Bayesovský znaménkový test

JAN ŘEHÁK
Ústav pro filozofii a sociologii ČSAV, Praha

BLANKA ŘEHÁKOVÁ
Ústav pro výzkum veřejného mínění FSÚ, Praha

V sociologickém výzkumu, v sekundárních analýzách, v komparativních úlohách i při vyhodnocování experimentálních uspořádání sběru dat se velmi často vyskytují diskrétní znaménková data, která byla popsána v práci Řehák, Řeháková [1978], uvádějící též řadu různých ilustrujících příkladů a metody klasické statistiky pro základní vyhodnocení těchto dat.

Ve stati předkládáme alternativní metodu [1] zpracování těchto dat, založenou na bayesovských statistických základech. Bayesovský statistický přístup je v našem sociologickém výzkumu málo používán, přestože tvoří značnou a neoddělitelnou součást moderní matematické statistiky; proto jej krátce charakterizujeme. Metoda je popsána v krocích, ve kterých ji aplikujeme v praxi a je též ilustrována numerickými příklady. Pro snadnou aplikaci metody uvádíme jednak základní tabulku A jednak pomocné tabulky normální distribuční funkce a normálních kvantilů (tabulky B a C). Uživatel v běžné výzkumné praxi vystačí většinou s těmito základními údaji a popisem postupu. Matematické formulace jsou určeny metodologům a statistikům.

Metoda má své přednosti i nedostatky. Její aplikabilita závisí na konkrétní situaci, konkrétním výzkumníkovi a jeho znalostech o zkoumané populaci. K jejím výhodám patří:

— má jednoduchý princip odrážející běžný gnoseologický postup;
— je velmi snadno aplikovatelná v běžném hodnocení četností;
— umožňuje přímé spojení apriorní informace, znalostí, přesvědčení, pilotážní či předvýzkumné informace, které jsou závislé na sběru dat, s informací získanou z výběrového souboru;
— umožňuje formálně prokázat souvislost obou hlavních hypotéz, které jsou předmětem zájmu této stati.

K nevýhodám patří:

— je často obtížné určit vstupní parametry, které charakterizují sumu naší apriorní evidence a výzkumného přesvědčení.

Jde o *metodu statistickou*, přesto však největší nároky pro uživatele jsou ve sféře *sociologického myšlení*, neboť určování apriorních vah i interpretace aposteriorních rozhodnutí je věcí meritorní, nikoli formálně matematickou či věcí standardní rutiny.

1. Hypotézy o rozložení znaménkových dat a postup bayesovského uvažování o nich

Znaménkový znak (označme jej S) charakterizujeme třemi hodnotami: „+" = kladná kategorie, „0" = neutrální kategorie, „−" = záporná kategorie. Je to tedy trichotomický znak; je ordinální a má pevný střed (neutrální bod). Při jeho realizaci na da-

[1] Metoda byla oznámena v přednášce Řehák, Řeháková [1972] a aplikována v otorinolaryngologickém výzkumu (Tománek [1971], [1972]). Tabulka A, která je základem metody byla spočtena na počítači MINSK 22.

ném výběrovém souboru dostáváme rozložení četností (absolutních či relativních, resp. procent) na těchto třech hodnotách. Dále budeme pracovat s relativními četnostmi, resp. s pravděpodobnostmi výskytu v populaci, ze které výběr pochází. Tak dostáváme tři základní parametry rozložení četností:

(1) $P\ (S = \text{„+“}) = p_+$
$P\ (S = \text{„0“}) = p_0$
$P\ (S = \text{„−“}) = p_-$

které splňují podmínku $p_+ + p_0 + p_- = 1$, $p_+ \geqq 0$, $p_0 \geqq 0$, $p_- \geqq 0$ (leží v tzv. simplexu třírozměrného prostoru).

Parametry charakterizují tedy obsazení polí (kategorií).

Ve výzkumné praxi nás zajímají především *hypotézy o existenci majoritní pravděpodobnosti*: [2]

(2) $H_1 : p_+ > \frac{1}{2}$
$H_2 : p_0 > \frac{1}{2}$
$H_3 : p_- > \frac{1}{2}$

Jestliže tyto hypotézy jsou příliš silné, pak nás zajímají slabší, ale meritorně též nemírně důležité *hypotézy asymetrie*:

(3) $H_4 : p_+ > p_-$ $\quad H_5 : p_+ < p_-$
$H_6 : p_+ > p_0$ $\quad H_7 : p_+ < p_0$
$H_8 : p_- > p_0$ $\quad H_9 : p_- < p_0$

Z nich nejdůležitější jsou H_4 a H_5, které znamenají převahu v obsazení kategorie „+“ nad kategorií „−“, nebo naopak. Tyto hypotézy porovnávají obsazení dvou kategorií bez ohledu na obsazení třetí. H_1, H_2, H_3 znamenají ovšem silnější výpověď, neboť přijímáme-li hypotézu majoritní pravděpodobnosti, automaticky přijímáme též příslušnou hypotézu asymetrie (Je-li např. $p_+ > \frac{1}{2}$, pak nutně musí být větší než p_- i p_0).

Bayesovský přístup spočívá ve třech krocích:

1. *Konstrukce modelu pro chování výběrových dat.*[3]
2. *Formulace apriorního názoru.*
3. *Matematické spojení obou předchozích kroků pomocí Bayesovy věty.*[4]

První krok je společný celé statistice jak klasické, tak Bayesovské, neboť každá statistická metoda je založena na jistých (buď velmi silných či velmi volných) předpokladech, které musíme při praktickém použití metod statistiky vždy prověřovat (nejsou-li splněny předpoklady, metoda dává nespravné podklady pro interpretaci). V našem případě je tento krok volen tak, aby metoda byla neparametrická a nebyla zatížená žádnými distribučními předpoklady.

Druhý krok znamená velmi obtížnou úvahu o tom, co víme či co si myslíme (apriori, aniž jsme viděli data) o parametrech našeho modelu. Je odpovědí na otázku: jaké váhy položíme na jednotlivé možnosti pro trojice (p_+, p_0, p_-). Některé z těchto trojic můžeme například vyloučit tím, že jim dáme váhu nula, či prakticky vyloučit tím, že jim dáme váhu velmi nepatrnou v poměru k jiným hodnotám.

---

[2] Někdy nás zajímají také negace těchto hypotéz:

$\overline{H}_1 : p_+ < 1/2$, $\overline{H}_2 : p_0 < 1/2$, $\overline{H}_3 : p_- < 1/2$

nebo zobecnění, jímž se zde nezabýváme, ale které je možné pro kteroukoli z pravděpodobností H: pro dané a, b platí $a < p < b$.

[3] V našem případě např. považujeme za vhodný model multinomického rozložení pravděpodobností; po získání dat a dosazení empirických hodnot za příslušné parametry modelu dostáváme tzv. věrohodnostní funkci.

[4] O Bayesovské statistice existuje velmi rozsáhlá literatura. Zájemcům doporučujeme obzvláště práce De Groot [1974], Savage [1968], Schmitt [1969]. O problému apriorních pravděpodobností viz též Řehák [1974].

V praxi je užitečné a výhodné používat tzv. *konjugovaného* (*přirozeně sdruženého, spřaženého*) *rozložení*, které znamená výběr z určité třídy různých funkcí — výběr jedné váhové funkce, která nejlépe odpovídá našim apriorním zkušenostem, resp. názorům. (Zde použijeme tzv. Dirichletova rozložení, jak je zřejmé z matematické formulace úlohy; o tom, jak je možno zjednodušit některé úvahy o apriorním rozložení v našem konkrétním případě, pojednáme níže).

Třetí krok znamená matematické odvození aposteriorního rozložení a z něho pak odvození pravděpodobnosti žádané hypotézy.

Klasická statistika pracuje pouze se zvoleným modelem chování dat, s věrohodnostní funkcí, která vzniká dosazením empirických hodnot do modelu a rozhoduje o hypotézách na základě této funkce.

Bayesovská statistika pracuje navíc ještě s apriorní informací o neznámých parametrech. Motivem pro to je především:

a) snaha modelovat postupnost vědeckého poznání a neustálého přibližování se k pravdě;
b) snaha rozhodovat o hypotézách na základě přímých indikací; zatímco klasická statistika rozhoduje na základě toho, jak pravděpodobným je výskyt obdržených dat za předpokladu dané hypotézy, bayesovská statistika rozhoduje na základě toho, jakou podporu přinášejí data daná hypotézou, na základě toho, jak vysokou pravděpodobnost hypotézy můžeme z dat odvodit, tj. jaké váhy určují data pro zkoumané konkurující si hypotézy.

O úspěšnosti přístupu rozhodne ovšem sama dlouhodobá praxe vědeckého poznávání. Bayesovský princip však podstatně zvyšuje nároky na uživatele a na jeho vztah ke statistické technice jakožto instrumentu poznání. Jednotlivé metody tohoto přístupu nejsou pouhými mechanickými nástroji, stávají se *prostředkem dialogu mezi uživatelem a daty*.

Vztah apriorního rozložení a empirických dat lze charakterizovat takto:

1. Jestliže máme málo pozorování, pak apriorní rozložení může být pouze málo ovlivněno, a tudíž v konečných závěrech se silné apriorní přesvědčení prosadí silněji než výzkumná datová evidence (to je ovšem paralela ke známému: věříme-li v určitou hypotézu, pak nás malý výběrový soubor nepřesvědčí o její neplatnosti; slabá empirická evidence nemůže zvrátit náš silný názor).
2. Čím větší je soubor empirických dat, tím menší je vliv našeho apriorního přesvědčení na konečný výsledek. Apriorní názor má v konečném rozhodnutí stále menší váhu s tím, jak roste počet pozorování.
3. Je-li naše přesvědčení jednoznačné, tj. apriorní váha je pro jednu z možných alternativ rovna jedné (a všechny ostatní možnosti mají tedy nulovou apriorní váhu), pak empirická evidence nemá na aposteriorní rozhodnutí žádný vliv.

V jednotlivých případech je možné na apriorní názor či evidenci abdikovat, potlačit jej, neuvažovat žádné preference předem. Také tehdy, je-li náš názor natolik neurčitý a natolik neformulovatelný, že nejsme schopni žádné preference formulovat, pak můžeme dát všem trojicím parametrů (p_+, p_0, p_-) stejnou váhu. [5] V takovém případě je závěr o hypotézách založen pouze na empirické evidenci.

Bayesovský přístup má různé varianty, z nichž některé jsou založeny na tom, že apriorní evidenci odhadujeme z jiných dat, některé se pokoušejí využít empirických četností z populace apod. Nebudeme zde rozlišovat na této obecné rovině, jaký je obsah apriorního vstupu: může to být jen vyjádření přesvědčení o platnosti hypotézy, stejně tak jako empirické údaje z pilotážního nebo předvýzkumného šetření.

---

[5] Tento postup je založen na Bayesově postulátu o rovnoměrné neznalosti: „jestliže neznáš apriorní rozložení, dej všem možnostem stavu reality apriori stejnou váhu"; v diskutovaném případě je to rovnoměrné rozložení na simplexu trojic (p_+, p_0, p_-).

2. Apriorní rozložení, význam jeho parametrů a jejich určení

Metoda vyhodnocení hypotéz o znaménkových datech je založena na tom, že rozhodujeme na základě tří čísel,[6] parametrů aposteriorního rozložení, jejichž význam vyplyne z následující úvahy.

Apriorní rozložení má stejný tvar jako aposteriorní, má stejné parametry, které se liší pouze číselnou hodnotou.
Aposteriorní hodnoty parametrů vznikají jako:

$$N_+ = m_+ + n_+,\ N_0 = m_0 + n_0,\ N_- = m_- + n_-, \tag{4}$$

kde m_+, m_0, m_- jsou parametry apriorního rozložení a n_+, n_0, n_- jsou empirické absolutní četnosti.

Tedy: *aposteriorní parametr* = *absolutní četnost* + *apriorní parametr*

Obojí parametry mají tedy roli četností (mají stejný rozměr). Apriorní názor vyjádříme pomocí rozložení četností:
a) je buď vzato z pilotáže, předvýzkumu, či je získáno nějakou drobnou sondou;
b) nebo je ekvivalentní našemu subjektivnímu zkušenostnímu pohledu na věc;
V tom případě je $m_+ : m_0 : m_-$ poměr obsazení jednotlivých kategorií, tedy vyjadřuje náš názor, že

$$p_+ : p_0 : p_- = m_+ : m_0 : m_- \tag{5}$$

Součet

$$m = m_+ + m_0 + m_- \tag{6}$$

pak charakterizuje předpokládanou „přesnost", s níž jsme přesvědčeni o platnosti poměru (5). m je ekvivalentní velikosti výběrového souboru poskytujícího empirickou evidenci, která by byla stejně silná jako náš apriorní názor.

Nakonec rozhodujeme na základě hodnot (4). Získáváme je jako součet apriorních a empirických (resp. věrohodnostních) hodnot. *Ze vzorce (4) vidíme, jak se příspěvky skládají ve statistický závěr, jak se posilují, či jak se vzájemně mohou potlačovat.*

Chceme-li vyjádřit svůj apriorní názor pomocí *postulátu rovnoměrné neznalosti*, tj. položit na všechny trojice (p_+, p_0, p_-) stejnou váhu, pak toho dosáhneme volbou:

$$m_+ = m_0 = m_- = 1 \tag{7}$$

Parametry m_i $(i = +, 0, -)$, a tedy i N_i, mohou být též necelá čísla. Pro praktické účely však stačí omezit se na celé kladné hodnoty. Parametry m_i musí však být větší než nula a nejsou-li, pak musíme za ně dosadit nenulové číslo.

Například používáme-li m_2 jako četnosti z pilotáže, pak je-li jedna hodnota nulová, je lépe ji zaměnit jedničkou. Praktické výsledky se tím nikterak podstatně neovlivní.

Obdobně jako v (6) si ještě označíme: počet empirických pozorování jako n a součet aposteriorních parametrů jako N:

$$n = n_+ + n_0 + n_-\ ;\quad N = N_+ + N_0 + N_- \tag{8}$$

---

[6] Matematicky orientovaného čtenáře odkazujeme na část 5 a na literaturu citovanou v poznámce 4. Poznamenejme též, že označení „parametry" má ve stati dvojí význam: a) parametry rozložení znaménkových dat: (p_+, p_0, p_-) jsou pravděpodobnostní obsazení, tj. jejich hodnoty jsou předmětem našeho výzkumného zájmu; b) parametry apriorního a aposteriorního rozložení (n_i, m_i) určují, jaké váhy klademe a jaké jsme odvodili pro všechny trojice (p_+, p_0, p_-).

Klasické statistické metody často selhávají v praxi pro malé výběrové soubory. I když známe kritické hladiny různých testů pro malé soubory, tyto testy jsou málo silné, poskytují malou obranu proti chybě druhého druhu. To se obzvláště projevuje u souborů, které nejsou zcela homogenní (což je právě případ sociologických zkoumání). V takových případech je baeysovský přístup obzvláště cenný, neboť umožňuje studium souladu názoru a empirických dat. Též při odhadu založeném na tvaru a vlastnostech aposteriorního rozložení dostáváme při malých souborech lepší výsledky.

3. Metoda testování hypotéz H_1—H_9 a odhad pravděpodobností p_+, p_0, p_-.

Testování hypotéz bude založeno na tom, že k daným číslům N_+, N_0, N_- (viz (4)) najdeme pravděpodobnost platnosti příslušné hypotézy:

$$P = P\,(H_i/N_+, N_0, N_-)$$

Nalezení hodnoty P se děje podle postupu níže popsaného. Je-li P dostatečně vysoká, pak H_i můžeme přijmout. Je-li P velice malá, pak můžeme přijmout negaci hypotézy H_i.

A. Testování hypotézy o majoritní pravděpodobnosti (H_1) a hypotézy asymetrie (H_4)

Krok 1:

Určíme hodnotu γ ($\frac{1}{2} < \gamma < 1$), která určuje rozhodovací pravidlo [7]

(9) pro $P \geqq \gamma$ přijmeme H_1 (tj. $p_+ > \frac{1}{2}$)
pro $P \leqq 1 - \gamma$ přijmeme $\bar{H}_1$ (tj. $p_+ < \frac{1}{2}$)
pro $1 - \gamma < P < \gamma$ nejsou data dostatečně informativní pro rozhodnutí o H_1 resp. $\bar{H}_1$

Krok 2:

Určíme *své* apriorní rozložení pomocí parametrů m_+, m_0, m_-, a to bez ohledu na data; nechceme-li nebo nejsme-li schopni určit tyto hodnoty, volíme $m_+ = m_0 = m_- = 1$.

Krok 3:
Zjistíme empirické četnosti n_+, n_0, n_-.

Krok 4:

Odvodíme aposteriorní parametry

$$N_+ = m_+ + n_+, \quad N_0 = m_0 + n_0, \quad N_- = m_- + n_-$$

Krok 5:

V tabulce A aposteriorních pravděpodobností nalezneme:

a) pro hypotézu H_1: pole v řádku s číslem N_+ a ve sloupci s číslem $(N - N_+)$
b) pro hypotézu H_4: pole v řádku s číslem N_+ a ve sloupci s číslem N_-

---

[7] Tento krok je stále častěji vynechávám. Výzkumník pak rozhoduje na základě aposteriorní hodnoty P: přímo uváží její hodnotu a podle ní přijme či nepřijme zkoumanou hypotézu. Tento postup je analogický tomu, že u klasických testů si nenecháme tisknout zprávu, zda např. chí-kvadrát je statisticky významný na určité hladině, ale necháme si vytisknout prostě samotnou hodnotu významnosti a podle ní se rozhodneme. To má vážné interpretační důsledky především tam, kde současně hodnotíme řadu obsahově příbuzných proměnných, kde se u žádné proměnné neprojeví signifikantní průkaznost hypotézy, ale u všech se hodnoty k hladině významnosti blíží. I když v takové situaci nemáme statisticky prokazatelné výsledky, máme novou interpretační základnu pro odhalení trendů, možných zákonitostí a důvod k další podrobnější analýze nebo ke zpřesňování měřicích nástrojů.

Jestliže hodnoty vstupních parametrů překročí rozměry tabulky, použijeme aproximace, kterou dává lemma 3 (viz část 5):

a) rozhodneme se buď pro Gram-Charlierovu nabo Camp-Paulsonovu metodu aproximace;
b) spočteme Y;
c) v tabulce distribuční funkce standardizovaného normálního rozložení (viz např. Janko [1958], Bolšev-Smirnov [1965], Kelley [1966]) nalezneme hodnotu $P = \Phi(Y)$, která znamená požadovanou pravděpodobnost pro rozhodnutí.

Tabulka B obsahuje základní údaje o této funkci pro rychlou orientaci v praxi. (Domníváme se, že ve většině sociologických analýz bude tato tabulka postačující.)

Krok 6:

Aplikujeme pravidlo (9) určené v kroku 1:
buď přijmeme hypotézu, nebo přijmeme její negaci, nebo nemůžeme o dané hypotéze provést žádný závěr.

Z uvedeného postupu (krok 5) vidíme, že hypotézy H_1 a H_4 se prověřují zcela stejným způsobem; $H_1 : p_+ > \frac{1}{2}$ znamená totéž co $p_+ > p_0 + p_-$, tj. kladná kategorie má vyšší obsazení než zbývající dvě dohromady.

H_4 prověřujeme tak, že po vynechání prostřední, neutrální kategorie „0", prověřujeme H_1 pro nové pravděpodobnostní obsazení

$$p'_+ = \frac{p_+}{1 - p_0} > p'_- = \frac{p_-}{1 - p_0};$$

protože součet nových parametrů $p'_+ + p'_- = 1$, je tato hypotéza stejná jako $H : p'_+ > \frac{1}{2}$.

Postup pochopitelně můžeme použít na kteroukoli hypotézu $H_1 - H_9$, a to prostou záměnou četností a použitím těch četností z trojice (N_+, N_0, N_-), které dané hypotéze odpovídají.

Pro rozhodování pomocí normální aproximace nemusíme nutně vyhledávat hodnotu funkce $\Phi(Y)$; stačí znát hodnotu Y a kvantily normální distribuční funkce Y^*. V tabulce C uvádíme nejdůležitější hodnoty pro taková porovnání. Pak přijmeme H, jestliže $Y \geqq Y^*$ (což je ekvivalentní $P \geqq \gamma$, kde $p = \Phi(Y^*)$).

B. Metoda odhadu parametrů

Metoda odhadu parametrů je založena na aposteriorních parametrech (N_+, N_0, N_-). Pro $i =$ „+", „0", „−" platí:

(10) odhad $p_i : \tilde{p}_i = N_i/N$

(11) variace odhadu: $\operatorname{var} \tilde{p}_i = \dfrac{N_i (N - N_i)}{N^2 (N + 1)}$

(12) směrodatná chyba: $\sigma_i = \sqrt{var\, \tilde{p}_i} = \sqrt{\dfrac{N_i (N - N_i)}{N^2 (N + 1)}}$

(13) kovariance dvou odhadů: $\operatorname{cov}(\tilde{p}_i, \tilde{p}_j) = -\dfrac{N_i N_j}{N^2 (N + 1)}$ (pro $i \neq j$)

Vzorce (10) určují nejlepší bayesovský *bodový odhad*. Vzorce (11) — (13) mohou být aplikovány pro *konfidenční intervaly*, pokud N_i jsou natolik velké, abychom aposteriorní rozložení mohli přibližně aproximovat normálním rozložením (přibližně, je-li

Tabulka A: Aposteriorní pravděpodobnosti $P(p_+ > p_-/N_+, N_-)$, $P(p_+ > \frac{1}{2}/N_+, N - N_+)$

	1	2	3	4	5
1	,50000000	,25000000	,12500000	,06250000	,03125000
2	,75000000	,50000000	,31250000	,18750000	,10937500
3	,87500000	,68750000	,50000000	,34375000	,22656250
4	,93750000	,81250000	,65625000	,50000000	,36328125
5	,96875000	,89062500	,77343750	,63671875	,50000000
6	,98437500	,93750000	,85546875	,74609375	,62304688
7	,99218750	,96484375	,91015625	,82812500	,72558594
8	,99609375	,98046875	,94531250	,88671875	,80615234
9	,99804688	,98925781	,96728516	,92700195	,86657715
10	,99902344	,99414063	,98071289	,95385742	,91021729
11	,99951172	,99682617	,98876953	,97131348	,94076538
12	,99975586	,99829102	,99353027	,98242188	,96159363
13	,99987793	,99908447	,99630737	,98936462	,97547913
14	,99993896	,99951172	,99790955	,99363708	,98455811
15	,99996948	,99974060	,99882507	,99623108	,99039459
16	,99998474	,99986267	,99934387	,99778748	,99409103
17	,99999237	,99992752	,99963570	,99871159	,99640131
18	,99999619	,99996185	,99979877	,99925518	,99782825
19	,99999809	,99997997	,99988937	,99957228	,99870026
20	,99999905	,99998951	,99993944	,99975586	,99922806
21	,99999952	,99999452	,99996698	,99986142	,99954474
22	,99999976	,99999714	,99998206	,99992174	,99973324
23	,99999988	,99999851	,99999028	,99995601	,99984463
24	,99999994	,99999923	,99999475	,99997538	,99991000
25	,99999997	,99999960	,99999717	,99998628	,99994814
26	,99999999	,99999979	,99999848	,99999239	,99997026
27	1,00000000	,99999989	,99999919	,99999578	,99998302
28	1,00000000	,99999995	,99999956	,99999768	
29	1,00000000	,99999997	,99999977		
30	1,00000000	,99999999			
31	1,00000000				

–	6	7	8	9	10
1	,01562500	,00781250	00390625	,00195313	,00097656
2	,06250000	,03515625	01953125	,01074219	,00585938
3	,14453125	,08984375	,05468750	,03271484	,01928711
4	,25390625	,17187500	,11328125	,07299805	,04614258
5	,37695313	,27441406	,19384766	,13342285	,08978271
6	,50000000	,38720703	,29052734	,21197510	,15087891
7	,61279297	,50000000	,39526367	,30361938	,22724915
8	,70947266	,60473633	,50000000	,40180969	,31452942
9	,78802490	,69638062	,59819031	,50000000	,40726471
10	,84912109	,77275085	,68547058	,59273529	,50000000
11	,89494324	,83384705	,75965881	,67619705	,58809853
12	,92826843	,88105774	,82035828	,74827766	,66818810
13	,95187378	,91646576	,86841202	,80834484	,73826647
14	,96821594	,94234085	,90537643	,85686064	,79756356
15	,97930527	,96082306	,93309975	,89498020	,84627188
16	,98669815	,97376060	,95343018	,92420519	,88523853
17	,99154973	,98265517	,96804267	,94612393	,91568123
18	,99468899	,98867208	,97835737	,96224065	,93896094
19	,99669462	,99268335	,98552037	,97388051	,95642072
20	,99796134	,99532235	,99042135	,98215093	,96928583
21	,99875304	,99703769	,99372952	,98794023	,97861302
22	,99924314	,99814042	,99593497	,99193760	,98527532
23	,99954388	,99884215	,99738856	,99466308	
24	,99972695	,99928454	,99833656		
25	,99983754	,99956104			
26	,99990391				

Tabulka A/2

–	11	12	13	14	15
1	,00048828	,00024414	,00012207	,00006104	,00003052
2	,00317383	,00170898	,00091553	,00048828	,00025940
3	,01123047	,00646973	,00369263	,00209045	,00117493
4	,02868652	,01757813	,01063538	,00636292	,00376892
5	,05923462	,03840637	,02452087	,01544189	,00960541
6	,10505676	,07173157	,04812622	,03178406	,02069473
7	.16615295	,11894226	,08353424	,05765915	.03917694
8	,24034119	,17964172	,13158798	,09462357	,06690025
9	,32380295	,25172234	,19165516	,14313936	.10501981
10	,41190147	,33181191	,26173353	,20243645	.15372813
11	,50000000	,41590596	,33881974	,27062809	,21217811
12	,58409405	,50000000	,41940987	,34501898	,27859855
13	,66118026	,58059013	,50000000	,42250949	,35055402
14	,72937191	,65498102	,57749051	,50000000	,42527701
15	,78782189	,72140146	,64944598	,57472299	,50000000
16	,83653021	,77896583	,71420591	,64446445	,57223222
17	,87610572	,82753577	,77087084	,70766764	,63994993
18	,90753333	,86753455	,81920269	,76343517	
19	,93197703	,89975578	,85947924		
20	,95063142	,92519361			
21	,96462223				

–	16	17	18	19	20
1	,00001526	,00000763	,00000381	,00000191	,00000095
2	,00013733	,00007248	,00003815	,00002003	,00001049
3	,00065613	,00036430	,00020123	,00011063	,00006056
4	,00221252	,00128841	,00074482	,00042772	,00024414
5	,00590897	,00359869	,00217175	,00129974	,00077194
6	,01330185	,00845027	,00531101	,00330538	,00203866
7	,02623940	,01734483	,01132792	,00731665	,00467765
8	,04656982	,03195733	,02164263	,01447964	,00957865
9	,07579482	,05387607	,03775935	,02611949	,01784907
10	,11476147	,08431877	,06103906	,04357928	,03071417
11	,16346979	,12389428	,09246667	,06802297	,04936857
12	,22103417	,17246422	,13246545	,10024421	,07480639
13	,28579409	,22912916	,18079730	,14052076	
14	,35553555	,29233235	,23656483		
15	42776777	,36005006			
16	,50000000				

–	21	22	23	24	25
1	,00000048	,00000024	,00000012	,00000006	,00000003
2	,00000548	,00000286	,00000149	,00000077	,00000040
3	,00003302	,00001794	,00000972	,00000525	,00000283
4	,00013858	,00007826	,00004399	,00002462	,00001372
5	,00045526	,00026676	,00015537	,00009000	,00005186
6	,00124696	,00075686	,00045612	,00027306	,00016246
7	.00296231	,00185958	,00115785	,00071545	,00043896
8	,00627048	,00406503	,00261144	,00166345	
9	,01205977	,00806240	,00533692		
10	,02138697	,01472469			
11	,03537777				

Tabulka A/3

	26	27	28	29	30
1	,00000001	,00000000	,00000000	,00000000	,00000000
2	,00000021	,00000011	,00000006	,00000063	,00000001
3	,00000152	,00000081	,00000043	,00000023	
4	,00000762	,00000422	,00000232		
5	,00002974	,00001698			
6	,00009610				

Tabulka B: Vybrané hodnoty $\Phi(Y)$ $(\Phi(-Y) = 1 - \Phi(Y))$

Y	$\Phi(Y)$	Y	$\Phi(Y)$	Y	$\Phi(Y)$	Y	$\Phi(Y)$
0	0,5000	1,0	0,8413	2,0	0,9773	3,0	0,9987
0,1	0,5398	1,1	0,8643	2,1	0,9821	3,1	0,9990
0,2	0,5793	1,2	0,8849	2,2	0,9861	3,2	0,9993
0,3	0,6179	1,3	0,9032	2,3	0,9893	3,3	0,9995
0,4	0,6554	1,4	0,9192	2,4	0,9918	3,4	0,9997
0,5	0,6915	1,5	0,9332	2,5	0,9938	3,5	0,9998
0,6	0,7558	1,6	0,9452	2,6	0,9953	3,6	0,9998
0,7	0,7580	1,7	0,9554	2,7	0,9965	3,7	0,9999
0,8	0,7881	1,8	0,9641	2,8	0,9974	3,8	0,9999
0,9	0,8159	1,9	0,9713	2,9	9,9981	3,9	0,9999

Tabulka C: Vybrané kvantily normální distribuční funkce pro rozhodování o překročení prahové hodnoty

γ	Y^*	Y^* s velkou přesností	γ	Y^*	Y^* s velkou přesností
0,5	0,000	0,0000 0000	0,995	2,576	2,5758 2930
0,6	0,253	0,2533 4710	0,999	3,090	3,0902 3231
0,7	0,524	0,5244 0051	0,9995	3,291	3,2905 2673
0,75	0,674	0,6744 8975	0,9999	3,719	3,7190 1649
0,80	0,842	0,8416 2123	0,9999 5	3,891	3,8905 919
0,85	1,036	1,0364 3339	0,9999 9	4,265	4,2648 908
0,90	1,282	1,2815 5157	0,9999 95	4,417	4,4171 734
0,95	1,645	1,6448 5363	0,9999 99	4,753	4,7534 243
0,975	1,960	1,9599 6398	0,9999 995	4,892	4,8916 385
0,99	2,326	2,3263 4787	0,9999 999	5,199	5,1993 376

Řádky odpovídají hodnotám prvního parametru (N_+), sloupce odpovídají hodnotám druhého parametru (N_- resp. $N - N_+$).

$N_i > 5$ a $N - N_i > 5$, chceme-li zjistit aproximativní konfidenční interval [8] pro p_i).

Konfidenční intervaly pro menší četnosti nutno hledat pomocí tabulek tzv. *neúplné beta funkce* (či odvozeně pomocí distribuční funkce binomického rozložení). Tento obtížný postup zde neuvádíme.

4. Příklad postupu rozhodování o hypotézách majoritní pravděpodobnosti a asymetrie a odhadu parametrů

Ilustrace postupu testování různých hypotéz v této části má sloužit jako vodítko pro uživatele. Příklad je vzat z malé skupiny (dílny) o 25 respondentech. Názory na vliv zavádění automatizace byly zjišťovány mj. otázkami:
„Jak podle Vašeho názoru ovlivňuje automatizace
a) pracovní spokojenost (zvyšuje, nemá vliv, snižuje)?
b) možnosti pro mladé dělníky (zvyšuje, nemá vliv, snižuje)?"

Při rozhodování o hypotézách budeme postupovat podle kroků popsaných v předcházející části. Budou nás zajímat hypotézy o existenci majoritní (nadpoloviční) pravděpodobnosti, tj. názoru zastávaného více než padesáti procenty osob, a dále hypotéza o asymetrii, tj. převaha názoru na „zvyšování" oproti „snižování" nebo naopak.

Postup ukážeme souběžně pro oba uvedené dotazy:

Krok 1:

Prahovou hodnotu pro přijetí rozhodnutí γ položíme rovnu $\gamma = 0.999$.

Krok 2:

V pilotážním šetření bylo dotazováno několik málo nahodile vybraných dělníků. Výsledky nemají žádnou analytickou cenu, ale můžeme je použít jako základ pro určení parametrů apriorního rozložení. Byla však položena pouze otázka o vlivu automatizace na pracovní spokojenost (otázka a)). Z pěti rozhovorů byli tři odpovědi v kategorii „+" a dvě v kategorii „0". Dosadíme proto $m_+ = 3$, $m_0 = 2$; vzhledem k tomu, že kategorie „−" nebyla obsazena, dosadíme automaticky $m_- = 1$.

Druhá otázka (b) o vlivu automatizace na „možnosti pro mladé dělníky" nebyla v pilotáži položena. Nemáme zformulován ani žádný názor o odpovědích respondentů. Proto použijeme Bayesův postulát „o stejnoměrné neznalosti", tj. přiřadíme apriori všem trojicím parametrů stejné váhy; toho docílíme tak, že položíme $m_+ = m_0 = m_- = 1$.

Krok 3:

Získání a sumarizace empirických dat — řízený rozhovor a třídění dat.

Krok 4:

Výpočet aposteriorních parametrů je patrný z tabulek 1 a 2. V případech některých hypotéz je hned na první pohled vidět, že nemá význam testování provádět. Uvádíme je zde jen pro ilustraci postupu.

V tabulce A vyhledáme apostriorní pravděpodobnosti pro jednotlivé hypotézy. Sumarizace je dána v tabulce 3.

---

[8] Konfidenční interval $100(1 - \alpha)$ % má tvar $p_i \pm z_{\alpha/2}\,\sigma_i$; hodnoty koeficientu $z_{\alpha/2}$ (což je horní $\alpha/2$ kvantil standardního normálního rozložení) nalezneme např. v tabulce C v práci Řehák, Řeková [1978]).

Tabulka 1: Názory na ovlivnění pracovní spokojenosti

Kategorie	Zhoršuje (−)	Zůstává stejná (0)	Zvyšuje (+)	Součet
Apriorní parametry $/m_i/$	1	2	3	6
Četnosti $/n_i/$	5	8	12	25
Aposteriorní parametry $/N_i/$	6	10	15	31

Tabulka 2: Názory na ovlivnění možností pro mladé dělníky

Kategorie	Zhoršuje (−)	Zůstává stejné (0)	Zvyšuje (+)	Součet
Apriorní parametry $/m_i/$	1	1	1	3
Četnosti $/n_i/$	3	1	20	24
Aposteriorní parametry $/N_i/$	4	2	21	27

K dané hypotéze (1. sloupec) nejprve určíme dvojici parametrů pro vyhledávání aposteriorní pravděpodobnosti (v tab. 3 je ve druhém sloupci uvedena obecně symbolicky, ve třetím sloupci pak jsou dány konkrétní hodnoty parametrů odvozené z tabulek 1 a 2). První parametr vystupuje jako číslo řádku tabulky A (čtvrtý sloupec) druhý parametr jako číslo sloupce tabulky A (viz pátý sloupec v tab. 3). V posledním sloupci je uvedena aposteriorní posloupnost, na jejímž základě rozhodujeme o příslušné hypotéze; najdeme ji v tabulce A v daném řádku a sloupci.

Na vybraných případech z tabulky 3 ukážeme oba způsoby aproximace aposteriorních pravděpodobností normálním rozložením uvedené v lemě 3 v části 5. V tabulce 4 a 5 je $\Phi = P(H)$. Aposteriorní parametry značíme N_1 a N_2.

A) *Aproximace Gram-Charlierovým rozvojem*

Tabulka 4 ukazuje tři kroky postupu aproximace: a) určení parametrů (sloupec 1), b) výpočet hodnoty Y (sloupec 2), c) nalezení hodnoty standardní normální distribuční funkce, [9] která aproximuje hledanou pravděpodobnost P. Ve čtvrtém sloupci uvádíme rozdíl mezi aproximovanou hodnotou a přesnou hodnotou získanou z tabulky A.

[9] Hodnoty funkce Φ byly získány z tabulek Janko [1958]. Proto byl výraz ve druhém sloupci vypočten na dvě desetinná místa (tedy značně nepřesně). Nebyla prováděná interpolace v tabulkách. Existují ovšem tabulky daleko přesnější, zde jsme však volili běžný a rutinní postup s nejsnáze dostupnými tabulkami u nás. Gram-Charlierova transformace je pro náš speciální případ ($p = 1/2$) totožná s původní Laplaceovou transformací, umožňuje však přesnější odhad maximální absolutní chyby. Přesnost je oproti údaji v lemě 3 ještě dále zvýšena, neboť pro $p = 1/2$ je binomické rozložení symetrické a k normálnímu se blíží velmi rychle.

Tabulka 3: Určení aposteriorních pravděpodobností vybraných hypotéz pro data tabulky 1 a tabulky 2

Hypotéza	Parametry aposteriorních rozložení	Hodnoty parametrů	Řádek tabulkyA	Sloupec tabulkyA	Aposteriorní pravděpodobnost
„Vliv automatizace na pracovní spokojenost"					
$p_+ > \frac{1}{2}$	$(N_+, N - N_+)$	(15,16)	15	16	.4278
$p_0 > \frac{1}{2}$	$(N_0, N - N_0)$	(10,21)	10	21	.0214
$p_- > \frac{1}{2}$	$(N_-, N - N_-)$	(6,25)	6	25	.0002
$p_+ > p_-$	(N_+, N_-)	(15,6)	15	6	.9793
„Vliv automatizace na možnosti pro mladé dělníky"					
$p_+ > \frac{1}{2}$	$(N_+, N - N_+)$	(21,6)	21	6	.999
$p_0 > \frac{1}{2}$	$(N_0, N - N_0)$	(2,25)	2	25	.0000004
$p_- > \frac{1}{2}$	$(N_-, N - N_-)$	(4,23)	4	23	.000044
$p_+ > p_-$	(N_+, N_-)	(21,4)	21	4	.99986

B) *Aproximace Camp-Paulsonova*

V tabulce 5 ilustrujeme nejpřesnější praktickou aproximaci binomického rozložení normálním. Postup je analogický tabulce 4; výpočet hodnoty Y však je proveden ve třech krocích (sloupce 2—4).

Tato transformace je přesnější právě u extrémních pravděpodobností.

Kdybychom chtěli v našem případě aplikovat (ekvivalentně) tabulku C (nejsou-li např. po ruce podrobnější tabulky funkce Φ), pak pro náš případ postupujeme jednoduše:

a) pro $\gamma = 0.999$ je $Y^*_{0.999} = 3.090$
b) pro $1 - \gamma = 0.001$ je $Y^*_{0.001} = -Y^*_{0.999} = -3.090$

Proto platí, že $P\,(=\Phi) \geqq .999$, právě když $Y \geqq 3.090$ a $P \leqq .001$, právě když $Y \leqq -3.090$.

Pro data v tabulce 5 je $Y = X\,(3\sqrt{Z})^{-1}$, a tudíž máme evidenci pro přijetí hypotézy $p_- < \frac{1}{2}$ pro „spokojenost" a hypotézy $p_0 < \frac{1}{2}$, $p_- < \frac{1}{2}$ pro „možnosti pro mladé dělníky."

Pomocí aproximace tabulky 4 docházíme ke stejnému výsledku. V obou případech se výsledky liší od přesných hodnot. V tabulce 3 přijímáme pro znak „možnosti pro mladé dělníky" i hypotézu $p_+ > \frac{1}{2}$.Tato ztráta informace je tedy způsobena numericky (!), použitím aproximačních formulí.

Implikace jsou dvě:

— nepoužívejme nikdy aproximaci tam, kde máme k dispozici jednoduchou a přímou metodu [10];

Tabulka 4: Postup výpočtu aproximace Gram-Charlierovým rozvojem (viz lema 3)

Parametry $(N_1, N - N_1)$	$\frac{2N_1 - N}{N - 1}$	Φ	Absolutní chyba*)
(15,16)	— 0,18	0,42858	0,00081
(10,21)	— 2,01	0,02222	0,00083
(6,25)	— 3,47	0,0002602	0,0000977
(21,6)	2,94	0,998359	0,000394
(2,25)	— 4,51	0,000003241	0,000002841
(4,23)	— 3,73	0,00009574	0,00005175

*) *absolutní chyba* = | *přesná hodnota* — *aproximovaná hodnota* |

Tabulka 5: Postup Camp-Paulsonovy transformace (viz lema 3)

Parametry $(N_1, N - N_1)$	X	Z	$X\,(3\sqrt{Z})^{-1}$	Φ	Absolutní chyba*)
(15,16)	— 0,1943	0,1264	— 0,182	0,42858	0,00081
(10,21)	— 2,0024	0,1086	— 2,025	0,02118	0,000207
(6,25)	— 3,4706	0,0688	— 4,410	0,000005169	0,000157
(21,6)	4,7590	0,2764	3,017	0,998736	0,000017
(2,25)	— 5,2975	0,1328	— 4,845	0,0000006173	0,00000022
(4,23)	— 4,0724	0,1214	— 3,896	0,00004810	0,00000411

*) *absolutní chyba* = | *přesná hodnota* — *aproximovaná hodnota* |

[10] To platí pro celou statistiku. Známý příklad je použití chí-kvadrátu místo přesného Fischerova testu u málo obsazených kontingenčních tabulek 2×2 nebo použití normálního z-testu místo Studentova t-testu v případě testování hodnoty průměru normálního rozložení při malém počtu pozorování.

Tabulka 6: Postup odhadu pravděpodobností p_+, p_0, p_- pro údaje z tabulek 1 a 2 („Spokojenost" a „Možnosti pro mladé")

p_i	hodnota parametru N_i	odhad p_i	směrodatná odchylka odhadu	přibližný interval spolehlivosti[11] s 95%
„Vliv automatizace na pracovní spokojenost"				
p_+	15	.484	.088	(.312, .656)
p_0	10	.323	.082	(.162, .484)
p_-	6	.194	.070	(.057, .331)
„Vliv automatizace na možnosti pro pladé dělníky"				
p_+	21	.778	.079	(.623, .933)
p_0	2	.074	.049	(nelze použít)
p_-	4	.148	.067	(nelze použít)

— je výhodnější pracovat s hodnotou P; i když Y nepřekročí odpovídající hranici. P může být tak blízko ke zvolenému γ, že hypotézu stejně přijmeme, přestože $P < \gamma$.

Krok 5:

Rozhodnutí pro $\gamma = 0.999$; přijímáme tyto výroky o parametrech:

a) znak „spokojenost": $\overline{H}_3 : p_- < \frac{1}{2}$
b) znak „možnosti pro mladé": $H_1 : p_+ > \frac{1}{2}$
$\overline{H}_2 : p_0 < \frac{1}{2}$
$\overline{H}_3 : p_- < \frac{1}{2}$
$H_4 : p_+ > p_-$

V tomto případě však jsou výroky $\overline{H}_2$, $\overline{H}_3$ a H_4 důsledkem výroku H_1. Můžeme tedy říci, že názor o zhoršení spokojenosti nedosáhne 50 %, zatímco názor o zvýšení možností pro mladé dělníky je majoritní (přesahuje 50 %).

Kdybychom na přijetí hypotézy nebyli tak přísní, mohli bychom přijmout i hypotézu asymetrie v případě znaku spokojenosti, tj. názor na zlepšení převažuje nad názorem na zhoršení.

Odhad parametrů provádíme jednoduchým dosazením do vzorců (10) a (12); postup je naznačen v tabulce 6.

5. Statistický model a teoretické základy metody

Tato část je věnována především matematicky se orientujícímu čtenáři: Znak S má tři parametry, které zde budeme značit p_1, p_2, p_3 (s korespondencí $p_1 = p_-$, $p_2 = p_0$ $p_3 = p_+$). Předpokládáme, že data vznikají jako n- násobné nezávislé realizace náhodného pokusu (n je dáno) a že hodnoty těchto realizací, S_1, S_2, ..., S_n mohou

[11] Přibližný interval spočítáme jako: odhad $p_i \pm 1{,}96 \times$ *směrodatná odchylka* odhadu.

být sumarizovány pomocí výsledných absolutních četností n_i (= počet pozorování v kategorii), $n_1 + n_2 + n_3 = n$. Za tohoto předpokladu má trojice (n_1, n_2, n_3) *multinomické rozložení pravděpodobností.*

$$P(n_1, n_2, n_3/p_1, p_2, p_3) = \frac{n!}{n_1!\, n_2!\, n_3!}\, p_1^{n_1}\, p_2^{n_2}\, p_3^{n_3} \tag{14}$$

Pro bayesovskou inferenci a parametrech p_i volíme konjugované (přirozeně sdružené) apriorní rozložení, které dává dostatečně širokou třídu dvourozměrných Dirichletových rozložení reprezentující pro praktické úlohy dostatečně široké spektrum tvarů.

Lema 1. Budiž rozložení trojice (n_1, n_2, n_3) multinomické, dané vzorcem (14). Apriorní rozložení nahodných veličin (p_1, p_2, p_3) nechť je dvourozměrné Dirichletovo rozložení s hustotou

$$f_D(p_1, p_2, p_3/m_1, m_2, m_3) = \frac{\Gamma(m_1 + m_2 + m_3)}{\Gamma(m_1)\,\Gamma(m_2)\,\Gamma(m_3)}\, p_1^{m_1-1}\, p_2^{m_2-1}\, p_3^{m_3-1} \tag{15}$$

na simplexu Q (viz (1)), $f_D = 0$ mimo Q. Přitom $\Gamma(x)$ je gamma funkce a m_i jsou reálné kladné parametry. Aposteriorní rozložení náhodných veličin (p_1, p_2, p_3) je opět Dirichletovo s parametry N_1, N_2, N_3, kde $N_i = n_i + m_i$ (pro $i = 1, 2, 3$), tedy

$$f_D(p_1, p_2, p_3/N_1, N_2, N_3) = \frac{\Gamma(N_1 + N_2 + N_3)}{\Gamma(N_1)\,\Gamma(N_2)\,\Gamma(N_3)} \cdot p_1^{N_1-1}\, p_2^{N_2-1}\, p_3^{N_1-1} \tag{16}$$

Základní vlastnosti Dirichletova rozložení popisuje S. S. Wilks [Wilks 1962]. Očekávané hodnoty, variace a covariance jsou dány formulemi (10)—(13), které je možno použít pro úlohu aposteriorního odhadu parametrů.

Pro testování majoritní pravděpodobnosti $H_1 : p_1 > \frac{1}{2}$ použijeme lemu 2.

Lema 2 (znaménkový test pro hypotézu existence majoritní pravděpodobnosti). Za předpokladů lemy 1 a vztahu (14) platí pro aposteriorní pravděpodobnosti

$$P(p_1 < \tfrac{1}{2}/N_1, N_2, N_3) = (\tfrac{1}{2})^{N-1} \sum_{i=N_1}^{N-1} \binom{N-1}{i} \tag{17}$$

$$P(p_1 > \tfrac{1}{2}/N_1, N_2, N_3) = (\tfrac{1}{2})^{N-1} \sum_{i=0}^{N_1-1} \binom{N-1}{i} \tag{18}$$

$$= P_B(X < N_1/\tfrac{1}{2}, N-1)$$

$$= F_{2(N-N_1),\, 2N_1}\left(\frac{N_1}{N-N_1}\right) \tag{19}$$

kde P_B je distribuční funkce binomického rozložení s parametry $\frac{1}{2}$ a $N - 1$ a $F_{g,h}(z)$ je distribuční funkce Fisher-Snedecorova F — rozložení s (g, h) stupni volnosti.

Pravděpodobnosti jsou dány v tabulce A pro $N \leqq 32$. Mimo rozsah tabulky můžeme buď použít speciálních tabulek binomického resp. F — rozložení, nebo některou z mnoha možných normálních aproximací, z nichž základní a nejpraktičtější uvádíme v lemě 3.

Lema 3. Binomickou pravděpodobnost (18) můžeme aproximovat některou z následujících metod (značíme $\Phi(x)$ kumulativní distribuční funkci normálního standardního rozložení $N(0,1)$):

A. Gram-Charlierův rozvoj s korekcemi pro spojitost

$$P(H_1/N_1, N - N_1) \doteq \Phi(Y), \tag{20}$$

kde $Y = \dfrac{2N_1 - N}{\sqrt{N-1}}$

absolutní chyba $< 0{,}112\,(N-1)^{-\frac{1}{2}}$

B. Camp-Paulsonova aproximace

(21) $P(H_1/N_1, N - N_1) \doteq \Phi(X(3\sqrt{Z})^{-1})$,

kde $X = \left(9 - \frac{1}{N_1}\right)\left(\frac{N_1}{N - N_1}\right)^{\frac{1}{3}} - 9 + \frac{1}{N - N_1}$

$$Z = \frac{1}{N_1}\left(\frac{N_1}{N - N_1}\right)^{\frac{2}{3}} + \frac{1}{N - N_1}$$

absolutní chyba $< 0{,}014\,(N-1)^{-\frac{1}{2}}$

Odkazy na jednotlivé metody zde neuvádíme a odkazujeme čtenáře na přehlednou práci [Johnson, Kotz 1969].

Camp-Paulsonova transformace je přesnější než Gram-Charlierova, je i přesnější než další běžné a známé transformace, například transformace arcsinová či Freeman-Tukeyho zlepšení arcsinové transformace.

Camp-Paulsonova transformace je zdánlivě komplikovaná, avšak za pomoci moderních kalkulaček je výpočet velice rychlý. Pro praxi tuto metodu doporučujeme, i když pro velké soubory je metoda Gram-Charlierova postačující.

Po výpočtu hodnoty Y (resp. $X\ (3\sqrt{Z})^{-1}$) jako argumentu funkce, nalezneme hodnotu funkce Φ v tabulkách normální distribuční funkce s průměrem 0 a rozptylem 1. Tyto tabulky jsou v téměř každé učebnici statistiky a ve statistických tabulkách (např. Janko [1958]). Základní hodnoty funkce viz tabulka B.

Věta (znaménkový test pro hypotézu asymetrie) Za předpokladů a značení lemy 1 a vztahu (14) s empirickými četnostmi n_1, n_2, n_3 platí pro aposteriorní pravděpodobnosti

$$(22)\quad P(p_1 > p_3/N_1, N_2, N_3) = \sum_{k=0}^{N_1-1}\binom{N_1 + N_3 - k - 2}{N_3 - 1}\left(\frac{1}{2}\right)^{N_1 + N_3 - 1 - k} =$$

$$= \sum_{j=0}^{N_1-1}\binom{N_3 - 1 + j}{N_3 - 1}\left(\frac{1}{2}\right)^{N_3 + j} =$$

$$(23)\quad = P_{nb}(X \leqq N_1 - 1/\tfrac{1}{2}, N_3 - 1) =$$

$$(24)\quad = P_B(X \geqq N_3/p = \tfrac{1}{2}, M - 1) =$$

$$(25)\quad = (\tfrac{1}{2})^{M-1}\sum_{k=N_3}^{M-1}\binom{M-1}{k},$$

kde $M = N_1 + N_3$
P_{nb} = distribuční funkce negativně binomického rozložení
P_B = distribuční funkce binomického rozložení.

Důkaz:

K důkazu věty použijeme základních vlastností Dirichletovy hustoty, neúplné beta funkce a negativního binomického rozložení.

Za platnosti předpokladů je apriorní hustota rozložení trojice náhodných veličin p_1, p_2, p_3 dvourozměrné Dirichletovo rozložení s parametry N_1, N_2, N_3 dané vztahem (8), $(N_i = m_i + n_i)$.

Pak

$$(34)\quad P = P(p_1 > p_3/N_1, N_2, N_3) = \int\limits_{\substack{p_1 > p_3 \\ p_1 > 1-p_3}} \int\int \frac{\Gamma(N)}{\Gamma(N_1)\,\Gamma(N_2)\,\Gamma(N_3)}\, p_1^{N_1-1}\, p_2^{N_2-1}\, p_3^{N_3-1}\, dp_1\, dp_2\, dp_3$$

Postup je založen na integraci výrazu (34) s využitím základních vztahů pro neúplnou beta funkci

$$P = \frac{\Gamma(N)}{\Gamma(N_1)\,\Gamma(N_2)\,\Gamma(N_3)} \int_0^{\frac{1}{2}} p_3^{N_3-1} \int_{p_3}^{1-p_3} p_1^{N_1-1}\, (1-p_1-p_3)^{N_2-1} dp_1\, dp_3$$

Označíme $\int\limits_{p_3}^{1-p_3} p_1^{M_1-1}\, (1-p_1-p_3)^{M_2-1}\, dp_1 = I_{M_1-1,\, M_2-1}$

Integrací per partes a postupným dosazením dostaneme

$$I_{N_1-1,\, N_2-1} = \frac{1}{N_2}\, p_3^{N_1-1}\, (1-2p_3)^{N_2} + \frac{N_1-1}{N_2} I_{N_1-2,\, N_2} =$$

$$= \sum \frac{[N_1-1]_k}{[N_2+k-1]_{k+1}}\, p_3^{N_1-1-k}\, (1-2p_3)^{N_2+k}$$

kde $[M]_k = M\,(M-1)\ldots(M-k+1)$
$[M]_0 = 1$

Použitím výsledku[12]

$$\int_0^{\frac{1}{2}} p_3^{N_1+N_3-2-k}\, (1-2p_3)^{N_2+k}\, dp_3 = \text{(per partes)}$$

$$= (\tfrac{1}{2})^{N_1+N_3-1-k}\, .\, Be\,(N_1+N_3-1-k,\, N_2+1+k) =$$

$$= \frac{(N_1+N_3-2-k)!\,(N_2+k)!}{(N-1)!} \left(\frac{1}{2}\right)^{N_1+N_3-1-k}$$

dostaneme dále[13]

$$\int_0^{\frac{1}{2}} p_3^{N_3-1} I_{N_1-1,\, N_2-1}\, dp_3 = \sum \int_0^{\frac{1}{2}} \frac{[N_1-1]_k}{[N_2+k]_{k+1}} \,.\, p_3^{N_1+N_3-k-2} \,.\, (1-2p_3)^{N_2+k}\, dp_3 =$$

$$= \sum \frac{[N_1-1]_k}{[N_2+k]_{k+1}} \cdot \frac{(N_1+N_3-2-k)!\,(N_2+k)!}{(N-1)\,!} \left(\frac{1}{2}\right)^{N_1+N_3-1-k} =$$

$$= \sum_{k=0}^{N_1-1} \frac{(N_1+N_3-2-k)!\,(N_2-1)!\,[N-1]_k}{(N-1)\,!} \left(\frac{1}{2}\right)^{N_1+N_3-1-k}$$

[12] *Be* je značení pro beta funkci.

[13] Následující sumy jsou provedeny od $k=0$ do $k = N_1 - 1$.

Pravděpodobnost P vznikne násobením tohoto výsledku konstantou výrazu (26)

$$P_N = \sum_{k=0}^{N_1-1} \frac{(N_1 + N_3 - 2 - k)!}{(N_1 - 1 - k)!\,(N_3 - 1)!} \left(\frac{1}{2}\right)^{N_1+N_3-1-k} =$$

$$= \sum_{k=0}^{N_1-1} \binom{N_1 + N_3 - 2 - k}{N_3 - 1} \left(\frac{1}{2}\right)^{N_1+N_3-1-k} =$$

$$= \sum_{t=0}^{N_1-1} \binom{N_3 + t - 1}{t} \left(\frac{1}{2}\right)^{N_3+t}$$

Pravděpodobnost P_N je distribuční funkce negativně binomického rozložení s parametry $p = \frac{1}{2}$ a N_3. Je to tedy pravděpodobnost, že při náhodném opakování jevu s pravděpodobností $p = \frac{1}{2}$ potřebujeme nejvýše $N_1 + N_3 - 1$ pokusů, abychom dosáhli N_3 úspěchů.

Vztahy (24) a (25) plynou z pravděpodobnostních identit vztahujících negativně binomické a binomické rozložení.

Poznámky k aplikaci věty 1

1. Aposteriorní pravděpodobnost H_4 je pro parametry N_1, N_2, N_3 stejná jako pravděpodobnost H_1 pro parametry N_1, 0, N_3. Pro rozhodnutí o hypotéze H_4 používáme stejný postup jako pro H_1, pouze redukujeme N o hodnotu N_2, tj. o „počet" odpovědí v neutrální kategorii.
2. Platí zde všechny poznámky u lemy 2 pro situaci, v níž N zaměníme hodnotou $M = N_1 + N_3$. Též aproximace z lemy 3 provádíme zcela obdobně.
3. Aposteriorní pravděpodobnost H_4 můžeme obdobně vyjádřit též distribucí Fisher-Snedecorovou.
4. Je zřejmé, že výsledek lze použít symetricky pro libovolnou z hypotéz $H_4 - H_9$, a to tak, že prostě zaměníme indexy u příslušných parametrů.
5. Kromě uvedených transformací binomického rozložení lze také použít transformaci inverzního hyperbolického sinu navrženou Anscombem [1948].

Závěry

Znaménkový test, který byl ve stati odvozen na základě bayesovského principu statistické inference, umožňuje tedy neparametrické vyhodnocování jednorozměrných distribucí diskrétního znaménkového znaku. Umožňuje testovat hypotézy asymetrie i existenci majoritní pravděpodobnosti. V případě potřeby může být zobecněn na obecné hypotézy tvaru $p_+ > c$, resp. na hypotézy $0 < p_+ < a$, $a < p_+ < b$, $b < p_+ < 1$, kde a, b jsou předem zvolené konstanty. Použití takovýchto hypotéz je však relativně málo časté, a proto příslušné testy neuvádíme.

Postupy popsané v této stati je ovšem možno použít i pro libovolné nominální znaky: jsou-li $p_1, p_2, \ldots, p_k$ pravděpodobnosti pro 1, 2, ..., k-tou kategorii nominálního znaku a chceme-li zjistit, zda platí hypotéza $p_i > \frac{1}{2}$ pro některé i nebo hypotéza $p_i > p_j$ pro některou dvojici kategorií, aplikujeme uvedený postup pro N_i, $N - N_i$, resp. pro N_i, N_j; tyto parametry vznikají obdobně jako u znaménkových dat: $N_i = m_i + n_i$, m_i charakterizují apriorní názor na pravděpodobnosti obsazení polí a n_i jsou empirické četnosti.

Stejně tak lze aplikovat i postup pro odhad parametrů p_i a jeho směrodatných chyb. Metodu lze použít na porovnávání příslušných kategorií rozšířeného znaménkového znaku jako alternativu pro rozklad chí-kvadrátového kritéria symetrie rozložení, viz [Řehák, Řeháková 1978].

Vhodnost metody a její přednosti i nedostatky ve srovnání s jinými postupy mohou být prověřeny pouze další praktickou zkušeností a kritickým metodologickým zhodnocením používání. Pro velké soubory dat budou výsledky klasického i bayesovského testování většinou shodné. Rozdíly se projeví především u malých datových souborů. A právě zde se může metoda vhodně doplňovat s postupy klasické statistiky. Je to však metoda obtížná právě tím, že při formulování apriorních vah vyžaduje hluboké uvažování zcela nerutinního charakteru, chceme-li ji využít v celé její síle.

Literatura

Anscombe, F. J.: *The Transformation of Poisson, Binomial and Negative Binomial Data.* Biometrika 35, 1948, s. 246–254.

Bolšev, L. N. – Smirnov, N. V.: *Tablicy matematičeskoj statistiki.* Moskva, Nauka 1965.

De Groot, M.: *Optimalnyje statističeskije rešenija.* Moskva, Mir 1974.

Janko, J.: *Statistické tabulky.* Praha, NČSAV 1958.

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

Kelley, T. L.: *Statističeskije tablicy.* VS AN SSSR, Moskva 1966.

Řehák, J.: *K použití subjektivních a objektivních pravděpodobností v lidské praxi.* Filosofický časopis 1974, s. 53–63.

Řehák J. – Řeháková, B.: *Nový přístup ke statistickému vyhodnocení dat znaménkového typu* (přednáška pro Českou otolaryngologickou společnost v Praze), 1972.

Řehák, J. – Řeháková, B.: *Základní statistické testy pro rozložení diskrétních znaménkových dat.* Sociologický časopis 1978, č. 1.

Savage, I. R.: *Statistics: Uncertainty and Behaviour.* Boston, Houghton Mifflin Co 1968.

Schmitt, S. A.: *Measuring Uncertainty: An Elementary Introduction to Bayesian Statistics.* Addison Wesley Publ. Comp., Reading, 1969.

Tománek, R.: *Unavitelnost sluchové a vegetativní rovnováhy* (vliv vegetativní rovnováhy na sluchovou únavu a její ovlivnění atropinem). Praha, FVL 1971.

Tománek, R.: *Ovlivnění sluchové únavy pilokarpinem* (přednáška pro Českou otolaryngologickou společnost v Praze), 1972.

Wilks, S. S.: *Mathematical Statistics.* New York, Wiley 1962.

Резюме

Ржегак Я. – Ржегакова Б.: Критерий знаков по Байесу

В статье выводится подход к оценке данных знакового типа по Байесу. Апостериорное распределение основано на функции правдоподобия модели мультиномического распределения и априорного распределения Дирихлея. Это распределение для p_+, p_o, p_- является основой для статистических выводов об этих параметрах. Дискретная знаковая переменная является трихотомией с положительной, нейтральной и отрицательной категорией; вероятность осуществления в этих категориях обозначается соответственно p_+, p_o, p_-. В статье описаны методы проверки гипотез о существовании преобладающей вероятности (напр., $p_+ > \frac{1}{2}$) и асимметрии ($p_+ > p_-$). Описана и роль оценки.

В первой части статьи приводится общая идея подхода Байеса, которая затем обсуждается в связи с данной проблемой. В следующей части объяснено значение априорного распределения и его параметров. Проверка гипотез для вышеприведенных задач описывается по шагам так, как она проводится на практике. В этой части приведена таблица апостериорных вероятностей (таблица А) и таблица избранных значений кумулятивной функции распределения стандартного нормального распределения (таблица В), далее важнейшие квантилы этого распределения (таблица С). В леме 3 приведены два способа нормальной апроксимации апостериорных вероятностей; апроксимация Кэмп-Паульсона рекомендуется как гораздо более точная.

Разные способы использования метода иллюстрируются на нумерическом примере, который должен облегчить пользование методом. Математическая формулировка метода основана на сводке и применении известных свойств в лемах 1–3 и в теореме, которая дает основной результат для проверки ассиметрии (приведено и ее сокращенное доказательство).

Метод критерия знаков по Байесу является альтернативой классического критерия знаков и других методов анализа знаковых данных, сводка которых содержится в работе Ржегак, Ржегакова (1978).

Summary

Řehák J. — Řeháková B.: A Bayesian Sign Test

In the paper, a Bayesian approach to the discrete sign data is developped. Using the two-dimensional Dirichlet prior distribution with parameters m_+, m_0, m_- and the multinominal model for data behaviour, the posterior distribution of p_+, p_0, p_- is utilized for inference about these unknown quantities. Having denoted probabilities of occurrence in positive, neutral and negative categoriy of the trichotomy representing discrete sign variable as p_+, p_0, p_-, the testing hypotheses $p_+ > \frac{1}{2}$, $p_0 > \frac{1}{2}$, $p_- > \frac{1}{2}$ and comparisons like $p_+ > p_-$ are treated. The estimation problem is handled as well.

The first part introduces the general idea of Bayesian inference: next, the prior distribution and possible meaning of its parameters is shown for the problem. The testing for asymmetry (eg. $p_+ > p_-$) and for the existence of majority probability (eg. $p_+ > \frac{1}{2}$) is exposed in steps for practical use. In this part the table of aposterior probabilities as function of two aposterior parameters (Table A) and selected values of standard normal cumulative distribution function and its selected quantiles (Tables B and C) are presented.

In lemma 3, the normal approximations of posterior probabilities are reviewed and the Camp-Paulson method recommended. The method is illustrated by a numerical example giving a guide to user.

Mathematical formulation consists in a summary and applications of well known statistical properties in lemmas 1 to 3 and in theorem which gives the results necessary for testing asymmetry (the abridged proof is presented).

The method is an alternative to the classical sign test and various other methods for analysis of sign data which were reviewed in Řehák, Řeháková [1978].