Кореляційний аналіз (взаємозалежність)

Формули

I. Вивчення стохастичних та кореляційних залежностей за допомогою умовних розподілів

Умовні розподіли

До умовного розподілу беремо кількість і значення ознаки однієї змінної, припускаючи, що друга змінна приймає конкретнє і постійне значення.

Параметри з умовних розподілів можна записати наприклад:

$\bar{Y}_{X=350}$ - середнє значення змінної Y, за припущенням, що характеристика X набуває вартость 350

$S\left(X\right)_{Y=3}$ - відхилення стандартової змінної X, за припущенням, що характеристика Y набуває вартость 3

Незалежність стохастична

Характеристики X та Y є стохастично незалежні, якщо всі їх середні та умовні дисперсії є рівні.

Незалежність кореляційна

Характеристики X та Y є кореляційно незалежні, якщо всі їх середні є рівні.

II. Тест незалежності хі-квадрат

1. формулюємо гіпотези:

H_0: характеристики X та Y є незалежні

H_1: характеристики X та Y є не незалежні

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

де r та k кількість можливих значень ознак X та Y, n_{ij} емпіричні числа у вибірці, $\hat{n}_{ij}$ це теоретичні числа, розраховані за формулою:

$$\hat{n}_{ij} = \frac{\textbf{сума емпіричних чисел}i\textbf{ цього рядка} \times \textbf{сума емпіричних чисел}j\textbf{ цього стовпчика}}{\textbf{загальний розмір вибірки}(n)}$$

3. Створюємо та рисуємо **праву** критичну область для розподілу хі-квадрат, для $(r-1)(k-1)$ ступенів свободи, де r та k кількість можливих значень ознак X та Y.

4. Перевіряємо, чи була статистика в критичній області. Якщо так – відкидаємо гіпотезу H_0 на користь альтернативної гіпотези H_1. Якщо ні – стверджуємо, що немає підстав до віхилення гіпотези H_0.

Увага

Для великої кількості ступенів свободи (понад 30) можна використовувати статистику:

$$Z = \sqrt{2\chi^2} - \sqrt{2\left(\textbf{кількість рядків}-1\right)\left(\textbf{кількість стовпчиків}-1\right)-1},$$

та двостронна критична область може бути прочитана з нормального розподілу.

III. Міри сили кореляції

III.1 Коефіцієнт збіжності Чупрова

Розраховуємо ту саму статистику, що й у тесті хі-квадрат:

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

Формула коефіцієнту Чупрова:

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

Коефіцієнт Чупрова коливається від 0 до 1. Чим ближче ці значення до 0, тим більше ознаки стохастично незалежні. Чим ближче до 1, тим більше залежні. 0 вказує на незалежність стохастичну, а 1 залежність функційну.

Коефіцієнт детермінації $T_{xy}^2 \cdot 100\%$ визначає нам, у якому відсотку зміни значення однієї ознаки змінюють значення іншої ознаки.

III.2 Індекси кореляції Пірсона

Розклад прикордонний

Прикордонний розподіл визначає значення однієї змінної незалежно від значення, припущеного перед іншим. В таблиці співвідношення чисел до них ми отримуємо їх шляхом підсумовування значень з рядків або стовпців.

- Індекс кореляції змінної Y відносно змінної X дорівнює:
  $e_{yx} = \frac{S\left(\overline{Y}_{x=i}\right)}{S\left(Y\right)}$, де $S\left(\overline{Y}_{x=i}\right)$ означає стандартове відхилення від середніх усіх умовних розподілів змінної Y.
- Індекс кореляції змінної X відносно змінної Y дорівнює:
  $e_{xy} = \frac{S\left(\overline{X}_{y=j}\right)}{S\left(X\right)}$, де $S\left(\overline{X}_{y=j}\right)$ є стандартним відхиленням середнього всіх умовних розподілів змінної X.

Індекси Пірсона мають значення від 0 до 1. Чим ближче ці значення до 0, тим більше одна ознака стохастично незалежна від іншої. Чим ближче 1, тим більше залежні.

Коефіцієнт детермінації $e_{xy}^2 \cdot 100\%, e_{yx}^2 \cdot 100\%$ визначають нам, на який процент змінюється значення однієї ознаки змінюють значення іншої ознаки.

III.3 Коефіцієнт кореляції лінійної Пірсона

Спочатку визначаємо коваріції:

$$\operatorname{cov}(X,Y)=\frac{\sum\sum\left(x_i-\overline{X}\right)\left(y_j-\overline{Y}\right)\cdot n_{ij}}{n}$$

Коефіцієнт кореляції лінійної Пірсона:

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

де $\operatorname{cov}(X,Y)$ означа коваріцію, а $S(X), S(Y)$ відхилення стандартове з розкладів прикордонних.

Коефіцієнт кореляції лінійної Пірсона приймує значення від -1 до 1. Чим ці значення ближчі 0, тим більше одна ознака є стохастичне незалежна від другої. Чим ближче 1, чи -1 тим більше залежні лінійно. Знак плюс показує на залежність додатню (збільшення однієї ознаки призводить до збільшення другої), знак мінус на залежність від'ємну (збільшення однієї ознаки призводить до зменшення другої).

Коефіцієнт детермінації $r_{xy}^2\cdot 100\%$ визначає нам, на який процент змінюється значення однієї ознаки змінюють значення іншої ознаки.