Analiza korespondencji

Dorota Celińska-Kopczyńska, Paweł Strawiński

Uniwersytet Warszawski

2 listopada 2018

Plan zajęć I

Wprowadzenie

- Analiza korespondencji należy do technik analizy wielowymiarowej zmiennych jakościowych.
- Polega na przeprowadzeniu operacji na **tabeli wielodzielczej**, czyli tabeli przedstawiającej rozkład obserwacji ze względu na kilka cech jednocześnie.
- Analiza ta dostarcza informacji na temat struktury powiązań pomiędzy zmiennymi, a jej graficzna prezentacja wyników umożliwia intuicyjne wnioskowanie odnośnie powiązań zachodzących pomiędzy kategoriami badanych zmiennych.

Przykład tabeli wielodzielczej

Płeć	Liczba wypalonych papierosów			Suma
	< 10	10-20	> 20	
Kobieta	15	20	5	40
Mężczyzna	10	30	20	60
Suma	25	50	25	100

Pojęcie niezależności

- Przed przeprowadzeniem analizy korespondencji należy upewnić się, że ta technika analizy jest właściwa
- Między badanymi zmiennymi musi zachodzić zależność: dążymy do odrzucenia hipotezy zerowej o niezależności badanych zmiennych
- Dwa zdarzenia są **niezależne**, jeśli prawdopodobieństwo wystąpienia ich iloczynu jest równe iloczynowi ich prawdopodobieństw brzegowych:

$$P(A \cap B) = P(A)P(B)$$

Test niezależności χ^2

- Porównuje się częstości zaobserwowanych z częstościami oczekiwanymi, przy założeniu prawdziwości hipotezy zerowej
- H_0 – zmienne są niezależne; H_1 – istnieje związek pomiędzy zmiennymi
- Częstości oczekiwane:

$$E_{ij} = \frac{\sum_{j=1}^{k} n_j \sum_{i=1}^{w} n_i}{\sum_{i=1}^{w} \sum_{j=1}^{k} n_{ij}} = \frac{suma\ wiersza * suma\ kolumny}{suma\ calkowita}$$

k – liczba kolumn; w – liczba wierszy

- Statystyka testowa:

$$\chi^2 = \sum \frac{(O_{ij} - E_{ij})^2}{E_{ij}} = \sum_{i=1}^{w} \sum_{j=1}^{k} \frac{(n_{ij} - E_{ij})^2}{E_{ij}}$$

O_{ij} – obserwowana częstość komórki

Ocena siły związku

- Test χ^2 służy do sprawdzenia, czy pomiędzy zmiennymi występuje zależność. Nie odpowiada natomiast na pytanie, jak silne jest to powiązanie.
- Wartości statystyki χ^2 nie można stosować do pomiaru siły związku, gdyż jest ona zależna od **liczebności próby i rośnie wraz z jej wzrostem**.
- Najpopularniejszymi miarami siły związku opartymi na statystyce χ^2 są:
  1. Współczynnik korelacji ϕ;
  2. Współczynnik zbieżności V-Cramera;
  3. Współczynnik kontyngencji Pearsona.

Współczynnik korelacji ϕ

$$\phi = \frac{n_{11}n_{22} - n_{12}n_{21}}{\sqrt{n_{11}n_{22}n_{12}n_{21}}} \; dla \; tabel \; 2x2$$

$$\phi = \sqrt{\frac{\chi^2}{n}} \; w \; p. \; p.$$

- W przypadku tablicy 2x2 równy jest współczynnikowi V-Cramera; przyjmuje wartości z przedziału (-1;1).
- W przypadku większych tablic przyjmuje wartości z przedziału (0;1).
- Wpływ wielkości próby jest eliminowany dzięki podzieleniu statystyki χ^2 przez liczeność próby.

Współczynnik zbieżności V-Cramera

$$V = \sqrt{\frac{\chi^2}{n * min(k-1; w-1)}}$$

- $V = 0$ zmienne są niezależne (brak korelacji).
- $V = 1$ pomiędzy zmiennymi występuje silna funkcyjna zależność.
- $0 < V < 1$ przedział możliwych wartości współczynnika V-Cramera dla tablic większych niż 2x2
- $-1 < V < 1$ przedział możliwych wartości współczynnika V-Cramera dla tablic 2x2.

Współczynnik kontyngencji Pearsona

$$P = \sqrt{\frac{\chi^2}{\chi^2 + n}}$$

- $P = 0$ zmienne są niezależne (brak korelacji).
- $0 < P < 1$ przedział możliwych wartości współczynnnika kontyngencji Pearsona.
- Im wartość współczynnika bliższa 1, tym silniejszy związek pomiędzy zmiennymi.

Motywacja

- Miary siły związku stanowią zaledwie punkt wyjścia w analizie zmiennych jakościowych. Nie mówią one nic o strukturze powiązań pomiędzy zmiennymi.
- Analiza korespondencji (a szczególnie jej graficzna prezentacja) umożliwia intuicyjne wnioskowanie o powiązaniach zachodzących pomiędzy kategoriami badanych zmiennych.

Obszar zastosowań

- Prezentacja graficzna zależności między zmiennymi jakościowymi.
- Metoda wspomagania i uzupełniania – nie jest zamiennikiem dla bardziej formalnych narzędzi statystycznych.
- Metoda raczej „eksploracyjna”, ułatwiająca stawianie hipotez dla dalszych etapów badania, niż „konfirmacyjna”.

Etapy analizy korespondencji

1. Wyznaczenie profili wierszowych i kolumnowych
2. Wyznaczenie masy wiersza i kolumny
3. Obliczenie odległości między wierszami (kolumnami) za pomocą metryk χ^2
4. Wyznaczenie przeciętnych profili wierszowych i kolumnowych
5. Redukcja wymiaru przestrzeni
6. Utworzenie i interpretacja wspólnego wykresu profili wierszowych i kolumnowych

Podstawowe macierze

- Najpierw obliczamy **macierz korespondencji**. Dzielimy liczebności w poszczególnych komórkach tabeli wielodzielczej przez liczebność całkowitą badanej próby.
- Następnie wyznaczamy **macierze profili wierszowych i kolumnowych**.
- Macierz profili wierszowych otrzymujemy dzieląc poszczególne elementy wierszy macierzy korespondencji przez sumę wszystkich elementów tego wiersza.
- Macierz profili kolumnowych otrzymujemy dzieląc poszczególne elementy kolumn macierzy korespondencji przez sumę wszystkich elementów tej kolumny.

Przeciętne profile kolumnowe i wierszowe

- **Masa wiersza** – suma elementów danego wiersza macierzy korespondencji.
- **Masa kolumny** – suma elementów danej kolumny macierzy korespondencji.
- **Przeciętny profil kolumnowy** – kolumna o elementach będących masami wierszy.
- **Przeciętny profil wierszowy** – wiersz o elementach będących masami kolumn.
- Przeciętny profil wierszowy można uzyskać, dzieląc podsumowujący wiersz w tabeli wielodzielczej przez ogólną liczebność.

- Tabela wielodzielcza

Płeć	Liczba wypalonych papierosów			Suma
	< 10	10-20	> 20	
Kobieta	15	20	5	40
Mężczyzna	10	30	20	60
Suma	25	50	25	**100**

- Macierz korespondencji

Płeć	Liczba wypalonych papierosów			Suma
	< 10	10-20	> 20	
Kobieta	**0,15**	0,20	0,05	**0,40**
Mężczyzna	**0,10**	0,30	0,20	**0,60**
Suma	0,25	0,50	0,25	1

- Macierz profili wierszowych

Płeć	Liczba wypalonych papierosów			Suma
	< 10	10-20	> 20	
Kobieta	0,38	0,50	0,13	0,40
Mężczyzna	0,17	0,50	0,33	0,60
Suma	0,25	0,50	0,25	1

Definicja odległości χ^2

$$\chi^2 = d^2(p, p') = \sum_{j=1}^{w(k)} \frac{(p_j - p'_j)^2}{\bar{p}_j}$$

- $d(p, p')$ – odległość między profilami p i p'
- p_j, p'_j – elementy profilu p i p' (częstości względne)
- $\bar{p}_j$ – elementy przeciętnego profilu
- Odległości te obliczane są zarówno dla profili wierszowych, jak i kolumnowych.
- Kategorie z relatywnie większą liczbą elementów wywierają mniejszy wpływ na odległość niż kategorie z mniejszą liczbą obserwacji.

Przykład – obliczenia dla profili wierszowych

- Macierz profili wierszowych

Płeć	Liczba wypalonych papierosów			P. profil kol.
	< 10	10-20	> 20	
Kobieta	0,38	0,50	0,13	0,40
Mężczyzna	0,17	0,50	0,33	0,60
P. profil wiersz.	0,25	0,50	0,25	1

$$\chi^2 = \frac{(0,38 - 0,17)^2}{0,25} + \frac{(0,5 - 0,5)^2}{0,5} + \frac{(0,13 - 0,33)^2}{0,25} = 0,34$$

Rozkład względem wartości osobliwych

- Każdy wiersz macierzy profili o wymiarach *wxk* może zostać przedstawiony jako punkt w przestrzeni k-wymiarowej, generowanej przez kolumny macierzy.
- Każda kolumna macierzy profili może zostać przedstawiona jako punkt w przestrzeni w-wymiarowej, generowanej przez wiersze tej macierzy.
- Korzystając z analizy korespondencji dążymy do przedstawienia analizowanego zbioru punktów w przestrzeni maksymalnie trójwymiarowej, przy zachowaniu jak największej informacji o zróżnicowaniu wierszy i kolumm.
- W tym celu korzysta się z **rozkładu względem wartości osobliwych** (*Singular Value Decomposition – SVD*).

Rozkład względem wartości osobliwych – cd

- Metoda SVD polega na przedstawieniu macierzy A rzędu r (o wymiarze w x k) w postaci iloczynu trzech macierzy:

$$A_{wxk} = U_{wxr} D_{rxr} V_{rxk}$$

- U, V – macierze ortonormalne ($U'U = I_{rxr}$ i $V'V = I_{kxk}$);
- D – macierz diagonalna utworzona z niezerowych wartości własnych macierzy $A'A$, uporządkowanych nierosnąco $\lambda_1 \geqslant \lambda_2 \geqslant \cdots \geqslant \lambda_r > 0$
- Kolumny macierzy U to wektory własne macierzy AA'
- Kolumny macierzy V to wektory własne macierzy $A'A$
- W praktyce z otrzymanego na podstawie tej metody układu współrzędnych do rzutowania interesują nas 2-3 wektory własne.

Pojęcie bewzładności (inercji)

- **Bezwładność (Inercja)** w analizie korespondencji odpowiada pojęciu wariancji.
- Całkowita bezwładność to miara roproszenia profili wokół odpowiednich przeciętnych profili. Całkowita bezwładność wierszy pokazuje, jak bardzo poszczególne profile wierszowe różnią się od przeciętnego profilu wierszowego.

$$\Lambda^2 = \sum m * d^2$$

gdzie m – masa wiersza (kolumny); d^2 – kwadrat odległości między profilem wiersza (kolumny) a odpowiednim przeciętnym profilem.

Inercja całkowita

- Bewzładność dla wierszy jest równa bezwładności dla kolumn. Dlatego najczęściej podaje się tylko jedną wartość nazywaną **bezwładnością (inercją) całkowitą**.
- $\chi^2 = \Lambda^2 n$; powiązanie inercji z wartością testu χ^2 wskazuje, że im mniejsza inercja, tym mniejsza szansa wystąpienia istotnego związku między wierszami i kolumnami tabeli wielodzielczej.
- Jeśli $\Lambda^2 = 0$ wtedy różnica między profilami a profilem przeciętnym jest niewielka, co oznacza niewielkie roproszczenie wokół profilu przeciętnego. Analogicznie wysoka wartość Λ^2 oznacza duże rozproszenie wokół profilu przeciętnego.
- Maksymalna wartość bezwładności to $min(k, w) - 1$.

Związek pomiędzy bezwładnością a wartościami własnymi

$$\Lambda^2 = \sum_{i=1}^{min(w,k)-1} \lambda_i^2$$

- Dzięki tej zależności możemy wybrać liczbę wymiarów odtwarzających jak najpełniejszą informację zawartą w wyjściowej tablicy kontyngencji
- Jeżeli $\frac{\lambda_1^2+\lambda_2^2}{\Lambda^2}$ przyjmuje wartość przekraczającą 0,75 przestrzeń dwuwymiarową można uznać za dobrą reprezentację początkowych danych.
- Najlepsza konfiguracja: dwie pierwsze kolumny macierzy V do reprezentacji kolumn i dwie pierwsze kolumny macierzy U do reprezentacji wierszy macierzy kontyngencji.

Interpretacja wyników

- Analizujemy położenie punktów obrazujących kategorie z wierszy i kolumn tablicy kontyngencji.
- Jeśli okazuje się, że dwuwymiarowe rozwiązanie zapewnia zadowalające dopasowanie, to kategorie wierszowe, które są bliskie sobie mają zbliżony rozkład (profil) w poszczególnych kolumnach. Analogicznie interpretujemy katogorie kolumnowe.
- Badamy rozmieszczenie punktów względem centrum rzutowania oraz punktów odpowiadających kategoriom różnych zmiennych względem siebie.
- Kategorie zmiennych położone na wykresie w niedalekiej odległości od siebie wskazują kombinacje pojawiające się częściej niż jest to oczekiwane przy założeniu niezależności między wierszami i kolumnami.