

Lechosław Stępień*

UOGÓLNIENIE STATYSTYKI χ^2
W WIELOWYMIAROWEJ ANALIZIE STATYSTYCZNEJ

1. Uwagi wstępne

W teorii statystyki oraz w jej zastosowaniach bardzo ważną rolę odgrywa χ^2 . Z nazwą tą wiąże się na ogół trzy różne terminy:

- rozkład χ^2 ,
- statystyka χ^2 , a więc zmienna losowa posiadająca rozkład χ^2 ,
- test χ^2 , jako metoda sprawdzenia hipotezy zgodności rozkładów lub niezależności zmiennych, wykorzystująca jako sprawdzian statystykę χ^2 o rozkładzie chi-kwadrat.

Rozkład χ^2 został po raz pierwszy znaleziony przez Hermerta w 1876 r. [5], a następnie niezależnie przez Pearsona w 1900 r. [6]. Teoria dotycząca rozkładu χ^2 oraz konstrukcji statystyk podlegających rozkładowi χ^2 jest na ogół opracowana w sposób wyczerpujący. Znaleźć ją można niemal we wszystkich znaczących publikacjach z zakresu rachunku prawdopodobieństwa i statystyki matematycznej. Nie jest celem niniejszego artykułu jej powielanie.

Rozpatrując natomiast problem weryfikacji hipotez nieparametrycznych przy wykorzystaniu statystyki (sprawdzianu) χ^2 , wyróżniamy trzy rodzaje testów istotności:

- test zgodności χ^2 ,

* Dr, adiunkt w Instytucie Ekonometrii i Statystyki UŁ.

- test χ^2 , jako test jednorodności prób,
- test niezależności χ^2 .

Pierwszy z nich - test zgodności χ^2 - służy sprawdzeniu czy badana cecha ma określony rozkład (np. rozkład normalny, dwumianowy ...). Dwuwariantową - przy znajomości lub nieznanomości parametrów rozkładu - konstrukcję tego testu znaleźć można m. in. w [2], [7].

Drugi rodzaj testów χ^2 - test jednorodności prób - służy sprawdzeniu hipotezy, że s prób jest pobrane z tej samej populacji, tzn. że są pod tym względem jednorodne.

Trzeci rodzaj, to testy niezależności χ^2 . Ich omówienie będzie m. in. tematem niniejszego opracowania. Zajmiemy się szczególnie problemem badania niezależności cech w wielowymiarowej analizie statystycznej.

2. Test niezależności w dwuwymiarowej tablicy kontyngencyjnej

Weryfikację hipotezy o niezależności cech w dwuwymiarowej tablicy kontyngencyjnej przeprowadza się na podstawie testu niezależności χ^2 . W literaturze test ten omawiany jest na ogół dla jednego tylko przypadku - nieznanymi rozkładów brzegowych $\pi_{i.}$ oraz $\pi_{.j}$. Należy jednak, jak potwierdzają to rozważania zawarte w [9], rozróżnić dwa przypadki:

- prawdopodobieństwa teoretyczne $\pi_{i.}$ oraz $\pi_{.j}$ w rozkładach brzegowych są znane,
- prawdopodobieństwa teoretyczne w rozkładach brzegowych są nieznane.

W pierwszym przypadku sumaryczną rozbieżność pomiędzy częstościami zaobserwowanymi i częstościami oczekiwanymi mierzymy za pomocą statystyki

$$(1) \quad \chi^2 = \sum_{i=1}^r \sum_{j=1}^s \frac{(n_{ij} - n_{..} \pi_{ij})^2}{n_{..} \pi_{ij}},$$

przy liczbie stopni swobody $rs - 1$.

Wyodrębnić tu możemy dwa składniki:

$$(2) \quad \chi_A^2 = \sum_{i=1}^r \frac{(n_{i.} - n_{..} \pi_{i.})^2}{n_{..} \pi_{i.}}$$

z liczbą stopni swobody $r - 1$ oraz

$$(3) \quad \chi_B^2 = \sum_{j=1}^s \frac{(n_{.j} - n_{..} \pi_{.j})^2}{n_{..} \pi_{.j}}$$

z liczbą stopni swobody $s - 1$.

Mierzą one rozbieżność pomiędzy zaobserwowanymi i oczekiwanymi częstościami brzegowymi. Za pomocą statystyk χ_A^2 oraz χ_B^2 zwerifikować można hipotezę zgodności dla rozkładów brzegowych. Pomniejszając wartość χ^2 daną wzorem (1) o χ_A^2 oraz χ_B^2 , otrzymujemy odchylenie od niezależności [9]:

$$(4) \quad \chi_{AB}^2 = \chi^2 - \chi_A^2 - \chi_B^2 =$$

$$= \sum_{i=1}^r \sum_{j=1}^s \frac{(n_{ij} - n_{i.} \pi_{.j} - n_{.j} \pi_{i.} + n_{..} \pi_{i.} \pi_{.j})^2}{n_{..} \pi_{i.} \pi_{.j}}$$

Aby ustalić liczbę stopni swobody wykorzystamy ogólne twierdzenie ([9], s. 203).

Twierdzenie. Jeżeli $Q = Q_1 + Q_2$, $Q \sim \chi^2(a)$, $Q_1 \sim \chi^2(b)$ oraz Q jest nieujemne to $Q_2 \sim \chi^2(a - b)^*$.

Ponieważ statystyka χ^2 ma $rs - 1$ stopni swobody, statystyki χ_A^2 , χ_B^2 odpowiednio $r - 1$, $s - 1$ stopni swobody, więc statystyka χ_{AB}^2 ma na mocy rozszerzonej wersji tego twierdzenia

$$(rs - 1) - (r - 1) - (s - 1) = (r - 1)(s - 1)$$

stopni swobody.

* Zapis $Q \sim \chi^2(a)$ oznacza, że zmienna losowa Q ma rozkład χ^2 o a stopniach swobody.

Statystyka (4) może wobec tego służyć jako sprawdzian przy weryfikacji hipotezy o niezależności cech przy znanych rozkładach brzegowych.

Drugi przypadek to badanie niezależności przy nieznanych rozkładach brzegowych. Trzeba je oszacować stosując estymatory największej wiarygodności:

$$\hat{\pi}_{i.} = \frac{n_{i.}}{n_{..}} \quad \text{oraz} \quad \hat{\pi}_{.j} = \frac{n_{.j}}{n_{..}}.$$

Dowód na to, że $\hat{\pi}_{i.}$, $\hat{\pi}_{.j}$ są estymatorami największej wiarygodności przeprowadzony jest m. in. w ([3], s. 474).

Uwzględniając te oszacowania otrzymamy wzór na χ^2 w następującej postaci:

$$(5) \quad \chi^2_B = \sum_{i=1}^r \sum_{j=1}^s \frac{(n_{ij} - \frac{n_{i.} \cdot n_{.j}}{n_{..}})^2}{\frac{n_{i.} \cdot n_{.j}}{n_{..}}}.$$

Ponieważ z próby wyznaczyliśmy $r + s - 2$ parametry, przeto statystyka (5) ma rozkład χ^2 o

$$(rs - 1) - (r + s - 2) = (r - 1)(s - 1)$$

stopniach swobody.

3. Test niezależności χ^2

w tablicach kontyngencyjnych trójwymiarowych

W niektórych badaniach statystycznych, szczególnie z zakresu nauk społecznych czy przyrodniczych, badacz interesuje czy istnieje łączna współzależność trzech cech statystycznych A, B, C - mających odpowiednio r, s, t wariantów. Należy wówczas zweryfikować hipotezę o niezależności między trzema cechami statystycznymi.

Podobnie jak w przypadku badania niezależności dwóch cech statystycznych, należy wyróżnić dwie sytuacje:

- brzegowe rozkłady $\pi_{ij.}$, $\pi_{.jk}$, $\pi_{i.k}$ oraz $\pi_{i..}$, $\pi_{.j.}$, $\pi_{...k}$ są znane,
- brzegowe rozkłady nie są znane.

Rozpatrzmy pierwszą z nich. Statystyka χ^2 mierzącą sumaryczną rozbieżność między częstościami zaobserwowanymi i częstościami oczekiwanymi ma postać:

$$(6) \chi^2 = \sum_{i=1}^r \sum_{j=1}^s \sum_{k=1}^t \frac{(n_{ijk} - n_{...} \pi_{i..} \pi_{.j.} \pi_{...k})^2}{n_{...} \pi_{i..} \pi_{.j.} \pi_{...k}}$$

Statystyka ta ma rozkład chi-kwadrat z liczbą stopni swobody $rst - 1$.

Wyodrębnimy tutaj trzy składniki, które analogicznie jak we wzorze (4) są następujące:

$$(7) \chi_{AB}^2 = \sum_{i=1}^r \sum_{j=1}^s \frac{(n_{ij.} - n_{..} \pi_{.j.} - n_{.j.} \pi_{i..} + n_{...} \pi_{i..} \pi_{.j.})^2}{n_{...} \pi_{i..} \pi_{.j.}}$$

z liczbą stopni swobody $(r - 1)(s - 1)$.

$$(8) \chi_{AC}^2 = \sum_{r=1}^r \sum_{k=1}^t \frac{(n_{i.k} - n_{i..} \pi_{...k} - n_{.k} \pi_{i..} + n_{...} \pi_{i..} \pi_{...k})^2}{n_{...} \pi_{i..} \pi_{...k}}$$

z liczbą stopni swobody $(r - 1)(t - 1)$.

$$(9) \chi_{BC}^2 = \sum_{j=1}^s \sum_{k=1}^t \frac{(n_{.jk} - n_{.j.} \pi_{...k} - n_{.k} \pi_{.j.} + n_{...} \pi_{.j.} \pi_{...k})^2}{n_{...} \pi_{.j.} \pi_{...k}}$$

z liczbą stopni swobody $(s - 1)(t - 1)$.

Wyodrębnimy również składniki:

$$(10) \chi_A^2 = \sum_{i=1}^r \frac{(n_{i..} - n_{...} \pi_{i..})^2}{n_{...} \pi_{i..}}$$

z liczbą stopni swobody $r - 1$.

$$(11) \quad \chi^2 = \sum_{j=1}^s \frac{(n_{.j.} - n_{...} \pi_{.j.})^2}{n_{...} \pi_{.j.}}$$

z liczbą stopni swobody $s - 1$.

$$(12) \quad \chi^2_C = \sum_{k=1}^t \frac{(n_{..k} - n_{...} \pi_{..k})^2}{n_{...} \pi_{..k}}$$

z liczbą stopni swobody $t - 1$.

Mierzą one rozbieżności pomiędzy zaobserwowanymi i oczekiwanymi częstościami brzegowymi zawartymi odpowiednio w tablicach brzegowych $A \times B$, $A \times C$, $B \times C$ oraz w rozkładach brzegowych ze względu na cechy A , B , C .

Odejmując (7), (8), (9), (10), (11), (12) od (6), otrzymujemy:

$$(13) \quad \chi^2_{ABC} = \chi^2 - \chi^2_{AB} - \chi^2_{AC} - \chi^2_{BC} - \chi^2_A - \chi^2_B - \chi^2_C.$$

Statystyka χ^2_{ABC} ma:

$(rst - 1) - (r - 1)(s - 1) - (r - 1)(t - 1) - (s - 1)(t - 1) - (r - 1) - (s - 1) - (t - 1) = (r - 1)(s - 1)(t - 1)$ stopni swobody. Może ona służyć do testowania niezależności między trzema zmiennymi łącznie, gdy znane są w populacji rozkłady brzegowe.

Przypuśćmy obecnie, że teoretyczne rozkłady brzegowe nie są znane. Oszacujemy je metodą największej wiarygodności - podobnie jak w przypadku tablic kontyngencyjnych dwuwymiarowych:

$$\hat{\pi}_{i..} = \frac{n_{i..}}{n_{...}}, \quad \hat{\pi}_{.j.} = \frac{n_{.j.}}{n_{...}}, \quad \hat{\pi}_{..k} = \frac{n_{..k}}{n_{...}}.$$

Statystyka χ^2 przyjmuje postać:

$$(14) \quad \chi^2 = \sum_{i=1}^r \sum_{j=1}^s \sum_{k=1}^t \frac{(n_{ijk} - \frac{n_{i..} \cdot n_{.j.} \cdot n_{...k}}{n_{...}})^2}{\frac{n_{i..} \cdot n_{.j.} \cdot n_{...k}}{n_{...}}}$$

z liczbą stopni swobody równą $(r-1)(s-1)(t-1)$.

4. Test niezależności χ^2

w tablicach kontyngencyjnych m-wymiarowych

Rozważania dotyczące testu niezależności χ^2 dla tablic kontyngencyjnych trójwymiarowych uogólnić można na m-wymiarowe tablice kontyngencyjne. Podobnie jak miało to miejsce w dwu- i trójwymiarowych tablicach kontyngencyjnych, wyróżnić można dwa przypadki dotyczące rozkładów brzegowych.

Statystyka χ^2 w przypadku, gdy rozkłady brzegowe są znane przybiera postać:

$$(15) \quad \chi^2 = \sum_{k_1=1}^{r_1} \sum_{k_2=1}^{r_2} \dots \sum_{k_m=1}^{r_m} \frac{(n_{k_1 k_2 \dots k_m} - n_{k_1.} n_{.k_2.} \dots n_{...k_m})^2}{n_{k_1.} n_{.k_2.} \dots n_{...k_m}}$$

z liczbą stopni swobody równą $r_1, r_2, \dots, r_m - 1$.

Statystykę (15) należy, analogicznie jak (13) pomniejszyć o wartości χ^2 dla rozkładów brzegowych odpowiednio $m-1, m-2, \dots, 2, 1$ - wymiarowych. Otrzymamy ostatecznie:

$$(16) \quad \chi_{A_1 A_2, \dots, A_m}^2 = \chi^2 - \left(\sum_{i_1} \chi_{A_{i_1}}^2 + \sum_{i_1 i_2} \chi_{A_{i_1 i_2}}^2 + \dots + \sum_{i_1 i_2, \dots, i_m} \chi_{A_{i_1 i_2, \dots, i_m}}^2 \right),$$

gdzie:

$$A = \{A_1, A_2, \dots, A_m\};$$

$$A_{i_1} = A \setminus A_{i_1}, \quad i_1 \in \{1, \dots, m\};$$

$$A_{i_1 i_2} = A \setminus (A_{i_1} \cup A_{i_2}), \quad i_1, i_2 \in \{1, \dots, m\};$$

$$A_{i_1 i_2, \dots, i_k} = A \setminus \bigcup_{p=1}^k A_{i_p}, \quad k = 2, \dots, m-1.$$

Statystyka $\chi^2_{A_1 A_2, \dots, A_m}$ ma: $(r_1 - 1)(r_2 - 1) \dots (r_m - 1)$ stopni swobody.

Gdy rozkłady brzegowe nie są znane szacujemy je według wzorów:

$$\hat{\pi}_{i_1} = \frac{n_{i_1}}{n}, \quad \hat{\pi}_{i_1 i_2} = \frac{n_{i_1 i_2}}{n}, \dots, \quad \hat{\pi}_{i_1 i_m} = \frac{n_{i_1 i_m}}{n}.$$

Statystyka χ^2 przyjmuje wówczas postać:

$$\chi^2 = \sum_{i_1=1}^{r_1} \dots \sum_{i_m=1}^{r_m} \frac{\left(\frac{n_{i_1 i_2, \dots, i_m}}{n^{m-1}} - \frac{n_{i_1} n_{i_2, \dots, i_m}}{n^m} \right)^2}{\frac{n_{i_1} n_{i_2, \dots, i_m}}{n^m} - \frac{n_{i_1}^2 n_{i_2, \dots, i_m}^2}{n^{2m-1}}}$$

z liczbą stopni swobody $(r_1 - 1)(r_2 - 1), \dots, (r_m - 1)$.

5. Przykład

Zaproponowane w poprzednich punktach rozważania teoretyczne zastosujemy obecnie do przykładu empirycznego. W trójwymiarowej tablicy kontyngencyjnej $2 \times 2 \times 2$ (tabl. 1) zawarte są dane z badania przeprowadzonego w określonym szpitalu klinicznym w Łodzi. Badaniem objęto 74 chorych poddanych operacji. Stan zdrowia przed

operacją (cecha A) i tuż po operacji (cecha B) oceniano w punktach od 9 do 10. Im wyższa liczba punktów, tym gorszy, zgodnie z oceną wystawioną przez lekarzy, stan zdrowia pacjenta. Trzecia cecha (C) oznacza wynik operacji (pozytywny bądź negatywny), przy czym negatywny wynik niekoniecznie oznacza zgon pacjenta podczas lub tuż po operacji (zgon nastąpił tylko w jednym przypadku).

T a b l i c a 1

Stan zdrowia przed operacją, tuż po operacji
i wynik operacji pacjentów
pewnego szpitala klinicznego w Łodzi

A	B	C		Ogółem
		C ₁	C ₂	
A ₁	B ₁	-	25	25
	B ₂	-	5	5
	n _{1.k}	-	30	30
A ₂	B ₁	1	24	25
	B ₂	6	13	19
	n _{2.k}	7	37	44
	B ₁	1	49	50
	B ₂	6	18	24
n _{..k}		7	67	74

Ź r ó d ł o: Obliczenia własne.

Dla uproszczenia zapisu w tablicy wyspecyfikujemy symbole nadane cechom i ich poszczególnym wariantom:

A - ocena stanu zdrowia pacjenta przed operacją (w punktach),
 A_1 - mniej niż 5 punktów, A_2 - 5 punktów lub więcej;

B - ocena stanu zdrowia pacjenta w czasie operacji (w punktach),
 B_1 - 0 punktów, B_2 - więcej niż 0 punktów;

C - wynik operacji, C_1 - pozytywny, C_2 - negatywny.

Aby rozstrzygnąć, czy istnieje współzależność między rozpatrywanymi trzema cechami w populacji generalnej operowanych pacjentów tego szpitala stawiamy hipotezę zerową o niezależności cech A, B, C, wobec hipotezy alternatywnej, że cechy są zależne.

Obliczona zgodnie ze wzorem (14) wartość statystyki χ^2 wynosi 24,392. Porównując tę wartość z χ^2_α odczytanymi z tablic dla $\alpha = 0,005$ oraz dla $\alpha = 0,001$ i dla jednego stopnia swobody otrzymujemy odpowiednio 7,879 i 10,828. Ponieważ $\chi^2 > \chi^2_\alpha$, na poziomach istotności 0,005 i 0,001, hipotezę o łącznej niezależności cech A, B, C należy odrzucić. Są one zatem zależne.

6. Podsumowanie

Przeprowadzona w opracowaniu analiza statystyki χ^2 dotyczyła głównie jej zastosowania do badania niezależności w tablicach kontyngencyjnych dwu- i trójwymiarowych z uogólnieniem na wielowymiarową analizę statystyczną. Jest to, jak nam się wydaje, zagadnienie dość istotne, szczególnie w naukach rolniczych, społecznych i medycznych, bowiem tam często badacz ma do czynienia z łącznym działaniem kilku cech niemierzalnych.

Braliśmy pod uwagę jedynie przypadki "czyste" - znanych lub nieznanymi rozkładów brzegowych. W praktyce mogą mieć miejsce sytuacje "mieszane". Już dla dwuwymiarowej tablicy kontyngencyjnej rozpatrujemy w niektórych badaniach przypadek jednego brzegu ustalonego, a drugiego nieznanego. Problem jeszcze bardziej się komplikuje, gdy przechodzimy do większej liczby wymiarów (w tablicy kontyngencyjnej trójwymiarowej jest już łącznie 8 przypadków). Może to być jednak przedmiotem innych opracowań.

Literatura

- [1] B l a l o o k H. M. (1975): Statystyka dla socjologów, PWN, Warszawa.
- [2] C r a m e r H. (1958): Metody matematyczne w statystyce, PWN, Warszawa.
- [3] F i s z M. (1967): Rachunek prawdopodobieństwa i statystyka matematyczna, PWN, Warszawa.
- [4] G ó r a l s k i A. (1974): Metody opisu i wnioskowania statystycznego w psychologii, PWN, Warszawa.
- [5] H e l m e r t F. R. (1976): Über die Wahrscheinlichkeit von Potenzsummen der Beobachtungsfehler etc., z.f. Math. U. Phys., 21.
- [6] P e a r s o n K. (1900): On the Criterion that a Given System of Deviations from the Probable etc., Phil. Mag., V, 50.
- [7] O k t a b a W. (1977): Elementy statystyki matematycznej i metodyka doświadczalnictwa, PWN, Warszawa.
- [8] K e n d a l l M. G., S t u a r t A. (1963): The Advanced Theory of Statistics, London.
- [9] R a o R. C. (1982): Modele liniowe statystyki matematycznej, PWN, Warszawa.
- [10] Y u l e G. U., K e n d a l l M. G. (1966): Wstęp do teorii statystyki, PWN, Warszawa.

Lechosław Stępień

GENERALIZATION OF STATISTIC χ^2 IN MULTIDIMENSIONAL
STATISTICAL ANALYSIS

Generalization of statistic χ^2 has been performed in the article and then the statistic has been applied to solving problems concerning the multidimensional statistical analysis.

Independency test χ^2 has been examined and its application has been illustrated by an empirical example derived from medical research.