



Vysoká škola báňská – Technická univerzita Ostrava



STATISTIKA II.

učební text

Radim Briš, Martina Litschmannová

Ostrava 2007

Recenze: Marcela Rabasová

Název: Statistika II.
Autor: Radim Briš, Martina Litschmannová
Vydání: první, 2007
Počet stran: 149

Studijní materiály pro studijní obor Výpočetní matematika- IKT (N-VM) fakulty FEI
Jazyková korektura: nebyla provedena.

Určeno pro projekt:

Operační program Rozvoj lidských zdrojů

Název: E-learningové prvky pro podporu výuky odborných a technických předmětů

Číslo: CZ.O4.01.3/3.2.15.2/0326

Realizace: VŠB – Technická univerzita Ostrava

Projekt je spolufinancován z prostředků ESF a státního rozpočtu ČR

© Radim Briš, Martina Litschmannová

© VŠB – Technická univerzita Ostrava

ISBN 978-80-248-1482-7

POKYNY KE STUDIU

STATISTIKA II.

Pro předmět 3. semestru oboru N-VM jste obdrželi studijní balík obsahující

- integrované skriptum pro distanční studium obsahující i pokyny ke studiu
- CD-ROM s doplňkovými animacemi vybraných částí kapitol
- harmonogram průběhu semestru a rozvrh prezenční části
- rozdělení studentů do skupin k jednotlivým tutorům a kontakty na tutorý
- kontakt na studijní oddělení

Prerekvizity

Pro studium tohoto předmětu se předpokládá absolvování předmětu Statistika I.

Cílem předmětu

je seznámení se základními pojmy teorie spolehlivosti. Po prostudování modulu by měl student být schopen orientace v základních statistických metodách používaných ve výzkumu i v praxi

Pro koho je předmět určen

Modul je zařazen do magisterského studia oboru Výpočetní matematika studijního programu Informační a komunikační technologie, ale může jej studovat i zájemce z kteréhokoliv jiného oboru, pokud splňuje požadované prerekvizity.

Skriptum se dělí na části, kapitoly, které odpovídají logickému dělení studované látky, ale nejsou stejně obsáhlé. Předpokládaná doba ke studiu kapitoly se může výrazně lišit, proto jsou velké kapitoly děleny dále na číslované podkapitoly a těm odpovídá níže popsána struktura.

Při studiu každé kapitoly doporučujeme následující postup:



Čas ke studiu: xx hodin

Na úvod kapitoly je uveden **čas** potřebný k prostudování látky. Čas je orientační a může vám sloužit jako hrubé vodítko pro rozvržení studia celého předmětu či kapitoly. Někomu se čas může zdát příliš dlouhý, někomu naopak. Jsou studenti, kteří se s touto problematikou ještě nikdy neseťkali a naopak takoví, kteří již v tomto oboru mají bohaté zkušenosti.



Cíl: Po prostudování tohoto odstavce budete umět

- popsat ...
- definovat ...
- vyřešit ...

Ihned potom jsou uvedeny cíle, kterých máte dosáhnout po prostudování této kapitoly – konkrétní dovednosti, znalosti.



Výklad

Následuje vlastní výklad studované látky, zavedení nových pojmů, jejich vysvětlení, vše doprovázeno obrázky, tabulkami, řešenými příklady, odkazy na animace.



Shrnutí pojmů 1.1.

Na závěr kapitoly jsou zopakovány hlavní pojmy, které si v ní máte osvojit. Pokud některému z nich ještě nerozumíte, vraťte se k nim ještě jednou.



Otázky 1.1.

Pro ověření, že jste dobře a úplně látku kapitoly zvládli, máte k dispozici několik teoretických otázek.



Úlohy k řešení 1.1.

Protože většina teoretických pojmů tohoto předmětu má bezprostřední význam a využití v databázové praxi, jsou Vám nakonec předkládány i praktické úlohy k řešení. V nich je hlavní význam předmětu a schopnost aplikovat čerstvě nabyté znalosti při řešení reálných situací hlavním cílem předmětu.



KLÍČ K ŘEŠENÍ

Výsledky zadaných příkladů i teoretických otázek výše jsou uvedeny v závěru učebnice v Klíči k řešení. Používejte je až po vlastním vyřešení úloh, jen tak si samokontrolou ověříte, že jste obsah kapitoly skutečně úplně zvládli.

Úspěšné a příjemné studium s touto učebnicí Vám přejí autoři výukového materiálu

Radim Briš a Martina Litschmannová

1. MODELY A MODELOVÁNÍ



Čas ke studiu: 0,5 hodiny



Cíl: Po prostudování této kapitoly budete umět:

- charakterizovat model jako nástroj pro zobrazení skutečnosti
- popsat proces modelování
- provést klasifikaci základních modelů
- vysvětlit pojem matematický model
- vysvětlit pojmy: stochastický a deterministický model
- popsat různé přístupy k modelování, jako dedukce, indukce a retrodukce



Výklad

1.1. Model

Pojem **model** se vyskytuje v odborné literatuře velmi často. Teorie modelů a modelování nabyla v souvislosti s rozvojem kybernetiky značného metodologického významu a modely nacházejí uplatnění v nejrůznějších oborech. Termín model může být chápán různě a modely mohou sloužit odlišným cílům. Problematika modelování zahrnuje velké množství různorodých otázek, takže jsme nuceni omezit se jen na ty, které přímo nebo nepřímo aspoň částečně souvisí s použitím statistických metod. Různé názory na obecnou podstatu modelů, na jejich obsah, klasifikaci a především funkci netvoří ani zdaleka ucelenou teorii s přesně vymezenou a jednotnou terminologií. Konstrukce modelu a pravidla této konstrukce jsou vázána na řešení konkrétních úloh teoretického i praktického rázu, a je proto zřejmé, že při posuzování metodologických otázek je třeba k této skutečnosti přihlídnout.

Při sledování jevů a procesů reálného světa si uvědomujeme, že je v naprosté většině případů nejsme schopni zcela vysvětlit. Jen velmi obtížně postihujeme zákonitosti jejich vzniku a ještě hůře pronikáme do jejich vazeb a souvislostí. **Modelování** je tvůrčí lidská činnost spočívající v idealizaci a zjednodušení dějů reálného světa. Většina autorů se shoduje v tom, že model musíme chápat jako určitou formu zobrazení skutečnosti. Rozdíly jsou pouze v tom, jaká je modelována skutečnost, jaké jsou modelovací prostředky a k jakému účelu model slouží.

Slovo model má svůj původ ve stavitelství, kde označuje míru, podle níž jsou vyjádřeny proporce stavby. Později dostal pojem model zásadně nový význam. Připouští se, že teorie nemusí být jen zobrazením skutečnosti její objektivní podobě, ale že může jít o její určitou idealizaci. Časté jsou případy, kdy je výhodnější operovat s modelem místo se skutečností z toho důvodu, že často ovládáme lépe pravidla modelovací techniky než pravidla nezachytitelné nebo přímo nepozorovatelné skutečnosti.

Gnozeologická podstata modelování vyplývá ze zákonů přírody a z historicky vzniklé schopnosti abstrahovat shodné vlastnosti různých objektů. Díky souvislostem, které mezi objekty existují, můžeme nepřímo sledovat některé objekty prostřednictvím jiných objektů. Přes mnohoznačnost pojmu model jej můžeme charakterizovat jako zjednodušenou formu zobrazení podstatných rysů zkoumaného úseku reality. Model je sestaven podle určitých pravidel, která dovolují napodobovat

chování a vlastnosti zobrazované reality. Model je nejen prostředkem získávání poznatků, ale pomocí modelu je také možno rozvinout teorii určité oblasti. Studium modelu umožňuje vyvodit některé poznatky o zobrazované skutečnosti jen v případě, pokud mezi skutečností a modelem existuje obdoba, která je pro poznávání skutečnosti nezbytná.

Činnost zaměřenou ke konstrukci modelu nazveme modelováním. Modelováním můžeme dojít k matematické teorii, která umožňuje vysvětlovat a objevovat souvislosti a částečně je i zobecňovat. Tento popis však nemůže opravovat nebo dokonce odstraňovat chyby způsobené nedokonalostí modelu samotného.

1.2. Jedna z možných klasifikací modelu

Samotné slovo model je tedy velmi mnohoznačné. Někteří autoři si dali práci a uvedli seznamy několika desítek výkladů významu pojmu model. Úplná definice modelu se dnes asi neobejde bez aparátu teorie množin a matematické logiky. Odlišné přístupy přitom najdeme v přírodních a technických vědách, logice a společenských vědách, jiné pojetí v kybernetice a jiných disciplínách. Východiskem při třídění modelů může být modelovaná skutečnost a prostředky modelování, jakož i charakter cílů, kterým konstrukce modelu slouží.

Velmi jednoduché je rozlišení **materiálních modelů** od **myšlenkových modelů**. Zatímco materiální modely zobrazují reálně existující objekty, modely druhé skupiny mají charakter spíše teoretický a existují jen v našem vědomí. Myšlenkové modely je možné dále třídit na **představové modely**, vytvářené hypotetickou konstrukcí nebo idealizací skutečnosti podle představ, a na **symbolické modely**, jejichž prvky jsou vytvářené symboly nebo znaky. Modely této skupiny mají velmi blízko k modelům, u kterých mají rozhodující význam logické a matematické vlastnosti, a nazývají se modely **logické** či **formální** nebo také *matematické*.

1.3. Matematické modely

Rovněž pojem matematický model lze chápat ve více významech. Většinou se matematickým modelem rozumí nějaká formalizovaná teorie, někdy i její matematické zobrazení, ale často se také (nepříliš šťastně) matematickým modelem označuje jakýkoli kvantifikovaný popis některých stránek skutečnosti. Úspěch matematického modelování závisí mimo jiné na našich schopnostech formalizovat teoretické i praktické poznatky o zkoumaném úseku reality. Jde o nalezení takového matematického aparátu, který odpovídá modelované skutečnosti a přitom respektuje účel, ke kterému byl model konstruován.

Matematický model musí (stejně jako každý jiný model) objektivním způsobem znázorňovat jevy a procesy reálného světa. Matematický model vyjadřuje zákonitosti jevů a procesů, a to jak v oblasti vědeckého poznávání, tak v oblasti praktické lidské činnosti. Je zajímavé, že i když matematické modely neobsahují žádné vztahy, které do nich nebyly vloženy, přesto poskytují poznání, které do nich nebylo vědomě dáno. Matematické modely mohou pomoci k poznání tím, že naznačí nebo dokonce umožní dokázat obecné výsledky, které byly obsaženy v souborech pozorování, ale nebyly z těchto souborů zřejmé. Mohou dávat podnět a inspiraci k budoucímu bádání.

Matematický model můžeme zjednodušeně definovat jako určitou formu zobrazení některých aspektů jevů a procesů reálného světa matematickými prostředky. Takovým prostředkem může být třeba soustava rovnic obsahující proměnné (veličiny) a konstanty (parametry).

1.4. Některé typy matematických modelů

Matematické modely lze třídit z různých hledisek. Za hlavní lze považovat odlišení **deterministických modelů** od **stochastických modelů**. Deterministické modely mají povahu

zákonitostí, jež při dodržení určitých předpokladů a podmínek vždy platí, neboli vyhovují každé konkrétní empirické situaci. Pro deterministické modely je charakteristické, že postavení všech veličin v modelu je nesporné a konkrétní hodnoty představují řadu pevně daných čísel. U deterministického modelu je známa nejen jeho struktura, která může být popsána třeba algebraickou nebo diferenciální rovnicí, ale nesporné jsou i hodnoty parametrů. Pro odlišení deterministických a stochastických vztahů není zatím podstatné, zda jsme k matematickému modelu došli logickým důkazem, kdy závěry vyplývají přímo z předpokladů, či zobecněním provedeným na základě empirických zkušeností.

Uvažujme Newtonův pohybový zákon: dráha y , kterou předmět na Zemi urazí za dobu t , je při určitých zjednodušeních dána rovnicí

$$y = vt - \frac{at^2}{2}$$

V této rovnici konstanty v , a představují počáteční rychlost a tíhové zrychlení. K této rovnici je možné dojít vhodnou úpravou diferenciální rovnice modelující pohyb tělesa na Zemi anebo zobecněním určitých pozorování, tedy induktivním (datově orientovaným) způsobem. Zatímco při deduktivní úvaze předpokládáme přesnou znalost hodnot v , a , vylučujeme vliv odporu vzduchu a provádíme některá další zjednodušení, při induktivní úvaze respektujeme chyby měření proměnných y, t (tyto proměnné se stávají náhodnými proměnnými Y, T) a do analýzy tím zahrnujeme i vliv některých dalších činitelů způsobujících, že platnost rovnice je pouze přibližná. Do rovnice vstoupil **prvek nejistoty** (náhody) a **hovoříme o modelu stochastickém**:

$$Y = vT - \frac{aT^2}{2}$$

Na rozdíl od **deterministického modelu** vyhovuje stochastický model konkrétním situacím jen přibližně a s určitou *pravděpodobností*. Stochastické modely bývají též označovány jako pravděpodobnostní a právě s nimi se v tomto textu budeme výhradně setkávat. V běžných úlohách různých vědních oborů existuje mnoho důvodů, proč získaná pozorování či měření mají charakter spíše náhodný než deterministický.

Pro stochastické modely je charakteristické, že dovolují poměrně přesnou matematickou manipulaci se vztahy mezi veličinami, i když ve skutečnosti platí tyto vztahy pouze přibližně.

Pro naše potřeby můžeme přijmout *pracovní* definici **stochastického modelu jako rovnice nebo soustavy rovnic obsahující náhodné veličiny, nenáhodné veličiny a parametry**.

Náhodné veličiny jsou proměnné, jejichž hodnoty předem neznáme, jsou dány provedením pokusu nebo pozorováním. **Nenáhodné veličiny** (někdy též označované jako *pevné* nebo *fixní* veličiny) jsou proměnné, jejichž hodnoty určujeme. Parametry jsou známé nebo častěji neznámé konstanty.

Potíže související s konstrukcí stochastického modelu vyplývají z nejistoty, která se týká i některých zcela základních otázek. Na prvním místě je třeba uvést nejistotu týkající se odlišení podstatných a nepodstatných veličin. Výběr proměnných, které by model měl obsahovat, je velmi složitý věcný i empirický problém. Nejistotou pociťujeme i kolem samotné matematické formy modelu. Informace teoretického rázu nemusí být dostatečné pro výběr konkrétní formy modelu. Nejistota se týká i oprávněnosti učiněných předpokladů, přesnosti měření (zjišťování), vhodnosti metody použité k odhadu parametrů atd. Matematické modelování je nepřetržitý proces srovnávání našich znalostí, předpokladů a úvah s výsledky zjišťování a s užitečností modelu z hlediska cílů, ke kterým byl sestaven.

Modely určené ke zkoumání vztahů mezi veličinami se obvykle dělí na modely funkční, modely pro účely řízení a modely predikční. Není třeba zdůrazňovat, že pokud známe skutečný *funkční vztah* mezi veličinami, jsme přímo v ideální situaci. Můžeme řídit, popř. kontrolovat i předpovídat hodnoty veličin, které jsou předmětem našeho zájmu. Případy, kdy máme podobné modely k dispozici, jsou

(odmyslíme-li si definiční vztahy) zcela výjimečné, přičemž funkční vztahy bývají většinou nelineární a obtížně interpretovatelné.

Znalost funkčního předpisu vyjadřujícího vztahy mezi veličinami nemusí ještě umožňovat řízení či kontrolu všech zúčastněných veličin. Užitečný model *řízení* může být někdy sestrojen jen tehdy, pokud jsou veličiny v úloze příčin zcela pod naší kontrolou a jsme schopni vypracovat podrobný a přesný plán experimentu.

Pokud nejsme z nejrůznějších důvodů schopni funkční model sestavit a plánovaný experiment nepřichází v úvahu, spokojujeme se většinou s modelem, který není v plné míře realistickým zobrazením skutečnosti a je pouze zjednodušujícím přiblížením k hlavním rysům chování a vztahu veličin. Modely této skupiny se někdy označují jako *predikční*. Důvodem k tomuto označení je zřejmě skutečnost, že právě úlohy související s předpovědí hodnot některých veličin na základě znalosti hodnot jiných veličin, se často řeší pomocí modelů, které jsou pouze zjednodušenou abstrakcí skutečnosti. Predikční modely jsou často užitečné a za určitých podmínek mohou naznačit vnitřní podstatu sledovaných jevů a procesů. Tyto modely bývají konstruovány především metodami regresní analýzy, což vyžaduje velkou obezřetnost vůči předpokladům a respekt k vypovídající schopnosti těchto metod.

Zjednodušující abstrakce je velmi často spojená s otázkou linearity, popř. se stupněm nelinearity modelu. Zkušenosti z různých vědních oborů ukazují, že většina systémů či procesů má nelineární charakter, což značně ztěžuje modelovací přístup. Částečným východiskem může být linearizující zjednodušení. Matematicky je možné problém linearizace řešit rozvojem nelineární funkce do Taylorovy řady se zanedbáním členů vyššího než prvního řádu. Jinou možností linearizace je zjednodušení, při kterém zanedbáme působení některých veličin a chování ostatních veličin do určité míry idealizujeme. Na jedné straně je tento přístup nebezpečný v tom smyslu, že lineární funkce bude příliš hrubým zobrazením skutečnosti, ale na druhé straně linearizace zjednodušuje interpretaci výsledků a zpracování dat.

Stochastické modely (a samozřejmě nejen ty) je možné dále třídit podle řady jiných hledisek. Například podle závislosti na čase rozlišujeme modely statické a dynamické, podle veličin v modelu na spojitě a nespojitě (diskrétní), atd.

1.5. Přístupy k modelování

Podle K. Pearsona (1938) jednota určité vědní disciplíny spočívá v samotných metodách této disciplíny a nikoli v oblasti, kde jsou tyto metody používány. Znamená to, že i když je třeba respektovat specifika různých vědních oborů, tak některé typy úsudků používané v jedné oblasti zkoumání svou podstatou nejsou zásadně odlišné od podobně utvořených úsudků v jiných oblastech.

Aristoteles uvádí tři typy vědeckých úsudků, deduktivní, induktivní a retroduktivní. Při deduktivní úvaze se postupuje od obecného k zvláštnímu a **dedukcí** se rozumí typ úsudků nebo metoda zkoumání, při níž podle určitých pravidel závěry jednoznačně vyplývají z předpokladů. Typickým příkladem je matematický důkaz nebo úsudek o realitě při znalosti modelu této reality, přičemž pravdivost výchozích tvrzení určuje i přesnost či pravdivost výsledků. V tomto smyslu paradoxně **teorie matematické** (říká se též **induktivní**) **statistiky** vyplývá z převážně deduktivních úvah, zatímco úvahy o *cílové populaci* na základě získaných výběrových údajů lze označit za induktivní úlohu. Při induktivní úvaze se postupuje ve srovnání s deduktivní úlohou obráceně, tedy od konkrétního k obecnému, od reality k modelu anebo od výběrových dat ke skutečným nebo hypotetickým populacím. Pro statickou indukci je charakteristické, že obecný závěr se vyvozuje na základě konkrétních pozorování.

Základním předpokladem vědeckého pokroku je neustálé hromadění poznatků získaných ze zkušeností. Podle cíle úlohy je znalost získaná tímto způsobem bohužel často jen *popisem*

napozorovaných skutečností, jindy navíc odborným či datově orientovaným *vysvětlením* různých okolností a zvláštností a jen někdy se výzkumník či zadavatel úlohy snaží *ovládnout* realitu poznáním a využitím vztahů, závislostí a souvislostí k prediktivním či zobecňujícím úvahám.

Nejspornější je **retroduktivní forma úsudku**, při které na základě zkušeností pouze vyvozujeme možnost výskytu určitého jevu nebo předpokládáme průběh určitého procesu a hledáme teoretické zdůvodnění nepozorovatelných skutečností. Tato oblast v souvislosti i s využíváním subjektivně pojímaných pravděpodobností bývá někdy v odsuzujícím významu označovaná až za *metastatistiku*. Úvahy tohoto typu jsou však nesporně potřebné a statistika v této oblasti zaznamenala nejen zásadní teoretický rozvoj, ale i mnoho užitečných využití.

Při konstrukci matematických modelů se setkáváme s různými přístupy. Přístup vycházející z *věcných znalostí* problematiky je velmi blízký deduktivní úvaze, při které předpokládáme že odpovídající modely jsou určitelné na základě obecných principů dané úlohy či příslušného vědního oboru. V poslední době se při modelování stále častěji doporučuje *kybernetický přístup*, při kterém je modelovaný systém pojímán jako známá či zcela fiktivní *skříňka* transformující určité vstupy (příčiny) na výstupy (důsledky). V publikaci *Statistical Science* 3/2001 byla popsána zajímavá debata o členění statistiků podle postoje k potřebě znalosti mechanismu této skříňky.

Nejednoznačnost takové transformace je způsobena neuvažovanými veličinami a předpokladem o náhodných složkách (poruchách) umožňujícím využít pravděpodobnostní principy. Pokud teorie zkoumaného úseku reality není dostatečně propracovaná a existují pouze hypotézy o chování jednotlivých veličin, používá se empirický přístup, který má značně subjektivní charakter a závisí na odborných znalostech i na intuici zpracovatele. Při empirickém modelování mají vytvořené modely často vztah pouze ke konkrétnímu souboru pozorování a zobecnění mimo obor hodnot vyskytujících se v souboru je problematické.



Shrnutí pojmů kapitoly 1.

Pojem **model** je velmi obecný a mnohoznačný. Přes mnohoznačnost pojmu model jej můžeme charakterizovat jako zjednodušenou formu zobrazení zkoumaného úseku reality. Model je sestaven podle určitých pravidel, která dovolují napodobovat chování a vlastnosti zobrazované reality. Model je nejen prostředkem získávání poznatků, ale pomocí modelu je také možno rozvinout teorii určité oblasti. Konstrukce modelu a pravidla této konstrukce jsou většinou vázána na řešení konkrétních úloh teoretického i praktického rázu. Činnost zaměřenou ke konstrukci modelu nazveme modelováním. **Modelování** je tvůrčí lidská činnost spočívající v idealizaci a zjednodušení dějů reálného světa.

Matematickým modelem se většinou rozumí nějaká formalizovaná teorie, někdy i její matematické zobrazení, ale často se jím také označuje jakýkoli kvantifikovaný popis některých stránek skutečnosti. Matematický model musí objektivním způsobem znázorňovat jevy a procesy reálného světa. Matematický model vyjadřuje zákonitosti jevů a procesů, a to jak v oblasti vědeckého poznávání, tak v oblasti praktické lidské činnosti. Matematický model lze zjednodušeně definovat jako určitou formu zobrazení některých aspektů jevů a procesů reálného světa matematickými prostředky. Takovým prostředkem může být třeba soustava rovnic obsahující proměnné (veličiny) a konstanty (parametry).

Matematické modely lze třídit z různých hledisek. Za hlavní lze považovat odlišení deterministických modelů od stochastických modelů. **Deterministické modely** mají povahu zákonitostí, jež při dodržení určitých předpokladů a podmínek vždy platí, neboli vyhovují každé konkrétní empirické situaci. Na rozdíl od *deterministického* modelu vyhovuje stochastický model konkrétním situacím jen přibližně a s určitou *pravděpodobností*. **Stochastické modely** bývají též označovány jako pravděpodobnostní. Pro ně je charakteristické, že dovolují poměrně přesnou matematickou manipulaci se vztahy mezi

veličinami, i když ve skutečnosti platí tyto vztahy pouze přibližně. Je pro ně charakteristická **nejistota**, kterou pocítujeme i kolem samotné matematické formy modelu. Zjednodušeně lze přijmout definici stochastického modelu jako rovnice nebo soustavy rovnic obsahující náhodné veličiny, nenáhodné veličiny (fixní, pevné) a parametry (konstanty). Nejjednodušší stochastické modely jsou **lineární**. Pro reálné složitější nelineární modely se používá linearizující zjednodušení.

Při konstrukci matematických modelů se setkáváme s různými přístupy. Přístup vycházející z *věcných znalostí* problematiky je velmi blízký **deduktivní** úvaze, při které předpokládáme že odpovídající modely jsou určitelné na základě obecných principů dané úlohy či příslušného vědního oboru. Při **induktivní** úvaze se postupuje ve srovnání s deduktivní úlohou obráceně, tedy od konkrétního k obecnému, od reality k modelu anebo od výběrových dat ke skutečným nebo hypotetickým populacím. Pro statickou indukci je charakteristické, že obecný závěr se vyvozuje na základě konkrétních pozorování. V poslední době se při modelování stále častěji doporučuje **kybernetický přístup**, při kterém je modelovaný systém pojímán jako známá či zcela fiktivní skříňka transformující určité vstupy (příčiny) na výstupy (důsledky).



Otázky 1.

1. Charakterizujte pojmy model a modelování.
2. Čím se odlišuje stochastický model od deterministického ?
3. Na čem jsou založeny logické procedury?

2. VYBRANÉ PRAVDĚPODOBNOSTNÍ MODELKY



Čas ke studiu: 1 hodina



Cíl: Po prostudování této kapitoly budete umět popsat a použít pro popis technických procesů:

- Erlangovo rozdělení
- Weibullovo rozdělení
- Logaritmicko – normální rozdělení
- Vícerozměrné normální rozdělení



Výklad

2.1. Erlangovo rozdělení

Určitým zobecněním exponenciální náhodné veličiny (doba do (první) poruchy) je náhodná veličina s Erlangovým rozdělením, která **popisuje dobu do výskytu k-té události v Poissonově procesu**.

Erlangovo rozdělení je **speciálním typem tzv. Gamma rozdělení pro k z množiny celých čísel**. (Tento vztah je vhodné znát, chceme-li k nalezení distribuční funkce, popř. hustoty pravděpodobnosti použít statistický software – některé statistické pakety mají implementováno pouze Gamma rozdělení a hodnoty Erlangova rozdělení pak získáme dosazením příslušných parametrů).

Erlangovo rozdělení má dva parametry: **k** – počet událostí (parametr tvaru, shape, α – v Gamma rozdělení), k nimž má dojít a rychlost výskytu těchto událostí λ (parametr měřítka, scale, β v Gamma rozdělení).

Má-li náhodná veličina X Erlangovo rozdělení, značíme to takto:

$$X_k \rightarrow \text{Erlang}(k, \lambda)$$

X_k = doba do výskytu k.události (na obr. k = 4)



Náhodnou veličinu s Erlangovým rozdělením si můžeme představit jako součet k nezávislých exponenciálních náhodných veličin (doba do výskytu k -té události je součtem dob mezi 0-tou a 1. událostí, 1. a 2. událostí, ..., $(k-1)$. a k . událostí).

Pro Erlangovo rozdělení s parametry k a λ platí tyto vztahy:

Hustota pravděpodobnosti:

$$f(t) = \lambda \cdot e^{-\lambda t} \cdot \frac{(\lambda t)^{k-1}}{(k-1)!}; \quad t > 0$$

Distribuční funkce:

$$F(t) = 1 - e^{-\lambda t} \cdot \sum_{j=0}^{k-1} \frac{(\lambda t)^j}{j!}$$

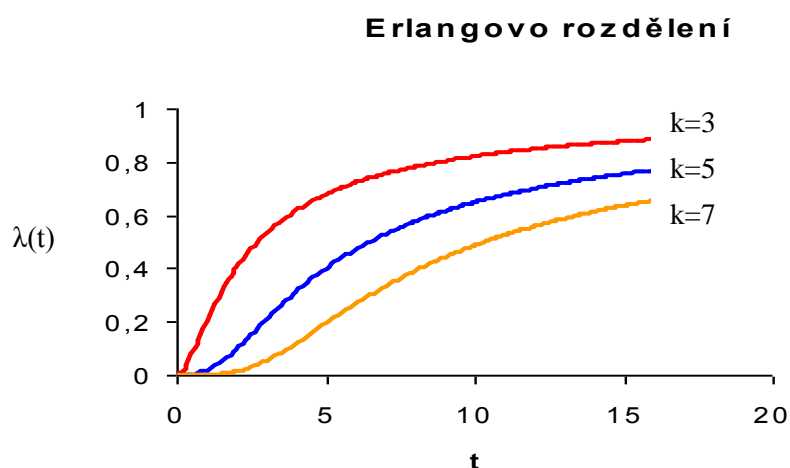
Intenzita poruch:

$$\lambda(t) = \frac{\lambda}{(k-1)! \sum_{j=0}^{k-1} \frac{1}{(k-1-j)! (\lambda t)^j}}$$

Střední hodnota: $EX_k = \frac{k}{\lambda}$

Rozptyl: $DX_k = \frac{k}{\lambda^2}$

Graf intenzity poruch Erlangova rozdělení pro $\lambda = 1$; $k = 3$; 5 ; 7



Intenzita poruch $\lambda(t)$ je v případě Erlangova rozdělení rostoucí funkce a proto je toto **rozdělení vhodné pro modelování procesů stárnutí**.



Průvodce studiem

Následující pasáž je určena pro zájemce o matematické pozadí používaných vztahů.

- **Odvození distribuční funkce Erlangova rozdělení**

Mějme:

X_k ... doba do výskytu k -té události v Poissonově procesu, $X_k \rightarrow \text{Erlang}(k; \lambda)$

N_t ... počet výskytu události v časovém intervalu $(0;t)$, $N_t \rightarrow \text{Po}(\lambda t)$

Platí, že v časovém intervalu $(0;t)$ nastane alespoň k událostí, právě když doba do výskytu k -té události je menší než t .

$$(N_t \geq k) \Leftrightarrow (X_k < t)$$

Z této ekvivalence lze odvodit distribuční funkci Erlangova rozdělení.

$$F(t) = P(X_k < t) = P(N_t \geq k) = 1 - P(N_t < k) = 1 - \sum_{j=0}^{k-1} \left(e^{-\lambda t} \cdot \frac{(\lambda t)^j}{j!} \right) = 1 - e^{-\lambda t} \cdot \sum_{j=0}^{k-1} \frac{(\lambda t)^j}{j!}$$

- **Odvození hustoty pravděpodobnosti**

Hustotu pravděpodobnosti získáme derivací distribuční funkce:

$$f(t) = \frac{dF(t)}{dt} = \lambda \cdot e^{-\lambda t} \cdot \sum_{j=0}^{k-1} \frac{(\lambda t)^j}{j!} + (-e^{-\lambda t}) \cdot \sum_{j=0}^{k-1} \frac{j \cdot (\lambda t)^{j-1} \cdot \lambda}{j!} =$$

$$= \lambda \cdot e^{-\lambda t} \cdot \sum_{j=0}^{k-1} \frac{(\lambda t)^j}{j!} - \lambda \cdot e^{-\lambda t} \cdot \sum_{j=1}^{k-1} \frac{(\lambda t)^{j-1}}{(j-1)!} =$$

$$= \lambda \cdot e^{-\lambda t} \cdot \left(\sum_{j=0}^{k-2} \frac{(\lambda t)^j}{j!} + \frac{(\lambda t)^{k-1}}{(k-1)!} \right) - \lambda \cdot e^{-\lambda t} \cdot \sum_{j=0}^{k-2} \frac{(\lambda t)^j}{j!} =$$

$$= \lambda \cdot e^{-\lambda t} \cdot \frac{(\lambda t)^{k-1}}{(k-1)!} + \lambda \cdot e^{-\lambda t} \cdot \frac{(\lambda t)^{k-1}}{(k-1)!} - \lambda \cdot e^{-\lambda t} \cdot \sum_{j=0}^{k-2} \frac{(\lambda t)^j}{j!} = \lambda \cdot e^{-\lambda t} \cdot \frac{(\lambda t)^{k-1}}{(k-1)!}$$

- **Odvození intenzity poruch**

$$\begin{aligned} \lambda(t) &= \frac{f(t)}{1 - F(t)} = \frac{\lambda \cdot e^{-\lambda t} \cdot \frac{(\lambda t)^{k-1}}{(k-1)!}}{e^{-\lambda t} \cdot \sum_{j=0}^{k-1} \frac{(\lambda t)^j}{j!}} = \frac{\lambda}{(k-1)! \cdot \sum_{j=0}^{k-1} \frac{(\lambda t)^j}{(\lambda t)^{k-1} \cdot j!}} = \frac{\lambda}{(k-1)! \cdot \sum_{j=0}^{k-1} \frac{(\lambda t)^{j-k+1}}{j!}} = \\ &= \frac{\lambda}{(k-1)! \cdot \sum_{j=0}^{k-1} \frac{1}{(\lambda t)^{k-1-j} \cdot j!}} = \frac{\lambda}{(k-1)! \cdot \sum_{j=0}^{k-1} \frac{1}{(\lambda t)^j \cdot (k-1-j)!}} \end{aligned}$$

- **Odvození střední hodnoty a rozptylu**

Mějme:

X_k ... doba do výskytu k-té události v Poissonově procesu, $X_k \rightarrow Erlang(k; \lambda)$

X ... doba do výskytu události v Poissonově procesu, $X \rightarrow E(\lambda)$

Je zřejmé, že Erlangova náhodná veličina (s parametry $k; \lambda$) je součtem k exponenciálních veličin (s parametrem λ):

$$X_k = \sum_{i=1}^k X_i$$

Z vlastností střední hodnoty víme, že střední hodnota součtu náhodných veličin je rovna součtu jejich středních hodnot:

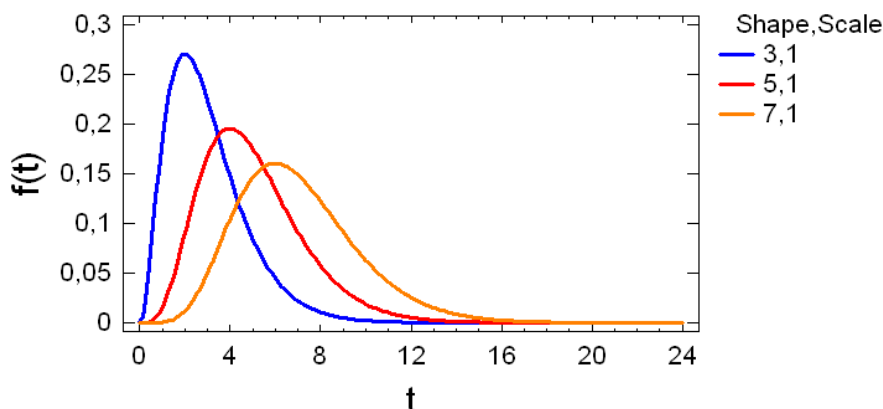
$$EX_k = \sum_{i=1}^k EX_i = \frac{1}{\lambda} + \frac{1}{\lambda} + \dots + \frac{1}{\lambda} = \frac{k}{\lambda}$$

Jednotlivé exponenciální náhodné veličiny jsou nezávislé a proto taktéž rozptyl součtu náhodných veličin je roven součtu jejich rozptylů:

$$DX_k = \sum_{i=1}^k DX_i = \frac{1}{\lambda^2} + \frac{1}{\lambda^2} + \dots + \frac{1}{\lambda^2} = \frac{k}{\lambda^2}$$

Na následujícím obrázku jsou **příklady hustoty Gamma rozdělení pro $\lambda = 1$ a různé hodnoty k** . Poznamenejme, že s rostoucím k roste rozptyl tohoto rozdělení a koeficient šikmosti se přibližuje nule (rozdělení je více symetrické).

Erlangovo rozdělení



2.2. Weibullovo rozdělení

Weibullovo rozdělení je velmi flexibilní (díky parametru β) a proto se jím zejména v teorii spolehlivosti popisují spojité náhodné veličiny definované jako **doba do poruchy** (doba bezporuchovosti). Používá se zejména při popisu komponent, které jsou **v období ranných poruch nebo v období stárnutí** (tj. tam kde se projevuje mechanické opotřebení nebo únava materiálu).

Weibullovo rozdělení má dva parametry: Θ – parametr měřítka (scale, $\Theta > 0$, závisí na materiálu, namáhání a podmínkách užívání) a β – parametr tvaru (shape, $\beta > 0$, na jeho hodnotě závisí tvar intenzity poruch a tím i vhodnost použití pro určité období doby života).

Má-li náhodná veličina X Weibullovo rozdělení, značíme to takto:

$$X \rightarrow W(\Theta, \beta)$$

Distribuční funkce:

$$F(t) = 1 - e^{-\left(\frac{t}{\Theta}\right)^\beta}; \quad t > 0; \Theta > 0; \beta > 0$$

Hustota pravděpodobnosti:

$$f(t) = \frac{\beta}{\Theta} \left(\frac{t}{\Theta}\right)^{\beta-1} e^{-\left(\frac{t}{\Theta}\right)^\beta}; \quad t > 0; \Theta > 0; \beta > 0$$

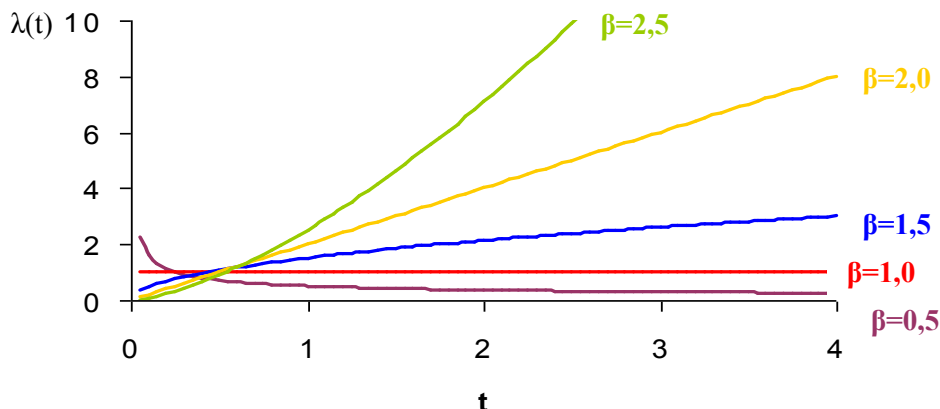
Intenzita poruch:

$$\lambda(t) = \frac{\beta}{\Theta} \left(\frac{t}{\Theta}\right)^{\beta-1}; \quad t > 0; \Theta > 0; \beta > 0$$

Ze vztahu pro intenzitu poruch Weibullova rozdělení je zřejmé, že:

$$\lambda(t) = konst \cdot t^{\beta-1}$$

a proto tvar intenzity poruch závisí na volbě parametru β .

Některé příklady intenzity poruch Weibullova rozdělení ($\Theta=1$):

Všimněme si, že pro $\beta=1$, přejde Weibullovo rozdělení v rozdělení exponenciální (konstantní intenzita poruch) s parametrem $\lambda = \frac{1}{\Theta}$.

$$\beta = 1 \Rightarrow W(\Theta;1) \rightarrow E\left(\frac{1}{\Theta}\right)$$

Z výše uvedeného grafu je rovněž zřejmé použití Weibullova rozdělení v závislosti na parametru β :

$0 < \beta < 1$	období dětských nemocí	$\lambda(t)$... klesající funkce
$\beta = 1$	období stabilního života	$\lambda(t) = konst. = \frac{1}{\Theta} = \lambda$ (exp. rozdělení)
$1 < \beta < 2$	období stárnutí	$\lambda(t)$... konkávní, rostoucí funkce
$\beta = 2$	období stárnutí	$\lambda(t)$... lineárně rostoucí funkce
$\beta > 2$	období stárnutí	$\lambda(t)$... konvexní, rostoucí funkce



CD-ROM

Na přiloženém CD-ROMu si můžete prohlédnout animace zobrazující vliv parametru tvaru Weibullova rozdělení na charakteristiky tohoto rozdělení.

2.3. Logaritmicko-normální rozdělení

Jestliže má náhodná veličina Y , $Y = \ln X$, normální rozdělení s parametry μ a σ^2 , pak náhodná veličina X má logaritmicko-normální rozdělení se stejnými parametry, což zapisujeme:

$$X \rightarrow LN(\mu; \sigma^2)$$

Z definice je zřejmé, že náhodná veličina s logaritmicko-normálním rozdělením může nabývat pouze kladných hodnot (definiční obor $\ln x$). Proto nachází uplatnění při **popisu náhodných veličin**

nabývajících pouze kladných hodnot a to zejména v případech, kdy hustota pravděpodobnosti je asymetrická (šikmost není nulová) s jedním vrcholem. Značný význam tohoto rozdělení tedy nacházíme v teorii spolehlivosti (různé parametry součástek nabývají pouze kladných hodnot – životnost, rozměry, tažnost, ...) a v ekonomii při popisu příjmů (příjmová rozdělení).

Hustota pravděpodobnosti:

$$f(x) = \begin{cases} \frac{1}{x\sigma\sqrt{2\pi}} \cdot e^{-\frac{(\ln x - \mu)^2}{2\sigma^2}}; & \text{pro } x > 0 \\ 0 & \text{pro } x \leq 0 \end{cases}$$

Distribuční funkce:

Distribuční funkci log.-normálního rozdělení nalezneme prostřednictvím distribuční funkce normovaného normálního rozdělení.

$$F(x) = \begin{cases} \Phi\left(\frac{\ln x - \mu}{\sigma}\right); & \text{pro } x > 0 \\ 0 & \text{pro } x \leq 0 \end{cases}$$

Střední hodnota: $EX = e^{\mu + \frac{\sigma^2}{2}}$

Rozptyl: $DX = e^{2\mu + \sigma^2} (e^{\sigma^2} - 1)$

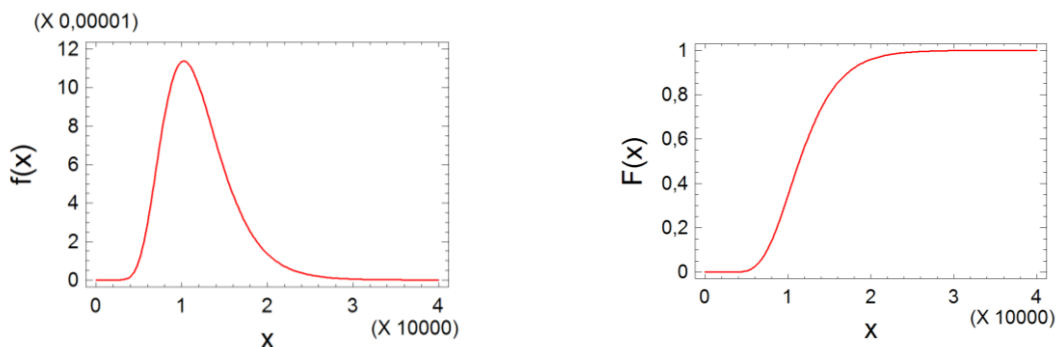
100p%-ní kvantil: $x_p = e^{\mu + \sigma \cdot z_p}$,

kde z_p je 100p%-ní kvantil normovaného normálního rozdělení

Grafické znázornění hustoty pravděpodobnosti a distribuční funkce:

X ... příjem zaměstnanců jisté firmy

$$X \rightarrow LN(12.000; 4.000^2)$$



Při praktickém používání tohoto rozdělení postupujeme tak, že náhodnou veličinu X nejdříve převedeme na $Y = \ln X$ a potom již postupujeme stejně jako u normálního rozdělení.



Průvodce studiem

A opět zde máme pasáž pro zájemce:

- **Odvození distribuční funkce logaritmicko-normálního rozdělení:**

Nechť:

$$Y = \ln X$$

$$X \rightarrow LN(\mu; \sigma^2) \Leftrightarrow Y \rightarrow N(\mu; \sigma^2)$$

$F_X(x)$ (resp. $F_Y(y)$) je distribuční funkce náhodné veličiny X (resp. Y)

$$\forall x > 0: F_X(x) = P(X < x) = P(e^Y < x) = P(Y < \ln x) = F_Y(\ln x) = \Phi\left(\frac{\ln x - \mu}{\sigma}\right)$$

$$\forall x \leq 0: F_X(x) = 0$$

- **Odvození hustoty pravděpodobnosti logaritmicko-normálního rozdělení:**

$f_X(x)$... hustota pravděpodobnosti náhodné veličiny X

$$\forall x > 0: f_X(x) = \frac{dF_X(x)}{dx} = \frac{d\Phi\left(\frac{\ln x - \mu}{\sigma}\right)}{dx} = \varphi\left(\frac{\ln x - \mu}{\sigma}\right) \cdot \frac{1}{x \cdot \sigma} = \frac{1}{x\sigma\sqrt{2\pi}} e^{-\frac{\left(\frac{\ln x - \mu}{\sigma}\right)^2}{2}}$$

$$= \frac{1}{x\sigma\sqrt{2\pi}} \cdot e^{-\frac{(\ln x - \mu)^2}{2\sigma^2}}$$

$$\forall x \leq 0: f_X(x) = 0$$

- **Odvození vztahu pro výpočet 100p%-ního kvantilu:**

$$P(X < x_p) = p$$

$$F(x_p) = p$$

$$\Phi\left(\frac{\ln x_p - \mu}{\sigma}\right) = p$$

$$\frac{\ln x_p - \mu}{\sigma} = z_p \quad (\Phi(z_p) = p)$$

$$\ln x_p = \sigma \cdot z_p + \mu$$

$$x_p = e^{(\mu + \sigma \cdot z_p)}$$



Řešený příklad

Nechť X je náhodná veličina s logaritmicko-normálním rozdělením s parametry: $\mu=2$; $\sigma^2=9$. Určete:

- pravděpodobnost, že náhodná veličina X je z intervalu $(0;30)$
- medián daného rozdělení
- střední hodnotu a rozptyl náhodné veličiny X

$$X \rightarrow LN(2;9)$$

ada) Pravděpodobnost, že náhodná veličina X je z intervalu $(0;30)$, můžeme určovat rovněž jako pravděpodobnost, že náhodná veličina X je menší než 30, neboť log.-normální náhodná veličina může nabývat pouze kladných hodnot.

Připomeňme si postup při určování distribuční funkce log.-normální náhodné veličiny:

$$F(x) = \begin{cases} \Phi\left(\frac{\ln x - \mu}{\sigma}\right); & \text{pro } x > 0 \\ 0 & \text{pro } x \leq 0 \end{cases}$$

A nyní již přejdeme k určení hledané pravděpodobnosti:

$$\underline{\underline{P(0 < X < 30) = F(30) - F(0) = \Phi\left(\frac{\ln 30 - 2}{\sqrt{9}}\right) - 0 = \Phi(0,47) = 0,681}}$$

nebo

$$\underline{\underline{P(0 < X < 30) = P(X < 30) = F(30) = \Phi\left(\frac{\ln 30 - 2}{\sqrt{9}}\right) = \Phi(0,47) = 0,681}}$$

adb) Pro určení mediánu můžeme použít vztah pro 100 p %-ní kvantil, který byl odvozen v Průvodci studiem:

$$x_p = e^{\mu + \sigma \cdot z_p}$$

$$z_{0,5} = 0 \quad \Rightarrow \quad \underline{\underline{x_{0,5} = e^{2 + \sqrt{9} \cdot 0} = e^2 \cong 7,4}}$$

adc) Střední hodnotu a rozptyl určíme na základě výše uvedených vztahů:

$$EX = e^{\mu + \frac{\sigma^2}{2}} \Rightarrow \underline{\underline{EX = e^{2 + \frac{9}{2}} = e^{\frac{13}{2}} \cong 665,1}}$$

$$DX = e^{2\mu + \sigma^2} (e^{\sigma^2} - 1) \Rightarrow \underline{\underline{DX = e^{2 \cdot 2 + 9} (e^9 - 1) \cong 3,6 \cdot 10^9}}$$

2.4. Vícerozměrné normální rozdělení

Uvažujme náhodný vektor \underline{X} . ($\underline{X} = (X_1, X_2, \dots, X_n)^T$). Vektor \underline{X} má n-rozměrné normální rozdělení pravděpodobnosti s parametry μ a Σ , jestliže jeho hustota pravděpodobnosti je:

$$f(x_1, x_2, \dots, x_n) = \frac{1}{(2\pi)^{n/2} |\Sigma|^{1/2}} \cdot e^{-\frac{1}{2}(\underline{x}-\underline{\mu})^T \cdot \Sigma^{-1} \cdot (\underline{x}-\underline{\mu})}$$

$$-\infty < x_j < \infty, j = 1, \dots, n,$$

kde: $\underline{\mu} = (\mu_1, \dots, \mu_n)^T$ je vektor n reálných čísel

$$\Sigma = \begin{pmatrix} \sigma_{11} & \sigma_{12} & \dots & \sigma_{1n} \\ \vdots & \vdots & \vdots & \vdots \\ \sigma_{n1} & \sigma_{n2} & \dots & \sigma_{nn} \end{pmatrix} \text{ je kovarianční matice } (\sigma_{ij} = \text{cov}(X_i, X_j))$$

(symetrická pozitivně definitní matice typu (n;n))

$|\Sigma|$ je determinant kovarianční matice Σ , $|\Sigma| \neq 0$

Σ^{-1} je inverzní matice k matici Σ

(protože matice Σ je pozitivně definitní, je $|\Sigma| > 0$ a inverzní matice Σ^{-1} existuje)

Má-li náhodná veličina n-rozměrné normální rozdělení pravděpodobnosti s parametry μ a Σ , značíme to takto:

$$X \rightarrow N_n(\mu, \Sigma)$$

□ Dvourozměrné normální rozdělení

Speciálním případem n-rozměrného normálního rozdělení je dvourozměrné normální rozdělení. Kovarianční matice Σ má v tomto případě tvar:

$$\Sigma = \begin{pmatrix} \sigma_1^2 & \rho\sigma_1\sigma_2 \\ \rho\sigma_1\sigma_2 & \sigma_2^2 \end{pmatrix}$$

Všimněte si, že podmínka nenulového determinantu kovarianční matice ($|\Sigma| \neq 0$) je splněna pro $|\rho| < 1$.

Hustota pravděpodobnosti náhodného vektoru \underline{X} ($\underline{X}=(X_1, X_2)^T$) s dvourozměrným normálním rozdělením je dána vztahem:

$$f(x_1, x_2) = \frac{1}{2\pi\sigma_1\sigma_2\sqrt{1-\rho^2}} \cdot e^{-\frac{1}{2(1-\rho^2)} \left[\left(\frac{x_1-\mu_1}{\sigma_1} \right)^2 - 2\rho \left(\frac{x_1-\mu_1}{\sigma_1} \right) \left(\frac{x_2-\mu_2}{\sigma_2} \right) + \left(\frac{x_2-\mu_2}{\sigma_2} \right)^2 \right]}$$

Věta:

Nechť $f(x_1, x_2)$ je hustota náhodného vektoru $\mathbf{X}=(X_1, X_2)^T$ s dvourozměrným normálním rozdělením $N_2(\mu, \Sigma)$, kde $\mu = (\mu_1, \mu_2)$ je vektor středních hodnot a Σ je kovarianční matice náhodného vektoru \mathbf{X} . Pak náhodné veličiny X_1 a X_2 mají normální rozdělení $N(\mu_1, \sigma_1^2)$ a $N(\mu_2, \sigma_2^2)$. Hustoty f_{x_1}, f_{x_2} nezávisí na korelačním koeficientu ρ .

**Průvodce studiem**

Pro zájemce o hlubší pochopení studované látky uvádíme důkaz předcházející věty:

Hustota pravděpodobnosti náhodného vektoru \underline{X} ($\underline{X}=(X_1, X_2)^T$ s dvourozměrným normálním rozdělením je dána vztahem:

$$f(x_1, x_2) = \frac{1}{2\pi\sigma_1\sigma_2\sqrt{1-\rho^2}} \cdot e^{-\frac{1}{2(1-\rho^2)}\left[\left(\frac{x_1-\mu_1}{\sigma_1}\right)^2 - 2\rho\left(\frac{x_1-\mu_1}{\sigma_1}\right)\left(\frac{x_2-\mu_2}{\sigma_2}\right) + \left(\frac{x_2-\mu_2}{\sigma_2}\right)^2\right]}$$

Marginální hustoty nalezneme takto:

$$\begin{aligned} f_{x_1}(x_1) &= \int_{-\infty}^{\infty} f(x_1, x_2) dx_2 = \\ &= \int_{-\infty}^{\infty} \frac{1}{2\pi\sigma_1\sigma_2\sqrt{1-\rho^2}} \cdot e^{-\frac{1}{2(1-\rho^2)}\left[\left(\frac{x_1-\mu_1}{\sigma_1}\right)^2 - 2\rho\left(\frac{x_1-\mu_1}{\sigma_1}\right)\left(\frac{x_2-\mu_2}{\sigma_2}\right) + \left(\frac{x_2-\mu_2}{\sigma_2}\right)^2\right]} dx_2 = \\ &= \frac{1}{2\pi\sigma_1\sigma_2\sqrt{1-\rho^2}} \cdot \int_{-\infty}^{\infty} e^{-\frac{1}{2(1-\rho^2)}\left[\left(\frac{x_1-\mu_1}{\sigma_1}\right)^2 - 2\rho\left(\frac{x_1-\mu_1}{\sigma_1}\right)\left(\frac{x_2-\mu_2}{\sigma_2}\right) + \left(\frac{x_2-\mu_2}{\sigma_2}\right)^2\right]} dx_2 \end{aligned}$$

Zavedeme si substituci: $y = \left(\frac{x_2 - \mu_2}{\sigma_2}\right)$

Pak: $dy = \frac{1}{\sigma_2} dx_2$

$$\begin{aligned}
f_{X_1}(x_1) &= \frac{1}{2\pi\sigma_1\sigma_2\sqrt{1-\rho^2}} \cdot \int_{-\infty}^{\infty} \sigma_2 \cdot e^{-\frac{1}{2(1-\rho^2)}\left[\left(\frac{x_1-\mu_1}{\sigma_1}\right)^2 - 2\rho\left(\frac{x_1-\mu_1}{\sigma_1}\right)y + y^2\right]} dy = \\
&= \frac{1}{2\pi\sigma_1\sqrt{1-\rho^2}} \cdot \int_{-\infty}^{\infty} e^{-\frac{1}{2(1-\rho^2)}\left[\left(\frac{x_1-\mu_1}{\sigma_1}\right)^2 - 2\rho\left(\frac{x_1-\mu_1}{\sigma_1}\right)y + y^2\right]} dy = \\
&= \frac{1}{2\pi\sigma_1\sqrt{1-\rho^2}} \cdot \int_{-\infty}^{\infty} e^{-\frac{1}{2(1-\rho^2)}\left[\left(\frac{x_1-\mu_1}{\sigma_1}\right)^2 - 2\rho\left(\frac{x_1-\mu_1}{\sigma_1}\right)y + y^2 + \rho^2\left(\frac{x_1-\mu_1}{\sigma_1}\right)^2 - \rho^2\left(\frac{x_1-\mu_1}{\sigma_1}\right)^2\right]} dy = \\
&= \frac{1}{2\pi\sigma_1\sqrt{1-\rho^2}} \cdot \int_{-\infty}^{\infty} e^{-\frac{1}{2(1-\rho^2)}\left[\rho^2\left(\frac{x_1-\mu_1}{\sigma_1}\right)^2 - 2\rho\left(\frac{x_1-\mu_1}{\sigma_1}\right)y + y^2 + (1-\rho^2)\left(\frac{x_1-\mu_1}{\sigma_1}\right)^2\right]} dy = \\
&= \frac{1}{2\pi\sigma_1\sqrt{1-\rho^2}} \cdot \int_{-\infty}^{\infty} e^{-\frac{1}{2(1-\rho^2)}\left[\rho^2\left(\frac{x_1-\mu_1}{\sigma_1}\right)^2 - 2\rho\left(\frac{x_1-\mu_1}{\sigma_1}\right)y + y^2\right]} \cdot e^{-\frac{1}{2(1-\rho^2)}\left[(1-\rho^2)\left(\frac{x_1-\mu_1}{\sigma_1}\right)^2\right]} dy = \\
&= \frac{1}{2\pi\sigma_1\sqrt{1-\rho^2}} \cdot e^{-\frac{1}{2}\left(\frac{x_1-\mu_1}{\sigma_1}\right)^2} \int_{-\infty}^{\infty} e^{-\frac{1}{2(1-\rho^2)}\left[y - \rho\left(\frac{x_1-\mu_1}{\sigma_1}\right)\right]^2} \cdot dy =
\end{aligned}$$

Nyní zavedeme substituci: $z = y - \rho\left(\frac{x_1 - \mu_1}{\sigma_1}\right)$

Pak: $dz = dy$

$$\begin{aligned}
f_{X_1}(x_1) &= \frac{1}{2\pi\sigma_1\sqrt{1-\rho^2}} \cdot e^{-\frac{1}{2}\left(\frac{x_1-\mu_1}{\sigma_1}\right)^2} \int_{-\infty}^{\infty} e^{-\frac{1}{2(1-\rho^2)}\left[y - \rho\left(\frac{x_1-\mu_1}{\sigma_1}\right)\right]^2} \cdot dy = \\
&= \frac{1}{2\pi\sigma_1\sqrt{1-\rho^2}} \cdot e^{-\frac{1}{2}\left(\frac{x_1-\mu_1}{\sigma_1}\right)^2} \int_{-\infty}^{\infty} e^{-\frac{1}{2(1-\rho^2)}z^2} \cdot dz
\end{aligned}$$

A nakonec zavedeme ještě jednu substituci: $t = \frac{1}{\sqrt{1-\rho^2}}z$

Pak: $dt = \frac{1}{\sqrt{1-\rho^2}}dz$

$$\begin{aligned}
 f_{X_1}(x_1) &= \frac{1}{2\pi\sigma_1\sqrt{1-\rho^2}} \cdot e^{-\frac{1}{2}\left(\frac{x_1-\mu_1}{\sigma_1}\right)^2} \int_{-\infty}^{\infty} e^{-\frac{1}{2(1-\rho^2)}z^2} \cdot dz = \\
 &= \frac{1}{2\pi\sigma_1\sqrt{1-\rho^2}} \cdot e^{-\frac{1}{2}\left(\frac{x_1-\mu_1}{\sigma_1}\right)^2} \int_{-\infty}^{\infty} \sqrt{1-\rho^2} \cdot e^{-\frac{1}{2}t^2} \cdot dt = \frac{1}{2\pi\sigma_1} \cdot e^{-\frac{1}{2}\left(\frac{x_1-\mu_1}{\sigma_1}\right)^2} \int_{-\infty}^{\infty} e^{-\frac{1}{2}t^2} \cdot dt = \\
 &= \frac{1}{2\pi\sigma_1} \cdot e^{-\frac{1}{2}\left(\frac{x_1-\mu_1}{\sigma_1}\right)^2} \cdot \sqrt{2\pi} = \frac{1}{\sqrt{2\pi}\sigma_1} \cdot e^{-\frac{1}{2}\left(\frac{x_1-\mu_1}{\sigma_1}\right)^2}
 \end{aligned}$$

Vidíme, že jde o hustotu náhodné veličiny s normálním rozdělením $N(\mu_1, \sigma_1^2)$. Obdobně bychom ukázali, že marginální hustota f_{X_2} náhodné veličiny X_2 odpovídá normálnímu rozdělení $N(\mu_2, \sigma_2^2)$.

Je-li $\rho = 0$, pak:

$$\begin{aligned}
 f(x_1, x_2) &= \frac{1}{2\pi\sigma_1\sigma_2} \cdot e^{-\frac{1}{2}\left[\left(\frac{x_1-\mu_1}{\sigma_1}\right)^2 + \left(\frac{x_2-\mu_2}{\sigma_2}\right)^2\right]} = \frac{1}{2\pi\sigma_1\sigma_2} \cdot e^{-\frac{1}{2}\left(\frac{x_1-\mu_1}{\sigma_1}\right)^2} \cdot e^{-\frac{1}{2}\left(\frac{x_2-\mu_2}{\sigma_2}\right)^2} = \\
 &= \frac{1}{\sqrt{2\pi}\sigma_1} \cdot e^{-\frac{1}{2}\left(\frac{x_1-\mu_1}{\sigma_1}\right)^2} \cdot \frac{1}{\sqrt{2\pi}\sigma_2} \cdot e^{-\frac{1}{2}\left(\frac{x_2-\mu_2}{\sigma_2}\right)^2} = f_{X_1}(x_1) \cdot f_{X_2}(x_2)
 \end{aligned}$$

a náhodné veličiny X_1, X_2 jsou tedy nezávislé.



Řešený příklad

Nechť náhodný vektor $\underline{X}=(X_1, X_2)^T$ má dvourozměrné normální rozdělení s parametry: $\mu_1 = 2, \mu_2 = (-1), \sigma_1^2 = 4, \sigma_2^2 = 16, \rho = (-0,8)$. Stanovte pravděpodobnosti:

- $P(1 < X_1 < 4)$
- $P(3 < X_2 < 6)$

Již výše jsme si ukázali, že náhodné veličiny X_1 a X_2 mají normální rozdělení $N(\mu_1, \sigma_1^2)$ a $N(\mu_2, \sigma_2^2)$.

$$X_1 \rightarrow N(2;4) \quad X_2 \rightarrow N(-1;16)$$

ada)

$$\begin{aligned}
 \underline{P(1 < X_1 < 4)} &= F(4) - F(1) = \Phi\left(\frac{4-2}{\sqrt{4}}\right) - \Phi\left(\frac{1-2}{\sqrt{4}}\right) = \Phi(1) - \Phi(-0,5) = \Phi(1) - (1 - \Phi(0,5)) \\
 &= 0,841 - (1 - 0,691) = \underline{\underline{0,532}}
 \end{aligned}$$

adb)

$$\begin{aligned} \underline{\underline{P(3 < X_2 < 6)}} &= F(6) - F(3) = \Phi\left(\frac{6 - (-1)}{\sqrt{16}}\right) - \Phi\left(\frac{3 - (-1)}{\sqrt{16}}\right) = \Phi(1,75) - \Phi(1) = 0,960 - 0,8 \\ &= \underline{\underline{0,119}} \end{aligned}$$

$$DX = e^{2\mu + \sigma^2} (e^{\sigma^2} - 1) \Rightarrow \underline{\underline{DX}} = e^{2 \cdot 2 + 9} (e^9 - 1) \cong \underline{\underline{3,6 \cdot 10^9}}$$



Otázky 2.

1. Popište náhodnou veličinu mající Erlangovo rozdělení
2. Popište náhodnou veličinu mající Weibulovo rozdělení
3. V čem spočívá flexibilita Weibullova rozdělení? (užití pro různá období intenzity poruch)
4. Popište náhodnou veličinu mající Logaritnicko – normální rozdělení
5. Popište náhodný vektor mající Vícerozměrné normální rozdělení

3. FUNKCE NÁHODNÉ VELIČINY



Čas ke studiu: 0,75 hodiny



Cíl: Po prostudování této kapitoly budete umět:

- transformovat náhodnou veličinu X na náhodnou veličinu Y , je-li mezi těmito náhodnými veličinami vzájemně jednoznačný vztah



Výklad

3.1. Funkce náhodné veličiny

V mnoha případech, kdy známe rozdělení náhodné veličiny X , potřebujeme určit rozdělení náhodné veličiny Y , která je funkcí X , tzn. $Y = h(X)$.

Je-li funkce $h(x)$ v oboru možných hodnot veličiny X **monotónní**, pak existuje inverzní funkce $h^{-1}(y)$, a jde o **vzájemně jednoznačný vztah mezi X a Y** .

Je-li v takovém případě $h(x)$ **rostoucí**, pak pro všechna $x_2 > x_1$ je $y_2 > y_1$, a distribuční funkci veličiny Y lze psát jako:

$$G(y) = P(Y < y) = P[X < h^{-1}(y)] = F[h^{-1}(y)]$$

Pro **klesající funkci $h(x)$** , tzn. pro všechna $x_2 > x_1$ platí $y_1 > y_2$, je distribuční funkce:

$$G(y) = P(Y < y) = P[X > h^{-1}(y)] = 1 - F[h^{-1}(y)]$$

Pro diskrétní náhodnou veličinu X je pravděpodobnostní funkce dána jako:

$$p_Y(y_i) = p_X(h^{-1}(y_i))$$

Je-li X spojitá náhodná veličina s hustotou pravděpodobnosti $f(x)$, přičemž $h^{-1}(y)$ má pro všechna y spojitou derivaci, pak pro rostoucí funkci $h(x)$ dostaneme hustotu pravděpodobnosti $g(y)$ veličiny Y jako:

$$g(y) = \frac{dG(y)}{dy} = f(h^{-1}(y)) \cdot \frac{dh^{-1}}{dy} = f(h^{-1}(y)) \cdot \frac{dx}{dy}$$

Podobně pro klesající funkci $h(x)$ dostaneme:

$$g(y) = \frac{dG(y)}{dy} = -f(h^{-1}(y)) \cdot \frac{dh^{-1}}{dy} = -f(h^{-1}(y)) \cdot \frac{dx}{dy}$$

Vzhledem k tomu, že v případě rostoucí funkce $h(x)$ je $\left(\frac{dx}{dy} > 0\right)$, zatímco v případě klesající funkce je $\left(\frac{dx}{dy} < 0\right)$, lze oba předchozí vztahy spojit do jednoho:

$$g(y) = \frac{dG(y)}{dy} = f(h^{-1}(y)) \cdot \left| \frac{dh^{-1}}{dy} \right| = f(h^{-1}(y)) \cdot \left| \frac{dx}{dy} \right|$$

Není-li $h(x)$ monotónní funkcí, pak mezi X a Y neexistuje vzájemně jednoznačný vztah a tedy ani inverzní funkce k $h(x)$. Distribuční funkce $G(y) = P(Y < y)$ je v takovém případě dána pravděpodobností, že náhodná veličina nabude hodnoty z kteréhokoliv intervalu, pro který $Y < y$.

Pak platí:

Pro diskrétní náhodnou veličinu X :
$$G(y) = \sum_{i: h(x_i) \leq y} p_i$$

Pro spojitou náhodnou veličinu X :
$$G(y) = \int_{h(x) \leq y} f(x) dx$$

Pro případ diskrétní náhodné veličiny X je pravděpodobnostní funkce p_Y veličiny Y dána vztahem:

$$p_Y(y) = \sum_{i: h(x_i) = y} p_X(x_i)$$

Nechť existuje konečný počet x_i takových, že $h(x_i) = y$. Nechť pro každé x_i existuje derivace $\frac{dh}{dx} \neq 0$. Pak existuje hustota pravděpodobnosti $g(y)$ náhodné veličiny Y :

$$g(y) = \sum_{i: h(x_i) = y} f(x_i) \left| \frac{dh}{dx} \right|_{x=x_i}^{-1}$$



Řešený příklad

Nechť veličina X má rovnoměrné rozdělení v intervalu $\left\langle -\frac{\pi}{2}; \frac{\pi}{2} \right\rangle$. Jaké rozdělení má veličina $y = \operatorname{tg} x$?

$$g(y) = f(h^{-1}(y)) \cdot \left| \frac{dx}{dy} \right|$$

Hustota pravděpodobnosti rovnoměrného rozdělení: $f(x) = \frac{1}{\frac{\pi}{2} - \left(-\frac{\pi}{2}\right)} = \frac{1}{\pi}$

$$h(x) = y = \operatorname{tg} x \Rightarrow h^{-1}(y) = x = \operatorname{arctg} y$$

$$\left| \frac{dx}{dy} \right| = \left| \frac{d(\operatorname{arctg} y)}{dy} \right| = \left| \frac{1}{1+y^2} \right| = \frac{1}{1+y^2}$$

Hustota pravděpodobnosti veličiny Y je tedy:

$$g(y) = f(h^{-1}(y)) \cdot \left| \frac{dx}{dy} \right| = \frac{1}{\pi(1+y^2)}, \quad y \in \mathbb{R}$$

Uvedené **rozdělení** se nazývá **Cauchyho**. Je příkladem rozdělení, které nemá konečný rozptyl:

$$\begin{aligned} DY &= \int_{-\infty}^{\infty} y^2 \cdot g(y) dy = \int_{-\infty}^{\infty} y^2 \cdot \frac{1}{\pi(1+y^2)} dy = \frac{1}{\pi} \int_{-\infty}^{\infty} \frac{y^2+1-1}{(1+y^2)} dy = \\ &= \frac{1}{\pi} \left[\int_{-\infty}^{\infty} 1 dy - \int_{-\infty}^{\infty} \frac{1}{(1+y^2)} dy \right] = \frac{1}{\pi} [\infty - \pi] = \infty \end{aligned}$$



Řešený příklad

Nechť veličina X má normální rozdělení $N(0;1)$. Jaké rozdělení má veličina $y = x^2$?

Pro nezáporná y existuje inverzní funkce $h^{-1}(y) : x = \pm\sqrt{y}$.

$$x = \pm\sqrt{y} \Rightarrow \left| \frac{dx}{dy} \right| = \frac{1}{2\sqrt{y}}$$

Pak hustota pravděpodobnosti nezáporné náhodné veličiny Y je:

$$y \geq 0 :$$

$$\begin{aligned} g(y) &= f(\pm\sqrt{y}) \cdot \left| \frac{dx}{dy} \right| = (f(\sqrt{y}) + f(-\sqrt{y})) \cdot \left| \frac{dx}{dy} \right| = \left(\frac{1}{\sqrt{2\pi}} \cdot e^{-\frac{y}{2}} + \frac{1}{\sqrt{2\pi}} \cdot e^{-\frac{y}{2}} \right) \cdot \frac{1}{2\sqrt{y}} = \\ &= \frac{1}{\sqrt{2\pi y}} \cdot e^{-\frac{y}{2}} \end{aligned}$$

Jde o hustotu rozdělení χ^2 s jedním stupněm volnosti.

3.2. Přibližné stanovení charakteristik funkce náhodné veličiny

V praxi je někdy k dispozici pouze jediná změřená hodnota veličiny X (odhad její střední hodnoty) a směrodatná odchylka měření σ_x (daná například udanou chybou měřícího přístroje). Pokud je variační koeficient mnohem menší než jedna $\left(\frac{\sigma_x}{\mu} \ll 1\right)$, lze přibližně odhadnout charakteristiky veličiny $y = h(x)$.

Předpokládejme, že náhodná veličina X je spojitá.

Střední hodnotu náhodné veličiny Y odhadneme na základě vztahu:

$$EY = \int h(x)f(x)dx = \int \left[h(EX) + h'(EX) \cdot (x - EX) + \frac{h''(EX)}{2} \cdot (x - EX)^2 + \dots \right] f(x)dx \cong \\ \cong h(EX) + \frac{h''(EX)}{2} \cdot DX \cong h(EX)$$

Rozptyl DY lze pak vyjádřit přibližně z lineárního členu Taylorova rozvoje:

$$DY = \int (h(x) - EY)^2 f(x)dx \cong \int (h(x) - h(EX))^2 f(x)dx \cong \left(\frac{dh}{dx}\right)_{x=EX}^2 \cdot DX$$



Otázky 3.

1. Necht' $Y=h(X)$. $h(x)$ je monotónní funkce. Nalezněte vztah mezi hustotou pravděpodobnosti náhodné veličiny Y a hustotou pravděpodobnosti náhodné veličiny X .



Úlohy k řešení 3.

1. F je distribuční funkce náhodné veličiny X , je spojitá a rostoucí. Náhodná veličina Y je definována vztahem: $Y = F(X)$. Určete rozdělení náhodné veličiny Y (hustotu pravděpodobnosti).
2. Náhodná veličina X má rovnoměrné rozdělení na intervalu $\langle 0;3 \rangle$. Určete rozdělení náhodné veličiny Y , $Y=2X+1$.
3. Náhodná veličina X má normální rozdělení $N(\mu; \sigma^2)$. Určete rozdělení náhodné veličiny Y , $Y = e^X$.
4. Náhodná veličina X má hustotu pravděpodobnosti: $f(x) = \lambda \cdot e^{-\lambda x}$. Určete rozdělení náhodné veličiny Y , $Y = -\ln X$.

4. ZÁKLADY TEORIE SPOLEHLIVOSTI

4.1. Teorie spolehlivosti



Čas ke studiu: 10 minut



Cíl: Po prostudování tohoto odstavce budete umět:

- popsat charakteristické rysy teorie spolehlivosti
- technické a matematické aspekty teorie spolehlivosti



Výklad

□ Co zkoumá teorie spolehlivosti ?

Teorie spolehlivosti se zabývá technickými a matematickými otázkami spolehlivosti. Technická problematika souvisí s konstrukcí, použitými materiály, technologií a organizací výroby, diagnostikou a strategií údržby.

Matematická teorie spolehlivosti se soustředí na prognózu, odhad a optimalizaci bezporuchového provozu výrobků. (Výrobkem rozumíme prvek, systém nebo jeho část.)

Hlavními nástroji jsou zde teorie pravděpodobnosti a matematická statistika. Typicky matematickou záležitostí je např. stanovení charakteristik spolehlivosti jako jsou zaručená doba života, střední doba bezporuchového provozu, střední doba mezi poruchami, průměrné náklady na údržbu a opravy aj.

Matematická statistika a teorie pravděpodobnosti nám umožňují popis jevů, jejichž podstatu dokonale neznáme, ale jejichž zákonitosti vzniku jsou pro stanovení spolehlivosti velmi důležité. Jsou to např. fyzikální zákonitosti a mechanismy poruch, procesy stárnutí, koroze, opotřebení a únavy materiálů, vzájemná souvislost různých poruch, vliv prostředí apod. Protože analýza těchto jevů z hlediska čistě fyzikálního nebo chemického je příliš složitá, nezbyvá než zjišťovat poruchovost větších celků nebo většího počtu výrobků v delším čase statisticky. To však většinou vyžaduje sběr, přenos a zpracování informací přímo z provozu, jako např. soustavné a pečlivé vedení záznamů o všech poruchách a jejich příčinách, době provozu, době oprav, podmínkách činnosti a jiných vlivech u zařízení, která jsou často rozptýlena na různých místech a pracují v různých podmínkách.

Spolehlivost jakožto obecnou vlastnost výrobku splňovat po určitou dobu a za určitých podmínek danou funkci, je nutno posuzovat též podle ekonomického hlediska. Aplikací výsledků teorie spolehlivosti lze též použít nejen při návrhu zařízení a jeho způsobu provozu na zadané úrovni spolehlivosti, která vyplývá z výše zmíněných ekonomických kritérií, ale též při vzájemném porovnávání různých alternativ řešení, dále pro kvantitativní předpovědi chování složitých zařízení v dalším provozu a k sestavení optimální strategie údržby těchto zařízení.

Příklad 4.1.1

Moderní výrobky (systémy) sestavené z mnoha prvků jsou vysoce spolehlivé, např. počítač. Jestliže chceme tuto spolehlivost dále zvyšovat, pak nelze jít pouze cestou zvyšování spolehlivosti prvků. Jestliže systém např. sestává ze 100 000 prvků, které pracují nezávisle na sobě a každý z nich se

s pravděpodobností 0.99999 po sledovanou dobu neporouhá, potom pravděpodobnost, že se systém po sledovanou dobu neporouhá (tj. bezporuchovost), je $(0.99999)^{100000} = 0.368$. Je proto nezbytné hledat jiné způsoby pro zvyšování bezporuchovosti – např. zálohování důležitých částí, aplikace údržby atd.



Shrnutí kapitoly 4.1.

Spolehlivost lze charakterizovat jako obecnou vlastnost výrobku splňovat po určitou dobu a za určitých podmínek danou funkcí.

Teorie spolehlivosti je vědní disciplína zodpovídající technické a matematické otázky spolehlivosti.

Hlavní nástroje pro zodpovězení matematických otázek teorie spolehlivosti jsou **teorie pravděpodobnosti a matematická statistika**.

Organizace výrobního procesu či technologie výroby (např. použití vhodných materiálů) souvisí s **technickými otázkami spolehlivosti**.



Otázky 4.1.

1. Co je to spolehlivost?
2. Čím se zabývá teorie spolehlivosti?
3. Jaké jsou nástroje teorie spolehlivosti?

4.2. Základní pojmy



Čas ke studiu: 20 minut



Cíl: Po prostudování tohoto odstavce budete umět

- definovat základní pojmy teorie spolehlivosti z hlediska technického
- definovat: bezporuchovost, životnost, opravitelnost, pohotovost, ...
- charakterizovat poruchy a klasifikovat je



Výklad

Nejprve vyložíme základní pojmy teorie spolehlivosti z hlediska technického, což nám poslouží jako motivace pro zavedení příslušných pojmů matematických. Pojem spolehlivosti je obvykle spojován s pojmem výrobku (neboli objektu). Výrobek od okamžiku, kdy je vyroben, má svou historii: doprava, skladování, příprava na využití, vlastní využití, údržba, oprava a vyřazení. V některých fázích historie výrobku budeme požadovat, aby byl **spolehlivý**.

□ Spolehlivost jako obecná vlastnost

Spolehlivostí rozumíme obecnou vlastnost spočívající ve schopnosti výrobku plnit po stanovenou dobu požadované funkce při zachování provozních parametrů daných technickými podmínkami.

Je charakterizována dalšími dílčími vlastnostmi, jako jsou: bezporuchovost, životnost, opravitelnost, udržovatelnost, skladovatelnost, bezpečnost a další.

□ Jaké jsou dílčí vlastnosti spolehlivosti ?

Technickými podmínkami přitom rozumíme souhrn specifikací technických a provozních vlastností výrobku spolu se způsoby jeho provozu, údržby a oprav. Jinými slovy je spolehlivost způsobilost výrobku uchovat svou kvalitu v daných podmínkách využívání.

Bezporuchovost je způsobilost výrobku plnit bez poruchy požadované funkce po stanovenou dobu a za stanovených podmínek.

Životnost je způsobilost výrobku plnit požadované funkce do mezního stavu stanoveného technickými podmínkami. Na konci období životnosti se u výrobků projeví takové rysy spojené s opotřebením a stárnutím, že jejich odstranění je neekonomické, nebo nemožné. Někdy může jít i o tzv. „morální opotřebení“.

Opotřebením znamená ve spolehlivosti postupné změny znaků výrobků, které jsou vyvolány zatížením způsobeným pouze provozními podmínkami.

Stárnutí znamená změny vzniklé zatížením mimo provoz.

Opravitelnost je vlastnost výrobku spočívající v možnosti odhalení poruchy, zjištění její příčiny a odstranění opravou.

Udržovatelnost je vlastnost výrobku spočívající ve způsobilosti k předcházení poruch předepsanou údržbou.

Skladovatelnost je schopnost výrobku zachovávat nepřetržitě bezvadný (a tedy provozuschopný) stav po dobu skladování a přepravy při dodržení předepsaných podmínek.

Bezpečnost je vlastnost výrobku neohrožovat lidské zdraví nebo životní prostředí při plnění předepsané funkce po stanovenou dobu a za stanovených podmínek.

Z provozního hlediska je důležitá **pohotovost** výrobku, tj. schopnost výrobku v určitém okamžiku vyhovovat technickým podmínkám. Pohotovost (neboli též provozuschopnost) je komplexní vlastnost objektu zahrnující bezporuchovost a opravitelnost objektu v podmínkách provozu.

□ Co je porucha a jak poruchy klasifikujeme ?

Důležitým a zdánlivě jednoduchým pojmem teorie spolehlivosti je pojem porucha. **Porucha** je částečná nebo úplná ztráta, případně změna vlastností výrobku, která podstatným způsobem snižuje schopnost nebo způsobuje nemožnost výrobku plnit požadovanou funkci. Pojem porucha je v mnoha případech relativní. V praxi je proto zapotřebí pojem porucha přesně vymežit.

Zhoršení schopnosti provozu, které ještě nezpůsobí poruchu, se označuje jako **závada**.

□ Klasifikace poruch

1. Podle podmínek vzniku se poruchy dělí na poruchy z vnějších a vnitřních příčin.

Porucha z vnějších příčin je porucha způsobená nedodržením stanovených provozních podmínek a předpisů pro zatěžování, obsluhu a údržbu.

Porucha z vnitřních příčin je porucha způsobená vlastní nedokonalostí výrobku při zachování stanovených provozních podmínek a předpisů. Mezi poruchy z vnitřních příčin patří především **časné poruchy** projevující se v počátečním období provozu. Jejich výskyt s rostoucím časem klesá. Příčinou časných poruch jsou nedostatky při návrhu a výrobě. Dále sem patří **poruchy dožitím** vznikající následkem opotřebení nebo stárnutí (viz dále).

2. Podle časového průběhu se poruchy dělí na náhlé a postupné.

Náhlá porucha je porucha projevující se prudkou změnou jednoho nebo více parametrů výrobku.

Postupná porucha je porucha projevující se jako postupná změna parametrů výrobku, např. v důsledku stárnutí nebo opotřebení.

Zatímco poruchy náhlé se obvykle předvídat nedají, je předvídaní postupných poruch častou úlohou teorie spolehlivosti.

3. V některých situacích je účelné dále klasifikovat poruchy na částečné a úplné.

Částečná porucha znamená odchýlení jednoho nebo více parametrů od úrovně stanovené technickými podmínkami, které však úplně nebrání výrobku plnit požadovanou funkci.

Úplná porucha je porucha, která zcela zabraňuje výrobku plnit požadovanou funkci.

Částečná či postupná porucha se nazývá též **degradační porucha**, náhlá a úplná porucha se nazývá **havarijní porucha**.

4. Podle souvislosti s jinými poruchami se poruchy dělí na nezávislé a závislé.

Závislá porucha vzniká následkem poruchy jiného prvku, nezávislá nikoli.

5. Podle doby trvání se rozlišují poruchy trvalé a poruchy dočasné.

Trvalou poruchu je možno odstranit pouze opravou nebo náhradou porouchaného prvku.

Dočasné poruchy mohou samovolně vymizet nebo trvají jen po dobu působení vnějšího vlivu.

Dělení poruch do tříd je často relativní. Náhlé poruše obvykle předcházejí skryté změny vlastností prvku, které by bylo možno dosti podrobným zkoumáním zjistit a poruchu označit jako postupnou. Dokonalá znalost všech fyzikálně chemických dějů probíhajících v materiálech prvku, přesná znalost postupu výroby a podmínek provozu by dovolila předpovědět dobu vzniku poruchy prvku. V takovém případě by se porucha označila jako nenáhodná. Omezená znalost těchto činitelů je důvodem pro označení poruchy prvku jako náhodné.

□ **Které dílčí vlastnosti spolehlivosti budeme kvantitativně určovat ?**

Všechny výše uvedené dílčí spolehlivostní vlastnosti lze charakterizovat též kvantitativně pomocí vhodně zvolených spolehlivostních ukazatelů nebo charakteristik. V dalším se budeme zabývat pouze kvantitativním vyjádřením bezporuchovosti a pohotovosti.

Bezporuchovost určujeme především u neobnovovaných (tj. neopravitelných) objektů a nebo tam, kde se zajímáme o činnost do první poruchy (Obecně se ovšem tento pojem zavádí i pro opravitelné objekty).

Pohotovost (provozní schopnost) určujeme u obnovovaných objektů. Obnovované objekty se po vzniku poruchy opraví a provoz pokračuje. Oprava se považuje za účelnou tehdy, když průměrná cena opravy a náhradních součástí je malá vůči pořizovací ceně zařízení. Provoz obnovovaného systému nebo obnovovaného prvku lze popsat jako posloupnost stavů bezporuchového provozu a oprav, přičemž okamžiky poruch a oprav jsou náhodné.



Shrnutí kapitoly 4.2.

Spolehlivost je obecná vlastnost projevující se prostřednictvím dílčích vlastností: bezporuchovost, životnost, opravitelnost, udržovatelnost, skladovatelnost, bezpečnost.

Pohotovost je komplexní vlastnost výrobku zahrnující bezporuchovost a opravitelnost v podmínkách provozu.

Porucha je částečná nebo úplná ztráta, případně změna vlastností výrobku, která podstatným způsobem snižuje schopnost nebo způsobuje nemožnost výrobku plnit požadovanou funkci. Poruchy dělíme podle různých hledisek, nejčastěji podle podmínek vzniku na poruchy z vnějších a vnitřních příčin.



Otázky 4.2.

1. V čem se liší pojmy „bezporuchovost“ a „pohotovost“ ?
2. U jakých objektů má smysl tyto pojmy kvantitativně určovat ?
3. Co je porucha a jak lze poruchy klasifikovat ?

4.3. Doba do poruchy



Čas ke studiu: 40 minut



Cíl: Po prostudování tohoto odstavce budete umět

- popsat dobu do poruchy pomocí distribuční funkce a funkce bezporuchovosti
- charakterizovat dobu do poruchy pomocí hazardní funkce (intenzity poruch)
- vyjádřit vztahy mezi jednotlivými popisnými funkcemi doby do poruchy
- charakterizovat dobu do poruchy pomocí základních číselných charakteristik



Výklad

□ Co je doba do poruchy a jak ji matematicky popsat ?

Neporouchaný výrobek (prvek, systém, část systému) začne pracovat v okamžiku $t = 0$ za určitých podmínek, o nichž budeme zatím předpokládat, že se v průběhu času nemění. V okamžiku $t = X$ se výrobek porouchá. Doba X po kterou výrobek pracoval bez poruchy, se nazývá **doba do poruchy**.

V dalším budeme předpokládat, že doba do poruchy X je nezáporná náhodná veličina s **distribuční funkcí**

$$F(t) = P(X < t)$$

Distribuční funkce doby do poruchy vyjadřuje pravděpodobnost toho, že na intervalu $(0, t)$ dojde k poruše. S distribuční funkcí doba do poruchy je úzce spojena funkcí:

$$R(t) = P(X \geq t) ,$$

kteřá se nazývá **funkcí bezporuchovosti** (pravděpodobnost bezporuchového provozu), resp. **funkcí spolehlivosti** (zkráceně spolehlivost). Tato funkce vyjadřuje pravděpodobnost toho, že na intervalu $(0, t)$ nedojde k poruše. $R(t)$ je nerostoucí funkce času, $F(t)$ je neklesající funkce času. Obě veličiny jsou nezáporná bezrozměrná čísla nejvýše rovna jedné. Zpravidla předpokládáme, že $R(0) = 1$, a $R(\infty) = 0$.

Ve spolehlivosti se poměrně často setkáváme s pojmem zaručená **doba bezporuchového provozu (100% - ní život) T_γ** . Četnostní interpretace je taková, že přibližně 100% výrobků bude bez poruchy fungovat alespoň do okamžiku T_γ .

$$P(X \geq T_\gamma) = \gamma \Rightarrow 1 - F(T_\gamma) = \gamma \Rightarrow F(T_\gamma) = 1 - \gamma \Rightarrow T_\gamma = x_{1-\gamma}$$

Je-li distribuční funkce $F(t)$ spojitá, nazývá se odpovídající hustota pravděpodobnosti $f(t)$:

$$f(t) = \frac{dF(t)}{dt} = -\frac{dR(t)}{dt}$$

též **hustota poruch**.

□ Hazardní funkce a její alternativní vyjádření

Nejčastěji se bezporuchovost neopravovaného výrobku udává **hazardní funkcí** (někdy označovanou jako **intenzita poruch**), definovanou jako poměr hustoty pravděpodobnosti poruchy a funkce bezporuchovosti:

$$\lambda(t) = \frac{f(t)}{R(t)} \quad R(t) > 0$$

Velichiny $f(t)$ a $\lambda(t)$ mají rozměr [1/čas], obvykle se udávají v jednotkách [1/hod] nebo [1/rok]. Každá ze 4 základních veličin $R(t)$, $F(t)$, $f(t)$, $\lambda(t)$ popisuje úplně stejně bezporuchovost neopravovaného objektu a z každé z nich je možno odvodit tři zbývající. Vzájemné převody udává následující tabulka.

	$R(t)$	$F(t)$	$f(t)$	$\lambda(t)$
$R(t)$	$R(t)$	$1 - F(t)$	$1 - \int_0^t f(x)dx$	$\exp\left[-\int_0^t \lambda(x)dx\right]$
$F(t)$	$1 - R(t)$	$F(t)$	$\int_0^t f(x)dx$	$1 - \exp\left[-\int_0^t \lambda(x)dx\right]$
$f(t)$	$-\frac{dR(t)}{dt}$	$\frac{dF(t)}{dt}$	$f(t)$	$\lambda(t) \cdot \exp\left[-\int_0^t \lambda(x)dx\right]$
$\lambda(t)$	$-\frac{d(\ln R(t))}{dt} = -\frac{\frac{dR(t)}{dt}}{R(t)}$	$\frac{\frac{dF(t)}{dt}}{1 - F(t)}$	$\frac{f(t)}{1 - \int_0^t f(x)dx}$	$\lambda(t)$

Tabulka: Matematické převodní vztahy mezi základními funkčními ukazateli bezporuchovosti

Důležitou úlohu při rozdělení doby do poruchy hrají číselné charakteristiky tohoto rozdělení, zejména vybrané momenty a kvantily (střední doba do poruchy, rozptyl doby do poruchy, koeficienty šikmosti a špičatosti, γ -procentní život neboli zaručená doba bezporuchového provozu atd.). Uvedeme zde několik z nich.

□ Jak kvantitativně určit základní číselné charakteristiky doby do poruchy ?

Střední doba provozu do poruchy, která je pro neobnovované objekty rovna **střední době do poruchy** (ustálená mezinárodní zkratka pochází z angličtiny MTTF = Mean Time To Failure), se definuje jako střední (očekávaná) hodnota náhodné veličiny, tj. doby do poruchy X

$$EX = \int_0^{\infty} t f(t) dt$$

Hodnota EX je integrální hodnota, která vyjadřuje bezporuchovost jediným údajem. Obvykle se udává v [hod].

Vlastnost: Necht' nezáporná náhodná veličina X má funkci bezporuchovosti $R(x)$ a necht' $EX^k < +\infty$, kde k je přirozené číslo (tedy necht' existují konečné obecné momenty všech řádů). Potom :

$$EX^k = k \int_0^{\infty} x^{k-1} R(x) dx$$

Důkaz lze provést užitím metody „per partes“.

Pro střední dobu do poruchy dostáváme užitím vztahu pro k -tý obecný moment (pro $k = 1$) důležitý vztah:

$$EX = \int_0^{\infty} R(x) dx$$

Pro **rozptyl** doby do poruchy platí

$$DX = EX^2 - (EX)^2 = 2 \int_0^{\infty} x R(x) dx - (EX)^2,$$

což dostaneme opět užitím vztahu pro k -tý obecný moment (pro $k = 2$).

Gama-procentní život T_γ je definován jako $100 \cdot (1 - \gamma)$ procentní kvantil rozdělení doby do poruchy.

$$F(T_\gamma) = 1 - \gamma \quad \text{neboli} \quad R(T_\gamma) = \gamma$$

Četnostní interpretace je taková, že přibližně 100γ % výrobků bude bez poruchy fungovat do okamžiku T_γ .



Shrnutí kapitoly 4.3.

Distribuční funkce doby do poruchy vyjadřuje pravděpodobnost toho, že na intervalu $(0, t)$ dojde k poruše. Doplněk distribuční funkce do jedničky se nazývá **funkcí bezporuchovosti**, která vyjadřuje pravděpodobnost toho, že na intervalu $(0, t)$ nedojde k poruše.

Hazardní funkce (intenzita poruch) je poměr hustoty pravděpodobnosti poruchy a funkce bezporuchovosti.

Střední dobu provozu do poruchy (MTTF) lze určit integrací z funkce bezporuchovosti přes interval $(0, +\infty)$.

Gama-procentní život T_γ určuje přibližně dobu, po kterou bude bez poruchy fungovat 100γ % výrobků.

Rozptyl doby do poruchy lze určit rovněž ze znalosti funkce bezporuchovosti.



Průvodce studiem

Poznámky k obnovovaným (opravitelným) výrobkům:

1. Pro obnovované výrobky je nezbytné vyšetřovat kromě doby do poruchy ještě další náhodnou veličinu: dobu opravy (nebo dobu do ukončení opravy), přičemž touto veličinou budeme v dalším rozumět celkovou dobu údržby po poruše až po obnovu výrobku. Jako každá náhodná veličina, i doba opravy je charakterizována základními popisnými funkcemi, jako jsou hustota pravděpodobnosti (hustota oprav) a distribuční funkce. Zcela analogicky jako intenzita poruch se také zavádí **intenzita oprav** a nejčastěji používanou číselnou charakteristikou této náhodné veličiny je její střední (očekávaná) hodnota, která v teorii spolehlivosti nese označení jako **střední doba do obnovy**, zkratka **MTTR** (z anglického Mean Time To Repair).
2. Používané ukazatele spolehlivosti pro obnovované výrobky jsou dále: **Funkce okamžité pohotovosti $A(t)$** , což je pravděpodobnost, že výrobek je ve stavu schopném plnit v daných podmínkách a v daném časovém okamžiku požadovanou funkci, za předpokladu, že požadované vnější prostředky jsou zajištěny. Dále je to **součinitel asymptotické pohotovosti A** , což je limita okamžité pohotovosti, pro účely modelování, existuje-li, pro čas jdoucí k nekonečnu. V případě potřeby se určuje i **součinitel střední pohotovosti $\bar{A}(t_1, t_2)$** , což je střední hodnota funkce okamžité

pohotovosti v daném časovém intervalu (t_1, t_2) :
$$\bar{A}(t_1, t_2) = \frac{1}{t_2 - t_1} \int_{t_1}^{t_2} A(t) dt .$$



Otázky 4.3.

1. Jaké jsou možnosti pro jednoznačný a úplný popis náhodné veličiny: doba do poruchy nějakého výrobku ?
2. Které jsou v praxi nejpoužívanější číselné charakteristiky této náhodné veličiny ?
3. Jak je definována hazardní funkce (intenzita poruch), popřípadě odvod'te, jak souvisí s ostatními popisnými funkcemi náhodné veličiny: doba do poruchy ?

**Úlohy k řešení 4.3.**

- Ventil vodovodního potrubí má zadánu funkci bezporuchovosti: $R(t) = e^{-0,001 \cdot t}$. Určete střední dobu do poruchy ventilu MTTF a dále určete rozptyl doby do poruchy ventilu DX . Dále určete 80%-tní život ventilu $T_{0,80}$
- Určete 90%-tní život $T_{0,90}$ pro výrobek, jehož doba do poruchy se řídí Weibullovým rozdělením, s lineárně rostoucí intenzitou poruch ($\beta = 2$) a s parametrem $\lambda = 10$ ($F(t) = 1 - e^{-(\lambda t)^\beta}$).
- Doba do vybití baterie se řídí exponenciálním rozdělením ($F(t) = 1 - e^{-\frac{t}{MTTF}}$).
 - Jaká je střední doba do vybití MTTF, víme-li, že 4000 hodin přežije 1% těchto baterií?
 - Je-li střední doba do vybití 3.150 hodin, kolik procent těchto baterií přežije 4000 hodin?

4.4. Intenzita poruch**Čas ke studiu:** 25 minut**Cíl:** Po prostudování tohoto odstavce budete umět

- vysvětlit intenzitu poruch pomocí pravděpodobnosti
- demonstrovat intenzitu poruch graficky
- vysvětlit jednotlivé fáze života výrobku

**Výklad****□ Jaká je pravděpodobnostní interpretace intenzity poruch ?**

Nechť $t > 0$, $\Delta t > 0$, $\Delta t \rightarrow 0$ a počítejme podmíněnou pravděpodobnost jevu, že se prvek porouchá (doba do poruchy je X) v časovém intervalu $(t, t + \Delta t)$ za podmínky, že pracoval bez poruchy do okamžiku t . Pro tuto podmíněnou pravděpodobnost dostaneme:

$$\begin{aligned}
 P(t < X < t + \Delta t | X \geq t) &= \frac{P(t < X < t + \Delta t, X \geq t)}{P(X \geq t)} = \frac{P(t < X < t + \Delta t)}{P(X \geq t)} = \\
 &= \frac{1}{1 - F(t)} \cdot \frac{F(t + \Delta t) - F(t)}{\Delta t} \cdot \Delta t
 \end{aligned}$$

pro $\Delta t \rightarrow 0$ dostáváme $\frac{F(t + \Delta t) - F(t)}{\Delta t} \rightarrow \frac{dF}{dt} = f(t)$,

takže:

$$P(t < X < t + \Delta t | X \geq t) \approx \frac{f(t)}{1 - F(t)} \Delta t = \lambda(t) \cdot \Delta t$$

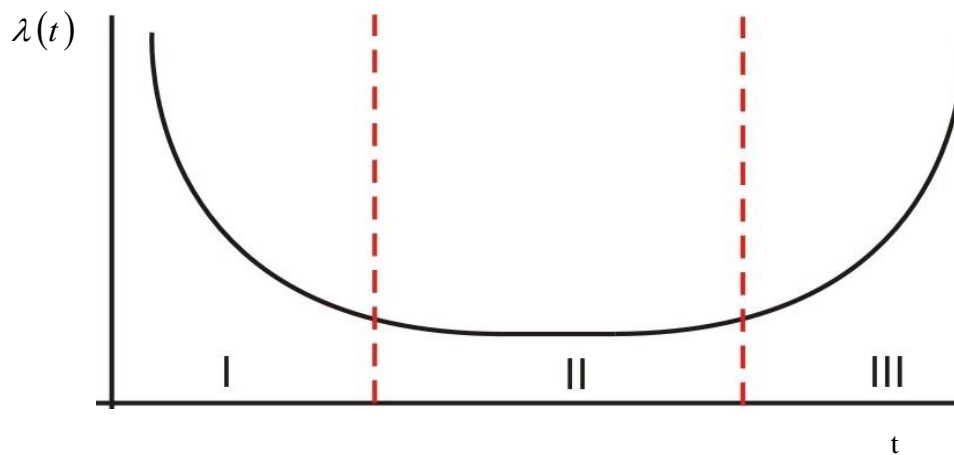
Intenzita poruch je tedy lokální charakteristikou spolehlivosti. Vyjadřuje přibližně pravděpodobnost toho, že prvek, který se neporouchal do okamžiku t , se porouchá v intervalu $(t, t + 1)$.

□ Jak vypadá nejčastější grafická interpretace intenzity poruch ?

Pokud zůstaneme u představy, že náhodná veličina X popisuje **dobu do poruchy** nějakého zařízení, pak typický tvar intenzity poruch je zobrazen na následujícím obrázku.

Křivka na tomto obrázku se nazývá **vanová křivka** a obvykle se dělí na tři úseky (I, II, III).

- I. V prvním úseku křivka intenzity poruch klesá. Odpovídající časový interval se nazývá období časných poruch (období záběhu, období počátečního provozu, období osvojování nebo období dětských nemocí podle analogie s úmrtnostní křivkou člověka). Příčinou zvětšené intenzity



poruch v tomto období jsou poruchy v důsledku výrobních vad, nesprávné montáže, chyb při návrhu nebo při výrobě apod.

- II. Ve druhém úseku dochází k běžnému využívání zaběhnutého výrobku, k poruchám dochází většinou z vnějších příčin, nedochází k opotřebení, které by změnilo funkční vlastnosti výrobku. Intenzita poruch je v tomto období přibližně konstantní. Příslušný časový interval se nazývá období normálního užití, či stabilního života.
- III. Ve třetím úseku procesy stárnutí a opotřebení mění funkční vlastnosti výrobku, projevují se nastřádané otřesy výrobku z období II (analogie s nesprávnou životosprávou člověka), trhliny materiálu a intenzita poruch vzrůstá. Příslušný časový interval se nazývá období poruch v důsledku stárnutí a opotřebení.

Poznámky:

1. Přestože uvedená intenzita poruch je typická pro mnoho průmyslových výrobků (a jakožto křivka úmrtnosti i pro člověka), lze ji těžko vyjádřit v elegantním analytickém tvaru pro všechna tři období najednou. Při vlastní analýze spolehlivosti musíme většinou aproximovat intenzitu poruch jednoduchými analytickými funkcemi vždy po jednotlivých obdobích.
2. U některých výrobků chybí období I, tj. období časných poruch. Je tomu např. u dobře kontrolovaných výrobků zaběhnutých přímo u výrobce. Jsou také výrobky, které „nestárnou“ - schází období III. To jsou např. výrobky vyřazené dříve než začnou stárnout. Velmi často, zejména při řešení spolehlivosti složitých systémů, budeme jednotlivé prvky sledovat pouze v období II, ve kterém je intenzita poruch přibližně konstantní.

3. Intenzitou poruch je úplně popsáno rozdělení doby do poruchy a naopak. Mezi funkcí

$$\text{bezporuchovosti a intenzitou poruch platí vztah: } R(t) = \exp\left(-\int_0^t \lambda(x)dx\right)$$

$$\lambda(t) = \frac{1}{R(t)} \frac{dR(t)}{dt}$$

□ Klasifikace monotónních intenzit poruch

V praxi vyšetřujeme intenzitu poruch po obdobích a tudíž se zabýváme studiem monotónních intenzit poruch. Proto se zavedly následující pojmy:

Rozdělení s distribuční funkcí $F(t)$ nazýváme **MIP rozdělením (RIP-rozdělením (anglicky IFR), KIP-rozdělením (anglicky DFR))**, jestliže odpovídající intenzita poruch je monotónní (neklesající, nerostoucí). Taktéž příslušné distribuční funkce budeme označovat MIP (RIP (IFR), KIP (DFR)).

Poznámka:

MIP ... monotónní intenzita poruch

RIP ... rostoucí intenzita poruch

KIP ... klesající intenzita poruch

□ Jaká je intenzita poruch systému složeného z n nezávislých prvků?

Věta:

Nechť se systém skládá z n nezávislých prvků s dobami do poruchy X_1, \dots, X_n a odpovídajícími intenzitami poruch $\lambda_1(t), \dots, \lambda_n(t)$, a necht' doba do poruchy systému je $t_{\min} = \min(X_1, \dots, X_n)$. Necht' $\lambda_{\min}(t)$ je intenzita poruch systému. Potom:

$$\lambda_{\min}(t) = \lambda_1(t) + \dots + \lambda_n(t)$$

Důkaz:

Necht' $R_{\min}(t)$ označuje funkci bezporuchovosti systému. Zřejmě:

$$R_{\min}(t) = \prod_{i=1}^n R_i(t),$$

kde $R_i(t)$ jsou funkce bezporuchovosti jednotlivých prvků.

Využitím vztahu $\lambda(t) = -\frac{d(\ln R(t))}{dt}$ dostáváme:

$$\lambda_{\min}(t) = -\frac{d(\ln R_{\min}(t))}{dt} = -\frac{d\left(\ln \prod_{i=1}^n R_i(t)\right)}{dt} = -\frac{d\left(\sum_{i=1}^n \ln R_i(t)\right)}{dt} = -\sum_{i=1}^n \frac{d(\ln R_i(t))}{dt} = \sum_{i=1}^n \lambda_i(t)$$

□ Reprodukční vlastnost Weibullova rozdělení

Jak jsme se dozvěděli, flexibilita Weibullova rozdělení umožňuje aproximovat širokou třídu rozdělení s monotónní intenzitou poruch. Takováto rozdělení se v technické praxi vyskytují poměrně často.

Weibullovo rozdělení má tuto reprodukční vlastnost:

Věta:

Nechť X_1, \dots, X_n jsou nezávislé stejně rozdělené náhodné veličiny s rozdělením $W(\Theta, \beta)$.

Potom náhodná veličina $X_{\min} = \min(X_1, \dots, X_n)$ má rozdělení $W\left(\frac{\Theta}{n^{\frac{1}{\beta}}}, \beta\right)$.

Důkaz:

Je-li $R_1(t)$ funkce bezporuchovosti náhodné veličiny X_1 a $R_{\min}(t)$ funkce bezporuchovosti náhodné veličiny X_{\min} , pak platí:

$$R_{\min}(t) = \prod_{i=1}^n R_i(t) = R_1^n(t)$$

V našem případě je tedy:

$$F_i(t) = 1 - e^{-\left(\frac{t}{\Theta}\right)^\beta}; \quad t > 0; \Theta > 0; \beta > 0 \quad \Rightarrow \quad R_i(t) = e^{-\left(\frac{t}{\Theta}\right)^\beta}; \quad t > 0; \Theta > 0; \beta > 0$$

$$\Rightarrow \quad R_{\min}(t) = R_i^n(t) = e^{-n \left(\frac{t}{\Theta}\right)^\beta} = e^{-\left(\frac{t}{\frac{\Theta}{n^{\frac{1}{\beta}}}}\right)^\beta}; \quad t > 0; \Theta > 0; \beta > 0,$$

což je funkce bezporuchovosti rozdělení $W\left(\frac{\Theta}{n^{\frac{1}{\beta}}}, \beta\right)$.



Řešený příklad

Životnost turbíny je dána životností funkčně nejslabší lopatky, protože moderní turbíny pracují s vysokými rychlostmi a porucha jedné lopatky má obvykle za následek zničení lopatkového kola, což je spojené s dalšími rozsáhlými škodami. Modelování životnosti lopatek má proto značný význam. Nechť doba do poruchy lopatky je náhodná veličina s Weibullovým rozdělením s parametrem tvaru 1,5 a parametrem měřítka 50. Jaké rozdělení má doba do poruchy turbíny (20 lopatek)?

Jestliže turbína má 20 lopatek s dobami do poruchy X_1, \dots, X_{20} , pak $X_{\min} = \min(X_1, \dots, X_{20})$ je doba do poruchy turbíny.

Do okamžiku poruchy pracují lopatky přibližně nezávisle na sobě, proto má doba do poruchy

turbíny Weibullovo rozdělení $W\left(\frac{\Theta}{n^{\frac{1}{\beta}}}, \beta\right)$.

$\Theta = 50$; $\beta = 1,5$; $n = 20 \Rightarrow$ Doba do poruchy turbíny Weibullovo rozdělení $W(6,8;1,5)$.



Shrnutí kapitoly 4.4.

Intenzita poruch je lokální charakteristikou spolehlivosti, je mírou **pravděpodobnosti** toho, že výrobek, který se neporouchal do okamžiku t , se porouchá v okamžiku bezprostředně následujícím po t . Intenzitou poruch je úplně popsáno pravděpodobnostní rozdělení doby do poruchy a naopak.

Vanová křivka je typická závislost intenzity poruch na čase. Na ní rozlišujeme tři charakteristická období života výrobku: období **časných poruch**, období **stabilního života** a období **stárnutí**.

Rozdělení s distribuční funkcí $F(t)$ nazýváme **MIP rozdělením (RIP-rozdělením (anglicky IFR), KIP-rozdělením (anglicky DFR))**, jestliže odpovídající intenzita poruch je monotónní (neklesající, nerostoucí). Taktéž příslušné distribuční funkce budeme označovat MIP (RIP (IFR), KIP (DFR)).

Intenzitu poruch systému složeného z n nezávislých prvků určíme podle následující věty:

Nechť se systém skládá z n nezávislých prvků s dobami do poruchy X_1, \dots, X_n a odpovídajícími intenzitami poruch $\lambda_1(t), \dots, \lambda_n(t)$, a nechť doba do poruchy systému je $t_{\min} = \min(X_1, \dots, X_n)$. Nechť $\lambda_{\min}(t)$ je intenzita poruch systému. Potom:

$$\lambda_{\min}(t) = \lambda_1(t) + \dots + \lambda_n(t)$$

Jako **reprodukční vlastnost Weibullova rozdělení** označujeme to, že jsou-li X_1, \dots, X_n nezávislé stejně rozdělené náhodné veličiny s rozdělením $W(\Theta, \beta)$. Potom náhodná veličina

$$X_{\min} = \min(X_1, \dots, X_n) \text{ má rozdělení } W\left(\frac{\Theta}{n^{\frac{1}{\beta}}}, \beta\right).$$



Otázky 4.4.

1. Charakterizujte intenzitu poruch pomocí pravděpodobnosti. Pravděpodobnost jakého jevu popisuje?
2. Co je to vanová křivka? Co je to období časných poruch?
3. Jaký je vztah mezi intenzitou poruch a funkcí bezporuchovosti?
4. Jak klasifikujeme pravděpodobnostní rozdělení na základě monotónní intenzity poruch?
5. Co je to reprodukční vlastnost Weibullova rozdělení?

4.5. Zálohování



Čas ke studiu: 15 minut



Cíl: Po prostudování tohoto odstavce budete umět

- charakterizovat podstatu zálohování
- rozlišit různé druhy zálohování a jednoduše je popsat
- formulovat základní zásadu pro zálohování



Výklad

□ Jaká je podstata zálohování a jaké druhy zálohování rozlišujeme ?

Zálohování je jedna ze základních metod zvyšování spolehlivosti, která umožňuje (alespoň teoreticky) neomezeně zvyšovat spolehlivost systémů. Podstata zálohování spočívá v tom, že se k prvku (tzv. hlavnímu) přidá jeden nebo více záložních prvků, které při poruše hlavního prvku tento prvek nahrazují.

Podle toho, v jakém režimu se nachází záložní prvek, dělíme zálohování do několika skupin. Jestliže záložní prvek pracuje ve stejném režimu jako prvek hlavní, mluvíme o **zatížené záloze („horké rezervě“)**. Jestliže záložní prvek plní svou funkci v mírnějším režimu než prvek hlavní, mluvíme o **odlehčené záloze**. Jestliže se záložní prvek nachází v režimu, ve kterém se nemůže porouchat, mluvíme o **nezatížené záloze („studené rezervě“)**. Ve většině skutečných zálohovaných systémů se setkáme s odlehčenou zálohou.

Důležitou součástí zálohovaných systémů je zařízení, které v případě poruchy hlavního prvku uvede do činnosti na místo hlavního prvku prvek záložní. Obecně se takové zařízení nazývá **přepínač**. V jednodušších modelech zálohování se předpokládá, že přepínač je absolutně spolehlivý. V reálných systémech však tomu tak nebývá, a proto při přesnější analýze je nutno v modelu počítat i s nespolehlivostí přepínačů.

□ Jak lze jednoduše popsat dva základní typy zálohování ?

Provedme nyní srovnání dob do poruchy zálohovaného systému se zatíženými a nezatíženými zálohami. Předpokládejme, že přepínač je absolutně spolehlivý, a že všechny prvky pracují na sobě nezávisle. Porouchaný prvek je okamžitě nahrazen prvkem záložním. Nechť X_1 je doba do poruchy hlavního prvku, a nechť X_2, \dots, X_n jsou doby do poruchy $n - 1$ záložních prvků.

Doba do poruchy zálohovaného systému se zatíženými zálohami je:

$$X_{\max} = \max (X_1, \dots, X_n)$$

a doba do poruchy zálohovaného systému s nezatíženými zálohami je:

$$X^{(n)} = X_1 + \dots + X_n.$$

Vzhledem k tomu, že $X_{\max} \leq X^{(n)}$, je zálohovaný systém s nezatíženými zálohami vždy výhodnější než zálohovaný systém se zatíženými zálohami.



Shrnutí kapitoly 4.5.

Základním problémem zálohování systémů je, zda zálohovat jednotlivé prvky systému nebo zda zálohovat celý systém identickým záložním systémem. Toto jsou extrémní případy, mezi kterými existuje široká škála možností zálohování. Některé bloky (tj. části systému) je možno zálohovat identickými bloky, jiné pak zálohovat po prvcích apod. Obecně lze snadno ukázat, že zálohování prvků vede vždy k vyšší spolehlivosti než zálohování bloků.

Podstata zálohování spočívá v tom, že se k prvku (tzv. hlavnímu) přidá jeden nebo více záložních prvků, které při poruše hlavního prvku tento prvek nahrazují. Záložní prvky mohou pracovat buď jako horké nebo studené rezervy.

Zálohovaný systém s nezatíženými zálohami vždy výhodnější (spolehlivější) než zálohovaný systém se zatíženými zálohami.

Zálohování prvků vede vždy k vyšší spolehlivosti než zálohování bloků.



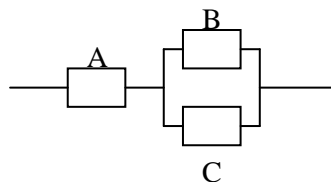
Otázky 4.5.

1. Charakterizujte podstatu zálohování.
2. Co je to horká rezerva? Co je to studená rezerva ?
3. Jaká jsou základní pravidla pro zálohování ?



Úlohy k řešení 4.5.

1. Systém na obrázku je funkční pokud funguje součástka A a nejméně jedna ze součástek B a C. Nechť pro jednotlivé součástky byly naměřeny následující doby do poruchy (A, B, C) = (400, 200, 300 hodin). Předpokládáme, že systém pracuje nezávisle na okolních podmínkách.
 - a) Nechť součástka C pracuje v režimu studená rezerva. Po kolika hodinách dojde k poruše systému ?
 - b) Nechť součástka C pracuje v režimu horká rezerva. Po kolika hodinách dojde k poruše systému ?



5. TEORIE ODHADU

Na úvod této kapitoly si zopakujme základní vlastnosti bodových odhadů.

5.1. Vlastnosti „dobrého“ bodového odhadu



Čas ke studiu: 25 minut



Cíl: Po prostudování tohoto odstavce budete:

- znát vlastnosti bodových odhadů
- rozumět pojmu dostatečná statistika a budete umět určit, zda vybraná statistika je dostatečnou



Výklad

Dobry“ (věrohodný) odhad musí splňovat určité vlastnosti. Mezi základní vlastnosti věrohodných odhadů patří:

- nestrannost (nevychýlenost, nezkreslenost)
- vydatnost (eficience)
- konzistence
- dostatečnost

□ Nestranný odhad

Řekneme, že odhad je **nestranný**, jestliže se jeho střední hodnota rovná hledanému parametru ($E\hat{\Theta} = \Theta$).

Znamená to, že tento odhad systematicky nenadhodnocuje ani nepodhodnocuje odhadovaný parametr.

Slabší formou nestrannosti je **asymptotická nestrannost**. Říkáme, že odhad je asymptoticky nestranný pokud: $\lim_{n \rightarrow \infty} E\hat{\Theta} = \Theta$

Příklady nestranných odhadů:

- \bar{X} je nestranným odhadem střední hodnoty (limitní věty)
- Výběrová relativní četnost p je nestranným odhadem relativní četnosti (podílu) π
- V případě náhodného výběru z normálního rozdělení je výběrový rozptyl s^2 nestranným odhadem rozptylu σ^2

Je třeba říci, že existuje mnoho dobrých odhadů, které nejsou nestranné.

□ Vydátný (eficientní) odhad

Nestrannost sama o sobě nezaručuje, že je odhad „dobrý“. Rádi bychom dosáhli také toho, aby bodové odhady byly rozloženy co nejtěsněji kolem odhadovaného parametru. Pokud budeme mít dva nestranné odhady $\hat{\theta}_1$ a $\hat{\theta}_2$, vybereme si ten, který bude mít menší rozptyl. Tato vlastnost se nazývá **vydátnost** (eficience).

Jestliže pro dva nestranné odhady $\hat{\theta}_1$ a $\hat{\theta}_2$ platí $D\hat{\theta}_1 > D\hat{\theta}_2$, potom je **relativní eficience** odhadu $\hat{\theta}_1$ vzhledem k odhadu $\hat{\theta}_2$ dána podílem $D\hat{\theta}_1 / D\hat{\theta}_2$, což je číslo mezi 0 a 1.

Nestranný odhad, jehož rozptyl je nejmenší mezi všemi nestrannými odhady příslušného parametru, se nazývá **nejlepší nestranný (eficientní) odhad**.

Příklady nejlepších nestranných odhadů:

- \bar{X} je nejlepším nestranným odhadem střední hodnoty (limitní věty)
- Výběrová relativní četnost p je nejlepším nestranným odhadem rel. četnosti (podílu) π
- V případě náhodného výběru z normálního rozdělení je výběrový rozptyl s^2 nejlepším nestranným odhadem rozptylu σ^2

□ Konzistentní odhad

Další žádoucí vlastností dobrého odhadu je konzistence. Odhad je konzistentní, pokud se s rostoucím rozsahem výběru (n) zpřesňuje, k čemuž dochází pokud:

- $\hat{\theta}$ je asymptoticky nestranný, tj. $E\hat{\theta} \rightarrow \theta$
- $\lim_{n \rightarrow \infty} D\hat{\theta} = 0$

Vlastnost b) říká, že se s rostoucím n (rozsahem výběru) rozdělení $\hat{\theta}$ zužuje kolem hledaného parametru.

Příklady konzistentních odhadů:

- \bar{X} je konzistentním odhadem střední hodnoty, protože $D\bar{X} = \frac{\sigma^2}{n} \rightarrow 0$ pro $n \rightarrow \infty$
- Výběrová relativní četnost p je konzistentním odhadem rel. četnosti (podílu) π , protože $Dp = \frac{\pi(1-\pi)}{n} \rightarrow 0$ pro $n \rightarrow \infty$

□ Dostatečný (postačující) odhad

Odhad parametru je **dostatečný**, jestliže obsahuje veškerou informaci o sledovaném parametru, kterou může výběrový soubor poskytnout. Znamená to, že žádný jiný parametr neobsahuje větší množství informace o výběrovém souboru.

Příklady dostatečných odhadů:

- \bar{X} je dostatečným odhadem střední hodnoty, protože pro jeho výpočet jsou použity všechny hodnoty výběrového souboru (nese největší informaci, srovnejte například s mediánem)
- Výběrová relativní četnost p je dostatečným odhadem rel. četnosti (podílu) π , protože pro její výpočet jsou použity všechny hodnoty výběrového souboru

Následující pasáže těchto materiálů (až po kapitolu věnovanou Rao-Cramerově nerovnosti) jsou z velké části inspirovány [Riečan, Lamoš, Lenárt: Pravděpodobnost a matematická statistika, Bratislava 1984].

□ Postačující statistika pro parametr Θ

Důkaz toho, zda je určitý odhad efektivní (nejlepší nestranný), není vždy jednoduchý. Abychom našli odhad, který má nejmenší rozptyl, je vhodné nahrazení celého výběru jednou statistikou, a to takovou, která bude obsahovat „veškerou“ informaci o parametru Θ .

Pokud je možné pomocí nějaké statistiky (může se jednat o vícerozměrnou statistiku) odhadnout neznámé parametry souvisejícího rozdělení, hovoříme o **postačující statistice**. Nejjednodušší postačující statistikou je podle definice samotný vektor náhodného výběru $\underline{X} = (X_1, \dots, X_n)$, taková postačující statistika však není příliš užitečná. Smysl má hledat takové postačující statistiky, které mají rozměr menší než n .

Definice:

Reálnou funkci $T_n(\underline{X})$ nazveme **postačující statistikou pro parametr θ** , jestliže sdružené rozdělení náhodného výběru $\underline{X} = (X_1, \dots, X_n)$ podmíněné jevem $T(\underline{X})=t$ není pro žádné t závislé na Θ .

Statistika $T(\underline{X}) = \sum_{i=1}^n T(x_i)$ představuje největší možnou redukci výsledků pozorování (nahrazení n pozorování menším počtem údajů). Proto se označuje jako **minimální postačující statistika**.

Jestliže pro parametrickou funkci $\tau(\Theta)$ existují nestranné odhady, pak nejlepší z nich (ve smyslu minimálního rozptylu) je funkcí minimálních postačujících statistik a je určen jednoznačně.

Sdružená pravděpodobnostní funkce u některých rozdělení:

Poissonovo rozdělení : $P(x; \lambda) = \exp(\ln \lambda \sum x_i - \lambda - \sum \ln(x_i!))$

a postačující statistikou pro parametr λ je výběrový úhrn $\sum x_i$.

Exponenciálního rozdělení je $f(\mathbf{x}; \delta) = \exp\left[-n \ln \delta - \frac{1}{\delta} \sum x_i\right]$

a opět postačující statistikou pro parametr δ je výběrový úhrn $\sum x_i$

Normálního rozdělení $N(0, \sigma^2)$, které má hustotu $f(x_i; 0; \sigma^2) = \exp\left(-\frac{1}{2} \ln 2\pi - \frac{1}{2} \ln \sigma^2 - \frac{x_i^2}{2\sigma^2}\right)$

$$\text{a sdružená hustota } f(x; 0; \sigma^2) = \exp\left(-\frac{n}{2} \ln 2\pi - \frac{n}{2} \ln \sigma^2 - \frac{1}{2\sigma^2} \sum_{i=1}^n x_i^2\right)$$

postačující statistikou pro parametr σ^2 je tedy $\sum x_i^2$

Jednoduchý postup při hledání postačujících statistik nabízí věta o faktorizaci. Tato věta zároveň umožňuje rychle rozhodnout o tom, zda je určitá statistika dostatečnou.

Věta o faktorizaci:

Nechť $\underline{X} = (X_1, \dots, X_n)$ je náhodný výběr z rozdělení $f(x; \Theta)$. $T(\underline{X})$ je postačující statistikou pro parametr Θ tehdy, jestliže sdružené rozdělení náhodného výběru je součinem dvou faktorů:

$$f(\underline{x}, \Theta) = g\{T(\underline{x}), \Theta\} \cdot h(\underline{x})$$



Řešený příklad

Nechť $\underline{X} = (X_1, \dots, X_n)$ je náhodný výběr z Poissonova rozdělení. Dokažme, že $T(\underline{X}) = \sum_{i=1}^n X_i$ je postačující statistikou pro parametr Θ Poissonova rozdělení ($\Theta = \lambda t$).

$$f(x, \Theta) = \frac{\Theta^x}{x!} \cdot e^{-\Theta} \quad x = 0, 1, 2, \dots, \Theta > 0$$

Sdružené rozdělení výběru má tvar:

$$\begin{aligned} f(\underline{x}, \Theta) &= P(X_1 = x_1, X_2 = x_2, \dots, X_n = x_n) = \prod_{i=1}^n P(X_i = x_i) = \prod_{i=1}^n \frac{\Theta^{x_i}}{x_i!} \cdot e^{-\Theta} = \\ &= e^{-n \cdot \Theta} \cdot \frac{\Theta^{\sum_{i=1}^n x_i}}{\prod_{i=1}^n x_i!}, \end{aligned}$$

kde $x_i = 0, 1, \dots, i = 1, 2, \dots,$

Sdružené rozdělení výběru můžeme faktorizovat, tj. můžeme jej zapsat jako součin dvou faktorů:

$$\begin{aligned} f(\underline{x}, \Theta) &= g(t, \Theta) \cdot h(x) \\ f(\underline{x}, \Theta) &= e^{-n \cdot \Theta} \cdot \Theta^t \cdot \frac{1}{\prod_{i=1}^n x_i!}, \end{aligned}$$

$$\text{kde } g(t, \Theta) = e^{-n\Theta} \cdot \Theta^t; \quad h(x) = \frac{1}{\prod_{i=1}^n x_i!}$$

$T(X) = \sum_{i=1}^n X_i$ je tedy postačující statistikou pro parametr Θ .



Shrnutí kapitoly 5.1.

„Dobry“ (věrohodný) odhad musí splňovat určité vlastnosti. Mezi základní vlastnosti věrohodných odhadů patří:

- **nestrannost** (nevychýlenost, nezkreslenost)
- **vydatnost** (eficience)
- **konzistence**
- **dostatečnost**

Reálnou funkci $T_n(X)$ nazveme **postačující statistikou pro parametr Θ** , jestliže sdružené rozdělení náhodného výběru $\underline{X} = (X_1, \dots, X_n)$ podmíněné jevem $T(\underline{X})=t$ není pro žádné t závislé na Θ .

Jednoduchý postup při hledání dostačujících statistik nabízí **věta o faktorizaci**. Tato věta zároveň umožňuje rychle rozhodnout o tom, zda je určitá statistika postačující.



Otázky 5.1.

1. Vyjmenujte a objasněte základní vlastnosti „dobrého“ bodového odhadu.
2. Co je to postačující statistika pro parametr Θ ?
3. Věta o faktorizaci – vysvětlete.

5.2. Konstrukce efektivních odhadů



Čas ke studiu: 25 minut



Cíl: Po prostudování tohoto odstavce budete:

- nalézt efektivní odhad reálné funkce parametru Θ (Θ je parametr rozdělení náhodného výběru)



Výklad

V této kapitole se seznámíme s Rao-Blackwellovou větou, která ukazuje praktický význam postačujících statistik pro výpočet efektivních (nejlepších nestranných) odhadů.

Rao-Blackwellova věta

Nechť $\underline{X} = (X_1, \dots, X_n)$ je náhodný výběr z rozdělení $f(x; \Theta)$. Nechť existuje postačující statistika $T(\underline{X})$ pro parametr Θ . Nechť $\tau(\Theta)$ je reálná funkce parametru Θ a $T^*(\underline{X})$ je nestranný odhad této charakteristiky. Potom platí:

1. Pro funkci $\tau(\Theta)$ existuje nestranný odhad $\tilde{T}(X) = \tilde{T}(T(X))$, který je funkcí postačující statistiky $T(\underline{X})$.

2. Nestranný odhad $\tilde{T}(\underline{X})$ má rozptyl menší nebo roven rozptylu odhadu $T^*(\underline{X})$:

$$D(\tilde{T}(\underline{X})) \leq D(T^*(\underline{X})) \quad \text{pro všechna } \Theta$$

3. $D(\tilde{T}(\underline{X})) = D(T^*(\underline{X})) \Leftrightarrow P(\tilde{T}(\underline{X}) = T^*(\underline{X})) = 1 \quad \text{pro všechna } \Theta$

**Průvodce studiem**

Pro důkaz Rao-Blackwellovy věty je nutné znát níže uvedenou vlastnost podmíněné střední hodnoty.

Je-li (X, Y) spojité náhodné vektory se sdruženou hustotou $f(x, y)$, definujeme podmíněnou střední hodnotu takto:

$$E(X|Y=y) = \int_{-\infty}^{\infty} x \cdot f(x|y) dx,$$

$$\text{kde } f(x|y) = \frac{f(x, y)}{f_Y(y)} = \frac{f(x, y)}{\int_{-\infty}^{\infty} f(x, y) dx}$$

Důležitou vlastností podmíněné střední hodnoty je, že:

$$\begin{aligned} E_Y[E(X|Y)] &= \int_{-\infty}^{\infty} E(X|Y) \cdot f_Y(y) dy = \int_{-\infty}^{\infty} f_Y(y) \int_{-\infty}^{\infty} x \cdot f(x|y) dx dy = \\ &= \int_{-\infty}^{\infty} f_Y(y) \int_{-\infty}^{\infty} x \cdot \frac{f(x, y)}{f_Y(y)} dx dy = \int_{-\infty}^{\infty} x \cdot f_X(x) dx = E(X) \end{aligned}$$

kde E_Y je střední hodnota vzhledem k náhodné veličině Y .

Důkaz:

ad1) Nechť $T^*(\underline{X})$ je libovolný nestranný odhad parametrické funkce $\tau(\Theta)$ a $T(\underline{X})$ je postačující statistika pro parametr θ .

Položme: $\tilde{T}(t) = E\{T^*(X) | T(X) = t\}$

Protože $T(\underline{X})$ je výběrová charakteristika, funkce $\tilde{T}(t)$ není funkcí θ . $\tilde{T}(T(\underline{X}))$ je statistika.

Dokážeme, že $\tilde{T}(T(\underline{X}))$ je nestranný odhad parametrické funkce $\tau(\Theta)$.

Pro každé Θ platí: $E_T(\tilde{T}) = E_T\{T^*(\underline{X}) | T(\underline{X}) = t\} = E(T^*(\underline{X})) = \tau(\Theta)$

ad2)

$$\begin{aligned} D_T(T^*(\underline{X})) &= E_T\{[T^*(\underline{X}) - E(T^*(\underline{X}))]^2\} = E_T\{[T^*(\underline{X}) - \tau(\Theta)]^2\} = \\ &= E_T\{[T^*(\underline{X}) - \tilde{T}(t)] + [\tilde{T}(t) - \tau(\Theta)]\}^2 = \\ &= E_T\{[T^*(\underline{X}) - \tilde{T}(t)]^2\} + 2E_T\{[T^*(\underline{X}) - \tilde{T}(t)] \cdot [\tilde{T}(t) - \tau(\Theta)]\} + E_T\{[\tilde{T}(t) - \tau(\Theta)]^2\} \end{aligned}$$

Střední hodnotu součinu $[T^*(\underline{X}) - \tilde{T}(t)] \cdot [\tilde{T}(t) - \tau(\Theta)]$ můžeme vyjádřit jako:

$$\begin{aligned} E_T\{[T^*(\underline{X}) - \tilde{T}(t)] \cdot [\tilde{T}(t) - \tau(\Theta)]\} &= E_T\{[T^*(\underline{X}) - \tilde{T}(t)]\} \cdot E_T\{[\tilde{T}(t) - \tau(\Theta)]\} = \\ &= E_T\{[T^*(\underline{X}) - \tilde{T}(t)]\} \cdot E\{[T^*(\underline{X}) - \tilde{T}(t)] | T(\underline{X}) = t\} = \\ &= E_T\{[T^*(\underline{X}) - \tilde{T}(t)]\} \cdot 0 = 0 \end{aligned}$$

Tedy:

$$\begin{aligned} D_T(T^*(\underline{X})) &= E_T\{[T^*(\underline{X}) - \tilde{T}(t)]^2\} + 2E_T\{[T^*(\underline{X}) - \tilde{T}(t)] \cdot [\tilde{T}(t) - \tau(\Theta)]\} + E_T\{[\tilde{T}(t) - \tau(\Theta)]^2\} = \\ &= E_T\{[T^*(\underline{X}) - \tilde{T}(t)]^2\} + 0 + D_T(\tilde{T}(t)) \end{aligned}$$

$$E_T\{[T^*(\underline{X}) - \tilde{T}(t)]^2\} \geq 0 \quad \Rightarrow \quad D_T(T^*(\underline{X})) \geq D_T(\tilde{T}(t))$$

ad3)

$$D_T(T^*(\underline{X})) \geq D_T(\tilde{T}(t)) \quad \Leftrightarrow \quad E_T\{[T^*(\underline{X}) - \tilde{T}(t)]^2\} = 0 \quad \Leftrightarrow \quad P\{T^*(\underline{X}) = \tilde{T}(t)\} = 1$$

Z Rao-Blackwellovy věty vyplývá, že při hledání nejlepších nestranných odhadů se můžeme omezit na odhady, které jsou funkcemi postačujících statistik. Tato věta nám ukazuje, jak v případě, že známe libovolný nestranný odhad, určit nestranný odhad, který je funkcí postačující statistiky.



Řešený příklad

Necht' $\underline{X} = (X_1, \dots, X_n)$ je náhodný výběr z Poissonova rozdělení:

$$f(x, \Theta) = \frac{\Theta^x}{x!} \cdot e^{-\Theta} \quad x = 0, 1, 2, \dots, \Theta > 0$$

Nalezněte nejlepší nestranný odhad pravděpodobnosti toho, že náhodná veličina X s Poissonovým rozdělením nabude hodnoty 0.

Hledáme nejlepší nestranný odhad parametrické funkce: $\tau(\Theta) = e^{-\Theta}$

Vzhledem k tomu, že $\tau(\Theta)$ je pravděpodobností toho, že náhodná veličina X s Poissonovým rozdělením nabude hodnoty 0, nabízí se jako vhodná postačující statistika relativní četnost nulových hodnot X_i ve výběru, tj.

$$T^*(\underline{X}) = T(X_1, \dots, X_n) = \frac{1}{n} \sum_{i=1}^n Y_i,$$

kde:

$$\begin{aligned} Y_i &= 0 && \text{pro } X_i \geq 1 \\ Y_i &= 1 && \text{pro } X_i = 0, \quad i = 1, 2, \dots, n \end{aligned}$$

Z předcházejícího řešeného příkladu víme, že pro parametr Θ je $T(\underline{X}) = \sum_{i=1}^n X_i$ postačující statistikou.

Nejlepší nestranný (efektivní) odhad $\tilde{T}(t)$ funkce $\tau(\Theta)$ budeme hledat následujícím způsobem:

1. Najdeme střední hodnotu $T^*(\underline{X})$ podmíněnou jevem $T(\underline{X}) = \sum_{i=1}^n X_i = t$

$$\begin{aligned} \tilde{T}(t) &= E\{T^*(\underline{X}) | T(\underline{X}) = t\} = E\left\{\frac{1}{n} \sum_{i=1}^n Y_i \mid \sum_{i=1}^n X_i = t\right\} = \frac{1}{n} \sum_{i=1}^n P\left\{Y_i = 1 \mid \sum_{i=1}^n X_i = t\right\} = \\ &= \frac{1}{n} \sum_{i=1}^n P\left\{X_i = 0 \mid \sum_{i=1}^n X_i = t\right\} = P\left\{X_1 = 0 \mid \sum_{i=1}^n X_i = t\right\} \end{aligned}$$

2. Postačující statistiku $T(\underline{X}) = \sum_{i=1}^n X_i$ můžeme zapsat ve tvaru:

$$T(\underline{X}) = X_1 + \sum_{i=2}^n X_i = X_1 + Z,$$

kde X_1 a Z jsou nezávislé náhodné veličiny s Poissonovým rozdělením:

$$E_{\Theta}(X_1) = \Theta \quad E_{\Theta}(Z) = (n-1) \cdot \Theta$$

Pak:

$$\tilde{T}(t) = \frac{P_{\Theta}(X_1 = 0) \cdot P_{\Theta}(Z = t)}{P_{\Theta}(X_1 + Z = t)} = \frac{e^{-\lambda - \lambda(n-1)} \cdot \frac{[\lambda \cdot (n-1)]^t}{t!}}{e^{-n\lambda} \cdot \frac{(n\lambda)^t}{t!}} = \left(\frac{n-1}{n} \right)^t$$

Efektivním odhadem parametrické funkce $\tau(\Theta) = e^{-\Theta}$ je odhad $\tilde{T}(t) = \left(\frac{n-1}{n} \right)^{\sum_{i=1}^n x_i}$.



Otázky 5.2.

1. Rao – Blackwellova věta, popište postup při hledání efektivních odhadů.

5.3. Fisherova míra informace



Čas ke studiu: 25 minut



Cíl: Po prostudování tohoto odstavce budete:

- znát pojem Fischerova míra informace
- umět nalézt Fischerovu míru informace $I(\Theta)$



Výklad

Důležitým ukazatelem kvality odhadu je jeho rozptyl (připomeňme, že každý odhad je náhodnou veličinou). Vyhovuje-li rozdělení, jehož parametr odhadujeme, jistým obecným předpokladům, lze ukázat, že není možné zkonstruovat odhad s rozptylem menším než jistá hodnota, tzv. Raova-Cramerova hranice. Mezi odhady s požadovanou vlastností se tedy vždy snažíme nalézt odhad, jehož rozptyl je roven této hodnotě. Pokud se to podaří, hovoříme o **odhadu s minimálním rozptylem**. Ve statistické literatuře se často operuje s **nestranným odhadem s minimálním rozptylem**.

K libovolnému rozdělení $f(\underline{x}, \Theta)$ a k libovolné parametrické funkci $\tau(\Theta)$ budeme hledat takovou funkci $C(\Theta)$, aby libovolný nestranný odhad $\tilde{T}(\underline{X})$, který splňuje podmínky regularity, měl rozptyl větší než $C(\Theta)$. Funkce $C(\Theta)$ tedy bude dolní mezí rozptylů pro všechny nestranné odhady parametrické funkce $\tau(\Theta)$.

Existují nestranné odhady, jejichž rozptyl je roven $C(\Theta)$. V některých případech je však tato hranice dosažitelná pouze asymptoticky (pro $n \rightarrow \infty$).

Dříve než se pustíme do hledání funkce $C(\Theta)$, definujeme si některé pojmy.

Definice:

Předpokládejme, že Θ je jednorozměrný parametr. Říkáme, že **systém hustot**

$$\{f(\underline{x}, \Theta), \Theta \in \Theta\}$$

je regulární právě když:

1. Θ je neprázdná otevřená množina
2. Množina $M = \{\underline{x}: f(\underline{x}, \Theta) > 0\}$ není funkcí Θ
3. Pro každé $\underline{x} \in M$ existuje konečná parciální derivace:

$$f'(\underline{x}, \Theta) = \frac{\partial f(\underline{x}, \Theta)}{\partial \Theta}$$

Pro každé $\Theta \in \Theta$, $Z(\underline{X}, \Theta) = \frac{\partial \ln f(\underline{X}, \Theta)}{\partial \Theta}$ platí:

4. $EZ = 0$

$$\begin{aligned} \int_M \frac{\partial \ln f(\underline{x}, \Theta)}{\partial \Theta} \cdot f(\underline{x}, \Theta) dx &= \int_M \frac{\partial \ln f(\underline{x}, \Theta)}{\partial \Theta} \cdot \frac{dF(\underline{x}, \Theta)}{d\underline{x}} dx = \int_M \frac{\partial \ln f(\underline{x}, \Theta)}{\partial \Theta} dF(\underline{x}, \Theta) = \\ &= \int_M \frac{f'(\underline{x}, \Theta)}{f(\underline{x}, \Theta)} dF(\underline{x}, \Theta) = \int_M \frac{f'(\underline{x}, \Theta)}{f(\underline{x}, \Theta)} f(\underline{x}, \Theta) dx = \int_M f'(\underline{x}, \Theta) dx = 0 \end{aligned}$$

5. $0 < DZ < \infty$ (konečný, kladný rozptyl)

$$\begin{aligned} DZ = I(\Theta) &= \int_M \left(\frac{\partial \ln f(\underline{x}, \Theta)}{\partial \Theta} - EZ \right)^2 dF(\underline{x}, \Theta) = \int_M \left(\frac{\partial \ln f(\underline{x}, \Theta)}{\partial \Theta} \right)^2 dF(\underline{x}, \Theta) = \\ &= \int_M \left(\frac{f'(\underline{x}, \Theta)}{f(\underline{x}, \Theta)} \right)^2 dF(\underline{x}, \Theta) = \int_M \left(\frac{f'(\underline{x}, \Theta)}{f(\underline{x}, \Theta)} \right)^2 f(\underline{x}, \Theta) dx \end{aligned}$$

Zjednodušeně často místo o regulárnosti systému hustot mluvíme o **regulárnosti rozdělení $f(\underline{x}, \Theta)$** .



Řešený příklad

Dokažte že hustota normálního rozdělení $N(\Theta, 1)$ je regulární.

Hustota normálního rozdělení $N(\Theta, 1)$:

$$f(x, \Theta) = \frac{1}{\sqrt{2\pi}} \cdot e^{-\frac{(x-\Theta)^2}{2}} \quad \text{pro } \Theta \in R$$

ad1) $\Theta \in R$, Θ je neprázdná otevřená množina

ad2) $M=R$, Množina $M = \{x: f(x, \Theta) > 0\}$ není funkcí Θ

$$\text{ad3) } \forall x \in R: f'(x, \Theta) = \frac{1}{\sqrt{2\pi}} \cdot e^{-\frac{(x-\Theta)^2}{2}} \cdot (x - \Theta) \quad \text{pro } \Theta \in R$$

ad4) Pro každé $\Theta \in \Theta$, $Z(X, \Theta) = \frac{\partial \ln f(X, \Theta)}{\partial \Theta}$ platí: $EZ = 0$

$$\begin{aligned} \int_{-\infty}^{\infty} f'(x, \Theta) dx &= \int_{-\infty}^{\infty} \frac{1}{\sqrt{2\pi}} \cdot e^{-\frac{(x-\Theta)^2}{2}} \cdot (x-\Theta) dx = \left| \begin{array}{l} t = x - \Theta \\ dt = dx \end{array} \right| = \frac{1}{\sqrt{2\pi}} \cdot \int_{-\infty}^{\infty} e^{-\frac{t^2}{2}} \cdot t dt = \\ &= \frac{1}{\sqrt{2\pi}} \cdot \left(\int_{-\infty}^0 e^{-\frac{t^2}{2}} \cdot t dt + \int_0^{\infty} e^{-\frac{t^2}{2}} \cdot t dt \right) = \frac{1}{\sqrt{2\pi}} \cdot (I_1 + I_2) = \frac{1}{\sqrt{2\pi}} \cdot (-1 + 1) \\ &= 0 \end{aligned}$$

kde:

$$\begin{aligned} I_1 &= \int_{-\infty}^0 e^{-\frac{t^2}{2}} \cdot t dt = \left| \begin{array}{l} u = -\frac{t^2}{2} \\ du = -t dt \end{array} \right| = - \int_{-\infty}^0 e^u du = - \lim_{v \rightarrow (-\infty)} [e^u]_v^0 = -(1 - 0) = -1 \\ I_2 &= \int_0^{\infty} e^{-\frac{t^2}{2}} \cdot t dt = \left| \begin{array}{l} u = \frac{t^2}{2} \\ du = t dt \end{array} \right| = \int_0^{\infty} e^{-u} du = - \lim_{v \rightarrow (\infty)} [e^{-u}]_0^v = -(0 - 1) = 1 \end{aligned}$$

ad5) Pro každé $\Theta \in \Theta$, $Z(X, \Theta) = \frac{\partial \ln f(X, \Theta)}{\partial \Theta}$ platí: ($0 < DZ < \infty$)

$$\begin{aligned} DZ &= I(\Theta) = \int_M \left(\frac{f'(x, \Theta)}{f(x, \Theta)} \right)^2 f(x, \Theta) dx = \int_{-\infty}^{\infty} (x - \Theta)^2 \cdot f(x, \Theta) dx = \\ &= \int_{-\infty}^{\infty} (x - EX)^2 \cdot f(x, \Theta) dx = DX = 1 \end{aligned}$$

Hustota normálního rozdělení $N(\Theta, 1)$ je tedy regulární.

Definice:

Nechť náhodná veličina X má regulární rozdělení $f(\underline{x}, \Theta)$. Integrál $I(\Theta)$ definovaný v podmínce 5 v definici regulárního systému hustot nazýváme **Fisherovou mírou informace**:

$$I(\Theta) = \int_M \left(\frac{\partial \ln f(\underline{x}, \Theta)}{\partial \Theta} \right)^2 dF(\underline{x}, \Theta) = \int_M \left(\frac{f'(\underline{x}, \Theta)}{f(\underline{x}, \Theta)} \right)^2 dF(\underline{x}, \Theta) = \int_M \left(\frac{f'(\underline{x}, \Theta)}{f(\underline{x}, \Theta)} \right)^2 f(\underline{x}, \Theta) dx$$

Fischerova míra informace je tedy střední hodnota náhodné veličiny definované jako:

$$\left(\frac{f'(X, \Theta)}{f(X, \Theta)} \right)^2$$

Můžeme ji tedy zapisovat také jako:

$$I(\Theta) = E \left\{ \left(\frac{f'(X, \Theta)}{f(X, \Theta)} \right)^2 \right\} = E \left\{ \left(\frac{\partial \ln f(\underline{x}, \Theta)}{\partial \Theta} \right)^2 \right\}$$

Následující věta nám uvádí důležitou vlastnost Fisherovy míry informace, využitelnou především při výpočtu informace v praktických příkladech.

Věta:

Nechť $f(\underline{x}, \Theta)$ je regulární hustota. Necht' pro všechna x a pro všechna Θ existuje druhá derivace $f(\underline{x}, \Theta)$:

$$f''(\underline{x}, \Theta) = \frac{\partial^2 f(\underline{x}, \Theta)}{\partial \Theta^2}$$

Necht' pro každé $\Theta \in \Theta$ platí:

$$\int_M \frac{f''(\underline{x}, \Theta)}{f(\underline{x}, \Theta)} dF(\underline{x}, \Theta) = 0$$

Potom:

$$I(\Theta) = - \int_M \frac{\partial^2 \ln f(\underline{x}, \Theta)}{\partial \Theta^2} dF(\underline{x}, \Theta)$$

**Řešený příklad**

Necht' náhodná veličina X má normální rozdělení $N(\mu, \sigma^2)$, kde σ^2 známe. Určete míru informace o parametru μ .

$$f(x, \mu) = \frac{1}{\sqrt{2\pi} \cdot \sigma} \cdot e^{-\frac{(x-\mu)^2}{2\sigma^2}} \quad \text{pro } x \in R$$

$$\ln f(x, \mu) = \ln \frac{1}{\sqrt{2\pi}} + \ln \frac{1}{\sigma} + \ln e^{-\frac{(x-\mu)^2}{2\sigma^2}} = -\frac{1}{2} \ln(2\pi) - \ln \sigma - \frac{(x-\mu)^2}{2\sigma^2}$$

$$\frac{\partial(\ln f(x, \mu))}{\partial \mu} = \frac{(x-\mu)}{\sigma^2}$$

$$\begin{aligned} \underline{\underline{I(\Theta)}} &= E \left\{ \left(\frac{\partial \ln f(x, \mu)}{\partial \mu} \right)^2 \right\} = E \left\{ \frac{(x-\mu)^2}{\sigma^4} \right\} = \frac{1}{\sigma^4} \cdot E \left\{ (x-\mu)^2 \right\} = \frac{1}{\sigma^4} \cdot DX = \frac{1}{\sigma^4} \cdot \sigma^2 = \\ &= \underline{\underline{\frac{1}{\sigma^2}}} \end{aligned}$$

**Řešený příklad**

Necht' náhodná veličina X má Poissonovo rozdělení $Po(\lambda)$. Určete míru informace o parametru λ .

$$f(k, \lambda) = \frac{\lambda^k}{k!} \cdot e^{-\lambda} \quad k = 0, 1, \dots, \quad \lambda > 0$$

$$EX = DX = \lambda$$

Je zřejmé, že $f(k, \lambda)$ je regulární hustota.

$$\ln f(k, \lambda) = \ln(\lambda)^k - \ln k! + \ln e^{-\lambda} = k \cdot \ln(\lambda) - \ln k! - \lambda$$

$$\frac{\partial(\ln f(k, \lambda))}{\partial \lambda} = \frac{k}{\lambda} - 1 = \frac{k - \lambda}{\lambda}$$

$$\underline{\underline{I(\Theta)}} = E \left\{ \left(\frac{\partial \ln f(k, \lambda)}{\partial \lambda} \right)^2 \right\} = E \left\{ \frac{(k - \lambda)^2}{\lambda^2} \right\} = \frac{1}{\lambda^2} \cdot E \{ (k - \lambda)^2 \} = \frac{1}{\lambda^2} \cdot DX = \frac{1}{\lambda^2} \cdot \lambda = \underline{\underline{\frac{1}{\lambda}}}$$



Otázky 5.3.

1. Definujte Fischerovu míru informace
2. Popište jak v praxi hledáme $I(\Theta)$

5.4. Rao – Cramerova nerovnost



Čas ke studiu: 25 minut



Cíl: Po prostudování tohoto odstavce budete:

- znát praktický význam Fischerovy míry informace
- umět nalézt dolní mez rozptylu nestranných odhadů - $C(\Theta)$



Výklad

V této kapitole si ukážeme jaký je vztah mezi Fisherovou mírou informace a dolní mezí rozptylu $C(\Theta)$ nestranných odhadů dané parametrické funkce.

Definice:

Nechť $\underline{X} = (X_1, \dots, X_n)$ je náhodným výběrem, který má regulární rozdělení $f(\mathbf{x}, \Theta)$, jenž je funkcí jednoho reálného parametru. Nechť $\tau(\Theta)$ je daná parametrická funkce taková, že její

$$\tau'(\Theta) = \frac{\partial \tau(\Theta)}{\partial \Theta}$$

existuje pro každé $\Theta \in \Theta$. **Nestranný odhad** $T(\underline{X})$ ($E(T) = \tau(\Theta)$) parametrické funkce $\tau(\Theta)$ nazveme **regulárním** právě když pro každé $\Theta \in \Theta$ platí:

$$\int_M T(\underline{x}) \cdot f'(\underline{x}, \Theta) dx = \frac{\partial}{\partial \Theta} \int_M T(\underline{x}) \cdot f(\underline{x}, \Theta) dx = \frac{\partial}{\partial \Theta} E(T(\underline{X})) = \frac{\partial \tau(\Theta)}{\partial \Theta} = \tau'(\Theta),$$

tzn. když jeho střední hodnotu $\left(E(T(\underline{X})) = \int_M T(\underline{x}) \cdot f(\underline{x}, \Theta) dx \right)$ můžeme derivovat podle Θ .

$$\begin{aligned} \int_M T(\underline{x}) \cdot f'(\underline{x}, \Theta) dx &= \int_M T(\underline{x}) \cdot \frac{f'(\underline{x}, \Theta)}{f(\underline{x}, \Theta)} \cdot f(\underline{x}, \Theta) dx = \int_M T(\underline{x}) \cdot \frac{f'(\underline{x}, \Theta)}{f(\underline{x}, \Theta)} dF(\underline{x}, \Theta) = \\ &= \int_M T(\underline{x}) \cdot \frac{\partial \ln f(\underline{x}, \Theta)}{\partial \Theta} dF(\underline{x}, \Theta) = \tau'(\Theta) \end{aligned}$$

Následující věta pak udává dolní mez rozptylu nestranného odhadu dané parametrické funkce $\tau(\Theta)$.

Rao – Cramerova věta

Nechť $\underline{X} = (X_1, \dots, X_n)$ je náhodným výběrem, který má regulární rozdělení $f(\underline{x}, \Theta)$. Nechť $\tau(\Theta)$ je daná parametrická funkce. Potom pro každý regulární nestranný odhad $T(\underline{X})$ funkce $\tau(\Theta)$ platí:

$$D(T(\underline{X})) \geq \frac{\{\tau'(\Theta)\}^2}{n \cdot I(\Theta)} = C(\Theta)$$

Odhad $T(\underline{X})$, jehož rozptyl je roven Rao – Cramerově dolní mezi rozptylu $C(\Theta)$, je efektivním odhadem.



Řešený příklad

Nechť $\underline{X} = (X_1, \dots, X_n)$ je náhodným výběrem z Poissonova rozdělení $Po(\lambda)$. Určete podle Rao – Cramerovy nerovnosti dolní mez rozptylu odhadu parametru λ .

Z předcházejícího řešeného příkladu víme, že Fisherova míra informace parametru λ je:

$$I(\lambda) = \frac{1}{\lambda}$$

Podle Rao – Cramerovy věty má každý regulární nestranný odhad parametru λ ($\tau(\lambda) = \lambda$) dolní mez rozptylu

$$C(\lambda) = \frac{\{\tau'(\lambda)\}^2}{n \cdot I(\lambda)} = \frac{1}{n \cdot \frac{1}{\lambda}} = \frac{\lambda}{n}$$

Rozptyl nestranného odhadu $T(\underline{X}) = \bar{X} = \frac{1}{n} \cdot \sum_{i=1}^n X_i$ je

$$D(T(\underline{X})) = D(\bar{X}) = \frac{\lambda}{n}$$

Pro parametr λ tedy existuje nestranný odhad s rozptylem rovnajícím se Rao – Cramerově dolní mezí.



Řešený příklad

Nechť $\underline{X} = (X_1, \dots, X_n)$ je náhodným výběrem z normálního rozdělení $N(\mu, \sigma^2)$, kde σ^2 známe. Určete pomocí Rao – Cramerovy nerovnosti dolní mez rozptylu odhadu parametru μ .

$$f(x, \mu) = \frac{1}{\sigma \cdot \sqrt{2\pi}} \cdot e^{-\frac{(x-\mu)^2}{2\sigma^2}}, \text{ kde } \sigma^2 \text{ známe}$$

$$\text{Chceme najít Rao-Cramerovu dolní mez rozptylu: } C(\mu) = \frac{\{\tau'(\mu)\}^2}{n \cdot I(\mu)}$$

Z předcházejících řešených příkladů víme, že Fisherova míra informace pro parametr μ je:

$$I(\mu) = \frac{1}{\sigma^2}$$

a příslušná parametrická funkce je: $\tau(\mu) = \mu$

Proto:

$$C(\mu) = \frac{\{\tau'(\mu)\}^2}{n \cdot I(\mu)} = \frac{1}{n \cdot \frac{1}{\sigma^2}} = \frac{\sigma^2}{n}$$

Víme, že nejlepším nestranným (efektivním) odhadem střední hodnoty μ je průměr $(T(\underline{X}) = \bar{X})$. Měl by tedy mít rozptyl roven Rao – Cramerově dolní mezi rozptylu.

Podle centrální limitní věty můžeme průměr aproximovat normálním rozdělením $N\left(\mu, \frac{\sigma^2}{n}\right)$, z čehož je zřejmé, že $D(T(\underline{X})) = \frac{\sigma^2}{n} = C(\mu)$. Čímž jsme dokázali že průměr je efektivním odhadem střední hodnoty normálního rozdělení.



Otázky 5.4.

1. Jak a proč určujeme dolní mez rozptylu nestranného odhadu?

5.5. Metoda momentů



Čas ke studiu: 15 minut



Cíl: Po prostudování tohoto odstavce budete umět:

- odhadovat parametry rozdělení pravděpodobnosti metodou momentů



Výklad

Pro odhad hodnot parametrů pravděpodobnostních rozdělení se nejčastěji používá **metoda maximální věrohodnosti** (maximum likelihood), nebo **metoda momentů**.

□ V čem spočívá princip metody momentů

Metoda momentů je principiálně jednoduchá metoda pro konstrukci bodových odhadů neznámých parametrů známých rozdělení, která spočívá v tom, že porovnávané výběrové momenty získaných dat s odpovídajícími teoretickými momenty předpokládaného rozdělení s hustotou $f(t)$. Metoda vede na řešení soustavy takového počtu rovnic, kolik je neznámých parametrů.

Máme-li k dispozici zaznamenaná data (náhodný výběr) $(t_1, \dots, t_n)^T$; pak:

k -tý výběrový obecný moment je dán vztahem:
$$M'_k = \frac{1}{n} \sum_{i=1}^n t_i^k$$

Podobně k -tý výběrový centrální moment:
$$M_k = \frac{1}{n} \sum_{i=1}^n (t_i - \bar{t})^k$$
, kde \bar{t} je průměr.

Odpovídající teoretické momenty jsou dány rovnicemi:

k -tý obecný moment:
$$\mu'_k = \int_0^{\infty} t^k f(t) dt$$

k -tý výběrový centrální moment:
$$\mu_k = \int_0^{\infty} (t - \mu'_1)^k f(t) dt$$

Metoda momentů:

Jestliže pravděpodobnostní rozdělení s hustotou $f(t)$ má r neznámých parametrů a jestliže soustava rovnic

$$\begin{aligned} M'_k &= \mu'_k, \quad k = 1, \dots, r \\ \text{resp.} \\ M_k &= \mu_k, \quad k = 1, \dots, r \end{aligned}$$

má jediné řešení, pak dává metoda momentů jednoznačně určené odhady r parametrů.



Řešený příklad

Je dán náhodný výběr $(t_1, \dots, t_n)^T$. Předpokládáme, že jde o výběr z exponenciálního rozdělení $E(\lambda)$. Metodou momentů odhadněte neznámý parametr λ .

Označme si hledaný odhad $\tilde{\lambda}$. $\tilde{\lambda}$ získáme jako řešení rovnice: $\mu'_1 = M'_1$

Hustota exponenciálního rozdělení je $f(t) = \lambda \cdot e^{-\lambda t}$, proto:

$$\begin{aligned} \mu'_1 &= \int_0^{\infty} t \lambda \cdot e^{-\lambda t} dt = \lambda \cdot \int_0^{\infty} t \cdot e^{-\lambda t} dt = \left[\begin{array}{l} u(t) = t \quad v'(t) = e^{-\lambda t} \\ u'(t) = 1 \quad v(t) = -\frac{1}{\lambda} \cdot e^{-\lambda t} \end{array} \right] = \lambda \cdot \left\{ \left[-\frac{1}{\lambda} \cdot t \cdot e^{-\lambda t} \right]_0^{\infty} + \frac{1}{\lambda} \int_0^{\infty} e^{-\lambda t} dt \right\} \\ &= \lambda \cdot \left\{ \left[-\frac{1}{\lambda} \cdot t \cdot e^{-\lambda t} \right]_0^{\infty} + \left[-\frac{1}{\lambda^2} \cdot e^{-\lambda t} \right]_0^{\infty} \right\} = \lambda \cdot \left\{ \left[-\frac{1}{\lambda} \cdot e^{-\lambda t} \left(t + \frac{1}{\lambda} \right) \right]_0^{\infty} \right\} = \\ &= \lambda \cdot \left\{ \lim_{t \rightarrow \infty} \left[-\frac{1}{\lambda} \cdot e^{-\lambda t} \left(t + \frac{1}{\lambda} \right) \right] + \frac{1}{\lambda^2} \right\} = \lambda \cdot \left\{ -\frac{1}{\lambda} \cdot \lim_{t \rightarrow \infty} \frac{t + \frac{1}{\lambda}}{e^{\lambda t}} + \frac{1}{\lambda^2} \right\} = -\lim_{t \rightarrow \infty} \frac{1}{\lambda \cdot e^{\lambda t}} + \frac{1}{\lambda} = \end{aligned}$$

$$M'_1 = \frac{\sum_{i=1}^n t_i}{n}.$$

Rovnice $\mu'_1 = M'_1$ přechází na rovnici: $\frac{1}{\lambda} = \frac{\sum_{i=1}^n t_i}{n}$ neboli $\tilde{\lambda} = \frac{n}{\sum_{i=1}^n t_i}$,

což je odhad neznámého parametru λ získaný metodou momentů.



Shrnutí kapitoly 5.5.

Metoda momentů je principiálně jednoduchá metoda pro konstrukci odhadů neznámých parametrů známých rozdělení, která spočívá v tom, že porovnáváme výběrové momenty získaných dat s odpovídajícími teoretickými momenty předpokládaného rozdělení s hustotou $f(t)$. Metoda vede na řešení soustavy takového počtu rovnic, kolik je neznámých parametrů.



Úlohy k řešení 5.5.

1. Necht' turbína elektrárny podléhá náhodným šokům, které splňují předpoklady Poissonových pokusů. Necht' při každém pátém šoku dojde k závažné poruše turbíny. Během dlouhodobého

sledování byly zaznamenány následující doby do poruch turbíny (v hodinách): (1020, 1100, 960, 1500, 1450, 1320, 1255, 1165, 1385, 1410).

- Určete pravděpodobnostní rozdělení pro dobu do poruchy turbíny.
- Určete odhad neznámého parametru zjištěného rozdělení metodou momentů.
- Určete hazardní funkci turbíny.
- Určete, ve které fázi svého životního cyklu se turbína nachází.

5.6. Metoda maximální věrohodnosti



Čas ke studiu: 15 minut



Cíl: Po prostudování tohoto odstavce budete umět:

- odhadovat parametry rozdělení pravděpodobnosti metodou maximální věrohodnosti



Výklad

□ Na čem je založena metoda maximální věrohodnosti

Odhady získané touto metodou se všeobecně vyznačují dobrými statistickými vlastnostmi.

Nechť $(t_1, \dots, t_n)^T$ je náhodný výběr z rozdělení s hustotou $f(t; \Theta)$, kde Θ je neznámý parametr. Naším problémem bude nalézt funkci (zvanou funkce věrohodnosti) danou

$$L(t_1, \dots, t_n; \Theta) = f(t_1; \Theta) \cdot f(t_2; \Theta) \dots f(t_n; \Theta) = \prod_{i=1}^n f(t_i; \Theta)$$

a z ní pak získat $\hat{\Theta}$ tak, aby $\hat{\Theta} = \hat{\Theta}(t_1, \dots, t_n)$ bylo co nejlepším odhadem pro Θ . Pravá strana rovnice je sdružená hustota pravděpodobnosti n -nezávislých proměnných (t_1, \dots, t_n) se stejným rozdělením.

Jelikož L je jednoduše funkcí neznámého parametru Θ , který je odhadován, metoda maximální věrohodnosti je založena na získání takové hodnoty Θ , která maximalizuje L .

Při praktických výpočtech se ukázalo jako výhodnější maximalizovat spíše funkci $\ln L$ namísto L , což je možné proto, že obě tyto operace jsou ekvivalentní a dávají stejné výsledky.

Podmínkou optimality je tedy rovnice:

$$\frac{\partial \ln L(t_1, \dots, t_n; \Theta)}{\partial \Theta} = 0$$

a hodnota parametru získaná z této podmínky se nazývá **maximálně věrohodný odhad parametru θ** .



Řešený příklad

Je dán náhodný výběr $(t_1, \dots, t_n)^T$. Předpokládáme, že jde o výběr z exponenciálního rozdělení $E(\lambda)$. Metodou maximální věrohodnosti odhadněte neznámý parametr λ .

Označme si hledaný odhad $\hat{\lambda}$.

Hustota exponenciálního rozdělení je $f(t) = \lambda \cdot e^{-\lambda t}$, proto funkce věrohodnosti pak bude dána výrazem:

$$L(t_1, \dots, t_n; \lambda) = (\lambda \cdot e^{-\lambda t_1})(\lambda \cdot e^{-\lambda t_2}) \dots (\lambda \cdot e^{-\lambda t_n}) = \lambda^n \cdot e^{-\lambda \sum_{i=1}^n t_i}$$

Logaritmováním získáme

$$\ln L(t_1, \dots, t_n; \lambda) = \ln \left(\lambda^n \cdot e^{-\lambda \cdot \sum_{i=1}^n t_i} \right) = \ln(\lambda^n) + \ln \left(e^{-\lambda \cdot \sum_{i=1}^n t_i} \right) = n \cdot \ln \lambda - \lambda \cdot \sum_{i=1}^n t_i$$

Zbývá vyřešit podmínku optimality:

$$\begin{aligned} \frac{\partial \ln L(t_1, \dots, t_n; \lambda)}{\partial \lambda} &= 0 \\ \frac{\partial \left(n \cdot \ln \lambda - \lambda \cdot \sum_{i=1}^n t_i \right)}{\partial \lambda} &= 0 \\ n \cdot \frac{1}{\lambda} - \sum_{i=1}^n t_i &= 0 \\ \hat{\lambda} &= \frac{n}{\sum_{i=1}^n t_i} \end{aligned}$$

Získali jsme maximálně věrohodný odhad parametru $\hat{\lambda}$.



Řešený příklad

Uvažujme dvouparametrické Weibullovo rozdělení s hustotou $f(t) = \frac{\beta}{\Theta} t^{\beta-1} e^{-\frac{t^\beta}{\Theta}}$.

Metodou maximální věrohodnosti odhadněte parametry β a Θ .

Funkce věrohodnosti L je dána

$$L(t_1, \dots, t_n; \beta, \Theta) = \frac{\beta}{\Theta} t_1^{\beta-1} e^{-\frac{t_1^\beta}{\Theta}} \dots \frac{\beta}{\Theta} t_n^{\beta-1} e^{-\frac{t_n^\beta}{\Theta}} = \left(\frac{\beta}{\Theta} \right)^n \prod_{i=1}^n t_i^{\beta-1} e^{-\frac{1}{\Theta} \sum_{j=1}^n t_j^\beta}$$

Logaritmováním získáme:

$$\begin{aligned} \ln L &= \ln \left(\frac{\beta}{\Theta} \right)^n + \ln \prod_{i=1}^n t_i^{\beta-1} + \ln e^{-\frac{1}{\Theta} \sum_{j=1}^n t_j^\beta} = n \ln \beta - n \ln \Theta + \sum_{i=1}^n \ln t_i^{(\beta-1)} - \frac{1}{\Theta} \sum_{i=1}^n t_i^\beta = \\ &= n \ln \beta - n \ln \Theta + (\beta - 1) \sum_{i=1}^n \ln t_i - \frac{1}{\Theta} \sum_{i=1}^n t_i^\beta \end{aligned}$$

Optimalizaci však provádíme s ohledem na oba neznámé parametry α, β , takže podmínka

optimality přechází v tomto případě na dvě následující rovnice:

$$\frac{\partial \ln L}{\partial \beta} = \frac{n}{\beta} + \sum_{i=1}^n \ln t_i - \frac{1}{\Theta} \sum_{i=1}^n t_i^\beta \ln t_i = 0$$

$$\frac{\partial \ln L}{\partial \Theta} = -\frac{n}{\Theta} + \frac{1}{\Theta^2} \sum_{i=1}^n t_i^\beta = 0$$

Z druhé rovnice můžeme snadno získat

$$\Theta = \frac{\sum_{i=1}^n t_i^\beta}{n}$$

zatímco z první rovnice dostaneme

$$\Theta = \frac{\sum_{i=1}^n t_i^\beta \ln t_i}{\frac{n}{\beta} + \sum_{i=1}^n \ln t_i}$$

Porovnáním pravých stran posledních dvou rovnic získáme jednu rovnici pro jednu neznámou β . Řešení je nutno provést numericky volbou vhodného iteračního procesu.



Shrnutí kapitoly 5.6.

Metoda maximální věrohodnosti je principiálně jednoduchá metoda pro konstrukci odhadů neznámých parametrů známých rozdělení pravděpodobnosti, která je založena na maximalizaci **věrohodnostní funkce**, což je sdružená hustota pravděpodobnosti daného náhodného výběru, brána ovšem jako funkce neznámých parametrů.



Úlohy k řešení 5.6.

- Doba do poruchy dieselgenerátoru se řídí exponenciálním rozdělením pravděpodobnosti. Během dlouhodobého sledování byly zaznamenány následující poruchové doby v hodinách: (150, 190, 165, 177, 203, 178, 162, 181, 194, 168).
 - Odhadněte parametr λ metodou maximální věrohodnosti,
 - Charakterizujte hazardní funkci dieselgenerátoru,
 - Odhadněte funkci bezporuchovosti v čase $t=100$ hodin,
 - Určete 90% -ní život dieselgenerátoru (zaručenou dobu bezporuchového provozu, tj. dobu do poruchy, která bude překročena s 90% pravděpodobností).
- Nechť turbína elektrárny podléhá náhodným šokům, které splňují předpoklady Poissonových pokusů. Nechť při každém pátém šoku dojde k závažné poruše turbíny. Během dlouhodobého sledování byly zaznamenány následující doby do poruch turbíny (v hodinách): (1020, 1100, 960, 1500, 1450, 1320, 1255, 1165, 1385, 1410).
 - Určete pravděpodobnostní rozdělení pro dobu do poruchy turbíny,
 - Určete odhad neznámého parametru zjištěného rozdělení metodou momentů,

- c) Určete hazardní funkci turbíny,
 d) Určete, ve které fázi svého životního cyklu se turbína nachází.



KLÍČ K ŘEŠENÍ

1a) $L(t_1, \dots, t_n; \lambda) = (\lambda \cdot e^{-\lambda t_1})(\lambda \cdot e^{-\lambda t_2}) \dots (\lambda \cdot e^{-\lambda t_n}) = \lambda^n e^{-\lambda \sum_{i=1}^n t_i}$ dává odhad parametru λ následovně:

$$\hat{\lambda} = \frac{n}{\sum_{i=1}^n t_i}, \text{ což po dosazení zadaných hodnot je } \hat{\lambda} = \frac{10}{1768} \cong 0,005656.$$

b) Hazardní funkce je konstantní, $\lambda(t) = 0,005656$, dieselgenerátor je tedy v období stabilního života.

c) Funkce bezporuchovosti v čase 100 hodin je: $R(t) = 1 - F(t) = e^{-\lambda t}$;
 $R(100) = e^{-0,5656} \cong 0,568$

d) Hledanou dobu určíme řešením rovnice: $P(X > T_{0,9}) = 0,9$, tj. $1 - F(T_{0,9}) = 0,9$;

$$T_{0,9} = 18,6 \text{ hodin}$$

2a) Doba do poruchy se řídí Gamma rozdělením s hustotou $f(t) = \frac{\lambda^5}{\Gamma(5)} \cdot t^4 \cdot e^{-\lambda t}$, s neznámým parametrem λ

b) λ odhadneme metodou momentů:

Rovnice $\mu'_1 = M'_1$ přechází na rovnici

$$\frac{5}{\lambda} = \frac{\sum_{i=1}^{10} t_i}{10} \quad \text{neboli} \quad \tilde{\lambda} = \frac{50}{\sum_{i=1}^{10} t_i} \cong 0,004, \text{ což je odhad neznámého parametru } \lambda$$

získaný metodou momentů.

c) Hazardní funkce je:

$$\lambda(t) = \frac{0,004}{24 \sum_{i=0}^4 \frac{1}{(4-i)! (0,004 t)^i}}$$

d) Turbína se nachází ve třetí fázi svého životního cyklu, tj. v období poruch v důsledku stárnutí a opotřebení.

$$Y_i = 0 \quad \text{pro } X_i \geq 1$$

$$Y_i = 1 \quad \text{pro } X_i = 0 \quad i = 1, 2, \dots, n$$

6. NEÚPLNÁ DATA



Čas ke studiu: 1,5 hodiny



Cíl: Po prostudování tohoto odstavce budete umět:

- Seznámíte se s různými typy cenzorování a naučíte se zapisovat výsledky zkoušek při těchto výběrových plánech. Ukážeme si použití metody maximální věrohodnosti pro neúplné výběry.



Výklad

6.1. Výběrové plány

Začneme-li za účelem zjištění charakteristik spolehlivosti v čase $t=0$ pozorovat určitý systém složený z n prvků stejného typu (majících stejné rozdělení doby do poruchy), **klasická statistická situace** nastává, jestliže pozorování provádíme dokud se všechny prvky neporouchají. Výsledkem takového experimentu je tzv. **úplný výběr X_1, \dots, X_n dob do poruchy**, tj. standardní náhodný výběr.

V praxi (zkoušky životnosti složitých systémů, klinické zkoušky, pojišťovnictví) se však často stává, že experiment je analyzován dříve než dojde k poruše všech jeho prvků. K předčasnému ukončení experimentu vedou většinou ekonomické a časové důvody (dlouhá doba do poruchy u některých prvků, resp. věcné důvody (předem stanovené termíny ...), při klinických pokusech může dojít např. k tomu, že pacient přestane spolupracovat (odstěhuje se, zemře z jiných než sledovaných příčin, ...). V těchto případech máme k dispozici pouze tzv. **neúplné výběry**.

V této kapitole se seznámíme se základními uspořádáními experimentu vedoucími k neúplným výběrům. V mnohém jsme čerpali z [Hurt, Teorie spolehlivosti, Praha 1984].

Mějme systém složený z n identických prvků. Necht' X_1, \dots, X_n jsou doby do poruchy jednotlivých prvků. Mluvíme-li o neúplných výběrech, znamená to, že ne všechny X_i (doby do poruchy) jsou opravdu pozorované.

V praxi se vyskytují neúplné výběry buď v podobě useknutých dat nebo cenzorovaných dat.

Useknutá data - víme, že X_i nad (resp. pod) určitým limitem se zcela ztrácí, avšak nejsme informováni o této ztrátě.

Cenzorovaná data - mimo naměřené X_i získáme i částečnou informaci o „špatně měřitelných hodnotách“.

□ Cenzorování I. typu (cenzorování časem)

Ke ztrátě dat dochází v tomto případě proto, že doba do poruchy některých prvků překročí dobu experimentu. Doba experimentu T (**časový cenzor**) je stanovena předem. Počet skutečně pozorovaných poruch je náhodná veličina, která může nabývat hodnot $0, 1, \dots, n$. Necht' $X_{(1)}, \dots, X_{(n)}$ označuje uspořádaný náhodný výběr X_1, \dots, X_n . Došlo-li během doby T k poruše r prvků, pak

výsledkem experimentu je prvních r hodnot pořádkových statistik $X_{(1)} \leq \dots \leq X_{(r)} \leq T$ a informace o tom, že $X_{(r+1)} > T$.

□ Cenzorování II. typu (cenzorování poruchou)

V tomto případě si předem definuje ukončení experimentu počtem prvků u nichž dojde k poruše (r). Na začátku experimentu si stanovíme přirozené číslo r ($r \leq n$) a v okamžiku $t=0$ zahájíme pozorování. Experiment ukončíme ve chvíli, kdy dojde k poruše r -tého prvku. **Výsledkem experimentu** je potom prvních r hodnot pořádkových statistik $X_{(1)} \leq \dots \leq X_{(r)}$. Doba trvání experimentu (doba do poruch r -tého prvku) je náhodná veličina $X_{(r)}$.

□ Náhodné cenzorování

Při hodnocení spolehlivosti složitých systému se většinou nedaří uspořádat experiment podle představ statistika a tak musíme mnohdy využít **provozní data** (tj. pozorování ze skutečného provozu). Často se časové cenzory jednotlivých prvků liší (ukončování pozorování u jednotlivých prvků bývá mnohdy náhodné). Mluvíme o náhodném cenzorování. Označíme-li T_1, \dots, T_n doby, v nichž bylo ukončeno pozorování 1. až n -tého prvku systému, nabízí se považovat data za cenzorována časovým cenzorem T , kde $T = \min \{T_1, \dots, T_n\}$. To se však ukázalo jako ekonomicky neúnosné. Výhodnější je vybudovat pro náhodné cenzorování matematický aparát.

Nechť X je náhodná veličina reprezentující dobu do poruchy a T je náhodná veličina reprezentující časový cenzor. U každého prvku pak pozorujeme buď X nebo T podle toho, zda dříve nastala porucha nebo zda bylo dříve sledování prvku ukončeno. **Výsledkem experimentu** je pak n dvojic $(W_1, I_1) \dots (W_n, I_n)$, kde

$$j = 1, \dots, n :$$

$$W_j = \min \{X_j, T_j\}$$

$$I_j = 1 \Leftrightarrow W_j = X_j$$

(byla pozorována porucha j -tého prvku, tj. j -té pozorování je necenzorované)

$$I_j = 0 \Leftrightarrow W_j = T_j$$

(nebyla pozorována porucha j -tého prvku, pozorování prvku bylo ukončeno v čase T_j ($T_j < X_j$), tj. j -té pozorování je cenzorované)

Výsledkem experimentu tedy je výběr z dvourozměrného rozdělení.



Řešený příklad

Máte k dispozici výsledky pozorování 6-ti prvků. Jejich doby do poruchy X_j jsou: 4, 2, 8, 6, 1, 9. Zapište výsledek experimentu

- při cenzorování I. typu s časovým cenzorem $T=5$
- při cenzorování II. typu se stanoveným $r=4$
- při náhodném cenzorování s časovými cenzory 3, 3, 4, 8, 3, 6

ada) Výsledkem experimentu je uspořádaný výběr dob do poruchy prvků, jejichž doba do poruchy není větší než časový cenzor, tj. $\{1, 2, 4\}$ a informace, že zbylé 3 prvky se porouchaly později než v okamžiku $T=5$.

adb) Výsledkem experimentu je uspořádaný výběr dob do poruchy prvních 4 prvků:
 $\{1,2,4,6\}$

adc) Výsledkem experimentu jsou dvojice (W_j, I_j)

X_j	4	2	8	6	1	9
T_j	3	3	4	8	3	6
W_j	3	2	4	6	1	6
I_j	0	1	0	1	1	0

Zkráceně se cenzorovaná pozorování označují znaménkem + : $\{3+,2,4+,6,1,6+\}$

Mnohý statistický software však cenzorovaná data označuje znaménkem - :

$\{-3,2,-4,6,1,-6\}$.

Dosud jsme se zabývali situací, kdy pozorování začínáme provádět v čase $t=0$ a až do okamžiku cenzorování můžeme získat skutečné doby do poruchy. O cenzorovaných datech pak víme, že jsou vyšší než doba experimentu. Mluvíme o datech (pozorováních) **cenzorovaných zprava**. Stává se však také, že nezachytíme údaje o poruchách, které nastaly před začátkem měření – získaná data jsou pak **cenzorovaná zleva** (nemusí jít pouze o poruchy, může jít také např. o koncentrace látky pod detekční hranicí, apod.). Jsou-li data cenzorovaná zprava i zleva, mluvíme o **dvojném cenzorování**, resp. o **intervalovém cenzorování**.

Nechť X_1, \dots, X_n jsou doby do poruchy, T_1, \dots, T_n časové cenzory cenzorování zprava a $L_1, \dots, L_n, L_i < T_i, i=1, \dots, n$ cenzory cenzorování zleva. Interval $\langle L_i, T_i \rangle$ je pak časovým intervalem pro i -té pozorování. Padne-li X_i mimo tento interval, jeho skutečnou hodnotu nemůžeme pozorovat. **Výsledkem experimentu při dvojném cenzorování** jsou dvojice $(Z_i, J_i), i=1, \dots, n$, kde

$$Z_i = \max \{ \min \{ X_i, T_i \}, L_i \}$$

$$J_i = 0 \Leftrightarrow Z_i = L_i \quad (\text{porucha nastala před začátkem pozorování, cenzorování zleva})$$

$$J_i = 1 \Leftrightarrow Z_i = X_i \quad (\text{porucha nastala v } \langle L_i, T_i \rangle, \text{ necenzurovaná hodnota})$$

$$J_i = 2 \Leftrightarrow Z_i = T_i \quad (\text{porucha nastala po konci pozorování, cenzorování zprava})$$

6.2. Zrychlené zkoušky životnosti

Ani cenzorování nemusí zaručit získání dat k dostatečně přesnému odhadu charakteristik vysoce spolehlivých systémů. Čím je prvek spolehlivější, tím je obtížnější měřit jeho spolehlivost. Jednou z možností, jak data pro vysoce spolehlivé prvky získat jsou **zrychlené zkoušky životnosti**.

Jejich myšlenka spočívá v tom, že sledované prvky vystavíme zatížení vyššímu, než v jakém pracují v běžném pracovním režimu. Musíme přitom znát vztah mezi odhadovanými parametry doby do poruchy a úrovní zatížení. (V praxi bývá nalezení těchto vztahů obtížné, je nutné vycházet z fyzikálních principů fungování prvku.)

Modelový příklad:

Nechť doba do poruchy kondenzátoru má exponenciální rozdělení s parametrem λ ($X \rightarrow E(\lambda)$). Víme, že pro některé typy kondenzátoru je

$$\lambda = \lambda(U) = \frac{U^P}{C},$$

kde $C > 0$ a P jsou neznámé konstanty a U je napětí. S rostoucím napětím λ roste (střední doba do poruchy klesá), proto můžeme při vyšších napětích pozorovat skutečné doby do poruchy a na jejich základě odhadnout parametry C a P . Potom můžeme odhadnout λ při požadovaném napětí U_0 .

Modelový příklad:

Nechť doba do poruchy polovodiče má exponenciální rozdělení s parametrem λ ($X \rightarrow E(\lambda)$). Víme, že pro polovodiče je

$$\lambda = \lambda(t) = e^{-\frac{B}{t-A}},$$

kde A, B jsou neznámé konstanty a t je teplota. Při vyšších teplotách je vyšší λ (střední doba do poruchy je nižší), a proto je při vyšších teplotách možné odhadnout hodnoty parametrů A a B . Potom můžeme odhadnout λ při požadované teplotě t_0 .

6.3. Metoda maximální věrohodnosti pro neúplné výběry

Připomeňme si nejdříve metodu maximální věrohodnosti v její klasické podobě. Metoda maximální věrohodnosti je principiálně jednoduchá metoda pro konstrukci odhadů neznámých parametrů známých rozdělení pravděpodobnosti, která je založena na maximalizaci věrohodnostní funkce, což je sdružená hustota pravděpodobnosti daného náhodného výběru, brána ovšem jako funkce neznámých parametrů.

Nechť $(t_1, \dots, t_n)^T$ je náhodný výběr z rozdělení s hustotou $f(t; \Theta)$, kde Θ je neznámý parametr. Věrohodnostní funkce je

$$L(t_1, \dots, t_n; \Theta) = f(t_1; \Theta) \cdot f(t_2; \Theta) \dots f(t_n; \Theta) = \prod_{i=1}^n f(t_i; \Theta)$$

Při praktických výpočtech se ukázalo jako výhodnější maximalizovat spíše funkci $\ln L$ namísto L , což je možné proto, že obě tyto operace jsou ekvivalentní a dávají stejné výsledky.

Podmínkou optimality je tedy rovnice:

$$\frac{\partial \ln L(t_1, \dots, t_n; \Theta)}{\partial \Theta} = 0$$

a hodnota parametru získaná z této podmínky se nazývá maximálně věrohodný odhad parametru θ .

□ Věrohodnostní funkce pro cenzorování II. typu

Začneme s odvozováním věrohodnostní funkce pro nejjednodušší případ – cenzorování II. typu (cenzorování poruchou, r předem dané). V tomto případě jde o výpočet hustoty prvních r pořádkových statistik.

Nechť $(X_{(1)}, \dots, X_{(r)})^T = \underline{X}$ jsou pozorované uspořádané doby do poruchy a necht' $0 < x_{(1)} < \dots < x_{(r)}$. Označme $(x_{(1)}, \dots, x_{(r)})^T = \underline{x}$. Necht' $\Delta > 0$ je takové, že $x_{(i)} + \Delta < x_{(i+1)}$ pro $i=1, \dots, r-1$ ($x_{(0)}=0$). Necht' $\underline{\Delta}$ je r -rozměrný vektor, který má všechny složky rovny Δ , a necht'

$$E = \{ \underline{x} < \underline{X} < \underline{x} + \underline{\Delta} \}.$$

Náhodný jev E tedy nastane právě když žádné pozorování není menší než $x_{(1)}$, právě jedno pozorování padne do intervalu $\langle x_{(1)}, x_{(1)} + \Delta \rangle$, žádné pozorování nepadne do intervalu $\langle x_{(1)} + \Delta, x_{(2)} \rangle$, právě jedno pozorování padne do intervalu $\langle x_{(2)}, x_{(2)} + \Delta \rangle$, ..., právě jedno pozorování padne do intervalu $\langle x_{(r)}, x_{(r)} + \Delta \rangle$ a $(n-r)$ pozorování je větších nebo rovných $x_{(r)} + \Delta$.

Proto

$$P(E) = \frac{n!}{0! \cdot 1! \cdot 0! \cdot \dots \cdot 1! \cdot (n-r)!} \cdot \prod_{i=1}^r (F(x_{(i)} + \Delta) - F(x_{(i)})) \cdot R^{n-r}(x_{(r)} + \Delta),$$

kde $R(x) = 1 - F(x)$ je funkce spolehlivosti.

Věrohodnostní funkce \underline{X} při cenzorování II. typu potom je

$$L(\underline{x}) = \lim_{\Delta \rightarrow 0} \frac{P(E)}{\Delta^r} = \frac{n!}{(n-r)!} \cdot \prod_{i=1}^r f(x_{(i)}) \cdot R^{n-r}(x_{(r)})$$

□ Věrohodnostní funkce pro cenzorování I. typu

Nalezení této věrohodnostní funkce je o něco složitější. Výsledkem experimentu je prvních r pořádkových statistik $X_{(1)} \leq \dots \leq X_{(r)} \leq T$ a informace o tom, že $X_{(r+1)} > T$, ..., $X_{(n)} > T$. Značení použijeme stejné jako v předcházejícím případě. Nechť $x_{(r)} + \Delta < T$. Podobně jako při cenzorování II. typu zjistíme, že pravděpodobnost, že žádné pozorování není menší než $x_{(1)}$, právě jedno pozorování padne do intervalu $\langle x_{(1)}, x_{(1)} + \Delta \rangle$, žádné pozorování nepadne do intervalu $\langle x_{(1)} + \Delta, x_{(2)} \rangle$, právě jedno pozorování padne do intervalu $\langle x_{(2)}, x_{(2)} + \Delta \rangle$, ..., právě jedno pozorování padne do intervalu $\langle x_{(r)}, x_{(r)} + \Delta \rangle$ a $(n-r)$ pozorování je větších než T , je

$$P(\underline{x} < \underline{X} < \underline{x} + \underline{\Delta}, X_{(r+1)} > T, \dots, X_{(n)} > T) = \frac{n!}{0! \cdot 1! \cdot 0! \cdot \dots \cdot 1! \cdot (n-r)!} \cdot \prod_{i=1}^r (F(x_{(i)} + \Delta) - F(x_{(i)})) \cdot R^{n-r}(T)$$

Sdružené rozdělení výsledku experimentu při cenzorování I. typu má věrohodnostní funkci

$$L(\underline{x}, r) = \frac{n!}{(n-r)!} \cdot \prod_{i=1}^r f(x_{(i)}) \cdot R^{n-r}(T)$$

$$\text{pro } 0 < x_{(1)} < \dots < x_{(r)} < T, \quad r = 0, 1, \dots, n$$

□ Věrohodnostní funkce pro náhodné cenzorování

K předpokladům uvedeným při hledání sdružené hustoty při cenzorování II. a I. typu předpokládejme, že časový cenzor T je náhodná veličina, která má spojité rozdělení s distribuční funkcí $G(x)$ a hustotou $g(x)$, obecně také závislé na neznámých parametrech. Předpokládejme, že X a I jsou nezávislé

náhodné veličiny. Odvodíme nejdříve rozdělení náhodného vektoru (W, I) , kde $W = \min \{X, T\}$ a $I = I(X < T)$, tj. I je indikátor jevu $\{X < T\}$. Pro $w > 0$ dostáváme

$$P(W < w, I = 1) = P(X < w, X < T) = \iint_{x < w, x < t} dF(x)dG(t) = \int_0^w (1 - G(x))dF(x) = F(w) - \int_0^w f(x)G(x)dx$$

Analogicky

$$P(W < w, I = 0) = G(w) - \int_0^w g(x)F(x)dx$$

Derivováním výše uvedených pravděpodobností podle w dostaneme **hustotu pravděpodobnosti náhodného vektoru (W, I)** .

$$h(w, i) = f(w) \cdot (1 - G(w)), \quad w > 0, \quad i = 1$$

$$h(w, i) = g(w) \cdot (1 - F(w)), \quad w > 0, \quad i = 0$$

$$h(w, i) = 0 \quad \text{jinak}$$

Věrohodností funkce při náhodném cenzorování je

$$L(\Theta) = \prod_{j=1}^n h(W_j, I_j)$$

Poměrně často se setkáváme se situací, kdy **doba do poruchy X a časový cenzor T neobsahují společné neznámé parametry**. Necht' Θ má nyní význam k -rozměrného vektoru neznámých parametrů obou rozdělení F a G , $\Theta = (\underline{\Theta}_1, \underline{\Theta}_2)$, kde $\underline{\Theta}_1$ je vektor neznámých parametrů rozdělení F a $\underline{\Theta}_2$ je vektor neznámých parametrů rozdělení G . Definujme množiny U , resp. C , kde U , resp. C , je množina indexů necenzorovaných, resp. cenzorovaných, pozorování.

$$U = \{j : I_j = 1\} \quad C = \{j : I_j = 0\}$$

Věrohodností funkce má potom tvar:

$$L(\underline{\Theta}) = \prod_{j \in U} f(X_j) \cdot \prod_{j \in U} (1 - G(X_j)) \cdot \prod_{j \in C} g(T_j) \cdot \prod_{j \in C} (1 - F(T_j))$$

Jestliže rozdělení doby do poruchy X a časového cenzoru T neobsahují společné parametry a neexistuje funkční závislost mezi $\underline{\Theta}_1$ a $\underline{\Theta}_2$, můžeme získat maximálně věrohodný odhad maximalizací věrohodností funkce

$$L(\underline{\Theta}_1) = \prod_{j \in U} f(X_j) \cdot \prod_{j \in C} (1 - F(T_j))$$

Takto získané odhady nezávisí na rozdělení G , avšak jejich rozdělení obecně na tomto rozdělení závisí.



Řešený příklad

Nechť doba do poruchy má exponenciální rozdělení $X \rightarrow E(\lambda)$. Určete maximálně věrohodný odhad parametru λ na základě

- úplného výběru x_1, \dots, x_n
- cenzorování I. typu
- cenzorování II. typu
- náhodného cenzorování

Označme si hledaný odhad $\hat{\lambda}$.

ada) Je dán náhodný výběr $(x_1, \dots, x_n)^T$.

Hustota exponenciálního rozdělení je $f(x) = \lambda \cdot e^{-\lambda x}$, proto funkce věrohodnosti pak bude dána výrazem:

$$L(x_1, \dots, x_n; \lambda) = (\lambda \cdot e^{-\lambda x_1})(\lambda \cdot e^{-\lambda x_2}) \dots (\lambda \cdot e^{-\lambda x_n}) = \lambda^n \cdot e^{-\lambda \sum_{i=1}^n x_i}$$

Logaritmováním získáme

$$\ln L(x_1, \dots, x_n; \lambda) = \ln \left(\lambda^n \cdot e^{-\lambda \cdot \sum_{i=1}^n x_i} \right) = \ln(\lambda^n) + \ln \left(e^{-\lambda \cdot \sum_{i=1}^n x_i} \right) = n \cdot \ln \lambda - \lambda \cdot \sum_{i=1}^n x_i$$

Zbývá vyřešit podmínku optimality:

$$\begin{aligned} \frac{\partial \ln L(x_1, \dots, x_n; \lambda)}{\partial \lambda} &= 0 \\ \frac{\partial \left(n \cdot \ln \lambda - \lambda \cdot \sum_{i=1}^n x_i \right)}{\partial \lambda} &= 0 \\ n \cdot \frac{1}{\lambda} - \sum_{i=1}^n x_i &= 0 \\ \hat{\lambda} &= \frac{n}{\sum_{i=1}^n x_i} = \frac{1}{\bar{X}} \end{aligned}$$

Maximálně věrohodný odhad parametru λ na základě úplného výběru x_1, \dots, x_n je $\hat{\lambda} = \frac{1}{\bar{X}}$

adb) Při cenzorování I. typu je věrohodnostní funkce

$$\begin{aligned} L(\underline{x}, r) &= \frac{n!}{(n-r)!} \cdot \prod_{i=1}^r f(x_{(i)}) \cdot R^{n-r}(T) \\ L(\underline{x}, \lambda, r) &= \frac{n!}{(n-r)!} \cdot \prod_{i=1}^r \lambda \cdot e^{-\lambda x_{(i)}} \cdot (e^{-\lambda T})^{n-r} \end{aligned}$$

Logaritmus funkce věrohodnosti tedy bude dán výrazem

$$\begin{aligned}\ln L(\underline{x}, \lambda, r) &= \ln \frac{n!}{(n-r)!} + \sum_{i=1}^r \ln (\lambda \cdot e^{-\lambda \cdot x_{(i)}}) + \ln (e^{-\lambda \cdot T})^{n-r} = \\ &= \ln \frac{n!}{(n-r)!} + \sum_{i=1}^r \ln \lambda - \lambda \cdot \sum_{i=1}^r x_{(i)} - (n-r) \cdot \lambda \cdot T = \\ &= \ln \frac{n!}{(n-r)!} + r \cdot \ln \lambda - \lambda \cdot \sum_{i=1}^r x_{(i)} - (n-r) \cdot \lambda \cdot T\end{aligned}$$

Podmínky optimality zapišeme jako

$$\frac{\partial \ln L(\underline{x}, \lambda, r)}{\partial \lambda} = 0$$

$$\frac{r}{\lambda} - \sum_{i=1}^r x_{(i)} - (n-r) \cdot T = 0$$

Maximálně věrohodný odhad při cenzorování I. typu je tedy

$$\hat{\lambda} = \frac{r}{\sum_{i=1}^r x_{(i)} + (n-r) \cdot T}$$

adc) Při cenzorování II. typu je věrohodnostní funkce

$$L(\underline{x}, \lambda) = \frac{n!}{(n-r)!} \cdot \prod_{i=1}^r f(x_{(i)}) \cdot R^{n-r}(x_{(r)})$$

$$L(\underline{x}, \lambda) = \frac{n!}{(n-r)!} \cdot \prod_{i=1}^r \lambda \cdot e^{-\lambda \cdot x_{(i)}} \cdot (e^{-\lambda \cdot x_{(r)}})^{n-r}$$

Logaritmus funkce věrohodnosti tedy bude dán výrazem

$$\begin{aligned}\ln L(\underline{x}, \lambda) &= \ln \frac{n!}{(n-r)!} + \sum_{i=1}^r \ln (\lambda \cdot e^{-\lambda \cdot x_{(i)}}) + \ln (e^{-\lambda \cdot x_{(r)}})^{n-r} = \\ &= \ln \frac{n!}{(n-r)!} + \sum_{i=1}^r \ln \lambda - \lambda \cdot \sum_{i=1}^r x_{(i)} - (n-r) \cdot \lambda \cdot x_{(r)} = \\ &= \ln \frac{n!}{(n-r)!} + r \cdot \ln \lambda - \lambda \cdot \sum_{i=1}^r x_{(i)} - (n-r) \cdot \lambda \cdot x_{(r)}\end{aligned}$$

Podmínky optimality zapišeme jako

$$\frac{\partial \ln L(x_1, \dots, x_n; \lambda)}{\partial \lambda} = 0$$

$$\frac{r}{\lambda} - \sum_{i=1}^r x_{(i)} - (n-r) \cdot x_{(r)} = 0$$

Maximálně věrohodný odhad při cenzorování II. typu je tedy

$$\hat{\lambda} = \frac{r}{\sum_{i=1}^r x_{(i)} + (n-r) \cdot x_{(r)}}$$

add) Při náhodném cenzorování budeme předpokládat, že parametry rozdělení náhodného cenzoru nemají s parametrem λ žádnou souvislost. Pro tento případ je věrohodnostní funkce

$$L(\underline{x}, \lambda) = \prod_{j \in U} f(x_j) \cdot \prod_{j \in C} (1 - F(T_j))$$

$$L(\underline{x}, \lambda) = \prod_{j \in U} \lambda \cdot e^{-\lambda \cdot x_j} \cdot \prod_{j \in C} (e^{-\lambda \cdot T_j}) = \lambda^{\sum_{j \in U} I_j} \cdot e^{-\lambda \cdot \sum_{j \in U} x_j} \cdot e^{-\lambda \cdot \sum_{j \in C} T_j}$$

Logaritmus funkce věrohodnosti tedy bude dán výrazem

$$\ln L(\underline{x}, \lambda) = \ln \lambda^{\sum_{j \in U} I_j} - \lambda \cdot \sum_{j \in U} x_j - \lambda \cdot \sum_{j \in U} T_j (1 - I_j) = \sum_{j \in U} I_j \cdot \ln \lambda - \lambda \cdot \sum_{j \in U} x_j - \lambda \cdot \sum_{j \in C} T_j$$

Podmínky optimality zapišeme jako

$$\frac{\partial \ln L(\underline{x}, \lambda)}{\partial \lambda} = 0$$

$$\frac{\sum_{j \in U} I_j}{\lambda} - \sum_{j \in U} x_j - \sum_{j \in C} T_j = 0$$

Maximálně věrohodný odhad při náhodném cenzorování je tedy

$$\hat{\lambda} = \frac{\sum_{j \in U} I_j}{\sum_{j \in U} x_j + \sum_{j \in C} T_j} = \frac{\sum_{j=1}^n I_j}{\sum_{j=1}^n W_j}$$



Řešený příklad

Nechť doba do poruchy má Weibullovo rozdělení $X \rightarrow W(\Theta, \beta)$. Určete maximálně věrohodný odhad neznámých parametrů Θ, β na základě

- úplného výběru x_1, \dots, x_n
- cenzorování I. typu
- cenzorování II. typu
- náhodného cenzorování

Poznámka: Jestliže je parametr β známý, můžeme Θ odhadnout na základě metod pro exponenciální rozdělení (odvozených v předcházejícím příkladě). (Návod: $X \rightarrow W(\Theta, \beta)$,

pak $Y = X^\beta \rightarrow E\left(\frac{1}{\Theta}\right)$. Aplikací této transformace na příslušný náhodný výběr dostaneme odpovídající výběr z exponenciálního rozdělení.) S případem, kdy parametr Θ je známý a parametr β máme odhadnout se v praxi nesetkáváme.

Hustota pravděpodobnosti Weibullova rozdělení je

$$f(x) = \frac{\beta}{\Theta} \left(\frac{x}{\Theta}\right)^{\beta-1} e^{-\left(\frac{x}{\Theta}\right)^\beta}; \quad x > 0; \Theta > 0; \beta > 0$$

ada) Je dán náhodný výběr $(x_1, \dots, x_n)^T$.

Funkce věrohodnosti pro úplný výběr bude dána výrazem:

$$L(\underline{x}; \Theta; \beta) = \prod_{i=1}^n \frac{\beta}{\Theta} \left(\frac{x_i}{\Theta}\right)^{\beta-1} e^{-\left(\frac{x_i}{\Theta}\right)^\beta}$$

Logaritmováním získáme

$$\begin{aligned} \ln L(\underline{x}; \Theta; \beta) &= \sum_{i=1}^n \ln \frac{\beta}{\Theta} + \sum_{i=1}^n \ln \left(\frac{x_i}{\Theta}\right)^{\beta-1} + \sum_{i=1}^n \ln e^{-\left(\frac{x_i}{\Theta}\right)^\beta} = \\ &= n \cdot \ln \frac{\beta}{\Theta} + (\beta - 1) \left(\sum_{i=1}^n \ln x_i - n \cdot \ln \Theta \right) - \left(\frac{1}{\Theta}\right)^\beta \sum_{i=1}^n (x_i)^\beta = \\ &= n \cdot \ln \beta - n \cdot \ln \Theta + (\beta - 1) \sum_{i=1}^n \ln x_i - n \cdot (\beta - 1) \cdot \ln \Theta - \Theta^{-\beta} \sum_{i=1}^n (x_i)^\beta = \\ &= n \cdot \ln \beta - n \cdot \ln \Theta + (\beta - 1) \sum_{i=1}^n \ln x_i - n \cdot \ln \Theta^\beta + n \cdot \ln \Theta - \Theta^{-\beta} \sum_{i=1}^n (x_i)^\beta = \\ &= n \cdot \ln \beta + (\beta - 1) \sum_{i=1}^n \ln x_i - n \cdot \ln \Theta^\beta - \Theta^{-\beta} \sum_{i=1}^n (x_i)^\beta \end{aligned}$$

Zavedme si substituci: $\alpha = \Theta^\beta$

Pak

$$\ln L(\underline{x}; \alpha; \beta) = n \cdot \ln \beta + (\beta - 1) \sum_{i=1}^n \ln x_i - n \cdot \ln \alpha - \frac{1}{\alpha} \cdot \sum_{i=1}^n (x_i)^\beta$$

Zbývá vyřešit podmínky optimality:

$$\frac{\partial \ln L(\underline{x}; \alpha, \beta)}{\partial \alpha} = 0 \quad \frac{\partial \ln L(\underline{x}; \alpha, \beta)}{\partial \beta} = 0$$

$$\frac{\partial \ln L(\underline{x}; \alpha, \beta)}{\partial \alpha} = -\frac{n}{\alpha} + \frac{1}{\alpha^2} \sum_{i=1}^n (x_i)^\beta$$

$$\frac{\partial \ln L(\underline{x}; \alpha, \beta)}{\partial \beta} = \frac{n}{\beta} + \sum_{i=1}^n \ln x_i - \frac{1}{\alpha} \sum_{i=1}^n ((x_i)^\beta \cdot \ln x_i)$$

Odhady parametrů Θ a β získáme řešením soustavy:

$$-\frac{n}{\alpha} + \frac{1}{\alpha^2} \sum_{i=1}^n (x_i)^\beta = 0$$

$$\frac{n}{\beta} + \sum_{i=1}^n \ln x_i - \frac{1}{\alpha} \sum_{i=1}^n ((x_i)^\beta \cdot \ln x_i) = 0$$

Z první rovnice můžeme snadno získat

$$-n \cdot \alpha + \sum_{i=1}^n (x_i)^\beta = 0 \Rightarrow \alpha = \frac{\sum_{i=1}^n (x_i)^\beta}{n}$$

zatímco z druhé rovnice dostaneme

$$\frac{n}{\beta} + \sum_{i=1}^n \ln x_i = \frac{1}{\alpha} \sum_{i=1}^n ((x_i)^\beta \cdot \ln x_i) \Rightarrow \alpha = \frac{\sum_{i=1}^n ((x_i)^\beta \cdot \ln x_i)}{\frac{n}{\beta} + \sum_{i=1}^n \ln x_i}$$

Porovnáním pravých stran posledních dvou rovnic získáme jednu rovnici pro jednu neznámou β . Řešení je nutno provést numericky volbou vhodného iteračního procesu. Poté určíme parametr α a zpětnou substitucí

$$\left(\alpha = \Theta^\beta \Rightarrow \ln \alpha = \ln \Theta^\beta \Rightarrow \ln \Theta = \frac{\ln \alpha}{\beta} \Rightarrow \Theta = e^{\frac{\ln \alpha}{\beta}} \right)$$

určíme parametr Θ .

adb) Při cenzorování I. typu je věrohodnostní funkce

$$L(\underline{x}, r) = \frac{n!}{(n-r)!} \cdot \prod_{i=1}^r f(x_{(i)}) \cdot R^{n-r}(T)$$

$$L(\underline{x}, \Theta, \beta, r) = \frac{n!}{(n-r)!} \cdot \prod_{i=1}^r \left(\frac{\beta}{\Theta} \left(\frac{x_{(i)}}{\Theta} \right)^{\beta-1} e^{-\left(\frac{x_{(i)}}{\Theta} \right)^\beta} \right) \cdot \left(e^{-\left(\frac{T}{\Theta} \right)^\beta} \right)^{n-r} =$$

$$= \frac{n!}{(n-r)!} \cdot \left(\frac{\beta}{\Theta} \right)^r \cdot \prod_{i=1}^r \left(\frac{x_{(i)}}{\Theta} \right)^{\beta-1} \cdot e^{-\frac{\sum_{i=1}^r (x_{(i)})^\beta}{\Theta^\beta}} \cdot \left(e^{-\left(\frac{T}{\Theta} \right)^\beta} \right)^{n-r}$$

Logaritmus funkce věrohodnosti tedy bude dán výrazem

$$\ln L(\underline{x}, \Theta, \beta, r) = \ln \frac{n!}{(n-r)!} + r \cdot \ln \frac{\beta}{\Theta} + \sum_{i=1}^r \ln \left(\frac{x_{(i)}}{\Theta} \right)^{\beta-1} + \ln e^{-\frac{\sum_{i=1}^r (x_{(i)})^\beta}{\Theta^\beta}} + \ln \left(e^{-\left(\frac{T}{\Theta} \right)^\beta} \right)^{n-r} =$$

$$= \ln \frac{n!}{(n-r)!} + r \cdot \ln \beta - r \cdot \ln \Theta + (\beta-1) \cdot \sum_{i=1}^r \ln \left(\frac{x_{(i)}}{\Theta} \right) - \frac{\sum_{i=1}^r (x_{(i)})^\beta}{\Theta^\beta} - (n-r) \left(\frac{T}{\Theta} \right)^\beta =$$

$$= \ln \frac{n!}{(n-r)!} + r \cdot \ln \beta - r \cdot \ln \Theta + (\beta-1) \cdot \left(\sum_{i=1}^r \ln x_{(i)} - r \cdot \ln \Theta \right) -$$

$$- \Theta^{-\beta} \left(\sum_{i=1}^r (x_{(i)})^\beta + (n-r) \cdot T^\beta \right) =$$

$$= \ln \frac{n!}{(n-r)!} + r \cdot \ln \beta - r \cdot \beta \cdot \ln \Theta + (\beta-1) \cdot \sum_{i=1}^r \ln x_{(i)} -$$

$$- \Theta^{-\beta} \left(\sum_{i=1}^r (x_{(i)})^\beta + (n-r) \cdot T^\beta \right)$$

Podmínky optimality zapišeme jako

$$\frac{\partial \ln L(\underline{x}; \Theta, \beta)}{\partial \Theta} = 0 \quad \frac{\partial \ln L(\underline{x}; \Theta, \beta)}{\partial \beta} = 0$$

$$\frac{\partial \ln L(\underline{x}; \Theta, \beta)}{\partial \Theta} = -\frac{r \cdot \beta}{\Theta} + \beta \cdot \Theta^{-\beta-1} \cdot \left(\sum_{i=1}^r (x_{(i)})^\beta + (n-r) \cdot T^\beta \right)$$

$$\begin{aligned}
\frac{\partial \ln L(\underline{x}; \Theta, \beta)}{\partial \beta} &= \frac{r}{\beta} - r \cdot \ln \Theta + \sum_{i=1}^r \ln x_{(i)} + \Theta^{-\beta} \cdot \ln \Theta \cdot \left(\sum_{i=1}^n (x_{(i)})^\beta + (n-r) \cdot T^\beta \right) - \\
&\quad - \Theta^{-\beta} \cdot \left(\sum_{i=1}^n \left((x_{(i)})^\beta \cdot \ln x_{(i)} \right) + (n-r) \cdot T^\beta \cdot \ln T \right) = \\
&= \frac{r}{\beta} - r \cdot \ln \Theta + \sum_{i=1}^r \ln x_{(i)} + \ln \Theta \cdot \left[\sum_{i=1}^n \left(\frac{x_{(i)}}{\Theta} \right)^\beta + (n-r) \cdot \left(\frac{T}{\Theta} \right)^\beta \right] - \\
&\quad - \left[\sum_{i=1}^n \left(\left(\frac{x_{(i)}}{\Theta} \right)^\beta \cdot \ln x_{(i)} \right) + (n-r) \cdot \left(\frac{T}{\Theta} \right)^\beta \cdot \ln T \right] = \\
&= \frac{r}{\beta} - r \cdot \ln \Theta + \sum_{i=1}^r \ln x_{(i)} + \sum_{i=1}^n \left[\left(\frac{x_{(i)}}{\Theta} \right)^\beta \cdot (\ln \Theta - \ln x_{(i)}) \right] + \\
&\quad + (n-r) \cdot \left(\frac{T}{\Theta} \right)^\beta (\ln \Theta - \ln T) = \frac{r}{\beta} - r \cdot \ln \Theta + \sum_{i=1}^r \ln x_{(i)} - \\
&\quad - \sum_{i=1}^n \left[\left(\frac{x_{(i)}}{\Theta} \right)^\beta \cdot \ln \frac{x_{(i)}}{\Theta} \right] - (n-r) \cdot \left(\frac{T}{\Theta} \right)^\beta \cdot \ln \frac{T}{\Theta}
\end{aligned}$$

Podmínky optimality dále upravíme

$$-\frac{r \cdot \beta}{\Theta} + \beta \cdot \Theta^{-\beta-1} \cdot \left(\sum_{i=1}^n (x_{(i)})^\beta + (n-r) \cdot T^\beta \right) = 0$$

$$\frac{r}{\beta} - r \cdot \ln \Theta + \sum_{i=1}^r \ln x_{(i)} - \sum_{i=1}^n \left[\left(\frac{x_{(i)}}{\Theta} \right)^\beta \cdot \ln \frac{x_{(i)}}{\Theta} \right] - (n-r) \cdot \left(\frac{T}{\Theta} \right)^\beta \cdot \ln \frac{T}{\Theta} = 0$$

Z první rovnice dostaneme

$$-r \cdot \beta \cdot \Theta^\beta + \beta \cdot \left(\sum_{i=1}^n (x_{(i)})^\beta + (n-r) \cdot T^\beta \right) = 0 \quad \Rightarrow \quad \Theta^\beta = \frac{\left(\sum_{i=1}^n (x_{(i)})^\beta + (n-r) \cdot T^\beta \right)}{r}$$

Po dosazení do druhé rovnice a úpravě dostaneme

$$\frac{\sum_{i=1}^n \left((x_{(i)})^\beta \cdot \ln x_{(i)} \right) + (n-r) \cdot T^\beta \cdot \ln T}{\sum_{i=1}^n (x_{(i)})^\beta + (n-r) \cdot T^\beta} - \frac{1}{\beta} - \frac{1}{r} \cdot \sum_{i=1}^r \ln x_{(i)} = 0$$

Tato rovnice již neobsahuje neznámý parametr Θ a jejím vyřešením tak (numerickými metodami) dostaneme maximálně věrohodný odhad $\hat{\beta}$. Dosadíme-li $\hat{\beta}$ do první podmínky optimality, dostaneme odhad $\hat{\Theta}$:

$$\hat{\Theta} = \left(\frac{\left(\sum_{i=1}^n (x_{(i)})^{\hat{\beta}} + (n-r) \cdot T^{\hat{\beta}} \right)}{r} \right)^{\frac{1}{\hat{\beta}}}$$

adc) Při cenzorování II. typu stačí v rovnicích pro odhady parametrů Θ a β nahradit symbol T symbolem $x_{(r)}$

Odhad parametru β tak získáme numerickým řešením rovnice:

$$\frac{\sum_{i=1}^n \left((x_{(i)})^{\hat{\beta}} \cdot \ln x_{(i)} \right) + (n-r) \cdot (x_{(r)})^{\hat{\beta}} \cdot \ln x_{(r)}}{\sum_{i=1}^n (x_{(i)})^{\hat{\beta}} + (n-r) \cdot (x_{(r)})^{\hat{\beta}}} - \frac{1}{\hat{\beta}} - \frac{1}{r} \cdot \sum_{i=1}^n \ln x_{(i)} = 0$$

a odhad parametru Θ získáme dosazením $\hat{\beta}$ do rovnice

$$\hat{\Theta} = \left(\frac{\left(\sum_{i=1}^n (x_{(i)})^{\hat{\beta}} + (n-r) \cdot (x_{(r)})^{\hat{\beta}} \right)}{r} \right)^{\frac{1}{\hat{\beta}}}$$

add) Při náhodném cenzorování dojdeme k rovnicím optimality podobně jako v předcházejících případech

$$\frac{\sum_{i=1}^n (W_i^{\hat{\beta}} \cdot \ln W_i)}{\sum_{i=1}^n W_i^{\hat{\beta}}} - \frac{1}{\hat{\beta}} - \frac{1}{r} \cdot \sum_{i \in U} \ln W_i = 0$$

$$\hat{\Theta} = \left(\frac{\left(\sum_{i=1}^n W_i^{\hat{\beta}} \right)}{r} \right)^{\frac{1}{\hat{\beta}}}$$

7. ZÁKLADY BAYESOVY INDUKCE



Čas ke studiu: 1,5 hodiny



Cíl: Po prostudování tohoto odstavce budete umět:

- Na úvod této kapitoly se seznámíte s odlišným pohledem na metodu maximální věrohodnosti a dále se pak budete věnovat základům Bayesovy indukce. Seznámíte se s pojmy apriorní a aposteriorní rozdělení, naučíte se nalézt Bayesův bodový a intervalový odhad.



Výklad

7.1. Metoda maximální věrohodnosti

V této kapitole se seznámíme s mírně odlišným pohledem na metodu maximální věrohodnosti, než jsme uvedli ve skriptech Statistika I. pro kombinované studium.

Nechť X_1, X_2, \dots, X_n je náhodný výběr z rozdělení s distribuční funkcí $F(x; \Theta)$, kde tvar distribuční funkce je znám a Θ je neznámý parametr. Obecně může distribuční funkce obsahovat více neznámých parametrů, které můžeme označit vektorově jako $\underline{\Theta}$. Problém bodového odhadu nyní spočívá v nalezení statistiky $\hat{\Theta}(X)$, jakožto funkce X_1, X_2, \dots, X_n , která by mohla být použita jako odhad $\underline{\Theta}$.

Tato statistika bývá často označována jako **estimátor** a její realizace, řekněme $\hat{\Theta}(x)$, jako odhad. Nechť $f(x; \underline{\Theta})$ je hustota pravděpodobnosti nebo pravděpodobnostní funkce náhodného výběru $\underline{X} (X_1, X_2, \dots, X_n)$.

Definice:

Pokud je hustota pravděpodobnosti nebo pravděpodobnostní funkce $f(x; \underline{\Theta})$ vyšetřovaná jako funkce $\underline{\Theta}$, nazveme ji **věrohodnostní funkcí** založenou na $\underline{x} = (x_1, \dots, x_n)$ a označíme ji jako $L(\underline{\Theta}; \underline{x})$.

Jestliže X_1, X_2, \dots, X_n je množina nezávislých náhodných pozorování z rozdělení s hustotami $f_i(x_i; \underline{\Theta})$, $i = 1, \dots, n$, pak věrohodnostní funkce může být získána jako:

$$L(\underline{\Theta}; x_1, \dots, x_n) = f_1(x_1; \underline{\Theta}) \cdot \dots \cdot f_n(x_n; \underline{\Theta})$$

Definice:

Nechť $L(\underline{\Theta}; \underline{x})$ je věrohodnostní funkce založena na náhodném výběru $\underline{X} = (X_1, X_2, \dots, X_n)$ z rozdělení $F(x; \underline{\Theta})$, kde $\underline{\Theta}$ je vektor neznámých parametrů, který nabývá hodnot z nějakého parametrického prostoru Θ . Pokud $\hat{\Theta} = \hat{\Theta}(\underline{x})$ je náhodný vektor, který maximalizuje $L(\underline{\Theta}; \underline{x})$ vzhledem k $\hat{\Theta} \in \Theta$, potom $\hat{\Theta}(\underline{x})$ budeme nazývat **maximálně věrohodný estimátor** Θ .

Pro konkrétní realizaci náhodného výběru $\underline{x} = (x_1, \dots, x_n)$ budeme $\hat{\Theta}(\underline{X})$ nazývat jako **maximálně věrohodný odhad Θ** . Pro tento odhad budeme používat zkratku MVO.

Tato tzv. metoda maximální věrohodnosti má hlavní výhodu ve své jednoduchosti a v tom, že dává odhady s velmi dobrými statistickými vlastnostmi.



Řešený příklad

MVO pro parametr exponenciálního rozdělení

Nechť $\underline{X} = (X_1, X_2, \dots, X_n)$ je náhodný výběr z exponenciálního rozdělení pravděpodobnosti.

Víme, že hustota pravděpodobnosti exponenciální náhodné veličiny má tvar:

$$f(x) = \lambda \cdot e^{-\lambda x}, \quad x > 0,$$

kde $\lambda > 0$ je neznámý parametr.

Chceme-li získat MVO pro parametr λ , zkonstruujeme nejdříve věrohodnostní funkci:

$$L(\lambda; \underline{x}) = \prod_{i=1}^n f(x_i) = \lambda^n e^{-\lambda \sum_{i=1}^n x_i}$$

V tomto případě je výhodné využít toho, že maximum kladné funkce se shoduje s maximum jejího logaritmu. Zaveďme si tedy funkci $L^*(\lambda; \underline{x})$, která bude logaritmem věrohodnostní funkce:

$$L^*(\lambda; \underline{x}) = \ln \left(\lambda^n e^{-\lambda \sum_{i=1}^n x_i} \right) = \ln(\lambda^n) + \ln \left(e^{-\lambda \sum_{i=1}^n x_i} \right) = n \cdot \ln \lambda - \lambda \sum_{i=1}^n x_i$$

Zbývá nalézt maximum $L^*(\lambda; \underline{x})$.

Bod podezřelý z extrému určíme tak, že první derivaci funkce položíme rovnu 0.

$$\frac{dL^*(\lambda; \underline{x})}{d\lambda} = \frac{n}{\lambda} - \sum_{i=1}^n x_i$$

$$\frac{n}{\lambda} - \sum_{i=1}^n x_i = 0 \Rightarrow \hat{\lambda} = \frac{n}{\sum_{i=1}^n x_i} = \frac{1}{\bar{x}}$$

Pomocí druhé derivace zjistíme, zda se skutečně jedná o maximum funkce (druhá derivace v tomto bodě musí být záporná).

$$\frac{d^2 L^*(\lambda; \underline{x})}{d\lambda^2} = -\frac{n}{\lambda^2}$$

$$\frac{d^2 L^*(\lambda; \mathbf{x})}{d\lambda^2}(\hat{\lambda}) = -n \cdot \bar{x}^{-2} < 0$$

Tímto jsme ukázali, že $\hat{\lambda} = \frac{1}{\bar{x}}$ maximalizuje $L^*(\lambda; \mathbf{x})$ a tím také věrohodnostní funkci. Tedy

$\frac{1}{\bar{x}}$, kde \bar{x} je výběrový průměr z výběru pocházejícího z exponenciálního rozdělení, je MVO parametru λ .

MVO očekávané hodnoty exponenciálního rozdělení, označený jako $\mu = \frac{1}{\lambda}$, může být odvozen podobným způsobem z minulého výsledku pomocí vlastnosti invariance maximálně věrohodných odhadů. Tato vlastnost říká [Nguyen H. T., Rogers G. S., 1989], že:

Pokud $\hat{\theta}$ je MVO parametru Θ , pak $g(\hat{\theta})$ je MVO funkce parametru $g(\Theta)$ za předpokladu, že $g(\cdot)$ je prostá funkce.

Nyní je jasné, že MVO očekávané hodnoty exponenciálního rozdělení je $\hat{\mu} = \bar{x}$.

7.2. Úvod do Bayesovy indukce

Parametr Θ , který je předmětem našeho zájmu, je v Bayesově přístupu vyšetřován jako náhodná veličina. Jde-li o parametr náhodné veličiny X s distribuční funkcí $F(x; \Theta)$, pak vždy, když pracujeme s touto funkcí, musíme mít na mysli podmíněnou distribuční funkci $F(x|\Theta)$.

7.3. Apriorní rozdělení

Uvažujme parametr Θ náhodné veličiny X s distribuční funkcí $F(x|\Theta)$. Necht' doposud není k dispozici žádné pozorování této náhodné veličiny. Necht' je k dispozici jakási předběžná zkušenost, či apriorní znalost o této populaci nebo jiná informace, která umožní zkonstruovat subjektivní pravděpodobnostní rozdělení pro parametr Θ . Takové rozdělení, které reflektuje nejistotu o hodnotě parametru z hlediska experimentátora ještě před pozorováním aktuálního výběru, se nazývá **apriorní rozdělení** parametru Θ . Dále jej budeme označovat $\pi(\Theta)$.

V mnoha situacích lze vyjádřit relativní pravděpodobnost toho, že parametr nabývá hodnot na nějaké množině Ω , pomocí vhodného rozdělení pravděpodobnosti. Úloha najít apriorní rozdělení pro zkoumaný parametr je všeobecně velmi obtížná. V některých případech bývá dokonce velmi výhodné zvolit jako apriorní hustotu takovou funkci, která nemusí být ani integrovatelná, a přesto po implementaci Bayesových metod dává rozumné výsledky (někdy však také vede k nesmyslným výsledkům). Taková hustota bývá označována jako tzv. **nevlastní apriorní hustota**. Bayesovy metody lze použít i v případě, že není dostupná žádná informace o vyšetřovaném parametru. Taková apriorní rozdělení, označována jako **neurčitá**, byla velmi intenzivně studována [Jeffreys, 1961] a jsou základem Bayesových metod vyvinutých autory [Box and Tiao, 1973].

Neexistuje žádný obecný návod, jak by měla být specifikována neurčitá apriorní hustota. Jelikož jsou pod tíhou různých argumentů použity různé definice neurčitých apriorních rozdělení, setkáme se často s různými nevlastními apriorními rozděleními [Zellner, 1977]. Pro účely nalezení neurčitého apriorního rozdělení použijeme jednu z nejpopulárnějších metod, navrženou v [Jeffreys, 1961].

Necht' $f(x|\Theta)$ je hustota pravděpodobnosti nebo pravděpodobnostní funkce pozorované náhodné veličiny X , kde vektor Θ je vektor neznámých parametrů. Za neurčité apriorní rozdělení Θ lze vzít:

$$\pi(\Theta) = [I(\Theta)]^{1/2},$$

kde $I(\Theta)$ je Fisherova informační matice (definována pomocí druhé derivace logaritmické věrohodnostní funkce).

$$I(\Theta) = -E\left(\frac{\partial^2}{\partial\Theta\partial\Theta'} \ln(f(X|\Theta))\right)$$



Řešený příklad

Uvažujme exponenciální rozdělení s hustotou $f(x|\lambda) = \lambda \cdot e^{-\lambda x}$, $x > 0$, kde $\lambda > 0$ je parametr, pro který je nutno zkonstruovat apriorní hustotu. Úkolem je nalézt neurčitou apriorní hustotu pro parametr λ a pro jeho převrácenou hodnotu $\mu = \frac{1}{\lambda}$.

$$I(\lambda) = -E\left(\frac{\partial^2}{\partial\lambda^2} \ln(\lambda \cdot e^{-\lambda x})\right) = -E\left(\frac{\partial^2}{\partial\lambda^2} [\ln(\lambda) - \lambda x]\right) = -E\left(-\frac{1}{\lambda^2}\right) = \frac{1}{\lambda^2}$$

Odtud vyplývá, že neurčitá (Jeffreysova) apriorní hustota pravděpodobnosti pro λ je nevlastní apriorní hustota, úměrná $\frac{1}{\lambda} \left(= \left(\frac{1}{\lambda^2}\right)^{1/2} \right)$:

$$\pi(\Theta) = \frac{1}{\lambda}$$

Pokud nás bude zajímat apriorní rozdělení pro převrácenou hodnotu λ , tj. μ , pak Fischerova matice bude:

$$\begin{aligned} I(\lambda) &= -E\left(\frac{\partial^2}{\partial\mu^2} \ln\left(\frac{1}{\mu} \cdot e^{-\frac{1}{\mu}x}\right)\right) = -E\left(\frac{\partial^2}{\partial\mu^2} \left[\ln\left(\frac{1}{\mu}\right) - \frac{1}{\mu}X\right]\right) = \\ &= -E\left(\frac{\partial}{\partial\mu} \left(-\mu \frac{1}{\mu^2} + \frac{1}{\mu^2}X\right)\right) = -E\left(\frac{\partial}{\partial\mu} \left(-\frac{1}{\mu} + \frac{1}{\mu^2}X\right)\right) = -E\left(\frac{1}{\mu^2} - 2\frac{1}{\mu^3}X\right) = \\ &= -\frac{1}{\mu^2} + \frac{2}{\mu^3} \cdot EX = -\frac{1}{\mu^2} + \frac{2}{\mu^3} \cdot \mu = \frac{1}{\mu^2} \end{aligned}$$

Tedy neurčitá apriorní hustota pro očekávanou hodnotu exponenciálního rozdělení μ je:

$$\pi(\mu) = \left(\frac{1}{\mu^2}\right)^{1/2} = \frac{1}{\mu}$$

7.4. Aposteriorní rozdělení

Nechť \underline{X} je náhodný vektor se sdruženou hustotou pravděpodobnosti nebo pravděpodobnostní funkcí $f(x|\Theta)$. Nechť $\pi(\Theta)$ je apriorní rozdělení náhodného vektoru $\underline{\Theta}$. Potom sdružené rozdělení \underline{X} a $\underline{\Theta}$ může být nalezeno jako:

$$h(x, \Theta) = f(x|\Theta) \cdot \pi(\Theta)$$

Za předpokladu, že Θ je absolutně spojitý náhodný vektor, marginální rozdělení \underline{X} může být nalezeno jako:

$$g(x) = \int_{\Theta} h(x; \Theta) d\Theta$$

A konečně podmíněné rozdělení $\underline{\Theta}$ při realizaci $\underline{X} = \underline{x}$ je:

$$\pi(\Theta|x) = \frac{h(x; \Theta)}{g(x)} \quad \text{pro } \Theta \in \Omega$$

Toto pravděpodobnostní rozdělení $\underline{\Theta}$ se nazývá **aposteriorní rozdělení $\underline{\Theta}$** . Toto podmíněné rozdělení parametru při daných datech x je takto nazváno zejména proto, že odráží představu experimentátora o zkoumaném parametru poté, co byl pozorován náhodný výběr z příslušné populace. Tedy aposteriorní rozdělení kombinuje předběžnou informaci obsaženou v apriorním rozdělení s informací o $\underline{\Theta}$ (obsaženou v datech – věrohodnostní funkce). Pokud budeme ignorovat konstantu úměrnosti, pak aposteriorní rozdělení může být zapsáno následovně:

$$\pi(\Theta|x) \propto f(x|\Theta) \cdot \pi(\Theta) \quad \text{pro } \Theta \in \Omega,$$

kde konstanta úměrnosti k může být nalezena tak, aby byla splněna normovací podmínka aposteriorní hustoty.

Pozn.: \propto označuje přímou úměrnost, tzn. $\pi(\Theta|x) \propto f(x|\Theta) \cdot \pi(\Theta)$ je zkráceným zápisem rovnice:

$$\pi(\Theta|x) = k \cdot f(x|\Theta) \cdot \pi(\Theta), \quad k \in R$$

Tedy:

aposteriorní rozdělení \propto (apriorní rozdělení \cdot věrohodnostní funkce)

Pokud aposteriorní rozdělení pravděpodobnosti patří do téže třídy rozdělení jako apriorní rozdělení, potom tuto třídu rozdělení nazýváme **přirozený konjugovaný systém rozdělení** pro rozdělení X . To znamená, že pokud apriorní rozdělení je konjugované vzhledem k výběrovému rozdělení, pak pro nalezení aposteriorního rozdělení potřebujeme aktualizovat pouze parametry apriorního rozdělení.



Řešený příklad

Nechť X_1, \dots, X_n je náhodný výběr z exponenciálního rozdělení s parametrem λ . Úkolem je nalézt aposteriorní rozdělení pro λ za předpokladu, že apriorní rozdělení je:

a) $\lambda \rightarrow \text{Gamma}(a; b)$, tj.

$$\pi(\lambda) = \frac{\left(\frac{1}{b}\right)^a}{\Gamma(a)} \cdot \lambda^{a-1} \cdot e^{-\frac{\lambda}{b}}, \quad \text{kde } \Gamma(a) = \int_0^{\infty} x^{a-1} e^{-x} dx, \quad E\lambda = ab$$

b) $\pi(\lambda) = \frac{1}{\lambda}$

Víme, že hustota pravděpodobnosti exponenciální náhodné veličiny má tvar:

$$f(x) = \lambda \cdot e^{-\lambda x}, \quad x > 0,$$

kde $\lambda > 0$ je neznámý parametr.

Zkonstruujeme nejdříve věrohodnostní funkci:

$$L(\lambda; \mathbf{x}) = \prod_{i=1}^n f(x_i) = \lambda^n e^{-\lambda \sum_{i=1}^n x_i} = \lambda^n e^{-\lambda n \bar{x}},$$

kde \bar{x} je výběrový průměr.

ada) Pokud ignorujeme konstantu úměrnosti, apriorní rozdělení $\pi(\lambda) \propto \lambda^{a-1} \cdot e^{-\frac{\lambda}{b}}$

aposteriorní rozdělení \propto (apriorní rozdělení \cdot věrohodnostní funkce)

$$\text{aposteriorní rozdělení} \propto \lambda^{a-1} \cdot e^{-\frac{\lambda}{b}} \cdot \lambda^n e^{-\lambda n \bar{x}} \Rightarrow \pi(\lambda|x) \propto \lambda^{n+a-1} \cdot e^{-\lambda \left(\frac{1}{b} + n \bar{x}\right)}$$

Lze snadno rozeznat, že se jedná o Gamma rozdělení:

$$(\lambda|x) \rightarrow \text{Gamma} \left(n + a; \frac{b}{nb \bar{x} + 1} \right)$$

Protože apriorní i aposteriorní rozdělení patří do téže třídy rozdělení, je evidentní, že tato třída (Gamma rozdělení) slouží jako přirozený konjugovaný systém pro výběrová exponenciální rozdělení.

adb) nevlastní apriorní rozdělení: $\pi(\lambda) = \frac{1}{\lambda}$

$$\text{aposteriorní rozdělení} \propto \frac{1}{\lambda} \cdot \lambda^n e^{-\lambda n \bar{x}} \Rightarrow \pi(\lambda|x) \propto \lambda^{n-1} \cdot e^{-\lambda n \bar{x}},$$

$$\text{takže: } (\lambda|x) \rightarrow \text{Gamma} \left(n; \frac{1}{n \bar{x}} \right)$$

a skutečnost, že apriorní rozdělení je nevlastní, zde nehraje významnou roli.

7.5. Bayesovy estimátory

Nechť $\pi(\Theta|x)$ je aposteriorní rozdělení pro parametr Θ , založeno na pozorováních \mathbf{x} náhodného vektoru \underline{X} , který má distribuční funkci $F(\underline{x}|\Theta)$. Cílem bude získat bodové a intervalové odhady pro Θ . Nejdříve uvažujme problém bodového odhadu.

□ Bayesův bodový odhad

Aposteriorní rozdělení v Bayesově přístupu hraje podobnou roli jako věrohodnostní funkce v kapitole 5 – výraz něčeho, s čím přichází veškerá informace o parametru $\underline{\Theta}$. Na rozdíl od věrohodnostní funkce však aposteriorní rozdělení obsahuje informaci jak v podobě apriorního rozdělení, tak i ve výběru samotném. Bodový odhad parametru může být získán obdobně jako v definici maximálně věrohodného estimátoru. Jestliže MVO pro parametr $\underline{\Theta}$ byl dán nalezením modu věrohodnostní funkce, pak obdobně můžeme definovat **zobecněný maximálně věrohodný estimátor** jako modus aposteriorního rozdělení $\pi(\Theta|x)$.

Odhad $\underline{\Theta}$ získaný touto metodou bude posuzován jako **odhad aposteriorním modem**.

Poznamenejme, že aposteriorní rozdělení je pravděpodobnostním rozdělením, zatímco věrohodnostní funkce, jako funkce $\underline{\Theta}$, nemusí nutně být pravděpodobnostním rozdělením. Modus aposteriorního rozdělení je speciální mírou polohy tohoto rozdělení. Lze tedy zavést i další míry polohy aposteriorního rozdělení, které mohou odhadnout $\underline{\Theta}$ stejně dobře. Tyto odhady budou posuzovány jako: **odhad aposteriorní očekávanou hodnotou**, resp. **odhad aposteriorním mediánem**. Z teoretického hlediska mohou být tyto alternativní odhady v jistém smyslu optimální, pokud je zadána jistá **ztrátová funkce odhadu**. Například aposteriorní očekávaná hodnota, resp. aposteriorní medián jsou Bayesovy odhady parametru $\underline{\Theta}$ za předpokladu **kvadratické ztrátové funkce**

$$L(\underline{\Theta}; \hat{\underline{\Theta}}) = (\underline{\Theta} - \hat{\underline{\Theta}})^2,$$

resp. ztrátové funkce ve tvaru: $L(\underline{\Theta}; \hat{\underline{\Theta}}) = |\underline{\Theta} - \hat{\underline{\Theta}}|$, kde $\hat{\underline{\Theta}}$ je odhad parametru $\underline{\Theta}$.

Abychom krátce popsali myšlenku Bayesových estimátorů, uvažujme obecnou ztrátovou funkci $L(\underline{\Theta}; \hat{\underline{\Theta}})$, pro níž existuje očekávaná hodnota vzhledem k aposteriornímu rozdělení $\pi(\Theta|x)$.

Potom estimátor $\hat{\underline{\Theta}}$ nazveme **Bayesovým estimátorem $\underline{\Theta}$** při uvažované ztrátové funkci $L(\underline{\Theta}; \hat{\underline{\Theta}})$, pokud $\hat{\underline{\Theta}}$ minimalizuje **aposteriorní očekávanou ztrátu**

$$E[L(\underline{\Theta}; \hat{\underline{\Theta}})|x] = \int_{\Theta} L(\underline{\Theta}; \hat{\underline{\Theta}}) \cdot \pi(\Theta|x) d\Theta$$

Například pro kvadratickou ztrátovou funkci může být aposteriorní očekávaná ztráta vyjádřena jako

$$E\left[(\underline{\Theta} - \hat{\underline{\Theta}})^2|x\right] = E\left[\left((\underline{\Theta} - E(\underline{\Theta}|x)) + (E(\underline{\Theta}|x) - \hat{\underline{\Theta}})\right)^2|x\right] = D(\underline{\Theta}|x) + [E(\underline{\Theta}|x) - \hat{\underline{\Theta}}]^2 \geq D(\underline{\Theta}|x)$$

Z poslední nerovnosti dále plyne, že pokud $\hat{\underline{\Theta}} = E(\underline{\Theta}|x)$, což je aposteriorní očekávaná hodnota $\underline{\Theta}$, pak aposteriorní očekávaná ztráta je rovna $D(\underline{\Theta}|x)$. Nyní je jasné, že aposteriorní očekávaná hodnota, jakožto odhad $\underline{\Theta}$, minimalizuje aposteriorní očekávanou ztrátu. Jinými slovy – aposteriorní očekávaná hodnota je Bayesův estimátor při uvažované kvadratické ztrátové funkci.



Řešený příklad

Uvažujme aposteriorní rozdělení dané $(\lambda|x) \rightarrow \text{Gamma} \left(n + a; \frac{b}{nb\bar{x} + 1} \right)$ pro parametr λ exponenciálního rozdělení. Z vlastnosti Gamma rozdělení bezprostředně plyne, že

$$\hat{\lambda} = \frac{b(n+a)}{1+nb\bar{x}}$$

je odhadem λ pomocí aposteriorní očekávané hodnoty. Všimněme si obzvláště, že pokud $a \rightarrow 0$ a $b \rightarrow \infty$, tedy při nevlastním apriorním rozdělení, Bayesův esimátor (při uvažované kvadratické ztrátové funkci) se redukuje na $\hat{\lambda} = \frac{1}{x}$, což je totéž jako

MVO pro λ .

Pokusme se nyní nalézt zobecněný maximálně věrohodný odhad λ (zobecněný MVO), založený na výběru X_1, \dots, X_n a při předpokladu apriorního rozdělení Gamma $G(a;b)$.

Je samozřejmé, že modus aposteriorního rozdělení bude ve stejném bodě jako maximum funkce

$$g(\lambda) = (n+a-1) \cdot \ln(\lambda) - \frac{\lambda(1+nb\bar{x})}{b},$$

Protože až na konstantu nezávislou na λ je $g(\lambda)$ totéž jako logaritmus aposteriorního rozdělení. Derivováním dostaneme:

$$\frac{dg}{d\lambda} = \frac{n+a-1}{\lambda} - \frac{1+nb\bar{x}}{b} \quad a \quad \frac{d^2g}{d\lambda^2} = -\frac{n+a-1}{\lambda^2}$$

Jelikož druhá derivace je záporná pro všechna $n+a > 1$, maximum $g(\lambda)$ může být nalezeno z 1. rovnice, kterou položíme rovnu 0. Odtud zobecněný MVO parametru λ je

$$\hat{\lambda} = \frac{b \cdot (n+a-1)}{1+nb\bar{x}}$$

7.6. Bayesův intervalový odhad

Předpokládejme, že pro Θ je nyní požadována konfidenční množina. Připomeňme, že na rozdíl od klasického přístupu, Θ je nyní náhodný vektor. Proto, na rozdíl od klasického přístupu, který vydává pravděpodobnostní výpověď o podmnožině základního prostoru s cílem nalezení konfidenční oblasti, zde provedeme pravděpodobnostní výpovědi týkající se přímo podmnožin parametrického prostoru. Konfidenční množiny, které získáme pomocí Bayesova přístupu, mají přímou pravděpodobnostní interpretaci. Bayesův analog vůči konfidenčnímu intervalu je přisuzován **Bayesově konfidenčním intervalu** nebo také pravděpodobnostnímu intervalu.

Definice:

Nechť $C(x)$ je podmnožina parametrického prostoru Ω taková, že

$$P(\Theta \in C(x)|x) = \int_{C(x)} \pi(\Theta|x) d\Theta = \gamma$$

Potom $C(x)$ se nazývá **100% ní pravdivostní množina pro Θ** , kde $\pi(\Theta|x)$ je aposteriorní rozdělení pro Θ .



Shrnutí kapitoly 7.

Parametr Θ , který je předmětem našeho zájmu, je v Bayesově přístupu vyšetřován jako náhodná veličina. Jde-li o parametr náhodné veličiny X s distribuční funkcí $F(x; \Theta)$, pak vždy, když pracujeme s touto funkcí, musíme mít na mysli podmíněnou distribuční funkci $F(x|\Theta)$.

Rozdělení, které reflektuje nejistotu o hodnotě parametru z hlediska experimentátora ještě před pozorováním aktuálního výběru, se nazývá **apriorní rozdělení** parametru Θ . Označujeme jej $\pi(\Theta)$.

Nechť $f(x|\Theta)$ je hustota pravděpodobnosti nebo pravděpodobnostní funkce pozorované náhodné veličiny X , kde vektor Θ je vektor neznámých parametrů. Za neurčité apriorní rozdělení Θ lze vzít:

$$\pi(\Theta) = [I(\Theta)]^{1/2},$$

kde $I(\Theta)$ je **Fischerova informační matice** (definována pomocí druhé derivace logaritmické věrohodnostní funkce).

$$I(\Theta) = -E \left(\frac{\partial^2}{\partial \Theta \partial \Theta'} \ln (f(X|\Theta)) \right)$$

Sdružené rozdělení X a Θ může být nalezeno jako:

$$h(x, \Theta) = f(x|\Theta) \cdot \pi(\Theta)$$

Za předpokladu, že Θ je absolutně spojitý náhodný vektor, marginální rozdělení X může být nalezeno jako:

$$g(x) = \int_{\Theta} h(x; \Theta) d\Theta$$

A konečně podmíněné rozdělení Θ při realizaci $X = x$ je:

$$\pi(\Theta|x) = \frac{h(x; \Theta)}{g(x)} \quad \text{pro } \Theta \in \Omega$$

Toto pravděpodobnostní rozdělení Θ se nazývá **aposteriorní rozdělení Θ** . Toto podmíněné rozdělení parametru při daných datech x je takto nazváno zejména proto, že odráží představu experimentátora o zkoumaném parametru poté, co byl pozorován náhodný výběr z příslušné populace.

aposteriorní rozdělení \propto (apriorní rozdělení \cdot věrohodnostní funkce)

Pokud aposteriorní rozdělení pravděpodobnosti patří do téže třídy rozdělení jako apriorní rozdělení, potom tuto třídu rozdělení nazýváme **přirozený konjugovaný systém rozdělení** pro rozdělení X .

Abychom krátce popsali myšlenku Bayesových estimátorů, uvažujme obecnou ztrátovou funkci $L(\underline{\Theta}; \hat{\underline{\Theta}})$, pro níž existuje očekávaná hodnota vzhledem k aposteriornímu rozdělení $\pi(\Theta|x)$.

Potom estimátor $\hat{\underline{\Theta}}$ nazveme **Bayesovým estimátorem $\underline{\Theta}$** při uvažované ztrátové funkci $L(\underline{\Theta}; \hat{\underline{\Theta}})$, pokud $\hat{\underline{\Theta}}$ minimalizuje **aposteriorní očekávanou ztrátu**

$$E[L(\underline{\Theta}; \hat{\underline{\Theta}})|x] = \int_{\Theta} L(\underline{\Theta}; \hat{\underline{\Theta}}) \cdot \pi(\Theta|x) d\Theta$$

Konfidenční množiny, které získáme pomocí Bayesova přístupu, mají přímou pravděpodobnostní interpretaci. Bayesův analog vůči konfidenčnímu intervalu je přisuzován **Bayesově konfidenčním intervalu** nebo také pravděpodobnostnímu intervalu.

Definice:

Nechť $C(x)$ je podmnožina parametrického prostoru Ω taková, že

$$P(\Theta \in C(x)|x) = \int_{C(x)} \pi(\Theta|x) d\Theta = \gamma$$

Potom $C(x)$ se nazývá 100γ % ní pravdivostní množina pro $\underline{\Theta}$, kde $\pi(\Theta|x)$ je aposteriorní rozdělení pro $\underline{\Theta}$.



Otázky 7.

1. Co je to apriorní rozdělení?
2. Definujte Fisherovu informační matici.
3. Co je to aposteriorní rozdělení?
4. Jaký je vztah mezi aposteriorním a apriorním rozdělením?
5. Co je to přirozený konjugovaný systém rozdělení?
6. Co je to ztrátová funkce?
7. Kdy mluvíme o Bayesově bodovém odhadu?

8. NÁHODNÉ PROCESY I



Čas ke studiu: 1,5 hodiny



Cíl: Po prostudování tohoto odstavce budete umět:

- Seznámíte se se základními pojmy z teorie náhodných procesů, Markovovými procesy, procesy růstů a zániků. Naučíte se popisovat vícestavové systémy pomocí pravděpodobností přechodů a pravděpodobností stavů.



Výklad

8.1. Náhodné procesy

Náhodným (stochastickým) procesem nazveme zobrazení, které každé hodnotě $t \in T$ přiřadí náhodnou veličinu $X(t)$. Proměnná t má ve většině případů význam času.

Realizací náhodného procesu rozumíme konkrétní pozorování náhodného procesu, tj. již nenáhodnou funkci, a značíme ji $x(t)$.

Dle povahy množiny T rozlišujeme:

- **náhodné procesy se spojitým časem** (náhodné funkce) – T je reálný interval,
- **náhodné procesy s diskrétním časem** (náhodné posloupnosti) – T je reálná diskrétní množina.

Hodnota $X(t)$ vyjadřuje stav pozorovaného objektu v čase t . Dle povahy náhodné veličiny $X(t)$ rozlišujeme:

- **náhodné procesy se spojitými stavy** - $X(t)$ je spojitá,
- **náhodné procesy s diskrétními stavy** - $X(t)$ je diskrétní.

Náhodný proces $\{X(t): t \geq 0\}$ se spojitým časem a s diskrétními stavy $0, 1, 2, \dots$ obvykle nazýváme **čítací proces**, protože zaznamenává počet nějakých událostí v čase. Hodnota $X(t)$ pak představuje počet daných událostí v intervalu $(0, t)$ a vzdálenosti jednotlivých okamžiků událostí od počátku $t = 0$ jsou náhodné veličiny.

8.2. Poissonův proces

Přibližme si nyní **Poissonův proces** jako příklad čítacího procesu, který se velmi často vyskytuje v aplikacích (například v teorii hromadné obsluhy).

Nechť $\{X(t): t \geq 0\}$ je čítací proces. Nechť navíc platí:

$$X(0) = 0,$$

délky intervalů mezi výskyty sledované události jsou nezávislé náhodné veličiny s exponenciálním rozdělením s hustotou

$$f(x) = \begin{cases} \lambda e^{-\lambda x} & \text{pro } x > 0, \\ 0 & \text{pro } x \leq 0, \end{cases}$$

kde $\lambda > 0$ je parametr (tzv. intenzita homogenního Poissonova procesu).

Pak tento proces nazveme **homogenním Poissonovým procesem**, přičemž $X(t)$ má Poissonovo rozdělení s parametrem λt , tedy

$$P(X(t) = i) = \frac{(\lambda t)^i}{i!} e^{-\lambda t}, \quad i = 0, 1, \dots$$

Střední počet výskytů události v intervalu $\langle 0, t \rangle$ je roven λt . Parametr λ tedy udává střední počet výskytů sledované události za jednotku času.

Protože intervaly mezi jednotlivými výskyty událostí jsou nezávislé, znalost okamžiků prvních n výskytů neovlivňuje předpověď doby čekání na další výskyt události. Také skutečnost, že sledovaná událost už po určité době nenastala, nemění pravděpodobnost jejího výskytu v dalším intervalu.

Příkladem Poissonova procesu by mohl být proces $\{X(t) : t \geq 0\}$, kde $X(t)$ udává počet poruch nějakého zařízení v časovém intervalu $\langle 0, t \rangle$.



Řešený příklad

Zdroj záření vysílá v průměru 1 impuls za 2 sekundy, přičemž impulsy tvoří Poissonův proces. Jaká je pravděpodobnost, že v každém z pěti intervalů o délce 5 sekund

(0s, 5s), (5s, 10s), ..., (20s, 25s)

budou registrovány nejméně 4 impulsy?

Protože $EX(t) = \lambda t$, spočteme z rovnice $1 = \lambda \cdot 2$ parametr $\lambda = 0,5$. Pro $t = 5$ pak získáme $EX(5) = 0,5 \cdot 5 = 2,5$. Pro pravděpodobnost, že během jednoho intervalu dojde k registraci alespoň 4 impulsů, platí

$$P(X(5) \geq 4) = \sum_{k=4}^{\infty} \frac{2,5^k}{k!} e^{-2,5}$$

a hledaná pravděpodobnost pro všech pět intervalů je tak rovna hodnotě

$$\left[\sum_{k=4}^{\infty} \frac{2,5^k}{k!} e^{-2,5} \right]^5.$$

8.3. Markovův proces

Nebude-li uvedeno jinak, $\{X(t): t \geq 0\}$ bude označovat náhodný proces se spojitým časem a diskrétní množinou stavů $I = \{0,1,2,\dots\}$ (stavy jsou pro jednoduchost označeny celými nezápornými čísly).

Proces $\{X(t): t \geq 0\}$ nazveme **Markovovým procesem**, splňuje-li tzv. markovskou vlastnost:

pro libovolná $0 \leq t_1 < t_2 < \dots < t_n < t \leq \tau$ a $i_1, \dots, i_n, i, j \in I$ platí:

$$P(X(\tau) = j | X(t) = i, X(t_n) = i_n, \dots, X(t_1) = i_1) = P(X(\tau) = j | X(t) = i).$$

Pravděpodobnost na pravé straně uvedené rovnosti nazýváme **pravděpodobnost přechodu**.

Je-li tedy t přítomný okamžik, potom chování Markova procesu v libovolném budoucím okamžiku $\tau \geq t$ závisí pouze na přítomném stavu a nikoli na stavech předchozích.

Markovův proces se nazývá **homogenní**, pokud pravděpodobnost přechodu z předchozího výkladu nezávisí na hodnotách t a τ , ale pouze na jejich rozdílu. Používáme pak značení

$$p_{ij}(\tau - t) = P(X(\tau) = j | X(t) = i).$$

Tedy v homogenním procesu závisí pravděpodobnosti přechodu pouze na rozdílu časových okamžiků a jsou navíc invariantní vůči posunutí v čase. Pro $\tau = t$ pak dostaneme

$$p_{ij}(0) = \begin{cases} 0 & \text{pro } i \neq j, \\ 1 & \text{pro } i = j. \end{cases}$$

Pro popis rozdělení Markovova procesu v čase t budeme v dalším textu užívat

$$p_i(t) = P(X(t) = i), i = 0, 1, 2, \dots$$

Při pravděpodobnostech $p_i(0)$, $i = 0, 1, 2, \dots$, se mluví o **počátečním rozdělení** Markovova procesu. Při velkém t je obvykle Markovův proces stabilizovaný a řídí se **stacionárním rozdělením** se stacionárními pravděpodobnostmi

$$\pi_i = \lim_{t \rightarrow \infty} p_i(t), i = 0, 1, 2, \dots$$

Jednoduchým příkladem homogenního Markovova procesu by mohl být homogenní Poissonův proces z předchozího příkladu. Počáteční rozdělení by mělo tvar $p_0(0) = 1$, $p_n(0) = 0$ pro $n = 1, 2, \dots$ a pro

pravděpodobnosti přechodu by platilo $p_{ij}(h) = \frac{\lambda h^{j-i}}{(j-i)!} e^{-\lambda h}$ pro $i, j \in N \cup \{0\}$, $j \geq i$.

V homogenním Markovově procesu platí

- $p_{ij}(t_1 + t_2) = \sum_{k=0}^{\infty} p_{ik}(t_1)p_{kj}(t_2), t_1, t_2 \geq 0, i, j = 0, 1, 2, \dots,$
- $p_i(t) = \sum_{j=0}^{\infty} p_j(0)p_{ji}(t), t \geq 0, i = 0, 1, 2, \dots.$

První rovnice se nazývá **Chapmanova-Kolmogorovova rovnice**.

Prospektivní Kolmogorovy diferenciální rovnice – necht' pro homogenní Markovův proces platí předpoklady:

1. existují limity $q_{ii} = \lim_{h \rightarrow 0_+} \frac{p_{ii}(h) - 1}{h}, i = 0, 1, 2, \dots,$
2. existují limity $q_{ij} = \lim_{h \rightarrow 0_+} \frac{p_{ij}(h)}{h}, i \neq j, i, j = 0, 1, 2, \dots,$
3. pro pevné j je konvergence vůči i v bodě 2 stejnoměrná.

Pak pro pravděpodobnosti přechodu platí

$$p'_{ij}(t) = \sum_{k=0}^{\infty} p_{ik}(t)q_{kj}, t > 0, i, j = 0, 1, 2, \dots$$

a pro pravděpodobnosti rozdělení procesu platí

$$p'_i(t) = \sum_{k=0}^{\infty} p_k(t)q_{ki}, t > 0, i = 0, 1, 2, \dots.$$

V dalším textu budeme používat značení $o(h)$, které se užívá pro libovolnou funkci argumentu h , pro kterou platí

$$\lim_{h \rightarrow 0} \frac{f(h)}{h} = 0.$$

Hodnoty $q_{ij}, i, j = 0, 1, 2, \dots,$ z poslední definice nazýváme **intenzity přechodu ze stavu i do stavu j** a dále platí

$$p_{ii}(h) = 1 + q_{ii}h + o(h), p_{ij}(h) = q_{ij}h + o(h), i \neq j.$$

8.4. Příklady

□ Poissonův proces

Už víme, že v tomto případě $X(t)$ udává počet výskytů sledované veličiny v intervalu $\langle 0, t \rangle$. V intervalu $(t, t + h)$, kde h je kladné číslo blízké nule, nastane nezávisle na počtu výskytů do času t sledovaná událost

- právě jednou s pravděpodobností $\lambda h + o(h)$,
- více než jednou s pravděpodobností $o(h)$,
- nenastane ani jednou s pravděpodobností $1 - \lambda h + o(h)$.

Tedy pravděpodobnosti přechodu se rovnají

$$p_{ii}(h) = 1 - \lambda h + o(h),$$

$$p_{i,i+1}(h) = \lambda h + o(h)$$

a

$$p_{ij}(h) = o(h), \quad j > i + 1.$$

Vidíme tedy, že pravděpodobnost jednoho výskytu události v krátkém intervalu je úměrná intenzitě procesu a délce intervalu. Dalším zjištěním je skutečnost, že pravděpodobnost dvou nebo více výskytů událostí klesá k nule s klesající délkou intervalu, a to rychleji než je délka intervalu.

Dle výsledků z minulé kapitoly pak spočteme

$$q_{ii} = \lim_{h \rightarrow 0_+} \frac{p_{ii}(h) - 1}{h} = \lim_{h \rightarrow 0_+} \frac{1 - \lambda h + o(h) - 1}{h} = \lim_{h \rightarrow 0_+} \frac{-\lambda h + o(h)}{h} = -\lambda,$$

$$q_{i,i+1} = \lim_{h \rightarrow 0_+} \frac{p_{i,i+1}(h)}{h} = \lim_{h \rightarrow 0_+} \frac{\lambda h + o(h)}{h} = \lambda,$$

a

$$q_{ij} = \lim_{h \rightarrow 0_+} \frac{p_{ij}(h)}{h} = \lim_{h \rightarrow 0_+} \frac{o(h)}{h} = 0, \quad j > i + 1.$$

Protože logicky nemůže nastat situace, že by byl počet výskytů události v intervalu $\langle 0, t \rangle$ větší než počet výskytů v intervalu $\langle 0, t + h \rangle$, položíme $p_{ij}(h) = 0$ pro $j < i$. Odtud pak máme $q_{ij} = 0$ pro $j < i$.

Nyní si již můžeme napsat Kolmogorovy diferenciální rovnice. Protože pro pevné i je q_{ki} nenulové pouze pro $k = i - 1$ a $k = i$, platí:

$$p_0'(t) = -\lambda p_0(t),$$

$$p_i'(t) = \lambda p_{i-1}(t) - \lambda p_i(t), \quad i = 1, 2, 3, \dots$$

Protože také požadujeme $X(0) = 0$, předepíšeme si počáteční podmínky:

$$p_0(0) = 1, \quad p_i(0) = 0, \quad i = 1, 2, 3, \dots$$

Řešení diferenciálních rovnic:

•

$$p_0'(t) + \lambda p_0(t) = 0, \quad p_0(0) = 1$$

Obecné řešení rovnice má tvar $p_0(t) = ce^{-\lambda t}$, $c \in \mathbb{R}$ a z počáteční podmínky plyne, že $c = 1$.

Řešením úlohy je tedy funkce $p_0(t) = e^{-\lambda t}$.

•

$$p_1'(t) + \lambda p_1(t) - \lambda e^{-\lambda t} = 0, \quad p_1(0) = 0$$

Řešením příslušné homogenní rovnice je funkce $p_1(t) = ce^{-\lambda t}$, $c \in \mathbb{R}$. Obecné řešení nalezneme pomocí variace konstanty. Dosazením do rovnice tedy dostaneme

$$c'(t)e^{-\lambda t} - \lambda c(t)e^{-\lambda t} + \lambda c(t)e^{-\lambda t} = \lambda e^{-\lambda t}.$$

Odtud plyne $c'(t) = \lambda$ a proto $c(t) = \lambda t + \tilde{c}$, $\tilde{c} \in \mathbb{R}$. Z počáteční podmínky plyne, že $\tilde{c} = 0$.

Řešením úlohy je tak funkce $p_1(t) = \lambda t e^{-\lambda t}$.

•

$$p_2'(t) + \lambda p_2(t) = \lambda^2 t e^{-\lambda t}, \quad p_2(0) = 0$$

Řešením příslušné homogenní rovnice je opět $p_2(t) = ce^{-\lambda t}$, $c \in \mathbb{R}$. Obecné řešení nalezneme pomocí variace konstanty. Po dosazení do rovnice obdržíme $c'(t) = \lambda^2 t$, a proto

$c(t) = \frac{\lambda^2 t^2}{2} + \tilde{c}$, $\tilde{c} \in \mathbb{R}$. Z počáteční podmínky znovu plyne, že $\tilde{c} = 0$. Řešením úlohy je

funkce $p_2(t) = \frac{\lambda^2 t^2}{2} e^{-\lambda t}$.

• ...

Výše uvedeným postupem získáme řešení ve tvaru

$$p_i(t) = \frac{(\lambda t)^i}{i!} e^{-\lambda t}, \quad i = 0, 1, 2, \dots$$

Vidíme tedy, že veličina $X(t)$ má Poissonovo rozdělení s parametrem λt . Jak jsme již zmínili, λ má zde význam středního počtu výskytů události za jednotku času. Číslo λ budeme nazývat **intenzitou Poissonova procesu**.

Díky vzájemné nezávislosti $X(t_1)$ a $X(t_2)$ při t_1 a t_2 velmi od sebe vzdálených můžeme psát

$$\lim_{h \rightarrow \infty} p_{ij}(h) = \pi_j.$$

Nyní vyjdeme z Chapmanovy-Kolmogorovovy rovnice:

$$p_{ij}(h_1 + h_2) = p_{ij}(h_1)p_{jj}(h_2) + \sum_{k \neq j} p_{ik}(h_1)p_{kj}(h_2)$$

a nechme $h_1 \rightarrow \infty$:

$$\pi_j = \pi_j p_{jj}(h_2) + \sum_{k \neq j} \pi_k p_{kj}(h_2).$$

Tedy

$$\pi_j (1 - p_{jj}(h_2)) = \sum_{k \neq j} \pi_k p_{kj}(h_2),$$

po vydělení h_2 a přechodem k $h_2 \rightarrow 0$ dostaneme

$$-\pi_j q_{jj} = \sum_{k \neq j} \pi_k q_{kj}, j = 0, 1, 2, \dots$$

Toto je soustava lineárních rovnic, kterou musí splňovat pravděpodobnosti π_j , pokud takové existují.

Pravděpodobnosti $\pi_j, j = 0, 1, \dots$, se nazývají **stacionární pravděpodobnosti**.

Poissonovým procesem modelujeme velmi často počet poruch na daném zařízení během určitého časového intervalu.

□ Proces růstu a zániku

Proces růstu a zániku je homogenní Markovův proces $\{X(t) : t \geq 0\}$ se stavy $0, 1, 2, \dots$. Veličina $X(t)$ udává četnost souboru (např. mikroorganismů, osob, ...) v čase t , přičemž během intervalu $(t, t+h)$, kde h je kladné číslo blízké nule, se soubor, který v čase t obsahuje i objektů, může

- zvětšit právě o jeden objekt s pravděpodobností $\lambda_i h + o(h)$, $i = 0, 1, \dots$,
- zmenšit právě o jeden objekt s pravděpodobností $\mu_i h + o(h)$, $i = 0, 1, \dots$,
- zmenšit nebo zvětšit o více než jeden objekt s pravděpodobností $o(h)$,
- nezmenšit ani nezvětšit s pravděpodobností $1 - \lambda_i h - \mu_i h + o(h)$.

Intenzity přechodu jsou pak rovny

$$q_{ii} = \lim_{h \rightarrow 0_+} \frac{p_{ii}(h) - 1}{h} = \lim_{h \rightarrow 0_+} \frac{1 - \lambda_i h - \mu_i h + o(h) - 1}{h} = \lim_{h \rightarrow 0_+} \frac{-\lambda_i h - \mu_i h + o(h)}{h} = -\lambda_i - \mu_i,$$

$$q_{i,i+1} = \lim_{h \rightarrow 0_+} \frac{p_{i,i+1}(h)}{h} = \lim_{h \rightarrow 0_+} \frac{\lambda_i h + o(h)}{h} = \lambda_i,$$

$$q_{i,i-1} = \lim_{h \rightarrow 0_+} \frac{p_{i,i-1}(h)}{h} = \lim_{h \rightarrow 0_+} \frac{\mu_i h + o(h)}{h} = \mu_i,$$

a

$$q_{ij} = \lim_{h \rightarrow 0_+} \frac{p_{ij}(h)}{h} = \lim_{h \rightarrow 0_+} \frac{o(h)}{h} = 0, \quad j > i + 1 \text{ nebo } j < i - 1.$$

Stacionární pravděpodobnosti $\pi_j, j = 0, 1, \dots$, pokud existují, jsou dány soustavou rovnic

$$\pi_0 \lambda_0 = \pi_1 \mu_1,$$

$$\pi_j (\lambda_j + \mu_j) = \pi_{j-1} \lambda_{j-1} + \pi_{j+1} \mu_{j+1}, \quad j = 1, 2, \dots$$

Tedy

$$\pi_j \mu_j - \pi_{j-1} \lambda_{j-1} = \pi_{j+1} \mu_{j+1} - \pi_j \lambda_j, \quad j = 1, 2, \dots$$

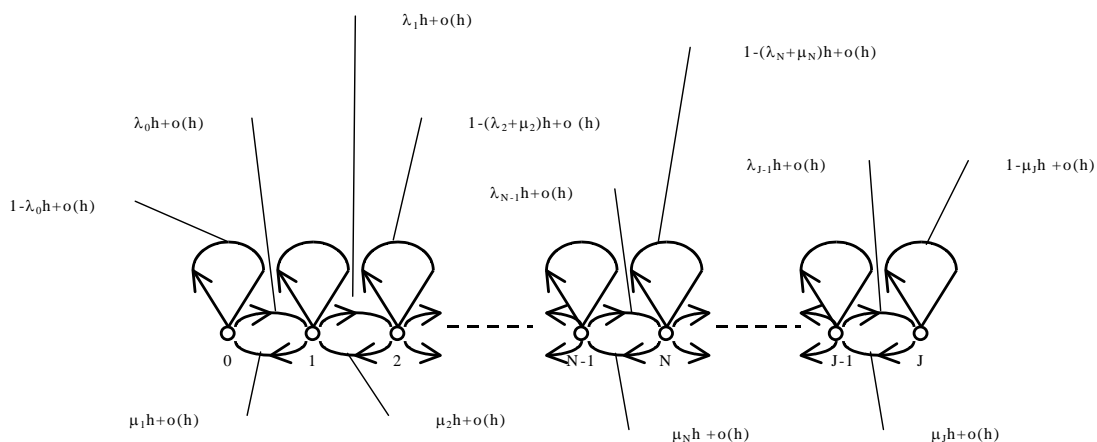
a dále pak

$$0 = \pi_1 \mu_1 - \pi_0 \lambda_0 = \pi_2 \mu_2 - \pi_1 \lambda_1 = \dots$$

Odtud

$$\pi_j \mu_j = \pi_{j-1} \lambda_{j-1}, \quad j \geq 1,$$

a proto platí rekurentní vztah



$$\pi_j = \frac{\lambda_{j-1}}{\mu_j} \pi_{j-1}.$$

Opakovaným užitím této rovnosti dostaneme

$$\pi_j = \pi_0 \prod_{k=1}^j \frac{\lambda_{k-1}}{\mu_k}.$$

Protože musí platit vztah $\sum_{j=0}^{\infty} \pi_j = 1$, obdržíme rovnost

$$\pi_0 + \pi_0 \frac{\lambda_0}{\mu_1} + \pi_0 \prod_{k=1}^2 \frac{\lambda_{k-1}}{\mu_k} + \dots = 1,$$

odtud

$$\pi_0 \left(1 + \sum_{j=1}^{\infty} \prod_{k=1}^j \frac{\lambda_{k-1}}{\mu_k} \right) = 1$$

a tedy konečně

$$\pi_0 = \left[1 + \sum_{j=1}^{\infty} \prod_{k=1}^j \frac{\lambda_{k-1}}{\mu_k} \right]^{-1}.$$

Může se stát, že bude řada ve výše uvedené rovnosti divergovat, tedy $\pi_0 = 0$ a $\pi_j = 0 \quad \forall j \in N$. Toto nastane například, pokud $\lambda_{j-1} > \mu_j \quad \forall j \in N$. Za takovýchto podmínek nebude existovat ustálený stav a soubor neustále poroste.

Neopomeňme ještě připomenout, že Poissonův proces je speciálním případem procesu růstu a zániku ($\mu_j = 0$ pro všechna j).

8.5. Markovovy řetězce

Obdobou Markovových procesů v diskrétním čase jsou Markovovy řetězce.

Nechť I označuje množinu $\{0,1,2,\dots\}$. Náhodná posloupnost $\{X_n : n = 0,1,2,\dots\}$ se nazývá **Markovův řetězec**, pokud platí

$$P(X_{n+1} = j | X_n = i, X_{n-1} = i_{n-1}, \dots, X_0 = i_0) = P(X_{n+1} = j | X_n = i)$$

pro libovolná $i_0, i_1, \dots, i_{n-1}, i, j \in I$ (markovská vlastnost).

Pokud pravděpodobnosti přechodu nezávisí na n , nazveme Markovův řetězec **homogenním** a píšeme

$$p_{ij} = P(X_{n+1} = j | X_n = i).$$

Pravděpodobnostmi přechodu vyšších řádů v homogenním Markovově řetězci rozumíme

$$p_{ij}(k) = P(X_{n+k} = j | X_n = i), k = 0, 1, 2, \dots$$

V homogenním Markovově řetězci platí (tzv. Chapmanovy-Kolmogorovovy rovnice)

$$p_{ij}(k_1 + k_2) = \sum_{k=0}^{\infty} p_{ik}(k_1) p_{kj}(k_2).$$

Tedy pravděpodobnost, že systém přešel ze stavu i do nějakého mezistavu k přes r přechodů a z mezistavu k se dostal do koncového stavu j v $(n-r)$ přechodech mezi stavy, je vyjádřena vztahem

$$p_{ij}(n) = \sum_{k=0}^{\infty} p_{ik}(r) p_{kj}(n-r).$$

Speciálně pro $n = 0$ platí

$$p_{kj}(0) = \begin{cases} 1, & k = j \\ 0, & \text{jinak} \end{cases},$$

a pro $n = 1$ platí

$$p_{ij}(1) = p_{ij}.$$

Mějme Markovův řetězec s m možnými stavy. matici $P = \{p_{ij}\}_{i,j=1}^m$ nazveme **maticí pravděpodobností přechodu**.

Vlastnosti matice P :

- P je čtvercová matice $m \times m$,
- $p_{ij} \in \langle 0, 1 \rangle$,
- součet prvků v každém řádku matice je jednotkový.

Protože (při $r = 1$) platí

$$p_{ij}(2) = \sum_k p_{ik} p_{kj} = (i, j)\text{-tý prvek matice } P^2,$$

$$p_{ij}(3) = \sum_k p_{ik} p_{kj}(2) = (i, j)\text{-tý prvek matice } P \times P^2 = P^3,$$

objasnili jsme následující tvrzení.

V homogenním Markovově řetězci platí

$$P(k) = P^k, k = 0, 1, 2, \dots$$

Stav j Markovova řetězce je dosažitelný ze stavu i , pokud pro nějaké $n \in N \cup \{0\}$ platí

$$p_{ij}(n) > 0.$$

Je-li každý stav řetězce dosažitelný z každého stavu, nazveme řetězec **neredukovatelným**.

Stav i je **periodický** s periodou $d > 1$, pokud $p_{ii}(n) > 0$ jen pro $n = d, 2d, 3d, \dots$, přičemž číslo d je nejmenší číslo s touto vlastností. Je-li $d = 1$, je stav **aperiodický**.

Pro každý stav i definujme pravděpodobnost $f_i(n)$, že první návrat do stavu i nastává po n přechodech mezi stavy, tedy

$$f_i(n) = p[x_n = i, x_k \neq i \text{ pro } 1, 2, \dots, n-1 | x_0 = i].$$

Nyní definujme pravděpodobnost, že se systém po opuštění stavu i do něj znovu vrátí, rovností

$$f_i = \sum_{n=1}^{\infty} f_i(n).$$

Je-li dále $f_i < 1$, nazývá se stav **tranzientní**, a je-li $f_i = 1$, nazveme stav **rekurentním**.

Stavy i a j spolu komunikují, pokud i je dosažitelný z j a naopak j je dosažitelný z i . Píšeme $i \leftrightarrow j$ ve smyslu ekvivalence s těmito vlastnostmi:

1. $i \leftrightarrow i$ pro každý stav i ,
2. $(i \leftrightarrow j) \Rightarrow (j \leftrightarrow i)$,
3. $[(i \leftrightarrow j) \text{ a } (j \leftrightarrow k)] \Rightarrow (i \leftrightarrow k)$.

Stav i nazveme **absorbující**, pokud již systém (po vstoupení do tohoto stavu) v tomto stavu zůstane až do konce, tj. $p_{ii} = 1$. Absorbující stav je ekvivalentní (ve výše uvedeném smyslu) pouze sám se sebou a je speciálním případem rekurentního stavu.

Nyní se budeme zabývat konečnými Markovovými řetězci s rekurentními a tranzientními stavy. Nechť tedy má Markovův řetězec m stavů, z nichž prvních r stavů je rekurentních a dalších $m - r$ stavů je tranzientních, přičemž rekurentní stavy tvoří jednu třídu ekvivalence a tranzientní druhou. Označme dále T množinu tranzientních stavů a T^c množinu rekurentních stavů. Pak matice pravděpodobností přechodu má tvar

$$P_{(m \times m)} = \begin{bmatrix} P_{1(r \times r)} & 0_{(r \times (m-r))} \\ R_{((m-r) \times r)} & Q_{((m-r) \times (m-r))} \end{bmatrix}.$$

Maticе P_1 je maticí pravděpodobností přechodů mezi rekurentními stavy. Protože platí tvrzení (i je rekurentní a $i \leftrightarrow j$) \Rightarrow (j je rekurentní), nemůže systém po vstupu do rekurentního stavu opustit rekurentní třídu a máme tak napravo od P_1 nulovou matici. Maticе R je maticí pravděpodobností přechodu z tranzientního stavu do rekurentního a maticе Q značí maticí pravděpodobností přechodu mezi tranzientními stavy. Při studiu těchto Markovových řetězců se budeme ptát zejména na tyto otázky:

- Začíná-li řetězec v tranzientním stavu i , jaký je průměrný počet návštěv tranzientního stavu j , než se konečně systém dostane do rekurentního stavu?
- Jaký je pak rozptyl počtu návštěv tranzientního stavu j ?
- Jaký je průměrný počet návštěv tranzientních stavů potřebný k opuštění tranzientní třídy při počátečním tranzientním stavu i ?
- Jaký je rozptyl počtu návštěv tranzientních stavů při počátečním tranzientním stavu i , než se systém dostane do rekurentního stavu?

Maticе M danou předpisem

$$M = (I - Q)^{-1}$$

nazveme **fundamentální** maticí.

Dá se ukázat, že $(I - Q)^{-1}$ existuje a platí

$$(I - Q)^{-1} = I + Q + Q^2 + \dots = \sum_{k=0}^{\infty} Q^k .$$

Necht' N_{ij} označuje náhodnou veličinu reprezentující počet návštěv tranzientního stavu $j \in T$ (při počátečním tranzientním stavu $i \in T$) před vstupem řetězce do rekurentního stavu. Označme dále $\overset{\text{ozn}}{\mu}_{ij} = EN_{ij}$. Pak pro každé $i, j \in T$ platí

$$\|\mu_{ij}\| = M ,$$

kde $\|\mu_{ij}\|$ označuje maticí s prvky μ_{ij} , $i, j = r + 1, r + 2, \dots, m$.

Označme

$$M_D \overset{\text{ozn}}{=} \begin{bmatrix} \mu_{r+1,r+1} & & & 0 \\ & \mu_{r+2,r+2} & & \vdots \\ & & \ddots & \\ 0 & \dots & & \mu_{m,m} \end{bmatrix}, M_2 \overset{\text{ozn}}{=} \|\mu_{ij}^2\| .$$

Necht' $\sigma_{ij}^2 = \text{Var}(N_{ij})$ značí rozptyl počtu návštěv tranzientního stavu j při počátečním tranzientním stavu i před vstupem do rekurentního stavu. Pak platí

$$\|\sigma_{ij}^2\| = M(2M_D - I) - M_2, \quad i, j \in T.$$

Dále značme N_i náhodnou veličinu reprezentující celkový počet navštívených tranzientních stavů při počátečním tranzientním stavu i před vstupem do rekurentního, tj. $N_i = \sum_{j \in T} N_{ij}$. Pak

$$EN_i = E\left(\sum_{j \in T} N_{ij}\right) = \sum_{j \in T} EN_{ij} = \sum_{j \in T} \mu_{ij}.$$

Značme $M_{\rho} = \left\| \sum_{j \in T} \mu_{ij} \right\|$, M_{ρ} je tedy sloupcový vektor, a $M_{\rho 2} = \left\| \left(\sum_{j \in T} \mu_{ij} \right)^2 \right\|$. Dá se dále ukázat platnost vztahu

$$\|\text{Var}(N_i)\| = (2M - I)M_{\rho} - M_{\rho 2}, \quad i \in T,$$

kde $\text{Var}(N_i)$ je rozptyl celkového počtu navštívených tranzientních stavů při počátečním tranzientním stavu i do opuštění tranzientní třídy stavů.

Necht' $f_{ij}(n)$ značí pravděpodobnost, že při počátečním tranzientním stavu i vstoupí řetězec do rekurentního stavu j v n krocích. Označme dále T_{ij} celkový počet navštívených tranzientních stavů před prvním vstupem do rekurentního stavu j při počátečním tranzientním stavu i , tj. $P(T_{ij} = n) = f_{ij}(n)$, $i \in T, j \in T^C$. Pravděpodobnost, že řetězec vstoupí do rekurentního stavu j , pak vyjádříme jako $f_{ij} = \sum_{n=1}^{\infty} f_{ij}(n)$. V maticovém zápisu pak

$$F(n) = \|f_{ij}(n)\|, \quad F = \|f_{ij}\|.$$

Platí následující tvrzení:

$$F(n) = Q^{n-1}R \quad \text{a} \quad F = MR.$$

Důkaz: Víme, že

$$f_{ij}(1) = p_{ij} \quad \text{a} \quad f_{ij}(n) = \sum_{k \in T} p_{ik} f_{kj}(n-1), \quad i \in T, j \in T^C.$$

V maticovém zápisu pak máme

$$F(1) = R \quad \text{a} \quad F(n) = QF(n-1).$$

Tedy

$$F(n) = Q^{n-1}R.$$

Dále

$$F = \sum_{n=1}^{\infty} F(n) = R + \sum_{n=2}^{\infty} Q^{n-1}R = (I + Q + Q^2 + \dots)R = MR.$$



Řešený příklad

Máme zadánu matici pravděpodobností přechodu P třístavového systému

$$P = \begin{bmatrix} 0,8 & 0,2 & 0 \\ 0 & 0,5 & 0,5 \\ 0 & 0 & 1 \end{bmatrix}.$$

Třetí stav je absorbuující ($p_{33} = 1$) a první a druhý stav jsou tranzientní. Nalezněte:

- průměrný počet návštěv tranzientních stavů při počátečním stavu 1 před vstupem do absorbuujícího stavu 3,
- průměrný počet návštěv stavu 2 při počátečním stavu 1 před vstupem do absorbuujícího stavu.

Submatice matice P vyjadřující pravděpodobnosti přechodu mezi tranzientními stavy má tvar

$$Q = \begin{bmatrix} 0,8 & 0,2 \\ 0 & 0,5 \end{bmatrix}.$$

Protože fundamentální matici M spočteme jako inverzní matici k matici $[I - Q]$, platí

$$M = \begin{bmatrix} 0,2 & -0,2 \\ 0 & 0,5 \end{bmatrix}^{-1} = \begin{bmatrix} 5 & 2 \\ 0 & 2 \end{bmatrix}.$$

Víme, že prvky μ_{ij} fundamentální matice M jsou průměrné počty návštěv tranzientního stavu j při počátečním tranzientním stavu i před vstupem do rekurentního stavu. Odtud průměrný počet návštěv tranzientních stavů při počátečním stavu i je dán součtem $\sum_{j \in T} \mu_{ij}$. V našem případě, hledáme-li řešení prvního úkolu, máme

$$\sum_{j=1}^2 \mu_{1j} = 7.$$

Odpověď na druhý úkol je již jednoduchá. Hledáme vlastně μ_{12} a to je rovno 2.

Pravděpodobnosti popisující rozdělení Markovova řetězce s m možnými stavy v čase n označme

$$p_i(n) = P(X_n = i), i = 1, 2, \dots, m.$$

V homogenním Markovově procesu platí

$$p_i(k) = \sum_{j=0}^{\infty} p_j(0) p_{ij}(k), k = 0, 1, 2, \dots$$

Jestliže označíme vektory

$$\tilde{P}(0) = (p_1(0), p_2(0), \dots, p_m(0)) \text{ a } \tilde{P}(n) = (p_1(n), p_2(n), \dots, p_m(n)),$$

pak, s využitím vztahu $P(n) = P^n$, můžeme dle předchozí věty psát

$$\tilde{P}(n) = \tilde{P}(0)P^n.$$

Nyní položíme $n \rightarrow \infty$. Je-li matice P regulární, existuje $\lim_{n \rightarrow \infty} \tilde{P}(n)$. Označíme-li pak tuto limitu Y , musí pro ni platit $Y = YP$. Y nazýváme vektorem stacionárních pravděpodobností. Vidíme, že Y nezávisí na počátečním rozdělení pravděpodobnosti Markovova řetězce.

Protože nás velmi často zajímá vektor rozdělení pravděpodobnosti po n krocích $\tilde{P}(n)$, je nutné vypočítat matici P^n , což nemusí být vždy jednoduché. Ukážeme si nyní 2 základní principy výpočtu.

Algebraický přístup.

Matice P řádu $m \times m$. Má-li matice P m různých vlastních čísel k_1, k_2, \dots, k_m , pak existuje regulární matice R taková, že $P = RDR^{-1}$, přičemž D je diagonální matice mající na diagonále postupně vlastní čísla matice P a i -tý sloupec matice R je tvořen vlastním vektorem příslušným vlastnímu číslu k_i . Dále platí rovnost $P^n = RD^nR^{-1}$. Protože D je diagonální, D^n se spočítá snadno.



Řešený příklad

System má matici pravděpodobností přechodu

$$P = \begin{bmatrix} 0,5 & 0,5 \\ 0,75 & 0,25 \end{bmatrix}.$$

Nalezněte $\tilde{P}(n)$ při počátečním rozdělení $\tilde{P}(0) = (1; 0)$.

Nejprve nalezneme vlastní čísla matice P . Z rovnosti $\det[\lambda I - P] = 0$ vypočteme vlastní čísla $\lambda_1 = 0,9797$, $\lambda_2 = -0,2297$. Pro vlastní vektor v_1 řeší rovnost $[\lambda_1 I - P] v_1 = 0$

platí $v_1 = \left(-\frac{0,5t}{0,4797}; t \right)$, $t \in \mathbb{R} - \{0\}$. Aby norma vektoru v_1 byla jednotková, zvolíme $t = 0,6923$ a tedy $v_1 = (0,7216; 0,6923)$. Stejným způsobem získáme druhý vlastní vektor $v_2 = (-0,5653; 0,8249)$. Máme tedy

$$D = \begin{bmatrix} 0,9797 & 0 \\ 0 & -0,2297 \end{bmatrix}, R = \begin{bmatrix} 0,7216 & -0,5653 \\ 0,6923 & 0,8249 \end{bmatrix}$$

Spočteme dále inverzní matici

$$R^{-1} = \begin{bmatrix} 0,8361 & 0,5729 \\ -0,7017 & 0,7314 \end{bmatrix}.$$

Platí tedy

$$P^n = \begin{bmatrix} 0,7216 & -0,5653 \\ 0,6923 & 0,8249 \end{bmatrix} \begin{bmatrix} 0,9797^n & 0 \\ 0 & -0,2297^n \end{bmatrix} \begin{bmatrix} 0,8361 & 0,5729 \\ -0,7017 & 0,7314 \end{bmatrix}$$

a ze vztahu $\tilde{P}(n) = \tilde{P}(0)P^n$ dopočítáme

$$\tilde{P}(n) = (0,6033 \cdot 0,9797^n - 0,3967 \cdot 0,2297^n; 0,4134 \cdot 0,9797^n + 0,4135 \cdot 0,2297^n).$$

□ Přístup Z-transformace

Protože platí sled rovností $\tilde{P}(n+1) = \tilde{P}(0)P^{n+1} = \tilde{P}(0)P^n P = \tilde{P}(n)P$, můžeme s užitím Z-transformace psát

$$\frac{Z\{\tilde{P}(n)\} - \tilde{P}(0)}{z} = Z\{\tilde{P}(n)\}P.$$

Odtud dostáváme

$$Z\{\tilde{P}(n)\}[I - zP] = \tilde{P}(0)$$

a dále

$$Z\{\tilde{P}(n)\} = \tilde{P}(0)[I - zP]^{-1}.$$

Porovnáním se vztahem $\tilde{P}(n) = \tilde{P}(0)P^n$ zjistíme, že

$$Z\{P^n\} = [I - zP]^{-1}.$$

Takže P^n obdržíme pomocí zpětné Z-transformace matice $[I - zP]^{-1}$.



Shrnutí kapitoly 8.

Náhodným (stochastickým) procesem nazveme zobrazení, které každé hodnotě $t \in T$ přiřadí náhodnou veličinu $X(t)$. Proměnná t má ve většině případů význam času.

Realizací náhodného procesu rozumíme konkrétní pozorování náhodného procesu, tj. již nenáhodnou funkci, a značíme ji $x(t)$.

Dle povahy množiny T rozlišujeme:

- **náhodné procesy se spojitým časem** (náhodné funkce),
- **náhodné procesy s diskrétním časem** (náhodné posloupnosti)

Hodnota $X(t)$ vyjadřuje stav pozorovaného objektu v čase t . Dle povahy náhodné veličiny $X(t)$ rozlišujeme:

- **náhodné procesy se spojitými stavy** - $X(t)$ je spojitá,
- **náhodné procesy s diskrétními stavy** - $X(t)$ je diskrétní.

Náhodný proces $\{X(t) : t \geq 0\}$ se spojitým časem a s diskrétními stavy $0, 1, 2, \dots$ obvykle nazýváme **čítací proces**, protože zaznamenává počet nějakých událostí v čase.

Nechť $\{X(t) : t \geq 0\}$ je čítací proces. Nechť navíc platí:

$$X(0) = 0,$$

délky intervalů mezi výskyty sledované události jsou nezávislé náhodné veličiny s exponenciálním rozdělením s hustotou

$$f(x) = \begin{cases} \lambda e^{-\lambda x} & \text{pro } x > 0, \\ 0 & \text{pro } x \leq 0, \end{cases}$$

kde $\lambda > 0$ je parametr (tzv. intenzita homogenního Poissonova procesu).

Pak tento proces nazveme **homogenním Poissonovým procesem**, přičemž $X(t)$ má Poissonovo rozdělení s parametrem λt .

Proces $\{X(t) : t \geq 0\}$ nazveme **Markovovým procesem**, splňuje-li tzv. markovskou vlastnost:

pro libovolná $0 \leq t_1 < t_2 < \dots < t_n < t \leq \tau$ a $i_1, \dots, i_n, i, j \in I$ platí:

$$P(X(\tau) = j | X(t) = i, X(t_n) = i_n, \dots, X(t_1) = i_1) = P(X(\tau) = j | X(t) = i).$$

V homogenním Markovově procesu platí

- $p_{ij}(t_1 + t_2) = \sum_{k=0}^{\infty} p_{ik}(t_1) p_{kj}(t_2), t_1, t_2 \geq 0, i, j = 0, 1, 2, \dots,$
- $p_i(t) = \sum_{j=0}^{\infty} p_j(0) p_{ji}(t), t \geq 0, i = 0, 1, 2, \dots.$

První rovnice se nazývá **Chapmanova-Kolmogorovova rovnice**.

Proces růstu a zániku je homogenní Markovův proces $\{X(t) : t \geq 0\}$ se stavy $0, 1, 2, \dots$.

Nechť I označuje množinu $\{0, 1, 2, \dots\}$. Náhodná posloupnost $\{X_n : n = 0, 1, 2, \dots\}$ se nazývá **Markovův řetězec**, pokud platí

$$P(X_{n+1} = j | X_n = i, X_{n-1} = i_{n-1}, \dots, X_0 = i_0) = P(X_{n+1} = j | X_n = i)$$

pro libovolná $i_0, i_1, \dots, i_{n-1}, i, j \in I$ (markovská vlastnost).



Otázky 8.

1. Vysvětlete pojem náhodný proces a popište typy náhodných procesů.
2. Definiujte Poissonův proces.
3. Definiujte Markovův proces.
4. Co jsou to procesy růstu a zániku?
5. Odvoďte stacionární pravděpodobnosti (pravděpodobnosti stavů).
6. Vysvětlete pojem Markovův řetězec.
7. Vysvětlete pojmy: neredukovatelný řetězec, periodický stav, aperiodický stav, tranzientní stav, rekurentní stav, absorbující stav.
8. K čemu slouží fundamentální matice?

9. NÁHODNÉ PROCESY II



Čas ke studiu: 1,5 hodiny



Cíl: Po prostudování tohoto odstavce budete umět:

- Naučíme se určovat pravděpodobnosti výskytu systému v jednotlivých stavech v daném čase t a seznámíme se s kongruentní metodou generování náhodných čísel.



Výklad

9.1. Spojitý parametr Markovovských řetězců

Nechť

$$\{X_t : t \geq 0\}$$

reprezentuje spojitý náhodný parametr Markovova řetězce s m diskrétními stavy. Pro $t > s > 0$ a stavy i a j nechť

$$P\{X_t = j\} = p_j(t),$$

což je pravděpodobnost, že proces je ve stavu j v čase t , a

$$P\{X_t = j | X_s = i\} = p_{ij}(s, t),$$

což je pravděpodobnost, že proces je ve stavu j v čase t a byl dán procesem, který byl ve stavu i v čase s .

Pravděpodobnost $p_{ij}(s, t)$ je nazývána **přechodovou pravděpodobnostní funkcí Markovova řetězce**. Markovův řetězec je homogenní nebo stacionární (v souvislosti s časem), jestliže $p_{ij}(s, t)$ závisí pouze na časovém intervalu $t' = t - s$. To vyhovuje Chapman-Kolmogorově rovnici, která je dána

$$p_{ij}(s, t) = \sum_{\text{stavy } k} p_{ik}(s, u) p_{kj}(u, t)$$

pro jakýkoliv čas $t > u > s \geq 0$ a stavy i a j . Po úpravě spočívající v nahrazení $t - s$ jako t' a $u - s$ jako u' se rovnice redukuje na

$$p_{ij}(t') = \sum_{\text{stavy } k} p_{ik}(u') p_{kj}(t' - u'),$$

dokud je proces homogenní.

Zde $p_{ij}(t)$ může být interpretována jako pravděpodobnost, že proces přejde ze stavu i do stavu j v časovém intervalu t . Jestliže

$$\lim_{t \rightarrow \infty} p_{ij}(t) = \begin{cases} 1, & \text{pro } j = i \\ 0, & \text{pro } j \neq i \end{cases},$$

říkáme, že Markovův řetězec je spojitý v $t = 0$. Budeme brát v úvahu pouze homogenní Markovovy řetězce.

Definujme dva přechody intenzit z hlediska derivací v $t = 0$, které hrají stejnou roli jako jeden krok přechodu pravděpodobností v diskrétním parametru Markovova řetězce.

Pro každý stav i předpokládáme

$$\left. \frac{d}{dt} p_{ii}(t) \right|_{t=0} = \lim_{t \rightarrow 0} \frac{1 - p_{ii}(t)}{0 - t},$$

která existuje a je konečná.

Pro všechna i a j , kde $i \neq j$, předpokládáme

$$\left. \frac{d}{dt} p_{ij}(t) \right|_{t=0} = \lim_{t \rightarrow 0} \frac{0 - p_{ij}(t)}{0 - t},$$

která existuje a je konečná.

Necht'

$$d_{ii} = \left. \frac{-d}{dt} p_{ii}(t) \right|_{t=0}$$

a pro $i \neq j$

$$d_{ij} = \left. \frac{d}{dt} p_{ij}(t) \right|_{t=0}.$$

Úpravami předchozích vztahů dostaneme

$$\frac{1 - p_{ii}(t)}{t} \Big|_{t=0} = d_{ii} + r_i(t),$$

kde $r_i(t) \rightarrow 0$ a $t \rightarrow 0$. Potom

$$1 - p_{ii}(t) = d_{ii}t + tr_i(t) = d_{ii}t + o(t).$$

Tato rovnice může být interpretována jako: pravděpodobnost přechodu ze stavu i do nějakého jiného stavu během časového intervalu t a je rovna $d_{ii}t + o(t)$.

Podobně můžeme psát

$$p_{ij}(t) = d_{ij}t + o(t).$$

Tento vztah může být interpretován jako pravděpodobnost přechodu ze stavu i do stavu j v časovém intervalu t , která je rovna $d_{ij}t + 0(t)$. Jako výsledek máme pro libovolné $t \geq 0$ a $h > 0$

$$p_{ij}(t+h) = p_{ij}(t)\{1 - (d_{ij}h + 0(h))\} + \sum_{k \neq j} p_{ik}(t)d_{kj}h + 0(h)$$

nebo

$$\frac{p_{ij}(t+h) - p_{ij}(t)}{h} = -d_{ij}p_{ij}(t) + \sum_{k \neq j} p_{ik}(t)d_{kj} + \frac{0(h)}{h}.$$

Předpokládáme, že $p'_{ij}(t)$ existuje a $h \rightarrow 0$, pak obdržíme

$$p'_{ij}(t) = \lim_{h \rightarrow 0} \frac{p_{ij}(t+h) - p_{ij}(t)}{h} = -d_{ij}p_{ij}(t) + \sum_{k \neq j} p_{ik}(t)d_{kj}.$$

Pro všechny stavy i a j tato rovnice dává systém diferenciálních rovnic, jejichž řešením získáme pravděpodobnosti přechodu.

K získání bezpodmínečných stavových pravděpodobností $p'_i(t)$ pro každý stav i přijmeme stejný důvod, který byl již použit v odvození předcházejících rovnic. Takže máme

$$p_i(t+h) = p_i(t)(1 - d_{ii}h) + \sum_{j \neq i} p_j(t)d_{ji}h + 0(h),$$

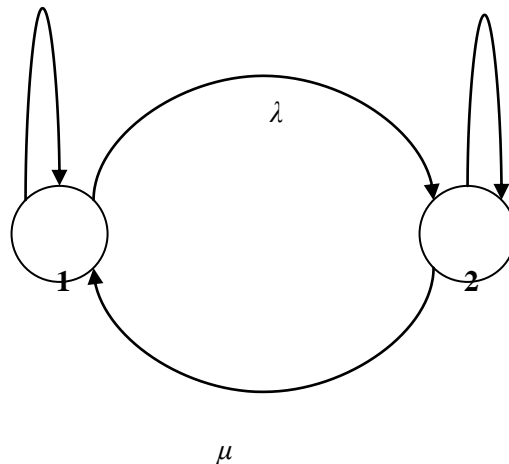
ze které získáme

$$p'_i(t) = -d_{ii}p_i(t) + \sum_{j \neq i} p_j(t)d_{ji}.$$

Tento výraz definuje systém rovnic, který je lineární z hlediska Laplaceovy transformace proměnných a může být řešen standardními technikami.

9.2. Ilustrace

Uvažujme systém zobrazený na obrázku, jenž se pohybuje mezi stavy 1 a 2. Rozložení přechodu v čase ze stavu 1 do stavu 2 je $\lambda e^{-\lambda t}$ a ze stavu 2 do stavu 1 $\mu e^{-\mu t}$. Úkolem bude určit pravděpodobnost, že systém bude ve stavu 1, popř. 2 v jakémkoliv daném čase t .



Z předchozího textu víme, že

$$d_{ii} = \left. \frac{-d}{dt} p_{ii}(t) \right|_{t=0}.$$

Výraz $p_{11}(t)$ je pro nás neznámý. Avšak může být určen z ekvivalentních relací.

Pravděpodobnost, že systém zůstane ve stavu i v $t = 0^+$, je dána tím, že systém je ve stavu i v $t = 0$

$$d_{ii} = -\frac{d}{dt}.$$

Podobně pravděpodobnost, že systém přejde ze stavu i do stavu j v $t = 0^+$ je dána tím, že je systém ve stavu i v $t = 0$

$$d_{ij} = \frac{d}{dt}.$$

Dále

$$d_{11} = -\left. \frac{d}{dt} (e^{-\lambda t}) \right|_{t=0} = \lambda$$

$$d_{12} = -\left. \frac{d}{dt} (1 - e^{-\lambda t}) \right|_{t=0} = \lambda$$

a podobně

$$d_{22} = \mu$$

$$d_{21} = \mu.$$

Dále můžeme psát

$$p'_{ij}(t) = -d_{ij} p_{ij}(t) + \sum_{k \neq j} p_{ik}(t) d_{kj},$$

použitím Laplaceovy transformace obdržíme

$$sP_{ij}(s) - P_{ij}(0) = -d_{ij} P_{ij}(s) + \sum_{k \neq j} P_{ik}(s) d_{kj},$$

kde

$$P_{ij}(0) = 1 \text{ pro } i = j \text{ a } P_{ij}(0) = 0 \text{ pro } i \neq j.$$

Potom

$$sP_{11}(s) - 1 = -\lambda P_{11}(s) + \mu P_{12}(s)$$

$$sP_{12}(s) = -\mu P_{12}(s) + \lambda P_{11}(s)$$

$$sP_{22}(s) - 1 = -\mu P_{22}(s) + \lambda P_{21}(s)$$

$$sP_{21}(s) = -\lambda P_{21}(s) + \mu P_{22}(s).$$

Zapsáno v maticové formě

$$\begin{pmatrix} s + \lambda & -\mu & 0 & 0 \\ -\lambda & s + \mu & 0 & 0 \\ 0 & 0 & s + \mu & -\lambda \\ 0 & 0 & -\mu & s + \lambda \end{pmatrix} \begin{pmatrix} P_{11}(s) \\ P_{12}(s) \\ P_{22}(s) \\ P_{21}(s) \end{pmatrix} = \begin{pmatrix} 1 \\ 0 \\ 1 \\ 0 \end{pmatrix}.$$

Po vyřešení a provedení inverzní transformace získáme

$$p_{11}(t) = \frac{\mu}{\lambda + \mu} + \frac{\lambda}{\lambda + \mu} e^{-(\lambda + \mu)t}$$

$$p_{12}(t) = \frac{\lambda}{\lambda + \mu} - \frac{\lambda}{\lambda + \mu} e^{-(\lambda + \mu)t}$$

$$p_{22}(t) = \frac{\lambda}{\lambda + \mu} + \frac{\mu}{\lambda + \mu} e^{-(\lambda + \mu)t}$$

$$p_{21}(t) = \frac{\mu}{\lambda + \mu} - \frac{\mu}{\lambda + \mu} e^{-(\lambda + \mu)t}$$

Nyní najdeme bezpodmínečné stavy pravděpodobností $p_i(t)$ pro každý stav i . Můžeme tedy psát

$$p'_j(t) = -d_{jj} p_j(t) + \sum_{i \neq j} p_i(t) d_{ij}.$$

Po použití Laplaceovy transformace na obou stranách rovnice získáme

$$sP_j(s) - p_j(0) = -d_{jj} P_j(s) + \sum_{i \neq j} d_{ij} P_i(s),$$

tj.

$$P_j(s) = \frac{1}{s + d_j} \left[p_j(0) + \sum_{i \neq j} P_i(s) d_{ij} \right].$$

Vezmeme $p_1 = 0$ a $p_2 = q = 1 - p$. Máme

$$d_{11} = \lambda$$

$$d_{12} = \lambda$$

$$d_{22} = \mu$$

$$d_{21} = \mu$$

$$P_1(s) = \frac{1}{s + \lambda} [p + \mu P_2(s)]$$

$$P_2(s) = \frac{1}{s + \mu} [q + \lambda P_1(s)]$$

na řešení,

$$P_1(s) = \frac{(s + \mu)p + \mu q}{s(s + \mu + \lambda)}$$

$$P_2(s) = \frac{q(s + \lambda) + \lambda p}{s(s + \mu + \lambda)}$$

po inverzní transformaci získáme

$$p_1(t) = \frac{1}{\lambda + \mu} \left[\mu - (\mu - (\mu + \lambda))pe^{-(\mu + \lambda)t} \right]$$

$$p_2(t) = \frac{1}{\lambda + \mu} \left[\lambda + (\mu - (\mu + \lambda))pe^{-(\mu + \lambda)t} \right].$$

9.3. Matice hustoty přechodu

Matice $A = \|d_{ij}\|$ se nazývá **matice hustoty přechodu** nebo **matice míry přechodu procesu**. Tato matice má následující vlastnosti:

1. její nediagonální prvky jsou nezáporné a diagonální prvky jsou záporné,
2. suma prvků v každém řádku je rovna nule, suma nediagonálních prvků je rovna sumě diagonálních prvků s opačným znaménkem.

Pro systém uvažovaný v předchozí ilustraci bude systém diferenciálních rovnic v maticové formě vypadat následovně:

$$\begin{pmatrix} p'_{11}(t), p'_{12}(t), p'_{22}(t), p'_{21}(t) \end{pmatrix} = \begin{pmatrix} p_{11}(t), p_{12}(t), p_{22}(t), p_{21}(t) \end{pmatrix} \begin{vmatrix} -\lambda & \lambda & 0 & 0 \\ \mu & -\mu & 0 & 0 \\ 0 & 0 & -\mu & \mu \\ 0 & 0 & \lambda & -\lambda \end{vmatrix},$$

kde

$$\begin{vmatrix} -\lambda & \lambda & 0 & 0 \\ \mu & -\mu & 0 & 0 \\ 0 & 0 & -\mu & \mu \\ 0 & 0 & \lambda & -\lambda \end{vmatrix}$$

je matice hustoty přechodu A .

Také pro bezpodmínečný stav pravděpodobností máme

$$\begin{pmatrix} p'_1(t), p'_2(t) \end{pmatrix} = \begin{pmatrix} p_1(t), p_2(t) \end{pmatrix} \begin{vmatrix} -\lambda & \lambda \\ \mu & -\mu \end{vmatrix}.$$

Zde je matice míry přechodu A

$$\begin{vmatrix} -\lambda & \lambda \\ \mu & -\mu \end{vmatrix}.$$

Matice $P = I + A$ hraje stejnou úlohu jako jeden krok v matici přechodu pravděpodobnosti u diskrétního parametru Markovových řetězců. Jestliže prvky matice A jsou konstanty, pak věty zmíněné v poslední části diskrétního parametru Markovova řetězce jsou také aplikovatelné na spojitý parametr Markovova řetězce s tou změnou, že počet návštěv daného stavu se stane celkovým časem, který byl ve stavu stráven.

Pro řetězec, který byl uvažován v ilustraci, máme

$$P = I + A = I + \begin{vmatrix} -\lambda & \lambda \\ \mu & -\mu \end{vmatrix} = \begin{vmatrix} 1-\lambda & \lambda \\ \mu & 1-\mu \end{vmatrix}.$$

Povšimněte si, že suma řádkových prvků v matici P je rovna 1 jako u matice přechodu pravděpodobnosti diskrétního parametru Markovova řetězce. Avšak zde P není matice pravděpodobnosti.



Řešený příklad

Mějme systém dán maticí P

$$P = \begin{array}{c} \begin{array}{ccc} 1 & 2 & 3 \end{array} \\ \begin{array}{ccc} 1-2\lambda & 2\lambda & 0 \\ \mu & 1-(\mu+\lambda) & \lambda \\ 0 & 0 & 1 \end{array} \end{array} \begin{array}{l} 1 \\ 2 \\ 3 \end{array}$$

Nalezněte:

- průměrný čas za který se systém dostane do konečného stavu, je-li počáteční stav 1,
- průměrný čas strávený ve stavu 2 předtím, než se dostane systém do konečného stavu (počáteční stav je 1).

Máme

$$Q = \begin{array}{c} \begin{array}{cc} 1 & 2 \end{array} \\ \begin{array}{cc} 1-2\lambda & 2\lambda \\ \mu & 1-\lambda-\mu \end{array} \end{array}$$

$$I - Q = \begin{array}{c} \begin{array}{cc} 1 & 0 \\ 0 & 1 \end{array} \end{array} - \begin{array}{c} \begin{array}{cc} 1-2\lambda & 2\lambda \\ \mu & 1-\lambda-\mu \end{array} \end{array} = \begin{array}{c} \begin{array}{cc} 2\lambda & -2\lambda \\ -\mu & \lambda+\mu \end{array} \end{array}$$

$$M = \|\mu_{ij}\| = [I - O]^{-1} = \frac{1}{2\lambda^2} \begin{vmatrix} \lambda + \mu & 2\lambda \\ \mu & 2\lambda \end{vmatrix}$$

a

$$M_{\rho} = \frac{1}{2\lambda^2} \begin{vmatrix} 3\lambda + \mu \\ 2\lambda + \mu \end{vmatrix}$$

A proto průměrný čas strávený před dostáním se do koncového stavu, začínáme-li stavem 1, je

$$\frac{3\lambda + \mu}{2\lambda^2}$$

a čas strávený ve stavu 2, než se systém dostane do koncového stavu začínajícího ze stavu 1, je

$$\mu_{12} = \frac{2\lambda}{2\lambda^2} = \frac{1}{\lambda}$$

9.4. Generování náhodných čísel

Náhodná čísla jsou požadována ve všech simulacích nebo ve studiích Monte Carlo. Přestože slova **simulace** a **Monte Carlo** jsou velmi často zaměnitelně používána, někteří výzkumníci mezi nimi rozeznávají rozdíly. Slovo simulace je používáno lidmi, kteří mluví o experimentálním modelu měnícím se v čase. Ve shodě s ostatními je termín Monte Carlo omezený na uměle vytvořené výběrové procedury, které využívají techniky redukující rozptyl. Ačkoliv je zde rozdíl mezi použitím těchto dvou termínů, náhodná čísla jsou v obou těchto metodách stále používána a je požadován mechanismus generování náhodných čísel. Toto bylo v úmyslu prezentovat v této části.

Skutečná sekvence náhodných čísel je ta, která se **nikdy** neopakuje a není zde žádný speciální předpoklad v jejím znovu užití. Metoda generování těchto čísel nebo proces mající skutečnou náhodnost vnitřních elementů by byl skutečný jev (úkaz). Techniky, které jsou používány na počítači, používají deterministické algoritmy, a proto je tzv. pseudostochastické (pseudo = nepravý, stochastický = náhodný) generování to nejlepší generování, které jsme schopni získat. Ke zdůraznění tohoto aspektu jsou náhodná čísla, která jsou generována na počítači, nazývána jako pseudonáhodná a jsou charakterizována délkou sekvence, po které je daná sekvence opakována.

Hodnota náhodné veličiny je závislá na výsledku náhodné události. Náhodná veličina X je dána vztahem

$$P\{X < x\} = F(x),$$

kde $x \in R$ a $F(x)$ je distribuční funkce. Hodnoty náhodné veličiny jsou známy jako **pseudonáhodná čísla**.

Dobrý generátor pseudonáhodných čísel by měl být schopen generovat náhodná čísla, která splňují následující vlastnosti:

1. čísla by měla být stejnoměrně rozložená,
2. statisticky nezávislá,

3. reprodukovatelná,
4. neopakující se do požadované délky,
5. vysoko-rychlostní generace,

Zároveň by generátor měl mít co nejmenší požadavky na paměť.

Jestliže náhodná veličina má stejnoměrné rozložení, pak také generovaná náhodná čísla jsou nazývána jako **stejnoměrně rozložená náhodná čísla**.

Předpokládejme, že čísla x_1, x_2, \dots, x_n jsou hodnoty náhodné veličiny X v nezávislém procesu. Pak se sekvence náhodných čísel $\{x_n\}$ nazývá **náhodná sekvence** a má stejnoměrné rozložení na jednotkovém intervalu $0 \leq x \leq 1$ za podmínky, že relativní frekvence sekvence $\{x_n\}$ na $\langle 0,1 \rangle$ má následující hodnotu

$$\lim_{n \rightarrow \infty} \frac{v_n(a, b)}{n} = b - a,$$

kde $v_n = v_n(a, b)$ je počet elementů v konečné subsekvenci x_1, x_2, \dots, x_n patřící do intervalu (a, b) vzaté z $\langle 0,1 \rangle$. To také znamená, že hodnota pro interval $\langle 0,1 \rangle$ ve výše uvedeném vzorci může být 1. Necht' náhodný vzorek ze standardního stejnoměrného rozdělení je vyjádřen pomocí $U(0,1)$. Teď zvážíme generování nezávislých náhodných proměnných podle $U(0,1)$.

Existuje několik metod, které generují náhodná čísla. Některé z nich jsou:

1. Inner Product Method (Von Neumann),
2. Lehmerova metoda,
3. Fibonacciho série metod,
4. Kongruentní metody.

9.5. Kongruentní metoda

Tato metoda je naprosto reprodukovatelná a požaduje minimum počítačové paměti. Dvěma celým číslům a a b říkáme **kongruentní modulo m** , jestliže jejich rozdíl je celé číslo a je násobkem m . Symbolicky můžeme psát $a = b \pmod{m}$. Velmi užitečný zdroj generace pseudonáhodných čísel je lineární kongruentní sekvence typu

$$x_{i+1} = \alpha x_i + \beta \pmod{m} \text{ pro } i = 1, 2, 3, \dots,$$

kde x_i, α, β jsou nezáporná celá čísla.

Pak

$$\begin{aligned} x_1 &= \alpha x_0 + \beta \pmod{m} \\ x_2 &= \alpha x_1 + \beta \pmod{m} = \alpha^2 x_0 + (\alpha + 1)\beta \pmod{m}, \end{aligned}$$

podobně

$$x_3 = \alpha x_2 + \beta \pmod{m} = \alpha^3 x_0 + (\alpha^2 + \alpha + 1)\beta \pmod{m}.$$

Obecně

$$x_i = \alpha^i x_0 + \frac{(\alpha^i - 1)}{(\alpha - 1)} \beta \pmod{m}.$$

Tedy známe x_0 (které je také nazýváno jako **počátek** náhodné sekvence čísel), α , β a m , můžeme tedy vyčíslit všechna čísla v sekvenci x_1, x_2, \dots, x_n . Musí být zaručeno, že $0 \leq x_i \leq m$ pro všechna i . Dokud jsou čísla produkována pomocí výše uvedeného vzorce, pak by měla ležet v intervalu $\langle 0, m \rangle$, k získání náhodných čísel mezi 0 a 1 musí být x_i děleno m , takže $x_i \rightarrow \frac{x_i}{m}$.

Pro daný generátor a dané x_0 (počátek) se délka nejkratší subsekvence, po které je sekvence opakována, nazývá **perioda počátku** pro daný generátor, zatímco největší perioda jakéhokoliv počátku x_0 se nazývá **perioda generátoru**.

Z tohoto důvodu je žádoucí volit α, β, m a x_0 tak, abychom maximalizovali periodu generovaných sekvencí. Pro $\beta = 0$ se tato metoda nazývá **Multiplikativní-kongruentní** metoda, jinak je známa jako **smíšená-kongruentní** metoda.

Pro jakékoliv programové simulace jsou z náhodných čísel obecně požadována velmi velká čísla. A proto je důležité mít velmi rychlé procedury generující náhodná čísla na počítačích. Toho může být dosaženo pouze tehdy, jestliže počítačový kód je napsán přímo ve strojovém jazyce. Avšak v roce 1979 Schrage ukázal, že kód pro pseudonáhodná čísla může být zapsán také v jazycích vyšší úrovně, které produkují stejné výsledky na jakémkoli počítači. Schrage použil multiplikativní kongruenční metodu ke generování náhodných čísel, která nebyla uspokojivě nalezena v extrémních rozložení.

Wichmann a Hill [Wichmann B.A. a I.D. Hill An Efficient and Portable Pseudo-random Number Generator in Applied Statistics Algorithms, pp. 238 – 142, vydáno Ellis Horwood Limited, Chichester, 1985] poskytuje účinné a statisticky spolehlivé multiplikativní kongruentní algoritmy generující pseudonáhodná čísla. Mají délku periody větší než $6,95 \cdot 10^{12}$, takže, i když budeme používat 1000 náhodných čísel za sekundu, sekvence se nebude opakovat dřív než za 200 let. Ve skutečnosti Wichmann a Hill používají tři jednoduché multiplikativní kongruentní generátory. Každý používá prvočísla pro svůj modul a základní kořen pro svůj násobitel, který garantuje kompletní cyklus. Poté jsou tyto tři výsledky sečteny a zanedbatelná část je odečtena. Před začátkem procedury jsou náhodně vybírána 3 celá čísla mezi (1, 30000), která jsou požadována k dodání do počítače. Wichmann a Hill rovněž poskytli kód zapsaný v jazyce FORTRAN 77, který je rovněž dostupný v uvedené literatuře.

Také můžeme generovat pseudonáhodná čísla podléhající jiným rozdělením - normálnímu, exponenciálnímu, gamma, atd.



Shrnutí kapitoly 9.

Nechť

$$\{X_t : t \geq 0\}$$

reprezentuje spojitý náhodný parametr Markovova řetězce s m diskrétními stavy. Pro $t > s > 0$ a stavy i a j nechť

$$P\{X_t = j\} = p_j(t),$$

což je pravděpodobnost, že proces je ve stavu j v čase t , a

$$P\{X_t = j | X_s = i\} = p_{ij}(s, t),$$

což je pravděpodobnost, že proces je ve stavu j v čase t a byl dán procesem, který byl ve stavu i v čase s .

Pravděpodobnost $p_{ij}(s, t)$ je nazývána **přechodovou pravděpodobnostní funkcí Markovova řetězce**.

Předpokládáme, že $p'_{ij}(t)$ existuje a $h \rightarrow 0$, pak:

$$p'_{ij}(t) = \lim_{h \rightarrow 0} \frac{p_{ij}(t+h) - p_{ij}(t)}{h} = -d_{ji} p_{ij}(t) + \sum_{k \neq j} p_{ik}(t) d_{kj}.$$

Pro všechny stavy i a j tato rovnice dává systém diferenciálních rovnic, jejichž řešením získáme pravděpodobnosti přechodu.

K získání bezpodmínečných stavových pravděpodobností $p'_i(t)$ pro každý stav i přijmeme stejný důvod, který byl již použit v odvození předcházejících rovnic. Takže máme

$$p_i(t+h) = p_i(t)(1 - d_{ii}h) + \sum_{j \neq i} p_j(t)d_{ji}h + o(h),$$

ze které získáme

$$p'_i(t) = -d_{ii}p_i(t) + \sum_{j \neq i} p_j(t)d_{ji}.$$

Tento výraz definuje systém rovnic, který je lineární z hlediska Laplaceovy transformace proměnných, a může být řešen standardními technikami.

Skutečná sekvence náhodných čísel je ta, která se **nikdy** neopakuje a není zde žádný speciální předpoklad v jejím znovu užití. Metoda generování těchto čísel nebo proces mající skutečnou náhodnost vnitřních elementů by byl skutečný jev (úkaz). Techniky, které jsou používány na počítači, používají deterministické algoritmy, a proto je tzv. pseudostochastické (pseudo = nepravý, stochastický = náhodný) generování to nejlepší generování, které jsme schopni získat.

Kongruentní metoda je naprosto reprodukovatelná a požaduje minimum počítačové paměti. Dvěma celým číslům a a b říkáme **kongruentní modulo m** , jestliže jejich rozdíl je celé číslo a je násobkem m .

Symbolicky můžeme psát $a = b \pmod{m}$. Velmi užitečný zdroj generace pseudonáhodných čísel je lineární kongruentní sekvence typu

$$x_{i+1} = \alpha x_i + \beta \pmod{m} \text{ pro } i = 1, 2, 3, \dots,$$

kde x_i, α, β jsou nezáporná celá čísla.



Otázky 9.

1. Co je to přechodová pravděpodobnostní funkce Markovova řetězce?
2. Jaké vlastnosti má matice hustoty přechodu?
3. Popište kongruentní metodu generování náhodných čísel.

10. DVOUFAKTOROVÁ ANALÝZA ROZPTYLU



Čas ke studiu: 1 hodina



Cíl: Po prostudování tohoto odstavce budete umět:

- Seznámíte se s dalšími možnostmi analýzy rozptylu, tj. s možnostmi ověření vlivů dvou faktorů (neinteragujících, interagujících) na sledovaný jev.



Výklad

10.1. Úvodní poznámky

Zmiňme úvodem alespoň zhruba základní myšlenku analýzy rozptylu. Máme-li k dispozici nějakou skupinu výsledků, kterou můžeme roztrždit podle několika různých hledisek, pak pomocí analýzy rozptylu rozdělíme také celkovou variabilitu mezi pozorováními na složky odpovídající jednotlivým hlediskům třídění. Testováním významnosti těchto složek dále určíme, která hlediska třídění se významně podílejí na celkové variabilitě mezi daty. Přitom pro data předpokládáme určitý model s danými parametry a vlastnostmi.

Třídění dat pak lze provádět podle jednoho či více hledisek (**faktorů**). Naším úkolem zde bude podrobněji prozkoumat problematiku analýzy rozptylu při třídění dle dvou faktorů. V rámci dvoufaktorové analýzy rozptylu si přiblížíme následující pojmy:

- **podtřídy** – jednotlivé kombinace úrovní obou faktorů,
- **pokusy bez opakování / s opakováním** – pro každou podtřídu bylo provedeno jediné / více pozorování,
- u pokusů s opakováním lze definovat **pokusy se stejným / různým počtem pozorování v každé podtřídě**,
- pokud se vlivy obou faktorů neskládají aditivně, říkáme, že existuje **interakce mezi faktory**.

10.2. Třídění dle dvou faktorů bez opakování

V této kapitole budeme předpokládat, že **interakce** mezi oběma faktory **neexistuje**.

Uvažujme případ, kdy první faktor je sledován na p úrovních a druhý faktor na q úrovních. Označíme-li jednotlivá pozorování symbolem x_{ij} , kde $i = 1, 2, \dots, p$ a $j = 1, 2, \dots, q$, pak $N = pq$ značí celkový počet pozorování. Pro každé pozorování dále předpokládejme, že

$$x_{ij} = \mu + \xi_i + \eta_j + \varepsilon_{ij},$$

kde μ je celková střední hodnota, ξ_i vliv prvního faktoru na i -té úrovni, η_j vliv druhého faktoru na j -té úrovni a ε_{ij} je náhodná chyba. Pro očekávanou hodnotu každého pozorování tedy platí vztah

$$E [x_{ij}] = \mu + \xi_i + \eta_j$$

a budeme uvažovat situaci, kdy jednotlivá pozorování jsou nekorelována a **normálně rozdělena** kolem středních hodnot daných výše uvedeným vzorcem se **společným rozptylem** σ^2 . Veličiny ξ_i a η_j tedy odlišují střední hodnoty v jednotlivých podtřídách a budeme dále předpokládat, že

$$\sum_{i=1}^p \xi_i = 0 \quad \text{a} \quad \sum_{j=1}^q \eta_j = 0.$$

Označme symbolem S celkový součet čtverců odchylek všech pozorování od celkového průměru

$$\bar{x} \stackrel{\text{def}}{=} \frac{1}{N} \sum_{i,j} x_{ij}, \text{ tj.}$$

$$S = \sum_{i,j}^{\text{ozn}} (x_{ij} - \bar{x})^2.$$

Lze ukázat, že platí následující rozklad

$$S = S_i + S_j + S_r,$$

kde

$$S_i = q \sum_{i=1}^p (\bar{x}_{i\cdot} - \bar{x})^2,$$

$$S_j = p \sum_{j=1}^q (\bar{x}_{\cdot j} - \bar{x})^2,$$

a

$$S_r = \sum_{i,j}^{\text{ozn}} (x_{ij} - \bar{x}_{i\cdot} - \bar{x}_{\cdot j} + \bar{x})^2,$$

přičemž $\bar{x}_{i\cdot}$, resp. $\bar{x}_{\cdot j}$ značí řádkový, resp. sloupcový průměr. Označíme-li dále symbolem $T_{i\cdot}$, resp. $T_{\cdot j}$ řádkové, resp. sloupcové součty, tj.

$$T_{i\cdot} = \sum_{j=1}^q x_{ij}, \quad \text{resp.} \quad T_{\cdot j} = \sum_{i=1}^p x_{ij},$$

a symbolem T celkový součet, tj.

$$T = \sum_{i,j}^{\text{ozn}} x_{ij},$$

pak řádkový, resp. sloupcový průměr dostaneme jako

$$\bar{x}_{i\cdot} = \frac{T_{i\cdot}}{q}, \text{ resp. } \bar{x}_{\cdot j} = \frac{T_{\cdot j}}{p}$$

a celkový průměr jako

$$\bar{x} = \frac{T}{N}.$$

Díky výše uvedeným vztahům pak jednoduše odvodíme, že

$$S = \sum_{i,j} x_{ij}^2 - 2\bar{x} \sum_{i,j} x_{ij} + \bar{x}^2 \sum_{i,j} 1 = \sum_{i,j} x_{ij}^2 - 2 \frac{T}{N} T + \frac{T^2}{N^2} N = \sum_{i,j} x_{ij}^2 - \frac{T^2}{N},$$

$$\begin{aligned} S_i &= q \sum_i \bar{x}_{i\cdot}^2 - 2q\bar{x} \sum_i \bar{x}_{i\cdot} + q\bar{x}^2 \sum_i 1 = q \sum_i \frac{T_{i\cdot}^2}{q^2} - 2q \frac{T}{N} \sum_i \frac{T_{i\cdot}}{q} + N\bar{x}^2 = \\ &= \frac{1}{q} \sum_i T_{i\cdot}^2 - 2 \frac{T^2}{N} + N \frac{T^2}{N^2} = \frac{1}{q} \sum_i T_{i\cdot}^2 - \frac{T^2}{N}, \end{aligned}$$

$$\begin{aligned} S_j &= p \sum_j \bar{x}_{\cdot j}^2 - 2p\bar{x} \sum_j \bar{x}_{\cdot j} + p\bar{x}^2 \sum_j 1 = p \sum_j \frac{T_{\cdot j}^2}{p^2} - 2p \frac{T}{N} \sum_j \frac{T_{\cdot j}}{p} + N\bar{x}^2 = \frac{1}{p} \sum_j T_{\cdot j}^2 - 2 \frac{T^2}{N} + N \frac{T^2}{N^2} = \\ &= \frac{1}{p} \sum_j T_{\cdot j}^2 - \frac{T^2}{N} \end{aligned}$$

a

$$S_r = \sum_{i,j} x_{ij}^2 - \frac{1}{q} \sum_i T_{i\cdot}^2 - \frac{1}{p} \sum_j T_{\cdot j}^2 + \frac{T^2}{N}.$$

Poslední rovnost není kvůli přehlednosti textu odvozována. Prostřednictvím těchto vzorců se velmi usnadní numerický výpočet požadovaných součtů. Jen doplňme, že jednotlivé součty nazýváme:

- S_i – součet čtverců mezi řádkovými průměry s $(p-1)$ stupni volnosti,
- S_j – součet čtverců mezi sloupcovými průměry s $(q-1)$ stupni volnosti,
- S_r – reziduální součet čtverců.

Označme symbolem r počet stupňů volnosti reziduálního součtu S_r . Protože celkový součet S má $(N-1)$ stupňů volnosti, platí pro r vztah

$$r = N - 1 - (p - 1) - (q - 1) = pq - p - q + 1 = q(p - 1) - (p - 1) = (p - 1)(q - 1).$$

Lze dokázat, že pro podíly $\frac{S_i}{p-1}$ a $\frac{S_j}{q-1}$ platí

$$E\left[\frac{S_i}{p-1}\right] = \sigma^2 + q \frac{\sum \zeta_i^2}{p-1}$$

a

$$E\left[\frac{S_j}{q-1}\right] = \sigma^2 + p \frac{\sum \eta_j^2}{q-1}.$$

Odtud, platí-li hypotézy

$$\zeta_i = 0 \text{ a } \eta_j = 0 \text{ pro } i = 1, \dots, p \text{ a } j = 1, \dots, q,$$

jsou tyto podíly nestrannými odhady rozptylu σ^2 . Poměr každého z těchto podílů a podílu $\frac{S_r}{(p-1)(q-1)}$ (tzv. reziduálního odhadu rozptylu) je za situace správnosti obou výše uvedených hypotéz hodnotou veličiny s Fisherovým-Snedecorovým rozdělením F o $[p-1, (q-1)(p-1)]$ a $[q-1, (q-1)(p-1)]$ stupňích volnosti. Testovací kritérium F se pak použije k testování obou hypotéz. Nyní si výše uvedené poznatky uspořádáme do tabulky.

Zdroj mělnivosti	Součet čtverců	Stupně volnosti	Podíl	Test. kritérium F
mezi řádky	S_i	$p-1$	$M S_i = \frac{S_i}{p-1}$	$F = \frac{M S_i}{M S_r}$
mezi sloupci	S_j	$q-1$	$M S_j = \frac{S_j}{q-1}$	$F = \frac{M S_j}{M S_r}$
uvnitř podtříd	S_r	$(p-1)(q-1)$	$M S_r = \frac{S_r}{(p-1)(q-1)}$	
celkem	S	$N-1$		

10.3. Třídění dle dvou faktorů s opakováním

Na rozdíl od předešlé podkapitoly předpokládejme, že **interakce** mezi oběma faktory **existuje**.

Budeme uvažovat situaci se **stejným počtem pozorování v každé podtřídě**. Označíme-li písmenem n konstantní počet pozorování v každé podtřídě, pak máme k dispozici celkový počet $N = npq$ pozorování. Jednotlivá pozorování tedy budeme značit symbolem x_{ijk} , kde $k = 1, \dots, n$. Z předpokladu interakce řádkových a sloupcových vlivů pak plyne rovnost

$$x_{ijk} = \mu + \xi_i + \mu_j + \lambda_{ij} + \varepsilon_{ijk},$$

kde vlivy λ_{ij} představují systematickou odchylku každého pozorování $x_{ij\cdot}$ od aditivního modelu střední hodnoty $\mu + \xi_i + \eta_j$. Opět předpokládejme nekorelovanost pozorování a jejich normální

rozdělení kolem středních hodnot s identickým rozptylem σ^2 . Podobně jako v minulé podkapitole budeme dále předpokládat, že platí

$$\sum_i \xi_i = \sum_j \eta_j = \sum_i \lambda_{ij} = \sum_j \lambda_{ij} = 0.$$

Rozklad celkového součtu S čtverců odchylek jednotlivých pozorování od celkového průměru, tj.

$$S = \sum_{i,j,k}^{ozn} (x_{ijk} - \bar{x})^2, \text{ má nyní tvar}$$

$$S = S_i + S_j + S_{ij} + S_r,$$

kde pro jednotlivé sčítance platí vztahy

$$S_i = nq \sum_{i=1}^p (\bar{x}_{i\bullet} - \bar{x})^2,$$

$$S_j = np \sum_{j=1}^q (\bar{x}_{\bullet j} - \bar{x})^2,$$

$$S_{ij} = n \sum_{ij}^{ozn} (\bar{x}_{ij} - \bar{x}_{i\bullet} - \bar{x}_{\bullet j} + \bar{x})^2$$

a

$$S_r = np \sum_{i,j,k}^{ozn} (\bar{x}_{ijk} - \bar{x}_{ij})^2,$$

přičemž značí-li T_{ij} součet pozorování v příslušné podtřídě, tj.

$$T_{ij} = \sum_{k=1}^n x_{ijk},$$

je \bar{x}_{ij} , tj. průměr n pozorování v této podtřídě, dán jako

$$\bar{x}_{ij} = \frac{T_{ij}}{n}.$$

Další vztahy jsou podobné těm z minulé podkapitoly. Označíme-li

$$T_{i\bullet} = \sum_{jk} x_{ijk}, \quad T_{\bullet j} = \sum_{ik} x_{ijk} \quad a \quad T = \sum_{ijk} x_{ijk},$$

pak platí

$$\bar{x}_{i\bullet} = \frac{T_{i\bullet}}{nq}, \quad \bar{x}_{\bullet j} = \frac{T_{\bullet j}}{np} \quad a \quad \bar{x} = \frac{T}{npq}.$$

Analogicky lze odvodit vzorce pro ulehčení numerických výpočtů ve tvaru

$$S = \sum_{i,j,k} x_{ijk}^2 - \frac{T^2}{N},$$

$$S_i = \frac{1}{nq} \sum_i T_{i\cdot}^2 - \frac{T^2}{N},$$

$$S_j = \frac{1}{np} \sum_j T_{\cdot j}^2 - \frac{T^2}{N},$$

$$S_{ij} = \frac{1}{n} \sum_{i,j} T_{ij}^2 - \frac{1}{nq} \sum_i T_{i\cdot}^2 - \frac{1}{np} \sum_j T_{\cdot j}^2 + \frac{T^2}{N}$$

a

$$S_r = \sum_{i,j,k} x_{ijk}^2 - \frac{1}{N} \sum_{i,j} T_{ij}^2.$$

Nyní opět vypočítáme počet stupňů volnosti r reziduálního součtu S_r . Protože S_i má $(p-1)$ stupňů volnosti, S_j má $(q-1)$ stupňů volnosti, S_{ij} má $pq - p - q - 1 = (p-1)(q-1)$ stupňů volnosti a S má $(N-1)$ stupňů volnosti, pak pro r platí

$$r = N - 1 - (p-1) - (q-1) - (p-1)(q-1) = N - p - p(q-1) = N - pq.$$

Protože

$$E\left[\frac{S_i}{p-1}\right] = \sigma^2 + nq \frac{\sum_i \xi_i^2}{p-1},$$

$$E\left[\frac{S_j}{q-1}\right] = \sigma^2 + np \frac{\sum_j \eta_j^2}{p-1}$$

a

$$E\left[\frac{S_{ij}}{(p-1)(q-1)}\right] = \sigma^2 + n \frac{\sum_{ij} \lambda_{ij}^2}{(p-1)(q-1)},$$

pak, platí-li hypotézy

$$\xi_i = 0, \eta_j = 0 \text{ a } \lambda_{ij} = 0 \text{ pro } i = 1, \dots, p \text{ a } j = 1, \dots, q,$$

jsou tyto podíly $\frac{S_i}{p-1}$, $\frac{S_j}{q-1}$ a $\frac{S_{ij}}{(p-1)(q-1)}$ nestrannými odhady rozptylu σ^2 . Poměr každého

z těchto podílů a podílu $\frac{S_r}{N-pq}$ je za situace správnosti tří výše uvedených hypotéz hodnotou

veličiny s Fisherovým-Snedecorovým rozdělením F o $[p-1, N-pq]$, $[q-1, N-pq]$ a

$[(q-1)(p-1), N-pq]$ stupních volnosti. Testovací kritérium F se pak znovu použije k testování všech tří hypotéz. Opět si můžeme vytvořit tabulku analýzy rozptylu.

Zdroj měnlivosti	Součet čtverců	Stupně volnosti	Podíl	Test. kritérium F
mezi řádky	S_i	$p-1$	$M S_i = \frac{S_i}{p-1}$	$F = \frac{M S_i}{M S_r}$
mezi sloupci	S_j	$q-1$	$M S_j = \frac{S_j}{q-1}$	$F = \frac{M S_j}{M S_r}$
interakce	S_{ij}	$(p-1)(q-1)$	$M S_{ij} = \frac{S_{ij}}{(p-1)(q-1)}$	$F = \frac{M S_{ij}}{M S_r}$
uvnitř podtříd	S_r	$N-pq$	$M S_r = \frac{S_r}{N-pq}$	
celkem	S	$N-1$		



Řešený příklad

V tabulce jsou uvedeny výsledky pokusu, při kterém byla stanovena koncentrace železa ve standardním roztoku železa obsahujícím 2,95% železa deseti různými studenty, přičemž každý student provedl vždy jedno stanovení pomocí pěti různých metod.

Student	Metoda				
	A	B	C	D	E
1	2,963	2,996	2,979	2,970	2,979
2	2,958	2,964	2,955	2,932	2,941
3	2,956	2,945	2,963	2,950	2,975
4	2,948	2,960	2,953	2,944	2,950
5	2,953	2,961	2,961	2,953	2,949
6	2,941	2,940	2,931	2,942	2,930
7	2,963	2,928	2,925	2,940	2,934
8	2,987	2,989	2,988	2,983	2,974
9	2,946	2,950	2,955	2,969	2,954
10	2,956	2,947	2,947	2,960	2,954

Budeme tedy provádět dvoufaktorovou analýzu rozptylu u pokusů bez opakování, protože každý student stanovil každou metodou hodnotu železa právě jednou. V našem případě vystupují faktory student a metoda. Budeme zkoumat, jak významně se oba faktory podílejí na různorodosti výsledků. Důležitými předpoklady jsou nekorelovanost jednotlivých měření a jejich normální rozdělení s identickým rozptylem. Budeme zde také předpokládat, že oba faktory navzájem neinteragují, tedy řídíme se předpokladem, že nepřesnost stanovení hodnoty železa nesouvisí například s tím, že by student některou z metod dostatečně neovládal nebo ji nerozuměl. Naším úkolem je odpovědět na otázku, zda variabilita mezi jednotlivými stanoveními je způsobena různou přesností některé z metod nebo také různou schopností přesného měření u studentů. Označíme vliv i -tého studenta symbolem ξ_i a vliv j -té metody η_j . Stanovme nejprve nulové hypotézy a alternativy:

$$H_0^\xi : \xi_1 = \xi_2 = \dots = \xi_{10} = 0, \quad H_0^\eta : \eta_1 = \eta_2 = \dots = \eta_{10} = 0,$$

$$H_A^\xi : \bar{H}_0^\xi, \quad H_A^\eta : \bar{H}_0^\eta,$$

tj. nulové hypotézy říkají, že vliv studentů v daných úrovních a vliv metod v daných úrovních je nulový. Provedeme analýzu dvěma způsoby, a to numericky a s použitím statistického softwaru.

Analýza pomocí numerických výpočtů.

Dílčí výsledky nebudeme zaokrouhlovat, neboť by tím docházelo k významnému zkreslení.

1. Položíme $p = 10$, $q = 5$ a vypočteme $N = pq = 50$.
2. Výpočet sumy čtverců jednotlivých pozorování: $\sum_{i=1}^{10} \sum_{j=1}^5 x_{ij}^2 = 436,857449$.
3. Výpočet řádkových součtů $T_{i\bullet} = \sum_{j=1}^5 x_{ij}$, $i = 1, \dots, 10$:

$T_{1\bullet}$	$T_{2\bullet}$	$T_{3\bullet}$	$T_{4\bullet}$	$T_{5\bullet}$	$T_{6\bullet}$	$T_{7\bullet}$	$T_{8\bullet}$	$T_{9\bullet}$	$T_{10\bullet}$
14,887	14,75	14,789	14,755	14,777	14,684	14,69	14,921	14,774	14,764

4. Výpočet sloupcových součtů $T_{\bullet j} = \sum_{i=1}^{10} x_{ij}$, $j = 1, \dots, 5$:

$T_{\bullet 1}$	$T_{\bullet 2}$	$T_{\bullet 3}$	$T_{\bullet 4}$	$T_{\bullet 5}$
29,571	29,58	29,557	29,543	29,54

5. Součet čtverců řádkových součtů přes všechny řádky: $\sum_{i=1}^{10} T_{i\bullet}^2 = 2184,268513$.
6. Součet čtverců sloupcových součtů přes všechny sloupce: $\sum_{j=1}^5 T_{\bullet j}^2 = 4368,437139$.
7. Celkový součet všech pozorování: $T = \sum_{i=1}^{10} T_{i\bullet} = \sum_{j=1}^5 T_{\bullet j} = 147,791$.

8. Součet čtverců S_i :

$$S_i = \frac{1}{5} \sum_{i=1}^{10} T_{i\bullet}^2 - \frac{T^2}{N} = \frac{1}{5} \cdot 2184,268513 - \frac{147,791^2}{50} = 1,01010898 \cdot 10^{-2}.$$

9. Součet čtverců S_j :

$$S_j = \frac{1}{10} \sum_{j=1}^5 T_{\bullet j}^2 - \frac{T^2}{N} = \frac{1}{10} \cdot 4368,437139 - \frac{147,791^2}{50} = 1,2028 \cdot 10^{-4}$$

10. Celkový součet čtverců S :

$$S = \sum_{i=1}^{10} \sum_{j=1}^5 x_{ij}^2 - \frac{T^2}{N} = 436,857449 - \frac{147,791^2}{50} = 1,385538 \cdot 10^{-2}.$$

11. Reziduální součet čtverců S_r :

$$S_r = S - S_i - S_j = 1,385538 \cdot 10^{-2} - 1,01010898 \cdot 10^{-2} - 1,2028 \cdot 10^{-4} = 3,62612 \cdot 10^{-3}$$

Vypočtené součty zapíšeme do tabulky analýzy rozptylu a spočteme zbývající položky v tabulce

Zdroj měnlivosti	Součet čtverců	Stupně volnosti	Podíl	Test. kritérium F	p-value
rozdíly v hodnotách při měření jednotlivých studentů	$1,0101 \cdot 10^{-2}$	9	$1,12322 \cdot 10^{-3}$	11,151	< 0,001
rozdíly v hodnotách při měření pomocí daných metod	$1,2028 \cdot 10^{-4}$	4	$3,007 \cdot 10^{-5}$	0,299	> 0,25
reziduální	$3,62612 \cdot 10^{-3}$	36	$1,00725 \cdot 10^{-4}$		
celkový	$1,385538 \cdot 10^{-2}$	49			

Z jednotlivých hodnot p-value usoudíme, že

- vliv jednotlivých studentů při měření hodnoty železa ve vzorku lze považovat za podstatný, je tedy patrný rozdíl v přesnosti měření studentů,
- avšak nelze tvrdit, že by užití metody podstatně ovlivňovalo výsledek měření.

Jinými slovy, různost výběrových průměrů při užití daných metod lze přičíst náhodnému kolísání, ale různost výběrových průměrů měření jednotlivých studentů je způsobena nestejnými středními hodnotami. Nezamítáme tedy hypotézu $\eta_1 = \eta_2 = \dots = \eta_5 = 0$ a zamítáme hypotézu $\xi_1 = \xi_2 = \dots = \xi_{10} = 0$.

Analýza pomocí statistického softwaru.

Díky dnešnímu statistickému softwaru lze velmi snadno obdržet tabulku analýzy rozptylu bez jakéhokoli námi provedeného numerického výpočtu. Pro naši úlohu jsme použili program JMP IN. Výsledky samozřejmě plně korespondují s předešlými numerickými výpočty. Získali jsme následující tabulku dvoufaktorové analýzy rozptylu (v o něco málo jiném tvaru než jsme si odvodili v textu):

Analysis of Variance				
Source	DF	Sum of Squares	Mean Square	F Ratio
Model	13	0,01022926	0,000787	7,8120
Error	36	0,00362612	0,000101	Prob>F
C Total	49	0,01385538		<,0001

Obdržíme i tabulku efektů obou faktorů:

Effect Test					
Source	Nparm	DF	Sum of Squares	F Ratio	Prob>F
student	9	9	0,01010898	11,1513	<,0001
metoda	4	4	0,00012028	0,2985	0,8770

Z posledního sloupce výše uvedené tabulky vyčteme hodnoty p -value. Je zřejmé, že učiníme naprosto shodný závěr jako při numerickém výpočtu. Nezamítneme hypotézu $\eta_1 = \eta_2 = \dots = \eta_5 = 0$ a zamítneme hypotézu $\xi_1 = \xi_2 = \dots = \xi_{10} = 0$.

10.4. Závěrečné poznámky

Uveďme zde ještě pár důležitých či zajímavých bodů týkajících se analýzy rozptylu:

- Při složitější analýze s výskytem více faktorů nám významně klesá reziduální rozptyl a tím roste síla testu.
- Při třídění dle dvou faktorů bez opakování nutně musíme předpokládat model bez interakcí. V tomto případě totiž neznáme odhad rozptylu σ^2 vznikajícího díky variabilitě mezi opakovanými výsledky uvnitř podtříd. Jestliže tedy při situaci bez opakování existuje interakce mezi oběma faktory, je její příspěvek plně zahrnut v reziduu.
- Případné odchylky od normality rozdělení dat, které mohou nastat ve většině experimentů, jen velmi slabě ovlivňují F-testy užívané při analýze rozptylu.
- Výsledky dosažené analýzou rozptylu závisí na přesnosti předpokladů týkajících se vlastností zavedeného modelu. Neopomenutelnými předpoklady jsou pak nekorelovanost a konstantní rozptyl všech pozorování.



Otázky 10.

1. V čem spočívá výhoda dvoufaktorové analýzy rozptylu oproti analýze jednofaktorové?
2. Co to znamená, mluvíme-li o pokusech s opakováním (bez opakování)?
3. Jaké předpoklady musí být splněny, chceme-li použít dvoufaktorovou analýzu rozptylu?



Úlohy k řešení 10.

1. Analytická laboratoř disponuje dvěma měřicími soustavami. V laboratoři pracují celkem 4 laborantky. Jeden vzorek o koncentraci léčiva 140 $\mu\text{l/ml}$ byl připraven všemi laborantkami a změřen na obou soustavách. Je třeba rozhodnout, zda kvalitu stanovení ovlivňují laborantky či přístroje.

Data: změřené obsahy léčiva [μg]:

	Soustava I	Soustava II
1. laborantka	140,05	140,28
2. laborantka	140,36	140,06
3. laborantka	139,95	140,10
4. laborantka	140,09	140,15

2. Ve firmě Life a.s. Hradec Králové proběhla v roce 1999 klinická studie, jejímž cílem bylo posoudit bioekvivalenci dvou uroxylových přípravků – UROXAN a UROLON. Studie probíhala na třech centrech – Hradec Králové, Brno a Praha. V každém z center bylo do studie zařazeno 16 dobrovolníků, kterým byly jednorázově podávány oba přípravky s wash-out periodou jeden týden. V každém centru vždy 6 dobrovolníků vzalo jako první přípravek UROXAN a druhý UROLON a 6 dobrovolníků vzalo tyto přípravky v opačném pořadí. Po každém podání přípravků byly dobrovolníkům mimo jiné měřeny hladiny hemoglobinu. Pomocí analýzy rozptylu určete, zda jsou hodnoty hemoglobinu v krvi po podání prvního přípravku ovlivněny podávaným přípravkem či centrem.

Data: Hladina hemoglobinu [g/l] po podání 1. přípravku:

Pořadí přípravku	Číslo dobrovolníka	Hradec Králové	Brno	Praha
UROXAN - UROLON	1	138	135	137
	2	126	174	123
	3	141	157	124
	4	151	136	152
	5	163	137	138
	6	139	140	144
	7	146	136	148
	8	144	144	141
UROLON - UROXAN	9	126	143	151
	10	132	142	142
	11	163	125	168
	12	145	155	167
	13	142	149	143
	14	159	153	139
	15	130	137	147
	16	139	133	131

**KLÍČ K ŘEŠENÍ**

1. Ani jeden z faktorů – laborantky ($p\text{-value}=0,824$) či soustavy ($p\text{-value}=0,830$) nemají vliv na kvalitu měření. Rovněž efekt interakce je nevýznamný.
2. Dvoufaktorová analýza rozptylu s opakováním prokázala, že druh podaného uroxylového přípravku (tzn. zda byl dobrovolníkovi podán UROXAN či UROLON) ($p\text{-value}=0,579$) ani jednotlivá centra ($p\text{-value}=0,981$) statisticky neovlivnila hladinu hemoglobinu v krvi dobrovolníků.

11. LOGISTICKÁ REGRESE A JEJÍ UŽITÍ PRO DISKRIMINACI



Čas ke studiu: 1,5 hodiny



Cíl: Po prostudování tohoto odstavce budete umět:

- V této kapitole se seznámíte s metodikou logistické regrese a s jejím užitím pro diskriminační analýzu.



Výklad

11.1. Úvod

V praxi jsme často postaveni před problém zařadit jisté objekty do předem vymezených skupin. K tomuto účelu máme k dispozici naměřené určité znaky na těchto objektech a naším úkolem je na základě znalosti hodnot těchto znaků zařadit předložený objekt do některé skupiny. K řešení tohoto problému lze přistupovat několika způsoby. Budeme předpokládat, že každý objekt patří do jedné ze dvou skupin (označme je 0 a 1). Problém diskriminace budeme řešit pomocí modelů logistické regrese (LR).

K sestavení rozhodovacího pravidla máme k dispozici obvykle n testovacích objektů, na kterých máme naměřeny příslušné znaky a o kterých buď víme anebo nevíme, do které skupiny patří (v závislosti na zvoleném modelu). Naměřené znaky nechť jsou reprezentovány p -rozměrnými náhodnými vektory $\underline{X}_1, \dots, \underline{X}_n$ a příslušnost i -tého objektu k dané skupině nechť je vyjádřena hodnotou náhodné veličiny Y_i , která nabývá hodnot 0 nebo 1, podle toho, do které skupiny daný objekt náleží. U nového objektu, který chceme zařadit na základě vytvořeného rozhodovacího pravidla, nechť jsou naměřené znaky reprezentovány p -rozměrným náhodným vektorem \underline{X} a rozhodnutí hodnotou náhodné veličiny Y .

□ Statistické rozhodovací funkce

Jednotlivé diskriminační procedury budou odvozeny na základě teorie statistických rozhodovacích funkcí, kterou na tomto místě stručně připomeneme.

K nalezení optimálního rozhodovacího pravidla bude využito bayesovského přístupu. Roli neznámého parametru, o jehož hodnotě chceme rozhodnout, bude hrát náhodná veličina $Y \in \{0;1\}$, která má pravděpodobnostní funkci $q(y)$. Rozhodnutí bude prováděno na základě hodnoty náhodného vektoru $\underline{X} \in R^p$, jenž má hustotu $r(x)$. Podmíněnou hustotu \underline{X} za podmínky $Y=y$ označíme $r(x|y)$. Nechť $\delta : R^p \rightarrow \{0;1\}$ je **rozhodovací funkce** a H je množina všech funkcí $\delta : R^p \rightarrow \{0;1\}$. **Ztrátovou funkci** zavedeme jako

$$L(Y, \delta(\underline{X})) = \begin{cases} 0, & \text{pokud } Y = \delta(\underline{X}) \\ 1, & \text{v opačném případě} \end{cases}$$

Riziková funkce je definována jako

$$R(Y, \delta) = E[L(Y, \delta(\underline{X}))|Y] = \int_{R^p} L(Y, \delta(\underline{x})) \cdot r(\underline{x}|y) d\underline{x}$$

Bayesovské riziko se určí jako

$$\rho(\delta) = ER(Y, \delta) = \sum_{y=0}^1 R(y, \delta) \cdot q(y)$$

Optimální rozhodovací funkci δ^* potom dostaneme jako

$$\delta^* = \arg \min_{\delta \in H} \rho(\delta)$$

Označme podmíněnou pravděpodobnostní funkci Y za podmínky $\underline{X}=\underline{x}$ jako $p(y|\underline{x})$. Necht' $p(1|\underline{x}) = P(Y = 1|\underline{X} = \underline{x}) = \pi(\underline{x})$ a $p(0|\underline{x}) = P(Y = 0|\underline{X} = \underline{x}) = 1 - \pi(\underline{x})$. Existuje-li pro všechna $\underline{x} \in R^p$

$$\hat{\delta}(\underline{x}) = \arg \min_{\delta \in H} E[L(Y, \delta(\underline{X}))|\underline{X} = \underline{x}] = \sum_{y=0}^1 L(Y, \delta(\underline{x})) \cdot p(y|\underline{x}),$$

lze snadno s pomocí Bayesovy věty ukázat, že $\hat{\delta} = \delta^*$. Přímým výpočtem lze dále nalézt vyjádření rizikové funkce a bayesovského rizika:

$$R(0, \delta) = P(\delta(\underline{X}) = 1|Y = 0),$$

$$R(1, \delta) = P(\delta(\underline{X}) = 0|Y = 1),$$

$$\rho(\delta) = P(\delta(\underline{X}) = 1, Y = 0) + P(\delta(\underline{X}) = 0, Y = 1).$$

Vidíme tedy, že bayesovské riziko lze v této situaci interpretovat jako pravděpodobnost špatného rozhodnutí o hodnotě Y , tj. o zařazení daného objektu do skupiny. Dále budeme bayesovské riziko nazývat jako **pravděpodobnost chyby**.

Logistická regrese

U lineárního modelu, kterým jsme se doposud zabývali, byla vysvětlovaná proměnná spojitá. Nyní se pokusíme vysvětlit chování 0-1 veličiny, která modeluje nevýskyt či výskyt sledovaného jevu. Stejně jako u lineárního modelu budeme vyjadřovat střední hodnotu vysvětlované proměnné jako funkci nezávisle proměnných. Tentokrát bude tato střední hodnota rovna pravděpodobnosti jedničky, tedy pravděpodobnosti výskytu sledovaného jevu.

11.2. Tvar závislosti

Uvažujme nezávislé náhodné veličiny Y_1, \dots, Y_n s alternativními rozděleními s parametry μ_i . Střední hodnoty jsou totožné s pravděpodobnostmi μ_i , ty mohou záviset na nějakých

nenáhodných doprovodných veličinách $\underline{x}_{i\bullet}$. Je zřejmé, že platí $DY_i = \mu_i \cdot (1 - \mu_i)$. To je první podstatný rozdíl v porovnání s normálním lineárním modelem, kde byl rozptyl konstantní.

Pokud bychom předpokládali, jako v lineárním modelu, závislost tvaru

$$\mu_i = EY_i = \beta' \underline{x}_{i\bullet},$$

bude problém s interpretací, protože nelze zaručit, že pro libovolné $\underline{x}_{i\bullet}$ bude μ_i ležet v intervalu (0;1). Hledejme tedy jiný interpretovatelný tvar závislosti a motivaci hledejme v odhadech maximální věrohodnosti.

Pravděpodobnosti dvou možných hodnot $Y_i=1$ a $Y_i=0$ lze souhrnně psát jako

$$P(Y_i = j) = \mu_i^j \cdot (1 - \mu_i)^{1-j}, \quad j = 0,1.$$

Logaritmickou věrohodnostní funkci lze tedy zapsat

$$\begin{aligned} \ell(\underline{\mu}) &= \ln \prod_{i=1}^n \mu_i^{Y_i} (1 - \mu_i)^{1-Y_i} \\ &= \sum_{i=1}^n (Y_i \cdot \ln \mu_i + (1 - Y_i) \cdot \ln (1 - \mu_i)) \\ &= \sum_{i=1}^n Y_i \cdot \ln \frac{\mu_i}{1 - \mu_i} + \sum_{i=1}^n \ln (1 - \mu_i) \end{aligned}$$

Jak je vidět, pozorované náhodné veličiny se v logaritmické věrohodnostní funkci projevují pouze v součinech s výrazy $\log(\mu_i/(1 - \mu_i))$. Podíl

$$\omega(\underline{x}_{i\bullet}) = \frac{\mu_i}{1 - \mu_i} = \frac{P_{\underline{x}_{i\bullet}}(Y_i = 1)}{P_{\underline{x}_{i\bullet}}(Y_i = 0)}$$

má bezprostřední interpretaci. Porovnává pravděpodobnost jedničky (výskyt sledovaného jevu) a nuly (nevýskyt jevu). Pro tento podíl se v angličtině používá výraz **odds**. Tomu odpovídá český termín **šance**. Samotné funkci $\eta = \ln(\mu/(1 - \mu))$ se říká **logit**.

Předpokládejme speciálně, že logit pravděpodobnosti je lineární funkcí neznámých parametrů

$$\eta_i(\beta) = \beta' \cdot \underline{x}_{i\bullet}.$$

Někdy se i v tomto obecném zápisu systematicky uvádí absolutní člen, protože, jak uvidíme, ne vždy jej budeme schopni odhadnout. Pak se místo regresní matice \mathbf{X} uvádí matice $(\mathbf{1}, \mathbf{X})$.

Střední hodnotu pak v našem modelu můžeme vyjádřit ve tvaru

$$\mu_i(\beta) = \frac{e^{\eta_i(\beta)}}{1 + e^{\eta_i(\beta)}} = \frac{e^{\beta' \cdot \underline{x}_{i\bullet}}}{1 + e^{\beta' \cdot \underline{x}_{i\bullet}}} = \frac{1}{1 + e^{-\beta' \cdot \underline{x}_{i\bullet}}},$$

což zaručí, že platí $0 < \mu_i < 1$ a odstraní tak jeden z naznačených problémů.

11.3. Odhad parametrů

Naznačme ještě odhad parametrů metodou maximální věrohodnosti. Protože platí

$$\frac{\partial}{\partial \eta_i} \ln(1 - \mu_i) = -\frac{\partial}{\partial \eta_i} \ln(1 + e^{\eta_i}) = -\frac{e^{\eta_i}}{1 + e^{\eta_i}} = -\mu_i$$

a logaritmickou věrohodnostní funkci jsme upravili na tvar

$$\ell(\beta_0, \beta) = \sum_{i=1}^n Y_i \eta_i(\beta_0, \beta) + \sum_{i=1}^n \ln(1 - \mu_i(\beta_0, \beta)),$$

jsou parciální derivace logaritmické věrohodnostní funkce rovny

$$\frac{\partial \ell}{\partial \beta} = \sum_{i=1}^n \frac{\partial \ell}{\partial \eta_i} \frac{\partial \eta_i}{\partial \beta} = \sum_{i=1}^n (Y_i - \mu_i(\beta_0, \beta)) \cdot \underline{x}_{i\bullet}. \quad (1)$$

Po malé úpravě zjistíme, že soustava normálních rovnic (nelineární v β) lze psát

$$\underline{X}'(\underline{Y} - \underline{\mu}(\beta)) = \underline{0}, \quad (2)$$

Snadno zjistíme, že platí

$$\frac{\partial \mu}{\partial \eta} = \frac{e^\eta}{(1 + e^\eta)^2} = \mu(1 - \mu)$$

odkud dostaneme

$$\frac{\partial \mu_i}{\partial \beta} = \mu_i(1 - \mu_i) \underline{x}_{i\bullet}$$

Když zavedeme diagonální matici rozptylů jednotlivých pozorování

$$D(\beta) = \text{diag} \{ \mu_1(1 - \mu_1), \dots, \mu_n(1 - \mu_n) \},$$

můžeme Fisherovu informační matici (vzhledem k ⁽¹⁾) zapsat jako

$$\begin{aligned} I(\beta) &= \underline{X}' D(\beta) \underline{X} \\ &= \sum_{i=1}^n \mu_i(\beta)(1 - \mu_i(\beta)) \underline{x}_{i\bullet} \cdot \underline{x}'_{i\bullet}. \end{aligned}$$

Vzhledem k tomu, že matice D je pozitivně definitní, je Fisherova informační matice přinejmenším pozitivně semidefinitní a v případě úplné sloupcové hodnosti matice \mathbf{X} dokonce pozitivně definitní. Tato skutečnost usnadňuje iterační řešení soustavy normálních rovnic.

Označme řešení normální rovnice (2) jako \underline{b} . Asymptotickou variační maticí je inverzní matice k Fisherově informační matici. V praxi při jejím výpočtu za neznámé parametry do $J(\underline{\beta})$ dosadíme odhady metodou maximální věrohodnosti, které jsou konzistentní, takže také $J(\underline{b})$ je konzistentním odhadem $J(\underline{\beta})$. Všimněme si, že na rozdíl od lineárního modelu v asymptotické matici nevystupuje parametr měřítka (rozptyl σ^2). Na druhé straně, jak jsme upozornili, závisí rozptyl odhadů na odhadovaných parametrech β .

11.4. Interpretace parametrů

Věnujme se interpretaci parametrů β_0, β_1 v modelu $\eta_i = \beta_0 + \beta_1 x_i$.

Binární nezávisle proměnná

Předpokládejme, že jednorozměrná veličina x nabývá právě dvou hodnot. Bez újmy na obecnosti to jsou hodnoty 0 a 1, takže x je umělá proměnná k dvouhodnotovému faktoru a vyjadřuje nepřítomnost či přítomnost nějakého jevu.

Pro $x=0$ jsou šance rovny:

$$\omega(0) = \frac{P(Y=1)}{P(Y=0)} = \frac{\frac{e^{\beta_0}}{1+e^{\beta_0}}}{\frac{1}{1+e^{\beta_0}}} = e^{\beta_0}$$

Parametr β_0 je tedy roven logitu pravděpodobnosti výskytu sledovaného jevu pro $x=0$:

$$\beta_0 = \ln \frac{P(Y=1)}{P(Y=0)}.$$

Pro $x=1$ je odpovídající šance rovna

$$\omega(1) = \frac{\frac{e^{\beta_0+\beta_1}}{1+e^{\beta_0+\beta_1}}}{\frac{1}{1+e^{\beta_0+\beta_1}}} = e^{\beta_0+\beta_1}.$$

Poměr šancí (odds ratio) pro dvě hodnoty x je pak roven

$$\frac{\omega(1)}{\omega(0)} = \frac{e^{\beta_0+\beta_1}}{e^{\beta_0}} = e^{\beta_1},$$

takže parametr β_1 je roven logaritmu poměrů šancí. Pokud pravděpodobnost sledovaného jevu na hodnotě x nezávisí, je poměr šancí roven jedné, tedy $\beta_1 = 0$.

Když známe odhad b_1 parametru β_1 i jeho asymptotický rozptyl σ_{11}^2 (označme $V = (J(b_0, b_1))^{-1}$ s řádky a sloupci číslovanými od nuly), můžeme testovat nulovou hypotézu $H_0: \beta_1 = 0$ pomocí statistiky

$$Z = \frac{b_1}{\sigma_{11}},$$

kteřá má za platnosti nulové hypotézy asymptoticky normované normální rozdělení $N(0;1)$. Některé statistické pakety zde předpokládají studentovo rozdělení s $(n - k - 1)$ stupni volnosti - t_{n-k-1} .

V případě binárního x nalezneme odhady b_0, b_1 snadno, přímo z odhadů šancí $\omega(0), \omega(1)$. Pro $x=i$ a $Y=j$ označme zjištěnou četnost jako N_{ij} . Celkem tedy máme $n_{i\cdot} = N_{i0} + N_{i1}$ pozorování s hodnotou $x=i$. Hledané odhady jsou

$$\hat{\omega}(0) = \frac{N_{01}/n_{0\cdot}}{N_{00}/n_{0\cdot}} = \frac{N_{01}}{N_{00}},$$

$$\hat{\omega}(1) = \frac{N_{11}/n_{0\cdot}}{N_{10}/n_{0\cdot}} = \frac{N_{11}}{N_{10}}.$$

Odtud snadno dostaneme

$$b_0 = \ln \frac{N_{01}}{N_{00}}, \quad b_1 = \ln \frac{N_{00} \cdot N_{11}}{N_{01} \cdot N_{10}}$$

Pokusme se explicitně vyjádřit rozptyl σ_{11}^2 . Diagonální matice $D(b_0, b_1)$ má pouze dvojí diagonální prvky, $n_{0\cdot}$ prvků s odhadem rozptylu pro $x=0$ a $n_{1\cdot}$ prvků s odhadem rozptylu pro $x=1$. Zmíněné odhady rozptylu závisle proměnné jsou rovny $N_{x0} \cdot N_{x1} / n_{x\cdot}^2$. Odhad Fisherovy informační matice má tedy tvar

$$J(b_0, b_1) = \begin{pmatrix} \left(\frac{N_{00} N_{01}}{n_{0\cdot}} + \frac{N_{10} N_{11}}{n_{1\cdot}} \right) & \frac{N_{00} N_{01}}{n_{0\cdot}} \\ \frac{N_{00} N_{01}}{n_{0\cdot}} & \frac{N_{00} N_{01}}{n_{0\cdot}} \end{pmatrix}.$$

Protože determinant této matice je roven $N_{00} N_{01} N_{10} N_{11} / (n_{0\cdot} n_{1\cdot})$, dostaneme příslušný prvek (vpravo dole) matice $V = (J(b_0, b_1))^{-1}$ jako

$$\sigma_{11}^2 = \frac{1}{N_{00}} + \frac{1}{N_{01}} + \frac{1}{N_{10}} + \frac{1}{N_{11}}.$$



Řešený příklad

Následující data podávají informaci o tom, zda matka kojila své dítě ještě ve 24. týdnu. Zabývejme se otázkou, zda tato skutečnost závisí na tom, zda bylo těhotenství plánováno.

		Koj24	
		0	1
Plan	0	35	6
	1	36	22

Z tabulky dostaneme snadno příslušné četnosti. Je zřejmé, že u plánovaných těhotenství kojilo ve 24. týdnu života dítěte relativně více matek, než u těhotenství neplánovaných. Provedme explicitní výpočty:

$$N_{00} = 35 \quad N_{01} = 6 \quad N_{10} = 36 \quad N_{11} = 22$$

$$b_0 = \ln \frac{N_{01}}{N_{00}} = \ln \frac{6}{35} = -1,764$$

$$b_1 = \ln \frac{N_{00} \cdot N_{11}}{N_{01} \cdot N_{10}} = \ln \frac{35 \cdot 22}{6 \cdot 36} = 1,271$$

$$\sigma_{11}^2 = \frac{1}{N_{00}} + \frac{1}{N_{01}} + \frac{1}{N_{10}} + \frac{1}{N_{11}} = \frac{1}{35} + \frac{1}{6} + \frac{1}{36} + \frac{1}{22} = 0,268$$

$$H_0 : \beta_1 = 0$$

$$H_A : \beta_1 \neq 0$$

$$Z = \frac{b_1}{\sigma_{11}} = \frac{1,271}{\sqrt{0,268}} = 2,453$$

$$p\text{-value} = 0,014$$

\Rightarrow Zamítáme hypotézu o tom, že kojení ve 24. týdnu nezávisí na plánování těhotenství.

11.5. Testování podmodelu pomocí rozdílů deviancí

K testování podmodelu lze použít test daný rozdílem deviancí, založený na odhadech \underline{b} v modelu a $\tilde{\underline{b}}$ v podmodelu. Test se provádí prostřednictvím tzv. **deviancí**, které nyní zavedeme.

Uvažujme nejprve nejhorší možný model, který má právě tolik parametrů, kolik je pozorování, tedy n . Přílehavější model (s větší hodnotou věrohodnostní funkce) neexistuje.

Tento nejbohatší model se nazývá **saturovaný**. Označme maximální hodnotu věrohodnostní funkce v saturovaném modelu symbolem ℓ_{\max} . Každý jiný představitelný model je podmodelem saturovaného modelu. Příléhavost běžného modelu můžeme posoudit pomocí **deviance**

$$D(\underline{b}) = 2(\ell_{\max} - \ell(\underline{b})).$$

Čím je náš model méně příléhavý, tím je hodnota deviance D větší, podobně, jako je větší reziduální součet čtverců v lineárním modelu pro méně výstižný model.

Pojem deviance jsme zde zavedli zejména proto, že se používá i v souvislosti s logistickou regresí, byť je zde hodnota ℓ_{\max} triviální. Saturovaný model má n parametrů μ_1, \dots, μ_n . Odhadem střední hodnoty μ_i je v případě logistické regrese přímo Y_i , takže je

$$\ell_{\max} = \sum_{i=1}^n (Y_i \cdot \ln Y_i + (1 - Y_i) \ln (1 - Y_i)) = 0.$$

Označíme-li odhady pravděpodobnosti jedničky v běžném modelu jako $\hat{\mu}_i = \mu(x_{i\cdot})$, devianci v modelu logistické regrese vyjádříme jako

$$\begin{aligned} D(\underline{b}) &= 2 \sum_{i=1}^n \left(Y_i \cdot \ln \frac{Y_i}{\hat{\mu}_i} + (1 - Y_i) \ln \left(\frac{1 - Y_i}{1 - \hat{\mu}_i} \right) \right) = \\ &= -2 \sum_{i=1}^n (Y_i \cdot \ln \hat{\mu}_i + (1 - Y_i) \ln (1 - \hat{\mu}_i)) \end{aligned}$$

Vraťme se k obecné situaci. V našem běžném modelu teď uvažujeme nějaký podmodel například po vyloučení části regresorů. Testovou statistiku danou rozdílem deviancí modelu a podmodelu (s odhady parametrů \tilde{b}_0, \tilde{b}) vyjádříme následně:

$$D(\tilde{b}) - D(\underline{b})$$

Tato testová statistika (rozdíl deviancí) má (za platnosti testovaného podmodelu) asymptotické rozdělení $\chi^2(f)$, kde f je rovno rozdílu počtu nezávislých parametrů v porovnávaných modelech. Hypotézu $H_0: \beta_1 = 0$ a podobné hypotézy o nulovosti jedné složky vektoru $\underline{\beta}$ lze testovat právě tímto testem rozdílu deviancí.

Podobnost deviance k reziduálnímu součtu čtverců vedla ke snaze rozšířit pojem koeficientu determinace také na logistickou regresí. K tomuto účelu nejprve zavedeme pojem **nulového modelu**. Jde o model, kde jsou všechny střední hodnoty $\mu_i = EY_i$ shodné. Hodnotu věrohodnostní funkce či deviance označíme ℓ_0 , resp. D_0 . Hodnotu logaritmicke věrohodnostní funkce normálního lineárního modelu lze vyjádřit jako

$$\ell(\underline{b}) = -\frac{n}{2}(1 + \ln(2\pi)) - \ln n - \frac{n}{2} \ln RSS,$$

kde

$$\hat{\sigma}^2 = \frac{1}{n} \cdot \sum_{i=1}^n (Y_i - \bar{Y})^2 = \frac{RSS}{n}$$

Koeficient determinace v lineárním modelu lze vyjádřit jako

$$\begin{aligned} R^2 &= 1 - \frac{RSS}{RSS_0} \\ &= 1 - e^{\left(-\frac{2}{n}(l(b)) - l_0\right)} \\ &= 1 - e^{\left(-\frac{2}{n}(D(b)) - D_0\right)} \end{aligned}$$

V uvedeném vztahu je uveden návod k výpočtu i pro případ logistické regrese. Přílehavější model, než je saturovaný, nalézt nelze. Deviance saturovaného modelu je, jak víme, rovna nule, takže koeficient R^2 nemůže překročit hodnotu

$$R_{\max}^2 = 1 - e^{\left(-\frac{1}{n}D_0\right)}$$

Po dosazení do vztahu pro D_0 dostaneme:

$$D_0 = -2 \left(\sum_{i=1}^n Y_i \ln \left(\sum_{i=1}^n Y_i \right) + \left(n - \sum_{i=1}^n Y_i \right) \ln \left(n - \sum_{i=1}^n Y_i \right) - n \ln n \right),$$

neboť pro všechna i je odhadem střední hodnoty relativní četnost jedniček, totiž $\sum_{i=1}^n (Y_i/n)$.

Nagelkerke (1991) proto navrhl upravit definici zobecněného koeficientu determinace na

$$\begin{aligned} R_N^2 &= \frac{R^2}{R_{\max}^2} \\ &= \frac{1 - e^{\left(-\frac{2}{n}(D(b)) - D_0\right)}}{1 - e^{\left(-\frac{1}{n}D_0\right)}} \end{aligned}$$

11.6. Modifikovaná logistická regrese – nástroj pro diskriminaci

Logistická regrese nebyla původně vytvořena pro účely diskriminace, ale jak si ukážeme, lze ji pro ni s úspěchem použít.

Model logistické regrese, který je upravený pro účely diskriminace, je definován následovně. Necht' Y_1, \dots, Y_n je posloupnost nezávislých náhodných veličin s alternativním rozdělením, jehož parametr splňuje

$$P(Y_i = 1 | \underline{X}_i = \underline{x}_i) = [1 + e^{(-\beta_0 - \beta' \underline{x}_i)}]^{-1},$$

$$P(Y_i = 0 | \underline{X}_i = \underline{x}_i) = [1 + e^{(\beta_0 + \beta' \underline{x}_i)}]^{-1},$$

kde (β_0, β') je neznámý $(p+1)$ rozměrný parametr a $\underline{X}_1, \dots, \underline{X}_n$ je posloupnost nezávislých náhodných veličin. Tento model má tzv. učicí fázi, ve které známe u každého objektu jak hodnoty \underline{X}_i , tak hodnoty Y_i (tj. víme, do které skupiny ten který objekt patří). Na základě této znalosti odhadneme parametry β_0, β a poté dostaneme odhad $\tilde{\pi}(\underline{x})$ funkce $\pi(\underline{x})$, kde

$$\pi(\underline{x}) = P(Y = 1 | \underline{X} = \underline{x}) = [1 + e^{(-\beta_0 - \beta' \underline{x})}]^{-1}.$$

Další objekt, u kterého neznáme jeho zařazení a u něhož jsme naměřili hodnotu \underline{X} pomocných znaků, přiřadíme do jedné ze skupin podle hodnoty rozhodovací funkce.

Zde sice neznáme apriorní hustotu veličiny Y ani podmíněnou hustotu náhodné veličiny \underline{X} za podmínky $Y=y$, ale pro výpočet optimální rozhodovací funkce nám postačí znalost podmíněné hustoty Y za podmínky $\underline{X}=\underline{x}$, která je určena hodnotou funkce $\pi(\underline{x})$. Pokud $\delta(\underline{x}) = j$, je totiž

$$E[L(Y, \delta(\underline{X})) | \underline{X} = \underline{x}] = \sum_{y=0}^1 L(y, \delta(\underline{x})) p(y | \underline{x}) = \sum_{y=0}^1 L(y, j) p(y | \underline{x}) =$$

$$= L(1 - j, j) p(1 - j | \underline{x}) = \begin{cases} \pi(\underline{x}), & j = 0, \\ 1 - \pi(\underline{x}), & j = 1. \end{cases}$$

Tedy

$$\min_{\delta \in D} E[L(Y, \delta(\underline{X})) | \underline{X} = \underline{x}] = \min \{ \pi(\underline{x}), 1 - \pi(\underline{x}) \}.$$

Toto minimum existuje $\forall x \in R^p$ a tudíž můžeme psát

$$\delta^*(\underline{x}) = \arg \min_{j=0,1} L(1 - j, j) p(1 - j | \underline{x}).$$

Tedy objekt, na němž jsme naměřili hodnotu \underline{X} pomocných znaků, zařadíme do první skupiny, pokud

$$\pi(\underline{X}) \geq 1 - \pi(\underline{X}), \quad tj.$$

$$\beta_0 + \beta' \underline{X} \geq 0$$

Tudíž objekt zařadíme do první skupiny, pokud $S(\underline{X}) \geq 0$ a do nulté skupiny, pokud $S(\underline{X}) < 0$. Přitom $S(\underline{x}) = \beta_0 + \beta' \underline{x}$. Dodejme, že pokud $S(\underline{X}) = 0$, tj. $\pi(\underline{X}) = \frac{1}{2}$, můžeme objekt zařadit do libovolné skupiny, aniž bychom zvýšili hodnotu pravděpodobnosti chyby. K vlastní diskriminaci však musíme použít odhad $\tilde{S}(\underline{x})$ funkce $S(\underline{x})$, ve kterém jsou neznámé parametry β_0, β nahrazeny odhady $\tilde{\beta}_0, \tilde{\beta}$. Pomocí metodiky kapitoly 11.3, tedy

$$\tilde{S}(\underline{x}) = \tilde{\beta}_0 + \tilde{\beta}' \underline{x}.$$

Hlavní výhodou tohoto modelu je fakt, že neklade žádné podmínky na rozdělení náhodných vektorů $\underline{X}_1, \dots, \underline{X}_n$.

Poznámky:

1. V regresních modelech bývají obvykle veličiny $\underline{X}_1, \dots, \underline{X}_n$, jež jsou naměřeny na objektech učící skupiny, nenáhodné, resp. jejich hodnoty jsou nastaveny experimentátorem. Může se také stát, že i v případě spojitého rozdělení veličin $\underline{X}_1, \dots, \underline{X}_n$ se některé z naměřených hodnot $\underline{X}_1, \dots, \underline{X}_n$ opakují. Nic z právě uvedeného však není na závadu. Stále můžeme na veličiny $\underline{X}_1, \dots, \underline{X}_n$ pohlížet jako na náhodné. Pro určení teoretické diskriminační funkce nepotřebujeme znát hustotu veličin $\underline{X}_1, \dots, \underline{X}_n$, postačuje nám znalost podmíněné hustoty veličin Y_i za podmínky $\underline{X}_i = \underline{x}_i$, $i=1, \dots, n$.

2. Prospektivní studie

Jak je uvedeno, mohou se naměřené hodnoty $\underline{X}_1, \dots, \underline{X}_n$ opakovat a být nastaveny experimentátorem, tj. mohou být nenáhodné (tzv. prospektivní studie). Nechť I je počet různých hodnot veličin $\underline{X}_1, \dots, \underline{X}_n$ v učící skupině a $\underline{x}_1, \dots, \underline{x}_n$ jsou tyto hodnoty. Nechť nyní $Y_{i,j}$, $i=1, \dots, I$, $j=1, \dots, m_i$, vyjadřují zařazení objektů do skupin. Přitom m_i je počet objektů s hodnotou vysvětlujících znaků \underline{x}_i , celkový počet objektů je tedy nyní roven $n = \sum_{i=1}^I m_i$. Nechť $Y_{i\bullet} = \sum_{j=1}^{m_i} Y_{i,j}$. Jestliže jsou hodnoty vysvětlujících veličin nenáhodné a nenáhodná jsou i čísla m_1, \dots, m_I , měli bychom při hledání maximálně věrohodných odhadů parametrů β_0 a β maximalizovat sdruženou hustotu veličin $Y_{1\bullet}, \dots, Y_{I\bullet}$ za podmínky $\underline{X}_1 = \underline{x}_1, \dots, \underline{X}_I = \underline{x}_I$. Rozdělení veličin $Y_{i\bullet}$ za podmínky $\underline{X}_i = \underline{x}_i$ je binomické s parametry m_i a $\pi(\underline{x}_i)$. Uvedená podmíněná sdružená hustota je potom rovna

$$f_{\beta_0, \beta}^*(y_{1\bullet}, \dots, y_{n\bullet} | \underline{x}_1, \dots, \underline{x}_n) = \prod_{i=1}^I \binom{m_i}{y_{i\bullet}} \pi(\underline{x}_i)^{y_{i\bullet}} (1 - \pi(\underline{x}_i))^{m_i - y_{i\bullet}}.$$

Logaritmická věrohodnostní funkce je tedy rovna

$$\begin{aligned} I^*(\beta_0, \beta) &= \sum_{i=1}^I \ln \binom{m_i}{Y_{i\bullet}} + \sum_{i=1}^I [Y_{i\bullet}(\beta_0 + \beta' \underline{X}_i) - m_i \ln(1 + e^{\beta_0 + \beta' \underline{X}_i})] = \\ &= \sum_{i=1}^I \ln \binom{m_i}{Y_{i\bullet}} + \sum_{i=1}^I \sum_{j=1}^{m_i} [Y_{i,j}(\beta_0 + \beta' \underline{X}_i) - \ln(1 + e^{\beta_0 + \beta' \underline{X}_i})] \end{aligned}$$

Vztah pro logaritmickou věrohodnost se tedy od vztahu uvedeného v kapitole 12.3 liší pouze o člen $\sum_{i=1}^I \ln \binom{m_i}{Y_{i\bullet}}$, který nezávisí na β_0 ani na β . Tudíž obě tyto funkce nabývají svého maxima ve stejném bodě.

11.7. Ověřování předpokladů logistické regrese

Proč testovat předpoklady modelů? Logistický model sice neklade žádné zvláštní požadavky na rozdělení náhodných veličin $\underline{X}_1, \dots, \underline{X}_n$, ale zato předpokládá specifický tvar pravděpodobnosti $P(Y = 1 | \underline{X} = \underline{x})$. Skutečnost, že opravdu platí

$P(Y = 1 | \underline{X} = \underline{x}) = [1 + e^{(-\beta_0 - \beta' \underline{x})}]^{-1}$, bychom měli ověřit nejlépe pomocí nějakého statistického testu.

Dále uvedeme některé testy dobré shody pro model logistické regrese.

□ Základní testy dobré shody pro model logistické regrese

Pro popis testů dobré shody použijeme značení zavedené v rámci poznámky 2. Tj. nechť učící skupina obsahuje n objektů. Nechť I je počet různých hodnot veličin $\underline{X}_1, \dots, \underline{X}_n$ v učící skupině a $\underline{x}_1, \dots, \underline{x}_I$ jsou tyto hodnoty. Jak již bylo dříve řečeno, veličiny $\underline{X}_1, \dots, \underline{X}_n$ nemusejí být nutně náhodné (obvyklý jev regresních modelů). Mohou se tedy některé z hodnot $\underline{X}_1, \dots, \underline{X}_n$ opakovat, i když jsou veličiny $\underline{X}_1, \dots, \underline{X}_n$ spojitě rozdělené. Celý model logistické regrese pracuje totiž s podmíněným rozdělením veličin Y_1, \dots, Y_n za podmínky $\underline{X}_1 = \underline{x}_1, \dots, \underline{X}_n = \underline{x}_n$. Hodnoty Y_1, \dots, Y_n vyjadřující zařazení i -tého objektu do jedné ze dvou skupin přeznačme na $Y_{i,j}$, $i = 1, \dots, I$, $j = 1, \dots, m_i$, kde m_i je počet objektů s hodnotou vysvětlujících znaků \underline{X}_i a $Y_{i,j}$, $j = 1, \dots, m_i$ označuje zařazení objektů, u nichž mají

vysvětlující znaky hodnotu $\underline{X}_i = \underline{x}_i$. Dále označme $n_1 = \sum_{i=1}^I \sum_{j=1}^{m_i} Y_{i,j}$, $n_0 = \sum_{i=1}^I \sum_{j=1}^{m_i} (1 - Y_{i,j})$.

Metodou maximální věrohodnosti získáme odhady $\tilde{\beta}_0$, $\tilde{\beta}$ parametrů β_0 , β . Pomocí těchto odhadů spočítáme odhady logistických pravděpodobností $\tilde{\pi}(\underline{x}_i) = \tilde{\pi}_i = [1 + e^{(-\tilde{\beta}_0 - \tilde{\beta}' \underline{x}_i)}]^{-1}$. Přímou z věrohodnostních rovnic plynou vztahy

$$n_1 = \sum_{i=1}^I \sum_{j=1}^{m_i} Y_{i,j} = \sum_{i=1}^I m_i \tilde{\pi}_i, \quad n_0 = \sum_{i=1}^I \sum_{j=1}^{m_i} (1 - Y_{i,j}) = \sum_{i=1}^I m_i (1 - \tilde{\pi}_i)$$

Popíšeme test dobré shody založený na Pearsonově χ^2 statistice.

Celou situaci lze popsat kontingenční tabulkou typu 2xI s danými marginálními sloupcovými četnostmi. Přitom i-tý sloupec tabulky reprezentuje binomické rozdělení s parametry m_i , $\pi(\underline{x}_i) = \pi_i$.

Tabulka odhadnutých četností má tvar:

		X			
		\mathbf{x}_1	...	\mathbf{x}_I	
Y	1	$m_i \tilde{\pi}_i$...	$m_i \tilde{\pi}_i$	n_1
	0	$m_i(1 - \tilde{\pi}_i)$...	$m_i(1 - \tilde{\pi}_i)$	n_0
		m_i	...	m_i	n

Tabulka empirických (pozorovaných) četností je tvaru:

		X			
		\mathbf{x}_1	...	\mathbf{x}_I	
Y	1	$Y_{i\cdot}$...	$Y_{i\cdot}$	n_1
	0	$m_i - Y_{i\cdot}$...	$m_i - Y_{i\cdot}$	n_0
		m_i	...	m_i	n

Protože máme dány marginální sloupcové četnosti, jsou hodnoty v naší tabulce vázány I podmínkami $(m_i \pi_i + m_i(1 - \pi_i) = m_i, \dots, m_i \pi_i + m_i(1 - \pi_i) = m_i)$. Tento údaj budeme potřebovat pro výpočet stupňů volnosti v Pearsonově testové statistice, která má tvar

$$Z^2 = \sum_{i=1}^I \frac{(Y_{i\cdot} - m_i \tilde{\pi}_i)^2}{m_i \tilde{\pi}_i} + \sum_{i=1}^I \frac{(m_i - Y_{i\cdot} - m_i(1 - \tilde{\pi}_i))^2}{m_i(1 - \tilde{\pi}_i)} = \sum_{i=1}^I \frac{(Y_{i\cdot} - m_i \tilde{\pi}_i)^2}{m_i \tilde{\pi}_i (1 - \tilde{\pi}_i)}.$$

Pomocí této statistiky lze testovat shodu dat v pozorované tabulce s tabulkou teoretickou, která je v našem případě založena na modelu logistické regrese. Při hypotéze H_0 : platí logistický model, má statistika Z^2 asymptoticky rozdělení χ^2 s následujícím počtem stupňů volnosti: velikost kontingenční tabulky – počet vazeb v teoretické tabulce – počet odhadnutých parametrů = $2I - I - (p+1) = I - (p+1)$. Tedy při platnosti H_0 je $Z^2 \rightarrow \chi^2(I - (p + 1))$. Samozřejmě musí být splněna podmínka $I > (p + 1)$.

Připomeňme však jeden problém, který je spojen s použitím výše uvedeného testu dobré shody. Rozdělení testové statistiky je získáno asymptoticky pro $n \rightarrow \infty$ a v praktických situacích (obzvláště v případech, kdy alespoň jedna složka vysvětlujících náhodných vektorů $\underline{X}_1, \dots, \underline{X}_n$ má spojitý charakter) je $I \approx n$. Tedy s rozsahem výběru roste též počet stupňů volnosti testových statistik. McCullagh a Nelder (1989) uvedli, že pro $I \approx n$ je při platnosti H_0 $EZ^2 < I - (p + 1)$. V roce 1989 však Hosmer a Lemeshow provedením rozsáhlých simulací potvrdili, že aproximace střední hodnoty statistiky Z^2 výrazem $I - (p + 1)$ je prakticky použitelná.

Dalším problémem, který je spojen s použitím Pearsonovy Z^2 statistiky, je požadavek na dostatečně velké teoretické četnosti, např. $m_i \tilde{\pi}_i \geq 5$, $m_i(1 - \tilde{\pi}_i) \geq 5$, $i = 1, \dots, I$, který též nebude obvykle splněn, pokud $I \approx n$.

Oba výše uvedené problémy lze vyřešit, pokud bude $I < n$ pevné. Popíšeme si testovou statistiku, navrženou již zmíněnými Hosmerem a Lemeshowem, která je založena právě na této myšlence.

□ Hosmerovy-Lemeshowovy testy

Statistiky vhodné pro testy dobré shody navržené Hosmerem a Lemeshowem jsou založeny na seskupení některých sloupců kontingenční tabulky uvedené v kapitole 11.6.1. Nejprve zvolíme $g < n$ počet požadovaných sloupců kontingenční tabulky. Pozorování přeznačíme tak, aby platilo $\tilde{\pi}_1 \leq \tilde{\pi}_2 \leq \dots \leq \tilde{\pi}_I$.

Výsledkem seskupení jsou sloupce obsahující přibližně stejný počet pozorování. Do prvního sloupce zařadíme přibližně n/g pozorování $Y_{1,1}, \dots, Y_{1,m_1}, \dots, Y_{n'_1,1}, \dots, Y_{n'_1,m_{n'_1}}$, kterým náleží

nejmenší odhadnuté pravděpodobnosti $\tilde{\pi}_i$, $i = 1, \dots, n'_1$. Naší snahou je, aby $m_1^* = \sum_{i=1}^{n'_1} m_i$ bylo

co možná nejbliže hodnotě n/g . Postupně vytváříme další sloupce, až konečně v g -tém sloupci je přibližně n/g pozorování $Y_{t,1}, \dots, Y_{t,m_t}, \dots, Y_{t,1}, \dots, Y_{t,m_t}$, kterým náleží největší

odhadnuté pravděpodobnosti $\tilde{\pi}_i$, $i = t, \dots, I$, $t = \sum_{k=1}^{g-1} n'_k + 1$, přitom n'_1, \dots, n'_g označují počty

různých hodnot vysvětlujících veličin X_1, \dots, X_I v jednotlivých sloupcích (tedy platí

$\sum_{k=1}^g n'_k = I$). Necht' $t_0 = 0$, $t_k = \sum_{j=1}^k n'_j$, $k = 1, \dots, g$ a necht' m_1^*, \dots, m_g^* jsou počty pozorování

v jednotlivých sloupcích, tedy splňují vztahy $m_k^* = \sum_{i=t_{k-1}+1}^{t_k} m_i$, $k = 1, \dots, g$. Snažíme se, aby m_k^*

bylo co nejbliže hodnotě $n/g \quad \forall k = 1, \dots, g$. Je-li $g=10$, nazývají se hodnoty odhadnutých pravděpodobností, jež oddělují jednotlivé sloupce, jako **decily rizika**. Samotné sloupce kontingenční tabulky budeme v naší práci nazývat **decilovými skupinami**¹.

Pro novou kontingenční tabulku typu $2 \times g$ nyní spočítáme odhadnuté teoretické a empirické četnosti. Odhadnutá teoretická četnost pro řádek $Y = 1$ a k -tý sloupec je

$c_k = \sum_{i=t_{k-1}+1}^{t_k} m_i \tilde{\pi}_i$, $k = 1, \dots, g$, pro řádek $Y = 0$ a k -tý sloupec je

$m_k^* - c_k = \sum_{i=t_{k-1}+1}^{t_k} m_i (1 - \tilde{\pi}_i)$, $k = 1, \dots, g$. Empirická četnost pro řádek $Y = 1$ a k -tý sloupec je

$o_k = \sum_{i=t_{k-1}+1}^{t_k} \sum_{j=1}^{m_i} Y_{i,j}$, $k = 1, \dots, g$, pro řádek $Y = 0$ a k -tý sloupec je

¹ O decilech se mluví i v situacích, kdy není v každém sloupci přesně desetina všech pozorování.

$$m_k^* - o_k = \sum_{i=t_{k-1}+1}^{t_k} \sum_{j=1}^{m_i} (1 - Y_{i,j}), \quad k = 1, \dots, g. \text{ Dále necht' } \bar{\pi}_k = \frac{1}{m_k^*} \sum_{i=t_{k-1}+1}^{t_k} m_i \tilde{\pi}_i = \frac{c_k}{m_k^*}, \quad k = 1, \dots, g$$

je odhad pravděpodobnosti $P(Y = 1 | X \in \{x_{t_{k-1}+1}, \dots, x_{t_k}\})$.

Tabulka odhadnutých teoretických četností má tedy tvar:

		X			
		1. sloupec	...	g-tý sloupec	
Y	1	$m_1^* \bar{\pi}_1$...	$m_g^* \bar{\pi}_g$	n_1
	0	$m_1^* (1 - \bar{\pi}_1)$...	$m_g^* (1 - \bar{\pi}_g)$	n_0
		m_1^*	...	m_g^*	n

Tabulka empirických (pozorovaných) četností je tvaru:

		X			
		1. sloupec	...	g-tý sloupec	
Y	1	o_1	...	o_g	n_1
	0	$m_1^* - o_1$...	$m_g^* - o_g$	n_0
		m_1^*	...	m_g^*	n

Testová statistika Hosmerova-Lemeshowova testu pro ověřování shody s modelem logistické regrese má tvar běžné Pearsonovy χ^2 statistiky pro ověřování shody teoretické a empirické tabulky, tedy

$$\hat{C} = \sum_{k=1}^g \frac{(o_k - m_k^* \bar{\pi}_k)^2}{m_k^* \bar{\pi}_k} + \sum_{k=1}^g \frac{(m_k^* - o_k - m_k^* (1 - \bar{\pi}_k))^2}{m_k^* (1 - \bar{\pi}_k)} =$$

$$= \sum_{k=1}^g \frac{(o_k - m_k^* \bar{\pi}_k)^2}{m_k^* \bar{\pi}_k (1 - \bar{\pi}_k)}$$

Užitím rozsáhlých simulací bylo ukázáno, že pro $I = n$ má při platnosti hypotézy H_0 statistika \hat{C} přibližně rozdělení χ^2 o $(g-2)$ stupních volnosti. Podle Hosmera a Lemeshowa lze při platnosti H_0 dobře aproximovat rozdělení statistiky \hat{C} rozdělením χ^2 o $(g-2)$ stupních volnosti též v situaci, kdy $I \approx n$.

Aby bylo možné použít výše uvedenou statistiku, měli bychom ještě ověřit $m_k \tilde{\pi}_k \geq 5$, $m_k (1 - \tilde{\pi}_k) \geq 5$, $k = 1, \dots, g$. Není-li tato podmínka splněna, měli bychom sloučit některé sloupce tabulky, a tedy snížit hodnotu čísla g . Autoři však tvrdí, že porušení této podmínky není příliš na závadu. Hosmer a Lemeshow dále doporučují volit $g \geq 6$, neboť pro $g > 6$ je již statistika \hat{C} málo citlivá na rozdíly mezi teoretickými a empirickými četnostmi a téměř vždy indikuje shodu s modelem.



Otázky 11.

1. Čemu se ve statistice říká diskriminace?
2. Srovnajte logistický a normální lineární regresní model.
3. Vysvětlete pojmy šance a logit.
4. Vysvětlete, jak se testují podmodely pomocí deviancí.
5. Jak se využívá logistická regrese pro diskriminaci?
6. Jaké předpoklady je třeba testovat u logistické regrese?
7. Princip Hosmer-Lemeshowových testů.

LITERATURA

-
- [1] ANDĚL J.: *Matematická statistika*, SNTL, 1978.
- [2] BOX G.E.P., TIAO G.C.: *Bayesian Inference in Statistical Analysis*; Reading, Mass.: Addison-Wesley, 1973.
- [3] BRIŠ R., LITSCHMANNOVÁ M.: *STATISTIKA I. pro kombinované a distanční studium*, Elektronické skriptum VŠB TU Ostrava, 2004.
- [4] BRIŠ R.: *Aplikovaná matematika*, Regionální centrum celoživotního vzdělávání, VŠB–Technická univerzita Ostrava, 2003, ISBN 80-248-0313-5.
- [5] BRIS R.: *Bayes approach in RDT using accelerated and long-term life data*, Reliability Engineering and System Safety, ELSEVIER, Vol.67 No.1 January 2000, ISSN: 0951-8320.
- [6] HEBÁK P., HUSTOPECKÝ J., JAROŠOVÁ E., PECÁKOVÁ I.: *Vícerozměrné statistické metody (1)*, Praha 2004 Informatorium, ISBN 80-7333-025-3.
- [7] HOSMER DW, LEMESHOW S: *Applied Logistic Regression*, New York John Wiley & Sons, Inc., 1989
- [8] HURT J.: *Teorie spolehlivosti*, SPN Praha 1984.
- [9] JEFREYS H.: *Theory of probability*, London: Cambridge University Press, 1961.
- [10] KELLNER A.: *Maximal Data Information Prior Distributions. In New Methods in the Applications of Bayesian Methods*, A.Aykac and C.Brumat (Eds.), Amsterdam: North Holland. 1977.
- [11] MARTZ H.F., WALLER R.A.: *Bayesian Reliability Analysis*, Wiley 1982, ISBN 0-471-86425-0.
- [12] MCGULLAGH P., NELDER J.A.: *Generalized linear Models*, Chapman Hall, 1989
- [13] MISRA K. B.: *Reliability Analysis and Prediction; A Methodology Oriented Treatment*, Elsevier, 1992, ISBN 0-444-89606-6.
- [14] NELSON W.: *Accelerated Testing, Statistical Models, Test Plans, and Data Analysis*; Wiley 1990, ISBN 0-471-52277-5.
- [15] NGUYEN H.T., ROGERS G. S.: *Fundamentals of Mathematical Statistics, Volume II: Statistical Inference*, Springer – Verlag New York, Inc., 1989, ISBN 0-387-97020-7.
- [16] SEGER J., HINDLS R., HRONOVÁ S.: *Statistika v hospodářství*, ETC Publishing, 1998, ISBN 80-86006-56-5
- [17] ŠTEPÁNEK, V. : *Matematická statistika v chemii*, skripta VŠCHT, SNTL, 1975.
- [18] WEERAHANDI S.: *Exact Statistical Methods for Data Analysis*, Springer-Verlag New York, Inc., 1995, ISBN 0-387-94360-9.
- [19] ZVÁRA K., *R & Regrese*, Skriptum MFF pro předmět STP094 Regrese, 2003.

OBSAH

1. MODEL Y A MODELOVÁNÍ	5
1.1. Model.....	5
1.2. Jedna z možných klasifikací modelu	6
1.3. Matematické modely	6
1.4. Některé typy matematických modelů	6
1.5. Přístupy k modelování	8
Shrnutí pojmů kapitoly 1	9
Otázky 1.	10
2. VYBRANÉ PRAVDĚPODOBNOSTNÍ MODEL Y	11
2.1. Erlangovo rozdělení	11
2.2. Weibullovo rozdělení	15
2.3. Logaritnicko-normální rozdělení.....	16
2.4. Vícerozměrné normální rozdělení	20
Otázky 2.	24
3. FUNKCE NÁHODNÉ VELIČINY	25
3.1. Funkce náhodné veličiny.....	25
3.2. Přibližné stanovení charakteristik funkce náhodné veličiny	28
Otázky 3.	28
Úlohy k řešení 3.	28
4. ZÁKLADY TEORIE SPOLEHLIVOSTI.....	29
4.1. Teorie spolehlivosti	29
Shrnutí kapitoly 4.1.	30
Otázky 4.1.	30
4.2. Základní pojmy.....	30
Shrnutí kapitoly 4.2.	33
Otázky 4.2.	33
4.3. Doba do poruchy	33
Shrnutí kapitoly 4.3.	35
Otázky 4.3.	36
Úlohy k řešení 4.3.	37
4.4. Intenzita poruch.....	37
Shrnutí kapitoly 4.4.	41
Otázky 4.4.	41
4.5. Zálohování.....	42
Shrnutí kapitoly 4.5.	43
Otázky 4.5.	43
Úlohy k řešení 4.5.	43
5. TEORIE ODHADU	44
5.1. Vlastnosti „dobrého“ bodového odhadu.....	44
Shrnutí kapitoly 5.1.	48
Otázky 5.1.	48
5.2. Konstrukce efektivních odhadů	48
Otázky 5.2.	52
5.3. Fisherova míra informace.....	52
Otázky 5.3.	56
5.4. Rao – Cramerova nerovnost	56
Otázky 5.4.	58
5.5. Metoda momentů.....	59
Shrnutí kapitoly 5.5.	60

Úlohy k řešení 5.5.	60
5.6. Metoda maximální věrohodnosti.....	61
Shrnutí kapitoly 5.6.	63
Úlohy k řešení 5.6.	63
6. NEÚPLNÁ DATA	65
6.1. Výběrové plány	65
6.2. Zrychlené zkoušky životnosti.....	67
6.3. Metoda maximální věrohodnosti pro neúplné výběry	68
7. ZÁKLADY BAYESOVY INDUKCE	79
7.1. Metoda maximální věrohodnosti.....	79
7.2. Úvod do Bayesovy indukce.....	81
7.3. Apriorní rozdělení	81
7.4. Aposteriorní rozdělení	83
7.5. Bayesovy estimátory	84
7.6. Bayesův intervalový odhad	86
Shrnutí kapitoly 7.	87
Otázky 7.	88
8. NÁHODNÉ PROCESY I	89
8.1. Náhodné procesy	89
8.2. Poissonův proces	89
8.3. Markovův proces.....	91
8.4. Příklady	93
8.5. Markovovy řetězce	97
Shrnutí kapitoly 8.	105
Otázky 8.	106
9. NÁHODNÉ PROCESY II.....	107
9.1. Spojitý parametr Markovovských řetězců.....	107
9.2. Ilustrace	109
9.3. Matice hustoty přechodu	112
9.4. Generování náhodných čísel.....	114
9.5. Kongruentní metoda	115
Shrnutí kapitoly 9.	117
Otázky 9.	118
10. DVOUFAKTOROVÁ ANALÝZA ROZPTYLU	119
10.1. Úvodní poznámky	119
10.2. Třídění dle dvou faktorů bez opakování.....	119
10.3. Třídění dle dvou faktorů s opakováním.....	122
10.4. Závěrečné poznámky	128
Otázky 10.	128
Úlohy k řešení 10.	129
11. LOGISTICKÁ REGRESE A JEJÍ UŽITÍ PRO DISKRIMINACI... 131	
11.1. Úvod.....	131
11.2. Tvar závislosti	132
11.3. Odhad parametrů	134
11.4. Intepretace parametrů	135
11.5. Testování podmodelu pomocí rozdílů deviancí.....	137
11.6. Modifikovaná logistická regrese – nástroj pro diskriminaci	139
11.7. Ověřování předpokladů logistické regrese	142
Otázky 11.	146