

Wielowymiarowe uogólnienie testu niezależności

W literaturze statystycznej znajdujemy głównie metody wnioskowania dotyczące jednej zmiennej. Jednak w badaniach statystycznych obiekty opisywane są za pomocą większej liczby cech. Podstawowym i często stosowanym narzędziem do badania związku między cechami są tablice wielodzielcze. Tablicę, która powstaje w wyniku podziału danych według dwóch cech nazywamy tablicą dwudzielczą. Kiedy uwzględniamy liczbę cech większą niż dwie, wówczas tworzymy tablice trójdzielcze, czterodzielcze, ..., n -dzielcze (Okta, 1974).

W publikacji Sulewskiego (2006) opisano procedurę generowania tablic dwudzielczych o zadanym związku między badanymi cechami oraz przedstawiono implementację komputerową testu niezależności dla dwóch cech zrealizowaną w edytorze VBA (*Visual Basic for Application*) arkusza kalkulacyjnego Excel.

W innym opracowaniu (Sulewski, 2007) zbadano 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.

Obecnie bardzo popularne wśród użytkowników komputerów są arkusze kalkulacyjne, w szczególności Microsoft Excel. Główne zalety tego programu to 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.

Kostki OLAP, stosowane np. w Narodowym Spisie Ludności i Mieszkań czy w Powszechnym Spisie Rolnym, są w swojej istocie niczym innym, jak tablicami wielodzielczymi.

W artykule przedstawiono tablicę trójdzielczą jako test niezależności chi-kwadrat wraz z jego implementacją komputerową w języku VBA. Zbadano też zdolność tablicy trójdzielczej do wykrywania związku między badanymi cechami, określaną jako moc testu.

Pierwszym celem pracy tego opracowania jest przedstawienie teorii tablic wielodzielczych, ze szczególnym uwzględnieniem tablic trójdzielczych. Autorowi wydaje się to zasadne, gdyż jedynym opracowaniem w języku polskim, prezentującym rozważania matematyczne dotyczące tablic wielodzielczych, jest wydana 40 lat temu książka W. Oktaby (1974). Drugim celem jest przedstawienie gotowych procedur i funkcji zapisanych w języku VBA, umożliwiających czytelnikowi samodzielne przeprowadzanie badań statystycznych w najbardziej popularnym

arkusza kalkulacyjnym, jakim jest Microsoft Excel. Celem trzecim jest zbadanie zdolności tablicy trójdzielczej do wykrywania związku między cechami.

TABLICA TRÓJDZIELCZA

Tablicę, która gromadzi wynik podziału próby według trzech cech A , B i C , gdzie n_{ij} jest liczebnością jednostek próby należących do i -tej kategorii cechy, A ($i=1, 2, \dots, w$), j -tej kategorii cechy B ($j=1, 2, \dots, k$), t -tej kategorii cechy C ($t=1, 2, \dots, p$), nazywamy tablicą trójdzielczą. Można ją sobie wyobrazić jako kostkę z w wierszami, k kolumnami i p płaszczyznami (wykr. 1).

Tablicę trójwymiarową można także zilustrować rozkładając płaszczyzny p obok siebie (zestawienie 1).

ZESTAWIENIE 1. TABLICA TRÓJDZIELCZA O WYMIARACH $w \times k \times p$

A/B	C							
	1				2			
	1	2	...	k	1	2	...	k
1	$X(1, 1, 1)$	$X(1, 2, 1)$	...	$X(1, k, 1)$	$X(1, 1, 2)$	$X(1, 2, 2)$	...	$X(1, k, 2)$
2	$X(2, 1, 1)$	$X(2, 2, 1)$	...	$X(2, k, 1)$	$X(2, 1, 2)$	$X(2, 2, 2)$	...	$X(2, k, 2)$
...	...	...	...	...	...	...	...	...
w	$X(w, 1, 1)$	$X(w, 2, 1)$	...	$X(w, k, 1)$	$X(w, 1, 2)$	$X(w, 2, 2)$	...	$X(w, k, 2)$

(dok.)

A/B	C							
	...				p			
	1	2	...	k	1	2	...	k
1	...	...	...	...	$X(1, 1, p)$	$X(1, 2, p)$	...	$X(1, k, p)$
2	...	...	...	...	$X(2, 1, p)$	$X(2, 2, p)$	...	$X(2, k, p)$
...	...	...	...	...	...	...	...	...
w	...	...	...	...	$X(w, 1, p)$	$X(w, 2, p)$	...	$X(w, k, p)$

Ź r ó ł o: opracowanie własne.

Wprowadźmy następujące oznaczenia liczby jednostek próby należących:

a) $Sw(i) = \sum_{j=1}^k \sum_{t=1}^p X(i, j, t)$ — do i -tej kategorii cechy A ,

b) $Sk(j) = \sum_{i=1}^w \sum_{t=1}^p X(i, j, t)$ — do j -tej kategorii cechy B ,

c) $Sp(t) = \sum_{i=1}^w \sum_{j=1}^k X(i, j, t)$ — do t -tej kategorii cechy C .

Wówczas:

$$n = \sum_{i=1}^w \sum_{j=1}^k \sum_{t=1}^p X(i, j, t) \quad (1)$$

jest całkowitą sumą tablicy trójdzielczej.

Oczekiwana liczebność podklas (i, j) w tablicy dwudzielczej AB , przy założeniu niezależności cech A i B w sumarycznym ujęciu cechy C , wynosi:

$$E_{AB}(i, j) = \frac{Sw(i) \cdot Sk(j)}{n} \quad i = 1, 2, \dots, w; \quad j = 1, 2, \dots, k \quad (2)$$

Wzór na statystykę χ^2 dla tablicy dwudzielczej AB ma postać (Oktała, 1974):

$$\chi_{AB}^2 = \sum_{i=1}^w \sum_{j=1}^k \frac{[S_{AB}(i, j) - E_{AB}(i, j)]^2}{E_{AB}(i, j)} \quad (3)$$

gdzie
$$S_{AB}(i, j) = \sum_{t=1}^p X(i, j, t) \quad i=1, 2, \dots, w; \quad j=1, 2, \dots, k \quad (4)$$

Gdy hipoteza zerowa H_0^{AB} o braku związku między cechami A i B jest słuszna, statystyka χ_{AB}^2 ma rozkład chi-kwadrat z $t = (w-1)(k-1)$ stopniami swobody. Na danym poziomie istotności α wyznaczamy wartość krytyczną $\chi_{\alpha, t}^2$. Jeśli $\chi_{AB}^2 \geq \chi_{\alpha, t}^2$, to związek między cechami występuje i można wyznaczyć siłę tej współzależności korzystając ze współczynnika zbieżności V Cramera w postaci (Józwiak, Podgórski, 1998):

$$V_{AB} = \sqrt{\frac{\chi_{AB}^2}{n \cdot (m-1)}} \quad (5)$$

gdzie $m = \min(w, k)$.

W zestawieniu 2 autor proponuje pewne zakresy wartości analizowanego współczynnika i przypisuje im odpowiednie stopnie współzależności cech.

ZESTAWIENIE 2. STOPNIE WSPÓLZALEŻNOŚCI CECH

Zakresy wartości współczynnika V Cramera	Stopnie współzależności cech
$(0, 0,33) \dots\dots\dots$	niski
$(0,33, 0,66) \dots\dots\dots$	średni
$(0,66, 0,99) \dots\dots\dots$	wysoki
1 $\dots\dots\dots$	funkcyjny

Źródło: jak przy zestawieniu 1.

Oczekiwana liczebność podklas (i, t) w tablicy dwudzielczej AC , przy założeniu niezależności cech A i C w sumarycznym ujęciu cechy B , wynosi:

$$E_{AC}(i, t) = \frac{Sw(i) \cdot Sp(t)}{n} \quad i=1, 2, \dots, w; \quad t=1, 2, \dots, p \quad (6)$$

Wzór na statystykę χ^2 dla tablicy dwudzielczej AC ma postać:

$$\chi_{AC}^2 = \sum_{i=1}^w \sum_{t=1}^p \frac{[S_{AC}(i,t) - E_{AC}(i,t)]^2}{E_{AC}(i,t)} \quad (7)$$

gdzie
$$S_{AC}(i,t) = \sum_{j=1}^k X(i,j,t) \quad i=1,2,\dots,w; \quad t=1,2,\dots,p \quad (8)$$

Gdy hipoteza zerowa H_0^{AC} o braku związku między cechami A i C jest słuszna, statystyka χ_{AC}^2 ma rozkład chi-kwadrat z $q_1=(w-1)(p-1)$ stopniami swobody. Na danym poziomie istotności α wyznaczamy wartość krytyczną χ_{α,q_1}^2 . Jeśli $\chi_{AC}^2 \geq \chi_{\alpha,q_1}^2$, to interakcja cech występuje i można wyznaczyć siłę tej współzależności (zestawienie 2) korzystając ze współczynnika zbieżności V Cramera (5).

Oczekiwana liczebność podklas (j, t) w tablicy dwudzielczej BC , przy założeniu niezależności cech B i C w sumarycznym ujęciu cechy A , wynosi:

$$E_{BC}(j,t) = \frac{Sk(j) \cdot Sp(t)}{n} \quad j=1,2,\dots,k; \quad t=1,2,\dots,p \quad (9)$$

Wzór na statystykę χ^2 dla tablicy dwudzielczej BC ma postać:

$$\chi_{BC}^2 = \sum_{j=1}^k \sum_{t=1}^p \frac{[S_{BC}(j,t) - E_{BC}(j,t)]^2}{E_{BC}(j,t)} \quad (10)$$

gdzie
$$S_{BC}(j,t) = \sum_{i=1}^w X(i,j,t) \quad j=1,2,\dots,k; \quad t=1,2,\dots,p \quad (11)$$

Gdy hipoteza zerowa H_0^{BC} o braku związku między cechami B i C jest słuszna, statystyka χ_{BC}^2 ma rozkład chi-kwadrat z $q_2=(k-1)(p-1)$ stopniami swobody. Na danym poziomie istotności α wyznaczamy wartość krytyczną χ_{α,q_2}^2 . Jeśli $\chi_{BC}^2 \geq \chi_{\alpha,q_2}^2$, to interakcja cech występuje i można wyznaczyć siłę tej współzależności (zestawienie 2) korzystając ze współczynnika zbieżności V Cramera (5).

Oczekiwana liczebność podklas (i, j, t) w tablicy trójdzielczej, przy założeniu niezależności cech A , B i C , wynosi:

$$E(i,j,t) = \frac{Sw(i) \cdot Sk(j) \cdot Sp(t)}{n^2} \quad i=1,2,\dots,w; \quad j=1,2,\dots,k; \quad t=1,2,\dots,p \quad (12)$$

Na mocy tej formuły wzór na statystykę χ^2 dla tablicy trójdzielczej ma postać (Oktaba, 1974):

$$\chi_{ABC}^2 = \sum_{i=1}^w \sum_{j=1}^k \sum_{t=1}^p \frac{[X(i,j,t) - E(i,j,t)]^2}{E(i,j,t)} - (\chi_{AB}^2 + \chi_{BC}^2 + \chi_{AC}^2) \quad (13)$$

Gdy hipoteza zerowa H_0^{ABC} o braku związku między cechami A , B i C jest słuszna, statystyka χ_{ABC}^2 ma rozkład chi-kwadrat z $q_3 = (w-1)(k-1)(p-1)$ stopniami swobody. Na danym poziomie istotności α wyznaczamy wartość krytyczną χ_{α, q_3}^2 . Jeśli $\chi_{ABC}^2 \geq \chi_{\alpha, q_3}^2$, to interakcja między cechami występuje i można wyznaczyć siłę tej współzależności (zestawienie 2). Współczynnik zbieżności V Cramera (5) dotyczy tylko tablic dwudzielczych, dlatego proponuję jego zmodyfikowaną postać:

$$V_{ABC} = \sqrt{\frac{\chi_{ABC}^2}{n \cdot (m-1)}} \quad (14)$$

gdzie $m = \min(w, k, p)$.

Badanie współzależności trzech cech można uogólnić na dowolną liczbę m cech. Wówczas odjemnik we wzorze (13) jest sumą wartości χ^2 wszystkich interakcji 2 cech, 3 cech, ... cech, $(m-1)$ cech. Odpowiednie liczby stopni swobody dla interakcji 2 cech to np. $q = (w-1)(k-1)$, dla 3 cech — $q = (w-1)(k-1)(p-1)$, ..., gdzie $w, k, p, r, s, \dots$ oznaczają liczby kategorii odpowiednich cech. Odjemna we wzorze (13) to m sum rozciągniętych na kategorie. Oczekiwaną liczebność $E(.)$ wyznacza się z relacji:

$$E(.) = \frac{\text{iloczyn } m \text{ sum brzegowych tablicy dwudzielczej}}{n^{m-1}} \quad (15)$$

gdzie n — liczebność próby.

Tak więc w przypadku tablicy czterodzielczej $2 \times 3 \times 3 \times 2$ z cechami A, B, C i D oczekiwana liczebność podklas (i, j, t, u) , przy założeniu niezależności cech, wynosi:

$$E(i, j, t) = \frac{Sw(i) \cdot Sk(j) \cdot Sp(t) \cdot Sr(u)}{n^3} \quad i=1, 2; \quad j=1, 2, 3; \quad t=1, 2, 3; \quad u=1, 2 \quad (16)$$

Na mocy powyższego wzór na statystykę χ^2 dla tablicy $2 \times 3 \times 3 \times 2$ ma postać:

$$\chi^2_{ABCD} = \sum_{i=1}^2 \sum_{j=1}^3 \sum_{t=1}^3 \sum_{u=1}^2 \frac{[X(i,j,t) - E(i,j,t)]^2}{E(i,j,t)} - \chi_T \quad (17)$$

$$\text{gdzie } \chi_T = (\chi^2_{AB} + \chi^2_{AC} + \chi^2_{AD} + \chi^2_{BC} + \chi^2_{BD} + \chi^2_{CD} + \chi^2_{ABC} + \chi^2_{ABD} + \chi^2_{ACD} + \chi^2_{BCD}) \quad (18)$$

Gdy hipoteza zerowa H_0^{ABCD} o braku związku między cechami A, B, C i D jest słuszna, statystyka χ^2_{ABCD} ma rozkład chi-kwadrat z $q=(2-1)(3-1)(3-1)(2-1)=4$ stopniami swobody. Na danym poziomie istotności α wyznaczamy wartość krytyczną $\chi^2_{\alpha,4}$. Jeśli $\chi^2_{ABCD} \geq \chi^2_{\alpha,4}$, to interakcja między cechami występuje i można wyznaczyć siłę tej współzależności korzystając ze współczynnika zbieżności V Cramera.

Implementację komputerową tablicy trójdzielczej jako testu niezależności utworzoną w edytorze VBA arkusza kalkulacyjnego Excel przedstawiono w pliku „test_3d.xls”, który umieszczono w Internecie pod adresem <http://www.utogim.eu/tablica.xls>. Składa się na nią procedura „TabTro”, która odwołuje się do funkcji „ChiKw2”, „ChiKw3”, „Cramer”. Aby ułatwić czytelnikowi zrozumienie użytych kodów zostały one wzbogacone o liczne komentarze umieszczone po apostrofach i zapisane kursywą, a ponadto w zestawieniu 3 zawarto opis zmiennych wykorzystanych przy tworzeniu implementacji komputerowej VBA.

ZESTAWIENIE 3. ZMIENNE WYKORZYSTYWANE W IMPLEMENTACJI KOMPUTEROWEJ WRAZ Z ICH LOKALIZACJĄ

Zmienne	Opis zmiennych	Lokalizacja zmiennych			
		TabTro	ChiKw2	ChiKw3	Cramer
X	liczebność komórek w tablicy dwudzielczej lub trójdzielczej	x	x	x	—
i	indeks wiersza	x	x	x	—
w	liczba wierszy	x	x	x	x
j	indeks kolumny	x	x	x	—
k	liczba kolumn	x	x	x	x
t	indeks płaszczyzny	x	—	x	—
p	liczba płaszczyzn	x	—	x	x
n	liczebność próby	x	x	x	x
Sk	suma realizacji w kolumnach tablicy dwudzielczej lub trójdzielczej	—	x	s	—
Sw	suma realizacji w wierszach tablicy dwudzielczej lub trójdzielczej	—	x	s	—
E	liczebność oczekiwana w komórkach tablicy dwudzielczej lub trójdzielczej	—	x	s	—
m	min (w, k) lub min (w, k, p)	—	—	—	x

**ZESTAWIENIE 3. ZMIENNE WYKORZYSTYWANE W IMPLEMENTACJI KOMPUTEROWEJ
WRAZ Z ICH LOKALIZACJĄ (dok.)**

Zmienne	Opis zmiennych	Lokalizacja zmiennych			
		TabTro	ChiKw2	ChiKw3	Cramer
<i>v</i>	współczynnik V Cramera	—	—	—	x
<i>Chi</i>	statystyka chi-kwadrat	—	x	x	x
<i>Sp</i>	suma realizacji w płaszczyznach tablicy trójdzielczej	—	—	x	—
<i>H0AB, H0AC, H0BC</i>	hipoteza zerowa dla tablic dwudzielczych	x	—	—	—
<i>H0ABC</i>	hipoteza zerowa dla tablicy trójdzielczej	x	—	—	—
<i>SwAB, SwAC, SwBC</i>	liczba stopni swobody dla tablic dwudzielczych	x	—	—	—
<i>SwABC</i>	liczba stopni swobody dla tablicy trójdzielczej	x	—	—	—
<i>WkAB, WkAC, WkBC</i>	wartość krytyczna dla tablic dwudzielczych	x	—	—	—
<i>WkABC</i>	wartość krytyczna dla tablicy trójdzielczej	x	—	—	—
<i>SAB</i>	suma realizacji według kategorii cechy <i>C</i>	x	—	—	—
<i>SAC</i>	suma realizacji według kategorii cechy <i>B</i>	x	—	—	—
<i>SBC</i>	suma realizacji według kategorii cechy <i>A</i>	x	—	—	—
<i>ChiAB, ChiAC, ChiBC</i>	statystyka chi-kwadrat dla tablic dwudzielczych	x	—	—	—
<i>ChiABC</i>	statystyka chi-kwadrat dla tablicy trójdzielczej	x	—	—	—
<i>Okno</i>	okno dialogowe	x	—	—	—
<i>Wyniki</i>	wyniki wyświetlane z tabl. 5	x	—	—	—
<i>s</i>	suma realizacji w komórkach tablicy trójdzielczej	x	—	—	—
<i>Po</i>	wartość poziomu istotności	x	—	—	—
<i>Stopień</i>	stopień współzależności cech	—	—	—	x

U w a g a. Znak „x” oznacza lokalizację.

Ź r ó d ł o: jak przy zestawieniu 1.

OBSŁUGA IMPLEMENTACJI TESTU NIEZALEŻNOŚCI

Po otwarciu pliku test_3d.xls do dyspozycji użytkownika są dwa przyciski. Przycisk „czyść dane” usuwa informacje z tabl. 5. Przycisk „start” uruchamia test niezależności dotyczący tablicy trójdzielczej, a zadaniem korzystającego z tego programu jest podanie w oknie dialogowym:

- wartości w , k , p opisujących rozmiar tablicy trójdzielczej,
- liczebności próby n ,
- poziomu istotności α ,
- wartości realizacji w poszczególnych komórkach tablicy trójdzielczej.

Jeżeli suma realizacji w komórkach tablicy trójdzielczej będzie różna od liczebności próby, to pojawi się okno dialogowe informujące o tym fakcie i kończące działanie procedury. Gdy dane są wprowadzone poprawnie, procedura wypełnia tabelę umieszczoną w arkuszu „test”.

W celu badania testów trzeba dysponować narzędziem do nadawania populacji generalnej określonej właściwości, jaką jest związek między cechami. Dlatego dane podlegające opracowaniu muszą być danymi z generatora liczb losowych, a nie danymi wziętymi z praktyki. Należy wiedzieć, jaki jest rzeczywisty związek między cechami, inaczej nie zbada się mocy testów.

Realizując to zaproponowano dwa przykłady liczbowe. Przykład 1 dotyczy generacji zawartości tablicy trójdzielczej za pomocą liczb losowych równomiernych (wylosowanych z rozkładu jednostajnego), gdy związek między cechami istnieje (losowanie zależne), natomiast przykład 2 dotyczy sytuacji, gdy związek między cechami nie istnieje (losowanie niezależne). Przechodzenie od zmiennej ciągłej do zmiennych kategorialnych uzyskano metodą „słupkową”.

Przykład 1

Posługując się generatorem liczb losowych równomiernych symulowano pobieranie $n=1000$ -elementowej próby z pewnej fikcyjnej populacji generalnej. Otrzymaną próbę podzielono według trzech cech A , B i C , z których każda miała trzy warianty. Na poziomie istotności 0,05 zbadano, czy istnieje współzależność między cechami A , B i C . Wyniki symulacji zestawiono w tablicy trójdzielczej $3 \times 3 \times 3$ (tabl. 1).

TABL. 1. WYNIKI SYMULACJI

Cechy	C_1			C_2			C_3		
	B_1	B_2	B_3	B_1	B_2	B_3	B_1	B_2	B_3
A_1	83	33	37	32	31	26	37	30	36
A_2	32	31	37	26	89	32	27	32	21
A_3	35	29	31	22	25	30	32	33	91

Źródło: opracowanie własne.

Tablicę trójdzielczą podzielono na tablice dwudzielcze 3×3 dla następujących par cech:

- A i B (tabl. 2),
- A i C (tabl. 3),
- B i C (tabl. 4).

Liczebność w tych tablicach otrzymano kolejno przez sumowanie liczebności dla poszczególnych kategorii zmiennej odpowiednio C , B i A .

**TABL. 2. TABLICA DWUDZIELCZA
DLA CECH A I B**

Cechy	B_1	B_2	B_3	Razem
A_1	152	94	99	345
A_2	85	152	90	327
A_3	89	87	152	328
Razem	326	333	341	1000

Źródło: jak przy tabl. 1.

**TABL. 3. TABLICA DWUDZIELCZA
DLA CECH A I C**

Cechy	C_1	C_2	C_3	Razem
A_1	153	89	103	345
A_2	100	147	80	327
A_3	95	77	156	328
Razem	348	313	339	1000

Źródło: jak przy tabl. 1.

TABL. 4. TABLICA DWUDZIELCZA DLA CECH B I C

Cechy	C_1	C_2	C_3	Razem
B_1	150	93	105	348
B_2	80	145	88	313
B_3	96	95	148	339
Razem	326	333	341	1000

Źródło: jak przy tabl. 1.

W sytuacji poprawnie wprowadzonych danych procedura „TabTro” wypełnia tablicę umieszczoną w arkuszu „test” (tabl. 5).

TABL. 5. WSPÓLZALEŻNOŚĆ CECH A , B I C

Charakterystyka	A i B	B i C	A i C	A , B i C
Stopnie swobody	4	4	4	8
Wartość krytyczna	9,488	9,488	9,488	15,507
Wartość statystyki	68,179	55,688	71,460	51,117
Związek między cechami	występuje	występuje	występuje	występuje
Współzależność	niska	niska	niska	niska

Źródło: jak przy tabl. 1.

Na podstawie uzyskanych wyników z prawdopodobieństwem popełnienia błędu I rodzaju równym 0,05 stwierdzono, że istnieje niska współzależność między cechami A , B i C oraz między każdą parą rozpatrywanych cech.

Przykład 2

Korzystając z generatora liczb losowych równomiernych z fikcyjnej populacji generalnej symulowano pobieranie $n=500$ -elementowej próby. Otrzymaną próbę podzielono według cech A , B i C , z których cechy A i C miały trzy warianty, a cecha B — cztery warianty. Sprawdzono na poziomie istotności 0,1 czy istnieje współzależność między cechami A , B i C . Wyniki symulacji zestawiono w tablicy trójdzielczej $2 \times 4 \times 2$ (tabl. 6).

TABL. 6. WYNIKI SYMULACJI

Cechy	C_1				C_2			
	B_1	B_2	B_3	B_4	B_1	B_2	B_3	B_4
A_1	24	31	41	38	24	33	35	24
A_2	31	31	33	24	34	35	28	34

Źródło: jak przy tabl. 1.

W celu uzyskania odpowiedniego wniosku tablicę trójdzielczą podzielono na tablice dwudzielcze:

- a) 2×4 dla pary cech A i B (tabl. 7),
- b) 2×2 dla pary cech A i C (tabl. 8),
- c) 4×2 dla pary cech B i C (tabl. 9).

Liczebność w tych tablicach otrzymano kolejno sumując liczebność poszczególnych kategorii zmiennej C , B i A , odpowiednio:

**TABL. 7. TABLICA DWUDZIELCZA
DLA CECH A I B**

Cechy	B_1	B_2	B_3	B_4	Razem
A_1	48	64	76	62	250
A_2	65	66	61	58	250
Razem	113	130	137	120	500

Źródło: jak przy tabl. 1.

**TABL. 8. TABLICA DWUDZIELCZA
DLA CECH A I C**

Cechy	C_1	C_2	Razem
A_1	134	116	250
A_2	119	131	250
Razem	253	247	500

Źródło: jak przy tabl. 1.

**TABL. 9. TABLICA DWUDZIELCZA
DLA CECH B I C**

Cechy	C_1	C_2	Razem
B_1	55	58	113
B_2	62	68	130
B_3	74	63	137
B_4	62	58	120
Razem	253	247	500

Źródło: jak przy tabl. 1.

W sytuacji poprawnie wprowadzonych danych procedura „TabTro” wypełnia tablicę umieszczoną w arkuszu „test” (tabl. 10).

TABL. 10. WSPÓLZALEŻNOŚĆ CECH A , B I C

Charakterystyka	A i B	B i C	A i C	A , B i C
Stopnie swobody	3	3	1	3
Wartość krytyczna	6,251	6,251	2,706	6,251
Wartość statystyki	4,364	1,301	1,800	3,109
Związek między cechami	nie występuje	nie występuje	nie występuje	nie występuje

Źródło: jak przy tabl. 1.

Na podstawie uzyskanych wyników z prawdopodobieństwem popełnienia błędu I rodzaju równego 0,1 stwierdzono, że współzależność między cechami A , B i C oraz między każdą parą rozpatrywanych cech nie występuje.

ZDOLNOŚĆ TABLICY TRÓJDZIELCZEJ DO WYKRYWANIA ZWIĄZKU MIĘDZY CECAMI

Znając wartości statystyki χ^2_{ABC} dla różnej liczebności próby n można wyznaczyć moc testu M , czyli prawdopodobieństwo odrzucenia hipotezy zerowej H_0^{ABC} o niezależności cech A , B i C , gdy nie jest ona prawdziwa:

$$M = \frac{R - U}{R} \quad (19)$$

U określa liczbę tych przypadków R , kiedy to wartość statystyki χ^2_{ABC} jest mniejsza od wartości krytycznej $\chi^2_{\alpha, q}$, gdzie α jest poziomem istotności oraz $q = (w-1)(k-1)(p-1)$ — liczbą stopni swobody.

W celu zbadania zdolności tablicy trójdzielczej do wykrywania związku między cechami A , B i C niezbędna jest generacja tablic trójdzielczych. Uwzględniając narzuconą siłę związku między cechami, do wypełnienia tablicy trójdzielczej skorzystano z metody „słupkowej” wykorzystującej rozkład równomierny. Metodę tę dla tablicy dwudzielczej przedstawiono w cytowanej już pracy P. Sulewskiego (2007).

W celu wypełnienia tablicy trójdzielczej o wymiarach $w \times k \times p$ liczbami losowymi równomiernymi, przedział $(0; 1)$ podzielono na wkp podprzedziałów o szerokościach liczbowo równych prawdopodobieństwu $p_{ijt} (i=1, 2, \dots, w; j=1, 2, \dots, k; t=1, 2, \dots, p)$, które spełnia warunek normalizacji:

$$\sum_{i=1}^w \sum_{j=1}^k \sum_{t=1}^p p_{ijt} = 1 \quad (20)$$

Każda z n wygenerowanych liczb losowych równomiernych „wpada” do jednego z wkp podprzedziałów i tym samym liczba obiektów w odpowiadającej temu podprzedziałowi komórce tablicy trójdzielczej zostaje zwiększona o jeden. Wielkości $X_{ijt} (i=1, 2, \dots, w; j=1, 2, \dots, k; t=1, 2, \dots, p)$ spełniające równość

$$\sum_{i=1}^w \sum_{j=1}^k \sum_{t=1}^p X_{ijt} = n \quad (21)$$

są liczebnością obiektów w poszczególnych komórkach tablicy trójdzielczej.

Przykład 3

Aby zbadać zdolność tablicy trójdzielczej $3 \times 3 \times 3$ do wykrywania związku między cechami A , B i C , zaproponowano cztery grupy wartości prawdopodobieństw $p_{ijt}(i, j, t=1, 2, 3)$ w celu generowania zawartości tablicy trójdzielczej metodą „słupkową”. Grupę I tworzą prawdopodobieństwa o równej wartości, co odpowiada sytuacji, gdy hipoteza zerowa H_0^{ABC} o braku związku między cechami jest słuszna. Grupy II–IV (tabl. 11) odzwierciedlają sytuację, gdy hipoteza zerowa H_0^{ABC} jest niesłuszna, zachodzi bowiem związek między cechami A , B i C . W trzech komórkach o współrzędnych $(1, 1, 1)$, $(2, 2, 2)$, $(3, 3, 3)$ — oznaczone pogrubioną czcionką — związek wszystkich realizacji stanowi odpowiednio: 20% (grupa II), 25% (grupa III) oraz 30% (grupa IV).

TABL. 11. WARTOŚCI PRAWDOPODOBIEŃSTW WEDŁUG GRUP I–IV

Cechy	C_1			C_2			C_3		
	B_1	B_2	B_3	B_1	B_2	B_3	B_1	B_2	B_3
Grupa I									
A_1	0,037	0,037	0,037	0,037	0,037	0,037	0,037	0,037	0,037
A_2	0,037	0,037	0,037	0,037	0,037	0,037	0,037	0,037	0,037
A_3	0,037	0,037	0,037	0,037	0,037	0,037	0,037	0,037	0,037
Grupa II									
A_1	0,067	0,033	0,033	0,033	0,033	0,033	0,033	0,033	0,033
A_2	0,033	0,033	0,033	0,033	0,067	0,033	0,033	0,033	0,033
A_3	0,033	0,033	0,033	0,033	0,033	0,033	0,033	0,033	0,067
Grupa III									
A_1	0,083	0,031	0,031	0,031	0,031	0,031	0,031	0,031	0,031
A_2	0,031	0,031	0,031	0,031	0,083	0,031	0,031	0,031	0,031
A_3	0,031	0,031	0,031	0,031	0,031	0,031	0,031	0,031	0,083
Grupa IV									
A_1	0,100	0,029	0,029	0,029	0,029	0,029	0,029	0,029	0,029
A_2	0,029	0,029	0,029	0,029	0,100	0,029	0,029	0,029	0,029
A_3	0,029	0,029	0,029	0,029	0,029	0,029	0,029	0,029	0,100

Źródło: jak przy tabl. 1.

W przypadku każdej liczebności próby $n \in \{100; 200; 400\}$, $R=1000$ razy obliczono wartość statystyki testowej χ_{ABC}^2 . Przyjmując poziom istotności $\alpha = 0,05$ otrzymano dla $t=(3-1)(3-1)(3-1)=8$ stopni swobody, wartość krytyczną $\chi_{0,05;8}^2=15,507$ i wyznaczono moc testu na podstawie wzoru (19).

Wykres 2 przedstawia zależność mocy testu od liczebności próby. W grupie I, gdy hipoteza zerowa H_0^{ABC} jest słuszna, moc testu oscyluje wokół założonego poziomu istotności $\alpha = 0,05$, co potwierdza prawidłowość przeprowadzonych

symulacji. W grupach II—IV moc testu rośnie wraz ze wzrostem liczebności próby i największe wartości przyjmuje w grupie IV, gdy niespełnienie hipotezy zerowej H_0^{ABC} jest największe.

Podsumowanie

Na podstawie wyników przeprowadzonych badań można stwierdzić, że w zakresie analizowanych prób tablica trójdzielcza ma zdolność do wykrywania związku między badanymi cechami, aczkolwiek zdolność ta bardzo silnie zależy od liczebności próby. W sytuacji gdy hipoteza zerowa o niezależności cech nie jest słuszna, moc testu rośnie wraz ze wzrostem liczebności próby i siły związku między cechami. Dysponując próbą $n=100$ mamy tylko 50% szans na odrzucenie ewidentnie niesłusznej hipotezy zerowej H_0^{ABC} dla IV grupy wartości prawdopodobieństw.

dr Piotr Sulewski — *Akademia Pomorska w Słupsku*

LITERATURA

- Józwiak J., Podgórski J. (1998), *Statystyka od podstaw*, PWE, Warszawa
Oktaba W. (1974), *Elementy statystyki matematycznej i metodyka doświadczalnictwa*, PWN, Warszawa
Sobczyk M. (1996), *Statystyka*, PWN, Warszawa

- Sulewski P. (2006), *Test niezależności dwóch cech realizowany za pomocą tablicy dwudzielczej*, „Śląskie Prace Matematyczno-Fizyczne”, nr 4, Śląsk
- Sulewski P. (2007), *Moc tablicy dwudzielczej jako test niezależności*, „Wiadomości Statystyczne”, nr 6, GUS

SUMMARY

This paper discusses the theory of contingency tables with particular emphasis on three-dimension-tables. Their use in statistical surveys is stressed in the article. The three-dimension-table is presented as a chi-square test of independence and its implementation in VBA (Visual Basic for Applications) language. In this way, users are provided ready-made procedures and functions for carrying out surveys in Microsoft Excel. It has been examined also the ability of the three-dimension-table to detect the relationship between the measured traits, referred to as the power of the test.

РЕЗЮМЕ

Статья характеризует теорию многораздельных таблиц с особым учетом трехраздельных таблиц. Было показано на их использование в статистических обследованиях. Статья представляет трехраздельную таблицу в качестве теста независимости χ^2 -квadrat вместе с его реализацией на языке VBA (visual Basic for Applications). Таким образом пользователям были предоставлены готовые процедуры и функции позволяющие проводить статистические обследования на калькуляционном листе в программе Microsoft Excel. Была исследована также способность трехразделительной таблицы для определения соотношения между обследуемыми признаками, называемая силой теста.