Metody nieparametryczne w statystyce

Tomasz Wójtowicz

Wydział Zarządzania
AGH Akademia Górniczo-Hutnicza im. S. Staszica w Krakowie

Metody rangowe

Współczynnik korelacji Spearmana

Gdy mamy dane pary (X_i, Y_i) dla $i = 1, \dots n$ opisujące cechy o rozkładzie ciągłym. Niech R_i będą rangami X_i, a Q_i niech będą rangami Y_i.

Współczynnik Spearmana ma postać:

$$r_s = 1 - \frac{6\sum_{i=1}^{n} d_i^2}{n(n^2-1)},$$

gdzie $d_i = R_i - Q_i$.

Badanie zależności dwóch cech

Własności współczynnika Spearmana:

- mierzy siłę zależności monotonicznej pomiędzy X i Y,
- jeżeli X i Y są niezależne, to $r_s = 0$,
- $-1 \leq r_s \leq 1$
- jeżeli $r_s \approx 1$, to wzrostowi wartości X odpowiada wzrost wartości Y,
- jeżeli $r_s \approx -1$, to wzrostowi wartości X odpowiada spadek wartości Y,
- Jeżeli X i Y są niezależne, to rozkład r_s **nie zależy** od rozkładu X i Y.

Badanie zależności dwóch cech

Gdy cechy (X, Y) pochodzą z rozkładów skokowych, wtedy wśród rang R_i, Q_i mogą wystąpić rangi wiązane. Wtedy współczynnik korelacji Spearmana ma postać:

$$r_s = \frac{\frac{1}{6}(n^3 - n) - \left(\sum_{i=1}^{n} d_i^2\right) - T_X - T_Y}{\sqrt{\left(\frac{1}{6}(n^3 - n) - 2T_X\right)\left(\frac{1}{6}(n^3 - n) - 2T_Y\right)}}$$

gdzie:

$$T_X = \frac{1}{12}\sum_{j=1}^{k}\left(t_j^3 - t_j\right) \qquad T_Y = \frac{1}{12}\sum_{j=1}^{l}\left(s_j^3 - s_j\right)$$

a t_j liczba wartości X_i posiadających tę samą rangę R_j, s_j - liczba Y_i posiadających tę samą rangę Q_j.

Badanie zależności dwóch cech

Na podstawie r_s można badać istotność korelacji pomiędzy cechami X i Y.

$$H_0: \rho = 0$$

$$H_1: \rho \neq 0$$

W przypadku małych prób do weryfikacji tych hipotez można bezpośrednio stosować statystykę r_s, której rozkład jest wtedy stablicowany.

Dla dużych n można zastosować statystykę:

$$t = \frac{r_s}{\sqrt{1 - r_s^2}} \sqrt{n - 2}$$

która ma asymptotycznie rozkład t-Studenta o $n - 2$ stopniach swobody.

Badanie zależności dwóch cech

Współczynnik Kendalla

$$r_K = P\left((X_i - X_j)(Y_i - Y_j) > 0\right) - P\left((X_i - X_j)(Y_i - Y_j) < 0\right)$$

czyli:

$$r_K = 2\frac{P - Q}{n(n-1)},$$

gdzie

P - liczba par zgodnych,

Q - liczba par niezgodnych

Para (X_i, Y_i) jest zgodna z parą (X_j, Y_j) jeżeli $\left(X_i - X_j\right)\left(Y_i - Y_j\right) > 0$

Para (X_i, Y_i) jest niezgodna z parą (X_j, Y_j) jeżeli $\left(X_i - X_j\right)\left(Y_i - Y_j\right) < 0$

Badanie zależności dwóch cech

Jeżeli X i Y są niezależne, to rozkład r_K **nie zależy** od rozkładu cech X i Y.

Wtedy

$$E(r_K) = 0$$

$$D^2(r_K) = \frac{2(2n+5)}{9n(n-1)}$$

Do badania istotności korelacji pomiędzy X i Y można wykorzystać:

- dla małych n: r_K bo jej wartości są stablicowane,
- dla dużych n: statystykę

$$t = \frac{r_K}{\sqrt{1-r_K^2}}\sqrt{n-2}$$

Porównanie rozkładów wielu cech

Rozważmy problem analogiczny do przypadku testu Wilcoxona. Tym razem badamy równość rozkładu k cech:

$$H_0 : F_1 = F_2 = \cdots = F_k$$

$$H_1 : \text{dla pewnych } i, j : \forall x \in \mathbb{R} \ \ F_i(x) \leq F_j(x) \text{ oraz } F_i \neq F_j$$

Każda i-ta cecha (spośród nadanych k cech) jest reprezentowana przez n_i-elementową próbę $Y_{i1}, \ldots, Y_{in_i}$, przy czym $n = n_1 + \cdots + n_k$.

W przypadku, gdy rozważane cechy mają rozkład normalny o takich samych odchyleniach standardowych, to powyższe hipotezy możemy weryfikować za pomocą analizy wariancji (ANOVA) przy pomocy statystyki:

$$F = \frac{\sum_{i=1}^{k} n_i (\bar{y}_i - \bar{y})^2}{\sum_{i=1}^{k} \sum_{j=1}^{n_i} \left(y_{ij} - \bar{y}_i \right)^2} \cdot \frac{n - k}{k - 1}$$

Porównanie rozkładów wielu cech

$$H_0\colon F_1 = F_2 = \cdots = F_k$$

$$H_1\colon F_i(x) < F_j(x) \text{ dla pewnych } i, j$$

Jeżeli założenia ANOVA nie są spełnione to możemy zastosować **test Kruskala-Wallisa.**

Każda i-ta cecha (spośród nadanych k cech) jest reprezentowana przez n_i-elementową próbę $Y_{i1}, \dots, Y_{in_i}$, przy czym $n = n_1 + \cdots + n_k$.

Niech $R_{i1}, \dots R_{in_i}$ oznaczają rangi elementów $Y_{i1}, \dots, Y_{in_i}$ i-tej próby wśród wszystkich elementów połączonych k-prób.

Niech $\bar{R}_i$ oznacza średnią arytmetyczną rang $R_{i1}, \dots R_{in_i}$

Porównanie rozkładów wielu cech

Statystyka **Kruskala-Wallisa** ma postać:

$$T = \frac{12}{n(n+1)} \sum_{i=1}^{K} n_i \left[\bar{R}_i - \frac{1}{2}(n+1) \right]^2 =$$

$$= \frac{12}{n(n+1)} \sum_{i=1}^{K} n_i \bar{R}_i^{\,2} - 3(n+1)$$

Statystyka ta mierzy odstępstwo średnich $\bar{R}_i$ od średniej $\frac{n+1}{2}$.

Stanowi monotoniczne odwzorowanie statystyki F (w ANOVA) obliczonej na podstawie rang.

Porównanie rozkładów wielu cech

Jeżeli badane cechy mają rozkłady ciągłe i prawdziwa jest hipoteza H_0, to:

- rozkład T nie zależy od rozkładów F_i,
- statystyka T ma asymptotycznie (gdy $\min\{n_1, \dots, n_k\} \to +\infty$) rozkład χ^2 o $k-1$ stopniach swobody.

Przybliżenie rozkład T rozkładem χ^2 o $k-1$ st. swobody stosuje się już:

- dla $k = 3$, gdy wszystkie $n_i > 5$,
- dla $k > 3$, gdy wszystkie $n_i > 4$.

Porównanie rozkładów wielu cech

W przypadku odrzucenia hipotezy głównej w teście Kruskala-Wallisa przeprowadzamy analizę *post-hoc* by za pomocą porównań wielokrotnych uszeregować rozkłady $F_1, \dots, F_k$ i odkryć, dla których z nich zachodzi istotna nierówność

$$F_i < F_j.$$

Zauważmy, że jeżeli $F_i(x) \leq F_j(x)$ dla wszystkich $x \in \mathbb{R}$, to taka sama nierówność zachodzi dla wartości oczekiwanych rang, tzn,:

$$E(R_{il}) \leq E(R_{jm})$$

gdzie R_{il} jest rangą pewnego Y_{il} z próby $Y_{i1}, \dots, Y_{in_i}$, a R_{jm} jest rangą pewnego Y_{jm} z próby $Y_{j1}, \dots, Y_{jn_j}$.

Porównanie rozkładów wielu cech

Aby sprawdzić nierówność

$$E(R_{il}) \leq E(R_{jm})$$

możemy skonstruować przedział ufności dla $E(R_{il}) - E(R_{jm})$.

Jeżeli przedział ten (wyznaczony dla współczynnika ufności $1 - \alpha$) będzie

leżał na prawo od 0, to na poziomie istotności α można będzie odrzucić hipotezę $F_i = F_j$ na rzecz hipotezy alternatywnej:

$$\forall x \in \mathbb{R} \;\; F_i(x) \leq F_j(x) \text{ oraz } F_i \neq F_j.$$

Porównanie rozkładów wielu cech

Przy założeniu prawdziwości hipotezy H_0 dla dowolnych $i, j \in \{1, \dots, k\}$:

$\bar{R}_i - \bar{R}_j$ ma w przybliżeniu rozkład normalny:

$$N\left(0, \sqrt{\frac{n(n+1)}{12} \cdot \left(\frac{1}{n_i} + \frac{1}{n_j}\right)}\right)$$

Na tej podstawie przedział ufności ma postać:

$$\left(\bar{R}_i - \bar{R}_j - u_\alpha \sqrt{\frac{n(n+1)}{12} \cdot \left(\frac{1}{n_i} + \frac{1}{n_j}\right)}, \bar{R}_i - \bar{R}_j + u_\alpha \sqrt{\frac{n(n+1)}{12} \cdot \left(\frac{1}{n_i} + \frac{1}{n_j}\right)}, \right)$$

Porównanie rozkładów wielu cech

Konstrukcja tego przedziału ufności jest równoważna przeprowadzeniu testu Manna-Whitneya-Wilcoxona dla każdej pary i, j.

Po odrzuceniu hipotezy głównej w teście Kruskala-Wallisa dokonujemy $p = \frac{k(k-1)}{2}$ porównań. W związku z tym, jeżeli test przeprowadzamy przy poziomie istotności α, to przy konstrukcji pojedynczego przedziału ufności należy zastosować poprawkę Boferroniego i zastąpić α przez α/p.

W przypadku dużego p to procedura może być mało efektywna, bo wtedy α/p jest małe.