Analiza Wielowymiarowa

Testy parametryczne i nieparametryczne

Maciej Nasiński, Paweł Strawiński

Uniwersytet Warszawski

Zajęcia 3
20 października 2022

Test t

- Najprostszym sposobem porównania średnich jest wykorzystanie testu opartego na statystyce o rozkładzie t-Studenta
- Niech zbiór $\mathbb{X}$ liczy n obserwacji, a zbiór $\mathbb{Y}$ m obserwacji
- Wówczas przy prawdziwej H_0 o równości średnich

$$t = \frac{\bar{X} - \bar{Y}}{\hat{var}} \sim t(n + m - 2)$$

- Gdzie $\hat{var}$ jest wariancją zmiennej w połączonych zbiorach
- Uwaga, gdy X lub Y nie ma rozkładu normalnego to rozkład statystyki testowej może różnić się od zakładanego

Test proporcji

- Test przeznaczony do weryfikowania hipotez o równości proporcji
- Niech zbiór $\mathbb{X}$ liczy n obserwacji, a zbiór $\mathbb{Y}$ m obserwacji
- Wówczas przy prawdziwej H_0 o równości proporcji

$$z = \frac{p_x - p_y}{\sqrt{\hat{p}(1-\hat{p})(\frac{1}{n}+\frac{1}{m})}} \sim N(0,1)$$

- gdzie $\hat{p}$ jest udziałem sukcesów w połączonych zbiorach

Test F

- Najprostszym sposobem porównania wariancji jest wykorzystanie statystyki o rozkładzie F
- Niech zbiór $\mathbb{X}$ liczy n obserwacji, a zbiór $\mathbb{Y}$ m obserwacji
- Wówczas przy prawdziwej H_0 o równości wariancji

$$F = \frac{S_x^2}{S_y^2} \sim F(n-1, m-1)$$

- Ale rozkład tej statystyki jest czuły na spełnienie założenia o normalności rozkładu

Test Levena

- Levene (1960) zaproponował test równości wariancji odporny na brak normalności rozkładu analizowanej cechy
- Brown i Forsythe (1974) zaproponowali by w teście średnią zastąpić medianą która jest bardziej odporną miarą tendencji centralnej
- Ta poprawka jest istotna w przypadku skośnych rozkładów zmiennych
- Statystyka oparta jest o odchylenia wartości zmiennych od średnich w grupach

Statystyka testowa

- Niech X_{ij} będzie obserwacją j w grupie i
- Niech $Z_{ij} =\mid X_{ij} - \bar{X}_i \mid$, gdzie $\bar{X}_i$ jest średnią wartością zmiennej w grupie i.

$$W_0 = \frac{\frac{\sum_i n_i (Z_i - \bar{Z})^2}{(g-1)}}{\frac{\sum_i (Z_{ij} - \bar{Z}_i)^2}{\sum_i (n_i - 1)}}$$

- gdzie g to liczba grup, a n_i oznacza liczebność grupy i

Tabela krzyżowa

- Tabela kontyngencji (krzyżowa) jest typem tabeli w formacie macierzy, która wyświetla (wielowymiarowy) rozkład częstości zmiennych
- Dostarcza podstawowego obrazu wzajemnych relacji między dwiema zmiennymi i może pomóc w znalezieniu interakcji między nimi
- Termin tabela kontyngencji został po raz pierwszy użyty przez Karla Pearsona w pracy z 1904 roku *On the Theory of Contingency and Its Relation to Association and Normal Correlation*
- Problemem jest znalezienie struktury (bezpośredniej) zależności leżącej u podstaw zmiennych zawartych w wielowymiarowych tablicach krzyżowych
- Tabela przestawna to sposób na tworzenie tabel krzyżowych z wykorzystaniem arkusza kalkulacyjnego

1978	Domestic	Foreign	Total
1	2	0	2
2	8	0	8
3	27	3	30
4	9	9	18
5	2	9	11
Razem	48	21	69

Test Kruskala-Wallisa

- Test Kruskala-Wallisa jest uogólnieniem testu Manna-Whitneya na większą liczbę grup
- Test wykorzystuje rangowanie obserwacji
- Wzór statystyki testowej jest skomplikowany. Jeżeli nie występują obserwacje o identycznych rangach to niech n będzie liczbą obserwacji, n_j liczbą obserwacji z w zbiorze j, a R_j będzie sumą rang w j-tym zbiorze:

$$KW = \frac{12}{n(n+1)} \sum_{j=1}^{J} \frac{R_j^2}{n_j} - 3(n+1) \ \sim_a \ \chi^2(J-1)$$

- Test może być również traktowany jako nieprarametryczny odpowiednik jednoczynnikowej analizy wariancji

Test Kołmogorowa-Smirnowa

- Jest wykorzystywany do porównywania rozkładów jednowymiarowych cech statystycznych.
- Test ma dwie wersję:
  - dla jednej grupy, służy do weryfikacji hipotezy czy dana zmienna na określony rozkład. Ta wersja nazywana jest testem zgodności Kołmogorowa.
  - dla dwóch grup, służący do weryfikacji hipotezy czy rozkład zmiennej w dwóch grupach jest identyczny