrétním kontextu může být název „Analýza přežití" poněkud zavádějící. Alternativní názvy, jako „Doba do selhání", jsou o něco lepší, ale také často nevyjadřují to, co chceme analyzovat, třeba dobu do klíčení (klíčení bereme typicky spíše jako úspěch než selhání). Proto budeme v této knize používat neutrální název: analýza času do události. Poznamenejme také, že podobné statistické metody (popřípadě rozšířené o některé fyzikální principy a úvahy ne tak běžně dostupné v biologických vědách) se vyskytují i v oblasti spolehlivosti průmyslových výrobků či složitých systémů, např. jako „Failure time analysis" nebo „Reliability".

Možná jste si vzpomněli, že v jednom z příkladů v prvním dílu MABD (Pekár & Brabec 2020) jsme také analyzovali čas do události. A přitom jsme použili GLM. Tak proč nyní tvrdíme, že jsou k tomu obecně vhodné jiné metody? Ano, čas do události lze správně analyzovat i pomocí GLM. Ale jenom za určitých (dosti omezujících) okolností, a to pokud jsme u všech sledovaných jedinců sledovanou událost zaznamenali. Nedošlo např. k tomu, že sledování skončilo dříve, než u některého/některých jedinců k události došlo a údaj o skutečném čase byl tedy neznámý (resp. bylo o něm známo jen to, že byl delší než doba sledování). Takovéto předpoklady byly splněny i v příkladu z prvního dílu MABD. V mnoha případech sledovaná událost u některých jedinců během doby

sledování prostě nenastane. Důvodů je několik, jak si ukážeme níže. Právě v takových situacích použijeme metody typické pro analýzu času do události. Leckoho možná napadne, že pokud je s „nenastalými" (tzv. cenzorovanými) událostmi problém, neměl by být problém je jednoduše vyřadit. Jak si povíme později, nebyl by to dobrý nápad, nejenže se tím sníží počet pozorování (to je ještě ten menší problém), ale vedlo by to ke zcela systematickému podhodnocení odhadů.

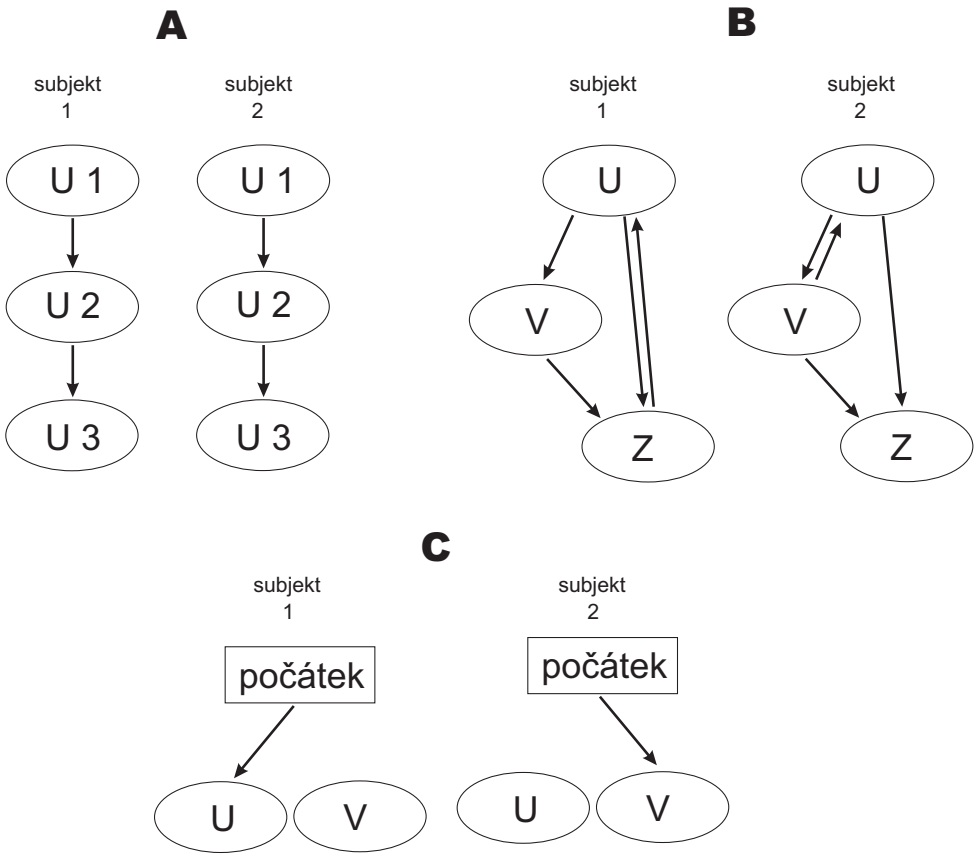
Podobně jako v jiných oblastech statistiky, existují parametrické, semiparametrické i neparametrické metody analýzy času do události. Pokud pracujeme s modely, které rozdělení doby do události jakožto závislé proměnné vysvětlují pomocí jedné nebo několika vysvětlujících proměnných, jde v podstatě o regresní problém (byť podstatně obecněji chápaný než např. v oblasti lineární regrese).

2.2 Co je událost

Jako událost (angl. event) je obecně brán jakýkoliv jev, o jehož sledování a modelování se zajímáme – několik jsme jich jmenovali výše a v příkladech tohoto dílu se seznámíte s dalšími. Někdy samotná událost trvá v řádu milisekund, jindy i několik hodin, přičemž nemusí být jednoduché stanovit, kdy událost vlastně nastala. Řekněme, že sledujeme mortalitu nějakého škodlivého hmyzu po aplikaci insekticidu. Hmyz neuhyne okamžitě, najednou. Úhyn je proces, který může trvat několik hodin nebo i dnů. Přímo po aplikaci toxické látky nejprve dochází k paralýze, která imobilizuje jedince a skutečný úhyn může nastat až za několik dnů. Přitom není snadné poznat přesnou dobu, kdy došlo k úhynu. Záleží na tom, jak úhyn definujeme. Úhyn bychom mohli definovat jako paralýzu, nebo jako zastavení dýchání či činnosti srdce. To první je snazší určit pozorováním pouhým okem. Pro ty další definice může být nezbytné použít přístroje. Ale i taková paralýza nenastane v jediný okamžik. Typicky, jedinec upadá do paralýzy několik minut. Čím déle událost trvá, tím větší bude rozptyl v měření času, kdy to nastalo, což je komplikace, které se chceme vyhnout. Proto je nutné si stanovit takový moment, který je dobře rozpoznatelný a charakterizuje sledovanou událost. I tak mnohdy dochází k nějakému zaokrouhlení a nepřesnosti.

Měření času do výskytu pouze jediné události je nejběžnější. Zejména pro události, které se z definice mohou vyskytnout u jednoho jedince či dané sledované jednotky pouze jednou. Jmenovitě, úmrtí nebo klíčení se může vyskytnout pouze jednou za život sledovaného jedince. Zato chřipka, páření, nebo kvetení se může vyskytnout opakovaně u stejného jedince v různých intervalech. V takových případech nás může zajímat výskyt a doba do několika typů událostí (doby mezi koncem minulé události a počátkem té současné). Tyto případy jsou sdruženy pod termínem **vícestavové analýzy** (angl. multi-state analysis), které zahrnují (obr. 2-1):

- opakovaná měření na stejném subjektu,
- výskyt různých událostí u stejného subjektu,
- výskyt několika druhů událostí u stejného subjektu.



Obr. 2-1. Příklady vícestavových situací. **A**. Opakovaná stejná událost U. **B**. Vícestavová mezi událostmi U, V a Z. **C**. Soutěžící riziko mezi událostmi U a V.

První typ je obdobou repeated-measures designu. Stejné události se vyskytují třeba při sledování ontogeneze (doba k svlékání při jednotlivých instarech hmyzu), kvetení (doba ke kvetení úborů na jedné rostlině) atp.

Druhý typ zahrnuje odlišné, dynamicky (i zpětno-vazebně) provázané události, jako jsou třeba různé fáze predačního chování, fenologie rostlin atd. A dále mohou mít buď pevně danou, nebo náhodnou posloupnost nebo být bez posloupnosti. Třeba fenologie má jasně definovanou posloupnost událostí. Složitější situace je např. při léčbě pacienta, kdy se postup léčby odvíjí od reakce pacienta, a tudíž může vést k jiným posloupnostem událostí u různých pacientů.

Třetí typ je dosti specifický, protože se vyskytuje, pokud existuje několik událostí, ke kterým subjekt může dospět, ale jenom jedna se může u daného subjektu skutečně objevit. Jde o model soutěžícího rizika (angl. competing risks) – tedy o model založený

na představě soutěže mezi různými typy událostí o to, která z nich nastane první (a je pak jedinou pozorovanou událostí, zatímco ostatní již nastat nemohly). Třeba samice kudlanky se může se samcem spářit, nebo jej může sežrat (před pářením).

Hlavní problém analýzy dat, ve kterých u daného jedince může nastat více než jedna událost, spočívá v tom, že naměřené doby na stejném jedinci typicky nejsou nezávislé. Měření na jednom jedinci si jsou často podobnější než měření získaná na jedincích různých. Tím vzniká korelace (obecněji závislost) mezi měřeními na stejném subjektu. To je problém trochu podobný tomu, co jsme pozorovali v případě opakovaných měření či obecně modelů pro spojitá korelovaná data. Přístupy k analýze takových problémů v kontextu časů do události jsou typologicky v zásadě podobné těm, se kterými jsme se setkali již dříve (jsou ale poměrně značně odlišné v technických detailech). Tedy založené na náhodných efektech (v kontextu analýzy přežití nazývaných frailty), na marginálním přístupu analogickém ke GEE (sendvičový odhad pro korektní výpočet kovarianční matice potřebné pro testy Waldova typu), nebo na složitějších strukturách (podmíněných pravděpodobnostech souvisejících s přechody mezi různými stavy).

2.3 Měření času do události

Pro dobu do události, která se může vyskytnout pouze jednou u sledovaného subjektu (třeba úmrtí), je jako začátek pozorování často zvolena doba narození/vylíhnutí apod. Podobně u události, které se mohou vyskytovat opakovaně u stejného subjektu (útok, páření) v průběhu studie, se první doba do každé události počítá od začátku (narození).

Přesné měření času do události není vůbec snadné, i když je událost dobře definována. Sledování vždy pokrývá jistý interval (to je důležitý předpoklad – například se nesmí stát, že interval sice sledujeme, ale zda událost v něm nastala či nikoli, nevíme tak úplně přesně). Vždy je nutné zvolit začátek a konec sledování, jednoduše proto, že nelze události sledovat od $-\infty$ do ∞. Mnohdy jsme omezeni časem z různých důvodů (např. máme jinou práci) a nemůžeme sledované subjekty pozorovat, dokud k události nedojde u všech jedinců. Proto je velice běžné, že sledovaná událost po dobu sledování u některých jedinců nenastala. Jinými slovy, u takovýchto jedinců nevíme přesně, kdy by bývala nastala (nastat ale mohla, typicky předpokládáme, že u všech jedinců nastane v konečném, ale různém čase). Možná hned poté, co jsme sledování ukončili, ale je také možné, že by k ní došlo až hodně dlouho po ukončení studie. Takovým jedincům je mnohdy „reflexivně" (bezmyšlenkovitě) přiřazována hodnota nejdelšího možného času (tedy délky sledování, během něhož událost nenastala). Třeba 100 hodin, pokud interval sledování trval 100 hodin a my jsme u daného jedince žádnou událost nezaznamenali. Ale pozor, my ve skutečnosti víme, že k události za 100 hodin ještě nedošlo (to je důležitý předpoklad: nesmí se stát, že událost proběhla a my jsme si ji jen „nějak nevšimli"). Takže tato hodnota nese úplně jinou informaci než pozorování času, kdy k události došlo. Říká jen a pouze to, že do doby 100 hodin se událost neobjevila a že tedy nastala v čase větším než 100 hodin.

Doba do události je inherentně kontinuální veličina. Jenomže my ji dokážeme zaznamenat pouze s konečnou přesností, třeba na hodinu, na minutu, na sekundu (ale nikoli s přesností nekonečnou). Lze sice využít přístroje s „kontinuálním záznamem", třeba videorekordér či datalogger, ale i takové přístroje mají konečnou vzorkovací frekvenci, a tedy i omezenou přesnost.

To, že dobu do události lze prakticky zaznamenat pouze s konečnou přesností, znamená, že data ve skutečnosti nejsou spojitá, ale diskrétní. V důsledku toho se mohou některé hodnoty (s nenulovou pravděpodobností) opakovat. Velice často se doba do nějaké události, třeba úmrtí, zaznamenává pouze jednou za týden, třeba v pátek. Pokud v pátek nalezneme několik uhynulých jedinců, pak nelze rozlišit, kdy přesně uhynuli, ani jestli uhynuli najednou nebo každý z nich v jiný den. Známe pouze týdenní interval, ve kterém k události došlo, nic více, nic méně. Pokud je ovšem interval (např. vzniklý nepřesností přístroje, četností kontrol apod.) hodně krátký vzhledem k celkové délce sledování a relativně pomalému procesu, který sledujeme (např. měříme délku útěkové vzdálenosti velkého kopytníka, která je typicky stovky metrů, s přesností na metr), není neobvyklé data považovat za spojitá. Je však dobré o přijatých aproximacích (např. spočívajících v náhradě díky zaokrouhlování ve skutečnosti diskrétních dat spojitým modelem) vždy přemýšlet a kriticky zvažovat, zda je použité řešení rozumné. Pokud není, je třeba použít přístup komplexnější, ale realističtější (např. zohlednit intervalové cenzorování v použité analýze či rovnou pracovat s alternativním modelem přežití v diskrétním čase). Naopak je nevhodné postupovat dle předpřipravené kuchařky či „určovacího klíče statistických metod" a říkat si, že pro každou analýzu dob do události musí být model spojitý apod.

I když máme nastavený dostatečně dlouhý interval a dostatečnou vzorkovací frekvenci, nemusí to automaticky znamenat, že zaznamenáme výskyt každé události. Událost může být těžko rozpoznatelná např. proto, že nebylo dostatek světla, nebo proto, že je její výskyt kryptický. To je fundamentální problém zcela odlišný od cenzorování, které by nastalo, pokud víme, že událost v průběhu sledování zajisté nenastala (např. když se sledovaný subjekt přestěhoval a událost do doby jeho výstupu ze studie nepochybně nenastala). Ve skutečnosti pak jde o data s dodatečnými klasifikačními chybami (tzv. misclassification) – nepozorovaná událost může být s určitou pravděpodobností jen nezaznamenaná skutečná událost, zatímco pozorované události jsou událostmi skutečnými, takže typicky máme co do činění s falešně negativními, ale ne falešně pozitivními záznamy událostí. Standardní modely analýzy přežití předpokládají, že ke klasifikačním chybám nedochází (všechny události správně zaznamenáme). Rozšíření na situaci, kdy tomu tak není, existují. Jsou ale komplexnější a často problémově-specifická. Pokud v datech nemáte všechny události zaznamenány (dochází k falešně negativním záznamům událostí), určitě si promluvte s profesionálním statistikem. Řešení bude záležet na pravděpodobnosti nesprávné klasifikace i dalších faktorech. Mějme ale na paměti, že je vždy jednodušší věnovat maximální pozornost experimentálnímu či observačnímu protokolu a mít data kvalitní, bezproblémová než se *ex post* problémy pokoušet odstranit pomocí sofistikovaných modelů (při troše smůly to ani jít nemusí).

2.4 Cenzorování

Ideálně máme přesná a bezchybná pozorování časů do událostí u všech subjektů. Jednou z (relativně snadno řešitelných a dobře prostudovaných) komplikací je cenzorování. Jde o obecný (a v praxi hodně častý) problém, který mezi studovanou veličinu (např. čas přežití, či čas do daného typu události) a data klade z různých důvodů (např. logistických, technických i principiálních) něco jako „tlusté sklo" (např. Therneau & Grambsch 2000, Hendl 2004). Obecně mluvíme o cenzorování I. a II. typu (Lawless 2003, Rausand & Hozland 2004). V biologických aplikacích je podstatně častější cenzorování I. typu, a proto se mu budeme věnovat podrobněji. Jde o cenzorování dané časovými omezeními (např. studie končí po předepsané době). Cenzorování II. typu je běžnější v experimentech pro ověření spolehlivosti v technických aplikacích (neznamená to však, že by jej nebylo možné stejně tak dobře použít v biologických studiích, např. při stanovování LD50) a jde o cenzorování dané omezeními na počet sledovaných subjektů (např. studie končí, pokud došlo k událostem u 10 jedinců z celkového počtu $N > 10$ do studie zařazených).

Z různých důvodů nemusí k události během studie dojít. Může se stát, že k události nedojde až do ukončení studie – pak mluvíme o **cenzorování (I. typu) zprava**. Kromě cenzorování zprava existují v rámci I. typu i jiná cenzorování. Může dojít k výskytu události už před počátkem sledování (a to, pokud subjekt nesledujeme od úplného „začátku"). Představme si to na konkrétní situaci. Chceme zjistit datum klíčení nějakého druhu rostliny v přírodě. Podle dostupných informací stanovíme interval doby klíčení, řekněme od 1. do 31. března, a naplánujeme si, že lokalitu budeme kontrolovat každý den. Jenomže realita pak může být i taková, že první den sledování (1. března) na lokalitě již nalezneme několik rostlin vyklíčených. U nich jsme prošvihli skutečný den klíčení. Víme pouze, že klíčení nastalo před 1. březnem. O takovéto časové hodnotě pak mluvíme jako o **zleva cenzorované** (víme, že doba do klíčení musela být menší než doba mezi táním a 1. březnem).

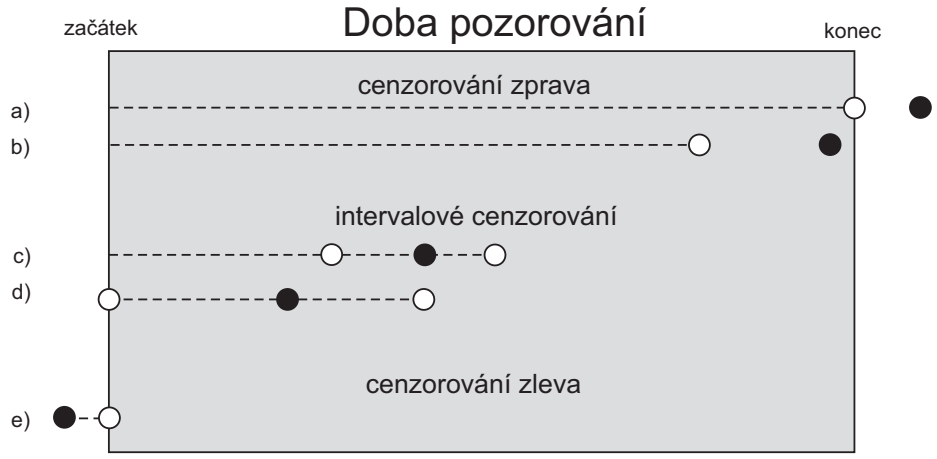
V praxi se ale můžeme snadno setkat i s komplexnějšími schématy, než je cenzorování zprava či zleva. Jmenovitě s tzv. **intervalovým cenzorováním**. To nastává, pokud víme, že daná událost nastala v nějakém (nezanedbatelně dlouhém) intervalu, ale už nevíme, kdy přesně. Například víme, že konkrétní zub na minulé pravidelné stomatologické prohlídce ještě nebyl patrný, ale na té současné (např. o rok později) už ano.

Je zřejmé, že cenzorování zleva (řekněme v čase A) je vlastně speciálním případem intervalového cenzorování – jde o intervalové cenzorování $(0, A)$, kdy víme pouze, že k události došlo do doby A. Také cenzorování zprava (řekněme v čase B) je speciálním případem – jde o intervalové cenzorování (B, ∞). Jsou to ale případy důležité a časté, takže se typicky řeší samostatně (v teorii i software, včetně typických R knihoven a procedur). V reálných datech máme typicky směs cenzorovaných a necenzorovaných (u těch je doba události známa přesně) dat. Leckdy se v datech vyskytuje i několik režimů cenzorování I. typu najednou. Existují i modely náhodného cenzorování – těm se ale v této knize věnovat nebudeme.

Na obr. 2-2 jsou zobrazeny různé případy cenzorování I. typu.

Kromě cenzorování může dojít také k ořezání či useknutí (angl. truncation). Jde o problém jiného typu než cenzorování. Váže se k apriorní selekci jedinců, kteří budou zařazeni do studie. Například musí splňovat nějakou podmínku, aby postoupili do studie (ve smyslu předem daných „inclusion criteria"). Podobně jako u cenzorování, může i u ořezávání nastat ořez zleva, ořez zprava a intervalový ořez. Rozdíl mezi levým cenzorováním a ořezáním je v tom, že v prvním případě všichni jedinci vstoupí do studie, ve druhém nikoliv. Například vstupní kritérium stanovující, že do studie jsou nabíráni jen jedinci s časem do události větším než A, povede k ořezávání zleva. Požadavek na vstup udávající maximální čas do události jako menší než B povede k ořezávání zprava. Požadavek na čas do události v intervalu $[A, B]$ povede k intervalovému ořezávání. Jedinci, kteří mají čas do události menší než A nebo větší než B, pak do studie nejsou vůbec zahrnuti (o jejich skutečných časech nevíme vůbec nic, nestudujeme je a nemůžeme se k nim nikterak vyjadřovat). Ořezávání často vzniká v důsledku různých technických omezení, např. nejsme schopni měřit hodnoty menší nebo větší než nějaké limity dané použitým přístrojovým vybavením, anebo v důsledku vědeckého zájmu soustředěného na jistý časový interval apod. V konkrétních datech může klidně docházet jak k cenzorování I. typu, tak k ořezávání.

Zdůrazněme, že čas cenzorování I. typu či čas ořezávání může, ale také nemusí být pro všechny jedince stejný (může se klidně lišit, např. v důsledku různých podmínek pozorování subjekt od subjektu). Důležité je, aby cenzorování bylo tzv. neinformativní. To



Obr. 2-2. Různé případy tří typů cenzorování. Plné kolečko označuje skutečný výskyt sledované události, bílé kolečko označuje kontrolní body v čase. Přerušované čáry vyznačují interval, kdy byl jedinec ve studii. a) Ke sledované události v době pozorování nedošlo. b) Jedinec (ve známém čase) utekl, k události u něj sice došlo, ale nezaznamenali jsme ji. c) Událost se vyskytla někdy mezi dvěma kontrolními daty. d) Události jsme si nevšimli, ale muselo k ní dojít, protože to ukázal jiný indikátor. e) Událost se vyskytla před započetím studie.

znamená, že doba cenzorování by měla být nezávislá na výskytu události. Drastickým příkladem nesplnění této podmínky je např. situace v klinické studii, kdy se po aplikaci léku zhorší stav pacientů natolik, že nejsou schopni přijít na kontrolu (Kartsonaki 2016) – takže cenzorování je totožné či téměř totožné se selháním.

Když to shrneme, ve vektoru času (který budeme zadávat jako jeden ze vstupních argumentů modelovací procedury) bude uveden čas pro každý sledovaný subjekt. U necenzorovaných subjektů to bude čas, kdy u nich nastala sledovaná událost, a u (zleva či zprava) cenzorovaných subjektů to bude čas cenzorování. Pro zleva cenzorované datum to bude čas první kontroly subjektu. Pro zprava cenzorované datum pak čas poslední kontroly subjektu. Závislá proměnná (t_i) tak bude párována s indikátorem cenzorování (δ_i). Jde o binární proměnnou, obsahující pouze hodnoty 0 anebo 1: 0 pro každého cenzorovaného jedince a 1 pro všechny necenzorované jedince.

To, jak cenzorování ovlivní odhad doby do události, si ukážeme na jednoduchém příkladu. Řekněme, že sledujeme dobu do nakladení vajec u nějakého druhu ptáka, a naměřili jsme 50, 70 a 100+ hodin. To poslední měření je cenzorované. Pokud takové měření vynecháme, bude průměr měření 60 hodin. Pokud jej budeme považovat za přesné měření, bude průměr 73.3 hodin. No, a pokud jej budeme brát jako cenzorované měření, pak bude průměr (za určitých předpokladů o typu rozdělení sledovaných časů) dokonce 90.6 hod. Vidíme, že odhad průměru je pro první dvě možnosti podstatně menší než po zahrnutí cenzorování. Pokud cenzorování zprava není správně ošetřeno, může to vést k hrubému systematickému vychýlení odhadů. Podobně je tomu u jiných typů cenzorování anebo ořezávání.

V některých situacích formálně podobných „cenzorování zprava" by šlo data analyzovat dvěma způsoby, nejprve kvalitativně a pak kvantitativně. To jest nejprve porovnat frekvenci výskytu událostí, např. mezi úrovněmi faktoru, formou logistické regrese. A pak pro případy, kdy k události došlo, porovnat dobu do události. Takovýto postup by byl validní, pouze pokud bychom dopředu věděli, že u cenzorovaných jedinců z nějakého důvodu událost ve skutečnosti ani nastat nemohla a tvoří tedy separátní, dobře definovanou skupinu. To je něco jiného než situace, na kterou cílí standardní analýza dob do události. Ta je totiž postavena na předpokladu, že k události musí u každého subjektu jednou dojít.

Odhady parametrů ve statistických modelech dob do události jsou typicky pořizovány metodou maximální věrohodnosti (maximum likelihood estimation), tedy optimalizací věrohodnostní funkce (King 1989). Pro standardní modely rozdělení dob do události je vyhodnocení věrohodnostní funkce kompletně pozorovaných dat (časy všech událostí pozorovány přesně a bez komplikací) přímočaré. Cenzorování a ořezávání vedou ke zcela odlišným modifikacím relevantních částí věrohodnostní funkce. Ořezávání s sebou často přináší větší numerické problémy při optimalizaci. Masivní cenzorování či ořezávání (vysoký podíl neúplných pozorování) vede k problémům statistickým i numerickým (parametry jinak dobrého modelu nemusí být identifikovatelné, nebo jsou odhadnutelné jen s gigantickou střední chybou apod.).

Je dobré si uvědomit, že principů odhadu pro modely časů do události lze s výhodou a leckdy vcelku přímočaře použít i v jiných, zdánlivě velmi vzdálených oblastech. Příkladem může být úporný, a i po mnoha letech od zavedení velké zmatky vyvolávající pojem Limit Of Detection (LOD, MacDougall et al. 1980). Jde o chemicko-analytický koncept, který se v konečných datech (např. při měření koncentrací toxických látek v životním prostředí) projevuje tím, že část hodnot je známa přesně, ale u druhé části měřených dat (s pozorovanou koncentrací nižší než laboratorně stanovený LOD) je známo jen to, že jsou pod mezí detekce. Pro pozorného čtenáře nebude těžké si představit, že (vždy kladné) koncentrace se chovají kvalitativně podobně jako „čas" (dokonce jejich pravděpodobnostní rozdělení bývá vpravo vyšikmené a mnohdy velmi podobné pravděpodobnostním modelům standardně používaným v analýze času do události) a že přítomnost dat pod mezí detekce odpovídá cenzorování zleva. Odtud je pak už jednoduché použít stabilní a dobře prostudované modely analýzy přežití i jejich stabilní softwarové implementace (např. v R) k analýze reálných chemicko-analytických výsledků namísto problematických, ale často i přímo nesprávných *ad hoc* metod objevujících se v chemické literatuře, ale bohužel stále i v praxi.

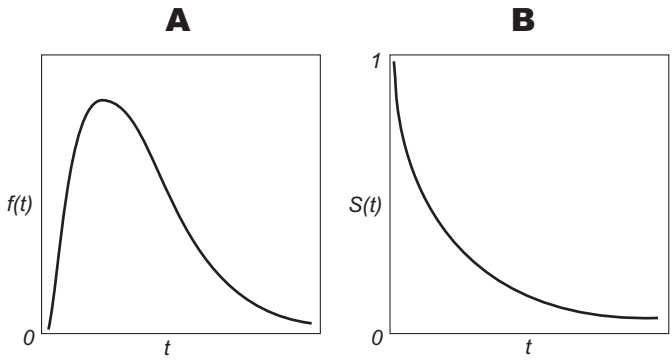
2.5 Trocha teorie

Podívejme se nyní na teorii, která je nezbytná k pochopení analýzy doby do události. Začneme definicemi. Předpokládáme, že měření času, T, je náhodná veličina. Hodnoty jsou vždy nezáporné a nenulové (tedy kladné)! Pokud jsme naměřili 0, pak to znamená, že jsme použili nepřesné měření času, protože z principu doba do události nemůže být nulová (muselo uběhnout alespoň pár milisekund). **Distribuční funkce**, $F(t)$, pozorovaného času, jakožto náhodné veličiny, se spočte z hustoty pravděpodobnosti, $f(t)$ (můžeme ji vnímat jako jakousi matematicky korektní formalizaci histogramu) podle vzorce:

$$F(t) = \int_0^t f(x)dx\,. \tag{2-1}$$

Rozdělení pozorovaných časů je typicky vyšikmené – má asymetrickou hustotu pravděpodobnosti (obr. 2-3A). Více si o konkrétních typech rozdělení řekneme v kap. 6.

Přestože událostí může být vcelku cokoliv, objasníme si teorii na úmrtí jakožto velice často sledované události. Budeme tedy mluvit o **křivkách přežití**, abychom byli ve shodě s ostatními učebnicemi. Tím jsme naznačili, co typický graf zobrazuje. Na ose y se vykresluje pravděpodobnost přežití delšího času než t (pravděpodobnost toho, že k úmrtí nedošlo do času t), tedy $S(t)$, a na ose x čas t (obr. 2-3B). Všimněte si, že přestože je událostí úmrtí, do grafu se vykresluje přežití, tedy opačný jev. Na to si musíme dát pozor, protože když budeme studovat např. dobu do páření, pak na ose y nebude (v default nastavení) pravděpodobnost páření (jak bychom asi očekávali), ale pravděpodobnost nepáření. To lze samozřejmě změnit použitím správných argumentů u funkce přežití či distribuční funkce.



Obr. 2-3. Příklad hustoty pravděpodobnosti (**A**) a funkce přežití (**B**).

Funkce přežití, $S(t)$, označuje pravděpodobnost, že u jedince dojde ke sledované události později než v čase t (jedinec zůstane v původním stavu až do času t, včetně). Tato funkce je vázána k distribuční funkci jednoduchým vztahem:

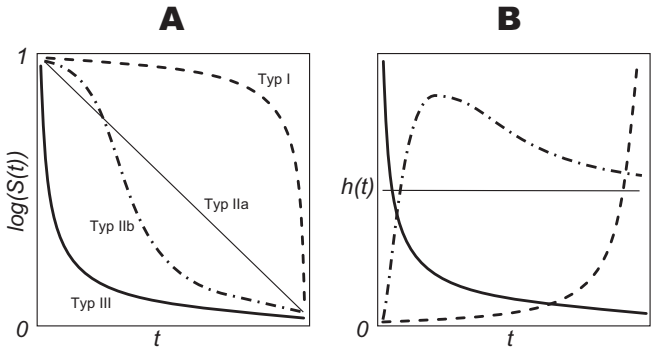
$$S(t) = 1 - F(t). \tag{2-2}$$

Tato funkce je klesající (přesněji nerostoucí), na rozdíl od distribuční funkce, která je rostoucí (neklesající) (obr. 2-3A). Na počátku studie, tedy v čase $t = 0$ je $S(0) = 1$ (nebo 100, pracujeme-li s procenty), tj. u žádného subjektu událost ještě nenastala. V průběhu studie k události u některých jedinců došlo, a jdeme-li v čase do nekonečna, pak se S blíží 0, tj. u všech subjektů k události musí dojít. To je ve skutečnosti jeden z předpokladů, se kterým budeme následující analýzy provádět. Pro jiné události než úmrtí to nemusí být ani zdaleka samozřejmě splněný předpoklad! Vždy se o použitých modelech a o tom, zda jsou pro daná data rozumná, vyplatí chvíli přemýšlet.

S funkcí přežití se mnozí z vás dozajista už setkali. V ekologii nebo v demografii se tato funkce běžně užívá k zobrazení přežití populace jedinců. V demografii se rozlišují tři základní tvary „křivek přežití" podle Pearle (1928) (obr. 2-4A):

- Typ I s největším rizikem úmrtí ve vyšším věku,
- Typ II s rizikem úmrtí více méně konstantním pro všechny věky,
- Typ III s největším rizikem úmrtí v nižším věku.

Obecně se mluví o **riziku** výskytu události neboli **funkci rizika** (angl. hazard rate), $h(t)$. Termíny riziko nebo hazard jsou v kontextu sledování přežití velmi hluboko zakořeněny a hojně používány. Lze je ale s úspěchem použít i v kontextu modelování jiných (kladných hodnot) náhodných proměnných. Pokud budeme studovat třeba čas do klíčení, pak lze snadno modelovat „riziko klíčení" (nyní již bez negativní konotace) apod.



Obr. 2-4. A. Tvar čtyř křivek přežití na logaritmické škále. **B.** Tvar funkce rizika pro čtyři typy křivek přežití v závislosti na čase (věku). Typ čáry v **B** odpovídá tomu v **A.**

Funkce rizika (angl. hazard function), $h(t)$, je definována jako:

$$h(t) = \frac{f(t)}{S(t)}, \tag{2-3}$$

tedy podíl hustoty pravděpodobnosti a funkce přežití. Formálně jde o podmíněnou pravděpodobnost výskytu události v (infinitezimálně) malém intervalu $(t,\ t + \Delta)$ za podmínky, že událost nenastala do času t. Na rozdíl od hustoty pravděpodobnosti, $f(t)$, která je nepodmíněnou (marginální) pravděpodobností výskytu události v (infinitezimálně) malém intervalu $(t,\ t + \Delta)$. Pragmaticky vzato, je riziko vlastně nástrojem typu „zvětšovacího skla". Skvělý nástroj k průzkumu pravého chvostu rozdělení (pro velké hodnoty t, kde je hodnota hustoty $f(t)$ již malá). Dělení (pro velká t) malou hodnotou $S(t) = 1 - F(t)$ umožňuje čitatel normalizovat/navýšit a studovat jeho chování pro velká t v detailu.

Funkce rizika může mít různý tvar. Pro Typ I se riziko monotónně zvětšuje, pro Typ IIa je konstantní a pro Typ III se monotónně zmenšuje (obr. 2-4B). Kromě těchto základních typů existuje celá řada složitějších průběhů, ve kterých funkce rizika nemusí být monotónní (Typ IIb).

Kromě okamžitého rizika nás může zajímat tzv. **kumulativní riziko**, $H(t)$, jež se spočte jako integrál rizika pro daný časový interval, tedy:

$$H(t) = \int_0^t h(x)dx. \tag{2-4}$$

Kumulativní riziko je rovno zápornému logaritmu funkce přežití:

$$H(t) = -\log\left(S(t)\right). \tag{2-5}$$

2.6 Empirické odhady

Podobně jako v předchozích dílech, ani zde se nebudeme podrobněji věnovat neparametrickým metodám. Nebudeme je používat v analýze, pouze v EDA. Proto si podrobněji představíme pouze ty funkce, které jsou vhodné k prvotní grafické analýze.

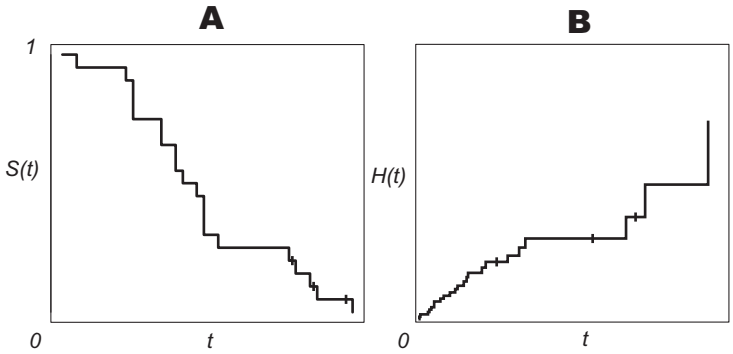
Jde o tzv. empirické odhady funkce přežití. Tyto křivky jsou jiné, než jsme viděli na obr. 2-4A. Mají sice podobný trend, ale jsou charakteristické schodovitým tvarem. A to proto, že vycházejí přímo z naměřených (empirických) hodnot, nikoli z teoretického kontinuálního modelu (jde o empirický odhad teoretického spojitého modelu). Jeden z nejpoužívanějších je **Kaplan-Meierův odhad**. Počítá se podle vztahu:

$$\hat{S}(t) = \prod_{t_i \le t} \left(\frac{n_i - d_i}{n_i} \right), \tag{2-6}$$

kde n_i je počet jedinců „v riziku", tedy jedinců, pro které událost v čase t_{i-1} ještě nenastala a nejsou cenzorovaní (tj. jsou živí a ne-cenzorovaní) a d_i je počet jedinců, pro které událost v časovém intervalu $[t_{i-1}, t_i)$ nastala. $t_1 < t_2 < \ldots < t_k$ jsou (podle velikosti seřazené) časy, ve kterých došlo alespoň k jedné události. To znamená, že přežije-li do druhého časového okamžiku 9 z 10 jedinců a do třetího 8 z 9 jedinců, je hodnota přežití rovna v t_2: $\frac{10-1}{10} = 0.9$ a v t_3: $\frac{9-1}{9} \times 0.9 = 0.8$.

Pro Kaplan-Meierův odhad křivka u (zprava) cenzorovaných událostí neklesá, ale nahrazuje se symbolem + a zůstává na předchozí hodnotě (obr. 2-5A). Obdobně lze spočíst (po částech konstantní) odhad funkce rizika, $h(t)$, konstantní v intervalu $[t_{i-1}, t_i)$, empiricky jako poměr počtu uhynulých (d_i) k počtu živých necenzorovaných (n_i) v čase t_i:

$$\hat{h}(t_i) = \frac{d_i}{n_i}. \tag{2-7}$$



Obr. 2-5. A. Kaplan-Meierova křivka přežití se třemi případy cenzorování (svislé čárky). **B.** Empirické kumulativní riziko se třemi případy cenzorování.