Grzegorz Kończak

Uniwersytet Ekonomiczny w Katowicach

O TESTOWANIU ISTOTNOŚCI WSPÓŁCZYNNIKÓW KORELACJI CZĄSTKOWEJ I WIELORAKIEJ DLA WIELOWYMIAROWYCH TABLIC WIELODZIELCZYCH

Wprowadzenie

Do najważniejszych zagadnień rozważanych w badaniach statystycznych należy analiza zależności. Dla zmiennych rejestrowanych na mocnych skalach pomiarowych wykorzystuje się współczynnik korelacji liniowej Pearsona oraz różnej postaci funkcje regresji. Dla wykluczenia wpływu zmiennych zakłócających wyznacza się współczynniki korelacji cząstkowej, a dla określenia łącznego wpływu kilku zmiennych na zmienną zależną współczynniki korelacji wielorakiej. W przypadku pomiarów dokonanych na skalach słabych należy skorzystać z innych narzędzi. Dla pomiarów na skali porządkowej wykorzystuje się współczynniki korelacji rang Spearmana i Kendalla. W przypadku pomiarów na skali nominalnej najczęściej wykorzystuje się różne współczynniki oparte na obliczeniu statystyki chi-kwadrat.

W artykule przedstawiono propozycję wyznaczania współczynników zależności cząstkowej dla zmiennych określonych na skalach nominalnych. Ze względu na konstrukcję współczynnika korelacji cząstkowej dla danych nominalnych, a w szczególności trudności w określeniu rozkładu estymatora tego współczynnika, zastosowano testy permutacyjne do weryfikacji hipotezy o istotności tych zależności.

1. Zależność dla zmiennych rejestrowanych na skalach nominalnych

W przypadku, gdy badaniem objęte są dwie zmienne X i Y przyjmujące wartości na skalach nominalnych, właściwym podejściem jest zastosowanie analiz związanych z tablicami kontyngencji, określanymi również jako tablice wielodzielcze. Jeśli warianty zmiennej X oznaczymy przez $x_1, x_2, \ldots, x_r$, a warianty zmiennej Y przez $y_1, y_2, \ldots, y_c$, gdzie r i c są odpowiednio liczbą wariantów zmiennych X i Y, to tablicę kontyngencji można przedstawić jak w tabeli 1.

Tabela 1

Układ danych w tablicy kontyngencji

Zmienna X	Zmienna Y				Sumy w wierszach
	y_1	y_2	…	y_c	
x_1	n_{11}	n_{12}	…	n_{1c}	$n_{1\bullet}$
x_2	n_{21}	n_{22}	…	n_{2c}	$n_{2\bullet}$
…	…	…	…	…	…
x_r	n_{r1}	n_{r2}	…	n_{rc}	$n_{r\bullet}$
Sumy w kolumnach	$n_{\bullet 1}$	$n_{\bullet 2}$	…	$n_{\bullet c}$	n

Wielkości n_{ij} ($i = 1, 2, \ldots, r$ oraz $j = 1, 2, \ldots, c$) są zaobserwowanymi liczebnościami realizacji jednocześnie x_i oraz y_j zmiennej dwuwymiarowej (X, Y).

Do analizy zależności pomiędzy zmiennymi X i Y zwykle wykorzystuje się różne mierniki, których konstrukcja opiera się na statystyce chi-kwadrat. Statystyka ta dla dwuwymiarowej tablicy wielodzielczej o wymiarach r x k przyjmuje postać:

$$\chi^2 = \sum_{i=1}^{r}\sum_{j=1}^{c}\frac{(n_{ij} - \hat{n}_{ij})^2}{\hat{n}_{ij}}, \qquad (1)$$

gdzie:

n_{ij} – liczebności obserwowane,

$\hat{n}_{ij} = \dfrac{n_{i\bullet} \cdot n_{\bullet j}}{n}$ – liczebności oczekiwane.

Statystyka (1) ma asymptotycznie rozkład chi-kwadrat o $(r - 1)(k - 1)$ stopniach swobody. Do testowania istotności zależności pomiędzy zmiennymi X i Y można wykorzystać wartości krytyczne z rozkładu chi-kwadrat, jeśli liczebności oczekiwane dla wszystkich komórek tabeli wynoszą przynajmniej 5 (por. np. Domański, 1990).

Statystyka (1) przyjmuje nieujemne wartości. Jest ona wykorzystywana do konstrukcji różnych współczynników, które przyjmują wartości z przedziału ograniczonego, co ułatwia interpretację poziomu zależności. Wzory (2) – (4) przedstawiają wybrane współczynniki siły zależności dla danych przedstawionych w tablicy wielodzielczej (Zeliaś et al., 2002).

Współczynnik kontyngencji *C* Pearsona:

$$C = \sqrt{\frac{\chi^2}{\chi^2 + n}} . \tag{2}$$

Współczynnik *V* Cramera:

$$V = \sqrt{\frac{\chi^2}{n \cdot \min(r-1, k-1)}} . \tag{3}$$

Współczynnik *T* Czuprowa:

$$T = \sqrt{\frac{\chi^2}{n \cdot \sqrt{(r-1)(k-1)}}} . \tag{4}$$

W dalszych rozważaniach będzie uwzględniony wyłącznie współczynnik (1), jednak wszystkie analizy mogą zostać rozszerzone na pozostałe przedstawione współczynniki zależności.

2. Pomiar zależności cząstkowych dla danych w wielowymiarowych tablicach wielodzielczych

J.H. Zar (2010) wskazuje na możliwość wyznaczania współczynników korelacji cząstkowej dla tablic wielodzielczych. Niech dana będzie tablica wielodzielcza skonstruowana na podstawie badania zależności pomiędzy trzema zmiennymi *X, Y* i *Z* przyjmującymi wartości na skalach nominalnych. Jeśli warianty zmiennej *X* oznaczymy przez $x_1, x_2, \ldots, x_r$, dla zmiennej *Y* przez $y_1, y_2, \ldots, y_c$, a dla zmiennej *Z* przez $z_1, z_2, \ldots, z_l$, gdzie *r*, *c* i *l* są odpowiednio liczbą wszystkich występujących wariantów zmiennych *X*, *Y* i *Z*, to wartość statystyki chi-kwadrat jest obliczana na podstawie wzoru:

$$\chi^2 = \sum_{i=1}^{r} \sum_{j=1}^{c} \sum_{k=1}^{l} \frac{(n_{ijk} - \hat{n}_{ijk})^2}{\hat{n}_{ijk}} , \tag{5}$$

gdzie:

n_{ijk} – liczebności obserwowane,

$\hat{n}_{ijk} = \frac{n_{i\bullet\bullet} \cdot n_{\bullet j\bullet} \cdot n_{\bullet\bullet k}}{n^2}$ – liczebności oczekiwane.

Przy założeniu niezależności zmiennych *X*, *Y* i *Z* statystyka (5) ma asymptotycznie rozkład chi-kwadrat o *rcl-r-c-l*+2 stopniach swobody (por. Sheskin, 2004).

Jeżeli hipoteza o niezależności nie jest odrzucona, to w konkluzji stwierdza się, że można przyjąć hipotezę o niezależności zmiennych. W przypadku odrzucenia hipotezy o niezależności zmiennych test nie informuje o występujących rodzajach zależności. Możliwe jest występowanie zależności pomiędzy wszystkimi zmiennymi, ale może występować zależność wyłącznie np. pomiędzy *X* i *Y*. W literaturze są rozważane różne możliwości odwołujące się do określenia siły zależności pomiędzy dwiema zmiennymi lub pomiędzy dwiema zmiennymi z wyłączeniem wpływu trzeciej zmiennej. Określenie siły takich zależności można zrealizować poprzez:

- zbadanie siły zależności pomiędzy x i y, x i z oraz pomiędzy y i z.
- obliczenie współczynników korelacji cząstkowej (por. Zar, 2010).

Tradycyjnie warianty zmiennych *X* i *Y* określa się jako „wiersze" i „kolumny", a jest to bezpośrednio związane z konstrukcją tablicy kontyngencji. D.J. Sheskin (2004) przyjmuje określenia wariantów zmiennej *Z* jako „warstwy". J.H. Zar (2010) proponuje wyznaczanie współczynników korelacji cząstkowej z wykorzystaniem modyfikacji obliczania liczebności oczekiwanych w komórkach tablicy wielodzielczej:

- Dla hipotezy, że wiersze są niezależne od łącznie kolumn i warstw

$$\hat{n}_{ijk} = \frac{n_{i\bullet\bullet} \cdot n_{\bullet jk}}{n} \text{ dla } i = 1, 2, \ldots, r, j = 1, 2, \ldots, c \text{ oraz } k = 1, 2, \ldots l.$$

  Liczba stopni swobody dla statystyki (5) wynosi:

  $v = (r - 1)(c - 1)(k - 1) + (r - 1)(c - 1) + (r - 1)(k - 1).$

- Dla hipotezy, że kolumny są niezależne od łącznie wierszy i warstw

$$\hat{n}_{ijk} = \frac{n_{\bullet j\bullet} \cdot n_{i\bullet k}}{n} \text{ dla } i = 1, 2, \ldots, r, j = 1, 2, \ldots, c \text{ oraz } k = 1, 2, \ldots l.$$

  Liczba stopni swobody dla statystyki (5) wynosi:

  $v = (r - 1)(c - 1)(k - 1) + (c - 1)(r - 1) + (c - 1)(k - 1).$

- Dla hipotezy, że warstwy są niezależne od łącznie wierszy i kolumn

$$\hat{n}_{ijk} = \frac{n_{\bullet\bullet k} \cdot n_{ij\bullet}}{n} \text{ dla } i = 1, 2, \ldots, r, j = 1, 2, \ldots, c \text{ oraz } k = 1, 2, \ldots l.$$

  Liczba stopni swobody dla statystyki (5) wynosi:

  $v = (r - 1)(c - 1)(k - 1) + (k - 1)(r - 1) + (k - 1)(c - 1).$

Występujące symbole $n_{i\bullet\bullet}, n_{\bullet j\bullet}, n_{\bullet\bullet k}$ oznaczają odpowiednio:

$$n_{i\bullet\bullet} = \sum_{j=1}^{c}\sum_{k=1}^{l} n_{ijk} \text{ , dla } i = 1, 2, \ldots, r$$

$$n_{\bullet j\bullet} = \sum_{i=1}^{r}\sum_{k=1}^{l} n_{ijk} \text{ , dla } j = 1, 2, \ldots, c$$

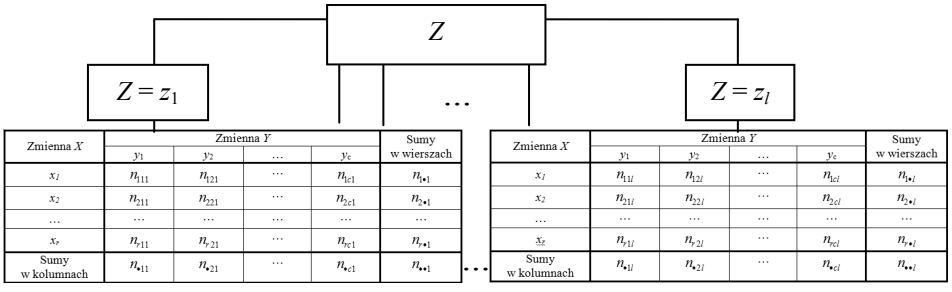
$$n_{\bullet\bullet k} = \sum_{i=1}^{r}\sum_{j=1}^{c} n_{ijk} \text{ , dla } k = 1, 2, \ldots, l.$$

Przedstawione wzory umożliwiają weryfikację hipotezy o łącznym wpływie dwóch zmiennych na trzecią. Ze względu na wykorzystanie rozkładu chi-kwadrat jest konieczne spełnienie założenia dotyczącego minimalnej liczebności oczekiwanej w komórkach tablicy wielodzielczej. W dalszej części opracowania przedstawiono inne możliwe podejście do testowania istotności występujących zależności cząstkowych dla danych w tablicy wielodzielczej. Prezentowane rozwiązanie odwołuje się do testu permutacyjnego (Good, 2005) i dlatego może być stosowane nawet w przypadku, gdy występują liczebności oczekiwane są mniejsze od 5.

3. Łączny wpływ dwóch zmiennych na trzecią zmienną

Weryfikacja hipotezy o niezależności 3 zmiennych może być przeprowadzona z wykorzystaniem statystyki (5). Takie podejście równoprawnie traktuje wszystkie trzy zmienne. W badaniach statystycznych często interesujący jest łączny wpływ kilku zmiennych na wyróżnioną zmienną oraz wyłączny wpływ określonej zmiennej (zmiennych) z pominięciem wpływu pozostałych zmiennych.

Niech będzie dana trójwymiarowa tablica wielodzielcza. Dane takie mogą być przedstawione w formie jak na rysunku 1.



Rys. 1. Zapis danych w trójwymiarowej tablicy kontyngencji

Zagadnienie badania łącznego wpływu zmiennych *X* i *Y* na zmienną Z (współczynnik korelacji wielorakiej) można formalnie zapisać za pomocą hipotez:

H_0: Brak łącznego wpływu zmiennych *X* i *Y* na zmienną *Z* (niezależność).

H_Z: Występuje zależność pomiędzy zmienną *Z* i zmiennymi *X* i *Y*.

Dla weryfikacji hipotezy H_0 wobec hipotezy alternatywnej H_Z nie może być bezpośrednio wykorzystana statystyka (5). Mogą w tym przypadku być wykorzystane wcześniej opisane współczynniki. Niech obliczona na podstawie wzoru (5) wartość statystyki będzie oznaczona przez T_0. W przypadku tablic wielowymiarowych, gdzie zmienne mogą przyjmować wiele wariantów, nie jest zazwyczaj spełniony warunek nałożony na liczebności oczekiwane w komórkach tablicy wielodzielczej. Nie ma w takich przypadkach możliwości skorzystania z wartości krytycznych wyznaczonych z rozkładu chi-kwadrat. Do przybliżenia rozkładu statystyki przy założeniu prawdziwości hipotezy H_0 można wykorzystać permutacje zmiennej *Z*. Ideę permutacji przedstawia rysunek 2

X	*Y*	*Z*
x_1	y_1	z_1
x_2	y_2	z_2
x_3	y_3	z_3
...	...	...
x_n	y_n	z_n

X	*Y*	*Z*
x_1	y_1	z_5
x_2	y_2	z_7
x_3	y_3	z_2
...	...	...
x_n	y_n	z_1

Rys. 2. Schemat permutowania zmiennej *Z* (po lewej zbiór wyjściowy, po prawej po jedna z możliwych permutacji zmiennej *Z*)

Jako współczynnik określający siłę zależności w dalszych rozważaniach może być dowolny z mierników (2) – (4), jak również statystyka (5). Niech współczynnik *T* zależności wyznaczony dla pierwotnych danych będzie oznaczony przez T_0. Dla każdej permutacji zmiennej *Z* jest obliczana wartość współczynnika T_i ($i = 1, 2, \ldots, N$). Takie postępowanie prowadzi do uzyskania empirycznego rozkładu statystyki *T* przy założeniu prawdziwości hipotezy H_0.

Dla podjęcia decyzji wykorzystuje się wartość *ASL* (Achieving Significance Level, empiryczna *p*-wartość, por. Efron, Tibshirani, 1993) zadaną wzorem:

$$ASL = P(T_i \geq T_0). \tag{6}$$

Wartość ta jest nieznana, a jej ocenę otrzymuje się na podstawie rozkładu empirycznego statystyki *T*:

$$\hat{ASL} = \frac{card(i : T_i \geq T_0)}{N}, \text{ gdzie } i = 0, 1, \ldots, N. \quad (7)$$

Jeżeli wartość *ASL* jest mniejsza od przyjętego poziomu istotności α, to hipoteza H_0 jest odrzucana na korzyść hipotezy alternatywnej H_Z. Podobne rozważania mogą być przeprowadzone dla odpowiednio sformułowanej hipotezy H_0 i hipotez alternatywnych H_Y i H_X.

Procedurę weryfikacji przedstawionej hipotezy na podstawie testu permutacyjnego można zapisać następująco:

1. Pobierana jest próbka losowa. Na podstawie próby losowej jest konstruowana tablica wielodzielcza.
2. Dla otrzymanej tablicy wielodzielczej jest obliczana wartość statystyki *T*. Otrzymaną wartość oznaczmy przez T_0.
3. Dla pobranej próbki zmienna *Z* jest losowo permutowana. Dla tak otrzymanej próby jest obliczana wartość statystyki *T*.
4. Krok 3 jest wykonywany *N* razy. Otrzymujemy wartości statystyki $T_1, T_2, \ldots, T_N$.
5. Obliczana jest wartość *ASL*.

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

3. Przykład empiryczny

Ideę zastosowania proponowanej metody przedstawiono na poniższym przykładzie. Dane o współwystępowaniu trzech zmiennych *X*, *Y* i *Z* zaprezentowano na rysunku 3. Testowanie istotności zależności cząstkowych z wykorzystaniem klasycznych metod dla tych danych przedstawia D.J. Sheskin (2004). Wyniki przeprowadzonych testów permutacyjnych przedstawiono na rysunku 4. Można na nim również znaleźć empiryczne rozkłady statystyki testowej *T* otrzymane na podstawie przeprowadzonych permutacji, a także wartość statystyki T_0.

Warstwa $Z = z_1$

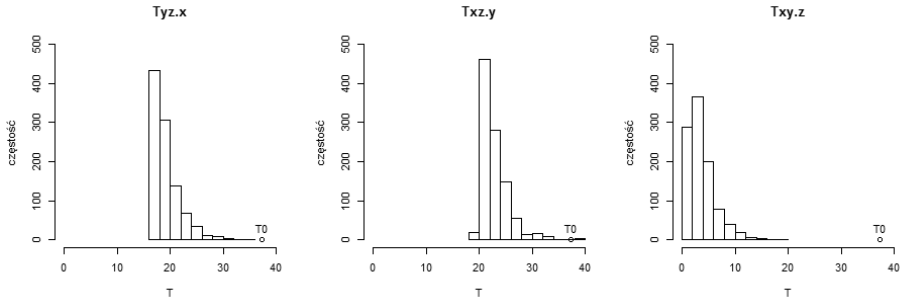
Zmienna *X*	Zmienna *Y*		Suma
	y_1	y_2	
x_1	10	15	25
x_2	25	45	70
Suma	35	60	95

Warstwa $Z = z_2$

Zmienna *X*	Zmienna *Y*		Suma
	y_1	y_2	
x_1	25	15	40
x_2	20	5	25
Suma	45	20	65

Rys. 3. Dane do przykładu empirycznego

Źródło: Na podstawie Sheskin (2004).



Rys. 4. Wyniki testu permutacyjnego

We wszystkich przeprowadzonych testach permutacyjnych przyjęto poziom istotności $\alpha = 0{,}05$. Przeprowadzenie testu permutacyjnego dla wszystkich możliwych przypadków łącznego wpływu dwóch ustalonych zmiennych na trzecią prowadzi do odrzucenia hipotezy H_0 przy przyjętym poziomie istotności α. Wartości *ASL* dla hipotez o niezależności zmiennych *X* oraz *Y* i *Z* łącznie, a także *Z* oraz *X* i *Y* łącznie wyniosła 0. W przypadku testowania hipotezy o niezależności zmiennej *Y* i zmiennych *X* i *Z* łącznie otrzymano $ASL = 0{,}002$. Dla wszystkich rozważanych przypadków został potwierdzony łączny wpływ dwóch zmiennych na pozostałą zmienną.

Podsumowanie

W analizie zależności szczególne miejsce zajmuje badanie siły wpływu pomiędzy zmiennymi na skalach nominalnych. Zwyczajowo takie dane przedstawiane są w tablicach wielodzielczych. Klasyczne metody takiej analizy wymagają spełnienia założenia dotyczącego minimalnej liczebności oczekiwanej w komórkach tablicy.

W opracowaniu przedstawiono propozycję testowania istotności wpływu ustalonej zmiennej na pozostałe w przypadku analizy trójwymiarowych tablic wielodzielczych. Ze względu na zastosowanie testu permutacyjnego nie jest konieczna znajomość rozkładu statystyki testowej, a weryfikację hipotezy można przeprowadzić nawet wówczas, gdy występują małe liczebności oczekiwane w komórkach tablicy.

Podziękowanie

Projekt został sfinansowany ze środków Narodowego Centrum Nauki przyznanych na podstawie decyzji numer DEC-2011/03/B/HS4/05630.

Literatura

Aczel A. (2000), Statystyka w zarządzaniu, WN PWN, Warszawa.

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

Domański Cz. (1990), Testy statystyczne, PWE, 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.

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

Zar J.H. (2010), Biostatical Analysis, Pearson Education, New Jersey.

Zeliaś A., Pawełek B., Wanat S. (2002), Metody statystyczne, PWE, Warszawa.

ON TESTING PARTIAL DEPENDENCY FOR DATA IN CONTINGENCY TABLES

Summary

The chi-square test of independence is used for data presented in contingency tables. The three dimensional contingency tables are analyzed in the paper. If the independence test leads to a significant result, then a researcher should conduct additional analysis to clarify the nature of the relationship between the three variables. The proposal of the partial independence test for data in contingency tables is presented in the paper. The proposal is based on the permutation test.