Statystyczna Analiza Danych – laboratorium

Znajdowanie zależności/współwystępowania

Dorota Celińska-Kopczyńska

Uniwersytet Warszawski

Zajęcia 9
4/5 maja 2023

Idea zajęć – co i po co będziemy robić?

- Zainteresowali się Państwo tematem analizy zależności/analizy korelacji
- W szczególności zauważyli Państwo, że nie każda para zmiennych może być analizowana z wykorzystaniem narzędzi z kursu
- Przedstawimy kilka dodatkowych narzędzi (materiał rozszerzający), w szczególności przeznaczonych do pracy ze zmiennymi jakościowymi

Rangowe współczynniki korelacji

- Próbę na wstępie rangujemy – każda wartość każdej cechy jest zastępowana jej pozycją (rangą) na uporządkowanej rosnąco liście wszystkich wartości tej cechy
- Są metodami nieparametrycznymi, mierzą siłę zależności monotonicznych
- Nie wymagają żadnych założeń dotyczących rozkładu w populacji
- Są odporne na występowanie obserwacji odstających
- Mogą stanowić alternatywę dla korelacji Pearsona (w przypadku zmiennych ilościowych)
- Przykłady:
  - korelacja rang Spearmana (dokładnie omówiona w trakcie wykładu!)
  - korelacja Kendalla-τ
  - korelacja Kruskalla-γ

Pary zgodne i niezgodne

- O każdej parze (x_i, y_i), (x_j, y_j) mówimy, że jest zgodna, jeżeli $(x_i - x_j)(y_i - y_j) > 0$.
- O każdej parze (x_i, y_i), (x_j, y_j) mówimy, że jest niezgodna, jeżeli $(x_i - x_j)(y_i - y_j) < 0$.
- Niech:
  - P – liczba par zgodnych
  - Q – liczba par niezgodnych
  - N – liczebność próby

Współczynnik korelacji Kendalla-τ

- Każdą parę (x_i, y_i), (x_j, y_j), dla której $(x_i - x_j)(y_i - y_j) = 0$ traktujemy jako niezgodną
- Statystyka testowa ma postać:

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

- Może być wykorzystany jako alternatywa dla Spearmana, gdy występuje wiele remisów (*tied ranks*)

Współczynnik korelacji Kruskalla-γ

- Remisy usuwamy,

$$G = \frac{P - Q}{P + Q}$$

- Statystyka testowa o postaci

$$t \approx G\sqrt{\frac{P + Q}{N(1 - G^2)}}$$

- Może być wykorzystany jako alternatywa dla Kendalla, gdy występuje wiele remisów (*tied ranks*)

Jak policzyć w R?

```
# wspolczynnik Pearsona jest domyslnie zaimplementowany

# wspolczynnik Spearmana

cor.test(data$V1, data$V2, method="spearman")

# wspolczynnik Kendalla-tau

cor.test(data$V1, data$V2, method="kendall")

# wspolczynnik Kruskalla-gamma
library(MESS)
gkgamma(tablica)
```

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 jakościowymi 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 kontyngencji Pearsona.
  3. Współczynnik zbieżności V-Cramera;

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 liczebność próby.

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.

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.
- Preferowany względem poprzednich miar

Jak policzyć w R?

```
# na poczatek wykonujemy test chi2 -- na tej podstawie wnioskujemy o niezaleznosci lub nie!

# wspolczynnik phi, wspolczynnik kontyngencji -- do policzenia recznie

# wspolczynnik V-Cramera
library(rcompanion)
cramerV(data)
cramerV(data, ci = TRUE)
```

Dla zainteresowanych – analiza korespondencji

- 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.

Dla zainteresowanych – dane nieprzekrojowe

- W tym kursie zajmujemy się danymi, które pochodzą z prób przekrojowych, każda z obserwacji może być wylosowana do próby niezależnie od pozostałych
- Analiza współwystępowania zmienia się jednak, gdy dopuścimy powiązania pomiędzy poszczególnymi obserwacjami (czasowe lub jakościowe)
  - Analiza szeregów czasowych
  - Analiza sieci społecznych
  - Statystyka przestrzenna
- ... i pewnie jeszcze wiele innych.