# Statistické ověřování reprezentativity: testy dobré shody

**BLANKA ŘEHÁKOVÁ**
Ústav pro výzkum veřejného mínění
při FSÚ, Praha

Otázka reprezentativity je jednou z klíčových otázek kvality sociologické informace a podmínkou upotřebitelnosti získaných výzkumných dat. Oprávněně je tomuto problému věnována velká pozornost, a to jak ve statistice (například rozsáhlá literatura o výběrových šetřeních z konečných populací), tak v sociologické metodologii. O reprezentativitě píše F. Zich[1] [1976]. J. Řehák [1978] provedl podrobnou diskusi pojmu a pokusil se o jeho vymezení[2], z něhož vyplývá řada praktických aspektů, kterými bychom se měli v každém výzkumném šetření zabývat.

Účelem této stati je navázat na uvedené práce a ukázat postupy statistického ověřování reprezentativity, tj. postupy, které ověřují shodu mezi procentním rozdělením daného znaku v celé populaci a ve výběrovém souboru. Tyto postupy patří ke statistickým technikám známým pod názvem „testy dobré shody chí-kvadrát". Pro potřeby sociologické výzkumné praxe rozvíjí článek metody znaménkového schématu odvozené pro speciální úlohu (zkrácené odvození je uvedeno v *Dodatku*).

Metody této stati jsou založeny na předpokladu prostého náhodného výběru, respektive k němu ekvivalentního postupu. Jestliže jsme získali data stratifikovaným nebo kvótním výběrem, musíme ověřovat shodu v každé oblasti zvlášť nebo použít techniky části 4.

## 1. Výběr znaků pro ověřování shody

Statisticky nemůžeme ověřovat reprezentativitu v plném rozsahu, jak ji vymezil J. Řehák [1978], neboť pro většinu dat nemáme kontrolní údaje. V tomto článku půjde pouze o kontrolu rozdělení těch znaků, jejichž populační rozdělení k danému datu je známo. Dat dostupných z ročenek a použitelných pro ověřování reprezentativity je relativně velmi málo. Jsou to například znaky: věk, vzdělání, pohlaví, rodinný stav, velikost místa bydliště, sociální skupina, stupeň ekonomické aktivity,

[1]) „Při výběru objektů zkoumání nám pak většinou jde o to, abychom ze souboru objektů vybrali takové, které reprezentují celou tuto množinu, tj. abychom poznatky získané od části mohli oprávněně extrapolovat na celek. Otázku, nakolik jsou výsledky výzkumné akce reprezentativní, si výzkumník musí klást prakticky vždy, nemá-li zůstat u deskripce jevu a chce-li získané poznatky zobecňovat." [F. Zich 1976, s. 191].

[2]) „Reprezentativnost (reprezentativita) výběrového šetření je stupeň přenositelnosti informace obsažené ve výběrovém souboru na cílovou populaci či její zvolenou část; tato zobecnitelnost je zprostředkována statistickým modelem výběrového uspořádání a znalostí odchylky předpokládaného modelu od jeho realizace. Zahrnuje stupeň obsahového pokrytí úlohy (počet a role těch znaků, jejichž informace je zobecnitelná), stupeň pokrytí statistických vlastností dat (vlastností statistických měr, distribucí a celkové struktury dat) a rozsah pokrytí populace (zahrnutí všech typů jednotek populace do výběru). Je to stupeň zachycení populační variability rozložení všech proměnných výzkumného instrumentu pomocí daného realizovaného výběrového souboru a daného modelu výběrového uspořádání." [J. Řehák 1978, s. 496].

národnost, povolání. Statistické populační údaje ovšem musí odpovídat výzkumnému základnímu souboru a nesmějí být zastaralé (nemůžeme např. procentní údaje o rodinném stavu, týkající se populace osob starších 15 let, brát za základ porovnání pro výzkumnou populaci 15—35 let).

Kontrolovat ovšem můžeme nejen rozdělení jednotlivých znaků, ale, jsou-li dostupné údaje, i rozdělení kombinací znaků, a to i na různých podsouborech. V praxi se dosti často stává, že rozdělení jednotlivých znaků jsou v dobré shodě, ale rozdělení jejich kombinace se už silně odlišuje.

Před hodnocením shody výzkumných výsledků s populačními údaji je nutné velmi pečlivě porovnat obsah kategorií používaných statistickými informačními zdroji s významem, který byl kategoriím přisouzen v daném výzkumu. Jako příklad můžeme uvést ekonomickou aktivitu či skupinu povolání: Jsou ženy na mateřské dovolené či pracující důchodci klasifikováni ve výzkumných instrukcích stejně jako v ročence? Jsou správně klasifikovány takové kategorie zaměstnání, jako zemědělský dělník, mistr, skladník? Jestliže se neshoda pro různé kategorie týká početně obsazených skupin, pak nemůžeme takové kategorie do porovnání zahrnout. Tento kvalitativní aspekt je třeba vždy považovat za základní podmínku pro hodnocení shody dále uvedenými matematicko-statistickými postupy.

Pochopitelně nebudeme ověřovat rozdělení těch znaků, které byly použity jako stratifikační či kvótní, neboť ty jsou kontrolované, proporce zastoupení v jednotlivých kategoriích jsou dány a nevznikají tudíž náhodným procesem. Výkyvy jsou nepatrné a mají vysvětlitelné příčiny. Předepisujeme-li však při výběrovém postupu distribuce jednotlivých kvótních znaků, můžeme kontrolovat, zda jsou reprezentativní rozdělení jejich dvojrozměrných kombinací. To je častá úloha ověřování shody, neboť zde jde o závažný faktor možného porušování reprezentativity.

V ověřování reprezentativity se ovšem nemusíme omezovat na proměnné, které tabelují statistické ročenky vydávané Federálním statistickým úřadem. Podle typu výzkumu to mohou být statistiky kulturní, školské, zdravotnické, sportovní, resortní, podnikové apod. K tomuto účelu můžeme zvolit jakékoli proměnné, o nichž máme objektivní údaje. Vždy však musíme prověřit, jak se kryjí kategorie výzkumné a kategorie údajů čerpaných z jiných zdrojů.

Pro ověřování reprezentativity je vhodné volit též nějaký faktický znak, jehož validita a reliabilita není příliš snížena, tj. znak, u něhož respondenti zkreslují odpovědi co nejméně a který je kontrolovatelný z nezávislých údajů.

Příklady:[3]

a) dotaz, zda se v rodině odebírá prostřednictvím PNS nějaký časopis či noviny; údaji o velikosti rodiny, resp. počtu členů rodiny patřících do zkoumané populace, a údaji o počtu abonovaných výtisků je možné kontrolovat reprezentativitu;
b) dotaz na vybavenost nějakým předmětem, resp. poměr vybavenosti dvěma předměty (automatická pračka, barevný televizor, stereo magnetofon, stereo přijímač apod.); je lépe volit takové předměty, které jsou relativně nové;
c) dotaz na členství v nějaké společnosti či zájmové organizaci.

Takto volené znaky nemusí nutně odpovídat tématu výzkumu, i když je pochopitelně lepší, podaří-li se nám takové znaky zařadit v souladu s výzkumnou úlohou, neboť tím současně přispíváme ke zhodnocení obsahové reprezentativity šetření.

Možnosti přímého ověřování reprezentativity jsou přes uvedené příklady relativně velmi malé. Právě pro důležitost, kterou reprezentativita má pro výzkum, je nutné dále hledat cesty k jejímu stále hlubšímu a přitom rutinnímu ověřování na základě známých metod i vyvíjet metody nové.

---

[3]) Příklady jsou vzaty z výzkumné praxe J. Řeháka (podle osobního sdělení).

Statistické hodnocení vztahu populačních a výzkumných rozdělení můžeme provádět dvěma způsoby:

1. *Testy dobré shody* ukazují, zda je či není porušena shoda mezi populačními a výzkumnými relativními četnostmi, tj. zda se v datech výběrového šetření ukazuje nějaký systematický nekontrolovaný vliv, či zda můžeme rozdíl interpretovat jako výsledek náhodných vlivů. Dalším krokem pak může být *znaménkové schéma*, které pomáhá specifikovat odchýlení.

2. *Koeficienty neshody*, které v intervalu (0, 1) ukazují na stupeň statistické neshody populačních a výzkumných rozdělení.[4]

## 2. Shoda jednorozměrného rozdělení

Nejprve zavedeme nutné symboly:

a) Zkoumaný *znak* má $K$ kategorií.

b) Jsou známé proporce $\pi_1, \pi_2, \ldots, \pi_K$, $(\Sigma\pi_i = 1,\ \pi_i \geqq 0)$ pro tu *populaci, kterou chceme reprezentovat.*

c) Velikost výběru je $n$; četnosti v jednotlivých kategoriích označíme postupně $n_1, n_2, \ldots, n_K$, $(\Sigma n_i = n)$.

d) $\frac{n_1}{n}, \frac{n_2}{n}, \ldots, \frac{n_K}{n}$ jsou relativní četnosti v jednotlivých kategoriích

$\left(\text{označíme si } f_i = \frac{n_i}{n}\right)$.

e) $p_1, p_2, \ldots, p_K$, $(\Sigma p_i = 1,\ p_i \geqq 0)$ jsou proporce v *populaci, kterou výběr skutečně reprezentuje.*

Úloha zní: testujte hypotézu, že skutečně reprezentovaná populace je totožná s tou, kterou jsme chtěli reprezentovat, proti alternativě, že se tyto dvě populace liší.

Postup:

1. *Formulace hypotézy:*

$H_0$: $p_1 = \pi_1,\ p_2 = \pi_2, \ldots, p_K = \pi_K$

$H_{alt}$: alespoň pro jednu kategorii $i$ je $p_i \neq \pi_i$

2. *Data:* máme k dispozici absolutní četnosti $n_1, n_2, \ldots, n_K$, získáné ve výzkumu náhodným výběrem, a populační rozdělení $\pi_1, \pi_2, \ldots, \pi_K$, získané z objektivních statistických zdrojů.

3. *Intuitivní model:* jestliže platí $H_0$, pak $\pi_i \doteq \frac{n_i}{n}$ neboli $n\pi_i \doteq n_i$ pro všechna $i$.

Při $H_{alt}$ bude tato přibližná rovnost jakýmkoli způsobem porušena.

4. *Testové kritérium* (viz např. Rao [1978]): spočítáme Pearsonovu statistiku

$$X^2 = \sum_{i=1}^{K} \frac{(n_i - n\pi_i)^2}{n\pi_i} = \sum_i \frac{n_i^2}{n\pi_i} - n = n\left(\sum_i \frac{f_i^2}{\pi_i} - 1\right), \tag{1}$$

která je rozdělena asymptoticky chí-kvadrát s $K$-1 stupni volnosti.

Test lze provést dvojím způsobem:

---

[4]) Tyto koeficienty budou obsahem jiného článku.

a) hodnotu $X^2$ porovnáme s kritickou hodnotou[5] $\chi^2_{\alpha,K-1}$ a je-li

$$X^2 \geqq \chi^2_{\alpha,K-1}\ ,\quad \text{zamítáme } H_0$$
$$X^2 < \chi^2_{\alpha,K-1}\ ,\quad \text{nemáme důvod } H_0 \text{ zamítnout;}$$

b) zjistíme (nalezneme v tabulkách[6]) nebo zjistíme na samočinném počítači hodnotu

$$P = \text{pravděpodobnost}\,(\chi^2_{K-1} \geqq X^2)\ ,$$

kterou nazýváme *významností* (signifikací), a tu porovnáme s předem zvolenou hladinou významností $\alpha$. Je-li

$$P \leqq \alpha$$

zamítáme $H_0$,

$$P > \alpha$$

nemáme důvod $H_0$ zamítnout.

Statistika $X^2$ má rozdělení chí-kvadrát pouze asymptoticky, a tak vzniká otázka, pro jak velké výběry (jak velká $n$) již lze taková rozdělení použít při testech dobré shody. Tento problém zajímá statistiky už dobrých padesát let. Dlouho doporučovane pravidlo, že všechny očekávané četnosti $n\pi_1, \ldots, n\pi_K$ by měly být větší než 5 (někteří statistici uváděli 10 i 20), se v posledních patnácti letech zmírnilo.

Yarnold [1970] dospěl na základě studia skutečného rozdělení statistiky $X^2$ při malých očekávaných četnostech k tomuto pravidlu: Je-li počet kategorií $K \leqq 3$ a jestliže $m$ označuje počet očekávaných četností menších než 5, pak minimální očekávaná četnost může být i $5m/K$.

Roscoe a Byars [1971] formulovali na základě rozsáhlé simulační studie následující doporučení:

I. Počet supňů volnosti = 1, pak používejte přesný test pomocí binomického rozdělení.

II. Počet stupňů volnosti > 1.

a) Je-li hypotetické rozdělení (tj. $\pi_1, \ldots, \pi_K$) rovnoměrné, pak očekávané četnosti mohou být i rovny jedné.
I toto pravidlo je přísné, neboť chí-kvadrátová aproximace je velmi robustní v testech dobré shody s rovnoměrným hypotetickým rozdělením a v běžných aplikacích se nemusíme příliš starat o minimální očekávanou četnost.

b) Pro $\alpha = 0.05$ a mírnou odchylku od rovnoměrného rozdělení je přijatelná aproximace při $\frac{n}{K} \geqq 1$.

c) Pro $\alpha = 0{,}05$ a velkou odchylku od rovnoměrného rozdělení je přijatelná aproximace při $\frac{n}{K} \geqq 2$.

---

[5]) Kritickou hodnotu pro rozdělení chí-kvadrát nalezneme ve většině statistických učebnic, viz též Janko (1958, 22. 118), Likeš, Laga [1978, Tab. 2].

[6]) Pro $X^2$ můžeme nalézt hodnotu distribuční funkce $P(\chi^2_{k-1} < X^2)$, např. v Janko (1958, str. 22, 116). Hodnota $P$ je pak

$$1 - P(\chi^2_{k-1} < X^2)$$

Tento postup se často používá u samočinných počítačů, kde se spočítá hodnota $P$ a tiskne se jako hodnota statistické významnosti.

d) Pro $\alpha = 0{,}01$ je nutno požadavky na průměrnou očekávanou četnost ($= \frac{n}{K}$) zdvojnásobit.

Toto doporučení potvrzují i simulace jiných autorů, například Larntze [1978], a není v rozporu s dílčími doporučeními roztroušenými v řadě publikací (např. Van der Waerden [1960]).

Proti vzrůstající tendenci ke zmírňování požadavků týkajících se minimální očekávané četnosti při použití chí-kvadrát testu dobré shody se postavili Tate a Hyer [1973]. Argumentují tím, že použití Pearsonovy statistiky $X^2$ pouze aproximuje přesný multinomický test a že tedy problém nestojí tak, jak dobrou aproximací rozdělení statistiky $X^2$ je rozdělení chí-kvadrát, ale že jde o to, jak dobrou aproximací kumulativních multinomických pravděpodobností jsou chí-kvadrátové pravděpodobnosti statistiky $X^2$. Závěrem jejich simulační studie je, že chí-kvadrát test dobré shody není uspokojující aproximací přesného multinomického testu, pokud všechny očekávané četnosti nejsou větší než 10. Dokonce i když převýší 10, mohou se objevit špatné aproximace. A proto se jim jeví jako nejspolehlivější doporučení Kendallovo [1966], že by žádná očekávaná četnost neměla být menší než 20.

Tento spor není nutno považovat pro praxi za důležitý. Statistika $X^2$ je sama o sobě samostatně stojícím testovým kritériem, které má svoji jednoduchou logiku a přijatelné heuristické zázemí. Proto je možné přijmout uvedené uvolněné podmínky pro aplikabilitu rozdělení chí-kvadrát k Pearsonově statistice $X^2$, aniž bychom si kladli za cíl přesně nahrazovat multinomický test.

Jestliže počet stupňů volnosti přesahuje rámec dostupných tabulek, můžeme použít tzv. Wilson-Hilfertyho aproximaci[7]

$$Y = \left[ \sqrt[3]{\frac{X^2}{K-1}} - 1 + \frac{2}{9(K-1)} \right] \sqrt{\frac{9(K-1)}{2}}, \qquad (2)$$

$$P = P(\chi^2_{K-1} > X^2) \doteq 1 - \Phi(Y) = \Phi(-Y)$$

kde $\Phi(Y)$ je distribuční funkce normálního rozdělení s průměrem nula a rozptylem jedna.

Aproximaci (2) můžeme využít dvěma postupy:

a) spočítáme $Y$, dosadíme do distribuční funkce $\Phi$ standardizovaného normálního rozdělení[8] a výsledné $P$ porovnáme s předem zvoleným $\alpha$.

b) spočítáme $Y$ a porovnáme ho se $100(1 - \alpha)$procentním kvantilem standardizovaného normálního rozdělení,[9]) resp. s kritickou hodnotou jednostranného normálního testu. Například pro

| | | |
|---|---|---|
| $\alpha = 0{,}05$ | zamítáme $H_0$, je-li | $Y \geqq 1{,}645$ |
| $\alpha = 0{,}01$ | zamítáme $H_0$, je-li | $Y \geqq 2{,}326$ |
| $\alpha = 0{,}001$ | zamítáme $H_0$, je-li | $Y \geqq 3{,}090$ |

Uspořádání dat vhodné pro test dobré shody ukazuje tabulka 1.

---

[7]) Viz Johnson, Kotz [1970, str. 176].

[8]) Viz např. Janko [1958, 111], Likeš, Laga [1978, Tab. 1A]

[9]) Viz např. Janko [1958, 113], Likeš, Laga [1978, Tab. 1B]. Tabulkové hodnoty se nazývají buď $100(1 - \alpha)$%ní kvantily (např. 95%-ní kvantil, nebo horní $100\alpha$%-ní kritický bod (např. 5%-ní).

*Tabulka 1. Uspořádání dat pro testování shody*

| kategorie | 1 | 2 | ... | $i$ | ... | $K$ | celkem |
|---|---|---|---|---|---|---|---|
| hypotetické relativní četnosti | $\pi_1$ | $\pi_2$ | ... | $\pi_i$ | ... | $\pi_K$ | 1 |
| očekávané četnosti | $n\pi_1$ | $n\pi_2$ | ... | $n\pi_i$ | ... | $n\pi_K$ | $n$ |
| výběrové četnosti | $n_1$ | $n_2$ | ... | $n_i$ | ... | $n_K$ | $n$ |

Jestliže odmítneme hypotézu $H_0$, zajímá nás většinou, ve kterých polích se projevují odchylky. Následující postup můžeme použít dvojím způsobem: pro test hypotézy o odchylce v předem dané určité kategorii a pro získání grafického znaménkového schématu.

*Test odchylky v jedné kategorii*

Jestliže jsme zamítli hypotézu dobré shody výběrových a zadaných očekávaných četností, můžeme formulovat další hypotézy pro předem danou vybranou kategorii ($I$-tou hodnotu znaku)

a) v $I$-té kategorii nastalo vychýlení;
b) v $I$-té kategorii je výběrová četnost větší než očekáváme;
c) v $I$-té kategorii je výběrová četnost menší než očekáváme.

Postup:

1. *Formulace hypotézy:*

$$H_0 : p_I = \pi_I ,$$

kde $I$ je předem určené

Možné alternativní hypotézy

$$H_a : p_I \neq \pi_I$$
$$H_b : p_I > \pi_I$$
$$H_c : p_I < \pi_I$$

Pro jednu z těchto alternativ se musíme rozhodnout, a to *před pohledem na data.*

2. *Data:* známe $n_I$ a $n$.

3. *Intuitivní model:* jestliže platí $H_0$, pak $\frac{n_I}{n} \doteq \pi_I$.

4. *Testové kritérium:* spočítáme statistiku

$$z_I = \frac{n_I - n\pi_I}{\sqrt{n\pi_I(1 - \pi_I)}} \tag{3}$$

která má pro $n\pi_I > 5$ přibližně standardní normální rozdělení.[10] Testování můžeme provést dvěma postupy:

[10]) Pro $n\pi_i \leqq 5$ lze volit jiné postupy, které zde neuvádíme a které vycházejí z binomického nebo Poissonova rozdělení (viz např. van der Waerden [1960]).

A. Zvolíme hladinu významnosti $\alpha$ (např. 0,05) a nalezneme v tabulkách[8] hodnotu distribuční funkce normálního rozdělení v bodě $z_I$, tj. $\Phi(z_I)$. Jestliže testujeme $H_0$ proti $H_a$, pak $H_a$ přijmeme, když

$$1 - \Phi(z_I) \leqq \frac{\alpha}{2}, \qquad \text{nebo} \quad \Phi(z_I) \leqq \frac{\alpha}{2}$$

Jestliže testujeme $H_0$ proti $H_b$, pak $H_b$ přijmeme, jestliže

$$1 - \Phi(z_I) < \alpha .$$

Jestliže testujeme $H_0$ proti $H_c$, pak $H_c$ přijmeme, když

$$\Phi(z_I) < \alpha .$$

B. Pro předem určené $\alpha$ porovnáme hodnotu $z_I$ s kvantilem normálního rozdělení $z^*$.[9,11]

Jestliže testujeme $H_0$ proti $H_a$, pak $H_a$ přijmeme, když

$$|z_I| > z^*, \quad z^* = 100\left(1 - \frac{\alpha}{2}\right)\,\% \text{ kvantil.}$$

Jestliže testujeme $H_0$ proti $H_b$, pak $H_b$ přijmeme, když

$$z_I > z^*, \quad z^* = 100(1 - \alpha)\,\% \text{ kvantil.}$$

Jestliže testujeme $H_0$ proti $H_c$, pak $H_c$ přijmeme, když

$$z_I < -z^*, \quad z^* = 100(1 - \alpha)\,\% \text{ kvantil.}$$

*Znaménkové schéma*

Hodnoty $z_i$, spočítané pro všechna pole jednorozměrné tabulky, můžeme použít pro znázornění odchylek tak, že každému poli přiřadíme znaménko podle pravidla, které ukazuje tabulka 2.

*Tabulka 2. Pravidlo pro tisk znaménkového schématu*

| hodnota $z_i$ | tisk |
|---|---|
| $z_i \leqq -3.29$ | − − − |
| $-3.29 < z_i \leqq -2.58$ | − − |
| $-2.58 < z_i \leqq -1.96$ | − |
| $-1.96 < z_i < 1.96$ | 0 |
| $1.96 \leqq z_i < 2.58$ | + |
| $2.58 \leqq z_i < 3.29$ | + + |
| $3.29 \leqq z_i$ | + + + |

Nutno ovšem poznamenat, že celý soubor znamének nemůžeme považovat za výsledek celkového statistického testu s hladinou $\alpha$. Vzhledem k závislosti pozorování

[11] Pro $\alpha = 0{,}05$ je $100\left(1 - \frac{\alpha}{2}\right)$%-ní kvantil = 1,960 a $100(1 - \alpha)$%-ní kvantil = 1,645. Pro $\alpha = 0{,}01$ je $100\left(1 - \frac{\alpha}{2}\right)$%-ní kvantil = 2,576 a $100(1 - \alpha)$%-ní kvantil = 2,326.

v jednotlivých polích jsou pravděpodobnostní hladiny posunuty a přesně neplatí. Jde tedy o grafickou pomůcku k interpretaci, nikoliv o přesná pravděpodobnostní rozhodnutí.

*Příklad 1:* Ve výzkumu sledujeme reprezentativitu u různých znaků a na různých podsouborech. U znaku „rodinný stav" na podsouboru žen[12]) dostaneme výsledky, které uvádí tabulka 3.

*Tabulka 3. Komparace populačního a výběrového rozdělení u znaku „rodinný stav" na podsouboru žen ČSSR od 15 let včetně*

| rodinný stav | vdaná | svobodná | rozvedená nebo vdova | celkem |
|---|---|---|---|---|
| procento v populaci | 62.51 % | 18.37 % | 19.12 % | 100.00 % |
| procento ve výběru | 54.04 % | 28.45 % | 17.51 % | 100.00 % |
| očekávaná četnost ve výběru | 371.31 | 109.12 | 113.57 | 594 |
| skutečná četnost ve vyběru | 321 | 169 | 164 | 594 |

Spočítáme statistiku $X^2$ pro testovani dobre shody mezi výběrem a populaci ČSSR od 15 let včetně vzhledem ke znaku rodinný stav u žen:

$$X^2 = \sum_{i=1}^{3} \frac{n_i^2}{n\pi_i} - n = \frac{321^2}{371{,}31} + \frac{169^2}{109{,}12} + \frac{104^2}{113{,}57} - 594 = 40{,}64$$

Kritická hodnota pro $\alpha = 0{,}05$ a počet stupňů volnosti = 2 je[5] 5,99. Protože $40{,}64 > 5{,}99$, zamítáme hypotézu dobré shody. Pro tento znak si též zjistíme znaménkové schéma (tabulka 4).

Například

$$z_2 = \frac{169 - 109{,}12}{\sqrt{594 \,.\, 0{,}1837 \,.\, (1 - 0{,}1837)}} = +6{,}345$$

a podle tabulky 2 je příslušný symbol „+ + +".

Interpretace: porušení reprezentativity pro soubor žen je způsobeno zvýšením procenta svobodných žen, a to na úkor žen vdaných. Statistický test zde pouze indikuje odchylky, skutečná příčina může být nalezena pouze sociologickým rozborem práce tazatelské sítě, vztahu respondentů k tématu výzkumu, porovnáním podobných výsledků u jiných znaků apod.

*Tabulka 4. Hodnoty statistiky $z_i$ ($i = 1, 2, 3$) a znaménkové schéma*

| rodinný stav | vdaná | svobodná | rozvedená nebo vdova |
|---|---|---|---|
| hodnota $z$ | −4.264 | +6.345 | −1.033 |
| znaménkové schéma | − − − | + + + | 0 |

[12]) Základním souborem byla populace ČSSR od 15 let včetně. Znak „pohlaví" byl kvótní, proto bylo nutné provést testování pro oba podsoubory mužů a žen zvlášť. Výsledek testu je validní za predpokladu, že jsou splněny podmínky prostého náhodného výběru.

### 3. Test dobré shody u rozdělení kombinace dvou znaků (žádný ze znaků není kontrolován)

Máme-li jeden znak o $K$ hodnotách a druhý o $M$ hodnotách, můžeme použít postup z části 2 na nový znak o $K \times M$ hodnotách, který vznikne kombinací obou. Postup je zcela stejný jako dříve, pouze zde máme $K \times M$ četností $n_{ij}$ a $K \times M$ hypotetických pravděpodobností $\pi_{ij}$. Statistika $X^2$ má tvar:

$$X^2 = \sum_{i=1}^{K} \sum_{j=1}^{M} \frac{(n_{ij} - n\pi_{ij})^2}{n\pi_{ij}} = \sum_{i=1}^{K} \sum_{j=1}^{M} \frac{n_{ij}^2}{n\pi_{ij}} - n \tag{4}$$

Počet stupňů volnosti je $(K \times M) - 1$.

Statistika $z$ pro test odchylky v jednom poli má pro pole $(i, j)$ tvar

$$z_{ij} = \frac{n_{ij} - n\pi_{ij}}{\sqrt{n\pi_{ij}(1 - \pi_{ij})}} \tag{5}$$

Též znaménkové schéma se dělá stejně.

Tyto postupy však platí jen v tom případě, když žádný z obou znaků není kontrolován (není stratifikačním nebo kvótním znakem). Postupy pro ostatní případy si uvedeme ve čtvrté části.

*Příklad 2.* Testujeme dobrou shodu mezi rozdělením v populaci a ve výběru u kombinace znaků „věková skupina" a „rodinný stav", a to na podsouboru mužů. Postup a nutné údaje jsou v tabulkách 5, 6, 7.

Celkové řádkové a sloupcové údaje se testování nezúčastní, proto jsou čísla naznačena v závorkách (to platí pro tabulky 5 a 6).[13]

Výpočet statistiky $X^2$ provedeme podle vzorce (4).

$$X^2 = \frac{31^2}{64{,}07} + \frac{99^2}{112{,}51} + \ldots + \frac{34^2}{17{,}81} - 528 = 39{,}06$$

*Tabulka 5. Testování dobré shody pro kombinaci znaků „věková skupina" a „rodinný stav"* — v prvním řádku je procentní ozdělení v populaci mužů ČSSR, v druhém řádku je procentní rozdělení ve výběru ($n = 528$)

| | věková skupina | | | | |
|---|---|---|---|---|---|
| rodinný stav | 15—29 | 30—44 | 45—59 | 60 a více | celkem |
| ženatý | 12.13 | 21.31 | 18.84 | 15.44 | (67.73) |
| | 5.87 | 18.75 | 19.70 | 18.18 | (62.50) |
| svobodný | 22.17 | 2.06 | 1.08 | 0.91 | (26.22) |
| | 24.24 | 2.27 | 1.32 | 0.57 | (28.41) |
| rozvedený nebo vdovec | 0.38 | 1.02 | 1.27 | 3.37 | (6.05) |
| | 0.19 | 0.95 | 1.52 | 6.44 | (9.09) |
| celkem | (34.69) | (24.39) | (21.20) | (19.73) | (100.00) |
| | (30.30) | (21.97) | (22.54) | (25.19) | (100.00) |

[13]) V tabulkách 5 a 6 všechny součty přesně nehrají z toho důvodu, že zde uvedená čísla jsou zaokrouhlená.

*Tabulka 6. Testování dobré shody pro kombinaci znaků „věková skupina" a „rodinný stav" u souboru mužů ČSSR.* V prvním řádku jsou očekávané četnosti ($n\pi_{ij}$) a v druhém řádku jsou skutečně zjištěné četnosti ($n_{ij}$)

| | věková skupina | | | | |
|---|---|---|---|---|---|
| rodinný stav | 15—29 | 30—44 | 45—59 | 60 a více | celkem |
| ženatý | 64,07 | 112.51 | 99.50 | 81.55 | (357.63) |
| | 31 | 99 | 104 | 96 | (330) |
| svobodný | 117.07 | 10.85 | 5.72 | 4.80 | (138.44) |
| | 128 | 12 | 7 | 3 | (150) |
| rozvedený nebo vdovec | 2.01 | 5.40 | 6.71 | 17.81 | (31.93) |
| | 1 | 5 | 8 | 34 | (48) |
| celkem | (183.14) | (128.77) | (111.92) | (104.16) | (528) |
| | (160) | (116) | (119) | (133) | (528) |

*Tabulka 7. Znaménkové schéma*[14] *pro kombinaci znaků „věková skupina" a „rodinný stav" pro soubor mužů (n = 528)*

| | Věková skupina | | | |
|---|---|---|---|---|
| rodinný stav | 15—29 | 30—44 | 45—59 | 60 a více |
| ženatý | —4.41 | —1.44 | 0.50 | 1.74 |
| | — — — | 0 | 0 | 0 |
| svobodný | 1.14 | 0.35 | 0.54 | —1.05 |
| | 0 | 0 | 0 | * |
| rozvedený nebo vdovec | —1.07 | —0.17 | 0.50 | 3.90 |
| | * | 0 | 0 | + + + |

Počet stupňů volnosti je 3 . 4 — 1 = 11.

Kritická hodnota rozdělení chí-kvadrát při 11 stupních volnosti pro $\alpha = 0{,}05$ je 19,7. Jelikož 39,06 je větší než 19,7, zamítáme hypotézu dobré shody ve prospěch hypotézy, která říká, že při výběru působila systematická příčina, která způsobila jeho vychýlení. Abychom se přiblížili k vysvětlení, sestrojíme znaménkové schéma (viz tabulka 7). Hodnoty statistiky z počítáme podle vzorce (5). Například pro pole (3, 2) dostaneme

$$z_{3,2} = \frac{5 - 5{,}40}{\sqrt{5{,}40(1 - 0{,}0095)}} = -0{,}17$$

Znaménkové schéma ukazuje, že porušení shody je dílem dvou polí: „mladí ženatí" a „staří rozvedení či vdovci". Ve výběru chybí muži první kategorie, což je kompenzováno přebytkem v druhé kategorii. Sociologické vysvětlení je otázkou zhodnocení celého souhrnu podobných testů dobré shody, znalosti tématu výzkumu, kvality tazatelské sítě, analýzy celého výběrového a zpracovatelského postupu.[15]

[14]) Ve dvou polích je uvedena místo běžných znamének hvězdička(*). V těchto polích, přestože jsme pro úplnost uvedli $z$-skóre, znaménko neuvádíme, protože zde $n\pi_{ij} < 5$ a aproximace pomocí normálního rozdělení není proto dobrá.

[15]) Z výzkumné praxe jsou známy i takové příčiny porušení reprezentativity, jako je založení krabice štítků ve výpočetním středisku.

Zde je ještě nutné upozornit, že výše uvedený postup neznamená analýzu kontingenční tabulky vztahu mezi rodinným stavem a věkem. Znaménkové schéma zde nemá nic společného s hypotézou nezávislosti mezi těmito znaky. Jde pouze o vztah mezi skutečnými populačními a výběrovými četnostmi. Data jsou uspořádána v tabulce pouze kvůli přehlednosti. Mohla být též uvedena v jednom řádku o 12 polích vedle sebe, každé pole označené dvojicí hodnot.

## 4. Test dobré shody pro rozdělení kombinace dvou znaků, jsou-li marginální četnosti kontrolovány

Jsou-li u dané kombinace znaků předepsány proporce pro jeden či oba znaky, pak pro kombinaci obou znaků očekáváme větší přesnost ve shodě hodnot $f_{ij}$ a $\pi_{ij}$. Náhoda vnesená výběrovým šetřením je zde menší, a proto i povolené odchylky $f_{ij} - \pi_{ij}$ vysvětlitelné náhodnými vlivy jsou nižší. To se při testování projeví snížením počtu „volných" parametrů, a tím snížením počtu stupňů volnosti.[16] Tabulka 8 udává, jak určujeme počet stupňů volnosti při testování dobré shody rozdělení kombinace znaků pro čtyři různé způsoby vzniku dat. Toto plyne z obecného postupu, který lze nalézt například v Yule, Kendall [1966]. Ve všech čtyřech případech se počítá $X^2$ stejně (podle vzorce (4)).

Statistika $z$ pro test odchylky v jednom poli je různá pro případy I až IV. Odvození nalezne čtenář v *Dodatku*. Vzorec pro *případ I* je uveden v (5).

*Tabulka 8. Počet stupňů volnosti při testování dobré shody rozdělení kombinace znaků* ($R$ = počet řádků, $S$ = počet sloupců)

| způsob vzniku dat | | počet nezávislých lineárních podmínek | počet stupňů volnosti |
|---|---|---|---|
| I | řádkové ($n_{i+}$) ani sloupcové součty ($n_{+j}$) nejsou předem určeny, je dán pouze počet pozorování $n$ | 1 | $RS - 1$ |
| II | řádkové součty ($n_{i+}$) jsou určeny ve výběru pevně předem | $R$ | $R(S - 1)$ |
| III | sloupcové součty ($n_{+j}$) jsou určeny ve výběru pevně předem | $S$ | $S(R - 1)$ |
| IV | řádkové ($n_{i+}$) i sloupcové součty ($n_{+j}$) jsou určeny ve výběru pevně předem | $R + S - 1$ | $(R - 1)(S - 1)$ |

[16]) Kdybychom aplikovali postup části 3 s $RS$-1 stupni volnosti na případ, kdy jsou jeden či oba znaky jednorozměrně kontrolovány a jejich správné proporce zajištěny, tak by platilo: a) jestliže $X^2$ je významná s $RS$-1 stupni volnosti, pak by byla ještě významnější pro případ kontrolovaných marginálií; b) jestliže $X^2$ není významná s $RS$-1 stupni volnosti, mohla by být významná pro případ kontrolovaných marginálií. Souhrnně řečeno, stejná hodnota $X^2$ je vždy významnější pro případ kontrolovaných marginálních četností. Dále platí, že stejná hodnota $X^2$ je významnější pro obě kontrolované marginálie než pro případ jedné kontrolované marginální distribuce. Tato odstupňovanost je způsobena stupněm povoleného náhodného vychýlení, tj. stupněm kontroly výběrového procesu. Čím více výběr kontrolujeme, tím menší odchylky připouštíme.

*Případ II.*

$$z_{ij} = \frac{n_{ij} - n\pi_{ij}}{\sqrt{n\pi_{ij}\left(1 - \frac{\pi_{ij}}{\pi_{i+}}\right)}} \tag{6}$$

kde $\pi_{i+} = \frac{n_{i+}}{n}$ je předem určená proporce pro řádek $i$.

*Případ III.*

$$z_{ij} = \frac{n_{ij} - n\pi_{ij}}{\sqrt{n\pi_{ij}\left(1 - \frac{\pi_{ij}}{\pi_{+j}}\right)}} \tag{7}$$

kde $\pi_{+j} = \frac{n_{+j}}{n}$ je předem určená proporce pro sloupec $j$.

*Případ IV.*

$$z_{ij} = \frac{n_{ij} - n\pi_{ij}}{\sqrt{n\pi_{ij}\left[1 - \pi_{ij} - \frac{\pi_{ij}(1 - \pi_{i+})(1 - \pi_{+j})(\pi_{i+} + \pi_{+j} - 2\pi_{ij})}{\pi_{i+}\pi_{+j}(1 - \pi_{i+})(1 - \pi_{+j}) - (\pi_{i+}\pi_{+j} - \pi_{ij})^2}\right]}} \tag{8}$$

kde $\pi_{i+} = \frac{n_{i+}}{n}$ je předem určená proporce pro řádek $i$ a $\pi_{+j} = \frac{n_{+j}}{n}$ je předem určená proporce pro sloupec $j$.[17]

## Závěr

V práci jsou uvedeny postupy, které statisticky prověřují shodu mezi výběrovým a známým populačním rozdělením. Testy dobré shody zde uváděné jsou založeny na klasické Pearsonově statistice chí-kvadrát. V článku je též navržen test pro odchylku v jednom poli, odvozený pro různé situace vzniku dat, který lze použít jako základ pro znaménkové schéma indikující odchylky.

Testy dobré shody chí-kvadrát slouží nejen pro zhodnocení reprezentativity, ale i v případech, kdy máme předem dané (stabilní a opakováním prověřené) pravděpodobnosti odpovídající zkušenosti z minulých výzkumů či pravděpodobnosti vycházející z nějakého teoretického názoru. Tak například můžeme za $\pi_i$, resp. $\pi_{ij}$, vzít výsledky výzkumů publikovaných jinde a ověřovat, zda je lze opakovat ve vlastním šetření, respektive zda výsledky z jiné populace platí i na populaci speciální. Vezmeme-li za základ pravděpodobnosti $\pi_i$ ($\pi_{ij}$) hodnoty odpovídající minulému stavu na téže populaci, pak můžeme tímto způsobem testovat, zda nastala v mezidobí

[17]) Výsledek případu IV dává také $z$-skóre pro znaménkové schéma a testování odchylky v jednom poli pro hypotézu nezávislosti (viz Řehák, Řeháková [1978]). Test nezávislosti je také odvozen za předpokladu podmíněných řádkových i sloupcových součtů. Předpokládáme-li navíc, že $\pi_{ij} = \pi_{i+} \cdot \pi_{+j}$ a dosadíme-li tyto vztahy do (8), dostaneme výraz pro test odchylky v poli $(i, j)$ v případě testování nezávislosti a pro znaménkové schéma indikující místo v tabulce, kde je nezávislost porušena.

nějaká změna. Zde nutno poznamenat, že přejímání výsledků z minulých výzkumů za pravděpodobnosti $\pi_i$ do testů dobré shody můžeme přijmout pouze v případě, že tyto proporce se při opakovaných šetřeních ukázaly jako stabilní. Tyto proporce totiž vznikají také náhodnými vlivy ve výzkumném šetření (výběrovým postupem). Proto pro porovnání je lépe použít test $X^2$ pro srovnání dvou nebo více výběrových souborů (viz například Řehák, Řeháková [1978]). Při komparaci s minulým stavem se při odmítnutí shody mohou objevit nové interpretační problémy, jako například otázka, zda jde o porušení reprezentativity nebo o odhalení vlivu časového údobí.

Ověřování reprezentativity by mělo být součástí rutiny každého výzkumu. Cílem článku je tuto praxi stimulovat a přispět k ní.

**Dodatek: Odvození testu pro odchylku v jednom poli**

Vzhledem k tomu, že článek uvádí možnosti aplikace metody znaménkového schématu, založené na uvedených vzorcích (5), (6), (7), (8) pro reziduální normalizované odchylky, uvádíme jeho zkrácené odvození. Metoda odvození je použitelná i pro jiné speciální případy znaménkových schémat, která mohou být užitečná v praxi analýzy sociologických dat.

*Případ I* je běžná normální aproximace binomického testu hypotézy $H_0 : p = \pi_{ij}$ proti alternativě $H_1$: $p = \pi_{ij}$, přičemž $n_{ij}$ je počet výskytů jevu pole $(i, j)$ a $n$ je počet pozorování binomické proměnné s parametrem $p$ (zde počet pozorování v tabulce).

*Případ II.* Předpokládáme (bez podmínky $\sum_j n_{ij} = n_{ij}$ dáno), že $n_{ij}$ mají sdružené normální rozdělení s $\mathrm{E}\, n_{ij} = n\pi_{ij}$, $\mathrm{var}\, n_{ij} = n\pi_{ij}(1 - \pi_{ij})$, $\mathrm{cov}\,(n_{ij}, n_{kl}) = -n\pi_{ij}\pi_{kl}$ (pro $(i, j) \neq (k, l)$, jak plyne z limitních vět pro multinomické rozdělení.

Zajímá nás rozdělení $n_{IJ}$ za podmínky, že $\sum_j n_{Ij} = n_{I+}$. Podle věty o podmíněných rozděleních sdruženého rozdělení (viz Rao [1978], str. 566, věta (V)) bude toto rozdělení též normální. Označme

$$n_{IJ} = X\,, \quad \sum_{j \neq J} n_{Ij} = Y\,, \quad U = X + Y\,.$$

Pak platí:

$$\mathrm{E}X = n\pi_{IJ} \qquad \mathrm{var}\, X = n\pi_{IJ}(1 - \pi_{IJ})$$
$$\mathrm{E}Y = n(\pi_{I+} - \pi_{IJ}) \qquad \mathrm{var}\, Y = n(\pi_{I+} - \pi_{IJ})(1 - \pi_{I+} + \pi_{IJ})$$
$$\mathrm{E}U = n\pi_{I+} \qquad \mathrm{var}\, U = n\pi_{I+}(1 - \pi_{I+})$$
$$\mathrm{cov}\,(X, Y) = -n\pi_{IJ}(\pi_{I+} - \pi_{IJ})\,, \qquad \mathrm{cov}\,(X, U) = n\pi_{IJ}(1 - \pi_{I+})$$

Nyní hledáme očekávanou hodnotu a varianci veličiny $X$ za podmínky $U = \mathrm{E}U = n\pi_{I+}$. Podle výše citované věty (Rao [1978]) je

$$\mathrm{E}\,(X \mid U = n\pi_{I+}) = \mathrm{E}X = n\pi_{IJ}$$

$$\mathrm{var}\,(X \mid U = n\pi_{I+}) = \mathrm{var}\, X - \frac{\mathrm{cov}^2\,(X, U)}{\mathrm{var}\, U} = n\pi_{IJ}\left(1 - \frac{\pi_{IJ}}{\pi_{I+}}\right)$$

Standardizovaná normální veličina $z_{IJ}$ je tedy za předpokladu modelu II rovna

$$z_{IJ} = \frac{n_{IJ} - n\pi_{IJ}}{\sqrt{n\pi_{IJ}\left(1 - \frac{\pi_{IJ}}{\pi_{I+}}\right)}}$$

což dává vzorec (6).

*Případ III* se dokazuje obdobně jako II.

*Případ IV.* Přijmeme značení z případu II a dále zavedeme

$$Z = \sum_{i \neq I} n_{iJ}\,, \quad V = X + Z\,.$$

Platí dále

$$\begin{aligned}
&EZ = n(\pi_{+J} - \pi_{IJ}) && \text{var } Z = n(\pi_{+J} - \pi_{IJ})(1 - \pi_{+J} + \pi_{IJ}) \\
&EV = n\pi_{+J} && \text{var } V = n\pi_{+J}(1 - \pi_{+J}) \\
&\text{cov } (X, Z) = -n\pi_{IJ}(\pi_{+J} - \pi_{IJ}) && \text{cov } (Y, X) = -n(\pi_{+J} - \pi_{IJ})(\pi_{I+} - \pi_{IJ}) \\
&\text{cov } (X, V) = \text{var } X + \text{cov } (X, Z) = n\pi_{IJ}(1 - \pi_{+J}) \\
&\text{cov } (U, V) = \text{var } X + \text{cov } (X, Y) + \text{cov } (X, Z) + \text{cov } (Y, Z) \\
&\qquad = n(\pi_{IJ} - \pi_{I+}\pi_{+J})
\end{aligned}$$

Hledáme rozdělení veličiny $X = n_{IJ}$ za podmínky $U = n\pi_{I+}$ a $V = n\pi_{+J}$. Podle výše uvedené věty (Rao [1978]) je toto rozdělení normální s očekávanou hodnotou a variancí

$$\begin{gathered}
\mathrm{E}\,(XIU = n\pi_{I+}, V = n\pi_{+J}) = \mathrm{E}X = n\pi_{IJ} \\
\text{var } (XIU = n\pi_{I+}, V = n\pi_{+J})
\end{gathered}$$

$$= \text{var } X - (\text{cov } (X, U), \text{cov } (X, V)) \begin{pmatrix} \text{var } U, & \text{cov } (U, V) \\ \text{cov } (U, V), & \text{var } V \end{pmatrix}^{-1} \begin{pmatrix} \text{cov } (X, U) \\ \text{cov } (X, V) \end{pmatrix}$$

Po provedení inverze, roznásobení a dosazení dostaneme, že

$$\text{var } (XIU, V) = n\pi_{IJ}\left[1 - \pi_{IJ} - \frac{\pi_{IJ}(1 - \pi_{I+})(1 - \pi_{+J})(\pi_{I+} + \pi_{+J} - 2\pi_{IJ})}{\pi_{I+}\pi_{+J}(1 - \pi_{I+})(1 - \pi_{+J}) - (\pi_{I+}\pi_{+J} - \pi_{IJ})^2}\right] = \sigma^2_{IJ}$$

Standardizovaná normální veličina $z_{IJ}$ je tedy za předpokladu modelu IV rovna

$$z_{IJ} = \frac{n_{IJ} - n\pi_{IJ}}{\sigma_{IJ}}.$$

Tím jsme dostali vzorec (8).

Q. E. D

## Literatura


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

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

Kendall, M. G. — Stuart, A.: *Těorija raspredělenij.* Moskva, Nauka 1966.

Larntz, K.: *Small — Sample Comparisons of Exact Levels for Chi — Squared Goodness — of — Fit Statistics.* Journal of the American Statistical Association 1978, 73, s. 253—263.

Likeš, J. — Laga, J.: *Základní statistické tabulky.* Praha, SNTL 1978.

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

Roscoe, J. T. — Byars, J. A.: *An Investigation of the Restraints with Respect to Sample Size Commonly Imposed on the Use of the Chi — Square Statistic.* Journal of the American Statistical Association 1971, 66, s. 755—759.

Řehák, J.: *K pojmu „reprezentativita" v sociologických výzkumech.* Sociologický časopis 1978, č. 4, s. 489—507.

Řehák, J. — Řeháková, B.: *Analýza kontingenčních tabulek: dva základní typy úloh a znaménkové schéma.* Sociologický časopis 1978, č. 6, s. 619—631.

Tate, M. W. — Hyer, L. A.: *Inaccuracy of the $X^2$ Test of Goodness of Fit When Expected Frequencies Are Small.* Journal of the American Statistical Association 1973, 68, s. 836—841.

Van der Waerden, B. L.: *Matěmatičeskaja statistika.* Moskva, Inostrannaja litěratura 1960.
Yarnold, J. K.: *The Minimum Expectation in $X^2$ Goodness of Fit Tests and the Accuracy of Approximations for the Null Distribution.* Journal of the American Statistical Association 1970, 65, s. 864—886.
Yule, G. U. — Kendall, M. G.: *Wstep do teorii statistiky.* Warszàwa, PWN 1966.
Zich, F.: *Sociologický výzkum.* Praha, Svoboda 1976.

## Резюме

**Б. Ржегакова: Статистическая проверка репрезентативности**

Проверка статистических гипотез является очень важной частью методологической разработки каждого исследования основанного на выборке и стремящегося к обобщению на некоторую данную целевую совокупность. В этой статье даны статистические техники для проверки согласия выборочных и популяционных данных, которые основаны на критерии хи-квадрат Пирсона. Кроме этих методов включены тоже новейшие результаты об ограничениях использования этого критерия (малые выборки, малые ожидаемые частоты).

В первой части дискутируются практические аспекты методологии проверки репрезентативности, подбор подходящих переменных и описывается задача в своих основных чертах. В следующих частьях дается описание статистических техник для проверки согласия распределения одной переменной и комбинации двух переменных. Последнее дискутируется как для случая нефиксированных, так и фиксированных (одного или обоих) маргинальных распределений. Эти случаи различáются определением степеней свободы.

В статье тоже даются методы проверки гипотези об отклонении в определенной категории знака (или комбинации знаков). Эта техника является удобной для спецификации альтернативной гипотезы, когда нулевая гипотеза отвергается. Добавление содержит краткий обзор доказательств новых результатов.

Все техники иллюстрируются на практических примерах, которые объясняют метод и вычислительный процесс. Они обоснованы на предположении простой случайной выборки или выборки, которая является ее математическим или эмпирическим эквивалентом.

## Summary

**B. Řeháková: Statistical Verification of Representativity**

Testing statistical representativity is an important part of methodological evaluation that is a part of any data analysis based on sampling and aiming at generalizations to a given target population. The paper presents statistical techniques for goodness--of-fit testing based on chi-square tests. A review of well-known Pearson chi-square goodness-of-fit test is given together with the review of the newest results on restrictions for its use (small sample sizes and small expected frequencies in cells).

The first part contains a discussion on the practical aspects of methodology, on the choice of variables convenient for these tasks and presents the basic characterization of the problem. In the next parts, statistical techniques for testing goodness-of-fit for a one-dimensional categorical distribution of research data are given. Then there is presented the testing of fit for combination of two variables without marginal restrictions, with the restrictions (given sample sizes) for one marginal and finally for both marginals. The technique is the same in all cases but the degrees of freedom must be adjusted accordingly.

For each case, the test for deviation in one cell is given. It can be used for preparing a sign scheme that is useful for specification of alternative hypothesis in the case of rejection of null-hypothesis.

The sketch of proof of these results is presented in Appendix. Eeach case is illustrated by a numerical example. All the techniques are derived on an assumption of simple random sampling or its mathematical or empirical equivalent.