WYKORZYSTANIE UOGÓLNIONEGO ROZKŁADU GAMMA DO GENEROWANIA TABLICY DWUDZIELCZEJ

ŚLĄSKI PRZEGLĄD STATYSTYCZNY
Nr 12(18)

Piotr Sulewski
Akademia Pomorska w Słupsku

ISSN 1644-6739

Streszczenie: Artykuł poświęcony jest generowaniu zawartości tablicy dwudzielczej (TD) $2 \times k$ z wykorzystaniem uogólnionego rozkładu gamma (URG). Opisano w nim generator liczb losowych URG oraz sposób tworzenia TD $2 \times k$ na podstawie wartości dystrybuanty doświadczalnej i dystrybuanty teoretycznej rozkładu wykładniczego, który jest szczególnym przypadkiem URG.

Słowa kluczowe: generacja tablic dwudzielczych, uogólniony rozkład gamma, liczby losowe o uogólnionym rozkładzie gamma, rozkład wykładniczy.

DOI: 10.15611/sps.2014.12.18

1. Wstęp

W podręcznikach statystycznych znaleźć można głównie metody wnioskowania dotyczące jednej zmiennej. Jednak obiekty opisywane są często za pomocą większej liczby zmiennych. Tablicę, która powstaje przez podział danych według dwóch zmiennych, nazywa się tablicą dwudzielczą (dwuwymiarową) i zalicza do podstawowych narzędzi statystycznych.

Tablica dwudzielcza jest podstawowym i często stosowanym narzędziem statystycznym do badania siły związku między cechami typu jakościowego. W drodze analitycznej trudno jest uzyskać informacje na temat wykrywania związku między cechami w tablicy dwudzielczej, na ile czułym jest ona narzędziem. Jedyny sposób osiągnięcia tego celu stanowi generowanie tablic dwudzielczych i badania symulacyjne. Generowanie tablic dwudzielczych, gdy nie ma związku między badanymi cechami, jest rzeczą prostą, gdyż w takiej sytuacji można skorzystać z generatorów liczb równomiernych i generować niezależnie przynależność do wiersza i kolumny. Zadaniem niewątpliwie trudniejszym wydaje się generowanie TD w sytuacji, gdy zachodzi związek między cechami.

W pracy [Sulewski 2007a] przedstawiono procedurę generowania zawartości TD 2×2 z wykorzystaniem rozkładu normalnego dwuwy-

miarowego. Metoda ta jednak nie sprawdziła się jako generator TD o większych rozmiarach, gdyż narożne komórki tablicy często były puste. Z tego powodu w artykule [Sulewski 2007b] zaproponowano inną metodę generowania zawartości TD wykorzystującą liczby losowe o rozkładzie równomiernym, którą określono mianem „metody słupkowej". W pracy [Sulewski 2009] do generowania zawartości TD 2 × 2 wykorzystano URG.

Tablica dwudzielcza (TD) jako narzędzie do badania siły związku między cechami jest testem niezależności wykorzystującym statystykę χ^2 z $(k-1)(w-1)$ stopniami swobody. W literaturze spotyka się różne warunki co do minimalnej liczby realizacji w komórkach tablicy dwudzielczej. W pracy [Sobczyk 1996] stwierdzono, że wszystkie liczebności empiryczne powinny być nie mniejsze niż 5, czyli $n_{ij} \geq 5$ dla każdego $i = 1, 2, \ldots, w, j = 1, 2, \ldots, k$. W pracy [Oktaba 1974] proponuje się, by wszystkie liczebności oczekiwane były nie mniejsze niż 10, czyli $\tilde{n}_{ij} \geq 10$ dla każdego $i = 1, 2, \ldots, w, j = 1, 2, \ldots, k$. Autor niniejszej pracy minimalną liczebność realizacji w komórkach opisuje nierównością $\tilde{n}_{ij} \geq 5$ dla każdego $i = 1, 2, \ldots, w, j = 1, 2, \ldots, k$ zaproponowaną w pracy [Jóźwiak, Podgórski 1998].

Celem niniejszej pracy jest przedstawienie metody generowania zawartości TD 2 × k z wykorzystaniem URG, gdy związek między cechami istnieje. W punkcie drugim opisano generator liczb losowych o URG. Punkt drugi dotyczy sposobu tworzenia TD 2 × k z uwzględnieniem wartości dystrybuanty doświadczalnej i dystrybuanty teoretycznej rozkładu wykładniczego, który jest szczególnym przypadkiem URG.

2. Generator liczb losowych o uogólnionym rozkładzie gamma

URG jest rozkładem o złożonej postaci analitycznej, która daje mu pożądaną elastyczność. Jego funkcja gęstości wyrażona jest wzorem [Stacy 1962]

$$f(z;a,b,c) = \frac{b}{a\Gamma(c)}\left(\frac{z}{a}\right)^{bc-1} \exp\left[-\left(\frac{z}{a}\right)^{b}\right] \quad (z > 0), \tag{1}$$

gdzie:
$b > 0$, $c > 0$ – parametry kształtu,
$a > 0$ – parametr skali.

Dystrybuantę URG można zapisać za pomocą niepełnej funkcji gamma

$$\Gamma_n(c,x)=\int_0^x u^{c-1}\exp(-u)\,du \tag{2}$$

w postaci [Stacy 1962]

$$G(z)=\frac{\Gamma_n\left[c,(z/a)^b\right]}{\Gamma(c)}. \tag{3}$$

Jeżeli $f(z; a, b, c)$ jest funkcją gęstości URG, to $f(x; a, 1, c)$ jest funkcją gęstości rozkładu gamma, która dla $c = 1$ staje się funkcją gęstości rozkładu wykładniczego. Między zmienną losową X o rozkładzie $f(x)$ i zmienną losową Z o rozkładzie $f(z)$ zachodzi związek [Wieczorkowski, Zieliński 1997]

$$X=\left(\frac{Z}{a}\right)^b \Rightarrow Z=a\cdot X^{1/b}, \tag{4}$$

więc wystarczy skonstruować generator realizacji zmiennej losowej X o rozkładzie gamma.

Najprostszy algorytm otrzymuje się wówczas, gdy c jest liczbą całkowitą. Niech X_1, X_2 będą zmiennymi losowymi niezależnymi. Jeżeli X_1 ma rozkład gamma z parametrem c_1 oraz X_2 ma rozkład gamma z parametrem c_2, to zmienna losowa $X_1 + X_2$ ma rozkład gamma z parametrem $c_1 + c_2$. Dla otrzymania zmiennej losowej o rozkładzie gamma z całkowitym parametrem c generuje się c realizacji zmiennych losowych o rozkładzie wykładniczym i oblicza ich sumę. Zatem

$$X=-\ln(U_1)-\ln(U_2)-\ldots-\ln(U_c)=-\ln\left(\prod_{i=1}^{c}U_i\right), \tag{5}$$

gdzie:
$U_1, U_2, \ldots, U_c$ – niezależne zmienne losowe o rozkładzie równomiernym $U(0; 1)$.

Jeżeli c nie jest liczbą całkowitą, realizację zmiennej losowej X o rozkładzie gamma generuje się na podstawie wzoru

$$X=X_1+X_2X_3, \tag{6}$$

gdzie:
X_1 – zmienna losowa o rozkładzie gamma z parametrem $n = [c]$ (część całkowita z c),

ŚLĄSKI PRZEGLĄD STATYSTYCZNY
Nr 12 (18)

$$X_1 = -\ln(U_1) - \ln(U_2) - \ldots - \ln(U_n) = -\ln\left(\prod_{i=1}^{n} U_i\right), \tag{7}$$

$U_1, U_2, \ldots, U_n$ – niezależne zmienne losowe o rozkładzie równomiernym $U(0; 1)$,
X_2 – zmienna losowa o rozkładzie gamma z parametrem $c = 1$ (rozkład wykładniczy),

$$X_2 = -\ln(U), \tag{8}$$

U – zmienna losowa o rozkładzie równomiernym $U(0; 1)$,
X_3 – zmienna losowa o rozkładzie beta z parametrami $(d, 1 - d)$, $d = c - [c] \in (0; 1)$.

Realizację zmiennej losowej X_3 otrzymano, stosując następujący algorytm:

a) generuje się realizację zmiennej losowej W o rozkładzie potęgowym z parametrem d

$$W = U^{1/d}, \tag{9}$$

U jest niezależną zmienną losową o rozkładzie równomiernym $U(0; 1)$;

b) generuje się realizację zmiennej losowej V o rozkładzie potęgowym z parametrem $1 - d$

$$V = U^{1/(1-d)}, \tag{10}$$

U jest niezależną zmienną losową o rozkładzie równomiernym $U(0; 1)$;

c) jeżeli $W + V > 1$, to powtarza się operacje a) i b); w wypadku przeciwnym

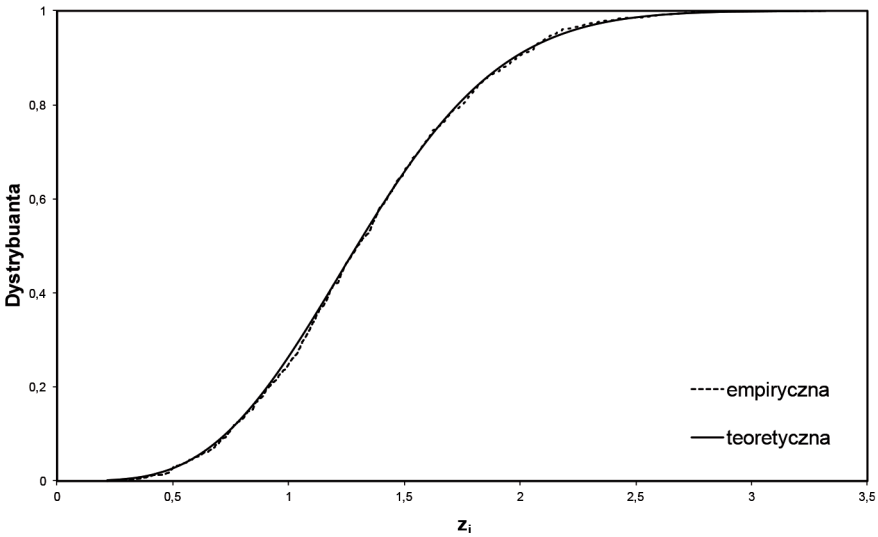
$$X_3 = \frac{W}{W+V}. \tag{11}$$

Generacje realizacji zmiennych losowych W (9) i V (10) wykonano metodą odwracania dystrybuanty rozkładu potęgowego.

W celu sprawdzenia poprawności działania generatora liczby losowe z_i^* posortowano, a następnie obliczono na podstawie (3) wartości dystrybuanty teoretycznej $T\left(z_{(i)}^*\right)$ oraz wartości dystrybuanty empirycznej danej wzorem

$$F_i = \frac{i}{n+1} \quad i = 1, \ldots, n. \tag{12}$$

Rysunek 1 przedstawia przebiegi dystrybuanty empirycznej i teoretycznej URG dla $a = 1$; $b = 2$; $c = 1{,}5$ oraz liczebności próby $n = 1000$.



Rys. 1. Dystrybuanta empiryczna i teoretyczna URG dla $n = 1000$ i $a = 1$; $b = 2$; $c = 1{,}5$

Źródło: opracowanie własne.

Jak wynika z rys. 1, przebiegi dystrybuanty teoretycznej URG i dystrybuanty empirycznej pokrywają się, co świadczy o tym, że liczby losowe z_i^* mają URG.

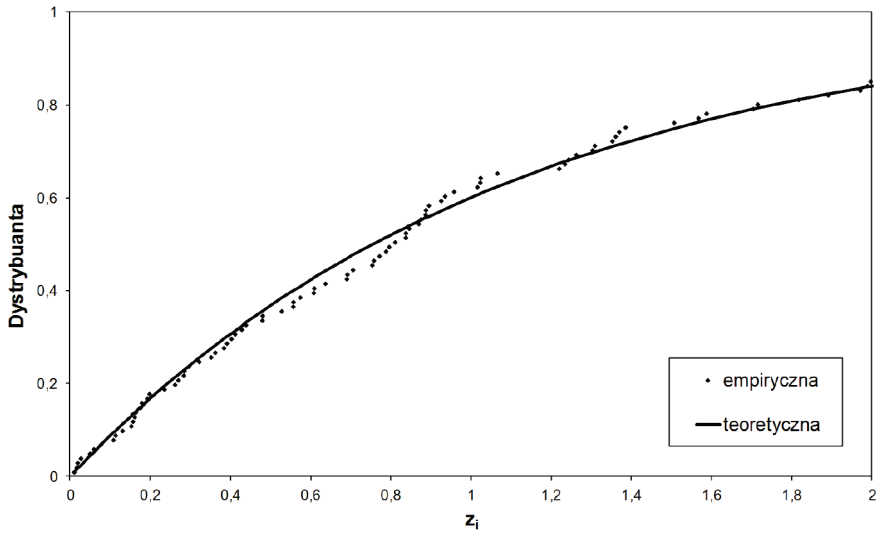
3. Generacja zawartości tablicy dwudzielczej 2 × k

Niech $z_{(i)}^*$ będą liczbami losowymi o URG posortowanymi rosnąco. Do utworzenia TD wykorzystano wartości dystrybuanty teoretycznej rozkładu wykładniczego

$$T\left(z_{(i)}^*; a^*\right) = 1 - \exp\left(-a^* \cdot z_{(i)}^*\right) \tag{13}$$

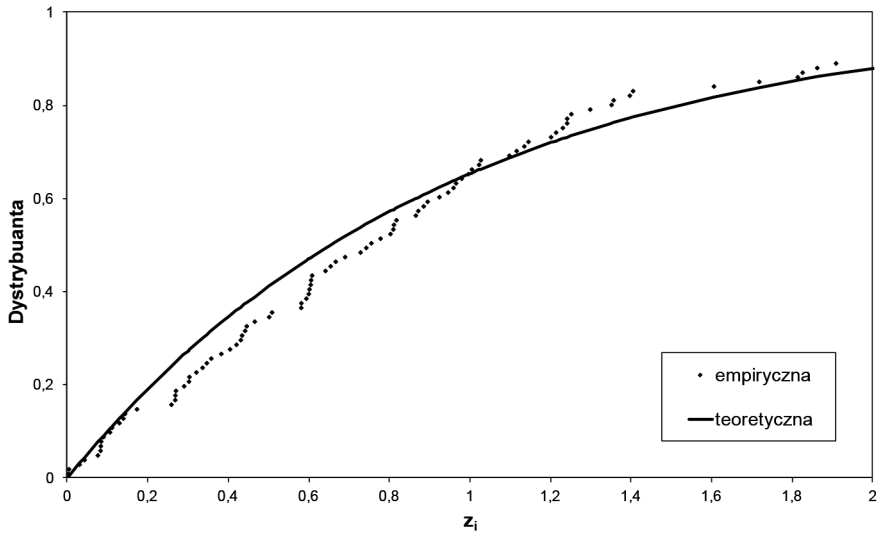
oraz wartości dystrybuanty empirycznej (12).

Do oszacowania nieznanej wartości parametru rozkładu wykładniczego skorzystano z metody najmniejszych kwadratów (MNK), dzięki której dystrybuanta empiryczna lepiej otacza dystrybuantę teoretyczną (rys. 2) niż ma to miejsce w metodzie największej wiarygodności (MNW) czy w metodzie momentów (MM) (rys. 3).



Rys. 2. Przebiegi dystrybuant, gdy parametr rozkładu wykładniczego szacowano MNK

Źródło: opracowanie własne.



Rys. 3. Przebiegi dystrybuant, gdy parametr rozkładu wykładniczego szacowano MM lub MNW

Źródło: opracowanie własne.

Jako oszacowanie a^* parametru a przyjęto wartość, która minimalizuje funkcję

$$M(a)=\sum_{i=1}^{n}\left[T\left(z_{(i)}^{*};a\right)-F_i\right]^2 . \tag{14}$$

Na podstawie (12) i (13) wyznaczono różnice dystrybuant

$$D^i = F_i - T\left(z_{(i)}^{*};a^{*}\right) \quad i=1,2,...,n \tag{15}$$

oraz wartości bezwzględne tych różnic uporządkowane w kolejności wzrastania

$$DP_{(j)} = \left|D^j\right| \quad j=1,2,...,n . \tag{16}$$

Znak D^i (i = 1, 2, …, n) decyduje o tym, do którego wiersza należy dana realizacja według zasady pokazanej w tab. 1. O przynależności do kolumny decydują wartości percentyli stopnia u/k (u = 1, 2. …, k – 1) obliczone ze wzoru

$$Per_{\frac{u}{k}} = \begin{cases} DP_{\left[\frac{(n+1)u}{k}\right]} & n-\text{nieparzyste} \\ \dfrac{DP_{[\frac{nu}{k}]}+DP_{[\frac{nu}{k}]+1}}{2} & n-\text{parzyste} \end{cases}, \tag{17}$$

gdzie [.] oznacza część całkowitą liczby.

Ze wzoru (17) wynika, że liczba percentyli jest o jeden mniejsza niż liczba kolumn tablicy dwudzielczej, którą zamierzamy wygenerować. W szczególności, gdy tablica ma cztery kolumny, wyznaczamy
kwartyl dolny $Q_1 = Per_{1/4}$,
medianę $M = Per_{1/2}$,
kwartyl górny $Q_1 = Per_{3/4}$.

Zasadę tworzenia tablicy 2 × 4 przedstawiono w tab. 1.

Tabela 1. Sposób postępowania przy tworzeniu tablicy 2 × 4

	Y_1	Y_2	Y_3	Y_4
X_1	$D^i > 0$ $D^i \le Q_1$	$D^i > 0$ $Q_1 < D^i \le M$	$D^i > 0$ $M < D^i \le Q_3$	$D^i > 0$ $D^i > Q_3$
X_2	$D^i \le 0$ $\left\|D^i\right\| \le Q_1$	$D^i \le 0$ $Q_1 < \left\|D^i\right\| \le M$	$D^i \le 0$ $M < \left\|D_i\right\| \le Q_3$	$D^i \le 0$ $\left\|D^i\right\| > Q_3$

Źródło: opracowanie własne.

Tabela 2 przedstawia TD 2 × 4 wygenerowaną za pomocą URG, gdy $a = 1$; $b = 1$; $c = 1$ (brak związku między X i Y). Tabela 3 przedstawia wygenerowaną za pomocą URG TD 2 × 4, gdy $a = 1$; $b = 1{,}01$; $c = 1$. Tabela 4 przedstawia wygenerowaną za pomocą URG TD 2 × 4, gdy $a = 1$; $b = 0{,}95$; $c = 1$.

Tabela 2. Tablica dwudzielcza wygenerowana za pomocą URG, gdy $a = 1$; $b = 1$; $c = 1$

	Y_1	Y_2	Y_3	Y_4	Razem
X_1	57	36	30	20	143
X_2	18	39	45	55	157
Razem	75	75	75	75	300

Źródło: opracowanie własne.

Tabela 3. Tablica dwudzielcza wygenerowana za pomocą URG, gdy $a = 1$; $b = 1{,}01$; $c = 1$

	Y_1	Y_2	Y_3	Y_4	Razem
X_1	24	30	26	34	114
X_2	51	45	49	41	186
Razem	75	75	75	75	300

Źródło: opracowanie własne.

Tabela 4. Tablica dwudzielcza wygenerowana za pomocą URG, gdy $a = 1$; $b = 0{,}95$; $c = 1$

	Y_1	Y_2	Y_3	Y_4	Razem
X_1	46	40	46	39	171
X_2	29	35	29	36	129
Razem	75	75	75	75	300

Źródło: opracowanie własne.

Dla $b = 1$, gdy między cechami nie ma związku, liczebności wierszy są podobne (tab. 2). Dla $b \neq 1$ uzyskuje się związek między cechami. Zwiększaj nieznacznie wartość parametru b ($b = 1{,}01$) większość elementów próby znajduje się w wierszu 2 (tab. 3). Gdy wartość parametru b się zmniejsza, ($b = 0{,}95$), większość elementów próby znajduje się w wierszu 1 (tab. 4).

4. Podsumowanie

Wykrywanie związku między cechami w tablicy dwudzielczej jest trudne na drodze analitycznej. Jedyny sposób osiągnięcia tego celu stanowi generowanie tablic dwudzielczych i badania symulacyjne.

Generowanie tablic dwudzielczych, gdy nie ma związku między badanymi cechami, nie przysparza trudności. Zadaniem niewątpliwie trudniejszym jest generowanie TD w sytuacji, gdy zachodzi związek między cechami.

W niniejszej pracy opisano sposób generowania zawartości TD $2 \times k$, do którego wykorzystano uogólniony rozkład gamma z parametrami a, b, c, którego szczególnym przypadkiem jest doskonale znany rozkład wykładniczy ($a = 1$, $b = 1$, $c = 1$). Jeżeli generuje się zawartość TD $2 \times k$, gdy związku między cechami nie ma, należy w symulacjach przyjąć $a = 1$, $b = 1$, $c = 1$. Jeżeli generuje się zawartość TD $2 \times k$, gdy związek między cechami jest, należy w symulacjach przyjąć $b \neq 1$.

Literatura

Jóźwiak J., Podgórski J., *Statystyka od podstaw*, PWE, Warszawa 1998.

Oktaba W., *Elementy statystyki matematycznej i metodyka doświadczalnictwa*, PWN, Warszawa 1974.

Sobczyk M., S*tatystyka*, PWN, Warszawa 1996.

Stacy E.W., *A generalization of the gamma distribution*, Annals of Mathematical Statistics 1962, Vol. 33.

Sulewski P., *Test niezależności dwóch cech realizowany za pomocą tablicy dwudzielczej*, Słupskie Prace Matematyczno-Fizyczne nr 4, Słupsk 2007a, s. 83–97.

Sulewski P., *Moc tablicy dwudzielczej jako test niezależności*, „Wiadomości Statystyczne" 2007b, nr 6, s. 14–23.

Sulewski P., *Two-by-two contingency table as a goodness-of-fit test*, „Computational Methods in Science and Technology" 2009, Vol. 15, No. 2, Poznań, s. 203–211.

Wieczorkowski R., Zieliński R., *Komputerowe generatory liczb losowych*, WNT, Warszawa 1997.

USING THE GENERALIZED GAMMA DISTRIBUTION TO GENERATE CONTINGENCY TABLES

Summary: The article is devoted to the generation of two-way table contents using the generalized gamma distribution (GG). It describes the generalized gamma random number generator and how to create a two-way table by means of the empirical distribution function and theoretical exponential distribution, which is a special case of GG.

Keywords: generation of two-way tables, generalized gamma distribution, generalized gamma random value, exponential distribution.