

Miary związku między cechami w tablicy trójdzielczej

Metody wnioskowania dotyczące jednej cechy są bardzo popularne w literaturze statystycznej. Jednak w badaniach statystycznych bardzo często obiekty opisuje się za pomocą więcej niż jednej cechy. Bardzo popularne narzędzie do badania związku między cechami stanowią tablice wielodzielcze. Jeżeli dane obejmują dwie cechy, to mówimy o tablicach dwudzielczych, trzy cechy — tablicach trójdzielczych, cztery cechy — tablicach czterodzielczych, ..., m cech — tablicach m -dzielczych (Okta, 1974). Tablicami wielodzielczymi są także kostki OLAP stosowane np. w Narodowym Spisie Ludności i Mieszkań czy w Powszechnym Spisie Rolnym.

W publikacji Sulewskiego (2007) opisano procedurę generowania tablic dwudzielczych o zadanym związku między cechami. Zbadano także moc testu, czyli zdolność tablicy dwudzielczej do wykrywania związku między badanymi cechami oraz obliczono jego siłę wykorzystując współczynnik V Cramera. Porównano rzeczywisty rozkład statystyki testowej dla tablic dwudzielczych z rozkładem teoretycznym, jakim jest rozkład chi-kwadrat o $(w-1)(k-1)$ stopniach swobody. Dzięki tej publikacji czytelnik może samodzielnie modelować tablice o rozmiarach, z którymi ma do czynienia w praktyce i może badać ich moc.

W innym opracowaniu (Sulewski, 2013) przedstawiono test niezależności chi-kwadrat dla tablicy trójdzielczej wraz z jego implementacją komputerową w języku VBA. Zbadano też zdolność tablicy trójdzielczej do wykrywania związku między badanymi cechami.

Wśród użytkowników komputerów bardzo popularne są arkusze kalkulacyjne, a w szczególności Microsoft Excel. Do zalet tej aplikacji zaliczamy m.in.: łatwe gromadzenie danych i na ich podstawie tworzenie tabel oraz wykresów, stosowanie różnorodnych obliczeń z zastosowaniem funkcji matematycznych i statystycznych, a także możliwość zwiększania operatywności programu za pomocą kodów VBA (*Visual Basic for Applications*).

W literaturze statystycznej bardzo popularne są narzędzia do pomiaru siły związku między cechami w tablicach dwudzielczych. Należą do nich m.in. współczynniki: V Cramera, T Czuprowa, kontyngencji C Pearsona, ϕ Yule'a, Goodmana-Kruskala (Goodman, Kruskal, 1954), κ Cohena (Cohen, 1960), Spearmana, η (*SPSS*), współczynnik punktowo-dwuseryjny (Sobczyk, 1996) i niepewności (*SPSS*). Jednak mechanizmy do pomiaru siły związku między cechami w tablicach trójdzielczych nie są już tak bardzo popularne, a w literaturze polskojęzycznej praktycznie nie występują.

W artykule zdefiniowano tablicę trójdzielczą wraz z jej charakterystykami oraz przedstawiono implementację komputerową dotyczącą pomiaru siły związku między trzema cechami za pomocą współczynników: Gray-Williamsa, Marcotorchino i Lombardo.

Pierwszym celem pracy jest przedstawienie teorii dotyczącej tablic trójdzielczych. Jest to niezbędne do lepszego zrozumienia zagadnień dotyczących pomiaru siły związku między trzema cechami. Drugim celem jest przedstawienie gotowych procedur i funkcji zapisanych w języku VBA, umożliwiających czytelnikowi samodzielne przeprowadzanie badań statystycznych w najbardziej popularnym arkuszu kalkulacyjnym, jakim jest Microsoft Excel.

TABLICA TRÓJDZIELCZA I JEJ CHARAKTERYSTYKI

Tablicę, która gromadzi wyniki podziału próby według trzech cech X , Y , Z nazywamy tablicą trójdzielczą. Można sobie ją wyobrazić jako kostkę z „ w ” wierszami, „ k ” kolumnami i „ p ” płaszczyznami. Tablicę trójdzielczą można także zilustrować rozkładając płaszczyzny „ p ” obok siebie (tabl. 1 i 2).

Gdy liczba wszystkich przyjmowanych przez cechy X , Y , Z wartości jest względnie mała, wpisuje się je w odpowiednie wiersze, kolumny i płaszczyzny. Z kolei gdy liczba możliwych wartości jest duża niezbędne okazuje się ich pogrupowanie w klasy.

Tablica trójdzielcza ułatwia odczytywanie zależności między cechami tak ilościowymi, jak i jakościowymi. Najprostszą jej postacią jest tablica $2 \times 2 \times 2$ (tabl. 1), która składa się z 8 liczebności n_{ijt} ($i, j, t = 1, 2$) rozkładu łącznego cech X , Y , Z .

TABL. 1. TABLICA TRÓJDZIELCZA $2 \times 2 \times 2$

Wyszczególnienie	Z_1		Z_2		Razem
	Y_1	Y_2	Y_1	Y_2	
X_1	n_{111}	n_{121}	n_{112}	n_{122}	$n_{1\bullet\bullet}$
X_2	n_{211}	n_{221}	n_{212}	n_{222}	$n_{2\bullet\bullet}$
Razem	$n_{\bullet 11}$	$n_{\bullet 21}$	$n_{\bullet 12}$	$n_{\bullet 22}$	n

Źródło: opracowanie własne.

Sumy brzegowe w tablicy trójdzielczej $2 \times 2 \times 2$ to:

$$n_{1\bullet\bullet} = n_{111} + n_{112}, \quad n_{12\bullet} = n_{121} + n_{122}, \quad n_{21\bullet} = n_{211} + n_{212}, \quad n_{22\bullet} = n_{221} + n_{222} \quad (1)$$

$$n_{1\bullet 1} = n_{111} + n_{211}, \quad n_{1\bullet 2} = n_{112} + n_{212}, \quad n_{2\bullet 1} = n_{121} + n_{221}, \quad n_{2\bullet 2} = n_{122} + n_{222} \quad (2)$$

$$n_{\bullet 11} = n_{111} + n_{211}, \quad n_{\bullet 12} = n_{112} + n_{212}, \quad n_{\bullet 21} = n_{121} + n_{221}, \quad n_{\bullet 22} = n_{122} + n_{222} \quad (3)$$

$$n_{1\bullet\bullet} = n_{111} + n_{112} + n_{121} + n_{122} = n_{11\bullet} + n_{12\bullet} \quad (4)$$

$$n_{2\bullet\bullet} = n_{211} + n_{212} + n_{221} + n_{222} = n_{21\bullet} + n_{22\bullet} \quad (5)$$

$$n_{\bullet 1\bullet} = n_{111} + n_{112} + n_{211} + n_{212} = n_{11\bullet} + n_{21\bullet} \quad (6)$$

$$n_{\bullet 2\bullet} = n_{121} + n_{122} + n_{221} + n_{222} = n_{12\bullet} + n_{22\bullet} \quad (7)$$

$$n_{\bullet\bullet 1} = n_{111} + n_{121} + n_{211} + n_{221} = n_{1\bullet 1} + n_{2\bullet 1} \quad (8)$$

$$n_{\bullet\bullet 2} = n_{112} + n_{122} + n_{212} + n_{222} = n_{1\bullet 2} + n_{2\bullet 2} \quad (9)$$

Tablicę trójdzielczą $w \times k \times p$, która składa się z $w \cdot k \cdot p$ liczebności n_{ijt} ($i = 1, \dots, w$; $j = 1, \dots, k$; $t = 1, \dots, p$) rozkładu łącznego cech X, Y, Z zaprezentowano w tabl. 2.

TABL. 2. TABLICA TRÓJDZIELCZA $w \times k \times p$

Wyszczególnienie	Z_1				...	Z_p				Razem
	Y_1	Y_2	...	Y_k	...	Y_1	Y_2	...	Y_k	
X_1	n_{111}	n_{121}	...	n_{1k1}	...	n_{11p}	n_{12p}	...	n_{1kp}	$n_{1\bullet\bullet}$
X_2	n_{211}	n_{221}	...	n_{2k1}	...	n_{21p}	n_{22p}	...	n_{2kp}	$n_{2\bullet\bullet}$
...	...	...	...	...	...	...	...	...	...	...
X_w	n_{w11}	n_{w21}	...	n_{wk1}	...	n_{w1p}	n_{w2p}	...	n_{wkp}	$n_{w\bullet\bullet}$
Razem	$n_{\bullet 11}$	$n_{\bullet 21}$	...	$n_{\bullet k1}$	...	$n_{\bullet 1p}$	$n_{\bullet 2p}$	...	$n_{\bullet kp}$	n

Ź r ó d ł o: jak przy tabl. 1.

Sumy brzegowe w tablicy trójdzielczej $w \times k \times p$ to:

$$n_{11\bullet} = \sum_{t=1}^p n_{11t}, \quad n_{12\bullet} = \sum_{t=1}^p n_{12t}, \quad \dots, \quad n_{wk\bullet} = \sum_{t=1}^p n_{wkt} \quad (10)$$

$$n_{1\bullet 1} = \sum_{j=1}^k n_{1j1}, \quad n_{1\bullet 2} = \sum_{j=1}^k n_{1j2}, \quad \dots, \quad n_{w\bullet p} = \sum_{j=1}^k n_{wjp} \quad (11)$$

$$n_{\bullet 11} = \sum_{i=1}^w n_{w11}, \quad n_{\bullet 12} = \sum_{i=1}^w n_{w12}, \quad \dots, \quad n_{\bullet kp} = \sum_{i=1}^w n_{wkp} \quad (12)$$

$$n_{1\bullet\bullet} = \sum_{j=1}^k \sum_{t=1}^p n_{1jt}, n_{2\bullet\bullet} = \sum_{j=1}^k \sum_{t=1}^p n_{2jt}, \dots, n_{w\bullet\bullet} = \sum_{j=1}^k \sum_{t=1}^p n_{wjt} \quad (13)$$

$$n_{\bullet 1\bullet} = \sum_{i=1}^w \sum_{t=1}^p n_{i1t}, n_{\bullet 2\bullet} = \sum_{i=1}^w \sum_{t=1}^p n_{i2t}, \dots, n_{\bullet k\bullet} = \sum_{i=1}^w \sum_{t=1}^p n_{ikt} \quad (14)$$

$$n_{\bullet\bullet 1} = \sum_{i=1}^w \sum_{j=1}^k n_{ij1}, n_{\bullet\bullet 2} = \sum_{i=1}^w \sum_{j=1}^k n_{ij2}, \dots, n_{\bullet\bullet p} = \sum_{i=1}^w \sum_{j=1}^k n_{ijp} \quad (15)$$

Wartość n jest sumą wszystkich liczebności tablicy trójdzielczej:

$$\begin{aligned} n &= \sum_{i=1}^w n_{i\bullet\bullet} = \sum_{j=1}^k n_{\bullet j\bullet} = \sum_{t=1}^p n_{\bullet\bullet t} = \sum_{i=1}^w \sum_{j=1}^k n_{ij\bullet} = \\ &= \sum_{i=1}^w \sum_{t=1}^p n_{i\bullet t} = \sum_{j=1}^k \sum_{t=1}^p n_{\bullet jt} = \sum_{i=1}^w \sum_{j=1}^k \sum_{t=1}^p n_{ijt} \end{aligned} \quad (16)$$

Jeżeli cechy X, Y, Z są cechami ciągłymi, to wartości X_i, Y_j, Z_t w tablicy trójdzielczej są środkami przedziałów.

Wprowadzenie pojęcia liczebności oczekiwanej oraz liczebności resztowych ułatwi formalizację zapisu tablicy trójdzielczej.

Liczebności oczekiwane i resztowe

Niech p_{ijt} ($i = 1, 2, \dots, w$; $j = 1, 2, \dots, k$; $t = 1, 2, \dots, p$) będzie prawdopodobieństwem łącznym wystąpienia zdarzenia polegającego na tym, że dany element próby będzie należał do i -tej kategorii według cechy X , do j -tej kategorii według cechy Y oraz do t -tej kategorii według cechy Z . Niezależność cech X, Y, Z oznacza, że dla wszystkich trójek wskaźników i, j, t zachodzi równość:

$$p_{ijt} = p_{i\bullet\bullet} \cdot p_{\bullet j\bullet} \cdot p_{\bullet\bullet t} \quad (17)$$

gdzie:

$p_{i\bullet\bullet}$ — prawdopodobieństwo, że obserwacja należy do i -tej kategorii według cechy X ;

$p_{\bullet j\bullet}$ — prawdopodobieństwo, że obserwacja należy do j -tej kategorii według cechy Y ;

$p_{\bullet\bullet t}$ — prawdopodobieństwo, że obserwacja należy do t -tej kategorii według cechy Z .

Nieznane prawdopodobieństwa p_{ijt} , $p_{i\bullet\bullet}$, $p_{\bullet j\bullet}$, $p_{\bullet\bullet t}$ oblicza się z następujących relacji:

$$n_{ijt} = n \cdot p_{ijt}, \quad n_{i\bullet\bullet} = n \cdot p_{i\bullet\bullet}, \quad n_{\bullet j\bullet} = n \cdot p_{\bullet j\bullet}, \quad n_{\bullet\bullet t} = n \cdot p_{\bullet\bullet t} \quad (18)$$

gdzie:

- n — liczebność próby,
- n_{ijt} — liczebności komórek,
- $n_{i\bullet\bullet}$ — sumy brzegowe w wierszach,
- $n_{\bullet j\bullet}$ — sumy brzegowe w kolumnach.

Liczebności oczekiwane wyznacza się ze wzoru:

$$e_{ijt} = n \cdot p_{i\bullet\bullet} \cdot p_{\bullet j\bullet} \cdot p_{\bullet\bullet t} \quad (19)$$

Stąd ostatecznie na mocy (18):

$$e_{ijt} = \frac{n_{i\bullet\bullet} \cdot n_{\bullet j\bullet} \cdot n_{\bullet\bullet t}}{n^2} \quad (i=1, 2, \dots, w, j=1, 2, \dots, k; t=1, 2, \dots, p) \quad (20)$$

Sumy brzegowe liczebności oczekiwanych e_{ijt} w tablicy $w \times k \times p$ są takie same, jak sumy brzegowe liczebności n_{ijt} . W szczególności sumy brzegowe dla wierszy w tablicy $2 \times 2 \times 2$ wynoszą:

$$\begin{aligned} e_{111} + e_{112} + e_{121} + e_{122} &= \frac{n_{1\bullet\bullet} \cdot n_{\bullet 1\bullet} \cdot n_{\bullet\bullet 1}}{n^2} + \frac{n_{1\bullet\bullet} \cdot n_{\bullet 1\bullet} \cdot n_{\bullet\bullet 2}}{n^2} + \\ &+ \frac{n_{1\bullet\bullet} \cdot n_{\bullet 2\bullet} \cdot n_{\bullet\bullet 1}}{n^2} + \frac{n_{1\bullet\bullet} \cdot n_{\bullet 2\bullet} \cdot n_{\bullet\bullet 2}}{n^2} = \\ &= \frac{n_{1\bullet\bullet} (n_{\bullet 1\bullet} \cdot n_{\bullet\bullet 1} + n_{\bullet 1\bullet} \cdot n_{\bullet\bullet 2} + n_{\bullet 2\bullet} \cdot n_{\bullet\bullet 1} + n_{\bullet 2\bullet} \cdot n_{\bullet\bullet 2})}{n^2} = \\ &= \frac{n_{1\bullet\bullet} [n_{\bullet 1\bullet} (n_{\bullet\bullet 1} + n_{\bullet\bullet 2}) + n_{\bullet 2\bullet} (n_{\bullet\bullet 1} + n_{\bullet\bullet 2})]}{n^2} = \\ &= \frac{n_{1\bullet\bullet} [n_{\bullet 1\bullet} \cdot n + n_{\bullet 2\bullet} \cdot n]}{n^2} = \frac{n \cdot n_{1\bullet\bullet} [n_{\bullet 1\bullet} + n_{\bullet 2\bullet}]}{n^2} = n_{1\bullet\bullet} \end{aligned} \quad (21)$$

$$\begin{aligned} e_{211} + e_{212} + e_{221} + e_{222} &= \frac{n_{2\bullet\bullet} \cdot n_{\bullet 1\bullet} \cdot n_{\bullet\bullet 1}}{n^2} + \frac{n_{2\bullet\bullet} \cdot n_{\bullet 1\bullet} \cdot n_{\bullet\bullet 2}}{n^2} + \\ &+ \frac{n_{2\bullet\bullet} \cdot n_{\bullet 2\bullet} \cdot n_{\bullet\bullet 1}}{n^2} + \frac{n_{2\bullet\bullet} \cdot n_{\bullet 2\bullet} \cdot n_{\bullet\bullet 2}}{n^2} = \dots = \\ &= \frac{n_{2\bullet\bullet} [n_{\bullet 1\bullet} \cdot n + n_{\bullet 2\bullet} \cdot n]}{n^2} = \frac{n \cdot n_{2\bullet\bullet} [n_{\bullet 1\bullet} + n_{\bullet 2\bullet}]}{n^2} = n_{2\bullet\bullet} \end{aligned} \quad (22)$$

Liczebności resztowe wyznacza się ze wzoru:

$$R_{ijt} = n_{ijt} - e_{ijt} \quad (i=1, 2, \dots, w; j=1, 2, \dots, k; t=1, 2, \dots, p) \quad (23)$$

Liczebności resztowe standaryzowane wyznacza się ze wzoru:

$$RN_{ijt} = \frac{n_{ijt} - e_{ijt}}{\sqrt{e_{ijt}}} \quad (i=1, 2, \dots, w; j=1, 2, \dots, k; t=1, 2, \dots, p) \quad (24)$$

Liczebności resztowe skorygowane wyznacza się ze wzoru:

$$RS_{ijt} = \frac{n_{ijt} - e_{ijt}}{\sqrt{e_{ijt} \left(1 - \frac{n_{i\bullet\bullet}}{n}\right) \left(1 - \frac{n_{\bullet j\bullet}}{n}\right) \left(1 - \frac{n_{\bullet\bullet t}}{n}\right)}} \quad (i=1, 2, \dots, w; j=1, 2, \dots, k; t=1, 2, \dots, p) \quad (25)$$

WSPÓŁCZYNNIKI KORELACJI TRZECH CECH

Współczynnik τ Goodmana-Kruskala jest popularnym współczynnikiem korelacji cech porządkowych dla tablic dwudzielczych. Numerycznymi rozszerzeniami współczynnika τ Goodmana-Kruskala dla tablic trójdzielczych są: współczynnik Gray-Williamsa, współczynnik Marcotorchino oraz współczynnik Lombardo.

Współczynnik Gray-Williamsa stosowany jest w sytuacji, gdy dwie cechy (np. Y, Z) są cechami objaśniającymi lub predyktorami wzajemnie zależnymi (*dependent predictor variables*), a trzecia cecha (np. X) jest cechą objaśnianą lub odpowiedzi (*response variable*) (wykr. 1A).

Współczynnik Marcotorchino stosowany jest w sytuacji, gdy dwie cechy (np. Y, Z) są cechami objaśniającymi lub predyktorami wzajemnie niezależnymi (*independent predictor variables*), a trzecia cecha (np. X) jest cechą objaśnianą lub odpowiedzi (*response variable*) (wykr. 1B).

Współczynnik Lombardo stosowany jest w sytuacji, gdy dwie cechy (np. X, Y) są cechami objaśnianymi lub odpowiedzi (*independent response variables*), a trzecia cecha (np. Z) jest cechą objaśniającą lub predyktorem (*predictor variable*) (wykr. 1C).

Współczynnik Gray-Williamsa

Jeżeli Y, Z są cechami objaśniającymi wzajemnie zależnymi, a X jest cechą objaśnianą, to niesymetryczny współczynnik Gray-Williamsa (Gray, Williams, 1975) ma postać:

$$\tau_{GW} = \frac{\sum_{i=1}^w \sum_{j=1}^k \sum_{t=1}^p p_{\bullet jt} \left(\frac{p_{ijt}}{p_{\bullet jt}} - p_{i\bullet\bullet} \right)}{1 - \sum_{i=1}^w p_{i\bullet\bullet}^2} \quad (26)$$

gdzie $p_x = n_x/n$.

Współczynnik ten przyjmuje wartości z przedziału $\langle 0, 1 \rangle$. Jeżeli cechy Y i Z nie dostarczają informacji o cesze X , wtedy $\tau_{GW} = 0$. Wraz ze wzrostem informacji dostarczanej przez cechy Y i Z o cesze X wartość tego współczynnika zbliża się do jedności.

Na podstawie próby n -elementowej pobranej z populacji generalnej zweryfikowano hipotezę zerową, że cechy Y i Z nie wpływają istotnie na cechę X . Jeżeli hipoteza zerowa jest prawdziwa, to statystyka (Light, Margolin, 1971)

$$C_{GW} = (n-1)(w-1)\tau_{GW} \quad (27)$$

ma asymptotyczny rozkład chi-kwadrat z Sw_{GW} stopniami swobody, tj.:

$$Sw_{GW} = (w-1)(k-1) + (w-1)(p-1) + (w-1)(k-1)(p-1) \quad (28)$$

Wartość p -value wyraża się wzorem:

$$pv_{GW} = 1 - F(C_{GW}, Sw_{GW}) \quad (29)$$

gdzie $F(\cdot)$ — dystrybuenta rozkładu chi-kwadrat dana wzorem:

$$F(x, k) = \frac{\Gamma_n(k/2, x/2)}{\Gamma(k/2)} \quad (30)$$

$\Gamma_n(\cdot)$ jest niepełną funkcją gamma daną wzorem:

$$\Gamma_n(c, x) = \int_0^x u^{c-1} \exp(-u) du \quad (31)$$

natomiast $\Gamma(\cdot)$ jest funkcją gamma daną wzorem:

$$\Gamma(c) = \int_0^\infty u^{c-1} \exp(-u) du \quad (32)$$

Jeżeli na poziomie istotności α mamy $pv_{GW} \leq \alpha$, to hipoteza H_0 zostaje odrzucona, gdy natomiast $pv_{GW} > \alpha$, to nie ma podstaw do odrzucenia H_0 .

Współczynnik Marcotorchino

Przyjmując, że Y, Z są cechami objaśniającymi wzajemnie niezależnymi, a X jest cechą objaśnianą, to niesymetryczny współczynnik Marcotorchino ma postać (Marcotorchino, 1984):

$$\tau_M = \frac{\sum_{i=1}^w \sum_{j=1}^k \sum_{t=1}^p p_{\bullet j \bullet} \cdot p_{\bullet \bullet t} \left(\frac{p_{ijt}}{p_{\bullet j \bullet} \cdot p_{\bullet \bullet t}} - p_{i \bullet \bullet} \right)^2}{1 - \sum_{i=1}^w p_{i \bullet \bullet}^2} \quad (33)$$

gdzie $p_x = n_x/n$. Zakładając, że $p_{ijt} = p_{i \bullet \bullet} p_{\bullet j \bullet} p_{\bullet \bullet t}$, to $\tau_M = 0$, natomiast $p_{ijt} = p_{i \bullet \bullet} p_{\bullet \bullet t}$, wtedy $\tau_M = 1$.

Na podstawie próby n -elementowej pobranej z populacji generalnej zweryfikowano hipotezę zerową, mówiącą że cechy Y i Z nie wpływają istotnie na cechę X . Przyjmując, że hipoteza zerowa jest prawdziwa, to statystyka (Beh i in., 2007)

$$C_M = (n-1)(w-1)\tau_M \quad (34)$$

ma asymptotyczny rozkład chi-kwadrat z Sw_M stopniami swobody, gdzie:

$$Sw_M = (w-1)(k-1) + (w-1)(p-1) + (k-1)(p-1) + (w-1)(k-1)(p-1) \quad (35)$$

Wartość p -value wyraża się wzorem:

$$pv_M = 1 - F(C_M, Sw_M) \quad (36)$$

gdzie $F(.)$ — dystrybuenta rozkładu chi-kwadrat dana wzorem (30). Jeżeli na poziomie istotności α mamy $pv_M \leq \alpha$, to hipoteza H_0 zostaje odrzucona, ale gdy $pv_M > \alpha$, wtedy nie ma podstaw do odrzucenia H_0 .

Współczynnik Lombardo

Zakładając, że X, Y są cechami objaśnianymi, a Z jest cechą objaśniającą, to niesymetryczny współczynnik Lombardo (współczynnik delta) ma postać (Lombardo, 2011):

$$\tau_L = \frac{\sum_{i=1}^w \sum_{j=1}^k \sum_{t=1}^p p_{\bullet\bullet t} \left(\frac{p_{ijt}}{p_{\bullet\bullet t}} - p_{i\bullet\bullet} \cdot p_{\bullet j\bullet} \right)^2}{1 - \sum_{i=1}^w \sum_{j=1}^k p_{i\bullet\bullet}^2 \cdot p_{\bullet j\bullet}^2} \quad (37)$$

gdzie $p_x = n_x/n$. Jeżeli $p_{ijt} = p_{i\bullet\bullet} p_{\bullet j\bullet} p_{\bullet\bullet t}$, to $\tau_L = 0$. Na przykład dla tablicy $2 \times 2 \times 2$ wartość maksymalna $\tau_L = 1$, gdy $p_{1jt} = 0,5$ dla pewnych $(j, t = 1, 2)$ i $p_{2bc} = 0,5$ dla pewnych $(b, c = 1, 2; b \neq j \wedge c \neq t)$.

Na podstawie próby n -elementowej pobranej z populacji generalnej zweryfikowano hipotezę zerową, że cecha Z nie wpływa istotnie na cechy X i Y . Jeżeli hipoteza zerowa jest prawdziwa, to statystyka (Lombardo, 2011)

$$C_L = (n-1)(w-1) \tau_L \quad (38)$$

ma asymptotyczny rozkład chi-kwadrat z Sw_L stopniami swobody, gdzie:

$$Sw_L = (w-1)(k-1) + (w-1)(p-1) + (k-1)(p-1) + (w-1)(k-1)(p-1) \quad (39)$$

Wartość p -value wyraża się wzorem:

$$pv_L = 1 - FC(C_L, Sw_L) \quad (40)$$

gdzie $F(\cdot)$ jest dystrybuantą rozkładu chi-kwadrat daną wzorem (30). Jeżeli na poziomie istotności α mamy $pv_L \leq \alpha$, to hipoteza H_0 zostaje odrzucona, jeżeli $pv_L > \alpha$, to nie ma podstaw do odrzucenia H_0 .

IMPLEMENTACJA KOMPUTEROWA

Implementację komputerową pomiaru siły związku między cechami w tablicy trójdzielczej utworzoną w edytorze VBA arkusza kalkulacyjnego Excel przedstawiono w pliku „mierniki3d.xls”, który umieszczono w Internecie pod adresem <http://www.utogim.eu/mierniki3d.xls>. Składają się na nią procedury i funkcje użytkownika umieszczone w zestawieniu (1). Aby ułatwić czytelnikowi zrozumienie użytych kodów wzbogacono je o liczne komentarze umieszczone po apostrofach i zapisane kursywą, a ponadto w zestawieniu (2) zawarto opis zmiennych wykorzystanych przy tworzeniu implementacji komputerowej.

ZESTAWIENIE (1) PROCEDUR I FUNKCJI TWORZĄCYCH IMPLEMENTACJĘ KOMPUTEROWĄ

Wyszczególnienie	Realizowane zadanie
Procedury	
WspKor3D	wyznaczanie współczynników korelacji cech X, Y, Z
Gray-Williams	obliczanie wartości współczynnika Gray-Williamsa
Marcotorchino	obliczanie wartości współczynnika Marcotorchino
Lombardo	obliczanie wartości współczynnika Lombardo
Funkcje	
Rozmiar3	ustalenie liczby kategorii cech X, Y, Z
GPijt	wyznaczanie wartości prawdopodobieństw p_{ijt}
SX3	obliczanie prawdopodobieństw brzegowych $p_{i\cdot}$
SY3	obliczanie prawdopodobieństw brzegowych $p_{\cdot j}$
SZ3	obliczanie prawdopodobieństw brzegowych $p_{\cdot t}$
SYZ3	obliczanie prawdopodobieństw brzegowych $p_{\cdot jt}$
TestChiKw	realizacja testu chi-kwadrat

Źródło: jak przy tabl. 1.

ZESTAWIENIE (2) ZMIENNYCH WYKORZYSTYWANYCH W IMPLEMENTACJI KOMPUTEROWEJ

Zmienne	Opis zmiennej
w, k, p	liczba kategorii w tablicy trójdzielczej
p_{ikk}	sumy brzegowe prawdopodobieństw p_i
p_{kjk}	sumy brzegowe prawdopodobieństw p
p_{kkt}	sumy brzegowe prawdopodobieństw p_t
p_{ijt}	wartości prawdopodobieństw p_{ijt}

**ZESTAWIENIE (2) ZMIENNYCH WYKORZYSTYWANYCH
W IMPLEMENTACJI KOMPUTEROWEJ (dok.)**

Zmienne	Opis zmiennej
<i>TauGw</i>	wartość współczynnika Gray-Williamsa
<i>TauM</i>	wartość współczynnika Marcotorchino
<i>TauL</i>	wartość współczynnika Lombardo
<i>SwGw</i>	liczba stopni swobody dla współczynnika Gray-Williamsa
<i>SwM</i>	liczba stopni swobody dla współczynnika Marcotorchino
<i>SwL</i>	liczba stopni swobody dla współczynnika Lombardo
<i>i, j, t, s</i>	zmienne pomocnicze
<i>Wynik</i>	okno dialogowe z wynikami
<i>n</i>	liczebność próby
<i>alfa</i>	poziom istotności
<i>Ho</i>	hipoteza zerowa
<i>CGw</i>	statystyka testowa dla współczynnika Gray-Williamsa
<i>CM</i>	statystyka testowa dla współczynnika Marcotorchino
<i>CL</i>	statystyka testowa dla współczynnika Lombardo
<i>PvGw</i>	wartość <i>p-value</i> dla współczynnika Gray-Williamsa
<i>PvM</i>	wartość <i>p-value</i> dla współczynnika Marcotorchino
<i>PvL</i>	wartość <i>p-value</i> dla współczynnika Lombardo

Ź r ó d ł o: jak przy tabl. 1.

Przykład 1

Zbadano wpływ miejsca pochodzenia (*Z*) i opieki (*Y*) (zmienne objaśniające zależne) na diagnozę dotyczącą choroby wątroby (*X*) (zmienna objaśniana). Otrzymane wyniki zestawiono w formie tablicy trójdzielczej $4 \times 2 \times 3$ (tabl. 3).

TABL. 3. TABLICA TRÓJDZIELCZA $4 \times 2 \times 3$

Diagnozy chorób wątroby	Miejsce pochodzenia diagnozy						razem
	północ		środek		południe		
	opieka medyczna						
	tak	nie	tak	nie	tak	nie	
Bezobjawowa	1	10	0	10	0	11	32
Zapalenie	43	64	31	29	89	66	322
Marskość	4	25	6	7	13	39	94
Zakażenie HCC	1	5	0	1	0	4	11
Razem	49	104	37	47	102	120	459

Ź r ó d ł o: Lombardo (2011).

Po utworzeniu tablicy trójdzielczej (tabl. 3) w arkuszu „tablica” (począwszy od komórki B2) uruchomiono procedurę „WspKor3D”. W arkuszu „wyniki” za pomocą pola kombi wybrano opcję „Gray-Williamsa” i wyznaczono wartość współczynnika Gray-Williamsa, gdy *Y*, *Z* są cechami objaśniającymi wzajemnie zależnymi, a *X* jest cechą objaśnianą. Następnie założono hipotezę zerową, że pocho-

dzenie (Z) i opieka (Y) nie wpływają istotnie na diagnozę dotyczącą choroby wątroby (X). Wprowadzono poziom istotności testu α . Skoro $pv_{GW} = 0,000 \leq \alpha$, zatem są podstawy do odrzucenia hipotezy zerowej, czyli pochodzenie i opieka wpływają istotnie na diagnozę dotyczącą choroby wątroby (zestawienie 3).

**ZESTAWIENIE (3) WYNIKÓW IMPLEMENTACJI KOMPUTEROWEJ
DLA WSPÓLCZYNNIKA GRAY-WILLIAMSZA**

Mienniki statystyczne	Charakterystyka współczynnika Gray-Williamsa
Wartość współczynnika	0,069
Poziom istotności	0,05
Hipoteza zerowa H_0	Y i Z nie wpływają istotnie na X
p -value	0,000
Wartość statystyki	94,41
Liczba stopni swobody	15

Są podstawy do odrzucenia hipotezy zerowej

Źródło: jak przy tabl. 1.

Przykład 2

Zbadano wpływ urbanizacji (Z) i położenia (Y) (zmienne objaśniające niezależne) na preferencję oliwek (X) (zmienna objaśniana) wyrażoną w sześciopunktowej skali porządkowej (A—F). Otrzymane wyniki zestawiono w formie tablicy trójdzielczej $6 \times 3 \times 2$ (tabl. 4).

TABL. 4. TABLICA TRÓJDZIELCZA $6 \times 3 \times 2$

Wyszczególnienie	Preferencja oliwek przez zamieszkujących						
	miasto			wieś			razem
	NW	NE	SW	NW	NE	SW	
A	20	18	12	30	23	11	114
B	15	17	9	22	18	9	90
C	12	18	23	21	20	26	120
D	17	18	21	17	18	19	110
E	16	6	19	8	10	17	76
F	28	25	30	12	15	24	134
Razem	108	102	114	110	104	106	644

Źródło: jak przy tabl. 3.

Po utworzeniu tablicy trójdzielczej (tabl. 4) w arkuszu „tablica” (począwszy od komórki B2) uruchomiono procedurę „WspKor3D”. W arkuszu „wyniki” za pomocą pola kombi wybrano opcję „Marcotorchino” i wyznaczono wartość współczynnika Marcotorchino, gdy Y i Z są cechami objaśniającymi wzajem-

nie niezależnymi, a X jest cechą objaśnianą. Następnie założono hipotezę zerową, że cechy urbanizacja (Z) i położenie (Y) nie wpływają istotnie na preferencję oliwek (X). Następnie wprowadzono poziom istotności testu α . Ponieważ $p_{v_M} = 0,004 \leq \alpha$, zatem są podstawy do odrzucenia hipotezy zerowej, czyli urbanizacja i położenie wpływają istotnie na preferencję oliwek (zestawienie 4).

**ZESTAWIENIE (4) WYNIKÓW IMPLEMENTACJI KOMPUTEROWEJ
DLA WSPÓŁCZYNNIKA MARCOTORCHINO**

Mienniki statystyczne	Charakterystyka współczynnika Marcotorchino
Wartość współczynnika	0,016
Poziom istotności	0,05
Hipoteza zerowa H_0	Y i Z nie wpływają istotnie na X
p -value	0,004
Wartość statystyki	50,46
Liczba stopni swobody	27

Są podstawy do odrzucenia hipotezy zerowej

Źródło: jak przy tabl. 1.

Przykład 3

Zbadano zadowolenie pacjentów danego oddziału (Z) (zmienna objaśniająca) z lekarzy (X) i pielęgniarek (Y) (zmienne objaśniane). Otrzymane wyniki zestawiono w formie tablicy trójdzielczej $3 \times 3 \times 4$ (tabl. 5).

TABL. 5. TABLICA TRÓJDZIELCZA $3 \times 3 \times 4$

Wyszczególnienie	Zadowolenie pacjentów z lekarzy/pielęgniarek oddziałów												
	chirurgia			ginekologia			wewnętrzny			ortopedia			razem
	0	1	2	0	1	2	0	1	2	0	1	2	
0	3	2	0	1	0	1	4	0	1	1	0	0	13
1	6	2	2	4	3	0	4	5	1	7	6	4	44
2	4	9	22	9	20	34	10	16	22	2	8	24	180
Razem	13	13	24	14	23	35	18	21	24	10	14	28	237

U w a g a. 0 — niezadowolony, 1 — prawie zadowolony, 2 — zadowolony.

Źródło: jak przy tabl. 3.

Po utworzeniu tablicy trójdzielczej (tabl. 5) w arkuszu „tablica” (począwszy od komórki B2) uruchomiono procedurę „WspKor3D”. W arkuszu „wyniki” za pomocą pola kombi wybrano opcję „Lombardo” i wyznaczono wartość współczynnika Lombardo, gdy Z jest cechą objaśniającą, a X i Y są cechami objaśnianymi. Następnie założono hipotezę zerową, że oddział (Z) nie wpływa istotnie

na liczbę pacjentów zadowolonych z lekarzy (X) i pielęgniarek (Y). Wprowadzono poziom istotności testu α . Ponieważ $pv_L = 0,95 > \alpha$, zatem nie ma podstaw do odrzucenia hipotezy zerowej, czyli oddział szpitalny nie wpływa istotnie na liczbę pacjentów zadowolonych z lekarzy (X) i pielęgniarek (Y) (zestawienie 5).

**ZESTAWIENIE (5) WYNIKÓW IMPLEMENTACJI KOMPUTEROWEJ
DLA WSPÓŁCZYNNIKA LOMBARDO**

Mienniki statystyczne	Charakterystyka współczynnika Lombardo
Wartość współczynnika	0,036
Poziom istotności	0,05
Hipoteza zerowa H_0	Z nie wpływa istotnie na X i Y
p -value	0,949
Wartość statystyki	16,99
Liczba stopni swobody	28

Nie ma podstaw do odrzucenia hipotezy zerowej

Źródło: jak przy tabl. 1.

Podsumowanie

Dysponując gotowym narzędziem do pomiaru siły związku między trzema cechami, które można pobrać z Internetu, czytelnik może samodzielnie przeprowadzić badania na podstawie danych zgromadzonych w formie tablicy trójdzielczej. Pomocne w tym są dokładnie opisane przykłady wraz z ich implementacją komputerową, w której istnieje możliwość wyboru jednego z trzech współczynników siły związku między cechami w tablicy trójdzielczej.

dr Piotr Sulewski — *Akademia Pomorska w Słupsku*

LITERATURA

- Beh E., Simonetti B., D'Ambra L. (2007), *Partitioning a non-symmetric measure of association for three-way contingency tables*, „Journal of Multivariate Analysis”, Vol. 98 (7)
- Cohen J. (1960), *A coefficient of agreement for nominal scales*, „Educational and Psychological Measurement”, Vol. 20 (1)
- Goodman L., Kruskal W. (1954), *Measures of Association for Cross Classifications*, „Journal of the American Statistical Association”, Vol. 49
- Gray L. N., Williams J. S. (1975), *Goodman and Kruskal's tau b: multiple and partial analogs*, [in:] *Proceedings of the Social Statistics Section*, American Statistical Association
- Light R. J., Margolin B. H. (1971), *An Analysis of Variance for Categorical Data*, „Journal of the American Statistical Association”, Vol. 66, No. 335

- Lombardo R. (2011), *Three-way association measure decompositions: The Delta index*, „Journal of Statistical Planning and Inference” No. 141
- Marcotorchino F. (1984), *Utilisation des Comparaisons par Paires en Statistique des Contin-
gences*, Partie (I), Etude IBM F069, France
- Oktaba W. (1974), *Elementy statystyki matematycznej i metodyka doświadczalnictwa*, PWN, War-
szawa
- Sobczyk M. (1996), *Statystyka*, PWN, Warszawa
- Sulewski P. (2007), *Moc tablicy dwudzielczej jako test niezależności*, „Wiadomości Statystyczne”,
nr 6, GUS
- Sulewski P. (2013), *Wielowymiarowe uogólnienie testu niezależności*, „Wiadomości Statystyczne”,
nr 12, GUS

SUMMARY

In the literature are popular statistical tool to measure the strength of the association between features in the 2-feature tables. However, the mechanisms to measure the strength of the association between the 3 feature in the tables are not too popular, and Polish literature practically does not exist. This paper describes an 3-featur array with its characteristics and provides 3-feature tables theory. The paper presents the computer implementation for measuring the strength of the relationship between the three modes with the Gray-Williams, Marcotorchino and Lombardo factors. The article also presents ready procedures and functions stored in VBA, enabling the reader to carry out independent surveys.

РЕЗЮМЕ

В статистической литературе популярными являются инструменты для измерения прочности связи между характеристиками в двойных таблицах. Однако механизмы для измерения прочности связи между характеристиками в тройных таблицах не очень популярны, а в литературе на польском языке практически не выступают.

В статье была определена тройная таблица вместе с ее характеристикой, а также была представлена теория тройных таблиц. Была представлена компьютерная реализация касающаяся измерения прочности связи между тремя характеристиками с использованием коэффициентов: Грэй-Уильямса, Маркаторчино и Ломбардо. Были разработаны готовые процедуры и функции представленные на языке VBA, позволяющие читателю самостоятельно проводить статистические обследования.