Grzegorz Kończak
Magdalena Chmielińska
Uniwersytet Ekonomiczny w Katowicach

ZASTOSOWANIE METOD SYMULACYJNYCH W ANALIZIE WIELOWYMIAROWYCH TABLIC WIELODZIELCZYCH

Wprowadzenie

W ostatnich latach w badaniach naukowych znacznie wzrosła rola metod symulacji komputerowej. Początki tych metod sięgają lat 40. XX w. Właśnie wtedy zespół pracujący w amerykańskiej bazie wojskowej w Los Alamos kierowany przez J. von Neumana zastosował metody Monte Carlo do symulacji funkcjonowania elektrowni jądrowych oraz wybuchów atomowych. Ze względu na brak możliwości zastosowania złożonych technik obliczeniowych dynamiczny wzrost zastosowań tych metod przypada jednak dopiero na końcowe lata XX w. Do najczęściej wykorzystywanych metod symulacyjnych w badaniach statystycznych należy zaliczyć bootstrap, algorytmy Gibbsa i Metropolisa oraz próbkowanie ważone. Wszystkie wymienione metody zasadniczo są związane z analizą danych o charakterze ilościowym. Stosunkowo rzadko metody symulacyjne są wykorzystywane do analizy danych o charakterze jakościowym, czyli danych rejestrowanych na skalach nominalnych lub porządkowych. W opracowaniu przedstawiono propozycję metody symulacyjnej pozwalającej na zbadanie zależności dla grupy zmiennych nominalnych. Proponowana metoda wykorzystuje test permutacyjny. Przeprowadzono również porównania otrzymanych rezultatów z wynikami klasycznego testu niezależności chi-kwadrat. W przedstawionych przykładach zastosowano proponowaną metodę do analizy danych pochodzących z Polskiego Generalnego Sondażu Społecznego[1].

[1] Badania prowadzone od 1992 r. w Instytucie Studiów Społecznych Uniwersytetu Warszawskiego (http://pgss.iss.uw.edu.pl).

1. Wnioskowanie statystyczne dla danych w tablicach wielodzielczych

Załóżmy, że badaniu poddano *n* jednostek ze względu na dwie cechy nominalne *X* i *Y* przyjmujące odpowiednio *r* oraz *s* wartości (wariantów cech). Wyniki takiego badania mogą zostać zapisane w tablicy kontyngencji o *r* wierszach oraz *s* kolumnach. Do sprawdzenia hipotezy o niezależności tych zmiennych można wykorzystać statystykę [Aczel 2000]:

$$\chi^2 = \sum_{i=1}^{r}\sum_{j=1}^{s}\frac{(O_{ij} - E_{ij})^2}{E_{ij}}, \tag{1}$$

gdzie:
O_{ij} – liczebności obserwowane,
E_{ij} – liczebności oczekiwane.

Na podstawie otrzymanych danych dwuwymiarowych można skonstruować tablicę wielodzielczą. Przykład tablicy dla danych pochodzących z badania Polski Generalny Sondaż Społeczny przedstawia tabela 1. Przedstawiono w niej rozkład odpowiedzi na pytanie „jak ważna jest własna rodzina i dzieci?" w zależności od stanu cywilnego ankietowanych. Cała tabela składa się z 35 komórek. W tabeli 1 dodatkowo szarym tłem wyróżniono komórki tablicy, dla których liczebności oczekiwane są mniejsze niż 5.

Statystyka (1) przy założeniu niezależności zmiennych *X* i *Y* ma asymptotycznie rozkład chi-kwadrat o $(r-1)\cdot(s-1)$ stopniach swobody. Wnioskowanie statystyczne jest uzasadnione, jeśli liczebności oczekiwane są nie mniejsze niż 5 we wszystkich komórkach tablicy wielodzielczej [Blalock 1975]. W tabeli 1 występuje aż 14 komórek o liczebnościach oczekiwanych mniejszych od 5. Jest to istotną przeszkodą w wykorzystaniu testu niezależności chi-kwadrat. W takich przypadkach zwykle łączy się wybrane klasy. Połączenie wierszy „rozwiedziony" oraz „separacja" zmniejszy rozmiar tablicy wielodzielczej i liczbę komórek z liczebnościami oczekiwanymi mniejszymi od 5 do 8. Wyróżnienie dwóch kategorii wagi rodziny: „nieważna" (wskazania 1-4) oraz „ważna" (wskazania 5-7) pozwoliłoby na uniknięcie komórek z liczebnościami oczekiwanymi mniejszymi od 5. Takie rozwiązanie jest jednak związane ze stratą informacji o natężeniu roli przypisywanej przez respondentów rodzinie.

Tabela 1

Jak ważna jest własna rodzina i dzieci w zależności od stanu cywilnego respondenta

Stan cywilny	Jak ważna jest własna rodzina i dzieci?						
	nieważne						bardzo ważne
	1	2	3	4	5	6	7
zamężna/żonaty/konkubinat	13	5	3	30	54	205	5609
wdowiec/wdowa	5	1	2	8	16	58	997
rozwiedziony(a)	3	0	3	6	15	21	301
separacja	1	0	1	2	2	4	80
kawaler/panna	20	6	12	38	71	152	1043

Źródło: Na podstawie danych z PGSS.

Powyższe rozważania można rozszerzyć na większą niż dwie liczbę zmiennych X_1, X_2, ..., X_h. W takim przypadku mamy do czynienia z tablicami wielodzielczymi wielowymiarowymi. W przypadku trzech zmiennych nominalnych ($h = 3$) przyjmujących odpowiednio r, s oraz t wartości tablica wielodzielcza jest faktycznie kostką trójwymiarową. W tym przypadku statystyka testowa przyjmie następującą postać [Sheskin 2004]:

$$\chi^2 = \sum_{i=1}^{r}\sum_{j=1}^{s}\sum_{k=1}^{t}\frac{(O_{ijk} - E_{ijk})^2}{E_{ijk}}, \tag{2}$$

gdzie:
O_{ijk} – liczebności obserwowane,
E_{ijk} – liczebności oczekiwane.

Dla niezależnych zmiennych statystyka (2) ma asymptotyczny rozkład chi-kwadrat o $rst - r - s - t + 2$ stopniach swobody. Zapis tablicy dla trzech wymiarów nie jest już tak naturalny jak w przypadku tablicy dwuwymiarowej. Dla danych zamieszczonych w tabeli 1 uwzględniając dodatkowo zmienną „płeć" odpowiednia tablica wielodzielcza miałaby dwie warstwy. Poszczególne warstwy takiej tablicy mogą zostać przedstawione w oddzielnych tablicach dwuwymiarowych. Wyniki można jednak również przedstawić w formie jak w tabeli 2.

W miarę wzrostu liczby zmiennych h i liczby kategorii w_i (i = 1, 2, ..., h) dla zmiennych coraz trudniej zapewnić spełnienie warunku, aby liczebności oczekiwane w każdej komórce tabeli wynosiły przynajmniej 5. W tablicy wielodzielczej uwzględniającej płeć respondenta (por. tabela 2) jest 70 komórek. Ponad połowa z nich (38 komórek) ma liczebności oczekiwane mniejsze od 5. Biorąc pod uwagę fakt, że badana próba liczy 8787 osób łatwo zauważyć, że dla mniejszych prób często praktycznie nie będzie możliwości odwołania się do testu niezależności chi-kwadrat.

Tabela 2

Tablica wielodzielcza dla trzech zmiennych klasyfikujących

Jak ważna jest własna rodzina i dzieci w zależności od stanu cywilnego i płci respondenta

Stan cywilny	Jak ważna jest własna rodzina i dzieci?													
	Kobieta							Mężczyzna						
	nieważna				bardzo ważna			nieważna				bardzo ważna		
	1	2	3	4	5	6	7	1	2	3	4	5	6	7
zamężna/ żonaty/ konkubinat	11	2	2	13	17	80	2915	2	3	1	17	37	125	2694
wdowiec /wdowa	3	1	2	7	10	46	841	2	0	0	1	6	12	156
rozwie-dziony(a)	1	0	1	2	8	8	222	2	0	2	4	7	13	79
separacja	0	0	0	0	1	1	58	1	0	1	2	1	3	22
kawaler/ panna	9	4	8	15	27	58	501	11	2	4	23	44	94	542

Źródło: Na podstawie danych z PGSS.

Jeżeli dla każdej ze zmiennych liczba kategorii jest jednakowa i wynosi w ($w_1 = w_2 = ... = w_h = w$), to liczba komórek w wielowymiarowej tablicy wielodzielczej wynosi w^h. Biorąc pod uwagę, że w każdej z tych kratek liczebność oczekiwana powinna wynosić przynajmniej 5, to minimalna liczebność próby wynosi $5w^h$. W tabeli 3 przedstawiono przykładowe wartości minimalnych liczebności próby dla ustalonej ilości zmiennych i wariantów dla każdej cechy.

Tabela 3

Minimalne liczebności dla wybranych rozmiarów tablic

Liczba zmiennych h	Liczba wariantów każdej zmiennej w			
	2	3	4	5
2	20	45	80	125
3	40	135	320	625
4	80	405	1280	3125
5	160	1215	5120	15625

Minimalne liczebności przedstawione w tabeli 3 dotyczą szczególnego przypadku, gdy realizacja każdego z wariantów dla wszystkich zmiennych jest jednakowo prawdopodobna. Jeżeli prawdopodobieństwa realizacji dla różnych kategorii nie są jednakowe, to minimalne liczebności próby będą większe niż przedstawione w tabeli 3. Na podstawie danych zamieszczonych w tabeli 3 widoczne jest, że już dla kilku zmiennych przy paru różnych wariantach każdej z cech praktycznie niemożliwe jest przeprowadzenie badania oraz otrzymania uogólnienia na całą populację poprzez bezpośrednie wykorzystanie testu niezależności chi-kwadrat.

Ze względu na potrzeby praktyczne analizy danych w tablicach wielodzielczych przy niespełnieniu warunku na liczebność oczekiwaną zaproponowano różne modyfikacje. W dużej mierze modyfikacje te dotyczą przypadku tablic o wymiarach 2 x 2. Jeżeli nie wszystkie liczebności oczekiwane wynoszą przynajmniej 5, to można wykorzystać modyfikacje statystyki chi-kwadrat uwzględniające poprawki na ciągłość Yatesa lub poprawkę Dandekara [Rao 1982]. Statystyka chi-kwadrat z poprawką Yatesa ma postać:

$$\chi^2 = \sum_{i=1}^{2}\sum_{j=1}^{2} \frac{\left(\left|O_{ij} - E_{ij}\right| - 0{,}5\right)^2}{E_{ij}}. \tag{3}$$

Postać statystyki zaproponowanej przez Dandekara jest następująca:

$$\chi_c^2 = \chi_0^2 - \frac{\chi_0^2 - \chi_{-1}^2}{\chi_1^2 - \chi_{-1}^2}(\chi_1^2 - \chi_0^2), \tag{4}$$

gdzie χ_0^2, χ_1^2, χ_{-1}^2 oznaczają wartości statystyki (1) wyznaczone po dodaniu odpowiednio 0, 1 lub -1 do liczebności n_{11} (pierwszy wiersz i pierwsza kolumna tablicy wielodzielczej).

Powyższe korekty nie powinny być jednak stosowane w przypadkach, gdy w więcej niż jednej komórce tablicy wielodzielczej liczebności oczekiwane są mniejsze od 5. Wyjściem w takich sytuacjach jest zastosowanie testu dokładnego Fishera [Agresti 1996]. Test ten, podobnie jak powyższe modyfikacje, dotyczy tablic o wymiarach 2 x 2. W tym teście są obliczane prawdopodobieństwa wystąpienia wszystkich możliwych układów liczebności w tablicy wielodzielczej przy zachowaniu ustalonych liczebności brzegowych. Rozważmy przypadek otrzymania wyników badania jak w tabeli 4.

Tabela 4

Przykład danych w tablicy o wymiarach 2*x*2

Zmienna X	Zmienna Y	
	Y_1	Y_2
X_1	4	0
X_2	0	4

Wprowadźmy oznaczenia dla poszczególnych liczebności w komórkach tabeli 4: $a = 4$, $b = 0$, $c = 0$, $d = 4$. Niech ponadto $r_1 = 4$, $r_2 = 4$ oznaczają liczebności brzegowe w wierszach, a $c_1 = 4$ oraz $c_2 = 4$ liczebności brzegowe w kolumnach. Przy założeniu ustalonych powyższych liczebności brzegowych wszystkich możliwych tablic wielodzielczych jest 5 (por. rys. 1).

4	0
0	4

3	1
1	3

2	2
2	2

1	3
3	1

0	4
4	0

Rys. 1. Wszystkie tablice wielodzielcze o liczebnościach brzegowych $r1 = r_2 = c_1 = c_2 = 4$

Uwzględniając jednak, które elementy zostały zakwalifikowane do poszczególnych komórek tablicy wielodzielczej liczbę wszystkich możliwych układów tworzących taką tablicę, można wyrazić następująco:

$$K = \sum_{x=0}^{4} \binom{4}{x} \cdot \binom{4}{4-x} = 70.$$

Prawdopodobieństwo wystąpienia realizacji tablicy wielodzielczej o kolejnych liczebnościach a, b, c i d w komórkach wyraża się wzorem:

$$p = \frac{r1!r2!c1!c2!}{n!a!b!c!d!}. \tag{5}$$

Wszystkie możliwe wartości liczebności a (przy ustalonym a i danych liczebnościach brzegowych wartości b, c i d są wyznaczone jednoznacznie) dla zmiennych niezależnych z przykładu z tabeli 4 przy zachowaniu liczebności brzegowych podaje tabela 5.

Tabela 5

Prawdopodobieństwa i wartości statystyki chi-kwadrat

a	p	χ^2
4	0,014	8,00
3	0,229	2,00
2	0,514	0,00
1	0,229	2,00
0	0,014	8,00

Wysokie wartości statystyki χ^2 świadczą przeciw hipotezie o niezależności zmiennych. Dla przedstawionego w tabeli 4 przypadku wartość statystyki χ^2 wynosi 8. Przy założeniu niezależności zmiennych prawdopodobieństwo wystąpienia tak dużej lub większej wartości wynosi 0,028. Dla poziomu istotności $\alpha = 0{,}05$ należy odrzucić hipotezę o niezależności zmiennych.

2. Weryfikacja hipotezy o niezależności h ($h > 2$) zmiennych nominalnych

Weryfikacja hipotezy o niezależności $h = 3$ zmiennych może być przeprowadzona z wykorzystaniem statystyki (2). W przypadku większej liczby zmiennych wzór (2) należy uzupełnić o odpowiednie sumy i indeksy liczebności. Po-

ważnym problemem jest zapewnienie odpowiednich liczebności oczekiwanych w komórkach nawet dla prób o stosunkowo dużych liczebnościach. W takich przypadkach skuteczna może być procedura wykorzystująca test permutacyjny [Good 2005]. Na podstawie wylosowanej próby jest obliczana wartość statystyki testowej T_0. Jako statystykę można przyjąć np. χ^2. Następnie zbiór danych jest N-1 razy permutowany i każdorazowo jest wyznaczana wartość statystyki T_i (i = 1, 2, ..., N-1). Wartość statystyki T_0 jest porównywana z kwantylem rzędu $1-\alpha$ empirycznego rozkładu statystyki T.

2.1. Procedura obliczeniowa

Przebieg procedury obliczeniowej zostanie przedstawiony dla przypadku tablicy trójwymiarowej. Wszystkie rozważania można w analogiczny sposób stosować do tablic h-wymiarowych. W obliczeniach zostanie wykorzystany test permutacyjny. Struktura analizowanych danych została schematycznie przedstawiona na rys. 2. X_{i1}, X_{i2}, …X_{in} oznaczają warianty zmiennej X. Odpowiednio zostały oznaczone warianty zmiennych Y i Z.

X	Y	Z
X_{i1}	Y_{i1}	Z_{i1}
X_{i2}	Y_{i2}	Z_{i2}
…	…	…
X_{in}	Y_{in}	Z_{in}

Rys. 2. Struktura danych wykorzystana w opisie algorytmu

Procedura algorytmu zastosowania testu permutacyjnego do analizy tablic wielodzielczych wielowymiarowych jest następująca:

1. Pobierana jest próbka losowa. Na podstawie próby losowej jest konstruowana tablica wielodzielcza.
2. Dla otrzymanej tablicy wielodzielczej jest obliczana wartość statystyki chi-kwadrat (2). Otrzymaną wartość oznaczmy przez T_0.
3. Dla pobranej próbki kolumny 2-3 (por. rys. 2) są losowo, niezależnie permutowane. Dla tak otrzymanej próby jest obliczana wartość statystyki chi-kwadrat.

4. Krok 3 jest wykonywany N razy. Otrzymujemy wartości statystyki T_1, T_2, ..., T_N.
5. Obliczana jest wartość *ASL* (ang. *achieving significance level*) [Efron i Tibshirani 1993]:

$$ASL = \frac{card(i : T_i \geq T_0)}{N}, \text{ gdzie } i = 0, 1, ..., N\text{-}1. \quad (6)$$

Jeżeli *ASL* jest mniejsze od przyjętego poziomu istotności α, to odrzucamy hipotezę H_0.

Przedstawiona procedura pozwala na weryfikację hipotez statystycznych dotyczących zależności pomiędzy pewną liczbą zmiennych nominalnych. Procedura symulacyjna pozwala na sprawne przeprowadzenie weryfikacji hipotezy nawet dla tablic wielodzielczych o bardzo dużych rozmiarach również w przypadku stosunkowo niewielkiej liczby obserwacji.

2.2. Wyniki analizy symulacyjnej

Dla porównania własności opisywanego testu permutacyjnego i klasycznego testu chi-kwadrat przeprowadzono analizy na podstawie danych pochodzących z Polskiego Generalnego Sondażu Społecznego (PGSS). Celem Polskiego Generalnego Sondażu Społecznego jest systematyczny pomiar trendów i skutków zmian społecznych w Polsce. Problematyka PGSS obejmuje badanie indywidualnych postaw, cenionych wartości, orientacji i zachowań społecznych, jak również pomiar zróżnicowania społeczno-demograficznego, zawodowego, edukacyjnego i ekonomicznego reprezentatywnych grup i warstw społecznych w Polsce [Cichomski et al. 2009]. Systematyczną analizę trendów społecznych umożliwia cykliczne powtarzanie badań, które zachowują porównywalne standardy metodologiczne i identyczne wskaźniki.

W tabeli 6 przedstawiono analizowane zbiory zmiennych oraz wartości statystyki chi-kwadrat. Zamieszczono również p-wartość dla testu chi-kwadrat. Wielkości te należy traktować wyłącznie orientacyjnie ze względu na niespełnienie założenia dotyczącego minimalnych wartości liczebności oczekiwanych w komórkach. W tabeli 6 przedstawiono również wartości *ASL* dla zastosowanego testu permutacyjnego.

Tabela 6

Charakterystyka analizowanych zmiennych i wyniki obliczeń

Lp.	Zmienne	Liczba komórek tablicy *rst*	Liczebność próby	Test chi-kwadrat		Test permutacyjny
				χ^2	*p* wartość	*ASL*
1	X_2, X_3	35	8787	459,7	*)	*)
2	X_1, X_2, X_3	70	8787	1046,5	*)	*)
3	X_4, X_5	30	15934	632,1	*)	*)
4	X_1, X_4, X_5	60	15924	734,0	*)	*)
5	X_6, X_7	40	2445	15,6	0,971	0,974
6	X_2, X_6, X_7	200	2435	193,7	0,297	0,318

*) – wartość mniejsza lub równa 0,001
X_1 – płeć (2 wartości: mężczyzna, kobieta)
X_2 – stan cywilny (5 wartości: żonaty/zamężna/konkubinat, wdowiec/wdowa, rozwiedziony(a), separacja, kawaler/panna)
X_3 – jak ważna jest własna rodzina i dzieci (7 wartości: 1 – nieważne, ..., 7 – bardzo ważne)
X_4 – zadowolenie z własnej sytuacji finansowej (3 wartości: 1 – zadowolony, 2 – mniej więcej zadowolony, 3 – niezadowolony)
X_5 – skala chęci do życia (10 wartości: 1 – „w ogóle nie chce mi się żyć", ..., 10 – „bardzo mocno chce mi się żyć")
X_6 – region zamieszkania (8 wartości)
X_7 – większy % podatku od bogatych (5 wartości: 1 – „znacznie większy %, ..., 5 – „znacznie mniejszy %")

Z analizy wyników przedstawionych w tabeli 6 można wysnuć wniosek, iż decyzje odnośnie do testowanej hipotezy o niezależności zmiennych przy zastosowaniu klasycznego testu niezależności chi-kwadrat (p-wartości) oraz testu permutacyjnego (wartości *ASL*) są w analizowanych przypadkach identyczne. Wyniki otrzymane z zastosowaniem testu niezależności chi-kwadrat nie mogą jednak być przedstawiane jako wiarygodne ze względu na niespełnienie założeń dotyczących minimalnych liczebności oczekiwanych w komórkach tablicy wielodzielczej. Metoda symulacyjna nie wymaga spełnienia takich założeń i może być stosowana nawet dla prób o niewielkich liczebnościach. Prezentowana metoda symulacyjna może okazać się szczególnie przydatna w przypadku wielo-

wymiarowych tablic wielodzielczych, gdzie wnioskowanie statystyczne o zależności pomiędzy badanymi zmiennymi na podstawie klasycznego testu niezależności chi-kwadrat może okazać się niezasadne ze względu na małe liczebności oczekiwane w komórkach tablicy.

Podsumowanie

Ograniczenia klasycznych metod statystycznych sprawiają, iż metody symulacyjne są coraz chętniej wykorzystywane w badaniach naukowych. Metody symulacyjne, mimo iż obecnie wykorzystywane są głównie do analizy danych o charakterze ilościowym, z powodzeniem mogą być również stosowane do analizy danych o charakterze jakościowym. Zaprezentowana w niniejszym artykule metoda oparta na teście permutacyjnym, pozwala na weryfikację hipotezy o niezależności dla grup zmiennych nominalnych. Zaletą omawianej metody symulacyjnej jest brak założeń dotyczących minimalnych liczebności oczekiwanych w komórkach tablicy wielodzielczej. Nabiera to szczególnego znaczenia w przypadku analizy zależności pomiędzy wieloma zmiennymi, z których każda posiada po kilka wariantów wartości. W takich przypadkach duża liczba zmiennych i kategorii dla zmiennych w praktyce uniemożliwia wnioskowanie o niezależności pomiędzy tymi zmiennymi za pomocą testu niezależności chi-kwadrat.

Literatura

Aczel A. (2000): *Statystyka w zarządzaniu*. Wydawnictwo Naukowe PWN, Warszawa.

Agresti A. (1996): *An Introduction to Categorical Data Analysis*. John Wiley & Sons, New York.

Blalock H.M. (1975): *Statystyka dla socjologów*. PWN, Warszawa.

Cichomski B., Jerzyński T., Zieliński M. (2009): *Polskie Generalne Sondaże Społeczne: struktura skumulowanych wyników badań 1992-2008*. Instytut Studiów Społecznych, Uniwersytet Warszawski, Warszawa,.

Efron B., Tibshirani R. (1993): *An Introduction to the Bootstrap*. Chapman & Hall, New York.

Good P. (2005): *Permutation, Parametric and Bootstrap Tests of Hypotheses*. Springer Science Business Media, New York.

Rao C.R (1982): *Modele liniowe statystyki matematycznej*. PWN, Warszawa.

Sheskin D.J. (2004): *Handbook of Parametric and Nonparametric Statistical Procedures*. Chapman & Hall/CRC, Boca Raton.

ON THE USE OF THE MONTE CARLO METHODS IN THE MULTIDIMENSIONAL CONTINGENCY TABLES ANALYSIS

Summary

Recently the role of computer simulation methods in scientific research has significantly increased. In this paper the proposal of the use simulation methods to test for independence of some categorical variables with contingency tables is presented. The permutation test is used in this method. The comparison of obtained results and results of classically chi-square test of independence has been done. In presented examples the data from Polish General Public Opinion Poll have been used.