

Czesław Domański, Lechosław Stępień

Uniwersytet Łódzki

OCENA OBCIĄŻENIA WSPÓŁCZYNNIKÓW WSPÓLZALEŻNOŚCI DLA TABLIC KONTYNGENCJI

1. Model statystyczny

Założmy, że interesuje nas zależność między pewnymi dwiema cechami niemierzalnymi X i Y . Symbole n_{ij} , $n_{i.}$, $n_{.j}$, n , p_{ij} , $p_{i.}$, $p_{.j}$, gdzie $i, j = 1, 2$ oznaczają odpowiednie liczebności i częstości dwuwymiarowej tablicy kontyngencyjnej zaobserwowane w próbie (tab. 1 i 2).

Tabela 1

X	Y		$n_{i.}$
	y_1	y_2	
x_1	n_{11}	n_{12}	$n_{1.}$
x_2	n_{21}	n_{22}	$n_{2.}$
$n_{.j}$	$n_{.1}$	$n_{.2}$	n

Tabela 2

X	Y		$p_{i.}$
	y_1	y_2	
x_1	p_{11}	p_{12}	$p_{1.}$
x_2	p_{21}	p_{22}	$p_{2.}$
$p_{.j}$	$p_{.1}$	$p_{.2}$	1

Częstości w populacji generalnej oznaczamy symbolami ρ_{ij} , $\rho_{i.}$, $\rho_{.j}$. Tworzą one tablicę kontyngencyjną dwuwymiarową (tab. 3).

Tabela 3

X	Y		$q_{i.}$
	y_1	y_2	
x_1	ρ_{11}	ρ_{12}	$\rho_{1.}$
x_2	ρ_{21}	ρ_{22}	$\rho_{2.}$
$q_{.j}$	$\rho_{.1}$	$\rho_{.2}$	1

Liczba ρ_{ij} oznacza prawdopodobieństwo zdarzenia polegającego na tym, że losowo wybrany z populacji generalnej element należy do i -tej klasy ze względu na cechę X i do j -tej klasy ze względu na cechę Y . Dlatego też liczby zawarte

w tab. 3 będziemy zwykle nazywali prawdopodobieństwami, częstości zaś w próbie p_{ij} – po prostu częstościami.

Rozważania teoretyczne oparte są na założeniu losowania niezależnego. Zatem liczebność populacji generalnej nie wpływa na sformułowane wnioski. Dlatego też nie czynimy żadnego założenia na temat liczebności populacji generalnej (może być ona przede wszystkim nieskończona).

Dalsza część opracowania dotyczyć będzie niektórych, spośród znanych w literaturze, mierników współzależności dla tablic kontyngencyjnych 2×2 ([2; 3; 4]), przy czym mierniki obliczone w próbie oznaczmy symbolami z daszkiem (ogólnie $\hat{w}$), a odpowiednim miernikom w populacji generalnej przyporządkujemy te same litery bez daszków (w).

Zgodnie z przedstawionymi założeniami, prawdopodobieństwa ρ_{ij} , ρ_i , ρ_j oraz wskaźniki w są wielkościami stałymi, charakterystykami liczbowymi rozkładu. Natomiast zaobserwowane w próbie liczebności n_{ij} , częstości p_{ij} oraz wskaźniki $\hat{w}$ są zmiennymi losowymi. W innej próbie o tej samej liczebności (o ile wylosowanie takiej próbki byłoby możliwe i celowe) mogłyby przyjąć inne wartości. Nie jest losową liczebność n .

2. Estymacja punktowa

Podstawowym problemem, na jaki natrafia badacz wnioskujący na podstawie próby o zależności dwóch cech jakościowych w populacji generalnej, jest możliwość oceny stopnia tej zależności. Nasze rozważania rozpoczniemy zatem od najprostszego sposobu takiej oceny, jakim jest estymacja punktowa, polegająca ogólnie na ocenie parametru populacji generalnej za pomocą pewnej funkcji wyników z próby, czyli pewnego estymatora.

Najbardziej uniwersalną metodą poszukiwania dobrych estymatorów jest metoda największej wiarygodności. Nie wnikając w teorię związaną z tą metodą, zastosujemy dwa znane w literaturze statystycznej fakty.

Estymatorem największej wiarygodności prawdopodobieństwa ρ_{ij} jest odpowiednia częstość z próbki.

$$p_{ij} = \frac{n_{ij}}{n}. \quad (1)$$

Poza tym skorzystamy z twierdzenia głoszącego, że estymatorem największej wiarygodności funkcji parametrów populacji generalnej jest taka sama funkcja estymatorów największej wiarygodności tych parametrów, o ile funkcja ta jest różniczkowalna. Ponieważ wskaźnik $\hat{w}$ z próby jest taką samą funkcją częstości p_{ij} jak odpowiedni wskaźnik w populacji generalnej

$$w = f(\rho_{11}, \rho_{12}, \rho_{21}, \rho_{22}), \quad (2)$$

$$\hat{w} = f(p_{11}, p_{12}, p_{21}, p_{22}), \quad (3)$$

więc $\hat{w}$ jest estymatorem największej wiarygodności w .

Metoda największej wiarygodności doprowadza do najbardziej „naturalnych” estymatorów mierników zależności. Własności tych estymatorów dla dużych prób są następujące:

- są one asymptotycznie nieobciążone,
- są zgodne,
- są asymptotycznie najefektywniejsze,
- mają asymptotyczny rozkład normalny.

Sprawą otwartą jest natomiast problem budowy estymatorów mierników zależności dla małych prób.

Dodajmy na koniec, że dla cech jakościowych zastosowanie popularnej i wygodnej metody momentów do budowy estymatorów się nie stosuje.

3. Ocena obciążenia współczynników z próby

W poprzednim punkcie zauważyliśmy, że mierniki z próby są asymptotycznie nieobciążonymi estymatorami odpowiednich parametrów w populacji generalnej. Innymi słowy, wielkość obciążenia maleje do zera, gdy liczność próby dąży do nieskończoności. Interesująca jest więc kwestia, jak duże jest obciążenie estymatorów dla małych prób.

Można częściowo odpowiedzieć na to pytanie (por. [4]), wykorzystując rozwinięcia mierników dla tablic 2×2 w szereg Taylora w otoczeniu punktu

$$\mu = (\mu_a, \mu_b, \mu_c, \mu_d), \quad (4)$$

gdzie

$$\mu_a = n\rho_a, \quad \mu_b = n\rho_b, \quad \mu_c = n\rho_c, \quad \mu_d = n\rho_d, \quad (5)$$

a

$$\rho_a = \rho_{11}, \quad \rho_b = \rho_{12}, \quad \rho_c = \rho_{21}, \quad \rho_d = \rho_{22}, \quad (6)$$

i podobnie oznaczamy

$$p_a = p_{11}, \quad p_b = p_{12}, \quad p_c = p_{21}, \quad p_d = p_{22}. \quad (7)$$

Rozwinięcie Taylora dla miernika z próby

$$\hat{w} = f(p_a, p_b, p_c, p_d) \quad (8)$$

przybiera postać

$$\begin{aligned} f(p_a, p_b, p_c, p_d) &= f(\mu_a + h_a, \mu_b + h_b, \mu_c + h_c, \mu_d + h_d) = \\ &= f(np_{11} + h_a, np_{12} + h_b, np_{21} + h_c, np_{22} + h_d). \end{aligned} \quad (9)$$

Dlatego

$$\begin{aligned}\hat{w} = & f(\boldsymbol{\mu}) + [f'_a(\boldsymbol{\mu})h_a + f'_b(\boldsymbol{\mu})h_b + f'_c(\boldsymbol{\mu})h_c + f'_d(\boldsymbol{\mu})h_d] + \\ & + \frac{1}{2!} [f''_{aa}(\boldsymbol{\mu})h_a^2 + f''_{bb}(\boldsymbol{\mu})h_b^2 + f''_{cc}(\boldsymbol{\mu})h_c^2 + f''_{dd}(\boldsymbol{\mu})h_d^2 + \\ & + 2f''_{ab}(\boldsymbol{\mu})h_a h_b + 2f''_{ac}(\boldsymbol{\mu})h_a h_c + f''_{ad}(\boldsymbol{\mu})h_a h_d + 2f''_{bc}(\boldsymbol{\mu})h_b h_c + \\ & + 2f''_{bd}(\boldsymbol{\mu})h_b h_d + 2f''_{cd}(\boldsymbol{\mu})h_c h_d +] + R_3.\end{aligned}\quad (10)$$

Zauważmy, że dla $x, y = a, b, c, d$

$$f(\boldsymbol{\mu}), f'_x(\boldsymbol{\mu}), f''_{xy}(\boldsymbol{\mu}) \quad (11)$$

są wielkościami stałymi, a

$$h_x = p_x - \mu_x \quad (12)$$

są zmiennymi losowymi, przy czym

$$Eh_x = E_{p_x} - \mu_x = \mu_x - \mu_x = 0 \quad (13)$$

oraz

$$Eh_x h_y = E(p_x - \mu_x)(p_y - \mu_y) = \text{cov}(p_x, p_y) \quad (14)$$

są kowariancjami par zmiennych losowych p_x, p_y , przede wszystkim zaś

$$Eh_x^2 = E(p_x - \mu_x)^2 = \text{var } p_x. \quad (15)$$

Na mocy definicji

$$\text{var } p_x = \frac{\rho_x(1 - \rho_x)}{n}, \quad (16)$$

$$\text{COV}(p_x, p_y) = \frac{\rho_x \rho_y}{n}. \quad (17)$$

Biorąc zatem pod uwagę oczekiwane wartości obu stron (10), po uwzględnieniu (14) i (15) oraz pominięciu reszty R_3 , otrzymujemy w przybliżeniu

$$\begin{aligned}E\hat{w} - f(\boldsymbol{\mu}) \approx & \left[f'_a(\boldsymbol{\mu})Eh_a + f'_b(\boldsymbol{\mu})Eh_b + f'_c(\boldsymbol{\mu})Eh_c + f'_d(\boldsymbol{\mu})Eh_d \right] + \\ & + \frac{1}{2!} [f''_{aa}(\boldsymbol{\mu})E(p_a - \mu_a)^2 + f''_{bb}(\boldsymbol{\mu})E(p_b - \mu_b)^2 + f''_{cc}(\boldsymbol{\mu})E(p_c - \mu_c)^2 + \\ & + f''_{dd}(\boldsymbol{\mu})E(p_d - \mu_d)^2 + 2f''_{ab}(\boldsymbol{\mu})E(p_a - \mu_a)(p_b - \mu_b) + 2f''_{ac}(\boldsymbol{\mu})E(p_a - \mu_a)(p_c - \mu_c) +\end{aligned}$$

$$+ 2f''_{ad}(\mu)E(p_a - \mu_a)(p_d - \mu_d) + 2f''_{bc}(\mu)E(p_b - \mu_b)(p_c - \mu_c) + \\ + 2f''_{bd}(\mu)E(p_b - \mu_b)(p_d - \mu_d) + 2f''_{cd}(\mu)E(p_c - \mu_c)(p_d - \mu_d)] . \quad (18)$$

Ponieważ $f(\mu) = w$, więc lewa strona (18) jest obciążeniem estymatora w . Zatem podstawiając po prawej stronie (18) w miejsce f'_x, f''_{xy} pierwsze i drugie pochodne rozważanych współczynników, otrzymujemy przybliżone wzory na ich obciążenie.

Obciążenie współczynnika Q -Yule'a danego wzorem

$$Q = \frac{\rho_a \rho_d - \rho_b \rho_c}{\rho_a \rho_d + \rho_b \rho_c} \quad (19)$$

obliczymy, podstawiając do wzoru (18) pierwsze i drugie pochodne funkcji $f(Q)$ względem p_a, p_b, p_c, p_d , gdzie np.

$$f'_a = \frac{2p_b p_c p_d}{(p_a p_d + p_b p_c)^2}, \quad f''_{aa} = \frac{-4p_b p_c p_d^2}{(p_a p_d + p_b p_c)^3}, \quad f''_{ab} = \frac{2p_c p_d (p_a p_d - p_b p_c)}{(p_a p_d + p_b p_c)^3}. \quad (20)$$

Po wykonaniu serii przekształceń otrzymamy

$$E\hat{Q} - Q \approx \frac{1}{2(p_a p_d + p_b p_c)^3 n} \{ -4p_b p_c p_d^2 p_a (1 - p_a) + \\ + 4p_a p_c^2 p_d p_b (1 - p_b) + 4p_a p_b^2 p_d p_c (1 - p_c) - 4p_a^2 p_b p_c p_d (1 - p_d) + \\ - 4p_a p_b p_c p_d (p_a p_d - p_b p_c) - 4p_a p_b p_c p_d (p_a p_d - p_b p_c) + \\ + 4p_a p_b p_c p_d (p_a p_d - p_b p_c) + 4p_a p_b p_c p_d (p_a p_d - p_b p_c) = \\ - 4p_a p_b p_c p_d (p_a p_d - p_b p_c) - 4p_a p_b p_c p_d (p_a p_d - p_b p_c) = \\ - 4p_a p_b p_c p_d (p_a p_d - p_b p_c) - 4p_a p_b p_c p_d (p_a p_d - p_b p_c) = \\ = \frac{2p_a p_b p_c p_d (-p_a + p_b + p_c - p_d)}{n(p_a p_d + p_b p_c)^3}. \quad (21)$$

W podobny sposób można obliczyć obciążenie innych współczynników zależności dla tablic czteropolowych. Podajmy efekty obliczeń dla mierników: W, R, R_L, D^*, R_c , zdefiniowanych następująco ([2; 3; 4]):

$$W = \frac{(\rho_{11} + \rho_{22}) - (\rho_{12} + \rho_{21})}{n}, \quad (22)$$

$$R = \frac{\rho_{11} \rho_{22}}{\rho_{12} \rho_{21}}, \quad (23)$$

$$L_R = \ln \frac{\rho_{11}\rho_{22}}{\rho_{12}\rho_{21}}, \quad (24)$$

$$D^* = \frac{\rho_{11}}{\rho_{11} + \rho_{21}} - \frac{\rho_{12}}{\rho_{12} + \rho_{22}}, \quad (25)$$

$$R_c = \frac{\rho_{11}(\rho_{12} + \rho_{22})}{\rho_{12}(\rho_{11} + \rho_{21})}. \quad (26)$$

Dlatego

$$E\hat{W} - W \approx 0, \quad (27)$$

$$E\hat{R} - R \approx \frac{1}{np_b p_c} \left(\frac{p_a p_d}{p_b} + \frac{p_a p_d}{p_c} \right), \quad (28)$$

$$E\hat{L}_R - L_R \approx \frac{1}{2n} \left(-\frac{1}{p_a} + \frac{1}{p_b} + \frac{1}{p_c} - \frac{1}{p_d} \right), \quad (29)$$

$$E\hat{D}^* - D^* \approx \frac{1}{2n} \left\{ \frac{3p_a p_c}{(p_a - p_c)^3} (-p_a + p_c) + \frac{p_b p_d}{(p_b + p_d)^3} (4 - p_b - 3p_d) \right\}, \quad (30)$$

$$E\hat{R}_c - R_c \approx \frac{p_a}{2np_b^2(p_a + p_c)^3} (p_a p_b p_c - p_b^2 p_c^2 - p_a p_b p_c p_d - 2p_b p_c^2 p_d + \\ + 2p_a p_c p_d + 2p_a^2 p_b p_d). \quad (31)$$

Występująca w mianownikach wzorów na obciążenie wartość n wskazuje, że oceny obciążeń zaprezentowanych współczynników dążą do zera, gdy n rośnie nieograniczenie.

Do ilustracji dokładności oszacowań tych obciążeń wybraliśmy miernik Q -Yule'a. W tabeli 4 dla trzech układów prawdopodobieństw $\rho_a, \rho_b, \rho_c, \rho_d$ oraz $n = 10, 15, 20, 30, 40, 60, 80, 100$ zestawiono wartości Q , $E\hat{Q}$, obciążenie $B\hat{Q} = E\hat{Q} - Q$, ocenę tego obciążenia $\widehat{B\hat{Q}}$ oraz jej stosunek do faktycznej wielkości obciążenia

$$q = \frac{\widehat{B\hat{Q}}}{B\hat{Q}} \cdot 100\%. \quad (32)$$

Tabela 4. Dokładność oszacowań obciążenia estymatora współczynnika Q -Yule'a

Prawdopodobieństwa (q_a, q_b, q_c, q_d)	n	Q	$E\hat{Q}$	$B\hat{Q} = E\hat{Q} - Q$	$\widehat{B\hat{Q}}$	$q = \frac{\widehat{B\hat{Q}}}{B\hat{Q}}$ (w %)
(0,4; 0,1; 0,1; 0,4)	10	0,88235	0,83648	-0,04587	-0,03908	85,2
	15	0,88235	0,85454	-0,02781	-0,02605	93,7
	20	0,88235	0,86209	-0,02026	-0,01954	96,4
	30	0,88235	0,86904	-0,01331	-0,01303	97,9
	40	0,88235	0,87242	-0,00993	-0,00977	98,4
	60	0,88235	0,87577	-0,00658	-0,00651	98,9
	80	0,88235	0,87743	-0,00492	-0,00489	99,4
	100	-	-	-	-	-
(0,70; 0,05; 0,20; 0,05)	10	0,55556	0,17734	-0,37822	-0,37811	99,9
	15	0,55556	0,20153	-0,35403	-0,25606	72,3
	20	0,55556	0,24653	-0,30903	-0,19204	62,1
	30	0,55556	0,33061	-0,22495	-0,12803	56,9
	40	0,55556	0,39144	-0,16412	-0,09602	58,5
	60	0,55556	0,46020	-0,09536	-0,06401	67,1
	80	0,55556	0,49192	-0,06364	-0,04801	75,4
	100	0,55556	0,50829	-0,04727	-0,03841	81,3
(0,85; 0,05; 0,05; 0,05)	10	0,88889	0,57145	-0,31744	-0,18656	58,7
	15	0,88889	0,53885	-0,35004	-0,12437	35,5
	20	0,88889	0,54885	-0,34004	-0,09328	27,4
	30	0,88889	0,61412	-0,27477	-0,06219	22,6
	40	0,88889	0,68663	-0,20226	-0,04664	23,1
	60	0,88889	0,78688	-0,10201	-0,03109	30,5
	80	0,88889	0,83401	-0,05488	-0,02332	42,5
	100	-	-	-	-	-

Źródło: obliczenia własne.

Na podstawie wyników obliczeń zawartych w tab. 4 można sformułować następujące wnioski:

- dla rozkładów symetrycznych, np. (q_a, q_b, q_c, q_d) = (0,4; 0,1; 0,1; 0,4), obciążenie estymatora współczynnika Q -Yule'a jest bardzo małe nawet dla małych prób (np.: $n = 15$, $q = 93,7\%$),
- dla asymetrycznych prawdopodobieństw brzegowych obciążenie estymatora $\hat{Q}$ nawet dla dużych prób jest wysokie,
- wartość obciążenia maleje w miarę wzrostu liczebności próby.

4. Uwagi końcowe

W artykule zaprezentowano jako przykładową metodę oceny obciążenia współczynnika zależności Q -Yule'a. W sposób podobny można to uczynić dla innych współczynników zależności w czteropolowych tablicach kontyngencyjnych.

Problem badania współzależności w czteropolowych tablicach kontyngencyjnych jest złożony, a prace związane m.in. z oceną obciążenia są w dalszym ciągu oczekiwane np. przez przedstawicieli nauk ekonomicznych. Autorzy tego opracowania zamierzają kontynuować badania z tym związane.

Literatura

- [1] Chambers R.L., Sted D.G. (2001), *Simple Methods for Ecological Inference in 2x2 Tables*, „Journal of the Royal Statistical Society”, A 164, part 1, s. 175-192.
- [2] Domański Cz. (red.) (2001), *Metody statystyczne. Teoria i zadania*, Wyd. UŁ, Łódź.
- [3] Goodman L.A., Kruskal W.M. (1972), *Measures of Association for Cross Classification, IV: Simplification of Asymptotic Variances*, „Journal of the American Statistical Association”, 67, s. 415-421.
- [4] Mosteller F. (1968), *Association and Estimation in Contingency Tables*, „Journal of the American Statistical Association”, 63, s. 1-28.
- [5] Stępień L. (1974), *Properties of Interdependence Measures in 2x2 Contingency Tables*, „Statistics in Transition”, vol. 1, no. 5, s. 703-714.

ESTIMATION OF THE BIAS OF CORRELATION COEFFICIENTS FOR CONTINGENCY TABLES

Summary

In the article the bias of six chosen correlation measures for 2x2 contingency tables is assessed with the use of the Taylor expansion. To illustrate obtained results, the estimated bias of the Q -Yule measure was presented in relation to the sample size and to the distribution of the probabilities of variables investigated. The results of calculations indicate that the bias of the estimator of the Q -Yule measure is very small for symmetric distributions whereas for asymmetric ones it is large. Values of the bias decrease together with the increase of the sample size.