

Analiza korelacji (współzależności)

Wzory

I. Badanie zależności stochastycznej i korelacyjnej przy pomocy rozkładów warunkowych

Rozkłady warunkowe

Do rozkładu warunkowego bierzemy liczebności i wartości cechy jednej zmiennej przy założeniu, że druga zmienna przyjmuje określoną i stałą wartość.

Parametry z rozkładów warunkowych można oznaczać np.:

$\bar{Y}_{X=350}$ - średnia ze zmiennej Y , przy założeniu, że cecha X przyjmuje wartość 350

$S(X)_{Y=3}$ - odchylenie standardowe ze zmiennej X , przy założeniu, że cecha Y przyjmuje wartość 3

Niezależność stochastyczna

Cechy X i Y są stochastycznie niezależne, jeśli ich wszystkie średnie i wariancje warunkowe są równe.

Niezależność korelacyjna

Cechy X i Y są korelacyjnie niezależne, jeśli ich wszystkie średnie są równe.

II. Test niezależność chi kwadrat

1. Formułujemy hipotezy:

H_0 : cechy X i Y są niezależne

H_1 : cechy X i Y są nie są niezależne

2. Obliczamy statystykę: $\chi^2 = \sum_{i=1}^k \sum_{j=1}^r \frac{(n_{ij} - \hat{n}_{ij})^2}{\hat{n}_{ij}}$

gdzie r i k to liczba możliwych wartości cech X i Y , n_{ij} to liczebności empiryczne w próbie, $\hat{n}_{ij}$ to liczebności teoretyczne liczone ze wzoru:

$$\hat{n}_{ij} = \frac{\text{suma liczebności empirycznych } i - \text{tego wiersza} \times \text{suma liczebności empirycznych } j - \text{tej kolumny}}{\text{ogólna liczebność próby } (n)}$$

3. Tworzymy i rysujemy obszar krytyczny **prawostronny** dla rozkładu chi-kwadrat, dla $(r-1)(k-1)$ stopni swobody, gdzie r i k to liczba możliwych wartości cech X i Y .

4. Sprawdzamy, czy statystyka znalazła się w obszarze krytycznym. Jeśli tak – odrzucamy hipotezę H_0 na rzecz hipotezy alternatywnej H_1 . Jeśli nie – stwierdzamy, że nie ma podstaw do odrzucenia hipotezy H_0 .

Uwaga

Dla dużej liczby stopni swobody (powyżej 30) możemy skorzystać ze statystyki:

$$Z = \sqrt{2\chi^2} - \sqrt{2(\text{liczba wierszy} - 1)(\text{liczba kolumn} - 1) - 1},$$

a obustronny obszar krytyczny odczytać z rozkładu normalnego.

III. Miary siły korelacji

III.1 Współczynnik zbieżności Czuprowa

Obliczamy tą samą statystykę co w teście chi kwadrat:

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

Wzór na współczynnik Czuprowa to:

$$T_{xy} = T_{yx} = \sqrt{\frac{\chi^2}{n\sqrt{(r-1)(k-1)}}$$

Współczynnik Czuprowa przyjmuje wartości od 0 do 1. Im te wartości są bliżej 0, tym bardziej cechy są stochastycznie niezależne. Im bliżej 1, tym bardziej są zależne. 0 oznacza niezależność stochastyczną, a 1 zależność funkcyjną.

Współczynnik determinacji $T_{xy}^2 \cdot 100\%$ określa nam, w jakim procencie zmiany wartości jednej cechy wpływają na zmianę wartości drugiej cechy.

III.2 Wskaźniki korelacyjne Pearsona

Rozkłady brzegowe

Rozkład brzegowy określa wartości jednej zmiennej niezależnie od wartości przyjmowanej przez drugą. W tabeli korelacyjnej liczebności do nich uzyskujemy je poprzez sumowanie wartości z wierszy lub kolumn.

- Wskaźnik korelacyjny zmiennej Y względem zmiennej X jest równy:

$$e_{yx} = \frac{S(\bar{Y}_{x=i})}{S(Y)}, \text{ gdzie } S(\bar{Y}_{x=i}) \text{ oznacza odchylenie standardowe ze średnich wszystkich}$$

rozkładów warunkowych zmiennej Y

- Wskaźnik korelacyjny zmiennej X względem zmiennej Y jest równy:

$$e_{xy} = \frac{S(\bar{X}_{y=j})}{S(X)}, \text{ gdzie } S(\bar{X}_{y=j}) \text{ oznacza odchylenie standardowe ze średnich}$$

wszystkich rozkładów warunkowych zmiennej X

Wskaźniki Pearsona przyjmuje wartości od 0 do 1. Im te wartości są bliżej 0, tym bardziej jedna cecha jest stochastycznie niezależna od drugiej. Im bliżej 1, tym bardziej są zależne.

Współczynniki determinacji $e_{xy}^2 \cdot 100\%$, $e_{yx}^2 \cdot 100\%$ określają nam, w jakim procencie zmiany wartości jednej cechy wpływają na zmianę wartości drugiej cechy.

III.3 Współczynnik korelacji liniowej Pearsona

Najpierw liczymy kowariancję:

$$\text{cov}(X, Y) = \frac{\sum \sum (x_i - \bar{X})(y_j - \bar{Y}) \cdot n_{ij}}{n}$$

Współczynnik korelacji liniowej Pearsona:

$$r_{xy} = r_{yx} = \frac{\text{cov}(X, Y)}{S(X)S(Y)},$$

gdzie $\text{cov}(X, Y)$ oznacza kowariancję, a $S(X), S(Y)$ odchylenia standardowe z rozkładów brzegowych.

Współczynnik korelacji liniowej Pearsona przyjmuje wartości od -1 do 1. Im te wartości są bliżej 0, tym bardziej jedna cecha jest stochastycznie niezależna od drugiej. Im bliżej 1, lub -1 tym bardziej są zależne liniowo. Znak dodatni wskazuje na zależność dodatnią (wzrost jednej cechy powoduje wzrost drugiej), znak ujemny na zależność ujemną (wzrost jednej cechy powoduje spadek drugiej).

Współczynnik determinacji $r_{xy}^2 \cdot 100\%$ określa nam, w jakim procencie zmiany wartości jednej cechy wpływają na zmianę wartości drugiej cechy.