

Badanie zgodności dwóch rozkładów - test serii, test mediany, test Wilcoxona, test Kruskala-Wallisa

Test serii (test Walda-Wolfowitza)

Założenie. Rozpatrywane rozkłady są ciągłe.

Mamy dwa uporządkowane ciągi danych statystycznych $x_1 \leq x_2 \leq \dots \leq x_n$ oraz $y_1 \leq y_2 \leq \dots \leq y_m$ łączymy je w jeden ciąg uporządkowany niemalejąco. W otrzymanym ciągu elementom z pierwszej sekwencji przyporządkowujemy 0 a elementom z drugiej sekwencji przyporządkowujemy 1. Obliczamy liczbę serii w otrzymanym ciągu zerojedynekowym.

Rozpatrujemy hipotezy:

H_0 (pobrane próbki pochodzą z populacji o tym samym rozkładzie),

H_1 (pobrane próbki pochodzą z populacji o różnych rozkładach),

Serie to podciągi złożone z jednakowych symboli.

Stosujemy statystykę:

$$U_{n,m} = \text{liczba serii}$$

Zbiór krytyczny:

$$K = (-\infty; k>$$

gdzie k odczytujemy z tablicy rozkładu serii dla poziomu istotności α i liczb n, m .

Decyzje:

Jeśli $u_{n,m} \in K$ to H_0 odrzucamy ,

Jeśli $u_{n,m} \notin K$ to nie ma podstaw do odrzucenia H_0 .

Uwaga.

Gdy n i m są większe od 20, to statystyka $U_{n,m}$ ma w przybliżeniu rozkład

$$N\left(\frac{2nm}{n+m}+1; \sqrt{\frac{2nm(2nm-(n+m))}{(n+m)^2(n+m-1)}}\right)$$

W naszym przypadku skorzystamy z powyższej uwagi i dla pobranej próby obliczamy wartość

$$u^* = \frac{u_{n,m} - \left(\frac{2nm}{n+m} + 1\right)}{\sqrt{\frac{2nm(2nm-(n+m))}{(n+m)^2(n+m-1)}}}$$

Statystyka ta ma rozkład $N(0, 1)$

Zbiór krytyczny:

$$K = (-\infty; -k>$$

gdzie k odczytujemy z tablicy dystrybuanty $N(0, 1)$ dla poziomu istotności α

$$\Phi(k) = 1 - \alpha .$$

TEST MEDIANY

Mamy dwa uporządkowane ciągi danych statystycznych $x_1 \leq x_2 \leq \dots \leq x_n$ oraz $y_1 \leq y_2 \leq \dots \leq y_m$ ($n \leq m$), łączymy je w jeden ciąg uporządkowany niemalejąco i wyznaczamy medianę m_e .

Rozpatrujemy hipotezy:

H_0 (pobrane próbki pochodzą z populacji o tym samym rozkładzie),

H_1 (pobrane próbki pochodzą z populacji o różnych rozkładach),

Tworzymy tablicę:

	X	Y	$n_{i\bullet}$
$>m_e$	n_{11}	n_{12}	$n_{1\bullet}$
$\leq m_e$	n_{21}	n_{22}	$n_{2\bullet}$
$n_{\bullet j}$	$n_{\bullet 1}$	$n_{\bullet 2}$	$n+m$

Na podstawie próby obliczamy wartość statystyki

$$(*) \quad u_n = \sum_{i=1}^2 \sum_{j=1}^2 \frac{(n_{ij} - \hat{n}_{ij})^2}{\hat{n}_{ij}}$$

(rozpatrywana statystyka ma rozkład Y_1)

gdzie

$$\hat{n}_{ij} = \frac{n_{i \cdot} \cdot n_{\cdot j}}{n + m} = \frac{(\text{suma } i\text{-tego wiersza}) \times (\text{suma } j\text{-tej kolumny})}{\text{liczebność próby}}$$

α - poziom istotności.

Zbiór krytyczny ma postać

$$K = \langle k; \infty) ;$$

gdzie

$$P(Y_1 \geq k) = \alpha$$

Jeśli $u_n \in K$ to H_0 odrzucamy, w przeciwnym przypadku nie ma podstaw do odrzucenia H_0 .

Uwaga.

Dla tablicy korelacyjnej

	1	2	
1	A	B	A+B
2	C	D	C+D
	A+C	B+D	n+m

Statystyka U_n ma postać:

$$U_n = \frac{(n+m)(AD-BC)^2}{(A+B)(A+C)(B+D)(C+D)}$$

i ma rozkład Y_1 .

Test Wilcoxona

Założenie. Rozpatrywane rozkłady są ciągłe.

Mamy dwa uporządkowane ciągi danych statystycznych $x_1 \leq x_2 \leq \dots \leq x_n$ oraz $y_1 \leq y_2 \leq \dots \leq y_m$ ($n \leq m$), łączymy je w jeden ciąg uporządkowany niemalejąco i nadajemy rangi.

Rozpatrujemy hipotezy:

H_0 (pobrane próbki pochodzą z populacji o tym samym rozkładzie),

H_1 (pobrane próbki pochodzą z populacji o różnych rozkładach),

Stosujemy statystykę:

U = suma rang elementów z I próbki

Zbiór krytyczny:

$$K = (-\infty; k_1] \cup [k_2; \infty)$$

gdzie k_1 , k_2 odczytujemy z tablicy rozkładu sumy rang dla poziomu istotności α i liczb n oraz m ,

Decyzje:

Jeśli $u \in K$ to H_0 odrzucamy ,

Jeśli $u \notin K$ to nie ma podstaw do odrzucenia H_0 .

Uwaga.

Gdy n i m są większe od 10, to suma rang elementów z I próbki ma w przybliżeniu rozkład

$$N\left(\frac{n(n+m+1)}{2}; \sqrt{\frac{nm(n+m+1)}{12}}\right)$$

W naszym przypadku skorzystamy z powyższej uwagi i dla pobranej próby obliczamy wartość

$$u^* = \frac{u - \frac{n(n+m+1)}{2}}{\sqrt{\frac{nm(n+m+1)}{12}}}$$

statystyki U^* , gdzie u jest wartością zmiennej losowej U (suma rang elementów z I próbki).

Zbiór krytyczny:

$$K = (-\infty; -k) \cup (k; \infty)$$

gdzie k odczytujemy z tablicy dystrybuanty rozkładu $N(0, 1)$

$$\Phi(k) = 1 - \frac{\alpha}{2}.$$

Uwaga o decyzji.

Wyznaczamy liczbę $\hat{\alpha}$ (krytyczny poziom istotności) spełniającą równość $P(X \geq |u_n^*|) = \frac{\hat{\alpha}}{2}$, gdzie $X \sim N(0, 1)$.

Podajemy decyzję:

nie ma podstaw do odrzucenia hipotezy H_0 , gdy $\hat{\alpha}$ jest istotnie większe od zera (standardowo większe od 0,05) odrzucamy hipotezę H_0 , gdy $\hat{\alpha}$ jest bliskie zeru.

Test (sumy rang) Kruskala-Wallisa

Jest to uogólnienie testu Wilcoxon.

Założenie. Rozpatrywane rozkłady są ciągłe.

Rozpatrujemy m cech w m populacjach o rozkładach wyznaczonych przez dystrybuanty $F_1, F_2, \dots, F_m$.

Mamy m ciągów danych statystycznych o liczebnościach n_i , ($n_1 + n_2 + \dots + n_m = n$), łączymy je w jeden ciąg uporządkowany niemalejąco i nadajemy rangi.

Rozpatrujemy hipotezy:

$$H_0(F_1 = F_2 = \dots = F_m),$$

$$H_1(\text{co najmniej dwie cechy mają różne rozkłady}),$$

Niech T_i - suma rang w poszczególnych próbach.

Stosujemy statystykę:

$$u = \frac{12}{n(n+1)} \sum_{i=1}^m \frac{T_i^2}{n_i} - 3(n+1)$$

dla $n_i \geq 20$ rozpatrywana statystyka ma rozkład Y_{m-1} .

Zbiór krytyczny:

$$K = <k; \infty)$$

Decyzje:

Jeśli $u \in K$ to H_0 odrzucamy ,

Jeśli $u \notin K$ to nie ma podstaw do odrzucenia H_0 .