

STUDIA METODOLOGICZNE

Piotr SULEWSKI

Wyznaczanie obszaru krytycznego przy testowaniu niezależności w tablicach wielodzielczych

Tablice wielodzielcze są podstawowym i często stosowanym narzędziem do badania związku między cechami. Tablicę, która powstaje w wyniku podziału danych według dwóch cech nazywamy dwudzielczą, trzech cech — trójdzielczą, ..., n cech — n -dzielczą. Przy testowaniu niezależności cech najbardziej popularną statystyką testową dla tablic dwudzielczych jest statystyka χ^2 Pearsona, która ma także swoje rozszerzenia dla tablic trójdzielczych i wyższych.

W przypadku tablic 2×2 i tablic $w \times k$ ($w, k > 2$) istnieją jednak ograniczenia co do stosowania statystyki χ^2 Pearsona, która ma asymptotyczny rozkład chi-kwadrat z $(w-1)(k-1)$ stopniami swobody. Także w przypadku tablic trójdzielczych $w \times k \times p$, jeżeli liczba realizacji w komórkach jest mała, rozkład statystyki testowej znacząco odbiega od rozkładu chi-kwadrat z $wkp - (w+k+p) + 2$ stopniami swobody. W celu zniesienia tych ograniczeń w artykule proponuję wyznaczanie wartości krytycznych za pomocą symulacji komputerowych metodą Monte Carlo. Także Lilliefors w teście Kołmogorowa dla rozkładu normalnego wyznaczył wartości krytyczne drogą symulacyjną, gdy parametry rozkładu były oszacowane z próby.

Celem artykułu jest dostarczenie czytelnikowi gotowej implementacji komputerowej napisanej w edytorze VBA (*Visual Basic for Applications*) arkusza kalkulacyjnego Excel, do badania niezależności dwóch i trzech cech w formie pliku zamieszczonego w Internecie.

Tablica dwudzielcza 2×2

Najprostszą postacią tablicy dwudzielczej jest tablica 2×2 , która składa się z 4 liczebności rozkładu łącznego cech X i Y oznaczonych literami a, b, c, d (tabl. 1). Wówczas ogólna liczba obserwacji $n = a + b + c + d$.

TABL. 1. TABLICA DWUDZIELCZA 2×2

Cecha X	Cecha Y		Razem
	Y_1	Y_2	
X_1	a	b	$a + b$
X_2	c	d	$c + d$
Razem	$a + c$	$b + d$	n

Źródło: opracowanie własne.

Przyjmując oznaczenia zawarte w tej tablicy statystyka χ^2 Pearsona ma postać:

$$X_{2 \times 2}^2 = \frac{n(ad - bc)^2}{(a + b)(c + d)(a + c)(b + d)} \quad (1)$$

Liczebność oczekiwana dla tej tablicy wynosi odpowiednio:

$$\begin{aligned} e_1 &= \frac{(a + b)(a + c)}{n} & e_2 &= \frac{(a + b)(b + d)}{n} & e_3 &= \frac{(a + c)(c + d)}{n} \\ e_4 &= \frac{(b + d)(c + d)}{n} \end{aligned} \quad (2)$$

Wzór (1) można stosować, gdy $n \geq 40$ oraz $e_i \geq 5$ dla każdego $i = 1, \dots, 4$ (Cochran, 1952). Jeżeli natomiast $e_i < 5$ dla pewnego $i = 1, \dots, 4$, to sugeruje się stosowanie poprawki Yatesa danej wzorem (Cochran, 1952):

$$X_Y^2 = \frac{n(|ad - bc| - 0,5n)^2}{(a + b)(c + d)(a + c)(b + d)} \quad (3)$$

Poprawka Yatesa na ciągłość ma na celu zapewnienie możliwości przyjmowania przez statystykę testową wszystkich wartości rzeczywistych zgodnie z istotą rozkładu chi-kwadrat. Należy jednak wspomnieć, że poprawka Yatesa nie jest jedynym narzędziem stosowanym przez statystyków. Inne typy rozwiązań w tym zakresie podali m.in. Yates i Campbell i są one następujące, jeżeli:

- 1) $e_1 < 5$ dla pewnego $i = 1, \dots, 4$, to należy stosować test Fishera-Yatesa (Yates, 1934). Cochran jednak zauważył, że warunek $e_1 < 5$ może w przyszłości wymagać modyfikacji;
- 2) $5 > e_i \geq 1$ dla pewnego $i = 1, \dots, 4$, to zdaniem Campbella (Campbell, 2007) dla tablic 2×2 należy stosować test ' $n-1$ ' chi-kwadrat w postaci (Pearson, 1947):

$${}_{n-1}X_{2 \times 2}^2 = X_{2 \times 2}^2 \cdot \frac{n-1}{n} = \frac{(n-1)(ad-bc)^2}{(a+b)(c+d)(a+c)(b+d)} \quad (4)$$

- 3) $e_1 < 1$ dla pewnego $i = 1, \dots, 4$, to należy stosować test Fishera-Yatesa (Campbell, 2007).

Gdy hipoteza zerowa H_0 o niezależności cech X i Y jest słuszna, statystyka $X_{2 \times 2}^2$ ma asymptotyczny rozkład chi-kwadrat z jednym stopniem swobody i jest miarą rozbieżności między rozkładem zaobserwowanym i hipotetycznym. Wykr. 1 przedstawia rozkład chi-kwadrat z jednym stopniem swobody oraz empiryczną funkcję gęstości rozkładu statystyki χ^2 uzyskaną za pomocą metody Parzena dla tablicy 2×2 i liczebności próby n , gdy H_0 jest słuszna. Dokładny opis tej metody znajduje się w innym opracowaniu pióra autora tegoż artykułu (Sulewski, 2013). Liczebność próby dobrano w taki sposób, aby pokazać:

- brak zgodności rozkładu statystyki χ^2 z rozkładem chi-kwadrat dla tablicy 2×2 ($n = 50$);
- zgodność rozkładu statystyki χ^2 z rozkładem chi-kwadrat dla tablicy 2×2 ($n = 500$).

Tablica dwudzielcza większa niż 2×2

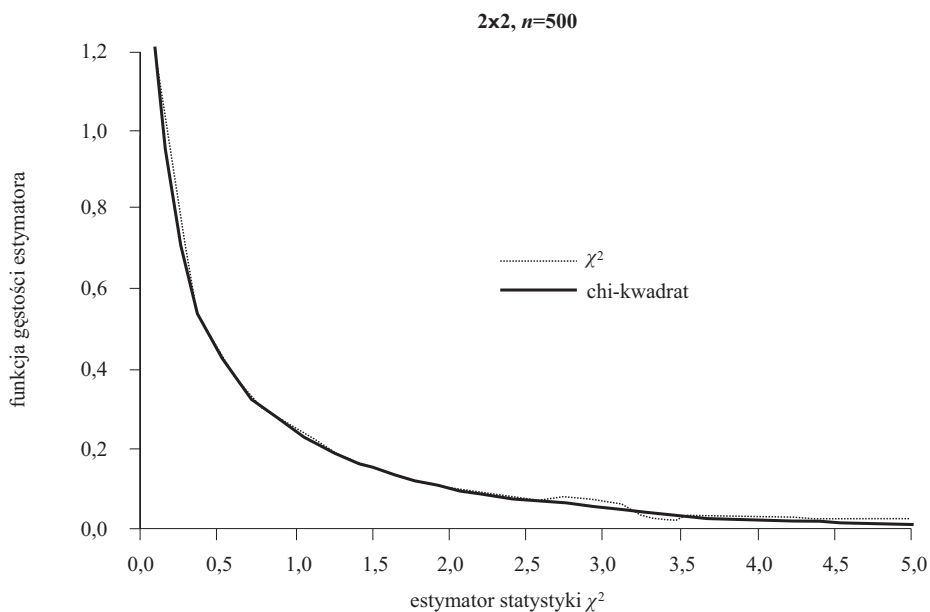
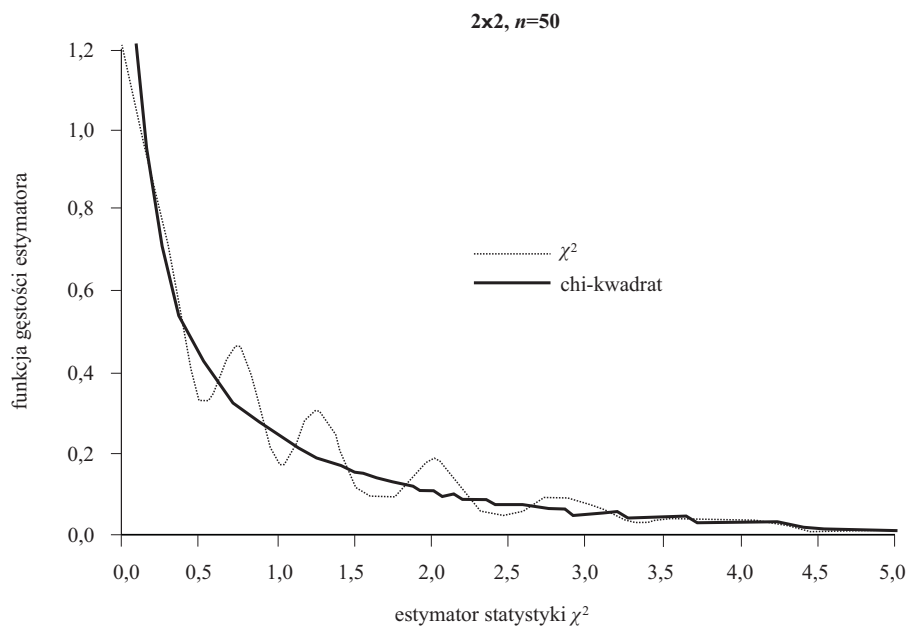
Tablica 2 przedstawia tablicę dwudzielczą $w \times k$, która składa się z $w \cdot k$ liczebności n_{ij} ($i = 1, \dots, w$; $j = 1, \dots, k$) rozkładu łącznego cech X i Y .

TABL. 2. TABLICA DWUDZIELCZA $w \times k$

Cecha X	Cecha Y				Razem
	Y_1	Y_2	...	Y_k	
X_1	n_{11}	n_{12}	...	n_{1k}	$n_{1\bullet}$
X_2	n_{21}	n_{22}	...	n_{2k}	$n_{2\bullet}$
...	...	...	...	...	...
X_w	n_{w1}	n_{w2}	...	n_{wk}	$n_{w\bullet}$
Razem	$n_{\bullet 1}$	$n_{\bullet 2}$	...	$n_{\bullet k}$	n

Źródło: jak przy tabl. 1.

Wykr. 1. ZGODNOŚĆ ROZKŁADU STATYSTYKI χ^2 PEARSONA Z ROZKŁADEM CHI-KWADRAT DLA TABLICY 2x2



Źródło: opracowanie własne.

Wartość n jest sumą wszystkich liczebności tablicy dwudzielczej:

$$n = \sum_{i=1}^w n_{i\bullet} = \sum_{j=1}^k n_{\bullet j} = \sum_{i=1}^w \sum_{j=1}^k n_{ij} \quad (5)$$

Statystyka χ^2 Pearsona do badania niezależności cech X , Y w tablicy dwudzielczej $w \times k$ ma postać (Pearson, 1900):

$$\chi_{XY}^2 = \sum_{i=1}^w \sum_{j=1}^k \frac{(n_{ij} - e_{ij})^2}{e_{ij}} \quad (6)$$

gdzie:

n_{ij} — liczebność zaobserwowana w i -tym wierszu i j -tej kolumnie,

e_{ij} — liczebność oczekiwana i -tego wiersza i j -tej kolumny dana wzorem:

$$e_{ij} = \frac{n_{i\bullet} \cdot n_{\bullet j}}{n} \quad (i = 1, 2, \dots, w; \quad j = 1, 2, \dots, k) \quad (7)$$

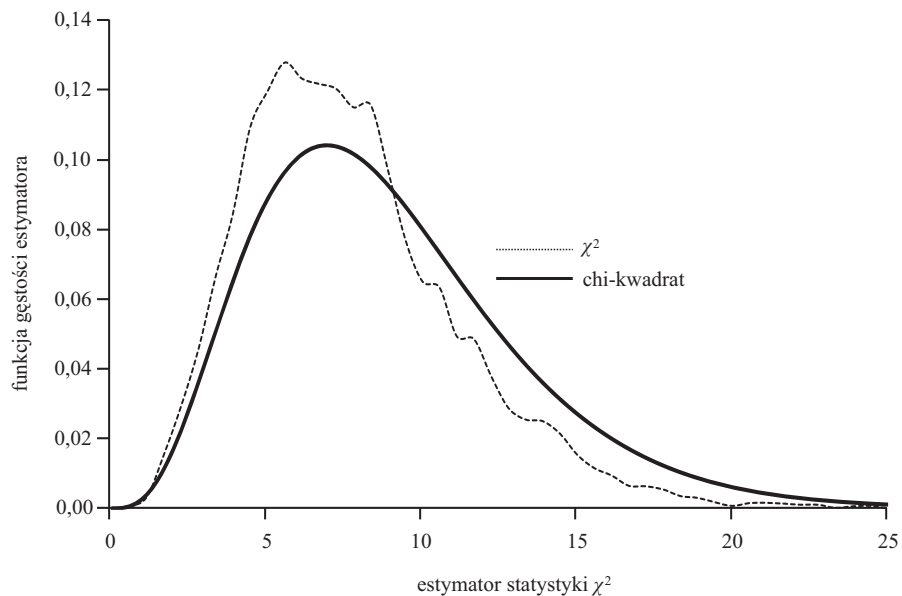
W przedstawionym teście niezależności liczebność próby musi być duża, aby w każdej komórce tablicy znalazła się możliwie duża liczba obserwacji. Zdaniem m.in. Yatesa, Moore'a, McCabe'a (Yates i in., 1999) oraz Shiera (Shier, 2004) wzór (6) należy stosować, gdy $e_{ij} \geq 1$ dla każdego $i = 1, \dots, w$; $j = 1, \dots, k$ oraz gdy nie więcej niż 20% liczebności oczekiwanej e_{ij} jest mniejsze niż 5. Zdaniem Cochran (Cochran, 1952) statystykę χ_{XY}^2 dla tablic większych niż 2×2 można stosować, gdy liczebność oczekiwana $e_{ij} > 5$ dla pewnego $i = 1, \dots, w$; $j = 1, \dots, k$.

Gdy hipoteza zerowa H_0 o niezależności cech X , Y jest słuszna, statystyka χ_{XY}^2 ma asymptotyczny rozkład chi-kwadrat z $df_1 = (w-1)(k-1)$ stopniami swobody i jest miarą rozbieżności między rozkładem zaobserwowanym i hipotetycznym. Wykr. 2 przedstawia rozkład chi-kwadrat z df_1 stopniami swobody oraz empiryczną funkcję gęstości rozkładu statystyki χ^2 uzyskaną za pomocą metody Parzena dla tablicy 4×4 i liczebności próby n , gdy H_0 jest słuszna. Dokładny opis tej metody znajduje się w opracowaniu Sulewskiego (2013). Liczebność próby w tym przypadku jest większa niż dla tablicy 2×2 , gdyż tym razem dysponujemy większą liczbą komórek. Liczebność próby dobrano w taki sposób, aby pokazać:

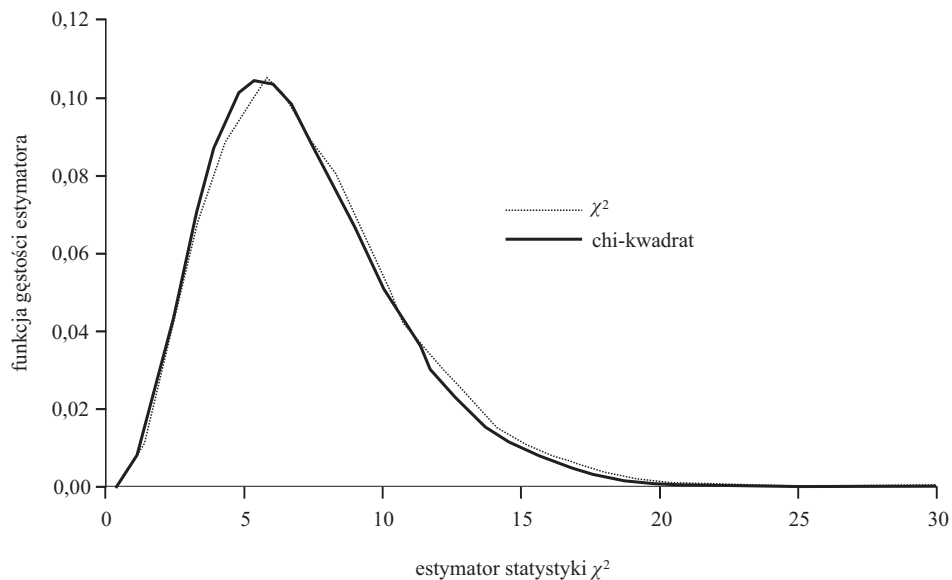
- brak zgodności rozkładu statystyki χ^2 z rozkładem chi-kwadrat dla tablicy 4×4 ($n = 160$);
- zgodność rozkładu statystyki χ^2 z rozkładem chi-kwadrat dla tablicy 4×4 ($n = 1000$).

Wykr. 2. ZGODNOŚĆ ROZKŁADU STATYSTYKI χ^2 PEARSONA Z ROZKŁADEM CHI-KWADRAT DLA TABLICY 4x4

4x4, n=160



4x4, n=1000



Źródło: jak w wykr. 1.

GENEROWANIE ZAWARTOŚCI TABLIC DWUDZIELCZYCH METODĄ SŁUPKOWĄ

W celu generowania zawartości tablicy dwudzielczej $w \times k$, gdy między cechami X , Y nie ma związku, przedział $(0; 1)$ podzielono na $w \cdot k$ podprzedziałów o szerokościach równych wartości prawdopodobieństwa $p_{ij} = 1/(w \cdot k) (i = 1, 2, \dots, w; j = 1, 2, \dots, k)$, które spełniają warunek normalizacji:

$$\sum_{i=1}^w \sum_{j=1}^k p_{ij} = 1 \quad (8)$$

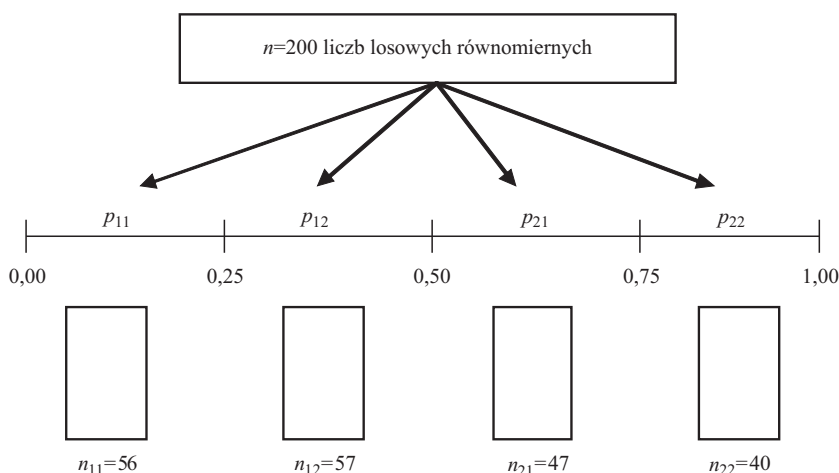
Każda z n wygenerowanych liczb losowych o rozkładzie równomiernym w przedziale $(0; 1)$ „wpada” do jednego z $w \cdot k$ podprzedziałów i tym samym liczba obiektów w odpowiadającej temu podprzedziałowi komórce tablicy zostaje zwiększona o jedną. Wielkości n_{ij} spełniające równość:

$$\sum_{i=1}^w \sum_{j=1}^k n_{ij} = n \quad (9)$$

są liczebnością obiektów w poszczególnych komórkach tablicy.

Na schemacie (1) przedstawiono wypełnianie komórek tablicy 2×2 dla liczebności próby $n = 200$, a tabl. 3 przedstawia odpowiadającą temu schematowi tablicę 2×2 .

SCHEMAT (1) WYPEŁNIANIA KOMÓREK TABLICY 2×2



Źródło: jak przy wykr. 1.

**TABL. 3. TABLICA DWUDZIELCZA 2×2
WYGENEROWANA ZA POMOCĄ METODY SŁUPKOWEJ**

Cecha X	Cecha Y		Razem
	Y_1	Y_2	
X_1	56	57	113
X_2	47	40	87
Razem	103	97	200

Źródło: jak przy tabl. 1.

WYZNACZANIE WARTOŚCI KRYTYCZNYCH DLA TABLIC DWUDZIELCZYCH

Wcześniej przedstawiono warunki, jakie muszą być spełnione, aby można było w testach niezależności dla tablic dwudzielczych stosować statystykę opisaną wzorami (1), (3), (4) i (6). W dobie coraz to szybszych komputerów można za pomocą stosownego oprogramowania znieść te ograniczenia i za pomocą symulacji wyznaczyć wartości krytyczne. Wyznaczono je tu odwołując się do oszacowań kwantyli, którymi są wartości stosownej statystyki pozycyjnej (David, 1970). W celu ustalenia wartości krytycznych metodą Monte Carlo dla danej tablicy dwudzielczej oraz liczebności próby, wyznaczono $R = 10^5$ wartości statystyki testowej (6) i za wartość krytyczną przyjęto $[(1 - \alpha)R]$ statystykę pozycyjną, gdzie α jest poziomem istotności oraz $[\cdot]$ oznacza część całkowitą liczby rzeczywistej. Tak duża liczba powtórzeń w obliczeniach statystyki testowej zapewnia uzyskanie dokładnego wyniku. Przy wyznaczaniu wartości krytycznych, gdy między cechami nie ma związku, zawartość tablic dwudzielczych generowano za pomocą metody słupkowej. Generowanie zawartości tych tablic innymi metodami, przy założeniu że między cechami nie ma związku, nie miało wpływu na uzyskane wielkości wartości krytycznych.

Przykład 1

Na podstawie danych przedstawionych w tablicach dwudzielczych 2×2 (tabl. 4) i 3×3 (tabl. 5), zbadano na poziomie istotności 0,05 istnienie związku między cechami X i Y korzystając ze statystyki χ^2 Pearsona.

**TABL. 4. TABLICA DWUDZIELCZA 2×2
Z ROZKŁADEM ŁĄCZNYM CECH X i Y**

Waga	Dieta		Razem
	stosowana	niestosowana	
Bez zmian.....	1	9	10
Utrata wagi	17	23	40
Razem	18	32	50

Źródło: jak przy tabl. 1.

TABL. 5. TABLICA DWUDZIELCZA 3×3 Z ROZKŁADEM ŁĄCZNYM CECH X I Y

Pochodzenie społeczne	Uczelnie wyższe			Razem
	uniwersytet	politechnika	akademia medyczna	
Chłopskie	18	7	2	27
Robotnicze	12	7	17	36
Inteligenckie	22	33	32	87
Razem	52	47	51	150

Źródło: jak przy tabl. 1.

ZESTAWIENIE (1) PROCEDUR I FUNKCJI WYKORZYSTANYCH DO PRZEPROWADZENIA TESTU NIEZALEŻNOŚCI DLA TABLICY DWUDZIELCZEJ

Wyszczególnienie	Typ	Realizowane zadania
Test2D	procedura	realizacja testu niezależności dla tablicy dwudzielczej
Rozmiar2	funkcja	ustalenie liczby kategorii cech X , Y
GNij	funkcja	pobranie liczebności tablicy dwudzielczej
ExpNij	funkcja	wyznaczanie liczebności oczekiwanej
SX2	funkcja	obliczanie sum brzegowych w wierszach
SY2	funkcja	obliczanie sum brzegowych w kolumnach
GenTab2	funkcja	generowanie zawartości tablicy dwudzielczej
Kwantyle2	funkcja	wyznaczanie wartości krytycznej
Sortu	funkcja	sortowanie
ChiKw2	funkcja	obliczanie wartości statystyki χ^2

Źródło: jak przy tabl. 1.

Implementację komputerową testu niezależności χ^2 Pearsona dla tablicy dwudzielczej utworzoną w edytorze VBA arkusza kalkulacyjnego Excel przedstawiono w pliku „TestyChi”, który umieszczono w Internecie pod adresem <http://www.utogim.eu/testychi.xlsm>. Składają się na nią procedury i funkcje przedstawione w zestawieniu 1. Po utworzeniu tablicy dwudzielczej w arkuszu „tablica2D” (tabl. 4) (począwszy od komórki B2) wprowadzono wartość poziomu istotności testu $\alpha = 0,05$ w arkuszu „wyniki” oraz uruchomiono procedurę „Test2D”. Efektem końcowym działania procedury jest tablica umieszczona w arkuszu „wyniki” ukazująca wyniki działania testu niezależności cech X i Y (tabl. 6).

TABL. 6. WYNIKI DZIAŁANIA TESTU NIEZALEŻNOŚCI CECH X I Y NA PODSTAWIE DANYCH Z TABL. 4

Hipoteza zerowa	Nie ma związku między cechami
Poziom istotności	0,05
Wartość krytyczna	3,848
Wartość statystyki	3,668

U w a g a. Nie ma podstaw do odrzucenia hipotezy zerowej.
Czas działania algorytmu 12 sekund.

Źródło: jak przy tabl. 1.

W uwadze (tabl. 6) podany jest czas działania algorytmu, na który główny wpływ ma sortowanie 10^5 wartości statystyki testowej (6).

Powtarzając te czynności dla tablicy dwudzielczej 3×3 (tabl. 5) otrzymano tablicę pokazującą wyniki działania testu niezależności cech X i Y (tabl. 7).

TABL. 7. WYNIKI DZIAŁANIA TESTU NIEZALEŻNOŚCI CECH X I Y NA PODSTAWIE DANYCH Z TABL. 5

Hipoteza zerowa	Nie ma związku między cechami
Poziom istotności	0,05
Wartość krytyczna	9,432
Wartość statystyki	20,951

U w a g a. Są podstawy do odrzucenia hipotezy zerowej. Czas działania algorytmu 25 sekund.

Ż r ó d ł o: jak przy tabl. 1.

STATYSTYKA χ^2 PEARSONA DLA TABLICY TRÓJDZIELCZEJ

Tablicę, która gromadzi wynik 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. 8 i 9).

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

TABL. 8. TABLICA TRÓJDZIELCZA $2 \times 2 \times 2$

Cecha X	Cecha Z				Razem
	Z_1		Z_2		
	cecha Y				
	Y_1	Y_2	Y_1	Y_2	
X_1	n_{111}	n_{121}	n_{112}	n_{122}	$n_{1..}$
X_2	n_{211}	n_{221}	n_{212}	n_{222}	$n_{2..}$
Razem	$n_{\cdot 11}$	$n_{\cdot 21}$	$n_{\cdot 12}$	$n_{\cdot 22}$	n

Ż r ó d ł o: jak przy tabl. 1.

Tabl. 9 przedstawia 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 .

TABL. 9. TABLICA TRÓJDZIELCZA $w \times k \times p$

Cecha X	Cecha Z									Razem
	Z_1				...	Z_p				
	cecha Y									
	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}	$n_{1..}$
X_2	n_{211}	n_{221}	...	n_{2k1}	...	n_{21p}	n_{22p}	...	n_{2kp}	$n_{2..}$
...	...	...	...	...	...	...	...	...	...	...
X_w	n_{w11}	n_{w21}	...	n_{wk1}	...	n_{w1p}	n_{w2p}	...	n_{wkp}	$n_{w..}$
Razem	$n_{\bullet 11}$	$n_{\bullet 21}$	...	$n_{\bullet k1}$	...	$n_{\bullet 1p}$	$n_{\bullet 2p}$	...	$n_{\bullet kp}$	$n_{\bullet ..}$

Ź r ó d ł o: jak przy tabl. 1.

Wartość n jest sumą wszystkich liczebności tablicy trójdzielczej, czyli:

$$n = \sum_{i=1}^w n_{i..} = \sum_{j=1}^k n_{\bullet j.} = \sum_{t=1}^p n_{\bullet \bullet t} = \sum_{i=1}^w \sum_{j=1}^k n_{ij.} = \sum_{i=1}^w \sum_{t=1}^p n_{i. t} = \sum_{j=1}^k \sum_{t=1}^p n_{\bullet j t} = \sum_{i=1}^w \sum_{j=1}^k \sum_{t=1}^p n_{ijt} \quad (10)$$

W analizie niezależności trzech cech można badać m.in. niezależność pełną, czyli czy cechy są parami niezależne. Statystyka χ^2 Pearsona do badania pełnej niezależności cech X, Y, Z w tablicy trójdzielczej $w \times k \times p$ ma postać:

$$\chi_{T3}^2 = \sum_{i=1}^w \sum_{j=1}^k \sum_{t=1}^p \frac{(n_{ijt} - e_{ijt})^2}{e_{ijt}} \quad (11)$$

gdzie:

n_{ijt} — liczebność zaobserwowana w i -tym wierszu, j -tej kolumnie, t -tej płaszczyźnie;

e_{ijt} — liczebność oczekiwana i -tego wiersza, j -tej kolumny, t -tej płaszczyzny dana wzorem:

$$e_{ijt} = \frac{n_{i..} \cdot n_{\bullet j.} \cdot n_{\bullet \bullet t}}{n^2} \quad (12)$$

Gdy hipoteza zerowa H_0 o niezależności cech X, Y, Z jest słuszna, statystyka χ^2_{T3} ma asymptotyczny rozkład chi-kwadrat z $df_2 = wkp - (w + k + p) + 2$ stopniami swobody i jest miarą rozbieżności między rozkładem zaobserwowanym i hipotetycznym. Wykr. 2 przedstawia rozkład chi-kwadrat z df_2 stopniami swobody oraz empiryczną funkcję gęstości rozkładu statystyki χ^2 uzyskaną za pomocą metody Parzena dla tablicy $2 \times 2 \times 2$ i liczebności próby n . Dokładny opis tej metody znajduje się w (Sulewski, 2013). Liczebność próby zależy oczywiście od liczby komórek tablicy i dobrano ją w taki sposób, aby pokazać:

- brak zgodności rozkładu statystyki χ^2 z rozkładem chi-kwadrat dla tablicy $2 \times 2 \times 2$ ($n = 80$);
- zgodność rozkładu statystyki χ^2 z rozkładem chi-kwadrat dla tablicy $2 \times 2 \times 2$ ($n = 1000$).

GENEROWANIE ZAWARTOŚCI TABLIC TRÓJDZIELCZYCH METODĄ SŁUPKOWĄ

W celu generowania zawartości tablicy $w \times k \times p$, gdy między cechami X, Y, Z nie ma związku, przedział $(0; 1)$ podzielono na $w \cdot k \cdot p$ podprzedziałów o szerokościach równych wartości prawdopodobieństwa $p_{ijt} = 1/(w \cdot k \cdot p)$ ($i = 1, 2, \dots, w; j = 1, 2, \dots, k; t = 1, 2, \dots, p$), które spełniają warunek normalizacji:

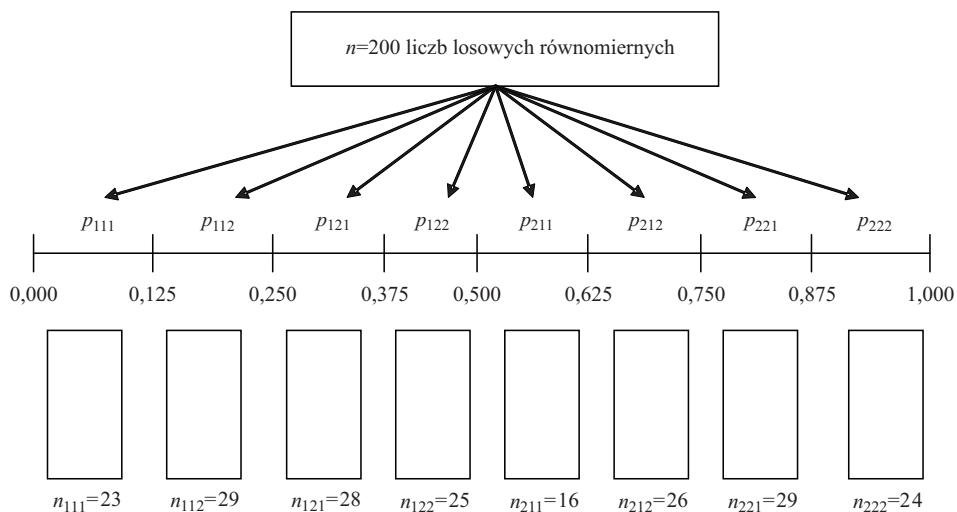
$$\sum_{i=1}^w \sum_{j=1}^k \sum_{t=1}^p p_{ijt} = 1 \quad (13)$$

Każda z n wygenerowanych liczb losowych o rozkładzie równomiernym w przedziale $(0; 1)$ „wpada” do jednego z $w \cdot k \cdot p$ podprzedziałów i tym samym liczba obiektów w odpowiadającej temu podprzedziałowi komórce tablicy zostaje zwiększona o jedną. Wielkości n_{ijt} spełniające równość:

$$\sum_{i=1}^w \sum_{j=1}^k \sum_{t=1}^p n_{ijt} = n \quad (14)$$

są liczebnością obiektów w poszczególnych komórkach tablicy.

SCHEMAT (2) WYPEŁNIANIA KOMÓREK TABLICZY $2 \times 2 \times 2$



Źródło: jak przy wykr. 1.

Schemat (2) przedstawia wypełnianie komórek tablicy $2 \times 2 \times 2$ dla liczebności próby $n = 200$, a tabl. 10 przedstawia odpowiadającą temu schematowi tablicę $2 \times 2 \times 2$.

TABL. 10. TABLICA TRÓJDZIELCZA $2 \times 2 \times 2$ WYGENEROWANA ZA POMOCĄ METODY SŁUPKOWEJ

Cecha X	Cecha Z				Razem
	Z_1		Z_2		
	cecha Y				
	Y_1	Y_2	Y_1	Y_2	
X_1	23	28	29	25	105
X_2	16	29	26	24	95
Razem	39	57	55	49	200

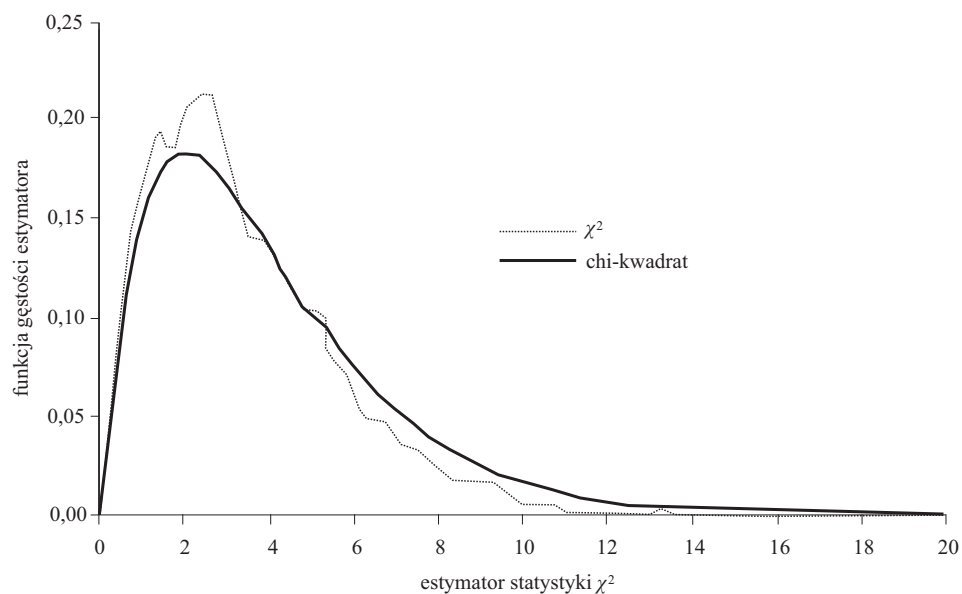
Źródło: jak przy tabl. 1.

WYZNACZANIE WARTOŚCI KRYTYCZNYCH DLA TABLIC TRÓJDZIELCZYCH

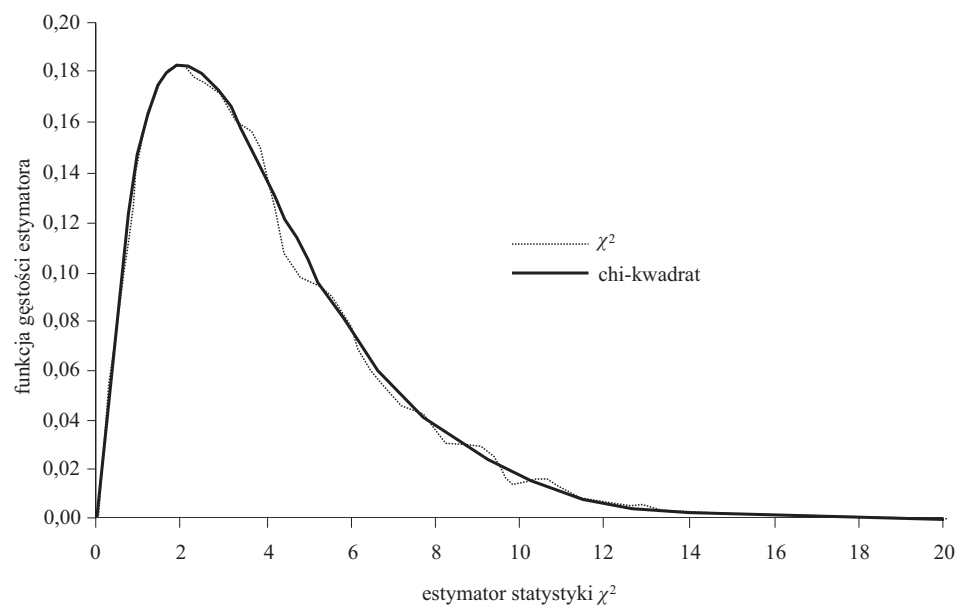
Z wykr. 3 wynika, że jeżeli próba nie jest dostatecznie duża, statystyka (11) nie ma rozkładu chi-kwadrat z df_2 stopniami swobody. Podobnie jak to miało miejsce w tablicach dwudzielczych można zrezygnować z rozkładu asymptotycznego i za pomocą symulacji wyznaczyć wartości krytyczne.

Wykr. 3. ZGODNOŚĆ ROZKŁADU STATYSTYKI χ^2 PEARSONA Z ROZKŁADEM CHI-KWADRAT DLA TABLICZY $2 \times 2 \times 2$

$2 \times 2 \times 2, n=80$



$2 \times 2 \times 2, n=1000$



Źródło: jak przy wykr. 1.

Wartości te wyznaczono odwołując się do oszacowań kwantyli, którymi są wartości stosownej statystyki pozycyjnej (David, 1970). W celu ustalenia wartości krytycznych metodą Monte Carlo dla danej tablicy trójdzielczej oraz liczebności próby wyznaczono $R = 10^5$ wartości statystyki testowej (11) i za wartość krytyczną przyjęto $[(1 - \alpha)R]$ statystykę pozycyjną, gdzie α jest poziomem istotności oraz $[\cdot]$ jest częścią całkowitą liczby. Przy wyznaczaniu wartości krytycznych, gdy między cechami nie ma związku, tablice trójdzielcze generowano za pomocą metody słupkowej.

Przykład 2

Na podstawie danych przedstawionych w tablicach trójdzielczych $2 \times 2 \times 2$ (tabl. 11) i $4 \times 3 \times 2$ (tabl. 12) zbadano na poziomie istotności 0,01 niezależność cech X, Y, Z korzystając ze statystyki χ^2 Pearsona.

TABL. 11. TABLICA TRÓJDZIELCZA $2 \times 2 \times 2$ Z ROZKŁADEM ŁĄCZNYM CECH X, Y, Z

Płeć	Miasta		Wieś		Razem
	grupy wiekowe mieszkańców				
	0—18	19—65	0—18	19—65	
Mężczyźni	3	15	11	33	62
Kobiety	10	16	2	10	38
Razem	13	31	13	43	100

Źródło: jak przy tabl. 1.

TABL. 12. TABLICA TRÓJDZIELCZA $4 \times 3 \times 2$ Z ROZKŁADEM ŁĄCZNYM CECH X, Y, Z

Grupy krwi	Mężczyźni			Kobiety			Razem
	wzrost w cm						
	160—170	171—180	181—190	160—170	171—180	181—190	
A	3	5	6	2	5	7	28
B	3	6	7	4	2	3	25
AB	4	6	1	9	6	5	31
0	0	2	4	1	2	7	16
Razem	10	19	18	16	15	22	100

Źródło: jak przy tabl. 1.

**ZESTAWIENIE (2) PROCEDUR I FUNKCJI WYKORZYSTANYCH DO PRZEPROWADZENIA
TESTU NIEZALEŻNOŚCI DLA TABLICY TRÓJDZIELCZEJ**

Wyszczególnienie	Typ	Realizowane zadanie
Test3D	procedura	realizacja testu niezależności dla tablicy trójdzielczej
Rozmiar3	funkcja	ustalenie liczby kategorii cech X, Y, Z
GNijt	funkcja	pobranie liczebności tablicy trójdzielczej
ExpNijt	funkcja	wyznaczanie liczebności oczekiwanej
SX3	funkcja	obliczanie sum brzegowych w wierszach
SY3	funkcja	obliczanie sum brzegowych w kolumnach
SZ3	funkcja	obliczanie sum brzegowych w płaszczyznach
GenTab3	funkcja	generowanie zawartości tablicy trójdzielczej
Kwantyle3	funkcja	wyznaczanie wartości krytycznej
Sortu	funkcja	sortowanie
ChiKw3	funkcja	obliczanie wartości statystyki χ^2

Źródło: jak przy tabl. 1.

Implementację komputerową testu niezależności χ^2 Pearsona dla tablicy trójdzielczej utworzoną w edytorze VBA arkusza kalkulacyjnego Excel przedstawiono w pliku „TestyChi”, który można znaleźć w Internecie pod adresem <http://www.utogim.eu/testychi.xlsm>. Składają się na nią procedury i funkcje przedstawione w zestawieniu 2. Po utworzeniu tablicy trójdzielczej w arkuszu „tablica3D” (Tabl. 9) (począwszy od komórki B2) wprowadzono wartość poziomu istotności testu $\alpha = 0,01$ w arkuszu „wyniki” oraz uruchomiono procedurę „Test3D”. Efektem końcowym działania procedury jest tablica umieszczona w arkuszu „wyniki” ukazująca wyniki działania testu niezależności cech X, Y, Z (tabl. 13).

**TABL. 13. WYNIKI DZIAŁANIA TESTU NIEZALEŻNOŚCI CECH X, Y, Z
NA PODSTAWIE DANYCH Z TABL. 11**

Hipoteza zerowa	Nie ma związku między cechami
Poziom istotności	0,01
Wartość krytyczna	13,217
Wartość statystyki	18,971

U w a g a. Są podstawy do odrzucenia hipotezy zerowej. Czas działania algorytmu 20 sekund.

Źródło: jak przy tabl. 1.

W uwadze (tabl. 13) podany jest czas działania algorytmu, na który główny wpływ ma sortowanie 10^5 wartości statystyki testowej (11).

Powtarzając czynności dla tablicy trójdzielczej $4 \times 3 \times 2$ (tabl. 12) otrzymano tablicę ukazującą wyniki działania testu niezależności cech X, Y, Z (tabl. 14).

**TABL. 14. WYNIKI DZIAŁANIA TESTU NIEZALEŻNOŚCI CECH X, Y, Z
NA PODSTAWIE DANYCH Z TABL. 12**

Hipoteza zerowa	Nie ma związku między cechami
Poziom istotności	0,05
Wartość krytyczna	32,984
Wartość statystyki	22,959

U w a g a. Nie ma podstaw do odrzucenia hipotezy zerowej.
Czas działania algorytmu 36 sekund.
Ż r ó d ł o: jak przy tabl. 1.

Podsumowanie

Wraz z rozwojem nauki pojawiają się w statystyce coraz to nowsze możliwości badania niezależności cech. Mimo to najważniejszą w tym kontekście jest nadal zaproponowana przez Pearsona w 1900 r. statystyka χ^2 , tym bardziej że ma ona swoje rozszerzenia także dla tablic trójdzielczych i wyższych. W przypadku tablic dwudzielczych $w \times k$ istnieją jednak ograniczenia co do stosowalności statystyki χ^2 , która ma asymptotyczny rozkład chi-kwadrat $(w-1)(k-1)$ stopniami swobody. Także dla tablic trójdzielczych $w \times k \times p$, jeżeli liczba realizacji w komórkach jest mała, rozkład statystyki testowej znacząco odbiega od rozkładu chi-kwadrat z $wkp - (w+k+p) + 2$ stopniami swobody. W obecnej dobie szybkiego rozwoju można znieść te ograniczenia wyznaczając wartości krytyczne dzięki symulacjom komputerowym generującym zawartość tablic wielodzielczych.

Lektura przedstawionej teorii dotyczącej tablic wielodzielczych oraz analiza zamieszczonych przykładów pozwoli wnikliwemu czytelnikowi przeprowadzić testy niezależności z wykorzystaniem statystyki χ^2 Pearsona przy dowolnej liczebności obiektów w poszczególnych komórkach tablicy wielodzielczej.

dr Piotr Sulewski — *Akademia Pomorska w Słupsku*

LITERATURA

- Campbell I. (2007), *Chi-squared and Fisher-Irwin tests of two-by-two tables with small sample recommendations*, „Statistics in Medicine”, Vol. 26
- Cochran W. G. (1952), *The χ^2 test of goodness of fit*, „Annals of Mathematical Statistics”, Vol. 23, No. 3

- David H. A. (1970), *Order statistics*, Wiley, New York
- Pearson E. S. (1947), *The choice of statistical tests illustrated on the interpretation of data classed in a 2×2 table*, „Biometrika”, Vol. 34
- Pearson K. (1900), *On the criterion that a given system of deviations from the probable in the case of a correlated system of variables is such that it can be reasonably supposed to have arisen from random sampling*, „Philosophy Magazine Series”, Series 5, Vol. 50
- Shier R. (2004), *The Chi-squared test for two-way tables*, Mathematics Learning Support Centre
- Sulewski P. (2013), *Modyfikacja testu niezależności*, „Wiadomości Statystyczne”, nr 10, GUS
- Yates F. (1934), *Contingency table involving small numbers and the χ^2 test*, „Journal of the Royal Statistical Society” (Supplement), No. 1
- Yates D., Moore D., McCabe G. (1999), *The Practice of Statistics* (1st Ed.), New York, W. H. Freeman

SUMMARY

In the study of the independence of characteristics in the multi-feature tables χ^2 Pearson's statistics are the most popular. For multi-feature arrays, there are certain limitations as to the applicability of the χ^2 Pearson's statistics, but in an era of rapidly developing computer setting can be abolished with the critical value of computer simulation to generate the contents of multi-feature tables. The aim of the study is to provide the reader with a ready computer implementation, written in VBA editor (Visual Basic for Applications). Reading the presented theory for multi-feature tables and analysis of the examples allow the reader to carry out independent tests using χ^2 Pearson's statistics at any number of objects in each multi-feature table cells.

РЕЗЮМЕ

В обследовании независимости признаков в многоразделительных таблицах самым популярным является статистика χ^2 Пирсона. Для многоразделительных таблиц существуют определенные ограничения в области возможностей использования статистики χ^2 Пирсона, но во время быстро развивающегося компьютерного оборудования можно их отменить определяя критические значения с помощью компьютерного моделирования генерируя содержание многоразделительных таблиц. Целью статьи является предоставление готовой компьютерной имплементации написанной в программе VBA (Visual Basic for Applications). Представленная теория касающаяся многоразделительных таблиц и анализ примеров позволят провести тесты независимости с использованием статистики χ^2 Пирсона для любого числа объектов в отдельных местах многоразделительной таблицы.