Analiza Wielowymiarowa

Analiza związku między zmiennymi

Dorota Celińska, Paweł Strawiński

Zajęcia 3
20 października 2015

1. Analiza korelacji
   - Korelacja
   - Miary zależności dwóch cech

2. Analiza zróżnicowania
   - Jednoczynnikowa analiza wariancji
   - Wieloczynnikowa analiza wariancji

Definicja korelacji

- Według definicji słownikowej korelacja oznacza współwystępowanie
- Korelacja może być traktowana jako miara wzajemnego „dopasowania” zmiennych losowych
- Analiza korelacji jest metodą wykrywania występowania statystycznej zależności między zmiennymi

Współczynnik korelacji

- Współczynnik korelacji jest unormowaną miarą kowariancji (wspólnej wariancji)
- Ogólny wzór na współczynnik korelacji

$$corr(X,Y) = \rho_{XY} = \frac{COV(X,Y)}{\sqrt{VarX}\sqrt{VarY}}$$

- Wzór ma sens wyłącznie dla zmiennych losowych o skończonych dwóch pierwszych momentach, czyli zmiennych losowych których rozkłady są stacjonarne w sensie słabym

Własności współczynnika korelacji

- Wartości współczynnika korelacji są ograniczone

$$-1 \leq \rho_{XY} \leq 1$$

- Wartość współczynnika korelacji jest niezmiennicza względem przekształcenia afinicznego zmiennych losowych

$$\rho_{XY} = \rho_{(a+bX)(\alpha+\beta Y)}$$

- Wartość bezwzględna współczynnika korelacji wynosi 1, gdy związek między zmiennymi losowymi jest liniowy

Test Chi2 Pearsona

- Może być wykorzystany do
  - sprawdzania, czy rozkład empiryczny zmiennej losowej jest zgodny z rozkładem teoretycznym
  - sprawdzenia, czy rozkłady zmiennych losowych przedstawione w formie tablicy kontyngencji (tablicy krzyżowej) są niezależne
- Statystyka testowa ma postać

$$\chi^2 = \sum_{i=1}^{I} \sum_{j=1}^{J} \frac{n_{ij} - En_{ij}}{En_{ij}} \sim \chi^2((I-1)(J-1))$$

gdzie:
i jest liczbą wierszy w tabeli krzyżowej
j jest liczbą kolumn w tabeli krzyżowej
n_{ij} liczbą obserwacji w komórce ij
En_{ij} oczekiwaną liczbą obserwacji w komórce ij

Współczynnik V-Cramera

- Statystyka V-Cramera jest unormowaną wersją statystyki χ^2
- Statystyka testowa ma postać

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

- Dla tablicy o wymiarach 2X2, stosowany jest inny wzór i wówczas statystyka testowa przyjmuje wartości od -1 do 1
- Dla tablicy o większych wymiarach statystyka testowa przyjmuje wartości od 0 do 1

Współczynnik korelacji Pearsona

- Służy do wyznaczania siły współzależności zmiennych o rozkładzie ciągłym
- Niech $\mathbb{X}$ oraz $\mathbb{Y}$ będą zmiennymi losowymi o rozkładzie ciągłym
- Niech $\bar{x} = \frac{\sum_{i=1}^{n} x_i}{n}$, $\bar{y} = \frac{\sum_{i=1}^{n} y_i}{n}$
- Współczynnik korelacji dany jest wówczas wzorem

$$r_{XY} = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_{i=1}^{n}(x_i - \bar{x})^2}\sqrt{\sum_{i=1}^{n}(y_i - \bar{y})^2}}$$

- Jego wartość informuje o sile i kierunku zależności liniowej
- Dla zmiennych o rozkładzie normalnym jest zgodnym estymatorem współczynnika korelacji
- Jego wartości mogą być zniekształcone przez występowanie obserwacji o wartościach skrajnych

Współczynnik korelacji Kendalla

- O każdej parze (x_i, y_i), (x_j, y_j) mówimy, że jest zgodna, jeżeli

$$(x_i - x_j)(y_i - y_j) > 0$$

- W przeciwnym przypadku mówimy, że jest niezgodna
- Niech
  - P będzie liczbą par zgodnych
  - Q będzie liczbą par niezgodnych
  - N będzie liczebnością próby
- Wówczas statystyka testowa ma postać

$$\tau = 2\frac{P - Q}{N(N-1)} \sim N(0,1)$$

Współczynnik korelacji Kendalla

- Mierzy siłę zależności monotonicznej
- Jest współczynnikiem nieparametrycznym, jego wartość nie zależy od rozkładu zmiennych
- Do obliczenia jego wartości wykorzystywane są rangi obserwacji, dzięki czemu wartości współczynnika są odporne na występowanie obserwacji odstających

Współczynnik korelacji rang Spearmana

- Jest to współczynnik korelacji Pearsona obliczony dla rang wartości zmiennych losowych
- Mierzy siłę zależności monotonicznej
- Jest miarą siły związku liniowego i wyłącznie do takich związków powinien być stosowany

Analiza zróżnicowania

- Analiza zróżnicowania, w języku angielskim *Analysis of variance (ANOVA)* to ogólna nazwa dla grupy modeli statystycznych używanych do analizy różnic w średnich wartościach cechy pomiędzy grupami
- Zróżnicowanie zmiennej jest dzielone na składowe, które można przypisać różnym czynnikom
- Historycznie została zaproponowana przez R. A. Fishera w celu analizy danych eksperymentalnych

Równość analizy wariancji

- Analiza wariancji oparta jest na równości analizy wariancji

$$TSS = ESS + RSS$$

- Niech: N jest liczebnością próby, J jest liczbą wartości w przypadku zmiennej nominalnej lub przedziałowej, dla zmiennej ciągłej $J = 1$
  - TSS jest całkowitą sumą kwadratów, ma $N - 1$ stopni swobody.
  - ESS jest wyjaśnioną sumą kwadratów, ma $J - 1$ stopni swobody.
  - RSS jest resztową sumą kwadratów, ma $N - J$ stopni swobody.

Jednoczynnikowa analiza wariancji (1)

- Jednoczynnikowa analiza wariancji jest testem statystycznym równości średnich w grupach
- Może być traktowana jako uogólnienie testu t na większą liczbę grup
- Wielokrotne przeprowadzenie testu t powodowałoby powstanie obciążenia Lovella
- Założenia analizy
    - cecha w każdej grupie ma rozkład normalny
    - grupy są niezależne
    - grupy mają cechy losowych prób prostych
    - wariancje w grupach są równe

Jednoczynnikowa analiza wariancji (2)

- Porównywana jest wariancja wewnątrzgrupowa i międzygrupowa
- Statystyka weryfikująca hipotezę o równości średnich jest ilorazem wariancji międzygrupowej (resztowej) i wewnątrzgrupowej (wyjaśnionej)

$$F = \frac{ESS}{RSS} \sim F(J, N - J)$$

- Nieparametrycznym odpowiednikiem jednoczynnikowej analizy wariancji jest test Kruskalla-Wallisa

Wieloczynnikowa analiza wariancji

- Celem wieloczynnikowej analizy wariancji jest zbadanie wpływu więcej niż jednego czynnika na całkowite zróżnicowanie
- Dodatkowo w analizie uwzględniane są efekty interakcji między czynnikami
- Hipotezą zerową jest nadal brak zróżnicowania średnich
- Wzory komplikują się, wymuszają obliczenie sum kwadratów związanych z każdym czynnikiem i każdą interakcją