Analýza údajů o výskytu jevů v časovém intervalu

Aplikace modelu Poissonova rozložení

JAN ŘEHÁK
Ústav pro filozofii a sociologii ČSAV, Praha

1. Popis typu dat a jejich záznam

Výskyt jevů v čase sledujeme a zaznamenáváme pro potřebu společenskovědní aplikace často. Popisujeme tak výsledky řady procesů, jež jsou charakterizovány realizacemi určitých činností či výskytem nějakých jevů, procesů vstupů, výstupů či průchodu prvků daným systémem, sledování stavů, procesů psané i mluvené řeči apod.

Cílem stati je seznámit zájemce s jednoduchou metodikou zpracování takových dat, která jsou založena na pravděpodobnostním modelu Poissonova rozložení četností:

(a) bude uveden základní model zpracování takovýchto dat za předpokladu, že zaznamenáváme výskyty jevů v určitých časových intervalech s daným počátkem a koncem sledování procesu;
(b) budou uvedeny nejjednodušší zpracovatelské techniky, jejichž použití není závislé na počítačích a programech; bude také uvedeno rozšíření známých technik o metodu detekce odchylek od modelu (znaménkové schéma);
(c) pomocí zjednodušených ilustračních příkladů bude naznačena možná aplikace postupu na data dotazníkových šetření (přitom pochopitelně nesmí vzniknout dojem, že jde o jediné možné použití v sociologii!).

Uveďme si několik charakteristických proměnných, které mají běžné výzkumné použití:

1. Počet určitých činností osoby tak, jak byl zjištěn podle časového deníku (výzkum čintostí volného času, pracovní operace, apod.).
2. Počet určitých činností nebo stavů osoby podle jejího vlastního vyjádření, resp. podle vyjádření někoho jiného. („Kolikrát jste byl v měsíci březnu v kině?", „Kolikrát Vaše dítě onemocnělo v tomto pololetí?")
3. Odhad počtu jevů posuzovatelem, resp. i subjektivní představa o výskytu jevů.
4. Počet nehod na pracovišti.
5. Počet absencí jednotlivce.
6. Počet vystoupení osoby v průběhu schůze.
7. Počet návštěvníků galerie za určité období (resp. počet prodaných vstupenek).
8. Počet výjezdů první pomoci za noc.
9. Počet použití nějakého zařízení (automatu, telefonu apod.).
10. Pozorovaný počet průjezdů automobilů určitým místem, resp. počet zastavení automobilů na nějakém místě.

Ve všech uvedených případech je dán pevný časový interval.

Původní záznam o výskytech jevů v čase není omezen na uvedený případ daných intervalů. Data lze získat třemi různými způsoby, z nichž každý má svoje vlastní zpracovatelské modely. Zároveň s těmito typy si uvedeme jako příklad typ otázky z dotazníku, který je pro ně charakteristický. Předmětem stati je případ první.

Způsoby záznamu:

a) Zjišťujeme počet výskytů daného jevu v intervalech předem zadaných délek.
(Otázka z dotazníku: „Kolikrát jste byl v měsíci březnu v kině?"

b) Určujeme body na časové ose, kdy zkoumaný jev nastal.
(Pokyn v dotazníku: „Uveďte dny, v nichž jste navštívil kino!")

c) Zjištění délky intervalů mezi výskyty.
(Otázka z dotazníku: „Před kolika dny jste byl v kině?")

Modely zpracování dat se liší nejen podle jejich geneze a podstaty jejich vzniku, ale také podle způsobu záznamu (typu informace, která je k dispozici, resp. kterou lze dostatečně spolehlivě získat).

Situace sběru dat o výskytovosti jevů v čase je užitečné rozdělit podle počtu a podle procesů stability či konstantnosti podmínek:

A) *Sledování jednoho procesu v jednom nebo více intervalech.* Často dělíme interval sledování na kratší dílčí (většinou stejně dlouhé) intervaly, v nichž zjišťujeme výskyty.

Základní úloha: zjištění, zda je proces v celém intervalu rovnoměrný či zda se vyskytují v některých částech odchylky, zda výskyt postupně nesílí či nezeslabuje se apod.

Příklad: Sledování úrazovosti; čas dělíme na roky, měsíce, týdny apod.

B) *Sledujeme jeden proces opakovaně nebo jeden proces opakujeme.* Přitom sledujeme cíl odhalení vlivu různých podmínek, které se při realizacích buď přirozeně nebo experimentálně mění.

Základní úloha: odhalení existence a síly vlivu jednotlivých podmínek.

Příklad: Sledování průjezdu křižovatkou v hodinových intervalech, a to v různých dnech, částech dne a za různých metod řízení dopravy.

C) *Sledování paralelních procesů*: zjišťování výskytu jevů v souběžných procesech (různé systémy, jednotlivci v dotazníkovém šetření).

Základní úloha: Zjištění, zda sledované procesy odpovídají stejnému pravděpodobnostnímu modelu (stejnému principu geneze dat), zda mají stejné intenzity výskytů (zda jsou data statisticky homogenní mezi sebou).

Příklady: — sledování úrazovosti v mnoha podnicích souběžně
— sledování průjezdů u několika křižovatek současně
— zjišťování počtu návštěv kina za měsíc březen

Ve výše uvedených (a mnoha podobných) praktických situacích považujeme čas za spojitý a výskyt jevu za možný kdekoliv uvnitř zkoumaného intervalu.

Sledovaný proces může však být založen i na jiném parametru, než je plynoucí fyzický čas, a tento parametr nemusí být ani spojitý. Příkladem je mluvený projev nebo psaný text, u něhož sledujeme posloupnost slov, vět, výrazů apod. Zde místo časového intervalu máme úsek určité délky (délka měřená např. v počtu slov). V obsahové analýze (resp. v diagnostice textů) můžeme sledovat počet výskytů určitého slova, fráze, gramatického jevu, interpunkčního znaménka, obsahové zmínky, krátkých slov, cizích slov apod.

Obdobně můžeme takovýto proces sledovat tam, kde nás zajímá například počet automobilů, které zastaví na určitém místě (či v určité situaci) ze všech automobilů, které projíždějí (zde místo časového intervalu máme určitý počet projíždějících automobilů).

Velká rozmanitost úloh naznačuje užitečnost rozvoje adekvátní metodologie analýzy dat. Znalost těchto úloh a seznámení se s možnostmi jejich řešení i praktickými aspekty metod zpracování má velký význam pro přípravu projektu výzkumu,

může obohatit záměry výzkumníka při formulování úlohy, ovlivnit formulace otázek v dotazníku i přípravu záznamových instrumentů pro sběr dat či plán sběru informace.

2. Rovnoměrný výskyt jevů v čase: model Polssonova rozložení

Jevy vyskytující se v čase mohou vznikat podle nejrůznějších pravděpodobnostních mechanismů. Budeme se zabývat pouze jedním z modelů, nejzákladnějším a nejjednodušším modelem, tzv. *Poissonovým procesem* a *Poissonovým rozložením dat.*[1])

Jeho vznik lze charakterizovat třemi vlastnostmi:

1· *Proces je stacionární*: výskyt jevů v intervalu $(t, t + T)$ se řídí stejnými pravděpodobnostmi bez ohledu na to, kde je položen počátek t na časové ose. Pravděpodobnost počtu výskytů je závislá na délce intervalu T, ale ne na tom, od kdy začínáme počet výskytů zjišťovat.

2. *Proces je nezávislý na své historii*: výskyt jevů v intervalu $(t, t + T$ je zcela nezávislý na tom, co se odehrálo před dobou t. Pro společenskovědní aplikace svázané s tradičním dotazníkovým šetřením můžeme tento aspekt charakterizovat jako: máme-li dva nepřekrývající se časové intervaly procesu, pak výskyt jevu v jednom z nich je nezávislý na druhém. Pojem nezávislosti se věcně liší ve třech výzkumných situacích A, B, C, popsaných výše. Například v případě C to znamená, že paralelní procesy (např. u jednotlivých respondentů) se neovlivňují (to nebude splněno například u aktivit volného času zjišťovaných u členů domácnosti, kde se jevy buď sbíhají, například společné plánování akcí, nebo doplňují, resp. vylučují, například problém hlídání dětí).

3. *Proces je ordinární*: v intervalu velmi krátké délky (vzhledem k výzkumně relevantní časové jednotce) pak pravděpodobnost, že se jev v takto krátkém intervalu vyskytne jednou, je úměrná délce intervalu $h : P = \lambda h$ a pravděpodobnost, že se jev vyskytne vícekrát v tomto intervalu, je zanedbatelně malá vzhledem k P.

Konstanta úměrnosti λ se nazývá *intenzita procesu* a hraje v modelu velmi důležitou roli!

Popis tří aspektů je proveden z uživatelského hlediska, matematicky lze formalizovat základní východiska pro odvození vztahu (1) (viz dále) stručněji. Zájemcům o matematickou stránku věci lze doporučit knihu Fellerovu [1967].

Třetí teoretický předpoklad je pro praktické uvažování nepříjemný, neboť nelze standardně určit, co rozumíme „velmi krátkým časovým intervalem". V matematické formalizaci používáme k tomu infinitesimálního počtu (interval délky dt). Pro výzkumné úvahy si pod tím můžeme představit interval, jehož délka je prakticky vzhledem k použité časové jednotce a významu jevu v problému zanedbatelná. Budeme-li sledovat návštěvu kina, pak 1 den až 2 dny může být velmi krátkým intervalem vzhledem k měsíci: pravděpodobnost, že respondent navštíví kino během dvou dnů, může být považována za dvojnásobnou vzhledem k jednomu dni; zároveň pravděpodobnost, že respondent navštíví kino dvakrát za den, je nepatrná vzhledem k pravděpodobnosti jedné návštěvy. Budeme-li uvažovat den vzhledem k počtu televizních sportovních přenosů, pak je den jednotkou a není oním „velmi krátkým časovým intervalem" a podobně.

Uvedené předpoklady jsou zformulovány tak, aby umožnily sestrojení matematického modelu. Poznamenejme že pro praxi výzkumných úloh budou platit vždy jen přibližně, vždy půjde o abstrakci a tím o zjednodušení složitější reality.

---

[1]) Při formulaci modelu se budeme pokud možno vyhýbat jakýmkoli matematickým výrazům a odvozením. Zájemce bude odkázán na speciální literaturu. V dalších částech jsou uvedeny pouze vzorce nutné pro popis výpočtů v jednotlivých úlohách.

Prozatím jsme uvažovali také nepřerušený časový tok. V praxi to ovšem tak být nemusí. Sledujeme-li návštěvnost kina, pak pro daného respondenta uvažujeme pouze tu část časové osy, kterou má k dispozici pro volný čas (vynecháváme pracovní čas, spánek apod.). Proto ve skutečnosti je uvažovaný spojitý časový interval složen z dílčích intervalů fyzického plynulého času.

Proces můžeme tedy vázat pouze k času, který bychom mohli nazvat *realizačním* (či jinak podle konkrétního problému).[2])

Výše uvedené příklady charakterizují tzv. Poissonovo rozložení [Feller 1967 : 431], u něhož je pravděpodobnost počtu výskytů v intervalu zadané délky T dána vzorcem:

$$(1) \quad P_{POISS}(Y = k) = P(\text{jev se vyskytl k-krát}) = P_k = e^{-\lambda T} \cdot \frac{(\lambda T)^k}{k!},$$

kde $k = 0, 1, 2, \ldots$ je počet výskytů jevu v intervalech délky T,

$\lambda > 0$ je parametrem rozložení, který nazýváme *intenzitou*.

Toto rozložení má dvě základní charakteristiky:

(2) očekávaný počet výskytů $= EY = \lambda T$

(3) rozptyl počtu výskytů $= \text{var } Y = \lambda T$

Rozptyl je tedy číselně roven očekávané hodnotě.

Vidíme také lépe (a využitelněji) význam parametru λ a důvod jeho názvu: bez ohledu na délku T je to průměrný (očekávaný) počet výskytů daného jevu za časovou jednotku. Tato vlastnost je tedy vyjádřením *rovnoměrnosti* procesu. Z hlediska očekávání můžeme pomocí tohoto ukazatele porovnat libovolné dva časové intervaly procesu (i překrývající se).

Použití Poissonova rozložení je širší, než je možno vyčerpat v této stati. Používá se například při analýze četností tzv. *vzácných jevů*,[3]) tj. jevů, jejichž četnosti výskytu jsou velmi malé a neplatí pro ně zpracovatelské modely, běžně aplikované pro četnostní data. Na druhé straně vidíme, že předpoklady Poissonova modelu jsou dosti striktní a že nemůžeme jejich platnost očekávat jako samozřejmost u všech proměnných diskutovaného typu. Proto při většině aplikací můžeme očekávat pouze přibližné splnění podmínek modelu.

Tři typy souborů dat A, B a C (viz část 1) se od sebe poněkud liší vzhledem k úlohám formulovaným v konkrétních aplikačních situacích. Liší se také v pojetí poissonovské homogenity, která představuje jeden ze základních aplikačních pojmů. *Poissonovskou homogenitou* v případě jednoho procesu (typ A) rozumíme tu vlastnost, že intenzita všech dílčích intervalů je stejná a variabilita výskytů v intervalech odpovídá pravděpodobnostem (1). Tak ji také ověřujeme.

Jde tu tedy o *rovnoměrnost v čase* při zachování poissonovských vlastností.

V případě dat o více procesech, resp. o více podmínkách, v nichž proces funguje, je úloha formulována jako homogenita mezi procesy, tj. poissonovskou homoge-

---

[2]) Přitom čas t měříme v libovolných jednotkách, a proto si můžeme zvolit za jednotku délku toho úseku, který byl základem zjišťování (1 měsíc, 1 rok, 1 týden, 1 čtvrtrok, 2 týdny apod.). Intenzita je pak vztažena ke zvolené jednotce: λ = průměrný počet výskytů za danou jednotku. Poznamenejme, že volba jednotky času v otázce dotazníku nemusí být pochopitelně jednotkou, s níž sociolog pracuje. Tak „1 měsíc" může například znamenat „tu část z 1 měsíce, která je k dispozici pro danou činnost" apod.

[3]) Termín *„vzácné jevy"* pochází z matematické statistiky (rare events) a má sloužit k charakteristice výskytů s nepatrnými pravděpodobnostmi. V sociologii je prozatím takováto aplikace spíše výjimkou. Široké uplatnění má však model zejména v lékařském výzkumu při zkoumání výskytu málo četných diagnóz, v sociálním lékařství apod.

nitou tu rozumíme *stejné intenzity u jednotlivých procesů* při platnosti předpokladů modelu.

Statistické zkoumání dat provádíme v obou případech stejně. V obou případech pro aplikaci statistických technik také musí být splněna podmínka *nezávislosti* jednotlivých údajů. Při porovnávání nesmí být jednotlivé procesy statisticky závislé, výsledek jednoho procesu nesmí ovlivňovat výsledek jiných procesů.[4]) Pro vzájemně závislé procesy musíme volit jiné techniky.

Uveďme ještě několik zcela praktických poznámek:

1. Časovou jednotku můžeme volit libovolně, a proto bude ve většině situací vhodné volit $T = 1$, abychom v modelu (1) eliminovali jeden parametr.

2. Analýza pomocí Poissonova modelu se značně komplikuje v případě, že nemáme k dispozici úplnou informaci o realizovaných četnostech. Jde na příklad o kategorizovaná data, o data, u nichž nemůžeme zjišťovat četnosti vyšší než určitý prah, data, u nichž slučujeme všechny vysoké četnosti do jednoho kódu apod. Proto se snažíme udržet původní záznamy jako počet výskytů a pracujeme s pokud možno podrobným rozložením četností. Kategorizace provádíme podle potřeby až v průběhu vlastní analýzy. I když je analýza při uvedených komplikacích v zásadě možná, vyžaduje speciální přístupy a náročné výpočty.

3. Při záznamu dat je nepřípustné „napravovat" chyby tím způsobem, že zapomeneme-li zaznamenat výskyt do příslušného intervalu, přiřadíme jej k následujícímu. Tato častá záznamová chyba vede sice ke stejnému průměru, ale porušuje informaci o variabilitě procesu v čase, a tudíž narušuje možnost ověření poissonovských vlastností.

V dalším textu bude diskutován případ paralelního zjišťování údajů o procesech charakterizujících jednotlivé respondenty v sociologickém šetření. Z hlediska zajištění všech předpokladů modelu a metody je to pravděpodobně nejkomplikovanější případ. Paralelně lze formulovat všechny úlohy i pro jiné typy procesů a výzkumných situací — převod je snadný. Též analogie pro typy A a B je zřejmá.

3. Sociologické úlohy spojené s daty subjektivních výpovědí respondentů

Předpoklady modelu jsou dosti silné a v praxi je jejich platnost silně omezena. Tyto předpoklady mají však důležitou aplikabilitu a závažné metodologické důsledky. Výzkum je totiž nutno připravit tak, aby do analýzy dat nevstupovaly metodologické šumy a tak místo meritorních interpretací nevznikaly artefakty způsobené metodologickými nedostatky (viz též poznámka 3).

---

[4]) Poučná chyba se mohla přihodit v jednom z výzkumů autorovi článku: Otázka „Kolikrát za poslední měsíc jste navštívil koncert vážné hudby? (uvažujte vždy poslední kalendářní měsíc, vyplňujete-li v dubnu, odpověď se týká měsíce března, apod.)". Zde mohly nastat hned dvě chyby, které by zkreslily analýzu: 1) měsíce jsou různě dlouhé (únor 28 dní, březen 31 dní), což je rozdíl délky sledovaného intervalu o 10 %; 2)z organizačních důvodů pokryl sběr dat období od února až do června — toto časové rozpětí nemělo vliv na velikou většinu šetřených proměnných, ale uvedená otázka tím byla zcela zkreslena, neboť sezónnost produkce vážné hudby (Pražské jaro apod.) tu mají nesmírně vysoký vliv. Soubor dat by tedy reprezentoval mnoho meritorně zcela nesrovnatelných procesů.

Proto u otázek podobného typu zjišťujeme i datum realizace dotazu. V daném případě byly zpracovávány příslušné podsoubory zvlášť. Znalost modelu zpracování a důsledky, které z něho vyplývají, tedy ovlivňují už i záznam dat, resp. i zjišťování dodatečné informace.

Obdobné chyby by bylo možno se dopustit například při dotazu na návštěvu sportovních podniků, ale i kin, neboť role návštěvy kina v létě je jiná než v zimě a nemůžeme proto předpokládat stejnou intenzitu. Tento aspekt vystupuje velmi silně při *komparaci výzkumů* či při *opakování výzkumu*. Rozdíly intenzity ve dvou výzkumech nemusí mít jen význam vývoje v čase, ale i význam vzhledem k umístění šetření v rámci ročního sezónního cyklu. Tyto chyba je velmi nebezpečná. Při plánování srovnávacích výzkumů je nutno s tím předem počítat. Jinak jsou data nesrovnatelná a šetření znehodnoceno.

Je zřejmé, že *volba časového úseku* je důležitá jednak z metodologických aspektů, jednak z důvodů podstaty problému. V případě otázek v dotazníku musíme respektovat tyto aspekty:

a) schopnost respondenta správně určit počet výskytů;
b) úsek musí splňovat požadavek rovnoměrnosti, nesmí v něm nastat zlom v intenzitě výskytů;
c) ve zkoumaném úseku musí jít o ekvivalentní roli pro všechny respondenty; pro různé respondenty může totiž tentýž časový úsek mít různé významy vzhledem k realizovatelnosti jevu: jsou to například dovolená, zkouškové období pro studenty, sezónní dostupnost, práce na směny pro večerní aktivity, mateřská dovolená, nemoc a podobně;
d) určení jevu samého vzhledem k časovému úseku: večerní sledování televize redukuje „měsíc" o třetinu u osob se střídavými pracovními směnami; stejná intenzita se tu projeví v daleko nižší sledovatelnosti

Všechny takovéto aspekty spolupůsobí na obsah pojmu „homogenní" a „rovnoměrný" v konkrétních výzkumných podmínkách. Pro každou sociologickou aplikaci musíme proto provést dobrou diskusi okolností a všech aspektů modelu. [5], [6])

Základní analytické úlohy spojené se souborem dat

A. Homogenita zkoumané skupiny jednotlivců

Poissonovská homogenita[7]) byla definována jako stav souboru, v němž všechny údaje vznikly na základě Poissonovských předpokladů a se stejnou intenzitou (pro daný časový úsek). I zde musíme počítat s vlivem náhodných chyb, nepřesností měření i záznamu, kódování, výpovědí osob. Vzhledem k těmto okolnostem i vzhledem k velmi rozrůzněné sociální realitě je potvrzení homogenity dosti vzácné. Sociální skupiny jsou vnitřně dále diferencovány nejrůznějšími faktory sociálními, biologickými, zájmovými i kulturními a je proto velmi obtížné najít soubory, pro něž by sociolog mohl vyslovit byť i hypotézu homogenity. O to důležitější je nalézt i menší skupiny, které už homogenní budou. Skupina vnitřně homogenní se stejnou intenzitou (je-li navíc odlišitelná od jiných skupin) se tak stává *typem*, jehož empi-

[5]) Podobný rozbor lze aplikovat i na zdánlivě odlehlý případ obecnější otázky: „Jak často chodíte do kina?", u níž jsou dány kategorizované odpovědi. Taková obecná otázka má smysl *pouze tehdy*, pokud se váže k časovému úseku, na němž je proces realizace dané aktivity respondentem rovnoměrný (ve smyslu stejné intenzity). Pokud se λ mění pro danou osobu, pak respondent neví, jak na ni odpovědět, nebo *výzkumník neví, ke které části procesu se respondentova odpověď váže*. Diskuse modelu má tedy i jednu velmi závažnou kvalitativní metodologickou konsekvenci: otázka výše uvedeného typu (i když často používaná), bývá chybně formulována a mívá smysl pouze pro odlišení kategorie „nikdy".

[6]) Zde je opět vidět dosti názorně, jak důležité je znát zpracovatelské techniky předem. Statistická metodologie je těsně spojena s kvalitativními rozhodnutími metodologie sběru dat, přípravy dotazníku. Formalizace, i když velice jednoduchá a používaná vlastně jen jako přepis heuristických základů, nám umožňuje uspořádat různé faktory vlivu na zpracování dat systematicky a tak se předem vyvarovat různých interpretačních chyb. Formalizace sama o sobě nemá pro sociologii význam, avšak i v tomto velmi jednoduchém případě ukazuje svoji roli a svoji užitečnost — její opravdová síla spočívá v sepětí s úvahami metodologickými. Proto není šťastné, oddělujeme-li metodologii „kvantitativní" a „kvalitativní". Matematizace sama o sobě, samoúčelně zaváděná, je prázdný efekt. Nepoužívání, potlačení, či dokonce vymýcení matematiky pak vede k oslabení vlastní metodologické možnosti a v důsledku toho k větší pravděpodobnosti výzkumných chyb, a tedy k nepřesným či chybným závěrům, interpretacím dat.

[7]) Nutno znovu zdůraznit, že zde mluvíme o specifickém typu homogenní skupiny. Budeme-li podobné úlohy formulovat vzhledem k jiným typům proměnných, budeme muset i pojem homogenity operacionálně vymezit jinak.

rickým projevem je právě intenzita procesu realizace daného jevu, rovnoměrnost a nezávislost v časových realizacích.

Ověření platnosti hypotézy homogenity pomocí metod matematické statistiky spočívá v tom, že zjistíme, zda rozložení dat v daném souboru má tvar Poissonova rozložení, to jest, zda až na náhodné a přípustné výjimky odpovídají empirické relativní četnosti vztahu (1) s nějakým, empiricky odvozeným nebo předem daným parametrem λ.

B. Odchýlení od homogenity

Není-li zkoumaná skupina homogenní vzhledem ke stejné intenzitě, tj. jestliže nenalezneme shodu mezi daty a modelem Poissonova rozložení, jistě nás bude zajímat, jakým způsobem se empirické četnosti liší od očekávaných.

Bude nás zajímat, zda můžeme odhalit směr odklonu od homogenity, tj. od předpokládaného modelu — je to otázka kvalitativních vlastností zkoumané populace vzhledem k jevu: Čím je heterogenita způsobena? Jaký proces vede ke zjištěné odchylce? Na takové otázky statistika nemůže přímo odpovědět, nicméně některé typy odchylek od Poissonova rozložení (1) jsou dobře interpretovatelné a jejich identifikace může podstatně pomoci k sociologické interpretaci zjištěných rozložení dat.

Některé modely alternativních situací jsou i matematicky zvládnutelné. Uvádíme ty situace, které jsou nejčastější a zároveň interpretačně nejzávažnější.

a) Proces v čase nemá ani u jedinců či u podskupin poissonovské vlastnosti, tj. rovnoměrnost, resp. nezávislost realizace jevu na předchozím průběhu. V takovém případě musíme hledat jiné modely vhodně popisující vznik dat (a matematická statistika jich má velké množství).

b) Část populace má tu vlastnost, že se u ní jev nevyskytuje nebo ani nemůže vyskytnout (tj. u části populace je pravděpodobnost výskytu nula). To znamená, že četnost údajů odpovídající nulovému výskytu je *směsí dvojí kvality*: ti, u nichž se jev nerealizuje nikdy, a ti, u nichž se nerealizuje pouze v daném zkoumaném intervalu. Tento případ je zřejmě dosti častý: naprostý nezájem o sport, o vážnou hudbu, nemožnost navštívit divadlo, část respondentů nemá televizor, a proto nesleduje dramatické pořady.

c) Existence dvou nebo více vnitřně poissonovsky homogenních částí souboru, z nichž však každá má jinou intenzitu.

d) Jednotlivé údaje se řídí Poissonovým rozložením, ale každý údaj má jinou intenzitu a tyto intenzity mají svoje vlastní rozložení v populaci nebo její části (které nás ovšem velice zajímá).

Jednotlivé případy se mohou různě kombinovat. Existují ještě další možné alternativní modely, uvedené však mají zcela jasnou sociologickou relevanci. Ověřování těchto modelů se provádí po odmítnutí shody s modelem (1). Abychom měli alespoň nějakou představu o alternativě, je možno využít *znaménkového schématu*, jehož výpočet bude uveden.

C. Odhad parametru

Jestliže je ověřena platnost modelu (1), zajímá nás pochopitelně, jaká je společná intenzita pro danou skupinu. Chceme odhadnout (bodově i intervalově) parametr λ pro reprezentovanou populaci. Společná intenzita je totiž agregovanou charakteristikou populace a může sloužit nejen jako základ pro její odlišení od jiných populací, ale také pro extrapolaci očekávaného jevu na různě dlouhé časové intervaly.

Jestliže se ukáže, že v některých souborech dat můžeme přijmout hypotézu jejich vnitřní homogenity, může nás zajímat, zda některé soubory nezůstanou homogenní i po sloučení, tj. mají-li *společnou intenzitu*. Jestliže ano, pak nás pochopitelně zajímá i jejich společná intenzita. Jde o testování hypotézy rovnosti λ_i v jednotlivých skupinách a při přijetí hypotézy o rovnosti o výpočet společného λ ze spojeného souboru dat. V případě zjištění rozdílnosti intenzit nás pak mohou zajímat přesnější komparativní metody, přinášející další sociologickou informaci, které však už v této stati neuvádíme.

4. Metody řešení úloh a příklady z výzkumné praxe

V této části postupně ukážeme metody řešení úloh A—D statistickými prostředky a předvedeme si je na příkladech z praxe analýzy dat. Dále budeme značit $n_i =$ = počet pozorování pro hodnotu $X = i$, $N = \Sigma n_i =$ součet validních dat, $f_i =$ = N_i/N jsou relativní četnosti výskytu.

A. Test shody pro Poissonovo rozložení

Úloha zní takto: je zjištěno rozložení četností (n_i) pro hodnoty $i = 0, 1, 2, \ldots$ atd. Odpovídá toto rozložení Poissonovu rozložení? Uvedeme dvě možná řešení problému. První řešení má za účel rychlou orientaci, je užitečné, i když může vést k nepřesnému závěru. Druhé řešení je založeno na testu chí-kvadrát.

Poissonovo rozložení má tu pozoruhodnou vlastnost, že jeho průměr je číselně stejný jako rozptyl (viz vzorec (2), (3)). Proto můžeme svoji rychlou orientaci založit na srovnání empirického průměru a rozptylu. Bohužel v případě rovnosti nemáme plně zajištěno, že jde o Poissonovo rozložení, typ rozložení může být i jiný (a tím platí i jiné vysvětlení vzniku dat). Test popíšeme:

A1. Fisherův index disperze

1. *Formulace hypotézy*:

$H_0 : \mathrm{E}X = \mathrm{var}\, X$
$H_A : \mathrm{E}X \neq \mathrm{var}\, X$, $H_{A'} : \mathrm{E}X < \mathrm{var}\, X$, $H_{A''} : \mathrm{E}X > \mathrm{var}\, X$

2. *Data*

rozložení četností $n_0, n_1, n_2, \ldots, n_i, \ldots$
$n_i =$ počet výskytů i jevů
$N = \Sigma n_i =$ velikost souboru
$R =$ maximální pozorovaná hodnota

3. *Intuitivní model*: jestliže H_0 platí, pak poměr empirické variance a empirického průměru bude přibližně roven 1. Odchylka od 1 znamená příklon k některé alternativní hypotéze.

4. *Testové kritérium* — postup výpočtu:

a) spočítáme Fisherův index disperse:

$$D = \frac{\sum_{i=0}^{R} n_i (i - \overline{X})^2}{\overline{X}} \tag{4}$$

$$= \frac{\Sigma n_i i^2 - \frac{(\Sigma n_i i)^2}{N}}{\overline{X}}$$

$$= \frac{N\Sigma n_i i^2 - (\Sigma n_i i)^2}{\Sigma n_i i},$$

kde $\bar{X} = \frac{1}{N} \sum_{i=0}^{R} n_i i$

b) D je rozloženo jako X o $(N - 1)$ stupních volnosti. To znamená, že pro malá N použijeme porovnání s distribucí chí-kvadrát a pro větší N použijeme normální aproximaci chí-kvadrátu.

α) Pro $N \leqq 30$ použijeme tohoto postupu H_0 *vs* H_A: pro předem danou významnost α s zjistíme horní a dolní kritickou hodnotu na hladině $\alpha/2$ odvozenou z rozložení chí-kvadrátu o $(N - 1)$ stupních volnosti. Označíme je D_d a D_h.

Je-li $D_d \leqq D \leqq D_h$ pak nemáme důvod zamítnout H_0.
Je-li $D \geqq D_d$ nebo $D \leqq D_d$ pak odmítáme H_0 ve prospěch H_A.

Kritické hodnoty nalezneme například v tabulkách Janko[8]) [1958 : 118, tabulka 5].
H_0 *vs* $H_{A'}$: pro danou významnost α nalezneme horní kritickou hodnotu pro rozložení chí-kvadrát o $(N - 1)$ stupních volnosti. Označíme ji D_H.[9])
Je-li $D < D_H$ pak nemáme důvod zamítnout H_0.
Je-li $D \geqq D_H$ pak odmítáme H_0 ve prospěch $H_{A'}$.
H_0 *vs* $H_{A''}$: obdobně nalezneme dolní kritickou hodnotu D_D na hladině α.
Je-li $D \geqq D_D$ pak nemáme důvod zamítnout H_0.
Je-li $D > D_D$ pak odmítáme H_0 ve prospěch H_A.

Tento postup budeme používat i pro $N > 30$, pokud máme k dispozici přesné tabulky pro chí-kvadrát distribuci s příslušným počtem stupňů volnosti.

β) Pro $N > 30$ (resp. pro N vybočující z rámce dostupných tabulek) použijeme Wilson-Hilfertovy aproximace normálním rozložením (viz Johnson-Kotz [1970]).
Spočítáme:

$$Z = \left[\left(\frac{D}{N-1}\right)^{\frac{1}{3}} - 1 + \frac{2}{9(N-1)}\right] \Big/ \sqrt{\frac{9(N-1)}{2}}$$

Dále porovnáme Z s kritickou hodnotou normálního rozložení. V tabulkách[10]) normálního rozložení nalezneme hodnoty Z_d, Z_h, Z_D, Z_H k danému α a testy provedeme zcela obdobně jako v předchozím případě.

Hypotézy H_A, $H_{A'}$, $H_{A''}$ musíme volit před samotným testováním, resp. před pohledem na data. $H_{A'}$ znamená, že očekáváme jako alternativu nějakou další rozptýlenost navíc kromě té, kterou povoluje náhodné chování našich dat za Poissonova modelu. Tento případ je ovšem častý a nejvíce nás zajímá. $H_{A''}$ spočívá v tom, že zde dále působí nějaký nenáhodný faktor, který snižuje náhodné rozptýlení — je to případ, při němž proti přirozené očekávané homogenitě modelu působí ještě dodatkový homogenizační faktor. Tím může být například restrinkce počtu výskytů jevu[11]) v určitých mezích. H_A pak je obecná alternativa. Používáme ji v případě,

[8]) Tabulka je uzpůsobena tak, že pro hladinu α volíme sloupce označené $\alpha/2 \,.\, 100$ a $(1 - \alpha/2)\, 100$, v nichž nalezneme D_h a D_d; řádek odpovídá hodnotě $(N - 1)$. Při programování postupu nám počítač nalezne hodnotu významnosti přímo.

[9]) V tabulkách Janko [1958 : 118, tab. 5] nalezneme D_H ve sloupci označeném 100α, v řádku označeném $(N - 1)$. D_D (viz test H_0 *vs* $H_{A''}$) nalezneme ve sloupci označeném $100(1 - \alpha)$.

[10]) Viz např. Janko [1958 : 113—114, tab. 2], nejobvyklejší hodnoty viz také Řehák, Řeháková [1978 : 87, tab. C; 82, tab. A]. Poznamenejme, že stačí vyhledat Z_h a Z_H, které jsou obvykle tabelovány, neboť ostatní hodnoty získáme pouhou změnou znaménka $Z_d = -Z_h$, $Z_D = -Z_H$. Připomeňme zde také, že uvedená aproximace se používá při ručních výpočtech. Při programování na samočinné počítače používáme přesnější vzorce.

[11]) Omezení počtu výskytů zkoumaného jevu se v praxi projeví téměř vždy. Například počet návštěv opery je omezen počtem představení za měsíc. Takové omezení je však pro málo četný jev pouze teoretické a v datech se neprojeví. $H_{A''}$ se prokáže především tehdy, když dané omezení působí jako součást procesu vzniku dat nebo záznamu dat, to jest, pokud se projeví přímý vliv na snížení hodnoty variance.

že u dat nemůžeme předem stanovit, zda očekáváme variabilitu navíc či variabilitu nižší.

Kromě statistického postupu Fisherova indexu disperze můžeme použít poměr rozptylu a průměru, tj. $D/(N-1)$ pro rychlou optickou kontrolu předpokládané rovnosti.

A2. Chí-kvadrát test dobré shody pro Poissonovo rozložení

Test chí-kvadrát je založen na porovnání shody celého rozložení, nikoli tedy jen na vybraných statistikách.

Postup:

1. *Formulace hypotézy*:

H_0: rozložení četností odpovídá Poissonovu modelu (jehož intenzita není známa)
H_A: rozložení četností se významně liší od Poissonova modelu.

2. *Data*:

rozdělení četností: $n_0, n_1, n_2, \ldots$; $N = \sum_{i=0}^{R} n_i$,
R = maximální pozorovaná hodnota;
předpokládáme $N \geqq 30$.

3. *Intuitivní model*:

Rozdělení relativních četností f_i/N $(i = 0, 1, 2, \ldots, R)$ by se nemělo příliš lišit od teoretického rozložení (1), v němž za λ dosadíme nejlepší empirický odhad, který nám statistická teorie poskytuje. Teoretické pravděpodobnosti P_{R+1}, P_{R+2}, by měly být už zanedbatelně malé.

4. *Testové kritérium* — postup výpočtu:

Postup výpočtů uvádíme paralelně ve dvou verzích: výpočty odhadnutých relativních četností P_k a výpočty očekávaných četností E_k, mezi nimiž platí ovšem jednoduchý vztah: $E_k = N \,.\, P_k$.

a) spočteme odhad $\hat{\lambda}$ pro λ:

$$\hat{\lambda} = \frac{1}{N} \sum_{i=0}^{R} i \,.\, n_i \tag{5}$$

b) spočteme hodnoty $\hat{P}_k$, resp. E_k, $k = 0, 1, 2, \ldots$

$$\hat{P}_k = \mathrm{e}^{-\hat{\lambda}} \,.\, \frac{(\hat{\lambda})^k}{k!}, \qquad E_k = N \,.\, \mathrm{e}^{-\hat{\lambda}} \frac{(\hat{\lambda})^k}{k!} \tag{6}$$

$k = 0, 1, 2, \ldots$

Postupujeme tak, že určíme[12])

$$\hat{P}_0 = \mathrm{e}^{-\hat{\lambda}}, \quad \text{nebo} \quad E_0 = N \,.\, \mathrm{e}^{-\hat{\lambda}} \tag{7}$$

[12]) Výpočet na kapesním kalkulátoru je velmi jednoduchý. Výraz (7) zjistíme buď z tabulky exponenciální funkce, nebo přímo na kalkulátoru a dále už každý další výraz vzniká násobením konstantou λ a dělením hodnotou k (8). V praxi je jednodušší počítat přímo očekávané četnosti E_k.

a dále postupně

$$\hat{P}_k = \hat{P}_{k-1} \cdot \frac{\hat{\lambda}}{k}, \quad \text{resp.} \quad E_k = E_{k-1} \frac{\hat{\lambda}}{k} \tag{8}$$

Postupujeme tak dlouho, pokud[13])

$$N - \sum_{k=0}^{K} E_k > m, \tag{9}$$

kde m je zadáno takto:

$$\begin{array}{llll} m = 5 & \text{pro} & K = 3 & \\ m = 3 & \text{pro} & K = 4 & \\ m = 2 & \text{pro} & K = 5 & \text{a} \quad K = 6 \\ m = 1 & \text{pro} & K \geqq 7 & \end{array}$$

Přitom chceme, aby bylo K co největší

Dále volíme

$$P_{K+1} = 1 - \sum_{k=0}^{K} P_K, \quad \text{resp.} \quad E_{K+1} = N - \sum_{k=0}^{K} E_k \tag{10}$$

c) K tomuto K spočteme

$$n_{K+1} = N - \sum_{k=0}^{K} n_i \tag{11}$$

d) je-li $K + 1 \geqq 2$, můžeme sestrojit tabulku A

a z ní spočítáme kritérium chí-kvadrát, které má počet stupňů volnosti $df = K$

$$X^2 = \sum_{i=0}^{K+1} \frac{(n_i - E_i)^2}{E_i} \tag{12}$$

$$= \sum_{i=0}^{K+1} \frac{n_i^2}{E_i} - N \tag{13}$$

e) Hodnotu X^2 porovnáme s tabulkovou kritickou hodnotou na zvolené hladině α [Janko 1958 : 118, tab. 5].

Pro většinu praktických případů postačí také údaje v článku Řehák, Řeháková [1978 : 93, tab. E].

Je-li $X^2 <$ nalezená kritická hodnota, nemáme důvod odmítnout hypotézu shody.

Je-li $X^2 \geqq$ nalezená kritická hodnota, hypotéza shody (a tedy příslušnost empirického rozložení k Poissonovským rozložením) zamítáme.

Příklad 1

Data jsou vybrána z výzkumu čtenářů časopisu *Mladý svět*. Data tabulky 1 odpovídají počtu navštívených koncertů populární hudby zjištěnému u části kontrolního souboru studentů středních škol. Zkoumaná část vznikla identifikací určitého typu čtenářského zájmu.[14])

---

[13]) Je tu voleno jednoduché pravidlo, které je odvozeno z různých simulačních prací [Řeháková 1979 : 618]. Konečné slovo ani jednoznačné pravidlo není zatím vyřčeno a jednotliví autoři se zatím neshodli. Zformulované pravidlo je platné jen do té doby, dokud nebude dalšími extenzívními simulačními šetřeními upřesněno či vyvráceno. Hranice pro daný speciální případ jsou voleny spíše konzervativně, aby aproximace platila i u vyšších hladin významnosti.

[14]) Šlo o skupinu respondentů, kteří byli určeni vysokou hodnotou faktoru „zájem o zábavné části časopisu", a to bez ohledu na to, jaké jiné čtenářské zájmy a postoje se u nich vyskytly.

Tabulka A. Tabulka empirických a očekávaných četností

Výskyt jevu	0krát	1krát	2krát	...	$(K+1)$krát	Součet
Empirické četnosti	n_0	n_1	n_2	...	n_{K+1}	N
Očekávané četnosti	E_0	E_1	E_2	...	E_{K+1}	N

a) Postup numerických výpočtů pro test rovnosti průměru a rozptylu pomocí *Fisherova indexu disperse.* Volíme předem hladinu významnosti $\alpha = 0\ 05$, zajímá nás dvoustranná alternativa: H_0 *versus* H_A.
Spočteme Fisherův index disperze a podíl variance a průměru:[15])

$$N = 184 \quad \Sigma n_i i = 237 \quad \Sigma n_i i^2 = 575$$

$$(N-1)\ \mathrm{var}\ X = \Sigma n_i i^2 - \frac{1}{N}(\Sigma n_i i)^2 = 269.73$$

$$\bar{X} = 1.29 \quad \mathrm{var}\ X = 1.47 \quad \frac{\mathrm{var}\ X}{\bar{X}} = 1.14$$

$$D = 209.41\ ; \quad \text{počet stupňů volnosti} = 183$$

Použijeme transformace (5):

$$Z = \left[(1.14)^{\frac{1}{3}} - 1 + \frac{2}{9(184-1)}\right] \cdot \sqrt{\frac{9(184-1)}{2}}$$

$$= [1.0460 - 1 + .0012] \times 28.6967$$
$$= 1.3539$$

Tabulka 1. Rozložení počtu návštěv koncertů populární hudby v měsíci březnu

Počet návštěv	0	1	2	3	4	5	6	7	Celkem
Studenti (zájmová skupina)	54	62	44	14	7	2	—	1	184

Při testování použijeme kritické hladiny normálního rozložení, která je pro náš případ ($\alpha = 0.05$, dvojstranná alternativa) rovna $Z_d = -1.960$, $Z_h = 1.960$. Protože skóre Z leží v přípustných hranicích, nemáme důvod k zamítnutí hypotézy o rovnosti.[16])
Podíl variance a průměru, který byl 1.14, se ukázal být nevýznamně odlišný od jedné. Nečekáme tedy žádnou heterogenitu navíc proti očekávané ani faktory, které by nám omezovaly výskyt jevu v dané skupině studentů. Výsledek testu nám však bohužel nic neříká o tom, zda princip realizace dat (tj. výskytu jevu) opravdu odpovídá poissonovskému principu.

b) Postup numerických výpočtů pro test dobré shody se provádí postupně podle vzorců (5) až (13).
Nejprve odhadneme $\hat{\lambda}$ a očekávané četnosti E_k:

$$\hat{\lambda} = \frac{1}{N}\Sigma n_i i = \frac{237}{184} = 1.29 \quad \text{(vzorec (5))}$$

$$E_0 = N \cdot e^{-\hat{\lambda}} = 184 \times 0.2758 = 50.75 \quad \text{(vzorec (7))}$$

[15]) Všechny výpočty uvádíme s omezeným počtem míst, výpočet však probíhal s maximální přesností, s níž mohl minikalkulátor pracovat. Proto například nepočítáme s $\bar{X} = 1.29$, ale $\bar{X} = 237/184$, kteroužto hodnotu uložíme v paměti s maximální přesností.

[16]) Ekvivalentně je možno použít ještě jiný způsob, nemáme-li k dispozici kritickou hranici normálního rozložení, ale máme k dispozici tabulky kritických hodnot chí-kvadrátu. V takovém případě používáme Z^2 a porovnáváme s kritickou hodnotou chí-kvadrátu s jedním stupněm volnosti, v našem případě $Z^2 = (1.3539)^2 = 1.8330$ porovnáváme s kritickou hodnotou: $3.84 = (1.96)^2$. (Výsledek je pochopitelně tentýž.)

Dále postupně podle vzorce (8):

$$E_1 = 50.75 \times \frac{1.29}{1} = 65.37$$

$$E_2 = 65.37 \times \frac{1.29}{3} = 42.10$$

$$E_3 = 42.10 \times \frac{1.29}{3} = 18.07$$

$$E_4 = 18.07 \times \frac{1.29}{4} = 5.82$$

$$E_5 = 5.82 \times \frac{1.29}{5} = 1.50$$

Vzhledem k tomu, že očekávané četnosti rychle klesají, budou další četnosti už malé, a proto zkontrolujeme, zda zbytek intervalů nebudeme slučovat. To se provede tak, že postupně zjistíme

$$\begin{aligned}
N - E_0 &= 184 - 50.75 = 133.25 \\
N - E_0 - E_1 &= 133.25 - 65.37 = 67.88 \\
N - E_0 - E_1 - E_2 &= 67.85 - 42.10 = 25.78 \\
N - E_0 - E_1 - E_2 - E_3 &= 25.78 - 18.07 = 7.71 \\
N - E_0 - E_1 - E_2 - E_3 - E_4 &= 7.71 - 5.82 = 1.89
\end{aligned}$$

Tento výpočet ukazuje, že na intervaly 1 a výše připadá očekávaná četnost 133.25, na intervaly 2 a výše připadá 67.88 atd. Vidíme, že na intervaly 5 a výše připadá 1.89, a tedy musíme připojit ještě předchozí interval, abychom splnili podmínku, že očekávaná četnost má být větší než $m = 5$ pro $K = 3$.

Můžeme tedy pro test chí-kvadrát sestrojit tabulku:

Tabulka 2. Empirické a očekávané četnosti pro Poissonovo rozdělení z tabulky 1

Výskyt jevu	0	1	2	3	4 a více	Celkem
Empirické	54	62	44	14	10	184
Očekávané	50.75	65.37	42.10	18.07	7.71	184

Testové kritérium X^2 spočteme podle vzorce (13):

$$X^2 = \frac{54^2}{50.75} + \frac{62^2}{65.37} + \frac{44^2}{42.10} + \frac{14^2}{18.07} + \frac{10^2}{7.71} - 184$$

$$= 2.06$$

Porovnáním s kritickou hodnotou ($df = 3$, $\alpha = 0.05$), která je rovna 7.81, vidíme, že není žádný důvod k odmítnutí shody empirického rozložení s Poissonovým rozložením.

Závěr tohoto statistického testu vede k důležitému poznatku. Vzhledem k dané činnosti je definovaná skupina respondentů homogenní. Má společnou intenzitu, patří z hlediska analyzovaného znaku ke stejnému typu (v relaci k definici skupiny pak ovšem můžeme provést další hlubší interpretace; výsledek je sociologicky pouze jednou součástí celého souhrnu poznatků, jejichž syntéza vyjeví nové výsledky).

B. Znaménkové schéma a testy pro odchylku četnosti v poli

V případě odmítnutí hypotézy homogenity, tj. hypotézy Poissonova rozložení četností, je důležité formulovat hypotézy alternativní, tj. hypotézy, které by nám pomohly interpretovat získaná data. K tomu lze použít v podstatě dvou přístupů:

a) formulovat jiné modely a pokusit se je ověřit nebo mezi nimi diskriminovat.[17])

[17]) K tomu účelu se používá ponejvíce negativně binomického rozložení, Poissonova rozložení s nulami, Fisherových nebo Neymanových epidemických rozložení, beta-binomických rozložení, rozložení logaritmické řady, nebo useknutých rozložení, popřípadě směsových rozložení. Každé z nich má své zvláštní vlastnosti odvozené z modelových základů a z genezo určující vznik měřených náhodných veličin.

b) zjistit znaménkové schéma, které by pomohlo identifikovat platný model.

Zde ukážeme pouze druhý přístup, který však ve speciálním případě může řešit i úkoly prvního typu.

Odvození znaménkového schématu znamená nalezení standardizovaných Z-skóre pro každé pole tabulky. Vysoká hodnota Z-skóre (u něhož předpokládáme normální standardizované rozložení) indikuje, že v daném poli se projevuje významná odchylka.

Výpočet znaménkového schématu provedeme tak, že určíme standardizované reziduální odchylky v jednotlivých polích a podle nich se orientujeme při upřesňování hypotézy. Reziduální standardizovanou odchylku určíme jako:

$$Z_j = \frac{n_j - E_j}{\sqrt{E_j \cdot V_j}}, \tag{14}$$

kde $$V_0 = 1 - \hat{P}_0 - \hat{P}_0 \frac{1}{i(\lambda)}$$

$$= 1 - \hat{P}_0 \left(1 + \frac{1}{i(\lambda)} \right)$$

$$V_j = 1 - \hat{P}_j - \left[\frac{\hat{P}_{j-1} - \hat{P}_j}{\hat{P}_j} \right]^2 \hat{P}_j \cdot \frac{1}{i(\lambda)} \tag{15}$$

$$= 1 - \hat{P}_j \left[1 + \left(\frac{j}{\bar{X}} - 1 \right)^2 \frac{1}{i(\lambda)} \right], \quad j = 1, 2, \ldots, K$$

$$V_{K+1} = 1 - \hat{P}_{K+1} - \frac{\hat{P}_K^2}{\hat{P}_{K+1}} \cdot \frac{1}{i(\lambda)}$$

$$= 1 - \hat{P}_{K+1} \left[1 + \frac{\hat{P}_K^2}{\hat{P}_{K+1}^2} \cdot \frac{1}{i(\lambda)} \right]$$

Fisherovu informační míru $i(\lambda)$ můžeme spočítat dvěma způsoby. Pro ruční výpočty postačí jednoduchý způsob I, při programování použijeme způsob II[18])

Způsob I:

$$i(\lambda) = \frac{1}{\bar{X}} \tag{16}$$

---

[18]) Oba způsoby se liší odvozením:

Způsob I:

1) odvodíme odhad parametru z původních četností n_i pro každou hodnotu veličiny; 2) odvozdíme tzv. Fisherovu informaci pro jedno pozorování; 3) sloučíme málo obsazené hodnoty od $(K + 1)$-ního počínaje tak, aby očekávaná četnost byla alespoň 5; 4) odvodíme znaménkové schéma.

Způsob II:

1) obdobně; 2) obdobně 3. kroku; 3) odvodíme Fisherovu informaci pro tento nově kategorizovaný znak; 4) odvodíme znaménkové schéma.

Oba způsoby se tedy liší pouze výpočtem Fisherovy informace. Numericky nedostáváme příliš rozdílné výsledky, způsob II odpovídá lépe modelu vzniku a zpracování dat, je však výpočetně náročnější.

Způsob II:

$$\text{označíme} \cdot \hat{R}_k = \sum_{j=0}^{k} \hat{P}_j$$

$$i(\lambda) = \hat{R}_K + (\frac{1}{\bar{X}} - 2)\,\hat{R}_{K-1} + \hat{R}_{K-2} + \frac{\hat{P}_K^2}{P_{K+1}} \tag{17}$$

Výpočty je možno provádět tak, že místo pravděpodobností P_i, $\hat{R}_i$ používáme očekávané četnosti, a výsledek dělíme N. Tento způsob je pohodlnější a i numericky vhodnější. Zjištěné standardizované reziduální odchylky Z_j můžeme použít k určení znamének podle zvoleného rozhodovacího pravidla [Řehák, Řeháková 1978 : 625, pozn. 11] a tím si zjednodušit orientaci v datech.

Postup lze využít také tam, kde bychom si předem určili pole, jehož odchylku chceme testovat.

Příklad 2

V tabulce 3 je uvedeno rozložení odpovědí na otázku: „Kolikrát jste v měsíci březnu navštívili koncert populární hudby?"

Tabulka 3. Počet návštěv koncertů populární hudby v měsíci březnu (procentní rozložení pro jeden z výběrových souborů čtenářů MS)

Počet návštěv	0	1	2	3	4	5	6	Celkem
Soubor čtenářů	342	140	31	8	3	2	2	528

Pro testování kritériem chí-kvadrát nejprve odvodíme průměr $\bar{X}$, očekávané četnosti a poslední z nich sloučíme tak, aby očekávaná četnost spojené kategorie byla alespoň 5.0 (stejně jako v příkladu 1b), pouze pro znaménkové schéma máme přísnější požadavek na četnost posledního intervalu).

Dostaneme postupně $\bar{X} = \frac{260}{528} = .4924$ a $X^2 = 13.2$ pro $df = 2$ (očekávané četnosti a sumarizace výsledků viz tabulka 4). Tato hodnota je vyšší než 5.99 (kritická hodnota pro $\alpha = 0.05$).

Vypočteme $i(\lambda)$:

Způsob (1): $i(\lambda) = \frac{1}{\bar{X}} = 2.0308$; $\frac{1}{i(\lambda)} = .4924$ (vzorec (16))

Způsob (2): $Ni(\lambda) = 520.70 + (2.0308 - 2)\,481.58 + 322.68 + \frac{39.12^2}{7.3} = 1067.8382$ (vzorec (17) vynásobený N)

$i(\lambda) = 1067.8382/\,528 = 2.0224$

$1/i(\lambda) = .4945$

Kumulativní četnosti viz tabulka 4. Dále vezmeme zá základ výpočtů výsledky druhého způsobu. Místo (15) použijeme N-násobku, který je uveden v rovnicích (18)

$$w_0 = Nv_0 = N - E_0\left(1 + \frac{1}{i(\lambda)}\right) \tag{18}$$

$$w_j = Nv_j = N - E_j\left[1 + (\frac{j}{\bar{X}} - 1)^2 \cdot \frac{1}{i(\lambda)}\right], \quad j = 1, 2, \ldots, K$$

$$w_{K+1} = Nv_{K+1} = N - E_{K+1}\left[1 + (\frac{E_K}{E_{K+1}})^2 \cdot \frac{1}{i(\lambda)}\right]$$

a vzorec (14) se změní na

$$Z_j = \frac{n_j - E_j}{\sqrt{\frac{E_j w_j}{N}}} \tag{19}$$

Smysl tohoto postupu je ve zkrácení a zjednodušení výpočtů (nemusíme vůbec zjišťovat P_j, stačí k tomu E_j).

Proto:

$$w_0 = 528 - 322.68(1 + 0.4945) = 45.75$$
$$w_1 = 528 - 158.90[1 + (2.0308 - 1)^2 \times 0.4945] = 285.61$$
$$w_2 = 528 - 39.12[1 + (2 \times 2.0308 - 1)^2 \times 0.4945] = 307.55$$
$$w_3 = 528 - 7.30\,[1 + (7.30/39.12)^2 \times 0.4945] = 520.57$$

a dále pak standardizované reziduální odchylky

$$Z_0 = \frac{342 - 322.68}{\sqrt{\frac{322.68 \times 45.75}{528}}} = 3.65$$

$$Z_1 = \frac{140 - 158.90}{\sqrt{\frac{158.90 \times 285.61}{528}}} = -2.04$$

$$Z_2 = \frac{31 - 39.12}{\sqrt{\frac{39.12 \times 307.55}{528}}} = -1.70$$

$$Z_3 = \frac{15 - 7.30}{\sqrt{\frac{7.30 \times 520.27}{528}}} = 2.87$$

Tyto hodnoty porovnáme s hranicemi ± 1.96, ± 2.58, ± 3.29: interval $(-1.96, 1.96)$ určuje ve schématu nulu a při postupném překračování hranic dostaneme $+$, $-$; $++$, $--$; $+++$, $---$. Shrnutí výsledků podává tabulka 4.

Interpretace výsledku:

Znaménkové schéma odhaluje, že vzhledem k modelu Poissonova rozložení máme mnohem více pozorování v kategorii $X = 0$ a v kategorii $X =$ (3 a více) a relativně se nedostává pozorování v kategorii $X = 1$.

Tabulka 4. Výpočet znaménkového schématu — souhrnná tabulka postupu pro data tabulky 3

Počet návštěv i	0	1	2	3 a více	Celkem
Četnosti n_i	342	140	31	15	528
Očekávané četnosti E_i	322.68	158.90	39.12	7.30	528
Kumulativní očekávané četnosti $F_i = NR_i$	322.68	481.58	520.70	528	—
Hodnoty w_i	45.77	285.61	307.55	520.27	—
v_i	.0867	.5409	.5825	.9859	—
Standardizované reziduální odchylky Z_i	3.65	−2.04	−1.70	2.87	—
Znaménkové schéma	+++	−	0	++	

$X^2 = 13.2$, df $= K = 2$, významné na hladině 0.05
$\overline{X} = .49$, var $X = .72$, $D = 767.57$, var $X/\overline{X} = 1.46$

Odchylky uvedené v tabulce 4 naznačují, že se soubor rozkládá na tři části: 1) část souboru, která nechodí na koncerty populární hudby; 2) část souboru, která má společnou intenzitu této návštěvní činnosti (kterou neznáme, ale která nemusí být příliš různá od $\bar{X} = .49$); 3) část souboru (není nikterak velká), která navštěvuje populární hudbu daleko intenzívněji než průměrný návštěvník souboru.

Skóre Z_i můžeme použít jako *testové kritérium v předem určeném poli*. Používáme toho především v testování odchylky v poli $X = 0$. Taková hypotéza je sociologicky zajímavá proto, že významně vyšší četnost v tomto poli může indikovat část populace, která se vůbec dané činnosti nevěnuje (tj. pravděpodobnost účasti na činnosti je nula).

Jiná hypotéza pro odchylku v poli se může týkat některé kategorie znamenající zaokrouhlenou odpověď (jako je pět, deset a podobně).[19])

C. Odhad parametru

Za předpokladu, že získané rozložení je Poissonovo, je odhadem intenzity λ aritmetický průměr $\bar{X}$. Můžeme však také snadno získat intervaly spolehlivosti pro λ jako:

$$\lambda \in (\lambda_D, \lambda_H) \tag{20}$$

Horní a dolní hranici pro interval se spolehlivostí $100(1-\alpha)\%$ určíme pro malé $N\lambda$ (řekněme $\leqq 5$) z interpolace distribuce chí-kvadrát.

Je-li $X = \Sigma i n_i$ větší než 5, pak použijeme vzorce[20])

$$\lambda_D = \bar{X} + \frac{1}{N}\left[\frac{1}{2}u^2 - u\sqrt{X + \frac{1}{4}u^2}\right] \tag{21}$$

$$\lambda_H = \bar{X} + \frac{1}{N}\left[\frac{1}{2}u^2 + u\sqrt{X + \frac{1}{4}u^2}\right], \tag{22}$$

kde u je kvantil normálního rozložení. Hodnoty u jsou v tabulce B:

Tabulka B. Koeficienty u v rovnicích (21), (22)

Spolehlivost	Hodnota u
90%	1.6449
95%	1.9600
99%	2.5758
99,9%	3.2905

Tento postup má význam jen v tom případě, že jde opravdu o Poissonovo rozložení. Například v příkladu 2 tento postup aplikovat nelze, neboť jsme v něm hypotézu o shodě empirického výskytu dat a modelu (1) odmítli.

Příklad 1 (pokračování)

Z dat tabulky 1 jsme odvodili platnost modelu Poissonova rozložení. Jeho parametr λ (= intenzita měsíční návštěvnosti koncertů populární hudby) byl odhadnut pomocí průměru $\bar{X}$ jako

[19]) V jedné aplikační situaci odhalil test přetížení kategorie 10; kontrola celého procesu práce s daty odhalila příčinu, která nebyla v zaokrouhlování údajů respondentů, jak bylo očekáváno, ale v tom, že kategorie 1 byla často omylem kódována jako 10.

[20]) Viz například Johnson, Kotz [1969 : 96]; Janko [1958 : 83] uvádějí nepatrně rozdílné vzorce s korekcemi pro spojitost, které však pro větší počet respondentů jsou bezpředmětné. Pro $X \leqq 5$ použijeme k určení intervalu podrobnějších tabulek rozložení chí-kvadrát [Janko 1958 : 83].

Základní knihovna ÚSAV

$\hat{\lambda} = 1.29$. Abychom určili interval spolehlivosti, potřebujeme k tomu ještě $X = 237$, $N = 184$.

$$\lambda_D = 1.29 + \frac{1}{184}\left[\frac{1.96^2}{2} - 1.96\sqrt{237 + \frac{1.96^2}{4}}\right] = 1.1361$$

$$\lambda_H = 1.29 + \frac{1}{184}\left[\frac{1.96^2}{2} + 1.96\sqrt{237 + \frac{1.96^2}{4}}\right] = 1.4648$$

95%ní konfidenční interval pro λ : (1.14, 1.46).

Pro spolehlivost 99% provedeme stejný výpočet, pouze dosadíme $u = 2.5758$: interval dostaneme jako (1.09. 1.52).

D. Porovnání dvou nebo více Poissonových rozložení

Dvě nebo více rozložení obvykle porovnáváme neparametricky tím, že aplikujeme chí-kvadrát test nezávislosti [Řehák - Řeháková 1978]. Jestliže však víme, že distribuce jsou poissonovské, porovnávání můžeme provést přesnějším způsobem při využití informace, kterou Poissonovo rozložení poskytuje. Přecházíme ke komparaci pomocí parametrických metod, které jsou jemnější a jsou platné (za předpokladu platnosti Poissonova modelu) i pro malé počty pozorování (zde i pro $N = 1$).

Testování homogenity Poissonových distribucí provádíme postupem:[21])

1. *Hypotéza* H_0: Nezávislé Poissonovy distribuce mají stejnou intenzitu:

$$\lambda_1 = \lambda_2 = \lambda_3 = \ldots = \lambda_R\,, \quad R = \text{počet porovnávaných distribucí}$$

Alternativní hypotéza: Alespoň jedna z intenzit λ_i je odlišná od ostatních.

2. *Uspořádání dat*:

Máme-li R rozložení četností, (n_{ij}), $i = 1, 2, \ldots, R$, zjistíme pro každé z nich:

$$\begin{aligned} \textit{počet pozorování} &= N_i = \Sigma n_{ij}\,, \quad i = 1, 2, \ldots, R \\ \textit{součet pozorování} &= X_i = \sum_j j n_{ij}\,, \quad i = 1, 2, \ldots, R \end{aligned}$$

n_{ij} = počet výskytů hodnoty j v i-tém souboru dat

$$N = \Sigma N_i \qquad X = \Sigma X_i$$

(23) $$\textit{očekávané součty} = S_i = \frac{N_i}{N} \cdot X$$

Údaje uspořádáme do tabulky C.

4. *Testové kritérium* — postup výpočtu

Testování provádíme kritériem chí-kvadrát dobré shody:

(24) $$X^2 = \sum_{i=1}^{R} \frac{(X_i - S_i)^2}{S_i}$$

(25) $$= \sum \frac{X_i^2}{S_i} - X$$

[21]) Postup je založen na vlastnosti: jsou-li n_i $(i = 1, \ldots, R)$ Poissonovy proměnné s parametry λt_i, pak rozložení $(n_1, n_2, \ldots, n_R)$ za podmínky, že $\Sigma n_i = n$ je multinomické s parametry n a $p_i = t_i/\Sigma t_i$.

$$= \frac{N}{X} \cdot \sum \frac{X_i^2}{N_i} - X \tag{26}$$

počet stupňů volnosti $df = R - 1$

Hypotézu H_0 zamítáme, je-li X významné na předem zvolené hladině významnosti α.

Poznámky k testu:

1. Test je použitelný za předpokladu, že každá z R distribucí se řídí modelem (1), pro jiné geneze náhodných veličin není platnost závěru zajištěna. Robustnost vzhledem k jiným modelům není známa.
2. Při zamítnutí nulové hypotézy je možné provést běžnou ďalší analýzu komparace procentních hodnot, jako je například znaménkové schéma pro generování alternativních hypotéz [Řeháková 1979] nebo simultánní inference pro řazení procent, testy pro poměr procent a podobně.
3. Test nemá omezení na N_i, N_i může být rovno jedné; platí však obvyklá omezení na testy dobré shody, a to vzhledem k hodnotám S_i.

Příklad 3

„Kolikrát jste v měsíci březnu navštívil kino?"
U třech nepřekrývajících se podsouborů byla zjištěna poissonovská vlastnost (shoda teoretického modelu a dat). U každého podsouboru byly zjištěny intenzity (viz tab. 5). Otázka zní, zda můžeme předpokládat, že se tyto intenzity vzájemně neliší.

Tabulka 5. Údaje o třech souborech

Soubory	Počet údajů (N_i)	Součet návštěv (X_i)	Intenzita (λ_i)
Soubor A	121	341	2.82
Soubor B	85	234	2.75
Soubor C	64	107	1.67
Celkem	270	682	2.53

Pozn.: intenzita sjednocení ($\lambda = 2.53$) by byla vhodným společným odhadem pouze v případě, že prokážeme rovnost jednotlivých dílčích intenzit.

Tyto údaje převedeme do tvaru tabulky C:

Tabulka C. Porovnání R Poissonových distribucí

Soubor	1	2	...	R	Celkem
Součet pozorování	X_1	X_2	...	X_R	X
Očekávaný součet	S_1	S_2	...	S_R	X

Podle vzorců (23), (25):

$$S_1 = \frac{121}{270} \cdot 682 = 305.64$$

$$S_2 = \frac{85}{270} \cdot 682 = 214.70$$

$$S_3 = \frac{64}{270} \cdot 682 = 161.66$$

$$\bar{X}^2 = \frac{341^2}{305.64} + \frac{234^2}{214.70} + \frac{107^2}{161.66} - 682 = 24.31$$

počet stupňů volnosti $df = 2$

Podle vzorce (26):

$$X^2 = \frac{270}{682}\left[\frac{341^2}{121} + \frac{234^2}{85} + \frac{107^2}{64}\right] - 682 = 24.31$$

Získaná hodnota je vysoce významná, proto hypotézu shody intenzit u všech tří rozložení považujeme za vyvrácenou.

Při postupu uváděném v příkladu 3 vidíme, že druhý postup (aplikace vzorce (26)) je jednodušší a rychlejší, není ani třeba vytvářet tabulku 6 (není třeba počítat očekávané četnosti). Přesto uvádíme oba postupy.

Tabulka 6. Porovnání tří Poissonových distribucí

Soubor	A	B	C	Celkem
Součet návštěv (X_i)	341	234	107	682
Očekávané četnosti	305.64	214.70	161.66	682

Komparaci Poissonových distribucí můžeme prohloubit ještě dále celou řadou dalších postupů, z nichž jmenujme především regresní přístup závislosti intenzity λ na nějakém jiném parametru (například na věku, vzdělání, časovém umístění atd.).

5. Závěr

Ve stati byly uvedeny jednoduché metody analýzy dat, které vypovídají o výskytu jevu v určitém časovém intervalu a vznikají jako realizace Poissonova procesu. Taková data vznikají v případech sledování různých procesů, jejichž intenzita, ale i mechanismy geneze, jsou pro společenskovědní interpretace velmi zajímavé a důležité. Příklady použití k numerické ilustraci jsou vzaty z výzkumné praxe a ukazují, že aplikace Poissonova rozložení je možná i v datech tradičního dotazníkového sociologického šetření. V této souvislosti se velmi silně prosazuje pojem poissonovské homogenity souboru dat, který vnáší kvalitativní informaci o možném způsobu geneze dat. Nalezení homogenních podsouborů však není jednoduché a nemůžeme očekávat, že takové hledání lze rutinizovat. Důležitou úlohou se jeví také identifikace odklonu od modelu; existuje řada jiných modelů pro analýzu výskytu jevů v čase a každý z nich zasluhuje samostatné prozkoumání vzhledem k použitelnosti v sociologii.

Výzkumníkům, kteří jsou zvyklí na rutinní zpracování pomocí kontingenčních tabulek, se může jevit podobná analýza jako zbytečná, zbytečně jemná či speciální, pátrání po platnosti parametrických modelů jako nepotřebné. A přesto taková snaha zbytečná není. Ověření platnosti určitého modelu má přinejmenším tyto podstatné rysy:

1. umožňuje daleko *přesnější informaci o parametrech* a dává parametrům zcela *specifický význam*, věcně z modelu interpretovatelný,
2. umožňuje získat informaci o *mechanismech geneze dat*, to jest informaci o kvalitě a podstatě procesu, který data produkuje,

3. umožňuje *zpětný tlak na metodologii* sběru dat a tím působí na kvalitu disponibilních datových základen,

4. umožňuje *přesnější posouzení souvislostí* měřených proměnných s jinými proměnnými, a to jak srovnání procesů stejného typu (zde poissonovského), tak porovnání kvalitativní s jinými typy procesů,

5. umožňuje *použití* jednoduchých modelů v modelech daleko komplexnějších, které souhrnně odrážejí širší společenské souvislosti a obsahují větší počet proměnných a kde vhodná formalizace jednotlivých složek podstatně zpřesňuje celkovou platnost modelu,

6. umožňuje *redukovat informaci* z celého rozložení do jedné nebo několika smysluplných charakteristik, které dostatečně soubor opisují.

Dále bychom mohli jmenovat další úlohy, jako zjednodušení sekundárních analýz, klasifikaci sociálních indikátorů, analýzu opakovaných šetření, analýzu časových řad atp. atp.

Metody uvedené ve stati jsou (až na znaménkové schéma) pro tento případ dobře známé a v citované literatuře dobře popsané. Cílem této stati bylo především upozornit na úlohu zkoumání poissonovské homogenity, a tak obohatit naše nejen analytické, ale především interpretační možnosti. Plný význam modelu se však projeví až ve spojitosti a v porovnání s jinými modely geneze dat.

Dodatek:

Standardizace reziduálních odchylek při testu dobré shody Poissonova rozložení

Přijmeme označení tak, jak bylo zavedeno ve stati (1) (7), (8) (10), n_j = počet pozorování v j-tém poli.

Podle teorie asymptotických odhadů [Rao 1978 : kap. 6] mají reziduální odchylky za předpokladu platnosti modelu (tj. nulové hypotézy dobré shody) normální rozložení s nulovým průměrem a rozptylem, který lze spočítat [Rao 1978 : 6b. 3; 437]. Standardizované reziduální Z-skóre pro pole j spočteme jako:

$$Z_j = \frac{n_j - E_j}{\sqrt{v_j E_j}},$$

$$v_j = 1 - P_j - \frac{1}{P_j}\left(\frac{\mathrm{d}P_j}{\mathrm{d}\lambda}\right)^2 \cdot \frac{1}{i}$$

$i = i(\lambda)$ = Fisherova míra informace o parametru λ, obsažené v jednom pozorování.
Dosazením (1) dostaneme:

$$v_0 = 1 - P_0 - P_0 \frac{1}{i}$$

$$v_j = 1 - P_j - \left(\frac{P_{j-1} - P_j}{P_j}\right)^2 P_j \cdot \frac{1}{i}, \quad j = 1, 2, \ldots, K$$

$$= 1 - P_j\left[1 + \left(\frac{j}{\lambda} - 1\right)^2 \frac{1}{i}\right]$$

$$v_{K+1} = 1 - P_{K+1} - \frac{P_K^2}{P_{K+1}} \cdot \frac{1}{i}$$

$$= 1 - P_{K+1}\left[1 - \frac{P_K^2}{P^2_{K+1}} \cdot \frac{1}{i}\right]$$

Zbývá tedy odhadnout Fisherovu informační míru. Obecně platí, že

$$i = i(\lambda) = \sum_{j \in J} \frac{1}{P_j} \left(\frac{\mathrm{d}P_j}{\mathrm{d}\lambda} \right)^2 ,$$

kde J je množina indexů kategorií pro dané rozložení. Budeme uvažovat nyní dva způsoby odvození informace:

Způsob I

Pro původní Poissonovo rozložení, u něhož jsou P_j dány vzorcem (1) a J je množina všech nezáporných celých čísel snadno zjistíme, že $i(\lambda)$ je dáno jako

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

Způsob II

Testování dobré shody provádíme pro kategorizované rozložení tak, že spojíme všechny kategorie od $(K + 1)$ní výše. Proto jsou P_j $(j = 0, \ldots, K)$ dány vzorcem (1), zatímco P_{K+1} je určeno jako:

$$P_{K+1} = \sum_{k=K+1}^{\infty} \mathrm{e}^{-\lambda} \cdot \frac{\lambda^k}{k!}$$

Označíme

$$R_i = \sum_{k=0}^{i} P_k$$

a spočteme Fisherovu informaci pro tento případ (v odvození jsou použity vztahy (18.3), (18.7) a (17.3) z práce Johnson - Kotz [1969]):

$$i(\lambda) = P_0 + \sum_{k=1}^{K} \frac{(P_{k-1} - P_k)^2}{P_k^2} \cdot P_k + \frac{P_K^2}{P_{K+1}}$$

$$= P_0 + \sum_{k=1}^{K} \left(1 - 2\frac{k}{\lambda} + \frac{k^2}{\lambda^2}\right) P_k + \frac{P_K^2}{P_{K+1}}$$

$$= R_K + \left(\frac{1}{\lambda} - 2\right)(R_{K-1} + R_{K-2} + \frac{P_K^2}{P_{K+1}}$$

Pro velká K se takto vypočtená informační míra limitně blíží k hodnotě odvozené pro Poissonovo rozložení (viz způsob I) bez slučování kategorií.
Asymptoticky má veličina Z_i normální standardizované rozložení. Proto ji lze aplikovat na testování odchylky v jednom předem určeném poli.
Při výpočtech a testování postupujeme tak, že za P_i dosazujeme očekávané hodnoty $\hat{P}_i$ a za λ dosadíme odhad $\bar{X}$.

Literatura

Cox, D. R. - Lewis, P. A. W.: *Statističeskij analiz posledovatěl nostěj sobytij*. Moskva, Mir 1969.

Feller, W.: *Vvěděnije v těoriju věrojatnostěj i jego priloženija*. Tom I. Moskva, Mir 1967.

Janko, J.: *Statistické tabulky*. Praha, Nakladatelství ČSAV 1958.

Johnson, N. L. - Kotz, S.: *Distributions in Statistics*: *Discrete Distributions*. Boston, Houghton Mifflin Company 1969.

Johnson, N. L. - Kotz, S: *Distributions in Statistics*: *Continuous Univariate Distributions — 1*. Boston, Houghton Mifflin Company 1970.

Kendall, M. G. - Stuart, A.: *Statističeskije vyvody i svjazy*. Moskva, Nauka 1973.

Rao, C. R.: *Lineární metody statistické indukce a jejich aplikace*. Praha, Academia 1978.

Řehák, J. - Řeháková, B.: *Analýza kontingenčních tabulek*: *rozlišení dvou typů úloh a znaménkové schéma*. Sociologický časopis 1978, č. 6, s. 619—631.

Řeháková, B.: *Statistické ověřování reprezentativity*: *testy dobré shody*. Sociologický časopis 1979, č. 6, s. 615—629.

Yarnold, J. K.: *The Minimum Expectation in χ^2 Goodness of Fit Tests and the Accuracy of Approximations For the Null Distribution*. Journal of the American Statistical Association 1970/65, pp. 864—886.

Резюме

Я. Ржегак: Анализ данных о появлении событий в временном промежутке Применение модели распределения Пуассона

Основная модель для случайных величин распределенных по закону Пуассона находит множество различных применений в социальных исследованиях. В статье рассмотрены основные статистические техники для работы с данными Пуассоновского типа: критерии проверки модели (хи-квадрат и индекс дисперсии Фишера), знаковая схема для спецификации алтернативной гипотези, доверительные интервалы для параметров распределения (представляющего интензивность потока событий) и проверки равенства интензивностей для нескольких пуассоновских распределений.

Основной целью статьи является введение проблемы, которая не так часто встречается в социологических исследованиях: искание пуассоновской однородности и существования пуассоновского процесса, чтобы получить важную качественную информацию о возникновении данных и о механизмах воздействующих на появления исследованных событий.

Нумерическая иллюстрация статистических техник показана на примерах из исследовательской практики — из того следует что проблема практична и ее более широкое введение в анализ социологических данных может принести пользу для интерпретации даже в данных анкеты и опросных листов.

В статье тоже дискутируется обратная связка требований вытекающих из применительности модели на методологию сбора данных и на формулировку отдельных анкетных вопросов. В приложении выведены формулы которые лежат в основе указанного в статье метода знаковой схемы.

Summary

Řehák J.: Analysis of Data Concerning an Occurrence of Events in a Time Interval. An Application of the Poisson Distribution Model

The basic model for Poisson distribution can be used in a number of situations of the social science data analysis. The paper offers the basic techniques for testing goodness-of-fit (chi-square test and Fishers dispersion index), the sign scheme that helps in formulation of more specific alternative hypothesis and serves as a rapid graphical aid for orientation in data, and shows the interval estimation of intensity parameter as well as the comparison of intensities under the assumption of Poisson distribution for each compared sample.

The main aim of the paper is, however, to introduce the problem that is not common in a routine analysis of research data: the search for the existence of Poisson proces and Poissonian homogeneity to obtain an important qualitative information on the genesis of data and mechanism that functions under investigated process. The methods of the paper are ilustrated by numerical examples that are, however, taken from the research practice; thus the problem is not artificial even in the traditional questionnaire surveys, from which data come.

In the paper, the important methodological feedback of a model on research methodology is disscussed. In the appendix, the derivation of the test for deviation in individual cells is used for the presented method of sign scheme.