NR 4 (659)
KWIECIEŃ 2016

WIADOMOŚCI STATYSTYCZNE

CZASOPISMO GŁÓWNEGO URZĘDU STATYSTYCZNEGO
I POLSKIEGO TOWARZYSTWA STATYSTYCZNEGO

STUDIA METODOLOGICZNE

***Piotr SULEWSKI, Antoni DRAPELLA***

# Wpływ nierównomierności wypełnienia tablicy dwudzielczej na wartość krytyczną statystyki testowej

**Streszczenie.** *Artykuł dotyczy tablic dwudzielczych 2×2. Gdy hipoteza $H_0$ o niezależności cech jest słuszna, bardzo często — za sprawą małych próbek — rozkład statystyki testowej odbiega od rozkładu chi-kwadrat. Kwantyl rozkładu chi-kwadrat nie jest zatem właściwą wartością krytyczną. Problemem nie jest, przy obecnej wydajności komputerów, wyznaczenie metodą modelowania statystycznego Monte Carlo właściwej wartości krytycznej, lecz modelowanie $H_0$. Modelowanie $H_0$ to generowanie takich tablic, w których wartości cechy przypisane wierszom są niezależne od wartości cechy przypisanej kolumnom. Odpowiednie do takiego modelowania są tablice — równomierna o jednakowym prawdopodobieństwie przynależności do komórek oraz nierównomierna mająca jednakowe prawdopodobieństwo we wszystkich wierszach danej kolumny lub we wszystkich kolumnach danego wiersza. Analiza wyników modelowania statystycznego ujawniła, że nawet gdy $H_0$ jest słuszna, rozkład statystyki testowej w istotny sposób zależy od nierównomierności tablicy. W artykule pokazano, że chcąc maksymalizować moc testu należy wartość krytyczną ustalać z uwzględnieniem miary nierównomierności tablicy. Finalnym efektem opracowania jest zaproponowane czytelnikowi gotowe narzędzie do samodzielnej weryfikacji $H_0$.*

**Słowa kluczowe:** tablica dwudzielcza, test niezależności, wartości krytyczne, metoda Monte Carlo.

Test niezależności i test jednorodności są zapewne najczęściej, obok testu Kołmogorowa i testu Behrensa-Fishera, stosowanymi „narzędziami" statystycznymi. Dane do testów niezależności i jednorodności aranżowane są w postaci tablic dwudzielczych, a szczególnie tablic 2×2. Statystyka testowa ma asymptotyczny rozkład chi-kwadrat. Warto jednak wiedzieć, że we frazie tej kryje się nieprzyjemna prawda. W praktyce bowiem bardzo często rozkład statystyki testowej nie podlega rozkładowi chi-kwadrat, co wynika z małych próbek, jakimi dysponujemy. Rachunek prawdopodobieństwa nie oferuje metod pozwalających na wyznaczenie dokładnych rozkładów na podstawie analizy. Tak naprawdę znajomość postaci analitycznej tych rozkładów nie jest nam potrzebna. Jedyne, co chcemy znać to wartości kwantyli z tzw. „ogona" tych rozkładów, najczęściej 90% i 95%. Do ich uzyskania służy metoda modelowania komputerowego Monte Carlo. Wyznaczanie kwantyla na podstawie 50000 powtórzeń testu zgodności trwa tu nie dłużej niż kilka minut i można tego dokonać nawet za pomocą komputera o przeciętnej wydajności procesora, jakim jest Intel i3. W przypadku procesora i7, realizującego 8 wątków obliczeniowych jednocześnie, będzie to trwać znacznie krócej.

Interesuje nas tutaj rozkład statystyki testowej, gdy hipoteza $H_0$ („zerowa") o niezależności lub jednorodności cech jest słuszna. Przy obecnej wydajności komputerów problemu nie stanowi wyznaczenie kwantyla, lecz modelowanie $H_0$. Modelowanie $H_0$ to generowanie takich tablic, w których wartości cechy przypisanej wierszom są niezależne od wartości cech przypisanych kolumnom. Odpowiednia do takiego modelowania jest tablica równomierna, o jednakowym prawdopodobieństwie przynależności do komórek. Jednak w modelowaniu nie można ograniczyć się tylko do niej. Zjawisko niezależności może wystąpić w każdej tablicy nierównomiernej, gdy prawdopodobieństwo według kolumn jest jednakowe we wszystkich wierszach lub prawdopodobieństwo według wierszy jest jednakowe we wszystkich kolumnach. Zasada ta dotyczy wszystkich tablic dwudzielczych, w tym także tablic 2×2.

Analiza wyników modelowania ujawniła, że rozkład statystyki testowej w istotny sposób zależy od nierównomierności tablicy dwudzielczej 2×2. Zbieżność rozkładu statystyki testowej do rozkładu chi-kwadrat okazuje się tym wolniejsza, im bardziej nierównomierna jest tablica. Główne przesłanie tego artykułu mówi, że chcąc maksymalizować moc testu należy ustalić wartość krytyczną z uwzględnieniem nierównomierności tablicy.

Tablica dwudzielcza 2×2 posiada ograniczenia w zakresie stosowania statystyki $\chi^2$ Pearsona, która ma asymptotyczny rozkład chi-kwadrat z jednym stopniem swobody. W celu zniesienia tych ograniczeń w pracy Sulewskiego (2015a) zaproponowano wyznaczanie wartości krytycznych na podstawie symulacji komputerowych metodą Monte Carlo. Także Lilliefors w teście Kołmogorowa dla rozkładu normalnego wyznaczył wartości krytyczne za pomocą symulacji, gdy parametry rozkładu były oszacowane z próby.

Artykuł ten jest kontynuacją poprzednich rozważań (Sulewski, 2015a). Dalsze badania autorów prezentowane w artykule pokazują, że wartość krytyczna w teście niezależności dla tablicy dwudzielczej 2×2 (gdy między cechami nie ma związku) zależy nie tylko od liczebności próby i poziomu istotności, ale także od stopnia nierównomierności danych.

Pierwsza część naszego opracowania jest teoretyczna, a druga — praktyczna, gdzie zdefiniowano najważniejsze miary siły związku w tablicy dwudzielczej 2×2 oraz podano wyrażenie na wyznaczanie wartości stopnia nierównomierności, dla którego na poziomie istotności $\alpha = 0{,}05$ i dla liczebności próby $n \in \{15; 20; 25; 30; 50; 100\}$ wyznaczono wartości krytyczne metodą symulacyjną Monte Carlo.

W części praktycznej podano dwa przykłady testów niezależności w tablicy dwudzielczej 2×2 wraz z ich implementacją komputerową napisaną w edytorze VBA (*Visual Basic for Applications*) arkusza kalkulacyjnego Excel. W celu ułatwienia czytelnikowi samodzielnego prowadzenia badań statystycznych, kopię tej implementacji umieszczono w Internecie[1]. Liczne komentarze dotyczące kodów źródłowych będą pomocne dla osób, które zechcą poznać mechanizm tworzenia tej implementacji.

Podstawowym celem artykułu jest przedstawienie teorii dotyczącej testów niezależności dla tablicy dwudzielczej 2×2 oraz wprowadzenie miary nierównomierności, a kolejnym — zaproponowanie czytelnikowi gotowego narzędzia w postaci pliku do samodzielnego prowadzenia badań statystycznych umieszczonego w Internecie.

## *MIARY SIŁY ZWIĄZKU MIĘDZY CECHAMI W TABLICY DWUDZIELCZEJ 2×2*

Przypomnijmy, że najprostszą postacią tablicy dwudzielczej jest tablica 2×2 z liczebnością poszczególnych komórek $a$, $b$, $c$, $d$ (tabl. 1). Wiadomo, że liczebność próby to $n=a+b+c+d$.

**TABL. 1. TABLICA DWUDZIELCZA 2×2 LICZEBNOŚCI**

| Cecha $X$ | Cecha $Y$ | | Razem |
|---|---|---|---|
| | $Y_1$ | $Y_2$ | |
| $X_1$ ..... | $a = n_{11}$ | $b = n_{12}$ | $a + b = n_{1\bullet}$ |
| $X_2$ ..... | $c = n_{21}$ | $d = n_{22}$ | $c + d = n_{2\bullet}$ |
| Razem ..... | $a + c = n_{\bullet 1}$ | $b + d = n_{\bullet 2}$ | $n$ |

Ź r ó d ł o: opracowanie własne.

[1] http://www.utogim.eu/cvchi.xls.

Tablicę dwudzielczą 2×2 można także przedstawić wykorzystując prawdopodobieństwo $p_{ij}$ $(i, j = 1, 2)$, przy założeniu że $p_{11} + p_{12} + p_{21} + p_{22} = 1$ (tabl. 2). W odniesieniu do tabl. 1 wartości te wyznaczane są ze wzorów:

$$p_{11} = a/n \qquad p_{12} = b/n \qquad p_{21} = c/n \qquad p_{22} = d/n \tag{1}$$

**TABL. 2. TABLICA DWUDZIELCZA 2×2 PRAWDOPODOBIEŃSTWA**

| Cecha $X$ | Cecha $Y$ | | Razem |
|---|---|---|---|
| | $Y_1$ | $Y_2$ | |
| $X_1$ ..... | $p_{11}$ | $p_{12}$ | $p_{11} + p_{12} = p_{1\bullet}$ |
| $X_2$ ..... | $p_{21}$ | $p_{22}$ | $p_{21} + p_{22} = p_{2\bullet}$ |
| Razem ..... | $p_{11} + p_{21} = p_{\bullet 1}$ | $p_{12} + p_{22} = p_{\bullet 2}$ | 1 |

Ź r ó d ł o: jak przy tabl. 1.

Przypomnijmy, że statystyka $\chi^2$ Pearsona dla tablicy dwudzielczej 2×2 ma postać (Pearson, 1900):

$$\chi^2 = \frac{n(ad - bc)^2}{(a+b)(c+d)(a+c)(b+d)} = \\ = \frac{n(p_{11}p_{22} - p_{12}p_{21})^2}{(p_{11} + p_{12})(p_{21} + p_{22})(p_{11} + p_{21})(p_{12} + p_{22})} \tag{2}$$

W badaniu populacji generalnej istotne są zarówno zależność między cechami, jak i siła tej zależności. Wartości statystyki $\chi^2$ zależą nie tylko od siły związku między zmiennymi, lecz także od liczebności próby. Fakt ten sprawia, że statystyka ta jest mało przydatna do mierzenia siły związku między cechami.

Miary siły związku, zwane także współczynnikami zależności (korelacji), powinny przyjmować wartości z przedziału $\langle 0; 1 \rangle$, gdzie wartość 0 oznacza, że nie ma związku, natomiast wartość 1 pokazuje związek doskonały. Jeżeli możliwe jest określenie nie tylko siły związku, ale także jej kierunku, wtedy wartości współczynników wahają się od –1 do +1. W przypadku wartości –1 otrzymuje się doskonałą korelację ujemną, zaś dla wartości +1 — doskonałą korelację dodatnią.

Wzory na wyznaczanie miar siły związku w tablicy dwudzielczej *w*×*k* są dobrze znane. Interesującą formę mają ich szczególne postaci dla tablicy dwudzielczej 2×2 oraz miary siły związku stosowane wyłącznie w tablicy dwudzielczej

2×2. Najprostszym sposobem wyznaczania siły związku między cechami jest współczynnik frakcji określony wzorem:

$$rf = \frac{a}{a+c} - \frac{b}{b+d} = \frac{p_{11}}{p_{11}+p_{21}} - \frac{p_{12}}{p_{12}+p_{22}}$$
$$-1 \le rf \le 1 \tag{3}$$

Współczynniki *v* Cramera, *t* Czuprowa oraz $\varphi$ Yule'a dla tablicy dwudzielczej 2×2 mają taką samą postać, a mianowicie:

$$vt\varphi = \frac{|ad-bc|}{\sqrt{(a+b)(a+c)(b+d)(c+d)}}$$
$$0 \le vt\varphi \le 1 \tag{4}$$

lub równoważnie:

$$vt\varphi = \frac{|p_{11}p_{22} - p_{12}p_{21}|}{\sqrt{(p_{11}+p_{12})(p_{11}+p_{21})(p_{12}+p_{22})(p_{21}+p_{22})}} \tag{4a}$$

Miary siły związku stanowią także współczynnik *q* Kendalla oraz skorygowany współczynnik *c* Pearsona. Wielkości te dane są wzorami:

$$q = \frac{ad-bc}{ad+bc} = \frac{p_{11}p_{22} - p_{12}p_{21}}{p_{11}p_{22} + p_{12}p_{21}}$$
$$-1 \le q \le 1 \tag{5}$$

$$c = \sqrt{\frac{2\chi^2}{\chi^2+n}}$$
$$0 \le c \le 1 \tag{6}$$

Na szczególną uwagę zasługuje asymetryczna miara siły związku między cechami, a mianowicie współczynnik $\tau$ Goodmana-Kruskala, mający także swoje rozszerzenia dla tablic trójdzielczych (Gray, Williams, 1975) oraz dla tablic czterodzielczych (D'Ambra, Crisci, 2013). Jest to miara dla tablicy dwudzielczej *w*×*k* powszechnie znana w dwóch wariantach (Goodman, Kruskal, 1954):

- jeśli cecha wierszowa jest zależna:

$$\tau_k = \frac{n\sum_{i=1}^{w}\sum_{j=1}^{k} n_{ij}^2 / n_{\bullet j} - \sum_{i=1}^{w} n_{i\bullet}^2}{n^2 - \sum_{i=1}^{w} n_{i\bullet}^2} = \frac{\sum_{i=1}^{w}\sum_{j=1}^{k} p_{ij}^2 / p_{\bullet j} - \sum_{i=1}^{w} p_{i\bullet}^2}{1 - \sum_{i=1}^{w} p_{i\bullet}^2} \tag{7}$$

$$\tau_k \in \langle 0; 1 \rangle$$

- jeśli cecha kolumnowa jest zależna:

$$\tau_w = \frac{n\sum_{j=1}^{k}\sum_{i=1}^{w} n_{ij}^2 / n_{i\bullet} - \sum_{j=1}^{k} n_{\bullet j}^2}{n^2 - \sum_{j=1}^{k} n_{\bullet j}^2} = \frac{\sum_{i=1}^{w}\sum_{j=1}^{k} p_{ij}^2 / p_{i\bullet} - \sum_{i=1}^{w} p_{\bullet j}^2}{1 - \sum_{i=1}^{w} p_{\bullet j}^2} \tag{8}$$

$$\tau_w \in \langle 0; 1 \rangle$$

Dla tablicy dwudzielczej 2×2 współczynnik $\tau$ Goodmana-Kruskala ma dużo prostszą postać:

$$\tau = \tau_k = \tau_w = \frac{(ad - bc)^2}{(a+b)(a+c)(b+d)(c+d)} \tag{9}$$

$$0 \le \tau \le 1$$

$$\tau = \frac{(p_{11}p_{22} - p_{12}p_{21})^2}{(p_{11} + p_{12})(p_{11} + p_{21})(p_{12} + p_{22})(p_{21} + p_{22})} \tag{9a}$$

Implementację komputerową tych współczynników w środowisku VBA, w skład której wchodzi wyznaczanie wartości tych współczynników oraz badanie ich istotności, opisano w monografii Sulewskiego (2014).

*MIARA NIERÓWNOMIERNOŚCI W TABLICY DWUDZIELCZEJ 2×2*

Jeżeli hipoteza zerowa $H_0$ o niezależności cech $X$ i $Y$ jest słuszna, to wartości krytyczne dla tablicy dwudzielczej 2×2 wyznaczane są metodą Monte Carlo. W procesie tym tablice dwudzielcze 2×2 generowano metodą słupkową, korzystając z prawdopodobieństwa $p_{ij} = 0{,}25\ (i, j = 1, 2)$. Dla tego prawdopodobieństwa wymienione miary siły związku między cechami przyjmują wartość zero. Dokładny opis metody słupkowej można znaleźć w artykułach Sulewskiego (2015a, b).

Dalsze badania symulacyjne wykazały, że wartości krytyczne zależą od wartości prawdopodobieństwa $p_{ij}$ $(i, j = 1, 2)$, dla którego miary siły związku między cechami wynoszą zero ($H_0$ o niezależności cech jest słuszna). W związku z tym autorzy tego tekstu proponują miarę nierównomierności w postaci:

$$mn = \sum_{i=1}^{2}\sum_{j=1}^{2}\frac{(p_{ij} - 0{,}25)^2}{0{,}25} = 4\sum_{i=1}^{2}\sum_{j=1}^{2}(p_{ij} - 0{,}25)^2 \tag{10}$$

która przyjmuje wartości z przedziału $\langle 0, 1)$. Zestawienie (1) przedstawia sytuację, w której $mn \in \{0, 1\}$. Dla prawdopodobieństw $p_{ij}$ $(i, j = 1, 2)$ $mn = 1$, jednak w tym przypadku nie można policzyć wartości statystyki $\chi^2$ Pearsona. W wyrażeniu (10) widoczne jest pewne podobieństwo do statystyki $\chi^2$ Pearsona dla tablic dwudzielczych $w \times k$.

**ZESTAWIENIE (1) WARTOŚCI PRAWDOPODOBIEŃSTWA $p_{ij}$ ORAZ WARTOŚCI MIARY $mn$**

| $mn = 0$ | |
|---|---|
| $p_{11} = 1/4$ | $p_{12} = 1/4$ |
| $p_{21} = 1/4$ | $p_{22} = 1/4$ |

| $mn = 1$ | |
|---|---|
| $p_{11} = 1/2$ | $p_{12} = 0$ |
| $p_{21} = 1/2$ | $p_{22} = 0$ |

| $mn = 1$ | |
|---|---|
| $p_{11} = 1/2$ | $p_{12} = 1/2$ |
| $p_{21} = 0$ | $p_{22} = 0$ |

Ź r ó d ł o: jak przy tabl. 1.

## *WYZNACZANIE WARTOŚCI KRYTYCZNYCH DLA TABLICY DWUDZIELCZEJ 2×2*

W opracowaniu Sulewskiego (2015a) przedstawiono warunki, jakie muszą być spełnione, aby można było w testach niezależności dla tablicy dwudzielczej 2×2 stosować statystykę (2) oraz jej odmiany. W dobie coraz szybszych komputerów można za pomocą stosownego oprogramowania znieść te ograniczenia i drogą symulacyjną — stosując metodę Monte Carlo i uwzględniając nierównomierność danych — wyznaczyć wartości krytyczne. Algorytm wyznaczania wartości krytycznych dla tablicy dwudzielczej 2×2 jest następujący:

1) wyznaczenie liczebności próby *n* oraz miary nierównomierności *mn* danej wzorem (10) na podstawie danych źródłowych;
2) ustalenie hipotezy zerowej $H_0$, że nie ma związku między cechami;
3) ustalenie poziomu istotności $\alpha$;
4) dla danej wartości $mn \in \{0; 0{,}01; ...; 0{,}99\}$, gdy $H_0$ jest słuszna, ustalenie wartości prawdopodobieństwa $p_{ij}$ $(i, j = 2)$, dla którego miary siły związku są równe zeru;
5) generowanie tablicy dwudzielczej 2×2 metodą słupkową na podstawie wartości $p_{ij}$ $(i, j = 2)$;
6) wyznaczenie wartości statystyki $\chi^2$ Pearsona danej wzorem (2);

7) $R = 5 \cdot 10^4$-krotne powtórzenie pkt. 5 i pkt. 6;
8) uporządkowanie w kolejności rosnącej wartości statystyki $\chi_i^2 \ (i = 1, ..., R)$;
9) obliczenie wartości dystrybuanty empirycznej $F_i^* = i/(R+1)$;
10) ustalenie wartości krytycznej $cv1_\alpha$ jako *i*-tej statystyki pozycyjnej, w przypadku której wartość dystrybuanty empirycznej wynosi $F_i^* = 1 - \alpha$ lub jest bardzo bliska tej wartości;
11) $u = 50$-krotne powtórzenie pkt. 5—10;
12) wyznaczenie wartości krytycznej $cv_\alpha = \sum_{i=l}^{50} cv1_\alpha / 50$.

Wartości prawdopodobieństwa $p_{ij} \ (i, j = 1, 2)$ dla danej wartości miary nierównomierności *mn* wyznaczono korzystając z narzędzia do analizy symulacji, jakim jest solver[2], przy założeniach:

$$\sum_{i=1}^{2}\sum_{j=1}^{2} p_{ij} = 1 \qquad p_{ij} \geq 0 \ (i, j = 1, 2) \qquad p_{11} = p_{21} \qquad p_{12} = p_{22} \tag{11}$$

Dla miary nierównomierności $mn = 0{,}1 \cdot i \ \ (i = 0, 1, ..., 9)$, gdy hipoteza zerowa $H_0$ o niezależności cech *X* i *Y* jest słuszna oraz miary siły związku są równe zeru (zestawienie 2), wyznaczono wartości krytyczne dla liczebności próby $n \in \{15; 20; 25; 30; 50; 100\}$ na poziomie istotności $\alpha = 0{,}05$. Uzyskane wyniki przedstawiono na wykresie, z którego wynika, że dla danej wielkości próby wartość krytyczna zmienia się wraz ze wzrostem stopnia nierównomierności danych i przyjmuje najmniejsze wartości, gdy ta nierównomierność jest największa. Odchylenia od wartości wyznaczanej „rutynowo" z rozkładu chi-kwadrat są znaczne.

**ZESTAWIENIE (2) WARTOŚCI MIARY NIERÓWNOMIERNOŚCI** *mn* **ORAZ WARTOŚCI PRAWDOPODOBIEŃSTW** $p_{ij}$

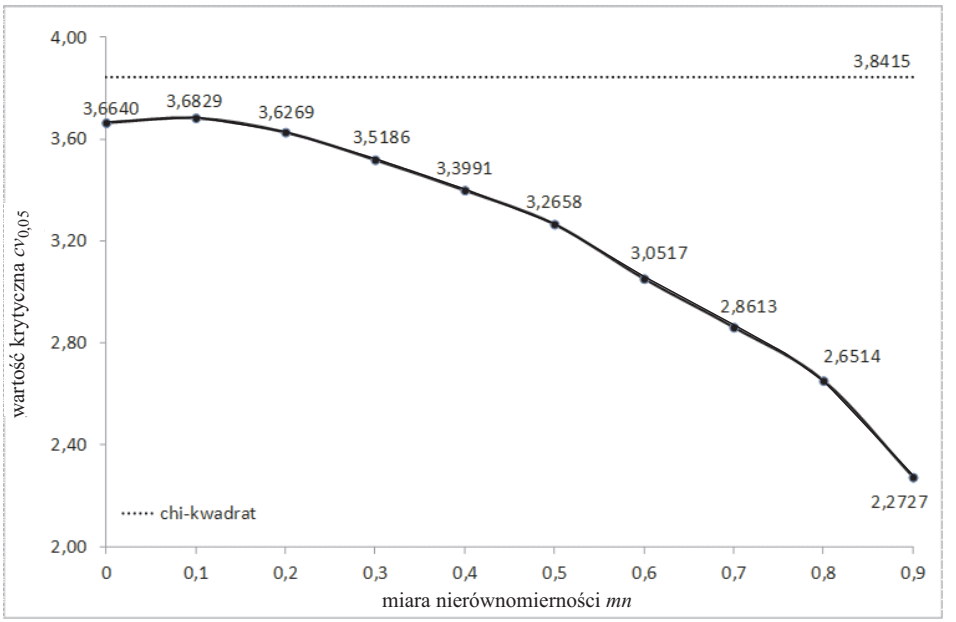
| $p_{11} = p_{21}$ lub $p_{11} = p_{12}$ | $p_{12} = p_{22}$ lub $p_{21} = p_{22}$ | *mn* |
|---|---|---|
| 0,250 | 0,250 | 0 |
| 0,171 | 0,329 | 0,1 |
| 0,138 | 0,362 | 0,2 |
| 0,113 | 0,387 | 0,3 |
| 0,092 | 0,408 | 0,4 |
| 0,073 | 0,427 | 0,5 |
| 0,056 | 0,444 | 0,6 |
| 0,041 | 0,459 | 0,7 |
| 0,026 | 0,474 | 0,8 |
| 0,013 | 0,487 | 0,9 |

Ź r ó d ł o: jak przy tabl. 1.

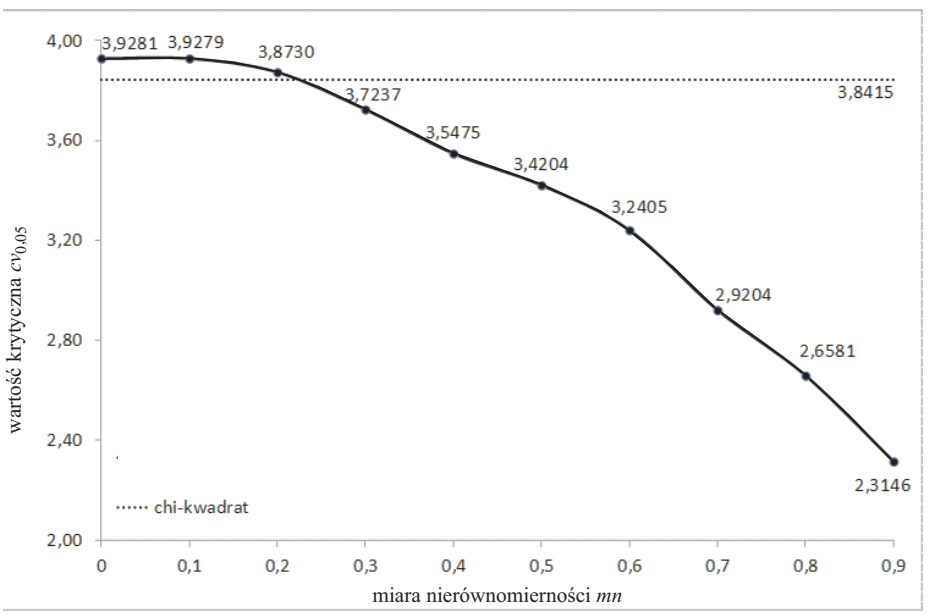
[2] Oprogramowanie ze stosownymi wskazówkami znajduje się w pliku internetowym w arkuszu „solver".

**WARTOŚCI KRYTYCZNE WEDŁUG LICZEBNOŚCI PRÓBY**

***n*=15**

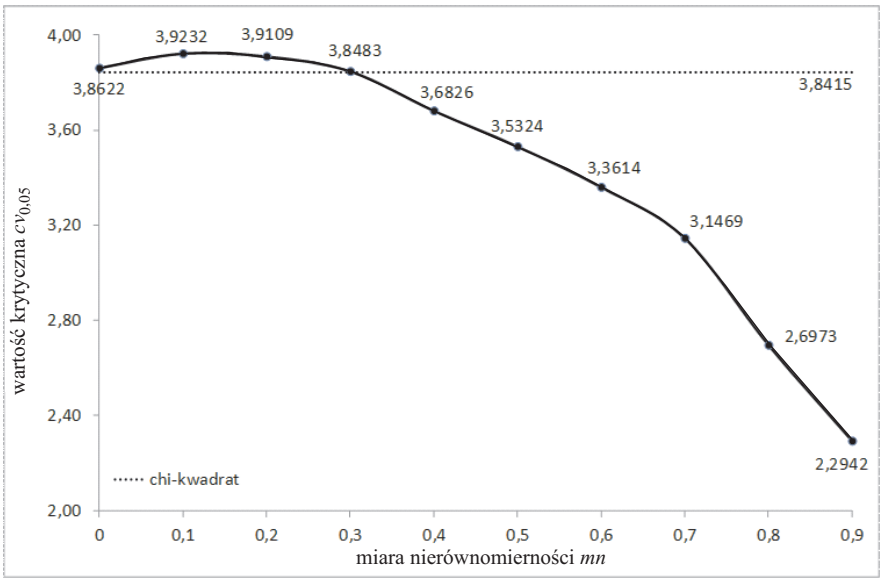



***n*=20**

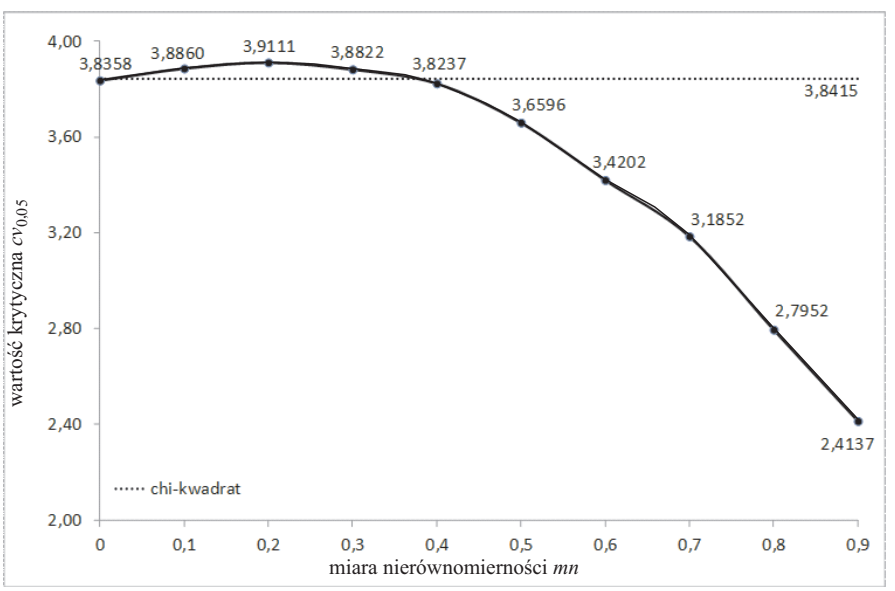


## WARTOŚCI KRYTYCZNE WEDŁUG LICZEBNOŚCI PRÓBY (cd.)

***n*=25**

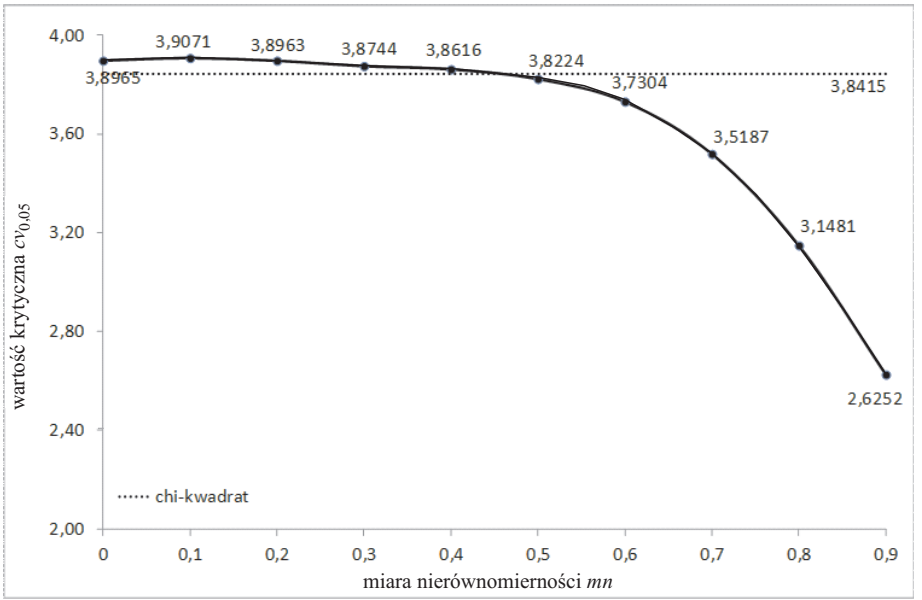



***n*=30**

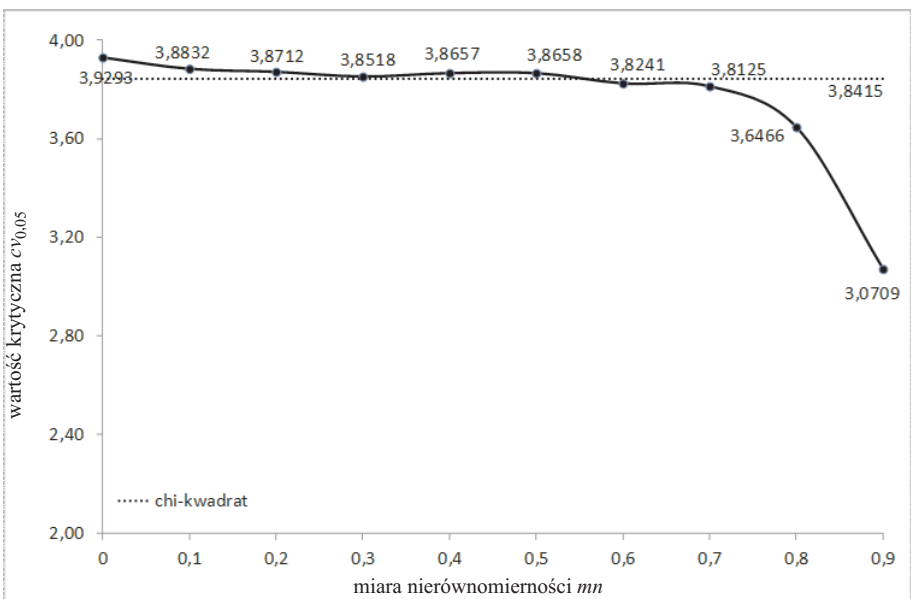


## WARTOŚCI KRYTYCZNE WEDŁUG LICZEBNOŚCI PRÓBY (dok.)

***n*=50**




***n*=100**




Ź r ó d ł o: opracowano na podstawie wyników przeprowadzonego badania.

## Przykład 1

W grupie 20 studentów matematyki przeprowadzono badanie ankietowe dotyczące wpływu stosowania diety na wagę ciała. Na podstawie danych przedstawionych w tablicy dwudzielczej 2×2 (tabl. 3) zbadano na poziomie istotności $\alpha = 0{,}05$ istnienie związku między cechami „waga ciała" i „dieta" korzystając ze statystyki $\chi^2$ Pearsona i wyznaczając wartość krytyczną testu na podstawie stopnia nierównomierności. Wyznaczono siłę związku wykorzystując opisane miary.

**TABL. 3. TABLICA DWUDZIELCZA 2×2 LICZEBNOŚCI Z ROZKŁADEM ŁĄCZNYM CECH „WAGA CIAŁA" I „DIETA"**

| Waga ciała | Dieta | | Razem |
|---|---|---|---|
| | stosowana | niestosowana | |
| **R a z e m** ..... | **8 (0,40)** | **12 (0,60)** | **20 (1,00)** |
| Bez zmian ..... | 1 (0,05) | 8 (0,40) | 9 (0,45) |
| Utrata wagi ..... | 7 (0,35) | 4 (0,20) | 11 (0,55) |

U w a g a. W nawiasach podano wartości prawdopodobieństw.

Ź r ó d ł o: jak przy tabl. 1.

Rozwiązanie. Implementację komputerową przykładu utworzoną w edytorze VBA arkusza kalkulacyjnego Excel przedstawiono w pliku[3], który można otworzyć w środowisku Microsoft Office począwszy od wersji 97-2003. W programie tym powinna być włączona opcja uruchamiająca wszystkie makra. Plik internetowy zawiera także arkusz w języku angielskim.

Autorzy artykułu podają tu algorytm według kolejnych kroków realizowania implementacji komputerowej. Podkreślono czynności, które należą do czytelnika — dotyczą one arkusza „tablica":

1) wypełnienie komórek A2, B2 nazwami odpowiednio cech *X* i *Y*;
2) wypełnienie komórek B4:B5 oraz C3:D3 wariantami odpowiednio cech *X* i *Y*;
3) wypełnienie komórek C4:D5 wielkościami *a*, *b*, *c*, *d*;
4) wyznaczenie sum brzegowych $(a+b)$, $(c+d)$, $(a+c)$, $(b+d)$;
5) wyznaczenie liczebności próby $n = a+b+c+d$;
6) wyznaczenie prawdopodobieństwa $p_{11}$, $p_{12}$, $p_{21}$, $p_{22}$;
7) założenie hipotezy zerowej $H_0$, że nie ma związku między cechami;
8) wprowadzenie wartości poziomu istotności $\alpha$ do komórki C9;
9) obliczenie wartości miary nierównomierności *mn* ze wzoru (10);
10) jeżeli $\alpha = 0{,}05$ i $mn = 0{,}1 \cdot i$ $(i = 0, 1, ..., 9)$ oraz $n \in \{15; 20; 25; 30; 50; 100\}$, to wartość krytyczna $cv_\alpha$ odczytana jest z wykresu;

[3] Plik ten umieszczono w Internecie pod adresem http://www.utogim.eu/cvchi.xls.

11) jeżeli $mn \neq 0{,}1 \cdot i \ \ (i = 0, 1, ..., 9)$, to wartość krytyczna $cv_\alpha$ wyznaczana jest symulacyjnie dla danej wartości miary nierównomierności *nm* zaokrąglonej do dwóch miejsc po przecinku;
12) wyznaczenie wartości statystyki $\chi^2$ ze wzoru (2);
13) jeżeli $\chi^2 < cv_\alpha$, to nie ma podstaw do odrzucenia $H_0$. Przejście do punktu 16;
14) jeżeli $\chi^2 \geq cv_\alpha$, to są podstawy do odrzucenia $H_0$;
15) wyznaczenie wartości wymienionych miar siły związku;
16) koniec algorytmu.

O działaniu algorytmu informuje licznik iteracji znajdujący się na pasku statusu.

Wizualizację przykładu 1 przedstawia zestawienie (3).

**ZESTAWIENIE (3) WYNIKÓW DZIAŁANIA TESTU NIEZALEŻNOŚCI DLA PRZYKŁADU 1**

**Segment A**

| $H_0$ | Nie ma związku między cechami |
|---|---|
| $\alpha$ | 0,05 |
| *mn* | 0,30 |
| $cv_\alpha$ | 3,72 |
| $\chi^2$ | 5,69 |
| Są podstawy do odrzucenia $H_0$ | |

**Segment B**

| Miary siły związku | Wartości siły związku |
|---|---|
| *rf* | –0,54 |
| *vtφ* | 0,53 |
| *q* | –0,87 |
| *c* | 0,44 |
| *τ* | 0,28 |

Ź r ó d ł o: jak przy tabl. 1.

## Przykład 2

W grupie 38 studentów informatyki przeprowadzono badanie ankietowe dotyczące szczepienia psów, z uwzględnieniem miejsca zamieszkania. Na podstawie danych przedstawionych w tablicy dwudzielczej 2×2 (tabl. 4) zbadano na poziomie istotności $\alpha = 0{,}1$ istnienie związku między cechami „miejsce zamieszkania" i „szczepienie psa" korzystając ze statystyki $\chi^2$ Pearsona i wyznaczając wartość krytyczną testu na podstawie stopnia nierównomierności. Wyznaczono, jak poprzednio, siłę związku wykorzystując wcześniej opisane miary.

**TABL. 4. TABLICA DWUDZIELCZA 2×2 Z ROZKŁADEM ŁĄCZNYM CECH „MIEJSCE ZAMIESZKANIA" I „SZCZEPIENIE PSA"**

| Miejsce zamieszkania | Szczepienie psa | | Razem |
|---|---|---|---|
| | tak | nie | |
| **R a z e m** | **10 (0,26)** | **28 (0,74)** | **38 (1,00)** |
| Miasta | 6 (0,16) | 22 (0,58) | 28 (0,74) |
| Wieś | 4 (0,10) | 6 (0,16) | 10 (0,26) |

U w a g a. W nawiasach podano wartości prawdopodobieństw.
Ź r ó d ł o: jak przy tabl. 1.

Rozwiązanie. Po wykonaniu algorytmu według kroków podanych w przykładzie 1, wizualizację tego przykładu przedstawiono w zestawieniu (4).

**ZESTAWIENIE (4) WYNIKÓW DZIAŁANIA TESTU NIEZALEŻNOŚCI DLA PRZYKŁADU 2**

**Segment A**

| $H_0$ | Nie ma związku między cechami |
|---|---|
| $\alpha$ | 0,10 |
| $mn$ | 0,58 |
| $cv_\alpha$ | 2,75 |
| $\chi^2$ | 1,31 |
| Nie ma podstaw do odrzucenia $H_0$ | |

**Segment B**

| Miary siły związku | Wartości siły związku |
|---|---|
| $rf$ | –0,19 |
| $vt\varphi$ | 0,19 |
| $q$ | –0,42 |
| $c$ | 0,07 |
| $\tau$ | 0,03 |

Ź r ó d ł o: jak przy tabl. 1.

## Podsumowanie

W badaniu niezależności cech w tablicach dwudzielczych najważniejsza jest zaproponowana przez Pearsona statystyka $\chi^2$. W celu zniesienia ograniczeń w stosowaniu tej statystyki wartości krytyczne wyznaczono symulacyjnie. W opracowaniu zwrócono uwagę na fakt, że wartości krytyczne wyznaczone symulacyjnie metodą Monte Carlo zależą także od stopnia nierównomierności danych. Zbieżność rozkładu statystyki testowej do rozkładu chi-kwadrat jest tym wolniejsza, im bardziej nierównomierna jest tablica. Główne przesłanie tego artykułu mówi, że chcąc maksymalizować moc testu należy wartość krytyczną ustalać z uwzględnieniem nierównomierności tablicy. Dzięki gotowej implementacji komputerowej czytelnik może samodzielnie badać niezależność cech w tablicy dwudzielczej 2×2.

**dr Piotr Sulewski** — *Akademia Pomorska w Słupsku*, **prof. dr hab. Antoni Drapella** — *Akademia Marynarki Wojennej w Gdyni*

**LITERATURA**

D'Ambra A., Crisci A. (2013), *Multiple TAU decomposition in mean effect and interaction term*, SIS Statistical Conference, Advances in Latent Variables. Methods, Models and Applications, Brescia.

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*, [w:] *Proceedings of the Social Statistics Section*, American Statistical Association.

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 reasonbly supposed to have arisen from random sampling*, „Philosophy Magazine Series", Series 5, Vol. 50.

Sulewski P. (2014), *Statystyczne badanie współzależności cech typu dyskretne kategorie*, Akademia Pomorska, Słupsk.

Sulewski P. (2015a), *Wyznaczanie obszaru krytycznego przy testowaniu niezależności w tablicach wielodzielczych*, „Wiadomości Statystyczne", nr 3.

Sulewski P. (2015b), *Ocena zdolności tablic dwudzielczych do wykrywania związku między uporządkowanymi cechami typu jakościowego*, „Wiadomości Statystyczne", nr 5.

**Summary.** *The article concerns two-way (2×2) contingency tables. When the $H_0$ — hypothesis of independence of features is correct, very often — because of the small sample — the distribution of the test statistics differ from the chi-square. Quantile of the chi-square is therefore not a correct critical value. With the current performance of computers, designation of critical value by statistical modeling of Monte Carlo method is not a problem, but a problem is $H_0$ modeling. The $H_0$ modeling is generating such arrays, which feature value assigned rows are independent of the characteristics of the assigned columns. Suitable for such modeling are tables — uniform of the same probability of belonging to cells and uneven having equal probability in all rows of a given column or in all columns of a given row. Analysis of the results of statistical modeling revealed that even when $H_0$ is right, the distribution of the test statistics significantly depends on the uneven array. The article shows that in order to maximize the power of the test should be set critical value, taking into account measures of inequality array. The final result of the study is offered the reader a ready tool for independent verification of the $H_0$ hypothesis.*

**Keywords:** two-way contingency tables, independence test, critical values, Monte Carlo method.

**Резюме.** *Статья рассматривает двухразделительные таблицы 2×2. Если гипотеза $H_0$ по независимости признаков является правильной, очень часто — с использованием небольших выборок – распределение тестовых статистик не подчиняется распределению хи-квадрат. Квантиль распределения хи-квадрат таким образом не является соответствующим критическим значением. Учитывая производительность сегодняшних компьютеров, проблемой не является обозначение по методу статистического моделирования Монте-Карло соответствующего критического значения, но моделирование $H_0$. Моделирование $H_0$ это разработка таких таблиц, в которых значения признака отнесены к строкам являются независимыми от величины признака отнесенного к столбцам. Соответствующими для такого моделирования таблицами являются — равномерная с одинаковой вероятностью принадлежности к клеткам, а также неравномерная имеющая одинаковую вероятность во всех строках данного столбца или во всех столбцах данной строки. Анализ результатов*

*статистического моделирования показал, что даже если $H_0$ является правильной, распределение тестовых статистик действительно зависит от неравномерности таблицы. В статье было показано, что для высокой мощности критерия следует определять критическое значение с учетом меры неравномерности таблицы. Конечным эффектом разработки является предложение читателю готового инструмента для самостоятельной проверки $H_0$.*

**Ключевые слова:** двухразделительная таблица, критерий независимости, критическое значение, Монте-Карло.