



Piotr Sulewski

Akademia Pomorska w Słupsku
Wydział Matematyczno-Przyrodniczy
Instytut Matematyki
piotr.sulewski@apsl.edu.pl

GENEROWANIE TABLIC DWUDZIELCZYCH Z WYKORZYSTANIEM DWUWYMIAROWEGO ROZKŁADU NORMALNEGO UCIĘTEGO

Streszczenie: W artykule zaprezentowano procedurę generowania liczb losowych o dwuwymiarowym rozkładzie normalnym uciętym (DRNU). Opisano algorytm generowania tablicy dwudzielczej, wykorzystując próbę realizacji zmiennej losowej DRNU. Przedstawioną teorię wzbogacono konkretnymi przykładami, zrealizowanymi w edytorze VBA arkusza kalkulacyjnego Microsoft Excel.

Słowa kluczowe: tablica dwudzielcza, metoda Monte Carlo, test niezależności, dwuwymiarowy rozkład normalny ucięty.

Wprowadzenie

Tablicę dwudzielczą (TD) $w \times k$ zalicza się do podstawowych, często stosowanych narzędzi statystycznych. Wykorzystuje się ją do badania mocy testów, czyli prawdopodobieństwa odrzucenia H_0 , gdy jest ona fałszywa. Do tego rodzaju badań niezbędne jest generowanie TD i określenie mocy testów za pomocą symulacji Monte Carlo. Terminem „generowanie TD” określono wypełnienie komórek TD liczbami pseudolosowymi.

Gdy związku między badanymi cechami nie ma, generowanie TD jest rzeczą prostą. Można skorzystać z generatorów liczb o rozkładzie równomiernym i losowo generować numer wiersza oraz kolumny tablicy, określając w ten sposób przynależność danej realizacji do komórki. Zadaniem znacznie trudniejszym jest generowanie TD, gdy związek między cechami istnieje.

Przystępując do badania właściwości testów statystycznych, trzeba dysponować narzędziem do modelowania populacji generalnej, czyli do nadawania

populacji generalnej określonej właściwości, jaką jest związek między cechami. W związku z czym dane podlegające opracowaniu, muszą być danymi pochodzącymi z generatora liczb losowych, a nie danymi wziętymi z praktyki. Należy wiedzieć, jaki jest rzeczywisty związek między cechami, inaczej zbadanie mocy testów będzie niemożliwe. W celu wyznaczenia empirycznej mocy testu niezbędne jest generowanie TD.

Do generowania TD można skorzystać z: łańcuchów Markowa (ang. *Markov Chain Monte Carlo*) [Diaconis, Sturmfels, 1998; Cryan, Dyer, 2003; Chen i in., 2005; Cryan i in., 2006; Fishman, 2012], sekwencyjnego próbkowania (ang. *Sequential Importance Sampling*) [Chen i in., 2005; Chen, Dinwoodie, Sullivant, 2006; Blitzstein, Diaconis, 2011; Yoshida i in., 2011], metody bootstrapowej (ang. *bootstrap method*) [Nandram, Bhatta, Bhadra, 2013], techniki „dziel i zwyciężaj” (ang. *probabilistic divide-and-conquer technique*) [Desalvo, Zhao, 2016]. Propozycję generowania TD dla ustalonych rozkładów brzegowych zaprezentowali R.B. Holmes i L.K. Jones [1996]. Procedurę generowania TD za pomocą metody słupkowej, wykorzystującą liczby losowe o rozkładzie równomiernym zaproponował w swojej pracy P. Sulewski [2014a]. Uogólniony rozkład *gamma* wykorzystano do generowania TD 2×2 [Sulewski, 2009] oraz TD $2 \times k$ [Sulewski, 2014b].

Przy testowaniu niezależności cech, najbardziej popularną statystyką testową dla TD jest statystyka χ^2 Pearsona, która ma także swoje rozszerzenia dla tablic trójdzielnych i wyższych. Dla TD $w \times k$ istnieją jednak ograniczenia stosowalności statystyki χ^2 Pearsona, która ma asymptotyczny rozkład chi-kwadrat z $(w-1)(k-1)$ stopniami swobody. W celu zniesienia tych ograniczeń wartości krytyczne można wyznaczać za pomocą symulacji komputerowych metodą Monte Carlo [Sulewski, 2015], a do tego niezbędne jest generowanie TD. Także Lilliefors w teście Kołmogorowa dla rozkładu normalnego wyznaczał wartości krytyczne drogą symulacyjną, gdy parametry rozkładu były oszacowane z próby.

Procedurę generowania TD z wykorzystaniem dwuwymiarowego rozkładu normalnego (DRN) oraz opis generatora liczb losowych o DRN przedstawiono w pracy P. Sulewskiego [2014a]. Metoda ta wykorzystywana dla TD $w \times k$, $(w, k > 2)$ posiada pewną wadę. Gdy się ją stosuje, w narożach tablicy pojawiają się komórki puste i w tej sytuacji nie można skorzystać, np. z niektórych statystyk chi-kwadrat (ang. *power divergence statistics*): statystyki G2 ilarazu wiarygodności, statystyki KL Kullbacka-Leiblera czy statystyki FT Free-mana-Tukey’a. Z tego też powodu w niniejszej pracy zaproponowano modyfikację wspomnianej metody, polegającą na zastosowaniu dwuwymiarowego rozkładu normalnego uciętego (DRNU). Powodem wyboru DRN do gene-

rowania TD była nie tylko jego wielka popularność, ale także to, że K. Pearson [Pearson, 1904] skorzystał z tego rozkładu w celu zdefiniowania współczynnika kontyngencji.

Celem artykułu jest zaproponowanie narzędzia do generowania TD o ustalonym z góry związku między cechami. Przedstawiono w nim funkcję gęstości dwuwymiarowego rozkładu normalnego zarówno pełnego, jak i uciętego. Wyznaczono wartość współczynnika korelacji DRNU. Zaprezentowano procedurę generowania liczb losowych o DRNU. Opisano generowanie TD $w \times k$ ($w, k > 2$), wykorzystując n elementową próbę realizacji zmiennej losowej DRNU. Przedstawioną teorię wzbogacono przykładami. Implementację komputerową wykonano w edytorze VBA arkusza kalkulacyjnego Excel.

1. Funkcja gęstości dwuwymiarowego rozkładu normalnego

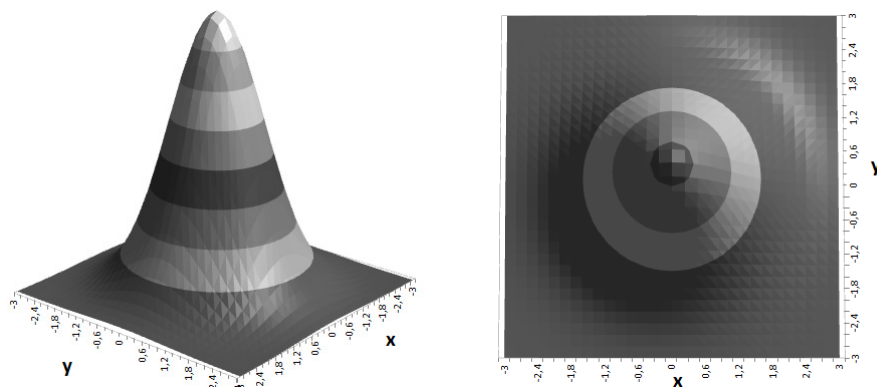
Zmienna losowa (X, Y) ma DRN, jeżeli jej funkcja gęstości dana jest wzorem [Kendall, Buckland, 1986]:

$$f(x, y) = \frac{1}{2\pi\sigma_1\sigma_2\sqrt{1-r^2}} \exp(-Q) \quad |r| \leq 1$$

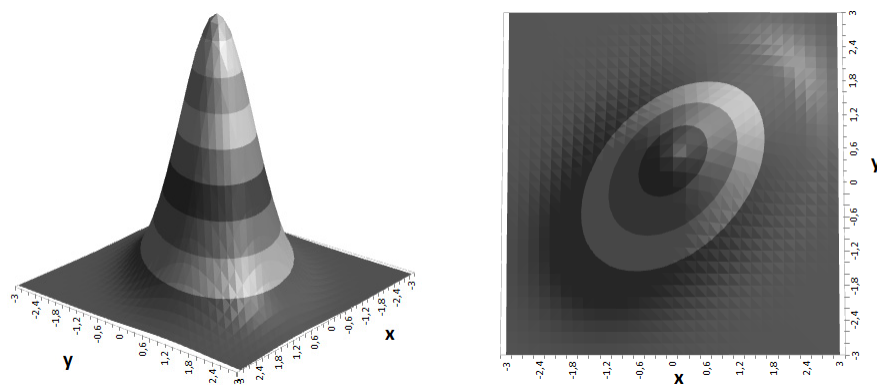
$$Q = \frac{1}{2(1-r^2)} \left[\left(\frac{x-\mu_1}{\sigma_1} \right)^2 - 2r \frac{x-\mu_1}{\sigma_1} \frac{y-\mu_2}{\sigma_2} + \left(\frac{y-\mu_2}{\sigma_2} \right)^2 \right] \quad (1)$$

gdzie r jest współczynnikiem korelacji liniowej Pearsona. Rozkład brzegowy zmiennej losowej X jest rozkładem normalnym $N(\mu_1, \sigma_1)$, zaś rozkład brzegowy zmiennej losowej Y – rozkładem normalnym $N(\mu_2, \sigma_2)$.

Rysunek 1 przedstawia wykres funkcji gęstości DRN, gdy zmienne X oraz Y są nieskorelowane ($r = 0$), natomiast rys. 2 – gdy zmienne X i Y są skorelowane ($r = 0,4$). Wartości pozostałych parametrów tego rozkładu to $\mu_1 = \mu_2 = 0$, $\sigma_1 = \sigma_2 = 1$.



Rys. 1. Wykres funkcji gęstości dwuwymiarowego rozkładu normalnego dla $\mu_1 = \mu_2 = 0$, $\sigma_1 = \sigma_2 = 1$, $r = 0$



Rys. 2. Wykres funkcji gęstości dwuwymiarowego rozkładu normalnego dla $\mu_1 = \mu_2 = 0$, $\sigma_1 = \sigma_2 = 1$, $r = 0,4$

Niech $f(x, y)$ będzie funkcją gęstości (1) dwuwymiarowej zmiennej losowej normalnej (X, Y) . Wówczas funkcja gęstości zmiennej losowej (X, Y) uciętej do obszaru $\langle a, b \rangle \times \langle c, d \rangle$ ma postać [Cramer, 1958]:

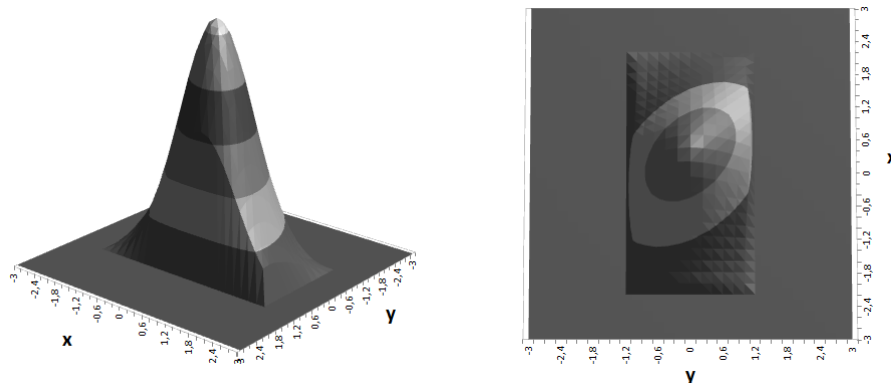
$$f_u(x, y, a, b, c, d) = \frac{f(x, y)}{\int_a^b \int_c^d f(x, y) dx dy} \quad (2)$$

dla wszystkich $(x, y) \in \langle a, b \rangle \times \langle c, d \rangle$ i jest równa zero poza tym obszarem. Wartości a, b, c, d mogą być także nieskończone.

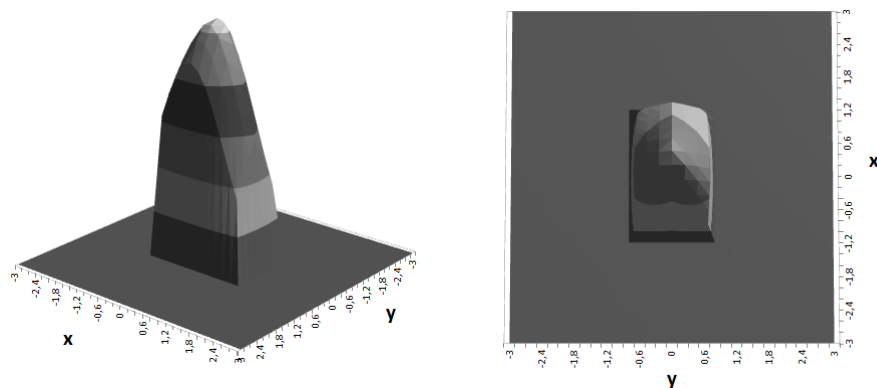
Na rys. 3 i 4 przedstawiono wykresy funkcji gęstości DRNU dla zestawów A i B wartości parametrów (tab. 1).

Tabela 1. Zestawy wartości parametrów dwuwymiarowego rozkładu normalnego uciętego

Zestaw	μ_1	σ_1	μ_2	σ_2	r	a	b	c	d
A	0	1	0	1	0,4	-2	2	-1	1
B	0	1	0	1	0,4	-1	1	-0,5	0,5



Rys. 3. Wykres funkcji gęstości dwuwymiarowego rozkładu normalnego uciętego dla zestawu A wartości parametrów



Rys. 4. Wykres funkcji gęstości dwuwymiarowego rozkładu normalnego uciętego dla zestawu B wartości parametrów

2. Współczynnik korelacji dwuwymiarowego rozkładu normalnego uciętego

We wzorze (1) r jest współczynnikiem korelacji liniowej Pearsona, gdyż [Cramer, 1958]:

$$\frac{\text{cov}(X, Y)}{\sqrt{V(X) \cdot V(Y)}} = r$$

Jeżeli $r = 0$, to funkcja gęstości DRNU ma postać $f(x, y) = f(x) \cdot f(y)$, czyli zmienne losowe X i Y są niezależne.

Badając testy niezależności, należy dysponować narzędziem do nadawania populacji generalnej określonej właściwości, jaką jest związek między cechami. Związek ten można różnicować za pomocą współczynnika korelacji, który dla DRNU wyraża się wzorem [Cramer, 1958]:

$$r_u = \frac{\alpha_{11} - \alpha_{10} \cdot \alpha_{01}}{\sqrt{\alpha_{20} - \alpha_{10}^2} \cdot \sqrt{\alpha_{02} - \alpha_{01}^2}} \quad (3)$$

gdzie

$$\alpha_{11} = \frac{\int_a^b \int_c^d xy f(x, y) dx dy}{\int_a^b \int_c^d f(x, y) dx dy}, \quad \alpha_{k0} = \frac{\int_a^b x^k f(x) dx}{\int_a^b f(x) dx}, \quad \alpha_{0k} = \frac{\int_c^d y^k f(y) dy}{\int_c^d f(y) dy} \quad (k = 1, 2)$$

Dla zestawów wartości parametrów (tab. 2) obliczono na podstawie (3) wartość współczynnika korelacji DRNU. Uzyskane wyniki przedstawiono w tab. 3.

Tabela 2. Zestawy wartości parametrów dwuwymiarowego rozkładu normalnego uciętego

Zestaw	μ_1	σ_1	μ_2	σ_2	r	a	b	c	d
C	0	1	0	1	0,5	-2	2	-2	2
D	0	2	0	2	0,5	-2	2	-2	2
E	0	2	0	2	0,5	-1,5	1,5	-1,5	1,5
F	0	2	0	2	0,7	-1,5	1,5	-1,5	1,5

Tabela 3. Wartości współczynnika korelacji rozkładu uciętego dla różnych zestawów wartości parametrów dwuwymiarowego rozkładu normalnego uciętego

Zestaw	r_u
C	0,407
D	0,178
E	0,110
F	0,207

Z tab. 2 i 3 wynika, że wraz ze wzrostem wartości parametrów σ_1 i σ_2 (zestaw C i D) przy stałej wartości pozostałych parametrów, korelacja rozkładu maleje. Zwiększając cięcie, czyli zmniejszając pole obszaru $\langle a, b \rangle \times \langle c, d \rangle$ (zestaw D i E), korelacja rozkładu maleje. Oczywiście wraz ze wzrostem korelacji rozkładu pełnego (zestaw E i F), rośnie także korelacja rozkładu uciętego.

3. Generowanie liczb losowych o dwuwymiarowym rozkładzie normalnym uciętym

Do generowania liczb losowych o DRNU niezbędne są liczby losowe o rozkładzie normalnym $N(0;1)$. Na podstawie centralnego twierdzenia granicznego zmienna losowa [Zieliński, Wieczorkowski, 1997]:

$$X = \sum_{i=1}^{12} R_i - 6$$

ma w przybliżeniu $N(0;1)$, gdzie R jest niezależną zmienną losową o rozkładzie równomiernym na przedziale $(0;1)$. Zauważmy, że $E(X) = 0$, $V(X) = 1$. Bardziej współczesnym sposobem jest generowanie liczb losowych o rozkładzie normalnym z wykorzystaniem transformacji Boxa-Mullera [Box, Muller, 1958]. Jeżeli R_1, R_2 są niezależnymi zmiennymi losowymi o rozkładzie równomiernym na przedziale $(0;1)$, to:

$$U = \sqrt{-2\ln(R_1)}\cos(2\pi R_2), V = \sqrt{-2\ln(R_1)}\sin(2\pi R_2) \quad (4)$$

są niezależne i mają $N(0,1)$. Jeżeli zmienna losowa X ma rozkład $N(0,1)$, to zmienna losowa $X \cdot s + m$ ma rozkład $N(m, s)$.

W celu otrzymania realizacji dwuwymiarowej zmiennej losowej normalnej (X, Y) o gęstości (1) należy skorzystać ze wzorów [Zieliński, Wieczorkowski, 1997]:

$$X = \sigma_1 U + \mu_1, Y = r\sigma_2 U + \sigma_2 V\sqrt{1-r^2} + \mu_2 \quad (5)$$

gdzie U, V są zmiennymi niezależnymi o $N(0;1)$.

W celu otrzymania realizacji dwuwymiarowej zmiennej losowej normalnej (X, Y) , uciętej do obszaru $a \leq X \leq b \wedge c \leq Y \leq d$ o gęstości (2) należy skorzystać z metody *rejection sampling* opisanego algorytmem [Zieliński, Wieczorkowski, 1997]:

1. Generuj U i V na podstawie (4).
2. Generuj X i Y na podstawie (5).
3. Jeżeli $a \leq X \leq b$ oraz $c \leq Y \leq d$, to zwróć (X, Y) . W sytuacji przeciwniej wróć do pkt. 1.

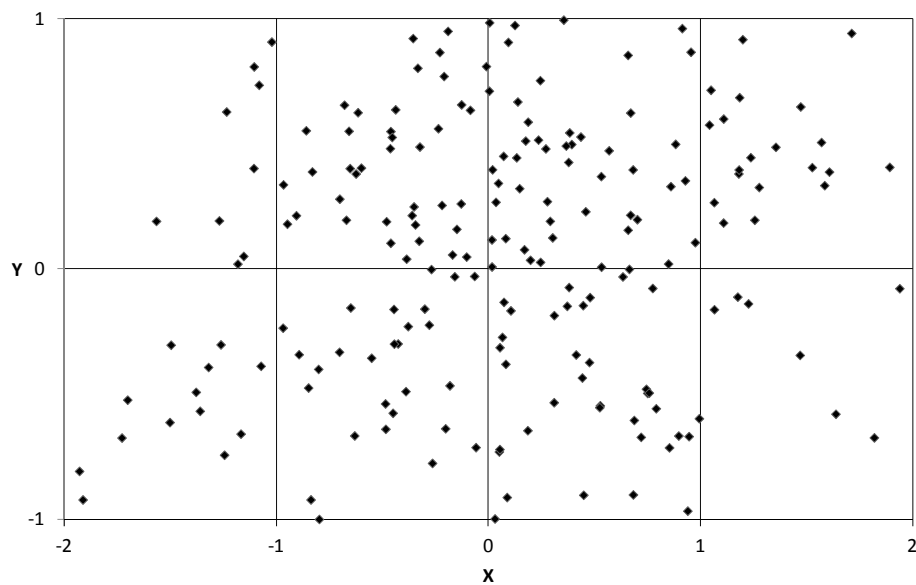
Implementację komputerową metody *rejection sampling* generowania prób z wielowymiarowego rozkładu normalnego uciętego w języku R przedstawiono w pracy S. Wilhelma i B.G. Manjunatha [2009]. Jeżeli tylko niewielki procent próbek wpada do obszaru $\langle a; b \rangle \times \langle c; d \rangle$, to można skorzystać z metody *Gibbs sampling*, w której 100% próbek wpada do obszaru $\langle a; b \rangle \times \langle c; d \rangle$ [Kotecha, Djuric, 1999; Horrace, 2005]. Główną wadą *Gibbs sampling* jest to, że próbki nie są niezależne, lecz skorelowane.

Badając właściwości testów statystycznych, należy nadać populacji generalnej określoną właściwość, jaką jest związek między cechami. Do wyznaczenia empirycznej mocy testów – w dobie szybko rozwijającej się komputeryzacji – metoda *rejection sampling* jest wystarczająca.

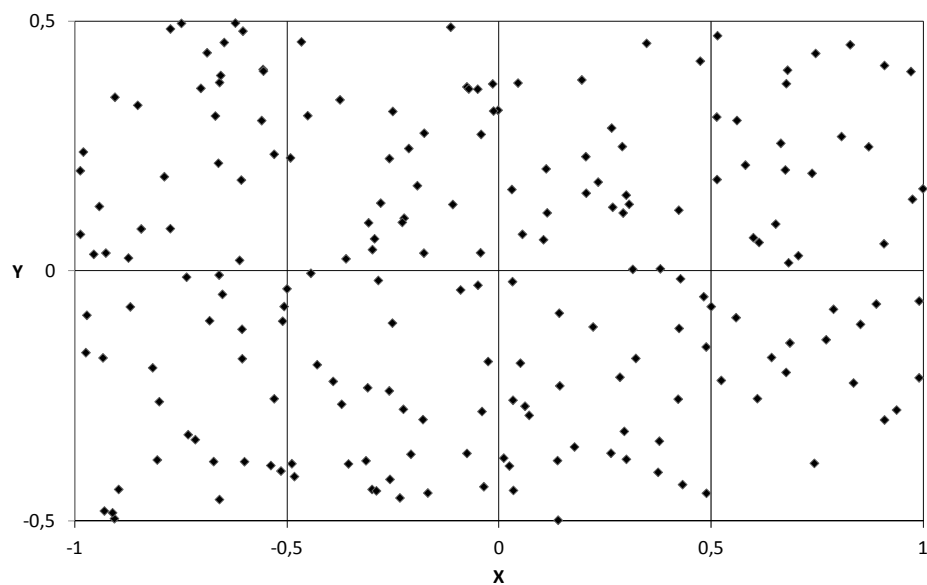
4. Generowanie tablic dwudzielczych $w \times k$ ($w, k > 2$)

W celu utworzenia TD $w \times k$ ($w, k > 2$), wygenerowano n realizacji zmiennej losowej o DRNU za pomocą wzorów (5) i metody *rejection sampling*. Otrzymane w ten sposób wartości (X_i, Y_i) ($i = 1, 2, \dots, n$) zapisano w układzie współrzędnych i na powstały w ten sposób obszar naniesiono TD $w \times k$.

Rysunek 5 przedstawia próbę $n = 200$ elementową, uzyskaną odpowiednio dla zestawu A wartości parametrów, natomiast rys. 6 – dla zestawu B (tab. 1). Próba ta tworzy obszar wartości dwuwymiarowej zmiennej losowej normalnej uciętej, który został nałożony na TD o wymiarach 2×4 .



Rys. 5. 200-elementowy obszar wartości zmiennej losowej o dwuwymiarowym rozkładzie normalnym uciętym dla zestawu A wartości parametrów, nałożony na TD 2×4



Rys. 6. 200-elementowy obszar wartości zmiennej losowej o dwuwymiarowym rozkładzie normalnym uciętym dla zestawu B wartości parametrów, nałożony na TD 2×4

W następnym kroku dla w wierszy i k kolumn obliczono szerokości poszczególnych klas za pomocą wzorów:

$$h_x = \frac{b-a}{w}, \quad h_y = \frac{c-d}{k}$$

oraz wyznaczono punkty podziału osi układu współrzędnych tworzące klasy:

$$Q_j = a + j \cdot h_x \quad (j = \overline{0, w}), \quad P_i = c + i \cdot h_y \quad (i = \overline{0, k})$$

Następnie stwierdzono, ile punktów o współrzędnych (X_i, Y_i) ($i = 1, 2, \dots, n$) należy do każdej z $w \cdot k$ klas TD. Dla tak wypełnionej TD wyznaczono rozkłady brzegowe. TD 2×4 odpowiadającą rys. 5 przedstawiono w tab. 4, natomiast TD 2×4 odpowiadającą rys. 6 – w tab. 5.

Tabela 4. Tablica dwudzielcza 2×4 dla zestawu A wartości parametrów dwuwymiarowego rozkładu normalnego uciętego

	Y_1	Y_2	Y_3	Y_4	Razem
X_1	9	39	49	20	117
X_2	13	27	36	7	83
Razem	22	66	85	27	200

Tabela 5. Tablica dwudzielcza 2×4 dla zestawu B wartości parametrów dwuwymiarowego rozkładu normalnego uciętego

	Y_1	Y_2	Y_3	Y_4	Razem
X_1	29	27	21	24	101
X_2	28	26	29	16	99
Razem	57	53	50	40	200

Z tab. 4 i 5 wynika, że najmniej realizacji mają skrajne kolumny, natomiast liczebność wierszy jest podobna. Dla tab. 5 obszar wartości zmiennej losowej o DRNU został zmniejszony i dzięki stałej wartości $r = 0,4$ w komórkach naróżnych pojawiło się więcej realizacji niż w tab. 4.

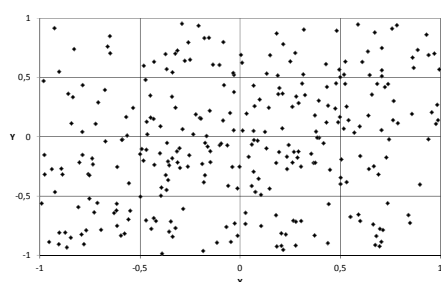
Przykład 1

Korzystając z DRNU, opisanego zestawami G i H wartości parametrów (tab. 6) wygenerowano TD 4×4 złożoną z 300 elementów. Wartości parametrów dobrano w taki sposób, by każda z 16 komórek TD była większa od zera.

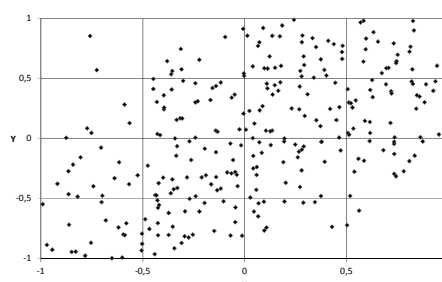
Tabela 6. Zestawy wartości parametrów dwuwymiarowego rozkładu normalnego uciętego

Zestaw	μ_1	σ_1	μ_2	σ_2	r	a	b	c	d
G	0	1	0	1	0,4	-1	1	-1	1
H	0	1	0	1	0,8	-1	1	-1	1

300 realizacji zmiennej losowej o DRNU, uzyskanych dla zestawów G i H wartości parametrów, naniesiono na układ współrzędnych (rys. 7A – zestaw G; rys. 7B – zestaw H). Następnie obliczono, ile realizacji wypełnia każdą z 16 klas, tworzących TD 4×4 . Dla tak wypełnionej TD wyznaczono rozkłady brzegowe. Uzyskane wyniki prezentują tab. 7 i 8.



A



B

Rys. 7. 300-elementowy obszar wartości zmiennej losowej o dwuwymiarowym rozkładzie normalnym uciętym nałożony na TD 4×4 **Tabela 7.** Tablica dwudzielcza 4×4 dla zestawu G wartości parametrów dwuwymiarowego rozkładu normalnego uciętego

	Y_1	Y_2	Y_3	Y_4	Razem
X_1	6	19	15	19	59
X_2	12	21	27	20	80
X_3	18	34	29	11	92
X_4	23	17	16	13	69
Razem	59	91	87	63	300

Tabela 8. Tablica dwudzielcza 4×4 dla zestawu H wartości parametrów dwuwymiarowego rozkładu normalnego uciętego

	Y_1	Y_2	Y_3	Y_4	Razem
X_1	2	10	31	24	67
X_2	5	28	28	32	93
X_3	13	31	24	15	83
X_4	19	26	10	2	57
Razem	39	95	93	73	300

Dla tab. 7 i 8 obszar wartości zmiennej losowej o DRNU jest stały, jednak różne są wartości współczynnika korelacji. Wraz z jego wzrostem zmniejsza się liczba realizacji w lewym górnym oraz prawym dolnym narożu.

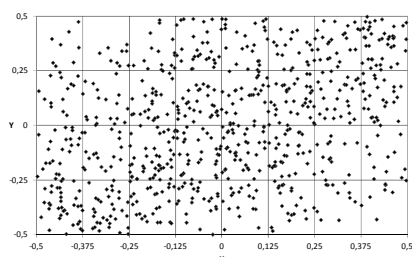
Przykład 2

Korzystając z DRNU, opisanego wartościami parametrów (tab. 9), wygenerowano zawartość TD 4×8 złożoną z 750 elementów. Wartości parametrów dobrano w taki sposób, by każda z 32 komórek TD była większa od zera.

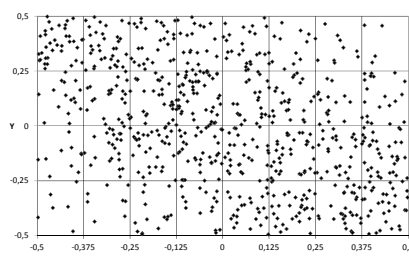
Tabela 9. Zestawy wartości parametrów dwuwymiarowego rozkładu normalnego uciętego

Zestaw	μ_1	σ_1	μ_2	σ_2	r	a	b	c	d
I	0	1	0	1	0,9	-0,5	0,5	-0,5	0,5
J	0	1	0	1	-0,9	-0,5	0,5	-0,5	0,5

750 realizacji zmiennej losowej o DRNU naniesiono na układ współrzędnych (zestaw I – rys. 8A, zestaw J – rys. 8B). Następnie obliczono, ile realizacji wypełnia każdą z 32 klas tworzących TD 4×8 . Dla tak wypełnionej TD, wyznaczono rozkłady brzegowe. Uzyskane wyniki prezentują tab. 10 i 11.



A



B

Rys. 8. 750-elementowy obszar wartości zmiennej losowej o dwuwymiarowym rozkładzie normalnym uciętym nałożony na TD 4×8 .

Tabela 10. Tablica dwudzielcza 4×8 dla zestawu I wartości parametrów dwuwymiarowego rozkładu normalnego uciętego

	Y_1	Y_2	Y_3	Y_4	Y_5	Y_6	Y_7	Y_8	Razem
X_1	8	9	16	29	22	19	24	40	167
X_2	15	18	23	30	29	34	26	26	201
X_3	25	17	44	30	38	30	12	11	207
X_4	35	33	26	24	17	20	12	8	175
Razem	83	77	109	113	106	103	74	85	750

Tabela 11. Tablica dwudzielcza 4×8 dla zestawu J wartości parametrów dwuwymiarowego rozkładu normalnego uciętego

	Y ₁	Y ₂	Y ₃	Y ₄	Y ₅	Y ₆	Y ₇	Y ₈	Razem
X ₁	37	34	27	20	27	12	9	4	170
X ₂	16	23	30	49	23	23	16	20	200
X ₃	10	23	32	35	29	34	26	34	223
X ₄	6	8	14	15	24	30	29	31	157
Razem	69	88	103	119	103	99	80	89	750

Z tab. 10 i 11 wynika, iż mimo silnej korelacji między zmiennymi X i Y komórki TD nie są puste, co miało miejsce w przypadku, gdy do generacji TD wykorzystano DRN. W tab. 9 najmniejsze wartości realizacji zmiennej losowej znajdują się w narożu lewym górnym oraz prawym dolnym (korelacja dodatnia). W tab. 10 najmniejsze wartości realizacji zmiennej losowej znajdują się w narożu lewym dolnym i prawym górnym (korelacja ujemna).

Przykład 3

Wyznaczono empiryczną moc testów na poziomie istotności $\alpha = 0,1$ dla tablicy dwudzielczej 3×3 oraz 4×4 , liczebności próby odpowiednio $n = 300$ oraz $n = 500$, korzystając ze statystyk chi-kwadrat. Tablice dwudzielcze generowano za pomocą dwuwymiarowego rozkładu normalnego uciętego dla zestawu K wartości parametrów (tab. 12). Wartości krytyczne wyznaczono symulacyjnie metodą Monte Carlo.

Tabela 12. Wartości parametrów dwuwymiarowego rozkładu normalnego uciętego

Zestaw	μ_1	σ_1	μ_2	σ_2	a	b	c	d	r
K	0	1	0	1	-1	1	-1	1	0; 0,05;...;0,7

Postaci statystyk chi-kwadrat [Cressie, Read, 1984] przedstawia tab. 13. Jak już to zostało wspomniane wcześniej, trzy z tych statystyk nie tolerują warunku $n_{ij} = 0$ ($i = 1, \dots, w$; $j = 1, \dots, k$).

Tabela 13. Statystyki chi-kwadrat (ang. *power divergence statistics*)

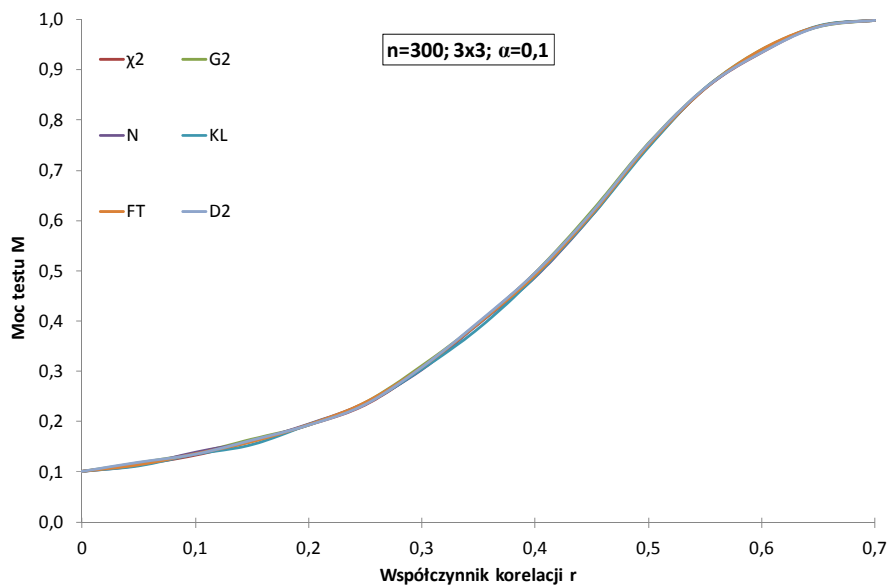
Nazwa statystyki	Postać statystyki
I	2
χ^2 Pearsona	$\chi^2 = \sum_{i=1}^w \sum_{j=1}^k \frac{(n_{ij} - e_{ij})^2}{e_{ij}}$
G^2 ilorazu wiarygodności	$G^2 = 2 \sum_{i=1}^w \sum_{j=1}^k n_{ij} \ln \left(\frac{n_{ij}}{e_{ij}} \right)$

cd. tabeli 13

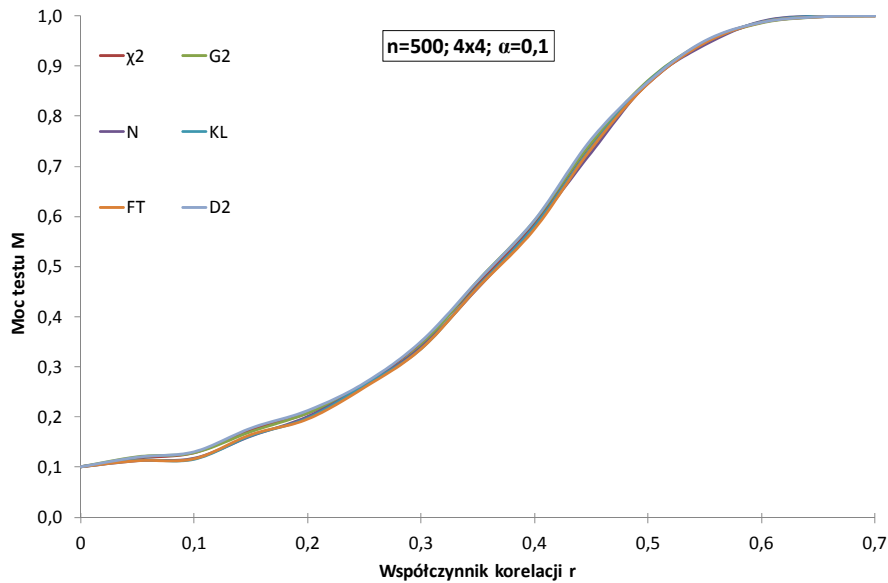
1	2
N Neymanna	$N = \sum_{i=1}^w \sum_{j=1}^k \frac{(n_{ij} - e_{ij})^2}{n_{ij}}$
KL Kullbacka-Leiblera	$KL = 2 \sum_{i=1}^w \sum_{j=1}^k e_{ij} \ln \left(\frac{e_{ij}}{n_{ij}} \right)$
FT Freemana-Tukeya	$FT = 4 \sum_{i=1}^w \sum_{j=1}^k \left(\sqrt{n_{ij}} - \sqrt{e_{ij}} \right)^2$
D ² Cressiego-Reada	$CR = \frac{9}{5} \sum_{i=1}^w \sum_{j=1}^k n_{ij} \left[\left(\frac{n_{ij}}{e_{ij}} \right)^{2/3} - 1 \right]$

Oznaczenia: e_{ij} – liczebność oczekiwana

Rysunki 9 i 10 przedstawiają empiryczne moce testów dla danych wartości parametrów DRNU (tab. 12). Wynika z nich, że testy niezależności wykorzystujące statystyki chi-kwadrat charakteryzują się podobną mocą dla wszystkich wartości współczynnika korelacji r i dla różnych rozmiarów TD. Analogiczna sytuacja miała miejsce w pracy P. Sulewskiego [2016], gdzie TD generowano metodą słupkową, a do wyznaczenia mocy testów skorzystano z innej miary nieprawdopodobieństwa H_0 .



Rys. 9. Empiryczne moce testów na poziomie istotności $\alpha = 0,1$, dla tablicy dwudzielczej 3×3 i liczebności próby $n = 300$



Rys. 10. Empiryczne moce testów na poziomie istotności $\alpha = 0,1$, dla tablicy dwudzielczej 4×4 i liczebności próby $n = 500$

Podsumowanie

Badając moc testów, należy wiedzieć, jaki jest rzeczywisty związek między cechami. W celu wyznaczenia empirycznej mocy testu niezbędne jest generowanie TD. Generowanie TD, gdy nie ma związku między badanymi cechami, nie przysparza trudności. Zadaniem niewątpliwie trudniejszym jest generowanie TD w sytuacji, gdy zachodzi związek między cechami. Generowanie TD z wykorzystaniem DRN napotyka na trudności związane z tym, iż jej komórki narożnikowe są puste. Zastosowanie DRNU sprawia, że nawet w sytuacji silnej korelacji między cechami, komórki TD nie są puste i zawierają wystarczającą liczbę realizacji.

Z przykładu 3 wynika, że zaproponowana metoda generacji TD jest alternatywą dla metody słupkowej i nie tylko.

Literatura

- Blitzstein J., Diaconis P. (2011), *A Sequential Importance Sampling Algorithm for Generating Random Graphs with Prescribed Degrees*, "Internet Mathematics", No. 6, s. 489-522.
- Box G.E.P., Muller M.E. (1958), *A Note on the Generation of Random Normal Deviates*, "The Annals of Mathematical Statistics", Vol. 29(2).

- Chen Y., Diaconis P., Holmes S.P., Liu J.S. (2005), *Sequential Monte Carlo Methods for Statistical Analysis of Tables*, "Journal of the American Statistical Association", No. 100, s. 109-120.
- Chen Y., Dinwoodie I.H., Sullivant S. (2006), *Sequential Importance Sampling for Multiway Tables*, "The Annals of Statistics", s. 523-545.
- Cramer H. (1958), *Metody matematyczne w statystyce*, PWN, Warszawa.
- Cressie N., Read T. (1984), *Multinomial Goodness-of-Fit Tests*, "Journal of the Royal Statistical Society: Series B (Statistical Methodology)", No. 46, s. 440-464.
- Cryan M., Dyer M. (2003), *A Polynomial-time Algorithm to Approximately Count Contingency Tables When the Number of Rows is Constant*, "Journal of Computer and System Sciences", No. 67, s. 291-310.
- Cryan M., Dyer M., Goldberg L.A., Jerrum M., Martin R. (2006), *Rapidly Mixing Markov Chains for Sampling Contingency Tables with a Constant Number of Rows*, "SIAM Journal on Computing", No. 36, s. 247-278.
- Desalvo S., Zhao J.Y. (2016), *Random Sampling of Contingency Tables via Probabilistic Divide-and-Conquer*, ArXiv preprint, arXiv:1507.00070v4.
- Diaconis P., Sturmfels B. (1998), *Algebraic Algorithms for Sampling from Conditional Distributions*, "The Annals of Statistics", No. 26, s. 363-397.
- Fishman G.S. (2012), *Counting Contingency Tables via Multistage Markov Chain Monte Carlo*, "Journal of Computational and Graphical Statistics", No. 21, s. 713-738.
- Holmes R.B., Jones L.K. (1996), *On Uniform Generation of Two-way Tables with Fixed Margins and the Conditional Volume test of Diaconis and Efron*, "The Annals of Statistics", Vol. 24(1).
- Horrace W.C. (2005), *Some Results on the Multivariate Truncated Normal Distribution*, "Journal of Multivariate Analysis", No. 94.
- Kendall M.G., Buckland W.R. (1986), *Słownik terminów statystycznych*, PWE, Warszawa.
- Kotecha J.H., Djuric P.M. (1999), *Gibbs Sampling Approach for Generation of Truncated Multivariate Gaussian Random Variables*, IEEE Computer Society.
- Nandram B., Bhatta D., Bhadra D. (2013), *A Likelihood Ratio Test of Quasi-independence for Sparse Two-way Contingency Tables*, "Journal of Statistical Computation and Simulation", Vol. 85(2), s. 284-304.
- Pearson K. (1904), *On the Theory of Contingency and its Relation to Association and Normal Correlation*, K. Pearson, Early Papers.
- Sulewski P. (2009), *Two-by-two Contingency Table as a Goodness-of-fit test*, "Computational Methods in Science and Technology", Vol. 15(2).
- Sulewski P. (2014a), *Statystyczne badanie współzależności cech typu dyskretne kategorie*, Wydawnictwo Naukowe Akademii Pomorskiej, Słupsk.
- Sulewski P. (2014b), *Wykorzystanie uogólnionego rozkładu gamma do generowania tablicy dwudzielczej*, "Śląski Przegląd Statystyczny", nr 12(18).

- Sulewski P. (2015), *Wyznaczanie obszaru krytycznego przy testowaniu niezależności w tablicach wielodzielczych*, „Wiadomości Statystyczne”, nr 3.
- Sulewski P. (2016), *Moc testów niezależności w tablicy dwudzielczej większej niż 2×2* , „Przegląd Statystyczny”, nr 63(2), s. 191-209.
- Wilhelm S., Manjunath B.G. (2009), *tmytnorm: A Package for the Truncated Multivariate Normal Distribution*, “The R Journal”, Vol. 2(1).
- Yoshida R., Xi J., Wei S., Zhou F., Haws D. (2011), *Semigroups and Sequential Importance Sampling for Multiway Tables*, arXiv preprint, arXiv:1111.6518.
- Zieliński R., Wieczorkowski R. (1997), *Komputerowe generatory liczb losowych*, WN-T, Warszawa.

GENERATING THE TWO-WAY CONTINGENCY TABLES USING THE TRUNCATED TWO-DIMENSIONAL NORMAL DISTRIBUTION

Summary: The procedure to generate random numbers of the truncated two-dimensional normal distribution (DRNU) was presented. The generating of two-way contingency table using random variable of DRNU was described. The presented theory has been enriched with specific examples, that were implemented in the VBA editor of spreadsheet Microsoft Excel.

Keywords: two-way contingency tables, generating of Monte Carlo, independence test, truncated two-dimensional normal distribution.