Statistické ověřování reprezentativity: koeficienty neshody

BLANKA ŘEHÁKOVÁ
Ústav pro výzkum veřejného mínění při FSÚ, Praha

Reprezentativita je beze sporu jedním z hlavních kritérií výzkumných dat a zároveň nutným předpokladem pro kvalitní interpretační závěry. Její porušení může vést ke zkreslení výsledků a tím následně k mylným praktickým opatřením či teoretickým úvahám. Rozbor pojmu a argumentace jeho důležitosti i základní úlohy s ním spojené byly prezentovány v práci J. Řeháka [1978]. Praktickými aspekty ověřování statistické reprezentativity se zabývala B. Řeháková [1979]. Cílem této stati je zavést nové metody při hodnocení reprezentativity a přispět tak k řešení tohoto naléhavého úkolu.

Při studiu statistických jevů nás zpravidla zajímají dva hlavní typy úloh: statistické testování existence jevu a měření síly (intenzity, stupně), s níž se daný jev projevuje. Proto často používáme postup, který odpovídá výzkumné logice: nejprve zjistíme, zda data indikují existenci jevu (statistickým testem), a je-li tomu tak, pak měříme intenzitu jevu.

V článku B. Řehákové [1979] byly popsány testy dobré shody, které jsou vhodné pro ověřování reprezentativity nominálních znaků. Nulová hypotéza znamenala shodu předpokládaného a získaného rozdělení četností. Zamítnutí nulové hypotézy vedlo k závěru, že vzhledem k danému znaku je šetření nereprezentativní a že se v něm projevují nenáhodné faktory, jejichž působení znamená vychýlení dat.

V tomto článku je navržen a popsán způsob měření stupně nereprezentativity (neshody) empiricky získaného rozdělení četností a předem daného (postulovaného, teoretického, skutečného populačního) rozdělení.

Zatímco testování shody provádíme většinou za předpokladu nominálního znaku (chí-kvadrát test je test shody proti omnibusové alternativě) a alternativy se snažíme specifikovat pomocí znaménkového schématu, stupeň neshody měříme pro různé typy znaků různě. Zde navržené koeficienty vycházejí z tak zvaného D-modelu, zavedeného v článku J. Řeháka a B. Řehákové [1978, 1979], čímž je dáno jednotné hledisko jejich interpretace pro různé typy znaků. Zároveň je též uvedena a diskutována Weilerova míra neshody [Weiler 1966] pro nominální znaky.

Dále jsou uvedeny směrodatné odchylky koeficientů pro určení konfidenčních intervalů a testy pro porovnání koeficientů ve dvou nebo více nezávislých souborech. V dodatku jsou stručně naznačeny důkazy hlavních vlastností.

I. Formulace úlohy

1. Je dán znak, který má K hodnot (kategorií).
2. Relativní četnosti hodnot znaku v cílové populaci[1] jsou známé a jsou rovny

[1] Nebo zde můžeme říci též: předpokládáme, že výběrové četnosti jsou rovny daným číslům. V tom případě nejde o úlohu statistického měření stupně odklonu od reprezentativity, ale o měření neshody empiricky zjištěných údajů od předpokladu, teorie, apriorního názoru apod. Obsah stati se bude dále soustřeďovat na problematiku reprezentativity, čtenář však může snadno úlohu přeformulovat pro potřeby hodnocení neshody předpokladu a zjištěné skutečnosti.

$$\pi_1, \pi_2, \ldots, \pi_K\,, \quad \Sigma\pi_i = 1\,, \quad \pi_i \geqq 0$$

pro všechna i.

3. Výběrový soubor obsahuje n pozorování tak, že v kategorii „i" je n_i pozorování ($i = 1, \ldots, K$), tj. rozdělení absolutních četností je

$$n_1, n_2, \ldots, n_K\,, \quad \Sigma n_i = n\,, \quad n_i \geqq 0$$

pro všechna i.

Výběrové relativní četnosti jsou

$$f_1 = \frac{n_1}{n}\,, \ldots, f_K = \frac{n_K}{n}\,, \quad \Sigma f_i = 1\,, \quad f_i \geqq 0$$

pro všechna i.

4. Výběrový soubor reprezentuje populaci, v níž jsou relativní četnosti znaku

$$p_1, p_2, \ldots, p_K\,, \quad \Sigma p_i = 1\,, \quad p_i \geqq 0$$

pro všechna i.

Hledáme vhodný koeficient ϱ, který by měřil odchylku rozdělení $(p_1, \ldots, p_K)$ od $(\pi_1, \ldots, \pi_K)$.

Vyjdeme ze vzdálenosti $D(\pi, p)$ dvou rozdělení π, p, která byla zavedena v článku J. Řeháka a B. Řehákové [1978] pro různé typy znaků. Koeficient neshody $\varrho = \varrho(\pi, p)$ mezi rozděleními π a p zavedeme tak, aby měl tyto jednoduché a rozumné vlastnosti:

a) $$\varrho(\pi, p) = 0$$

právě když $D(\pi, p) = 0$[2].

b) $$\varrho(\pi, p) = 1$$

právě když se rozdělení π a p liší maximálně[3] vzhledem ke vzdálenosti odpovídající typu znaku.

c) $$\varrho(\pi, p)$$

je rostoucí funkcí vzdálenosti rozdělení π a p. Tyto logické požadavky splňuje koeficient

(1) $$\varrho = \varrho(\pi, p) = \frac{D(\pi, p)}{\max\limits_{p} D(\pi, p)}$$

Jestliže neznáme rozdělení p, pak koeficient ϱ odhadneme[4] pomocí

(2) $$R = \varrho(\pi, f) = \frac{D(\pi, f)}{\max\limits_{f} D(\pi, f)}$$

Chceme-li zjišťovat interval spolehlivosti pro ϱ, pak musíme předpokládat, že výběrový soubor byl získán prostým náhodným výběrem, resp. postupem k němu

---

[2] Pro nominální a ordinální znak platí, že $D(\pi, p) = 0$ právě když $\pi_i = p_i$ pro všechna i, tj. populační a výběrové rozdělení je totožné. Pro kategorizovaný kardinální znak je $D(\pi, p) = 0$ právě když průměry rozdělení π a p jsou shodné.

[3] Maximální možná vzdálenost je zde chápána vzhledem k danému populačnímu rozdělení π.

[4] Tento odhad je vychýlený, přičemž pokus o opravu vychýlení vede k jiným nežádoucím vlastnostem. Vychýlení však klesá s rostoucí hodnotou ϱ a s počtem pozorování. Vzhledem k tomu, že koeficient budeme používat pro velké soubory, nebudeme se tímto nedostatkem zabývat.

ekvivalentním[5] (případ stratifikovaného výběru bude rozebrán později) a že alespoň pro jedno i je $\pi_i \neq p_i$.

100(1 — α)procentní interval spolehlivosti pro koeficient ϱ je

$$\left(R - z\sqrt{\frac{v(R)}{n}}\ ,\ R + z\sqrt{\frac{v(R)}{n}}\right) \tag{3}$$

kde n je výběrová velikost, $\sqrt{v(R)/n}$ je odhad směrodatné odchylky koeficientu R a z je 100(1 — $\alpha/2$)procentní kvantil normálního rozdělení.[6] Tento interval spolehlivosti určuje se spolehlivostí 100(1 — α) procent rozmezí, v jakém se pohybuje hodnota koeficientu $\varrho(\pi, p)$.

II. Koeficienty neshody

Uvedeme si nyní vzorce a příklady výpočtů koeficientů neshody pro tři základní typy znaků: nominální, ordinální a kategorizovaný kardinální. Zcela obdobně lze koeficient zavést i pro jiné typy znaků.

A. Nominální znak

Koeficient neshody $\varrho(\pi, p)$, kde π a p jsou rozdělení nominálního znaku, vychází ze vzdálenosti

$$D(\pi, p) = \sqrt{\sum_{i=1}^{K} (\pi_i - p_i)^2}\ , \tag{4}$$

přičemž

$$\max_p D(\pi, p) = \sqrt{(1 - \pi_{\min})^2 + \sum_{i=1}^{K} \pi_i^2 - \pi_{\min}^2} = \tag{5}$$

$$= \sqrt{1 - 2\pi_{\min} + \sum_{i=1}^{K} \pi_i^2}\ ,$$

kde $\pi_{\min} = \min(\pi_1, \pi_2, \ldots, \pi_K)$.

Podle (1), (2), (4), (5) je tedy

$$\varrho = \sqrt{\frac{\sum_{i=1}^{K} (\pi_i - p_i)^2}{1 - 2\pi_{\min} + \sum_{i=1}^{K} \pi_i^2}} \tag{6}$$

[5] Ekvivalentním postupem zde chápeme buď matematicky ekvivalentní výběr (např. některé případy systematického výběru), nebo výběr, který se empiricky chová vzhledem k výběrovým chybám jako prostý náhodný (např. některé případy kvótního výběru).

[6] Pro 95procentní spolehlivost je $z = 1{,}96$, pro 99procentní je $z = 2{,}58$, pro 99,9procentní je $z = 3{,}29$.

(7) $$R = \sqrt{\frac{\sum_{i=1}^{K} (\pi_i - f_i)^2}{1 - 2\pi_{\min} + \sum_{i=1}^{K} \pi_i^2}}$$

Odhad směrodatné odchylky koeficientu R je

(8) $$\sqrt{\frac{v(R)}{n}} = \sqrt{\frac{\sum_{i=1}^{K} f_i(f_i - \pi_i)^2 - \left[\sum_{i=1}^{K} f_i(f_i - \pi_i)\right]^2}{n\left[\sum_{i=1}^{K} (f_i - \pi_i)^2\right]\left[1 - 2\pi_{\min} + \sum_{i=1}^{K} \pi_i^2\right]}}$$

Příklad 1.

Ve výzkumu obyvatel ČSSR od 15 let byla zkoumána statistická reprezentativita šetření vzhledem ke znaku rodinný stav u žen. Data jsou uvedena v tabulce 1. Pomocí testu chí-kvadrát byla zamítnuta hypotéza shody.[7]

Tabulka 1. Výpočet R a jeho směrodatné odchylky pro znak „rodinný stav" u žen (n = 594)

Rodinný stav	Vdaná	Svobodná	Rozvedená nebo vdova	Součet
Populační proporce (π_i)	0,625 1	0,183 7	0,191 2	1,000 0
Výběrová proporce (f_i)	0,540 4	0,284 5	0,175 1	1,000 0
π_i^2	0,390 750	0,033 746	0,036 557	0,461 053
$(f_i - \pi_i)^2$	0,007 174	0,010 161	0,000 259	0,017 594
$f_i(f_i - \pi_i)^2$	0,003 877	0,002 891	0,000 045	0,006 813
$f_i(f_i - \pi_i)$	—0,045 772	0,028 678	—0,002 819	—0,019 913

(Údaje v posledních čtyřech řádcích jsou proti uvedeným v textu zaokrouhlené.)

Podle (7) je

$$R = \sqrt{\frac{0{,}01759394}{1 - 2 \,.\, 0{,}1837 + 0{,}46105314}} = \sqrt{\frac{0{,}01759394}{1{,}09365314}} = 0{,}1268 \doteq 0{,}13$$

a podle (8) je

$$\sqrt{\frac{v(R)}{n}} = \sqrt{\frac{0{,}00681297 - 0{,}00039654}{594 \,.\, 0{,}01759394 \,.\, 1{,}09365314}} = \sqrt{\frac{0{,}00641643}{11{,}42955063}} = 0{,}02369366$$

Podle (3) 95procentní interval spolehlivosti pro ϱ je $R \pm 1{,}96 \,.\, \sqrt{\frac{v(R)}{n}} = 0{,}1268 \pm 1{,}96 \,.\, 0{,}02369366 = 0{,}1268 \pm 0{,}0464 = (0{,}0804;\ 0{,}1732) \doteq (0{,}08;\ 0{,}17)$.

E. Ordinální znak

Pro ordinální znak definujeme vzdálenost dvou rozdělení jako

[7] Příklad byl uveden v článku B. Řehákové [1979].

(9) $$D(\pi, p) = \sqrt{2 \sum_{i=1}^{K-1} (\Pi_i - P_i)^2}$$

kde Π a P jsou distribuční funkce pro rozdělení π a p

$$(\Pi_i = \sum_{j=1}^{i} \pi_j,\ P_i = \sum_{j=1}^{i} p_j)\ .$$

Maximální hodnotu nabývá tato vzdálenost pro rozdělení $(0, 0, \ldots, 0, 1)$ nebo $(1, 0, \ldots, 0)$, tedy

(10) $$\max_p D(\pi, p) = \max\left(\sqrt{2 \sum_{i=1}^{K-1} \Pi_i^2}\ ,\ \sqrt{2 \sum_{i=1}^{K-1} (1 - \Pi_i)^2}\ \right).$$

Koeficient neshody ϱ je tedy

(11) $$\varrho = \sqrt{\frac{\sum_{i=1}^{K-1} (\Pi_i - P_i)^2}{\max\left(\sum_{i=1}^{K-1} \Pi_i^2,\ \sum_{i=1}^{K-1} (1 - \Pi_i)^2\right)}}$$

a jeho odhad R

(12) $$R = \sqrt{\frac{\sum_{i=1}^{K-1} (\Pi_i - F_i)^2}{\max\left(\sum_{i=1}^{K-1} \Pi_i^2,\ \sum_{i=1}^{K-1} (1 - \Pi_i)^2\right)}}$$

Odhad směrodatné odchylky koeficientu R je

(13) $$\sqrt{\frac{v(R)}{n}} =$$

$$= \sqrt{\frac{\sum_{k=1}^{K-1} (F_k - \Pi_k)^2 F_k(1 - F_k) + 2 \sum\sum_{k<j} (F_k - \Pi_k)(F_j - \Pi_j) F_k(1 - F_j)}{n\left[\sum_{k=1}^{K-1} (F_k - \Pi_k)^2\right]\left\{\max\left[\sum_{k=1}^{K-1} \Pi_k^2,\ \sum_{k=1}^{K-1} (1 - \Pi_k)^2\right]\right\}}}$$

Příklad 2.

Ve výzkumu obyvatel ČSSR od 15 let zkoumáme statistickou reprezentativitu šetření vzhledem ke znaku vzdělání. Data jsou uvedena v tabulce 2. Pomocí testu chí-kvadrát byla zamítnuta hypotéza shody.[8]

Podle vzorce (12) je

$$R = \sqrt{\frac{0{,}05091738}{2{,}54056819}} = 0{,}14$$

[8] Příklad byl uveden v článku B. Řehákové [1979].

Dosazením do (13) zjistíme, že

$$\sqrt{\frac{v(R)}{n}} = \sqrt{\frac{0{,}0114803534 + 0{,}0091623671}{1126 \,.\, 0{,}05091738 \,.\, 2{,}54056819}} = \sqrt{\frac{0{,}0206427205}{145{,}6583195}} = 0{,}0119046277$$

Podle (3) je 95procentní interval spolehlivosti pro koeficient neshody ϱ (za z dosadíme 1,96) roven (0,12; 0,16).

Tabulka 2. Výpočet koeficientu neshody R a jeho směrodatné odchylky pro znak „vzdělání" (n = 1126, K = 5)

Vzdělání	Základní nevyučen	Základní vyučen	Další bez maturity	Střední s maturitou	Součet
Kumulativní populační proporce (Π_k)	0,562 7	0,773 9	0,831 2	0,966 5	—
Kumulativní výběrové proporce (F_k)	0,372 1	0,673 2	0,766 5	0,950 3	—
$1 - F_k$	0,627 9	0,326 8	0,233 5	0,049 7	—
Π_k^2	0,316 631	0,598 921	0,690 893	0,934 122	2,540 568
$(1 - \Pi_k)^2$	0,191 231	0,051 121	0,028 493	0,001 122	0,271 968
$F_k - \Pi_k$	—0,190 6	—0,100 7	—0,064 7	—0,016 2	—
$(F_k - \Pi_k)^2$	0,036 328	0,010 140	0,004 186	0,000 262	0,050 917

(Údaje v posledních čtyřech řádcích jsou proti uvedeným v textu zaokrouhlené.)

C. Kategorizovaný kardinální znak

Distanční analýza kardinálních znaků je ekvivalentní práci s průměry. Jestliže jsou $x_1, \ldots, x_K$ číselné hodnoty znaku (např. středy třídících intervalů) uspořádané podle velikosti, pak

(14) $$D(\pi, p) = |\bar{X}_\pi - \bar{X}_p| \,,$$

kde $\bar{X}_\pi = \sum_{i=1}^{K} \pi_i x_i$, $\bar{X}_p = \sum_{i=1}^{K} p_i x_i$

Maximální dosažitelná vzdálenost je tedy

(15) $$\max_p D(\pi, p) = \max\,(|\bar{X}_\pi - x_1|, |x_K - \bar{X}_\pi|)$$
$$= \max\,(\bar{X}_\pi - x_1, x_K - \bar{X}_\pi)$$

Koeficient neshody, který měří neshodu průměrů rozdělení π a p, je tedy

(16) $$\varrho = \frac{|\bar{X}_\pi - \bar{X}_p|}{\max\,(\bar{X}_\pi - x_1, x_K - \bar{X}_\pi)}$$

a jeho odhad R

(17) $$R = \frac{|\bar{X}_\pi - \bar{X}_f|}{\max\,(\bar{X}_\pi - x_1, x_K - \bar{X}_\pi)}$$

Odhad směrodatné odchylky koeficientu R je[9]

$$\sqrt{\frac{v(R)}{n}} = \frac{\sqrt{\sum_{k=1}^{K} x_k^2 f_k - (\sum_{k=1}^{K} x_k f_k)^2}}{\sqrt{n}\ \max(\bar{X}_\pi - x_1, x_k - \bar{X}_\pi)} \tag{18}$$

Příklad 3.

Ve výzkumu obyvatel ČSSR od 18 do 29 let zkoumáme statistickou reprezentativitu šetření vzhledem ke znaku věk. Data jsou uvedena v tabulce 3. Pomocí testu chí-kvadrát byla zamítnuta hypotéza shody.[10]

Tabulka 3. Výpočet koeficientu neshody R a jeho směrodatné odchylky pro znak „věk" (n = 2400)

Věk	18—19	20—24	25—29	Celkem
Středy (x_i)	18,5	22	27	—
Populační proporce (π_i)	0,128 4	0,454 7	0,416 9	1
Výběrové proporce (f_i)	0,125 0	0,500 0	0,375 0	1
$\pi_i x_i$	2,375 4	10,003 4	11,256 3	23,635 1
$f_i x_i$	2,312 5	11,000 0	10,125 0	23,437 5
$f_i x_i^2$	42,781 25	242,000 00	273,375 00	558,156 25

Dosazením do vzorců (17) a (18) zjistíme, že

$$R = \frac{|23{,}6351 - 23{,}4375|}{5{,}1351} = 0{,}04$$

$$\sqrt{\frac{v(R)}{n}} = \frac{\sqrt{8{,}83984375}}{\sqrt{2{,}400} \cdot 5{,}1351} = 0{,}0118186$$

1,96 . 0,0118186 = 0,02 a dosazením do (3) zjistíme, že 95procentní interval spolehlivosti pro koeficient neshody ϱ je (0,02; 0,06).

III. Alternativní koeficient neshody

Weiler [1966] zavedl koeficient neshody φ na základě podobného principu normalizace distanční míry, avšak za míru vzdálenosti vzal statistiku chí-kvadrát. Pro úplnost je zde tento koeficient uveden a pro jeho odhad $\bar{\varphi}$ je též odvozen odhad směrodatné odchylky $\sqrt{\frac{v(\bar{\varphi})}{n}}$.

$$D(\pi, p) = \sqrt{\sum_{i=1}^{K} \frac{(\pi_i - p_i)^2}{\pi_i}} \tag{19}$$

[9] Stejný tvar vyplývá z odvození, při němž používáme obvyklou teorii pro průměr, založenou na přímé aplikaci centrálních limitních vět pro jakékoli rozdělení či pro situaci, v níž má zkoumaný znak hypoteticky normální rozdělení.

[10] Hodnota statistiky $X^2 = 21{,}15$, což je při 2 stupních volnosti významné na hladině $\alpha = 0{,}05$ (kritická hodnota je 5,99).

(20) $$\max_{p} D(\pi, p) = \sqrt{\frac{1 - \pi_{\min}}{\pi_{\min}}}, \quad \pi_{\min} = \min(\pi_1, \ldots, \pi_K)$$

(21) $$\varphi = \sqrt{\frac{\pi_{\min}}{1 - \pi_{\min}} \sum_{i=1}^{K} \frac{(\pi_i - p_i)^2}{\pi_i}}$$

(22) $$\bar{\varphi} = \sqrt{\frac{\pi_{\min}}{1 - \pi_{\min}} \sum_{i=1}^{K} \frac{(\pi_i - f_i)^2}{\pi_i}}$$

(23) $$\sqrt{\frac{v(\bar{\varphi})}{n}} = \sqrt{\frac{\pi_{\min}}{1 - \pi_{\min}} \frac{\sum_{i=1}^{K} \frac{f_i(f_i - \pi_i)^2}{\pi_i^2} - \left[\sum_{i=1}^{K} \frac{f_i(f_i - \pi_i)}{\pi_i}\right]^2}{n \sum_{i=1}^{K} \frac{(f_i - \pi_i)^2}{\pi_i}}}$$

Weiler [1966] rovněž podrobně diskutuje vychýlení $\bar{\varphi}$, zabývá se korekcemi a potížemi, které tyto korekce přinášejí. Uvádí též nomogramy pro přibližné určení intervalů spolehlivosti.

Pro ilustraci výpočtů použijeme dat z příkladu 1. Postup ukazuje tabulka 4.

Tabulka 4. Výpočet $\bar{\varphi}$ a směrodatné odchylky pro znak „rodinný stav" u žen *($n = 594$)*

Rodinný stav	Vdaná	Svobodná	Rozvedená nebo vdova	Celkem
Populační proporce (π_i)	0,625 1	0,183 7	0,191 2	1
Výběrová proporce (f_i)	0,540 4	0,284 5	0,175 1	1
$f_i(f_i - \pi_i)^2/\pi_i^2$	0,009 921 6	0,085 661 3	0,001 241 6	0,096 824 5
$(f_i - \pi_i)^2/\pi_i$	0,011 476 7	0,055 311 0	0,001 355 7	0,068 143 4

(Údaje v posledních dvou řádcích jsou proti uvedeným v textu zaokrouhlené.)

Dosazením do (22) a (23) zjistíme

$$\bar{\varphi} = \sqrt{\frac{0{,}1837}{0{,}8163} \cdot 0{,}0681434592} = 0{,}12$$

$$\sqrt{\frac{v(\bar{\varphi})}{n}} = \sqrt{\frac{0{,}1837}{0{,}8163} \cdot \frac{0{,}096824555 - 0{,}004643531}{594 \cdot 0{,}0681434592}} = 0{,}0226383689$$

$$1{,}96 \cdot 0{,}0226383689 = 0{,}04$$

95procentní interval spolehlivosti pro $\bar{\varphi}$ je (0,08; 0,16).

Když porovnáme tento výsledek ($\bar{\varphi} = 0{,}12$) s výsledkem u 1. příkladu ($R = 0{,}13$), vidíme že jsou velmi podobné. Vždy však tomu tak nemusí být. Bude-li $\pi_{\min}$ hodně malé, pak $\bar{\varphi}$ je přibližně rovno tomu f_i, které odpovídá $\pi_{\min}$, bez ohledu na to, jak velké jsou rozdíly mezi π_j a f_j v ostatních kategoriích, zatímco R (viz vzorec (7)) bere v úvahu i tyto rozdíly.

Budeme to ilustrovat na příkladu nominálního znaku s populačním rozdělením $\pi = (\pi_1, \pi_2, \pi_3)$ a s výběrovým rozdělením $f = (f_1, f_2, f_3)$, které udává tabulka 5.

Tabulka 5. Výpočet $\bar{\varphi}$ a R pro nominální znak

Hodnota znaku	1	2	3	Celkem
Populační proporce (π_i)	0,001	0,299	0,700	1
Výběrové proporce (f_i)	0,010	0,390	0,600	1
π_i^2	0,000 001	0,089 401	0,49	0,579 402
$(\pi_i - f_i)^2$	0,000 081	0,008 281	0,01	0,018 362
$\frac{(\pi_i - f_i)^2}{\pi_i}$	0,081	0,027 695 65	0,014 285 71	0,122 981 37

$$\bar{\varphi} = \sqrt{\frac{0{,}001}{1 - 0{,}001} \cdot 0{,}1229813663} = 0{,}01$$

$$R = \sqrt{\frac{0{,}018362}{1 - 0{,}002 + 0{,}579402}} = 0{,}11$$

($\bar{\varphi}$ počítáme podle (22), R podle (7)).

Vidíme, že R je mnohem větší než $\bar{\varphi}$, které se při zaokrouhlení na dvě desetinná místa rovná přesně hodnotě f_1, která odpovídá $\pi_{\min} = \pi_1$.

IV. Případ stratifikovaných výběrových postupů

Postupy části II a III lze používat pro případy, v nichž se výběrový mechanismus chová jako prostý náhodný výběr. V případě stratifikovaného výběru (resp. kvótního výběru, u nějž předpokládáme náhodné chování v částech určených kvótami) nemůžeme tyto postupy použít a musíme hledat jiné cesty.

Především je možné použít postup pro každou danou oblast zvlášť, čímž získáme vlastně podrobnější informace pro vyhodnocení reprezentativity. Zároveň můžeme navrhnout a zpracovat složitější statistické modely pro vyhodnocení reprezentativity celku. Nejjednodušší spočívá v tom, že zprůměrujeme koeficienty pro jednotlivé oblasti. Jsou-li $\varrho_1, \ldots, \varrho_M$ koeficienty neshody pro tentýž znak v oblastech $1, \ldots, M$, pak definujeme

$$\bar{\varrho} = \sum_{i=1}^{M} w_i \varrho_i , \quad \sum_{i=1}^{M} w_i = 1 , \quad w_i \geqq 0 , \quad i = 1, \ldots, M , \tag{24}$$

kde w_i jsou váhy (např. relativní četnosti zastoupení jednotlivých oblastí v populaci či ve výběru, nebo $w_i = 1/M$ pro všechna i).

Odhad směrodatné odchylky výběrového analogu $\bar{R}$ je

$$\sqrt{\frac{v(\bar{R})}{m}} = \sqrt{\sum_{i=1}^{M} w_i^2 \frac{v(R_i)}{m_i}} , \tag{25}$$

kde m_i jsou předem určené velikosti pro jednotlivé oblasti,

$$m = \sum_{i=1}^{M} m_i \quad \text{a} \quad \frac{v(R_i)}{m_i}$$

je druhá mocnina odhadu směrodatné odchylky koeficientu R_i, $i = 1, \ldots, M$. Hodnota (25) slouží k určení intervalu spolehlivosti koeficientu $\bar{\varrho}$ (viz vztah (3)).

Příklad 4.

V příkladu 1 jsme spočetli R a odhad jeho směrodatné odchylky pro znak rodinný stav u žen. Obdobně bychom spočetli tyto charakteristiky pro muže. Výsledky jsou souhrnně uvedeny v tabulce 6.

Tabulka 6. Údaje potřebné k výpočtu koeficientu $\bar{R}$ pro znak „rodinný stav" (w_1 a w_2 jsou relativní četnosti žen a mužů v populaci ČSSR od 15 let)

Ženy	Muži
$m_1 = 594$	$m_2 = 527$
$R_1 = 0{,}126\,835\,770\,4$	$R_2 = 0{,}049\,546\,207\,1$
$\frac{v(R_1)}{m_1} = 0{,}000\,561\,389\,5$	$\frac{v(R_2)}{m_2} = 0{,}000\,364\,204$
$w_1 = 0{,}519\,589\,078\,2$	$w_2 = 0{,}480\,410\,921\,8$

Výběrový soubor byl stratifikován, tj. soubory mužů a žen byly vybírány nezávisle a s předem danými požadavky na velikost. Koeficient nesbody $\bar{R}$ pro znak rodinný stav tedy spočítáme podle (24) s využitím údajů z tabulky 6.

$$\bar{R} = w_1R_1 + w_2R_2 = 0{,}08970502 \doteq 0{,}09$$

Podle (25) je

$$\sqrt{\frac{v(\bar{R})}{m}} = \sqrt{0{,}0002356162} = 0{,}0153497959$$

Dále $1{,}96\sqrt{\frac{v(\bar{R})}{m}} = 0{,}03$, takže 95procentní interval spolehlivosti pro $\bar{\varrho}$ je (0,06; 0,12).

V. Porovnání koeficientů neshody ve dvou nebo více nezávislých souborech

Úloha porovnání koeficientů neshody ve dvou nebo více nezávislých souborech je dosti častá, a to především pro metodologické cíle zhodnocení příčin nereprezentativnosti. Předpokládejme, že máme dva výzkumné (na sobě nezávislé) výběrové soubory, které se liší způsobem organizace tazatelských instrukcí, nebo výběrem tazatelů, či jiným zásahem do běžné rutiny, a že nás zajímá vliv tohoto zásahu na reprezentativitu.

Pro oba soubory a dané znaky[11] známe výběrové analogy $R_1 = \varrho_1(\pi_1, f_1)$, $R_2 = \varrho_2(\pi_2, f_2)$ koeficientů neshody $\varrho_1 = \varrho_1(\pi_1, p_1)$, $\varrho_2 = \varrho_2(\pi_2, p_2)$ a odhady směrodatných odchylek

$$\sqrt{\frac{v(R_1)}{n_1}}; \quad \sqrt{\frac{v(R_2)}{n_2}}.$$

Chceme testovat hypotézu

$$H_0 : \varrho_1(\pi_1, p_1) = \varrho_2(\pi_2, p_2)$$

proti alternativě

$$H_1 : \varrho_1(\pi_1, p_1) \neq \varrho_2(\pi_2, p_2)$$

[11] Porovnávat totiž můžeme i dva koeficienty, které nepatří stejnému znaku, pokud má taková úloha ovšem smysl. Nutné však je, aby oba koeficienty patřily k různým nezávisle vzniklým souborům. Rozdělení π_1 tedy může být jiné než π_2.

Testová statistika je

$$Z = \frac{R_1 - R_2}{\sqrt{\frac{v(R_1)}{n_1} + \frac{v(R_2)}{n_2}}} \tag{26}$$

Hodnotu $|Z|$ porovnáme s kritickou hodnotou normálního rozdělení z_α.[12]
Je-li $|Z| \geqq z_\alpha$, zamítáme hypotézu shody obou koeficientů na hladině významnosti α,

$|Z| < z_\alpha$, přijímáme hypotézu shody.

Jestliže místo alternativy H_1 vezmeme apriori jednostrannou alternativu[13]

$$H_2 : \varrho_1(\pi_1, p_1) > \varrho_2(\pi_2, p_2)$$

pak je-li

$$Z \geqq z_\alpha ,$$

zamítáme hypotézu shody H_0 ve prospěch alternativy H_2,

$$Z < z_\alpha ,$$

přijímáme H_0.

Jestliže vezmeme apriori jednostrannou alternativu[13]

$$H_3 : \varrho_1(\pi_1, p_1) < \varrho_2(\pi_2, p_2)$$

pak je-li

$$-Z \geqq z_\alpha ,$$

zamítáme H_0 ve prospěch H_3,

$$-Z < z_\alpha ,$$

přijímáme H_0.

U alternativ H_2, H_3 jsou z_α hodnoty pro jednostranný normální test.[14]

Příklad 5

Ve dvou výzkumech byla zkoumána reprezentativita vzhledem k několika znakům. Výsledky udává tabulka 7.[15] Výpočet testové statistiky Z si ukážeme pro znak E.

$$Z = \frac{0{,}085 - 0{,}052}{\sqrt{0{,}000091 + 0{,}000085}} = 2{,}49$$

Protože $2{,}49 > 1{,}96$, zamítáme hypotézu shody ϱ_1 a ϱ_2 ve prospěch alternativy $\varrho_1 \neq \varrho_2$ na hladině významnosti 0,05.

---

[12] $z_{0{,}05} = 1{,}96$, $z_{0{,}01} = 2{,}58$, $z_{0{,}001} = 3{,}29$.

[13] Zde je podstatné, že jednostranná alternativa musí být formulována před pohledem na data.

[14] Pro jednostranný test $z_{0{,}05} = 1{,}65$, $z_{0{,}01} = 2{,}33$, $z_{0{,}001} = 3{,}09$.

[15] U znaku D byl v 1. souboru chí-kvadrát nevýznamný na hladině významnosti 0,05, zatímco ve 2. souboru byl významný. Nebylo proto třeba provádět testování. Někdy se však může stát, že v tomto případě vyjde test na shodu koeficientů nevýznamně.

Tabulka 7. Porovnání 2 koeficientů neshody pro stejné znaky a různé soubory

Znak	Soubor 1		Soubor 2		Z	Výsledek dvoustranného testu při $\alpha = 0{,}05$
	R_1	$\frac{v(R_1)}{n_1}$	R_2	$\frac{v(R_2)}{n_2}$		
A	0,124	0,000 27	0,115	0,000 31	0,37	nevýznamný
B	0,151	0,000 32	0,183	0,000 35	—1,24	nevýznamný
C	0,143	0,000 25	0,189	0,000 22	—2,12	významný
D	0,011	0,000 090	0,046	0,000 18	—2,13	významný
E	0,085	0,000 091	0,052	0,000 085	2,49	významný
F	0,153	0,000 075	0,140	0,000 080	1,05	nevýznamný

Obdobně lze využít pro hledání příčin nereprezentativnosti testování rovnosti koeficientů pro dvě oblasti téhož výzkumu. Například pro data z příkladu 1 a 4 můžeme porovnat odklon od reprezentativity u mužů a žen. Potřebné údaje k provedení testu nalezneme v tabulce 6. Dosazením do vzorce (26) zjistíme, že $Z = 2{,}54 > 1{,}96$, a proto zamítáme hypotézu rovnosti koeficientů neshody u znaků rodinný stav pro ženy a rodinný stav pro muže.

Chceme-li porovnat více nezávislých koeficientů neshody (např. porovnáme několik výzkumů či několik oblastí v rámci jednoho výzkumu), pak to můžeme provést následujícím postupem:

Nulová hypotéza

$$H_0 : \varrho_1(\pi_1, p_1) = \varrho_2(\pi_2, p_2) = \ldots = \varrho_M(\pi_M, p_M) .$$

Alternativní hypotéza

$$H_1 : \text{alespoň jedno z } \varrho_i \text{ je různé od ostatních} .$$

Spočteme vážený průměr

$$\bar{R} = \sum_{i=1}^{M} \frac{R_i}{\frac{v(R_i)}{m_i}} \Bigg/ \sum_{i=1}^{M} \frac{1}{\frac{v(R_i)}{m_i}} = \sum_{i=1}^{M} \frac{m_i R_i}{v(R_i)} \Bigg/ \sum_{i=1}^{M} \frac{m_i}{v(R_i)}$$

a testovací statistiku

$$H = \sum_{i=1}^{M} \frac{(R_i - \bar{R})^2}{\frac{v(R_i)}{m_i}} = \sum_{i=1}^{M} \frac{m_i (R_i - \bar{R})^2}{v(R_i)} = \tag{27}$$

$$= \sum_{i=1}^{M} \frac{m_i}{v(R_i)} R_i^2 - \left(\sum_{i=1}^{M} \frac{m_i}{v(R_i)} R_i \right)^2 \Bigg/ \sum_{i=1}^{M} \frac{m_i}{v(R_i)}$$

Statistika H má za nulové hypotézy H_0 přibližně rozdělení chí-kvadrát s $M - 1$ stupni volnosti. Proto je-li

$H \geqq$ kritická hodnota chí-kvadrát pro α a $M - 1$, zamítáme H_0 na hladině významnosti α,

$H <$ kritická hodnota chí-kvadrát pro α a $M - 1$, přijímáme H_0 na hladině významnosti α.

Jako příklad porovnáme čtyři koeficienty neshody.

Příklad 6.

Ve čtyřech výzkumech ($m_1 = 1872$, $m_2 = 1757$, $m_3 = 1895$, $m_4 = 1821$) byly zjištěny koeficienty neshody pro vybrané znaky. Výsledky jsou uvedeny v tabulce 8.

Tabulka 8. Porovnání čtyř koeficientů neshody pro stejné znaky a různé soubory

Znak	Soubor 1		Soubor 2		Soubor 3		Soubor 4		H	Významné $\alpha = 0{,}05$
	R_1	$v_1(R_1)$	R_2	$v_2(R_2)$	R_3	$v_3(R_3)$	R_4	$v_4(R_4)$		
A	0,15	0,1334	0,14	0,1589	0,12	0,1438	0,16	0,1495	11,26	ano
B	0,08	0,1283	0,08	0,1301	0,07	0,1085	0,09	0,1311	3,11	ne
C	0,12	0,1825	0,13	0,1638	0,11	0,1732	0,11	0,1756	2,94	ne

Výpočet statistiky H si ukážeme na znaku A.

$$H = \frac{1872}{0{,}1334}\,0{,}15^2 + \ldots + \frac{1821}{0{,}1495}\,0{,}16^2 -$$

$$- \left(\frac{1872}{0{,}1334}\,0{,}15 + \ldots + \frac{1821}{0{,}1495}\right)^2 \Big/ \left(\frac{1872}{0{,}1334} + \ldots + \frac{1821}{0{,}1495}\right) =$$

$$= 1034{,}051568 - \frac{51598713{,}82}{50448{,}87927} = 11{,}26$$

Kritická hodnota chí-kvadrát pro $\alpha = 0{,}05$ a tři stupně volnosti je $7{,}81 < 11{,}26$, proto zamítáme hypotézu rovnosti koeficientů $\varrho_1, \ldots, \varrho_4$ pro znak A.

Závěr

Cílem této stati je seznámit zainteresovanou sociologickou veřejnost s možností měření stupně neshody dvou rozdělení četností téhož znaku, z nichž jedno je předem dané a druhé je získané empiricky, přičemž jsou rozlišeny znaky nominální, ordinální a kategorizované kardinální.

Použití zavedených koeficientů neshody není omezeno jen na ověřování reprezentativity výběrových šetření, ale lze jimi měřit též stupeň neshody empiricky zjištěných údajů a předpokladu, teorie či apriorního názoru.

Intervaly spolehlivosti pro koeficienty neshody jsou odvozeny za předpokladu prostého náhodného a stratifikovaného výběru.

Metody této stati spolu s metodami z článku B. Řehákové [1979] dávají dobrou možnost ověřování reprezentativity výběrových šetření u těch znaků, pro něž jsou známé populační charakteristiky. Mohou též přispět k odhalování příčin porušování reprezentativity. Hodí se především v těch případech, kdy výzkumné závěry zobecňujeme na konkrétní populace (např. obyvatelé ČSSR od 18 do 55 let).

Žádné, nebo jen velmi omezené uplatnění mají v případových studiích, ve výzkumech závodů, ve výzkumech populací, které nelze přesně vymezit. Rovněž je nelze použít u malých výběrových souborů.

Metody zkoumání reprezentativity zařazujeme před vlastní analýzu dat pro sociologickou interpretaci. Jejich důležitost spočívá

— ve zhodnocení reprezentativity alespoň u těch znaků, u nichž máme podklady pro testování shody,
— v indikaci nutnosti použití systému vah při výpočtu statistických charakteristik, aby bylo alespoň částečně odstraněno vychýlení,

— v metodologické interpretaci výsledků, která slouží jako základ pro zlepšení metodologie plánování výběrových šetření a sběru dat v dalších výzkumech.

Hodnocení reprezentativity má tedy dvě podstatné role: skýtá podklad pro konkrétní opatření v analýze dat a je součástí nutné metodologické reflexe výzkumného procesu.

V typech výzkumů, kde lze uvedených metod použít, by se mělo hodnocení reprezentativity stát součástí metodologické analýzy dat.

Vzhledem k důležitosti formulované úlohy byly metody hodnocení reprezentativity shrnuty do programu HOR (autoři Řeháková B., Dvořák P., Herzmann J.) a stávají se částí rutinního zpracování výzkumů prováděných Ústavem pro výzkum veřejného mínění.

Dodatek

V tomto technickém dodatku jsou uvedeny důkazy vlastností, na nichž jsou založeny metody článku.

Lema 1.

$$\max_{p}\left[\sqrt{\sum_{i=1}^{K}(\pi_i - p_i)^2}\right] = \sqrt{1 - 2\pi_{\min} + \sum_{i=1}^{K}\pi_i^2}\,,$$

kde $\pi = (\pi_1, \ldots, \pi_K)$, $p = (p_1, \ldots, p_K)$ jsou rozdělení četností a $\pi_{\min} = \min(\pi_1, \ldots, \pi_K)$.

Důkaz:

$$\sum_{i=1}^{K}(\pi_i - p_i)^2 = \sum_{i=1}^{K}\pi_i^2 + \sum_{i=1}^{K}p_i^2 - 2\sum_{i=1}^{K}\pi_i p_i \leqq \sum_{i=1}^{K}\pi_i^2 + \sum_{i=1}^{K}p_i - 2\pi_{\min}\sum_{i=1}^{K}p_i =$$

$$= \sum_{i=1}^{K}\pi_i^2 + 1 - 2\pi_{\min}$$

Přitom rovnost nastává, právě když pro j určené vztahem $\pi_j = \pi_{\min}$ platí $p_j = 1$.

Q. E. D.

Lema 2.

Nechť Π a P jsou kumulativní distribuční funkce pro rozdělení

$$\pi = (\pi_1, \ldots, \pi_K) \quad \text{a} \quad p = (p_1, \ldots, p_K)\,, \quad \Pi_i = \sum_{j=1}^{i}\pi_j\,, \quad P_i = \sum_{j=1}^{i}p_j\,.$$

Pak

$$\max_{p}\left[\sqrt{2\sum_{i=1}^{K-1}(\Pi_i - P_i)^2}\right] = \max\left(\sqrt{2\sum_{i=1}^{K-1}\Pi_i^2},\ \sqrt{2\sum_{i=1}^{K-1}(1-\Pi_i)^2}\right).$$

Důkaz:

Buď $\mathscr{P}$ množina všech $(P_1, \ldots, P_K)$ takových, že $P_i \leqq P_{i+1}$, $P_K = 1$. Tato množina je zadaná lineárními nerovnostmi, tvoří tedy konvexní polyedr s vrcholy $V_1 = (1, 1, \ldots, 1)$, $V_2 = (0, 1, \ldots, 1), \ldots, V_K = (0, 0, \ldots, 0, 1)$.

$D^2(P) = \sum_{i=1}^{K-1}(\Pi_i - P_i)^2$ je konvexní funkce na $\mathscr{P}$, proto může nabývat maxima pouze ve vrcholech polyedru V_i.

Označme

$$S_i = D^2(V_i) = \sum_{j=1}^{i-1}\Pi_j^2 + \sum_{j=i}^{K-1}(1 - \Pi_j)^2\,,$$

$$i = 2, \ldots, K-1,$$

$$S_K = D^2(V_K) = \sum_{j=1}^{K-1} \Pi_j^2, \quad S_1 = \sum_{j=1}^{K-1} (1 - \Pi_j)^2$$

Jelikož tedy $\max_{i=1,\ldots,K} D^2(P) = \max_{i=1,\ldots,K} S_i$, budeme dále hledat $\max_{i=1,\ldots,K} S_i$.

Pro $i = 1, \ldots, K-1$ platí, že

$$S_i = \sum_{j=1}^{K-1} \Pi_j^2 + 2\left[\frac{K-i}{2} - \sum_{j=i}^{K-1} \Pi_j\right]$$

$$S_{i+1} = S_i + 2(\Pi_i - \tfrac{1}{2})$$

Nyní mohou nastat dva případy

a) $\Pi_1 > \frac{1}{2} \Rightarrow \Pi_i > \frac{1}{2}$ pro všechna $i \Rightarrow S_1 < S_2 < \ldots < S_K$, a tedy $\max_{i=1,\ldots,K} S_i = S_K$.

b) $\Pi_1 \leqq \frac{1}{2} \Rightarrow$ existuje $1 < k \leqq K$ tak, že

$$\Pi_k > \tfrac{1}{2} \Rightarrow S_1 \geqq S_2 \geqq \ldots \geqq S_k < S_{k+1} < \ldots < S_K$$

a tedy

$$\max_{i=1,\ldots,K} S_i = \max(S_1, S_K).$$

Q. E. D.

Lema 3.

Nechť $x_1 \leqq \ldots \leqq x_K$ jsou číselné hodnoty znaku X a $\pi = (\pi_1, \ldots, \pi_K)$, $p = (p_1, \ldots, p_K)$ jsou dvě rozdělení tohoto znaku.

Označme

$$\bar{X}_\pi = \sum_{i=1}^{K} \pi_i x_i, \quad \bar{X}_p = \sum_{i=1}^{K} p_i x_i.$$

Pak

$$\max |\bar{X}_\pi - \bar{X}_p| = \max(\bar{X}_\pi - x_1, x_K - \bar{X}_\pi).$$

Důkaz:

$$|\bar{X}_\pi - \bar{X}_p| = \begin{cases} \bar{X}_\pi - \bar{X}_p & \text{pro } \bar{X}_\pi \geqq \bar{X}_p \\ \bar{X}_p - \bar{X}_\pi & \text{pro } \bar{X}_\pi \leqq \bar{X}_p \end{cases}$$

$$\bar{X}_\pi - \bar{X}_p \leqq \bar{X}_\pi - x_1, \quad \bar{X}_p - \bar{X}_\pi \leqq x_K - \bar{X}_\pi$$

a odtud již je tvrzení zřejmé.

Q. E. D.

Lema 4.

Označíme-li ϱ a R po řadě jako v [(6), (7)], [(11), (12)], [(16), (17)], a předpokládáme-li, že $\pi_i \neq p_i$ alespoň pro jedno i, pak asymptotické rozdělení veličin

$$\sqrt{\frac{n}{v(R)}} \cdot (R - \varrho)$$

je při $v(R) \neq 0$ normální s nulovým průměrem a jednotkovým rozptylem, kde $v(R)$ je po řadě určeno ze vztahů (8), (13), (18).

Důkaz tohoto tvrzení přesahuje účel článku. Je založen na asymptotické teorii rozdělení četností (viz Rao [1978], kapitola 6 a .2).

Literatura

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

Řehák, J.: *K pojmu „reprezentativita" v sociologických výzkumech.* Sociologický časopis 1978, č. 5, s. 489–507.

Řehák, J. - Řeháková, B.: *Základní charakteristiky pro měnných s konečným počtem hodnot a distanční analýza jejich rozložení.* Sociologický časopis 1979, č. 2, s. 214—233.
Viz též: *Basic Characteristics for Finite — Valued Variables and a Distance Analysis of Their Distributions*, příspěvek na IX. světovém sociologickém kongresu, Uppsala, 1978.
Řeháková, B.: *Statistické ověřování reprezentativity: testy dobré shody.* Sociologický časopis 1979, č. 6, s. 615—629.
Weiler, H.: *A Coefficient Measuring the Goodness of Fit.* Technometrics 1966, č. 2, s. 327—334.

Резюме

Б. Ржегакова: Статистическая проверка репрезентативности: коэффициенты несогласия

В данной статье предлагается новый метод анализа статистической репрезентативности, в основе которого лежат коэффициенты несогласия между популяционным и выборочным распределением. Эти коэффициенты выведены из общей теории для переменных с конечным числом значений [Řehák - Řeháková 1979].

В первой части дана конструкция коэффициентов несогласия. Формула основана на нормализации дистанции в отношении к достижимому максимуму для данного типа переменной и для проверяемого популяционного распределения. Коэффициент R (1) достигает нижнюю грань «ноль» в случае полного согласия (в смысле данного типа переменной) и, наоборот, верхнюю грань «один» достигает в случае максимального несогласия. В практической исследовательской работе используются выборочные коэффициенты (2). Для всех случаев даны интервалы достоверности (3).

Во второй части выведены из совместной общей формулы три специальные коэффициента:

А) для номинальных переменных ((4)—(8))
Б) для ординальных переменных ((9)—(13))
В) для кардинальных (численных) переменных ((14)—(18))

В части III дискутируется коэффициент, введенный Вайлером [Weiler 1966], учитывая некоторые его неудовлетворительные особенности.

Часть IV содержит решение для случая стратификационной выборки.

В части V указываются статистические критерии, при помощи которых можно сравнивать два или больше коэффициентов. Эти критерии ведут к нормальному распределению или к распределению хи-квадрат.

В подавлении даны четыре лемы, которые обеспечивают точный смысл для интерпретации введенных коэффициентов.

Все методы проилюстрированы посредством нумерических примеров. Таким образом статья может служить также руководством для практической работы социолога.

Summary

B. Řeháková: Statistical Verification of Representativity: Coefficients for Badness of Fit

The new method for evaluating statistical reprezentativity is presented that is based on coefficients measuring the badness of fit between populational and sample distributions. These coefficients follow a theory for finite-valued variables derived in Řehák, Řeháková [1979]. The general construction of coefficients that is presented in part I stems from a distance measure for a given type of variable that is normalized against its attainable maximum. Coefficient R (1) is defined in a way to attain zero in a case of a perfect fit and one in a case of the maximum deviation (both in sense of a given metrics or semimetrics). A sample analogue is used for practical work (2). In all cases the asymptotic standard error and confidence bounds (3) are given.

Derived from the common approach are three different measures for three different types of variables:

A) for nominal variables ((4)—(8))
B) for ordinal variables ((9)—(13))
C) for cardinal variables ((14)—(18))

in part II.

In part III, the measure by Weiler [1966] is discussed and some its undesireble properties are displayed. While the confidence limits in part II were derived for a simple random sample, in part IV the stratified random sample case is solved. Part V deals with the comparison of two or more coefficients. The suggested statistical tests result in normal and chi-square distributions.

Appendix contains four lemmas assuring the meaningfulness of presented coefficients.

A number of numerical illustrations are supposed to help in orientation and as a guide for a practical work of sociologists.