Metody nieparametryczne w statystyce

Tomasz Wójtowicz

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

Metody rangowe

Metody rangowe

Niech X i Y będą dwoma cechami o rozkładach określonych przez dystrybuanty F i G.

Chcemy porównać te rozkłady na podstawie prób losowych $X_1, \dots, X_{n_1}$ i $Y_1, \dots, Y_{n_2}$.

Czyli chcemy zweryfikować hipotezę, że:

$$F(x) = G(x) \text{ dla wszystkich } x \in \mathbb{R}.$$

Metody rangowe

Jeżeli: X ma rozkład $N(m_1, \sigma)$ i Y ma rozkład $N(m_2, \sigma)$ (czyli mają to samo odchylenie stand., ale różnią się wartościami oczekiwanymi) równość rozkładów X i Y jest równoważna równości:

$$m_1 = m_2$$

hipoteza alternatywna może mieć jedną z poniższych postaci:

$$m_1 \neq m_2$$

$$m_1 < m_2$$

$$m_1 > m_2$$

Metody rangowe

Do weryfikacji zestawu hipotez

$$H_0: m_1 = m_2$$

$$H_1: m_1 < m_2$$

można zastosować statystykę t-Studenta:

$$t = \frac{\bar{X} - \bar{Y}}{\sqrt{s_p^2 \left(\frac{1}{n_1} + \frac{1}{n_2}\right)}}$$

gdzie

$$s_p^2 = \frac{1}{n_1 + n_2}\left((n_1 - 1)S_X^2 + (n_2 - 1)S_Y^2\right)$$

Metody rangowe

Posługując się pojęciem dystrybuanty, hipotezy:

$$H_0: m_1 = m_2$$

$$H_1: m_1 < m_2$$

można zapisać w równoważnej postaci jako:

$$H_0: F(x) = G(x) \text{ dla wszystkich } x \in \mathbb{R}.$$

$$H_1: F(x) \geq G(x) \text{ dla wszystkich } x \in \mathbb{R} \text{ i } F \neq G$$

gdzie F jest dystrybuantą X, a G jest dystrybuantą Y.

Metody rangowe

Z kolei hipotezy:

$$H_0: F(x) = G(x) \text{ dla wszystkich } x \in \mathbb{R}.$$

$$H_1: F(x) \geq G(x) \text{ dla wszystkich } x \in \mathbb{R} \text{ i } F \neq G$$

można zapisać dokładniej jako

$$H_0: F(x) = G(x) \text{ dla wszystkich } x \in \mathbb{R}.$$

$$H_1: \mathrm{G}(x) = F(x - \delta) \text{ dla wszystkich } x \in \mathbb{R} \text{ i } \delta > 0$$

Jeżeli rozkłady cech X i Y odbiegają od rozkładu normalnego lub nie są znane do weryfikacji powyższych hipotez możemy posłużyć się **testem Wilcoxona**.

Test Wilcoxona

Procedura testu Wilcoxona

- mam dane: $x_1, \ldots, x_{n_1}$ i $y_1, \ldots, y_{n_2}$
- łączymy je w jedną próbę połączoną: $x_1, \ldots, x_{n_1}, y_1, \ldots, y_{n_2}$
- nadajemy rangi wszystkim obserwacjom w próbie połączonej
- niech $r_1, \ldots, r_{n_2}$ będą rangami $y_1, \ldots, y_{n_2}$ w próbie połączonej
- statystyka Wilcoxona jest równa

$$W = \sum_{i=1}^{n_2} r_i$$

- hipotezę H_0 odrzucamy, gdy statystyka W jest odpowiednio duża

Test Wilcoxona

Jeżeli $F = G$ i dystrybuanta F jest **ciągła**, to:

$$\frac{n_2(n_2+1)}{2} \leq W \leq n_1 n_2 + \frac{n_2(n_2+1)}{2}$$

Ponadto:

1. rozkład statystyki Wilcoxona W nie zależy od dystrybuanty F
2. $E(W) = \frac{n_2(n_2+n_1+1)}{2}$, $D^2(W) = \frac{n_1 n_2(n_2+n_1+1)}{12}$
3. dla dowolnego $x \in \mathbb{R}$:

$$P\left(\frac{W - EW}{\sqrt{D^2(W)}} \leq x\right) \to \Phi(x)$$

gdy $\min(n_1, n_2) \to +\infty$, a Φ jest dystrybuantą rozkładu $N(0,1)$.

Test Wilcoxona

Test Wilcoxona można też stosować do weryfikacji hipotez:

$$H_0\colon F(x) = G(x) \text{ dla wszystkich } x \in \mathbb{R}.$$

$$H_1\colon F(x) \leq G(x) \text{ dla wszystkich } x \in \mathbb{R} \text{ i } F \neq G$$

(czyli rozkład X jest bardziej na lewo względem Y)

co można też zapisać:

$$H_0\colon F(x) = G(x) \text{ dla wszystkich } x \in \mathbb{R}.$$

$$H_1\colon \mathrm{G}(x) = F(x - \delta) \text{ dla wszystkich } x \in \mathbb{R} \text{ i } \delta < 0$$

Test Wilcoxona

lub też do hipotez:

$$H_0\colon F(x) = G(x) \text{ dla wszystkich } x \in \mathbb{R}.$$

$$H_1\colon F(x) \neq G(x) \text{ dla wszystkich } x \in \mathbb{R} \text{ i } F \neq G$$

(czyli rozkład X jest przesunięty względem Y)

co można też zapisać:

$$H_0\colon F(x) = G(x) \text{ dla wszystkich } x \in \mathbb{R}.$$

$$H_1\colon \mathrm{G}(x) = F(x - \delta) \text{ dla wszystkich } x \in \mathbb{R} \text{ i } \delta \neq 0$$

Test Wilcoxona

Test Manna-Whitneya

- służy do weryfikacji tych samych hipotez co test Wilcoxona,
- statystyka U jest równa liczbie takich par (x_i, y_j), że $y_j > x_i$
- $E(U) = \frac{n_1 n_2}{2}$, $D^2(U) = \frac{n_1 n_2 (n_2 + n_1 + 1)}{12}$
- można pokazać, że:

$$U = W - \frac{n_2(n_2 + 1)}{2}$$

- czyli statystyki U i W są sobie równoważne.

Test Wilcoxona

Test Wilcoxona można stosować również do weryfikacji hipotezy głównej, że rozkład G powstaje przez przesunięcie rozkładu F w prawo o $\Delta > 0$, tzn.:

$$G(x) = F(x - \Delta)$$

Wtedy odpowiednia hipoteza ma postać:

$$H_0: \Delta = \Delta_0.$$

W poprzednich przypadkach mieliśmy:

$$H_0: \Delta = 0.$$

Tę nową hipotezę można zweryfikować stosując test Wilcoxona do danych: $x_1, \dots, x_{n_1}$ i $y_1 - \Delta_0, \dots, y_{n_2} - \Delta_0$.

Metody rangowe

Jeżeli rzeczywiście byłoby

$$G(x) = F(x - \Delta)$$

to mediana rozkładu zmiennej $\left(Y_j - \Delta\right) - X_i$ byłaby równa 0.

Z tego wynika, że naturalnym estymatorem parametru przesunięcia Δ jest mediana z próby

$D_{ji} = Y_j - X_i$ dla $i = 1, \dots, n_1, j = 1, \dots, n_2$.

Jest to **estymator Hodgesa-Lehmanna** wielkości przesunięcia w problemie dwóch prób.

Metody rangowe

Stosując analogiczne rozumowanie jak wcześniej możemy wyznaczyć przedział ufności dla Δ.

Do zadanego α dobierzmy w_α takie, by:

$$P(w_\alpha \leq U \leq n_1 n_2 - w_\alpha) = 1 - \alpha$$

gdzie U jest statystyką Mana-Whitneya. Ponieważ U przyjmuje tylko wartości naturalne, to w_α jest liczbą naturalną.

Wtedy przedział ufności dla Δ ma postać:

$$\left[D_{(w_\alpha)}, D_{(n_1 n_2 - w_\alpha + 1)}\right]$$

gdzie $D_{(k)}$ oznacza k-tą wartość w uporządkowanym ciągu D_{ij}.

Test Wilcoxona

Dotychczasowe rozważania o testowaniu równości rozkładów cech X i Y było prowadzone przy założeniu, że F jest ciągła. Przy tym założeniu mamy

$$P(X = Y) = 0.$$

Jeżeli dopuścimy, by rozkłady X i Y nie były ciągłe wtedy w próbie $x_1, \dots, x_{n_1}, y_1, \dots, y_{n_2}$ niektóre wartości mogą się powtarzać (tzw. obserwacje związane). Wtedy należy jako rangę każdej z powtarzających się obserwacji przyjmujemy średnią arytmetyczną ich rang.

Przykład

obserwacje	-1	-1	-1	4	4	4	4	11	11
nr	1	2	3	4	5	6	7	8	9
rangi	2	2	2	5,5	5,5	5,5	5,5	8,5	8,5

Test Wilcoxona dla par

Rozważmy sytuację, w której mamy pary obserwacji $(X_1, Y_1), \ldots, (X_n, Y_n)$, przy czym pary są od siebie wzajemnie niezależne, ale zmienne w parze mogą być zależne. Załóżmy, że pary mają ten sam rozkład dwuwymiarowy.

Wtedy, by zbadać równość rozkładów X i Y badamy rozkład różnicy $Y - X$.

Jeżeli X i Y mają taki sam rozkład, to $Y - X$ ma rozkład symetryczny względem 0. Wtedy hipotezę główną można zapisać w postaci:

$$H_0: F(x) = 1 - F(-x) \text{ dla każdego } x \in \mathbb{R}.$$

Hipotezy alternatywne mogą mieć wtedy postać:

$H_1: 1 - F(-x) \leq F(x)$ dla każdego $x \in \mathbb{R}$. (lewostronna skośność)

$H_1: 1 - F(-x) \geq F(x)$ dla każdego $x \in \mathbb{R}$. (prawostronna skośność)

Test Wilcoxona dla par

Statystyka testowa Wilcoxona W^+ dla par obserwacji jest zdefiniowana jako suma rang wartości bezwzględnych różnic odpowiadających różnicom dodatnim.

Jeżeli dystrybuanta F jest ciągła i $F(x) = 1 - F(-x)$ dla każdego $x \in \mathbb{R}$, to:

1. Rozkład statystyki W^+ nie zależy od dystrybuanty F.
2. $E(W^+) = \frac{n(n+1)}{4}$, $D^2(W^+) = \frac{n(n+1)(2n+1)}{24}$
3. dla dowolnego $x \in \mathbb{R}$:

$$P\left(\frac{W^+ - EW^+}{\sqrt{D^2(W^+)}} \leq x\right) \to \Phi(x)$$

4. gdy $\mathrm{n} \to +\infty$, a Φ jest dystrybuantą rozkładu $N(0,1)$

Test Wilcoxona dla par

Jeżeli uważamy, że rozkład X jest przesunięty względem rozkładu Y o pewną stałą Δ, tzn. $F(x) = F_0(x - \Delta)$, gdzie F_0 jest dystrybuantą rozkładu symetrycznego względem 0 i $\Delta > 0$. (odpowiada to sytuacji gdy wartości w pierwszej grupie systematycznie przewyższają wartości w drugiej grupie.

To powyższe przypuszczenie możemy sprawdzić weryfikując hipotezę:

$$H_0: \Delta = 0$$

za pomocą statystyki W^+ analogicznie jak wcześniej.

Mediana wartości $B_{ij} = \frac{1}{2}(D_i + D_j)$ jest **estymatorem Hodgesa-Lehmanna** wielkości przesunięcia dla par obserwacji