Analiza korespondencji

Maciej Beręsewicz

Kompetencje Studentów WIGE UEP na dobry start!

Poznań 2015

Analiza korespondencji w badaniach rynku

Analiza korespondencji w badaniach rynku

- Testy niezależności oparte na statystyce χ^2
- Analiza korespondencji – ujęcie teoretyczne

Test χ^2 niezależności

Test χ^2 niezależności

- Punktem wyjścia w analizie korespondencji jest dwuwymiarowa tablica kontyngencji $\mathbf{N} = [n_{ij}]$, gdzie n_{ij} oznaczają liczebności empiryczne w i–tym wierszu oraz j–tej kolumnie ($i = 1, \ldots, I; j = 1, \ldots, J$).

Tabela 1. Tablica kontyngencji dla dwóch zmiennych X i Y

Kategorie zmiennej X	Kategorie zmiennej Y				Suma wierszy
	Y_1	Y_2	$\ldots$	Y_J	
X_1	n_{11}	n_{12}	$\ldots$	n_{1J}	$n_{1.}$
X_2	n_{21}	n_{22}	$\ldots$	n_{2J}	$n_{2.}$
$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$
X_I	n_{I1}	n_{I2}	$\ldots$	n_{IJ}	$n_{I.}$
Suma kolumn	$n_{.1}$	$n_{.2}$	$\ldots$	$n_{.J}$	n

Źródło: opracowanie własne

Powyższa tabela przedstawia tablicę kontyngencji dla dwóch zmiennych X i Y, z których pierwsza ma I a druga J wariantów.

Test χ^2 niezależności

Test χ^2 niezależności

- W praktyce najczęściej wykorzystywanym narzędziem pozwalającym wykryć zależność między zmiennymi X i Y jest test χ^2 niezależności. Hipoteza zerowa – H_0 zakłada brak związku między zmiennymi, czyli zmienne X i Y są niezależne, wobec hipotezy alternatywnej – H_1, że zmienne X i Y nie są niezależne. Sprawdzianem hipotezy zerowej jest statystyka

$$\chi^2 = \sum_{i=1}^{I} \sum_{j=1}^{J} \frac{\left(n_{ij} - \hat{n}_{ij}\right)^2}{\hat{n}_{ij}}, \tag{1}$$

gdzie n_{ij} oraz $\hat{n}_{ij}$ oznaczają odpowiednio empiryczne i teoretyczne liczebności tablicy kontyngencji. Liczebności teoretyczne $\hat{n}_{ij}$ wyznaczamy ze wzoru

$$\hat{n}_{ij} = \frac{n_{i.}\, n_{.j}}{n}, \tag{2}$$

gdzie n oznacza liczebność próby, a $n_{i.}$, $n_{.j}$ liczebności brzegowe. Statystyka (1) przy założeniu prawdziwości hipotezy H_0 ma graniczny rozkład χ^2 z $(I-1)(J-1)$ stopniami swobody. Obszar krytyczny w teście χ^2 niezależności buduje się prawostronnie:

$$P\left(\chi^2 \geq \chi^2_\alpha\right) = \alpha. \tag{3}$$

Inne testy niezależności oparte na statystyce χ^2

Inne testy niezależności oparte na statystyce χ^2

- Do badania niezależności zmiennych X i Y można również wykorzystać statystykę największej wiarogodności L^2 oraz statystykę Cressie-Reada (CR).

$$L^2 = 2\sum_{i=1}^{I}\sum_{j=1}^{J} n_{ij} ln\left(\frac{n_{ij}}{\hat{n}_{ij}}\right), \tag{4}$$

$$CR = \frac{2}{\lambda(\lambda+1)}\sum_{i=1}^{I}\sum_{j=1}^{J} n_{ij}\left[\left(\frac{n_{ij}}{\hat{n}_{ij}}\right)^{\lambda} - 1\right]. \tag{5}$$

Statystyka χ^2 określona wzorem (1) oraz statystyka L^2 są szczególnymi przypadkami statystyki wyprowadzonej przez Cressie i Reada. Gdy $\lambda = 1$ otrzymujemy statystykę χ^2, a gdy $\lambda \to 0$ otrzymujemy statystykę L^2. Cressie i Read proponują, aby za wartość parametru λ przyjmować $\lambda = \frac{2}{3}$. Podobnie jak w klasycznym teście χ^2 niezależności obszar krytyczny buduje się prawostronnie:

$$P\left(L^2 \geq \chi^2_\alpha\right) = \alpha, \tag{6}$$

$$P\left(CR \geq \chi^2_\alpha\right) = \alpha. \tag{7}$$

Wybrane miary siły zależności

Wybrane miary siły zależności

- Współczynnik Φ Yule'a

$$\Phi = \sqrt{\frac{\chi^2}{n}}, \tag{8}$$

- Współczynnik kontyngencji C Pearsona

$$C = \sqrt{\frac{\chi^2}{\chi^2 + n}}, \tag{9}$$

- Współczynnik zbieżności Czuprowa

$$T = \sqrt{\frac{\chi^2}{n\sqrt{(I-1)(J-1)}}}, \tag{10}$$

- Współczynnik V Cramera

$$V = \sqrt{\frac{\chi^2}{n \cdot min\{I-1, J-1\}}}. \tag{11}$$

Wszystkie wymienione współczynniki przyjmują wartość 0 jeżeli nie ma zależności między badanymi zmiennymi. Im są one większe tym zależność jest silniejsza. Współczynniki Φ Yule'a oraz kontyngencji C Pearsona są miarami nieunormowanymi, natomiast współczynniki T Czuprowa i V Cramera przyjmują wartości z przedziału $[0, 1]$.

Analiza korespondencji – zastosowania

Analiza korespondencji – zastosowania

- Analiza korespondencji jest metodą statystycznej analizy wielowymiarowej pozwalającą na analizę danych mierzonych na słabych skalach pomiaru. Jej celem jest wskazanie wzajemnych powiązań kategorii zmiennych nominalnych zawartych w tablicy kontyngencji we wspólnej przestrzeni, zazwyczaj w dwóch lub trzech wymiarach, w postaci tzw. mapy percepcji.
- Analiza korespondencji jest metodą powstałą na gruncie psychologii, a swoje początki zawdzięcza Pearsonowi. Nazwa tej metody została spopularyzowana przez Hilla, a sama metoda stała się popularna dzięki pracom Greenacre'a, który przedstawił techniczną stronę tej metody wraz z wieloma przykładami jej zastosowań.
- Analiza korespondencji ma wiele zastosowań, przede wszystkim w socjologii, psychologii, biologii, medycynie a także w badaniach ekonomicznych.
- W badaniach marketingowych analiza korespondencji jest wykorzystywana na szeroką skalę w procesie:
  - segmentacji rynku,
  - określania pozycji produktu na rynku,
  - monitorowania skuteczności kampanii reklamowej,
  - rozpoznawania luki na rynku.

Macierz korespondencji **P**

Macierz korespondencji **P**

- Analiza korespondencji oparta jest na tzw. macierzy korespondencji **P**, która zdefiniowana jest jako macierz elementów tablicy kontyngencji n_{ij} podzielonych przez ogólną liczebność próby n, tj. $\mathbf{P} = \left[\frac{n_{ij}}{n}\right]$.

Tabela 2. Macierz korespondencji **P**

Kategorie zmiennej X	Kategorie zmiennej Y				Suma wierszy
	Y_1	Y_2	$\ldots$	Y_J	
X_1	p_{11}	p_{12}	$\ldots$	p_{1J}	$p_{1.}$
X_2	p_{21}	p_{22}	$\ldots$	p_{2J}	$p_{2.}$
$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$
X_I	p_{I1}	p_{I2}	$\ldots$	p_{IJ}	$p_{I.}$
Suma kolumn	$p_{.1}$	$p_{.2}$	$\ldots$	$p_{.J}$	1

Źródło: opracowanie własne

- Punktem wyjścia w analizie korespondencji jest:
  - obliczenie mas wierszy oraz mas kolumn,
  - obliczenie profilów wierszy oraz profilów kolumn,
  - wyznaczenie odległości między kategoriami zmiennych (punktów) za pomocą metryki χ^2.

Masy wierszy i kolumn

Masy wierszy i kolumn

- Wprowadźmy następujące oznaczenia. Niech $\mathbf{r}$ oraz $\mathbf{c}$ oznaczają wektory sum wierszy i kolumn macierzy korespondencji $\mathbf{P}$. Wektory te można przedstawić w postaci:

$$\mathbf{r} = \mathbf{P}\mathbf{1}_c = \left[\frac{n_{i.}}{n}\right], \tag{12}$$

$$\mathbf{c} = \mathbf{P}^{\mathbf{T}}\mathbf{1}_r = \left[\frac{n_{.j}}{n}\right], \tag{13}$$

  gdzie $\mathbf{1}_c$ jest wektorem jednostkowym o wymiarach $J \times 1$, a $\mathbf{1}_r$ jest wektorem jednostkowym o wymiarach $I \times 1$. W analizie korespondencji elementy wektora $\mathbf{r}$ oraz $\mathbf{c}$ nazywane są odpowiednio masami wierszy oraz masami kolumn.
- Z definicji mas wierszy oraz mas kolumn wynika, że masy wierszy to częstości brzegowe wierszy w tablicy kontyngencji, natomiast masy kolumn to częstości brzegowe kolumn.

Profile wierszy i kolumn

Profile wierszy i kolumn

- Niech r_i oznacza masę i–tego wiersza, p_{ij} częstości obserwowane w i–tym wierszu oraz j–tej kolumnie macierzy korespondencji $\mathbf{P}$, a $\mathbf{D}_r$ macierz diagonalną zawierającą masy wierszy $\mathbf{r}$. Macierz profili wierszy definiujemy w następujący sposób:

$$\mathbf{R} = \mathbf{D}_r^{-1}\mathbf{P} = \left[\frac{p_{ij}}{r_i}\right]. \tag{14}$$

  Profile wierszy powstają przez podzielenie elementów tablicy kontyngencji zawierającej liczebności empiryczne n_{ij} przez liczebności brzegowe wierszy $n_{i.}$.

- Niech c_j oznacza masę j–tej kolumny, p_{ij} częstości obserwowane w i–tym wierszu oraz j–tej kolumnie macierzy korespondencji $\mathbf{P}$, a $\mathbf{D}_c$ macierz diagonalną zawierającą masy kolumn $\mathbf{c}$. Macierz profili kolumn definiujemy w następujący sposób:

$$\mathbf{R} = \mathbf{D}_c^{-1}\mathbf{P}^T = \left[\frac{p_{ij}}{c_j}\right]. \tag{15}$$

  Profile kolumn powstają przez podzielenie elementów tablicy kontyngencji zawierającej liczebności empiryczne n_{ij} przez liczebności brzegowe kolumn $n_{.j}$.

- Wyznaczone profile wierszy oraz kolumn mogą być przedstawione w postaci punktu w wielowymiarowej przestrzeni. Profile bardziej podobne będą w przestrzeni punktami położonymi bliżej siebie. Natomiast profile bardziej różniące się od siebie będą przedstawiane w przestrzeni jako punkty leżące daleko od siebie. W analizie korespondencji początek układu rzutowania wyznaczają tzw. centroidy oznaczające środki ciężkości. Wektor $\mathbf{r}$ wyznacza centroidę dla wierszy, natomiast $\mathbf{c}$ centroidę dla kolumn. Jeśli dany profil różni się znacznie od średniego profilu, to taki punkt będzie leżał daleko od początku układu rzutowania. Natomiast punkty, których profile są zbliżone do średniego profilu, będą położone bliżej środka ciężkości.

Rozkład macierzy według wartości osobliwych

Rozkład macierzy według wartości osobliwych

- W analizie korespondencji bardzo użytecznym narzędziem jest rozkład macierzy według wartości osobliwych (SVD – singular value decomposition). Zastosowanie tej metody umożliwia wyznaczenie współrzędnych punktów (kategorii zmiennych), co w konsekwencji pozwoli na naniesienie punktów reprezentujących kategorie zmiennych nominalnych na tzw. mapę percepcji.
- Rozkład macierzy $\mathbf{A}$ o wymiarach $I \times J$ według wartości osobliwych polega na jej dekompozycji na iloczyn trzech macierzy:

$$\mathbf{A} = \mathbf{U}\boldsymbol{\Gamma}\mathbf{V}^T, \tag{16}$$

gdzie $\boldsymbol{\Gamma}$ jest macierzą diagonalną o wymiarach $K \times K$ o wartościach dodatnich, uporządkowanych w sposób malejący tak, że $\gamma_1 \geq, \ldots, \geq \gamma_K$. Elementy macierzy diagonalnej $\boldsymbol{\Gamma}$ nazywane są wartościami osobliwymi macierzy $\mathbf{A}$. Macierz $\mathbf{U}$ o wymiarach $I \times K$ utworzona jest z wektorów własnych macierzy $\mathbf{A}\mathbf{A}^T$, odpowiadających wartościom własnym $\gamma_1^2, \ldots, \gamma_K^2$. Z kolei macierz $\mathbf{V}$ o wymiarach $J \times K$ utworzona jest z wektorów własnych macierzy $\mathbf{A}^T\mathbf{A}$, odpowiadających tym samym wartościom własnym. Macierze $\mathbf{U}$ oraz $\mathbf{V}$ są ortogonalne, tzn. $\mathbf{U}^T\mathbf{U} = \mathbf{V}^T\mathbf{V} = \mathbf{I}$. Kolumny macierzy $\mathbf{U}$ nazywane są lewymi wektorami osobliwymi, kolumny zaś macierzy $\mathbf{V}$ prawymi wektorami osobliwymi.

Uogólniona metoda rozkładu macierzy według wartości osobliwych

Uogólniona metoda rozkładu macierzy według wartości osobliwych

- W analizie korespondencji najczęściej, celem stworzenia mapy percepcji, stosowana jest tzw. uogólniona metoda rozkładu macierzy według wartości osobliwych (GSVD), która pozwala na wyznaczenie najlepszej przestrzeni rzutowania. Znalezienie optymalnej bazy podprzestrzeni rzutowania wymaga dokonania rozkładu macierzy $\mathbf{P} - \mathbf{r}\mathbf{c}^T$ według wartości osobliwych. Macierz $\mathbf{A}$ można przedstawić za pomocą formuły:

$$\mathbf{A} = \mathbf{D}_r^{-1/2} \left(\mathbf{P} - \mathbf{r}\mathbf{c}^T \right) \mathbf{D}_c^{-1/2}. \tag{17}$$

Macierz $\mathbf{A}$ nazywana jest macierzą różnic standaryzowanych. Współrzędne kategorii zmiennych nominalnych X oraz Y są wyznaczane zgodnie z następującymi wzorami:

$$\mathbf{X} = \mathbf{D}_r^{-1/2} \mathbf{U} \boldsymbol{\Gamma}, \tag{18}$$

$$\mathbf{Y} = \mathbf{D}_c^{-1/2} \mathbf{V} \boldsymbol{\Gamma}. \tag{19}$$

Odległość chi-kwadrat i inercja

Odległość chi-kwadrat i inercja

- Najczęściej stosowaną odległością mającą duże znaczenie w dalszej analizie jest odległość chi-kwadrat. Odległość między kategoriami i oraz i' zmiennej nominalnej X wyznacza się za pomocą wzoru:

$$d\left(i,i'\right)=\sqrt{\sum_{j=1}^{J}\frac{1}{c_j}\left(\frac{p_{ij}}{r_i}-\frac{p_{i'j}}{r_{i'}}\right)^2}, \tag{20}$$

odległość zaś chi-kwadrat między kategoriami j oraz j' zmiennej nominalnej Y wyznacza się za pomocą wzoru:

$$d\left(j,j'\right)=\sqrt{\sum_{i=1}^{I}\frac{1}{r_i}\left(\frac{p_{ij}}{c_j}-\frac{p_{ij'}}{c_{j'}}\right)^2}. \tag{21}$$

Odległość chi-kwadrat i inercja

Odległość chi-kwadrat i inercja

- Inercja całkowita (bezwładność) jest miarą stopnia rozproszenia punktów wokół odpowiednich środków ciężkości. Im wyższa wartość inercji (bezwładności), tym silniejszy jest związek (zależność) między zmiennymi zawartymi w tablicy kontyngencji. Wartość inercji bliska 0 oznacza brak zależności między zmienną X oraz Y. Bezwładność jest statystyką χ^2 podzieloną przez liczbę obserwacji i obliczaną według wzoru:

$$\text{Inercja} = \sum_{i=1}^{I}\sum_{j=1}^{J}\frac{(p_{ij} - r_i c_j)^2}{r_i c_j} = \frac{\chi^2}{n}, \tag{22}$$

  gdzie χ^2 to wartość statystyki chi-kwadrat dla zadanej tablicy kontyngencji.
- Dla dwudzielczych tablic kontyngencji minimalna liczba wymiarów wyznaczana jest w następujący sposób:

$$k = min\left\{I - 1, J - 1\right\}. \tag{23}$$

Odległość chi-kwadrat i inercja

Odległość chi-kwadrat i inercja

- Stosując współrzędne punktów (kategorii zmiennych), można wyznaczyć wartości własne dla każdego wymiaru, tj.:
  - dla wierszy

$$\lambda_k^2 = \sum_{i=1}^{I} r_i f_{ik}^2, \tag{24}$$

    gdzie f_{ik} oznacza współrzędną punktu i–tego na k–tym wymiarze,
  - dla kolumn

$$\lambda_k^2 = \sum_{j=1}^{J} c_j g_{jk}^2, \tag{25}$$

    gdzie g_{jk} oznacza współrzędną punktu j–tego na k–tym wymiarze.

Odległość chi-kwadrat i inercja

Odległość chi-kwadrat i inercja

- Wartość całkowitej inercji może być również dekomponowana na poszczególne kategorie zmiennych, odpowiednio:
  - dla wierszy

$$\lambda_i = \sum_{k=1}^{K} r_i f_{ik}^2, \tag{26}$$

  - dla kolumn

$$\lambda_j = \sum_{k=1}^{K} c_j g_{jk}^2. \tag{27}$$

Interpretacja wyników

Interpretacja wyników

- W celu głębszej analizy danych zawartych w początkowej tablicy kontyngencji można posłużyć się dodatkowymi statystykami, tj. udziałem punktów w wymiarach (*contribution of points to dimensions*) oraz udziałem wymiarów w punktach (*contributions of dimensions to points*).
- Udział punktów w wymiarach (bezwzględny udział) wskazuje proporcjonalną inercję wyjaśnianą przez każdą kategorię w stosunku do głównych osi. Najważniejsze dla rozważań są punkty o istotnym udziale w kształtowaniu wymiaru. Suma udziałów dla każdego z wymiarów jest równa 1. Udział punktów w wymiarach jest obliczany odpowiednio według wzorów:
  - dla wierszy
$$ctr_{ik} = \frac{r_i f_{ik}^2}{\lambda_k^2}, \tag{28}$$
  - dla kolumn
$$ctr_{jk} = \frac{c_j g_{jk}^2}{\lambda_k^2}. \tag{29}$$
- Bezwzględny udział określa, jak istotny jest punkt w definiowaniu kierunku głównej osi, oraz służy jako klucz interpretacji poszczególnych osi.

Interpretacja wyników

Interpretacja wyników

- Udział wymiarów w punktach (kwadrat korelacji) dostarcza informacji o tym, jaka część inercji punktu jest wyjaśniona przez wymiar. Te udziały są obliczane odpowiednio według następujących wzorów:
  - dla wierszy
$$cor_{ik} = \frac{r_i f_{ik}^2}{\lambda_h}, \tag{30}$$
  - dla kolumn
$$cor_{jk} = \frac{c_j g_{jk}^2}{\lambda_j}. \tag{31}$$

Literatura

Literatura

- Gatnar E., Walesiak M. (red.) (2004), *Metody statystycznej analizy wielowymiarowej w badaniach marketingowych*, Wydawnictwo Akademii Ekonomicznej im. Oskara Langego we Wrocławiu, Wrocław 2004.
- Górniak J., Wachnicki J. (2008), *Pierwsze kroki w analizie danych*, Wydawnictwo SPSS.
- Stanisz A. (2007), *Przystępny kurs statystyki z zastosowaniem STATISTICA PL na przykładach z medycyny*, Tom 3. Analizy wielowymiarowe, StatSoft, Kraków.
- Walesiak M. (1996), *Metody analizy danych marketingowych*, Wydawnictwo Naukowe PWN, Warszawa.
- Walesiak M., Gatnar E. (red.) (2009), *Statystyczna analiza danych z wykorzystaniem programu R*, Wydawnictwo Naukowe PWN, Warszawa.

Dziękuje za uwagę