PRACE NAUKOWE UNIWERSYTETU EKONOMICZNEGO WE WROCŁAWIU nr 141
Ekonometria 29 2010

Wojciech Gamrot
Uniwersytet Ekonomiczny w Katowicach

PORÓWNANIE WŁASNOŚCI WYBRANYCH GENERATORÓW WIELOWYMIAROWYCH BINARNYCH LICZB PSEUDOLOSOWYCH

Streszczenie: W artykule rozważono dwie popularne procedury generujące ciągi wektorów pseudolosowych o wielowymiarowym rozkładzie zero-jedynkowym z korelacją, zaproponowane przez Leischa i in., a także przez Parka i in. Zaprezentowano wyniki symulacji przeprowadzonych w celu porównania obu procedur pod względem zgodności momentów rozkładu generowanego ciągu z momentami zadanymi. Pokazano, że dla pierwszej z nich mogą wystąpić istotne rozbieżności między rozkładem generowanym i rozkładem zadanym.

Słowa kluczowe: liczby pseudolosowe, wielowymiarowy rozkład zero-jedynkowy, wartość oczekiwana, współczynnik korelacji liniowej.

1. Wstęp

Problem generowania wektorów liczb pseudolosowych o wielowymiarowym rozkładzie zero-jedynkowym z zadaną macierzą korelacji ma istotne znaczenie praktyczne. Sztuczne dane wygenerowane przy użyciu odpowiedniej procedury numerycznej znajdują bowiem zastosowanie w tak różnorodnych dziedzinach, jak segmentacja danych marketingowych, symulacyjne analizy toksykologiczne, modelowanie wystąpień rozmaitych schorzeń w sąsiadujących organach (stomatologia, oftalmologia), konstrukcja schematów losowania prób w metodzie reprezentacyjnej, losowanie podprób dla estymacji wariancji w pewnych wariantach metody *bootstrap* czy też symulacyjne badanie własności estymatorów dla złożonych mechanizmów opisujących niekompletność danych w próbie statystycznej. Od procedur realizujących to zadanie wymaga się, z jednej strony, prostoty implementacji i znacznej szybkości działania, a z drugiej, spełnienia przez generowany ciąg szeregu własności, wśród których należy wymienić długi okres, brak autokorelacji i zgodność generowanego rozkładu z zadaną specyfikacją, a szczególnie równość

momentów rozkładu generowanego ciągu z momentami zadanymi przez użytkownika. Konstrukcja procedur generujących sztuczne dane o wielowymiarowym rozkładzie dwupunktowym jest zadaniem bardziej złożonym niż w przypadku takich rozkładów, jak rozkład normalny, ze względu na to, iż wariancje (kowariancje) poszczególnych zmiennych losowych tworzących składowe wektora zależą od ich wartości oczekiwanych. Zagadnienie to było już w literaturze przedmiotu wielokrotnie omawiane i doczekało się licznych prób rozwiązania, wśród których warto wymienić m.in. prace takich autorów, jak Bahadur [1961], Emrich oraz Piedmonte [1991], Lee [1993], a także Gange [1995]. Obecnie do popularnych rozwiązań należą algorytmy, które zaproponowali Leisch i in. (1998), a także Park i in. [1996]. Pierwszy z nich opiera się na odpowiednim przekształceniu wektora losowego o wielowymiarowym rozkładzie normalnym i zyskał popularność ze względu na dużą szybkość działania oraz dostępność implementacji w pakietach statystycznych, a zwłaszcza w pakiecie R. Zaletą drugiego z wymienionych algorytmów, opierającego się na funkcjach sumy zmiennych losowych o rozkładzie Poissona, jest brak konieczności przybliżonego rozwiązywania na drodze iteracyjnej układu równań lub zadania optymalizacji, dzięki czemu algorytm ten powinien się charakteryzować niewielką skalą błędów numerycznych oraz relatywnie niską złożonością obliczeniową, zwłaszcza w sytuacji, gdy wymaga się generowania ciągów wektorów pseudolosowych o każdorazowo różnych parametrach rozkładu. W niniejszej pracy obie wymienione procedury porównano pod względem zgodności generowanych ciągów liczb pseudolosowych z zadanym rozkładem na drodze symulacji komputerowej.

2. Procedury generujące liczby pseudolosowe

Wektor binarnych zmiennych losowych $\mathbf{X} = [X_1, ..., X_k]' \in \{0,1\}^k$ może w ogólnym przypadku przyjąć 2^k różnych wartości. Dla dużych k jest to ogromna liczba, co utrudnia wyczerpującą specyfikację i generowanie realizacji wektora **X** przy użyciu klasycznych metod odwracania dystrybuanty lub równomiernego rozbicia przedziału opisywanych m.in. przez Zielińskiego i Wieczorkowskiego [1997]. Dlatego też zazwyczaj do celów generacji określa się jedynie wektor prawdopodobieństw brzegowych $\mathbf{m} = [m_1, ..., m_k]$, gdzie $m_i = \mathrm{P}\{X_i = 1\}$ dla $i=1, ..., k$, czyli $\mathrm{E}(\mathbf{X}) = \mathbf{m}$, oraz macierz łącznych prawdopodobieństw $\mathbf{M} = [m_{ij}]$ o wymiarach $k \times k$, gdzie $m_{ij} = \mathrm{P}\{X_iX_j\} = 1$ dla $i, j = 1, ..., k$. Macierz ta zawiera drugie momenty mieszane wektora **X**, czyli $\mathrm{E}(\mathbf{X}'\mathbf{X}) = \mathbf{M}$. Alternatywne sformułowanie powyższej specyfikacji możliwe jest poprzez określenie macierzy kowariancji

$$\mathbf{C} = [c_{ij}] = \mathrm{E}((\mathbf{X} - \mathbf{m})' (\mathbf{X} - \mathbf{m})) = \mathbf{M} - \mathbf{m}\mathbf{m}' \tag{1}$$

lub macierzy korelacji $\mathbf{R} = [r_{ij}]$, gdzie

$$r_{ij} = \frac{m_{ij} - m_i m_j}{\sqrt{m_i(1-m_i)m_j(1-m_j)}}. \tag{2}$$

Macierz **C** musi przy tym być symetryczna i nieujemnie określona [Goldberger 1975], a współczynniki korelacji r_{ij} powinny spełniać relacje [Emrich, Piedmonte 1991]:

$$r_{ij} \le \sqrt{\frac{m_i(1-m_j)}{m_j(1-m_i)}} \tag{3}$$

dla $i, j = 1, ..., k$. Zgodnie z procedurą zaproponowaną przez Leischa i in. [1998], każdą pseudolosową realizację $\mathbf{X} = [X_1, ..., X_k]$ wektora **X** wyznacza się na podstawie realizacji $\mathbf{Y} = [Y_1, ..., Y_k]$ wektora pseudolosowego **Y** o k-wymiarowym rozkładzie normalnym z pewnym wektorem wartości oczekiwanych $\mathbf{m}_N$ i macierzą kowariancji $\mathbf{C}_N$, przyjmując $X_i = 1$, gdy $Y_i > 0$ oraz $X_i = 0$ w przeciwnym wypadku. Działanie algorytmu opiera się na doborze $\mathbf{m}_N$ i $\mathbf{C}_N$ w taki sposób, aby spełnione były warunki:

$$m_i = \mathrm{P}(Y_i > 0) = \int_{Z_i} \varphi(\mathbf{y}) d\mathbf{y}, \tag{4}$$

$$m_{ij} = \mathrm{P}(Y_i > 0, Y_j > 0) = \int_{Z_{ij}} \varphi(\mathbf{y}) d\mathbf{y}, \tag{5}$$

gdzie: $\mathbf{y} = [y_1, .., y_k]' \in \mathrm{R}^k$,

$\varphi(\mathbf{y})$ – funkcja gęstości k-wymiarowego rozkładu normalnego o wektorze wartości oczekiwanych $\mathbf{m}_N$ i macierzy kowariancji $\mathbf{C}_N$,

$Z_i = \{\mathbf{y} \in \mathrm{R}^k : y_i > 0\}$,

$Z_{ij} = \{\mathbf{y} \in \mathrm{R}^k : y_i > 0, y_j > 0\}$.

Wartości $\mathbf{m}_N$ i $\mathbf{C}_N$ otrzymuje się z wykorzystaniem iteracyjnej procedury, przy czym dla różnych ich kombinacji wartości powyższych wyrażeń wyznaczane są drogą numerycznego całkowania funkcji $\varphi(\mathbf{y})$. Gdy generowanych jest wiele realizacji binarnej zmiennej losowej o tym samym rozkładzie, wystarczy że $\mathbf{m}_N$ i $\mathbf{C}_N$ zostaną wyznaczone tylko jednokrotnie. W przedstawionych dalej obliczeniach wykorzystano implementację tej procedury *'rmvbin'* dostępną w module '*bindata*' pakietu statystycznego R. Będzie ona skrótowo oznaczana symbolem LWH.

Procedura Parka i in. [1996] wykorzystuje znany fakt, że zmienna losowa o rozkładzie Poissona może zostać wyrażona jako suma innych zmiennych o rozkładzie Poissona. Niech $I_A(i) = 1$, gdy $i \in A$ oraz $I_A(i) = 0$ w przeciwnym wypadku. Generowanie wektora binarnego $\mathbf{X} = [X_1, ..., X_k]'$ o wektorze prawdopodobieństw brzegowych **m** oraz macierzy korelacji **R** o nieujemnych elementach odbywa się poprzez sekwencyjną realizację następujących kroków:

Krok 0: Oblicz $\alpha_{ij} = \log(1 + r_{ij}(1-m_i)^{0.5} m_i^{-0.5}(1-m_j)^{0.5} m_j^{-0.5})$ dla $1 \le i,j \le k$.
Niech $e = 0$.

Krok 1: Niech $e = e + 1$.
Niech $T_e = \{\alpha_{ij}: \alpha_{ij} > 0,\ 1 \le i, j \le k\}$.
Niech β_e będzie najmniejszym elementem zbioru T_e.
Niech S_e będzie najliczniejszym zbiorem indeksów zawierających $\{r, s\}$, dla którego $\alpha_{ij} > 0$ dla $i, j \in S_e$.

Krok 2: Dla wszystkich $\alpha_{ij} > 0$ zastąp α_{ij} przez $\alpha_{ij} - \beta_e$. Gdy $\alpha_{ij} = 0$ dla $1 \le i, j \le k$, przejdź do
kroku 3, a w przeciwnym wypadku do kroku 1.

Krok 3: Wygeneruj $Y_1(\beta_1)$, ..., $Y_e(\beta_e)$ o rozkładzie Poissona z parametrami β_1, ..., β_e.
Dla $i = 1, ..., k$ niech $Z_i = \sum_{j=1}^{e} Y_j(\beta_j) I_{S_j}(i)$ oraz $X_i = I_{\{0\}}(Z_i)$.

Procedura ta będzie dalej skrótowo oznaczana symbolem PPS. Na potrzeby eksperymentów symulacyjnych została ona zaimplementowana w pakiecie statystycznym R.

3. Wyniki symulacji

Celem przeprowadzonego eksperymentu symulacyjnego było porównanie własności ciągów wektorów pseudolosowych będących wynikiem działania procedur LWH i PPS oraz ocena ich zgodności z zadanymi parametrami generatorów, reprezentowanymi przez arbitralnie dobrany wspólny wektor wartości oczekiwanych $\mathbf{m} = [0.4, 0.5, 0.6]'$ i arbitralnie dobraną, wspólną macierz kowariancji:

$$\boldsymbol{R} = \begin{bmatrix} 1 & 0.2 & 0.3 \\ 0.2 & 1 & 0.4 \\ 0.3 & 0.4 & 1 \end{bmatrix}. \tag{6}$$

Pożądane prawdopodobieństwa łącznych wystąpień jedynek dla poszczególnych par zmiennych opisuje omawiana wcześniej macierz $\mathbf{M} = [m_{ij}]$, której elementy na podstawie (2) można wyznaczyć jako:

$$m_{ij} = m_i m_j + r_{ij}\sqrt{m_i(1-m_i)m_j(1-m_j)}. \tag{7}$$

Macierz ta przyjmuje postać:

$$\boldsymbol{M} = \begin{bmatrix} 0.4000000 & 0.2489898 & 0.3120000 \\ 0.2489898 & 0.5000000 & 0.3979796 \\ 0.3120000 & 0.3979796 & 0.6000000 \end{bmatrix}. \tag{8}$$

Eksperyment symulacyjny polegał na wygenerowaniu ciągu obejmującego 1,8 miliona wektorów przy użyciu każdej z dwóch omawianych procedur. Na podstawie rozkładu empirycznego obu procedur zliczono następnie wystąpienia wartości 1 dla każdej ze składowych generowanego wektora, otrzymując wektor częstości:

$$\mathbf{w}_{\mathrm{LWH}} = [719201\ 900289\ 1080980] \tag{9}$$

dla procedury LWH oraz

$$\mathbf{w}_{\mathrm{PPS}} = [720293\ 900180\ 1080487] \tag{10}$$

dla procedury PPS, co odpowiada wektorom frakcji jedynek w rozkładzie empirycznym każdej ze zmiennych równym odpowiednio:

$$\mathbf{f}_{\mathrm{LWH}} = [0.3995561\ 0.5001606\ 0.6005444] \tag{11}$$

dla procedury LWH oraz

$$\mathbf{f}_{\mathrm{PPS}} = [0.4001628\ 0.5001000\ 0.6002706] \tag{12}$$

dla procedury PPS. Dla każdej ze składowych generowanego wektora zweryfikowano następnie, z wykorzystaniem dokładnego testu parametrycznego opartego na wzorze Bernoulliego, prawdziwość hipotezy stanowiącej o tym, że wartość oczekiwana w rozkładzie empirycznym jest równa odpowiednio 0.4, 0.5 i 0.6. *P*-wartości (istotności) tego testu odpowiadające obustronnemu obszarowi krytycznemu tworzą wektor:

$$\mathbf{p}_{\mathrm{LWH}} = [0.2244122,\ 0.6671444,\ 0.1361553] \tag{13}$$

dla procedury LWH oraz:

$$\mathbf{p}_{\mathrm{PPS}} = [0.6557524,\ 0.7890205,\ 0.4591870] \tag{14}$$

dla procedury PPS, co oznacza, że wartości oczekiwane poszczególnych składowych generowanego wektora nie różniły się istotnie od zakładanych. Następnie zliczono przypadki łącznego wystąpienia wartości 1, prezentując je w formie macierzy $\mathbf{W} = [w_{ij}]$, gdzie w_{ij} oznacza zarejestrowaną liczbę zdarzeń polegających na wystąpieniu jedynki dla *i*-tej oraz *j*-tej składowej wektora $\mathbf{X}$ równocześnie (w_{ii} odpowiada oczywiście omawianej wyżej liczbie jedynek dla *i*-tej zmiennej). Dla procedur LWH i PPS otrzymano odpowiednio macierze:

$$\mathbf{W}_{\mathrm{LWH}} = \begin{bmatrix} 719201 & 447078 & 560287 \\ 447078 & 900289 & 718675 \\ 560287 & 718675 & 1080980 \end{bmatrix} \tag{15}$$

oraz

$$\mathbf{W}_{\text{PPS}} = \begin{bmatrix} 720293 & 448072 & 561808 \\ 448072 & 900180 & 715951 \\ 561808 & 715951 & 1080487 \end{bmatrix}, \tag{16}$$

którym odpowiadają frakcje jedynek w rozkładzie empirycznym odpowiednio równe:

$$\mathbf{F}_{\text{LWH}} = \begin{bmatrix} 0.3995561 & 0.2483767 & 0.3112706 \\ 0.2483767 & 0.5001606 & 0.3992639 \\ 0.3112706 & 0.3992639 & 0.6005444 \end{bmatrix} \tag{17}$$

oraz

$$\mathbf{F}_{\text{PPS}} = \begin{bmatrix} 0.4001628 & 0.2489289 & 0.3121156 \\ 0.2489289 & 0.5001000 & 0.3977506 \\ 0.3121156 & 0.3977506 & 0.6002706 \end{bmatrix}. \tag{18}$$

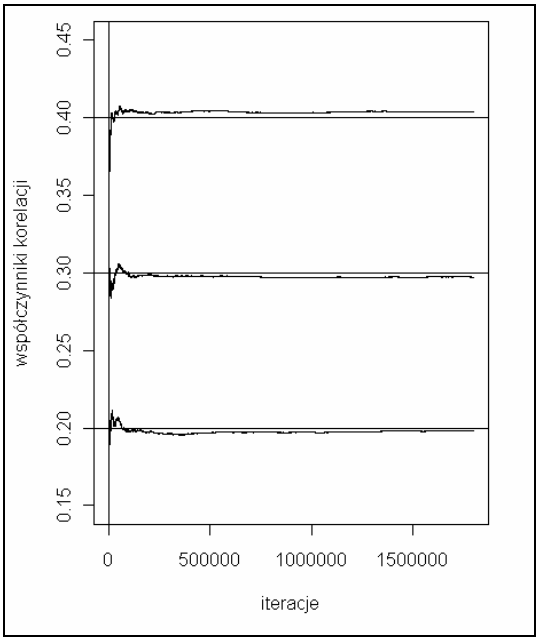
Za pomocą dokładnego testu parametrycznego opartego na wzorze Bernoulliego zweryfikowano następnie prawdziwość hipotezy stanowiącej, że frakcje równoczesnych wystąpień jedynek dla każdej pary zmiennych są równe odpowiednim elementom macierzy **M**. Uzyskane *p*-wartości testu dla obustronnego obszaru krytycznego wynoszą odpowiednio:

$$\mathbf{P}_{\text{LWH}} = \begin{bmatrix} 0.2244122 & 0.0571643 & \mathbf{0.0346601} \\ 0.0571643 & 0.6671444 & \mathbf{0.0004330} \\ \mathbf{0.0346601} & \mathbf{0.0004330} & 0.1361553 \end{bmatrix} \tag{19}$$

dla procedury LWH oraz

$$\mathbf{P}_{\text{PPS}} = \begin{bmatrix} 0.6557524 & 0.8509721 & 0.7379097 \\ 0.8509721 & 0.7890205 & 0.5304156 \\ 0.7379097 & 0.5304156 & 0.4591870 \end{bmatrix}. \tag{20}$$

Tak więc dla procedury PPS nie wykryto istotnych różnic, natomiast dla procedury LWH zarejestrowane częstości łącznych wystąpień jedynek różnią się istotnie od pożądanych dla pierwszej i trzeciej zmiennej, a także dla drugiej i trzeciej zmiennej. Aby to zjawisko dokładniej zilustrować, na rys. 1 przedstawiono wartości empirycznych ocen $\breve{r}_{12}, \breve{r}_{13}, \breve{r}_{23}$ współczynników korelacji liniowej Pearsona między poszczególnymi zmiennymi obliczane w kolejnych iteracjach eksperymentu symulacyjnego dla wszystkich wygenerowanych wcześniej danych. Wykres obejmuje trzy serie danych dla iteracji 4, 5, 6, ...1 800 000, czyli łącznie ponad 5 milionów punktów (dane z pierwszych trzech iteracji pominięto ze względu na zerowe wartości odchyleń standardowych zmiennych). Zaznaczono też pożądane wartości współczynników korelacji odpowiednio równe $r_{12} = 0.2$, $r_{13} = 0.3$, $r_{23} = 0.4$.



Rys. 1. Współczynniki korelacji $\breve{r}_{12}, \breve{r}_{13}, \breve{r}_{23}$ między zmiennymi binarnymi generowanymi za pomocą procedury LWH, wyznaczane w kolejnych iteracjach eksperymentu

Źródło: opracowanie własne.

Wykres na rys. 1 wyraźnie wskazuje na to, że współczynniki korelacji rejestrowane w kolejnych iteracjach bardzo szybko stabilizują się na poziomie wyraźnie różnym od zadanych wartości parametrów, co jest najbardziej widoczne dla ciągu ocen $\breve{r}_{23}$, znacząco większych niż zadana wartość 0.4.

4. Podsumowanie

Przedstawione wyniki symulacji wskazują, że procedura PPS generuje ciągi wektorów pseudolosowych o parametrach rozkładu, które nie różnią się istotnie od zadanych wartości. W przypadku procedury LWH stwierdzono natomiast istotne różnice między niektórymi zaobserwowanymi parametrami rozkładu empirycznego (współczynnikami korelacji) i ich zadanymi wartościami. Wydaje się, że zniekształcenia te mogą być konsekwencją numerycznego wyznaczania wartości całek niewłaściwych przez tę procedurę. Efekt taki zaobserwowano także po ponowieniu symulacji dla tej samej procedury w połączeniu z innymi kombinacjami zadanych parametrów rozkładu, choć jego nasilenie i kierunek obserwowanych zniekształceń były zmienne. Dla procedury PPS w powtórzonych dla innych wartości parametrów symulacjach nie obserwowano tak silnych odstępstw od zadanych wartości parametrów.

Warto dodać, że przy generowaniu długich ciągów wektorów o jednakowym rozkładzie prawdopodobieństwa procedura LWH dostępna darmowo w pakiecie R charakteryzuje się kilkakrotnie większą wydajnością od testowanej implementacji procedury PPS, co pozwala realizować obliczenia w znacznie krótszym czasie. Wynik porównania czasu pracy algorytmów zależy od rozmaitych uwarunkowań technicznych, należy go więc traktować z dużą ostrożnością. Szczególnie zaś spostrzeżenia co do różnicy wydajności algorytmów mogłyby się okazać mylące w wypadku generacji wektorów pseudolosowych o każdorazowo różnych parametrach rozkładów.

Z praktycznego punktu widzenia istotnym ograniczeniem procedury PPS jest brak możliwości generowania wektorów o składowych skorelowanych ujemnie. Od ograniczenia tego wolna jest procedura LWH.

Literatura

Bahadur R.R., *A representation of the joint distribution of responses to n dichotomous items*, [w:] *Studies in Item Analysis and Prediction*, Stanford Mathematical Studies in the Social Sciences VI, Ed. H. Solomon, Stanford University Press, Stanford 1961, s. 158-168.

Emrich L.J., Piedmonte M.R., *A method for generating high-dimensional multivariate binary variables*, „American Statistician" 1991, vol. 45, s. 302-304.

Gange S.J., *Generating multivariate cathegorical variates using the iterative proportional fitting algorithm*, „American Statistician" 1995, vol. 49, s. 134-138.

Goldberger A.S., *Teoria ekonometrii*, PWE, Warszawa 1975.

Lee A.J., *Generating random binary deviates having fixed marginal distributions and specified degrees of association*, „American Statistician" 1993, 47, s. 209-215.

Leisch F., Weingessel A., Hornik K., *On the generation of correlated artificial binary data*, Working Paper Series SFB, Adaptive Information Systems and Modelling in Economics and Management Science, Vienna University of Economics 1998.

Park C.G., Park T., Shin D.W., *A simple method for generating correlated binary variates*, „The American Statistician" 1996, 50(4), s. 306-310.

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

COMPARISON OF SOME PROCEDURES FOR GENERATING MULTIVARIATE PSEUDO-RANDOM BINARY DATA

Summary: This paper deals with the problem of generating sequences of correlated pseudo-random binary numbers. Two popular procedures proposed by Leisch et al and by Park et al are considered. Results of a simulation study aimed at assessing their consistency are presented. It is shown that significant discrepancies between desired parameters and parameters of the empirical distribution of random vectors occur for the first procedure.