Analiza DISCO

Jolanta Grala-Michalak

Zakład Rachunku Prawdopodobieństwa i Statystyki Matematycznej

Wydział Matematyki i Informatyki UAM

KLASYCZNA ANALIZA WARIANCJI

ANOVA-JEDNOZMIENNA ANALIZA WARIANCJI
MANOVA-WIELOZMIENNA ANALIZA WARIANCJI

Pobieramy K niezależnych prób losowych z K populacji, gdzie $K \geq 2$. Zakładamy, że pomiary obarczone są losowymi błędami, nieskorelowanymi ze sobą, podlegającymi rozkładowi normalnemu z zerową wartością oczekiwaną (zerowym wektorem wartości oczekiwanych) i stałą wariancją (macierzą wariancji-kowariancji).

Stąd wynika, że badana cecha (lub cechy) podlega (podlegają) w tych populacjach rozkładowi normalnemu z wartością oczekiwaną μ_i w i-tej populacji oraz jednakowej dla wszystkich populacji, znanej, dodatniej (dodatnio określonej) wariancji (macierzy wariancji-kowariancji).

Testujemy hipotezę zerową o równości wartości oczekiwanych (wektorów wartości oczekiwanych) μ_i w i-tej populacji

$$H_0 : \mu_1 = \mu_2 = \ldots = \mu_K$$

za pomocą analizy zmienności (mierzonej wariancją) wyników pomiarów.

Jeśli w postępowaniu testowym odrzucimy H_0, to wnioskujemy, że przynajmniej dwie spośród rozpatrywanych K populacji różnią się istotnie pod względem średniej.

METODY WERYFIKACJI HIPOTEZY ZEROWEJ

PARAMETRYCZNE

Klasyczny test F-Fishera

NIEPARAMETRYCZNE

test Kruskala-Wallisa oparty na rangach

test Mooda oparty na medianie

METODY WERYFIKACJI HIPOTEZY ZEROWEJ

PARAMETRYCZNE

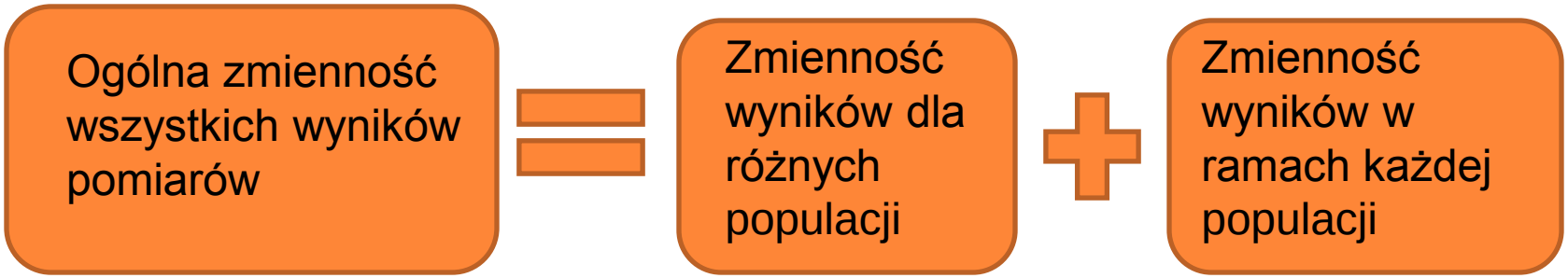
Klasyczny test F-Fishera

NIEPARAMETRYCZNE

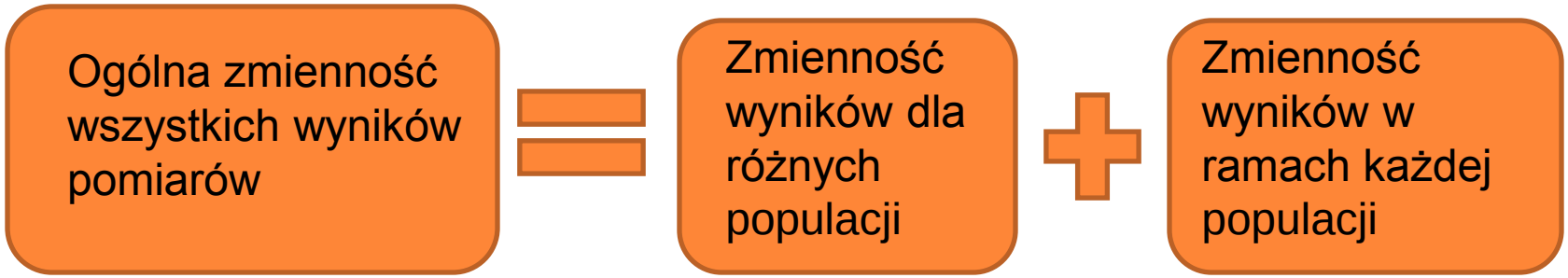
test Kruskala-Wallisa oparty na rangach

test Mooda oparty na medianie

Klasyczny test F-Fishera



Klasyczny test F-Fishera



mierzona sumą kwadratów odchyleń wyników poszczególnych obserwacji od średniej ogólnej

mierzona sumą kwadratów odchyleń średnich wyników dla populacji od średniej ogólnej

mierzona sumą kwadratów odchyleń wyników obserwacji od średniej dla populacji, z której pochodzi wynik

Klasyczny test F-Fishera

Ogólna zmienność wszystkich wyników pomiarów

=

Zmienność wyników dla różnych populacji

+

Zmienność wyników w ramach każdej populacji

mierzona sumą kwadratów odchyleń wyników poszczególnych obserwacji od średniej ogólnej

mierzona sumą kwadratów odchyleń średnich wyników dla populacji od średniej ogólnej

mierzona sumą kwadratów odchyleń wyników obserwacji od średniej dla populacji, z której pochodzi wynik

$$\sum_{i=1}^{K}\sum_{j=1}^{n_i}\left(x_{ij}-\bar{x}\right)^2 = \sum_{i=1}^{K}\sum_{j=1}^{n_i}\left(\bar{x}_i-\bar{x}\right)^2 + \sum_{i=1}^{K}\sum_{j=1}^{n_i}\left(x_{ij}-\bar{x}_i\right)^2$$

i jest numerem populacji, j jest numerem obserwacji w tej populacji

Klasyczny test F-Fishera

Ogólna zmienność wszystkich wyników pomiarów = Zmienność wyników dla różnych populacji + Zmienność wyników w ramach każdej populacji

$$\sum_{i=1}^{K}\sum_{j=1}^{n_i}\left(x_{ij}-\bar{x}\right)^2 = \sum_{i=1}^{K}\sum_{j=1}^{n_i}\left(\bar{x}_i-\bar{x}\right)^2 + \sum_{i=1}^{K}\sum_{j=1}^{n_i}\left(x_{ij}-\bar{x}_i\right)^2$$

$$SS = SST + SSE$$

$$\sum_{i=1}^{K}\sum_{j=1}^{n_i} d_E^2\left(x_{ij},\bar{x}\right) = \sum_{i=1}^{K}\sum_{j=1}^{n_i} d_E^2\left(\bar{x}_i,\bar{x}\right) + \sum_{i=1}^{K}\sum_{j=1}^{n_i} d_E^2\left(x_{ij},\bar{x}_i\right)$$

d jest odległością Euklidesa w $\Re^p$

Klasyczny test F-Fishera

Hipotezę zerową H_0 odrzucimy (wnioskując na poziomie istotności α), jeżeli suma kwadratów dla zmienności międzygrupowej SST będzie istotnie większa niż suma kwadratów dla zmienności wewnątrzgrupowej SSE. Wartością graniczną (od której rozpoczynamy odrzucenie H_0) dla ilorazu tych wielkości będzie wartość wyrażenia

$$\frac{N-K}{K-1}F(K-1,N-K,1-\alpha),$$

gdzie N jest łączną ilością obserwacji we wszystkich K populacjach, a F *jest* kwantylem rzędu (1-α) z rozkładu F o $(K-1)$ i $(N-K)$ stopniach swobody.

Klasyczny test F-Fishera

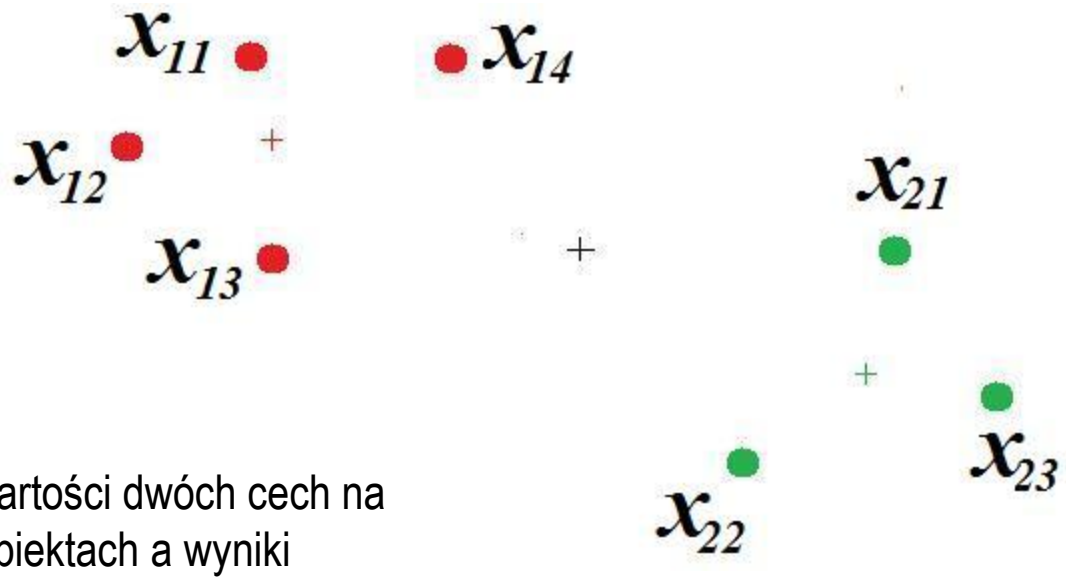
Hipotezę zerową H_0 odrzucimy (wnioskując na poziomie istotności α), jeżeli suma kwadratów dla zmienności międzygrupowej SST będzie istotnie większa niż suma kwadratów dla zmienności wewnątrzgrupowej SSE, a wartością graniczną będzie wielkość

$$\frac{N-K}{K-1}F(K-1,N-K,1-\alpha),$$

gdzie N jest łączną ilością obserwacji we wszystkich K populacjach, a *F jest* kwantylem rzędu (1-α) z rozkładu F o *(K – 1)* i (*N – K)* stopniach swobody.

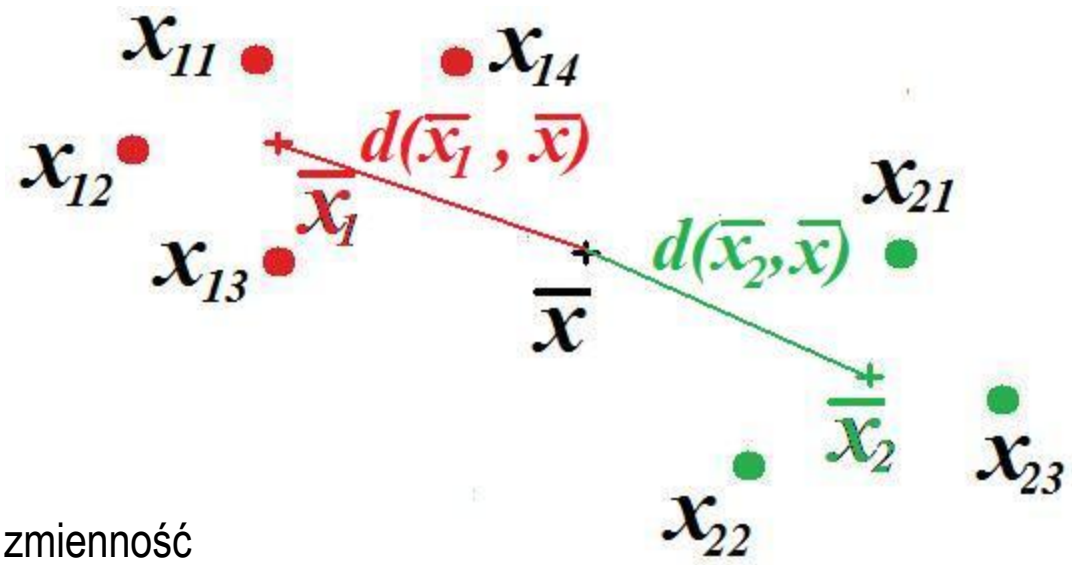
$$\frac{SST}{SSE} = \frac{\sum_{i=1}^{K}\sum_{j=1}^{n_i}\left(\bar{x}_i-\bar{x}\right)^2}{\sum_{i=1}^{K}\sum_{j=1}^{n_i}\left(x_{ij}-\bar{x}_i\right)^2} \geq \frac{N-K}{K-1}F(K-1,N-K,1-\alpha),$$

Przykład: $K = 2, N = 4 + 3$



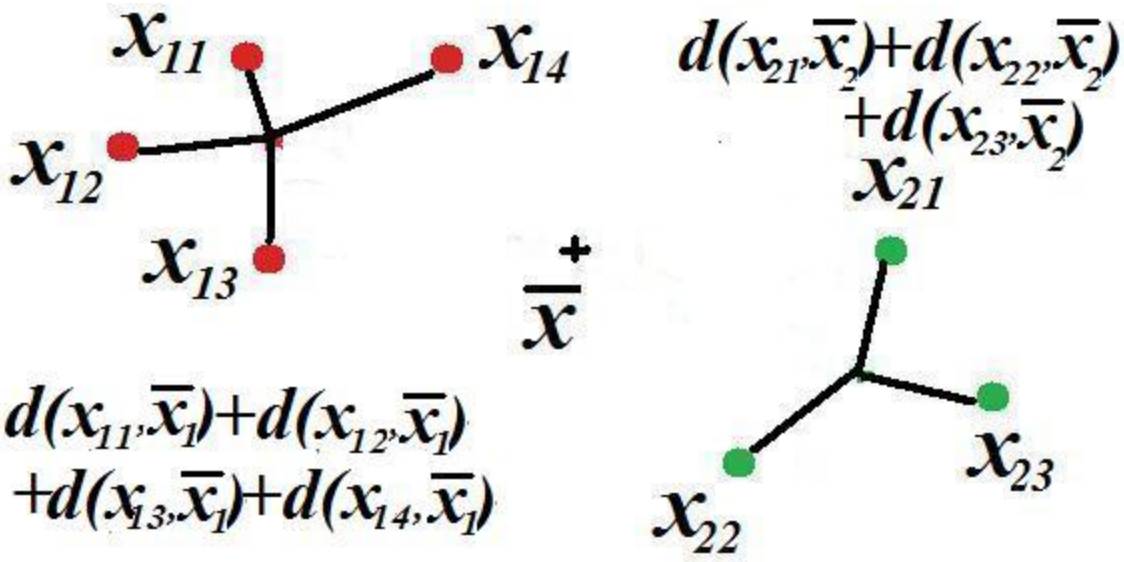
Mierzymy wartości dwóch cech na badanych obiektach a wyniki prezentujemy w postaci punktów na płaszczyźnie $\Re^2$

Przykład: $K = 2, N = 4 + 3$



Mierzymy zmienność pomiędzy grupami obserwacji otrzymanych dla różnych populacji za pomocą odległości d

Przykład: $K = 2, N = 4 + 3$



Mierzymy zmienność wewnątrz grup dla wszystkich populacji za pomocą odległości d

Przykład: $K = 2, N = 4 + 3$

Za pomocą procedury testowej oceniamy wartość ilorazu

$$\frac{\sum_{i=1}^{K}\sum_{j=1}^{n_i} d_E^2\left(\bar{x}_i, \bar{x}\right)}{\sum_{i=1}^{K}\sum_{j=1}^{n_i} d_E^2\left(x_{ij}, \bar{x}_i\right)}$$

i po porównaniu z wartością krytyczną testu decydujemy, czy odrzucić hipotezę zerową, czy też nie ma do tego podstaw.

METODA DISCO
DISTANCE COMPONENTS' METHOD

Metoda składowych odległości

Ogólna zmienność w próbie = Zmienność pomiędzy próbami + Zmienność wewnątrz prób

$$T_\alpha = S_\alpha + W_\alpha$$

Def.1

Odległością pomiędzy wynikami dla dwóch prób $A=\{a_1,\ldots,a_{n_1}\}$ oraz $B=\{b_1,\ldots,b_{n_2}\}$ nazywać będziemy wyrażenie

$$d_\alpha(A,B)=\frac{n_1 n_2}{n_1+n_2}[2g_\alpha(A,B)-g_\alpha(A,A)-g_\alpha(B,B)]$$

gdzie

$$g_\alpha(A,B)=\frac{1}{n_1 n_2}\sum_{i=1}^{n_1}\sum_{m=1}^{n_2}\left\|a_i-b_m\right\|^\alpha$$

Def.1

Odległością pomiędzy wynikami dla dwóch prób $A=\{a_1,\ldots,a_{n_1}\}$ oraz $B=\{b_1,\ldots,b_{n_2}\}$ nazywać będziemy wyrażenie

$$d_\alpha(A,B)=\frac{n_1 n_2}{n_1+n_2}[2g_\alpha(A,B)-g_\alpha(A,A)-g_\alpha(B,B)]$$

gdzie

$$g_\alpha(A,B)=\frac{1}{n_1 n_2}\sum_{i=1}^{n_1}\sum_{m=1}^{n_2}\left\|a_i-b_m\right\|^\alpha$$

- Indeks α jest liczbą z przedziału (0,2].
- Jest to szczególny przypadek średniej odległości Giniego.
- **Jeśli $\alpha = 2$, to $d_2(A,B) = 2\,SS$ w klasycznej analizie wariancji.**

Def.2

Jeśli $A_1,\ldots, A_K$ są *p*-wymiarowymi próbami, gdzie p >1, liczącymi, odpowiednio, $n_1,\ldots, n_K$ elementów oraz $N = n_1+\ldots+ n_K$, *to definiujemy zmienność z próby na próbę* S_α *w sposób następujący:*

$$S_\alpha(A_1,\ldots,A_K) = \sum_{1\le j<m<K}\left(\frac{n_j+n_m}{2N}\right) d_\alpha(A_j,A_m) =$$

$$= \sum_{1\le j<m<K}\left\{\left(\frac{n_j+n_m}{2N}\right)[2g_\alpha(A_j,A_m)-g_\alpha(A_j,A_j)-g_\alpha(A_m,A_m)]\right\}$$

gdzie, jak poprzednio, mamy

$$g_\alpha(A,B) = \frac{1}{n_1 n_2}\sum_{i=1}^{n_1}\sum_{m=1}^{n_2}\left\|a_i - b_m\right\|^\alpha .$$

Def.2

Jeśli $A_1,\ldots, A_K$ są *p*-wymiarowymi próbami, gdzie p >1, liczącymi, odpowiednio, $n_1,\ldots, n_K$ elementów oraz $N = n_1+\ldots+ n_K$, *to definiujemy zmienność z próby na próbę* S_α *w sposób następujący:*

$$S_\alpha(A_1,\ldots,A_K) = \sum_{1\le j<m<K}\left(\frac{n_j+n_m}{2N}\right) d_\alpha(A_j,A_m) =$$

$$= \sum_{1\le j<m<K}\left\{\left(\frac{n_j+n_m}{2N}\right)[2g_\alpha(A_j,A_m)-g_\alpha(A_j,A_j)-g_\alpha(A_m,A_m)]\right\}$$

gdzie, jak poprzednio, mamy

$$g_\alpha(A,B) = \frac{1}{n_1 n_2}\sum_{i=1}^{n_1}\sum_{m=1}^{n_2}\left\|a_i - b_m\right\|^\alpha .$$

∗ **Jeśli K=2, p=1, α=2, to S_2 = ½ $d_2(A_1, A_2)$ = *SS w klasycznej analizie wariancji.***

Def.3

Jeśli $A_1,\ldots, A_K$ są *p*-wymiarowymi próbami, gdzie *p* >1, liczącymi, odpowiednio, $n_1,\ldots, n_K$ elementów oraz $N = n_1+\ldots+ n_K$, *to definiujemy zmienność ogólną w łącznej próbie* T_α *w sposób następujący:*

$$T_\alpha(A_1,\ldots, A_K) = \frac{N}{2}\, g_\alpha(A, A)$$

przy czym

$$A = \bigcup_{i=1}^{K} A_i .$$

Def.4

Jeśli $A_1,\ldots, A_K$ są *p*-wymiarowymi próbami, gdzie *p* >1, liczącymi, odpowiednio, $n_1,\ldots, n_K$ elementów oraz $N = n_1+\ldots+ n_K$, *to definiujemy zmienność wewnątrz łącznej próby* W_α *w sposób następujący:*

$$W_\alpha(A_1,\ldots, A_K) = \sum_{i=1}^{K} \frac{n_i}{2}\, g_\alpha(A_j, A_j).$$

Twierdzenie o rozkładzie odległości na składowe

(**Maria L. Rizzo, Gábor J. Székely, 2010**)

Dla wszystkich całkowitych $K \geq 2$ ogólna zmienność T_α dla K prób z K populacji rozkłada się na dwie składowe według wzoru

$$T_\alpha(A_1,\ldots,A_K) = S_\alpha(A_1,\ldots,A_K) + W_\alpha(A_1,\ldots,A_K)$$

gdzie $S_\alpha \geq 0$ *i* $W_\alpha \geq 0$.

Twierdzenie o rozkładzie odległości na składowe

(**Maria L. Rizzo, Gábor J. Székely, 2010**)

Dla wszystkich całkowitych $K \geq 2$ ogólna zmienność T_α dla K prób z K populacji rozkłada się na dwie składowe według wzoru

$$T_\alpha(A_1,\ldots,A_K) = S_\alpha(A_1,\ldots,A_K) + W_\alpha(A_1,\ldots,A_K)$$

gdzie $S_\alpha \geq 0$ *i* $W_\alpha \geq 0$.

* **Dla p = 1 oraz indeksu α = 2 metoda DISCO jest uogólnieniem rozkładu zmienności ogólnej na składowe w klasycznej analizie wariancji.**

Estymacja

Zakładając, że X i X' są niezależnymi zmiennymi losowymi o jednakowym rozkładzie:

* średnia Giniego $\mathrm{g}_\alpha(\mathrm{A}, \mathrm{A})$ jest obciążonym estymatorem parametru ξ,
* wyrażenie $(n/(n\text{-}1))\cdot \mathrm{g}_\alpha(\mathrm{A}, \mathrm{A})$ jest nieobciążonym estymatorem parametru ξ,
* wartość oczekiwana S_α wynosi $\xi\cdot(K\text{-}1)/2$,
* wartość oczekiwana W_α wynosi $\xi\cdot(N\text{-}K)/2$,

$$\xi = \mathrm{E}\left\| X - X' \right\|^{\alpha}.$$

Estymacja

Zakładając, że X i X' są niezależnymi zmiennymi losowymi o jednakowym rozkładzie:

* średnia Giniego $g_\alpha(A, A)$ jest obciążonym estymatorem parametru ξ,
* wyrażenie $n/(n\text{-}1)\, g_\alpha(A, A)$ jest nieobciążonym estymatorem parametru ξ,
* wartość oczekiwana S_α wynosi $\xi(K\text{-}1)/2$,
* wartość oczekiwana W_α wynosi $\xi(N\text{-}K)/2$,

$$\xi = \mathrm{E}\left\| X - X' \right\|^{\alpha}.$$

Weryfikacja hipotezy zerowej

Odrzucamy hipotezę zerową H_0 dla dużych wartości nieujemnej statystyki

$$D_{n,\alpha} = \frac{S_\alpha}{W_\alpha} \cdot \frac{N-K}{K-1}.$$

Graniczny rozkład statystyki testowej

Dla wszystkich $0 < \alpha < 2$, przy prawdziwości H_0:

* $S_\alpha/(K\text{-}1)$ dąży według rozkładu do formy kwadratowej scentrowanych rozkładów gaussowskich,
* $W_\alpha/(N\text{-}K)$ dąży według prawdopodobieństwa do stałej,
* $D_{n,\alpha}$ dąży według rozkładu do formy kwadratowej Q zbudowanej na niezależnych standardowych rozkładach normalnych.

Székely i Bakirov (2003) pokazali, że $\mathrm{E}(Q) = 1$ oraz

$$P(Q \geq (\Phi^{-1}(1-\frac{\alpha_0}{2})) \leq \alpha_0$$

dla $\alpha_0 \leq 0{,}215$.

Empiryczny poziom istotności (wartość p) testu uzyskuje się na podstawie procedury permutacyjnej (Davidson, Hinkley, 1997):

(1) Oblicz wartość statystyki testowej $F_\alpha(A)$ na podstawie wyników próby.
(2) W każdym r-tym powtórzeniu punktu (2) (r = 1,…,R) wygeneruj losową permutację danych i oblicz wartość $F_\alpha^r(A)$.
(3) Zlicz, ile razy wartości statystyki $F_\alpha^r(A)$ dla danych po permutacji były większe niż $F_\alpha(A)$. Tę wielkość oznacz jako κ.
(4) Wyznacz empiryczny poziom istotności (wartość p) jako

$$\hat{p} = \frac{1+\kappa}{1+R}.$$

Dostateczną dokładność można uzyskać dla przynajmniej 99 permutacji.

Zalety testu DISCO

1. Dla indeksu $0 < \alpha < 2$ test DISCO jest odporny wobec wszystkich rozkładów alternatywnych ze skończonym drugim momentem.
2. Dla $\alpha = 2$ tak nie jest, test nie zawsze wykrywa różnice w skali lub innych charakterystykach.
3. Test DISCO można stosować również dla rozkładów, w których nie istnieje pierwszy moment, ale istnieje moment rzędu ε $(0 < \varepsilon < 1)$. Wtedy wybieramy indeks $\alpha = \varepsilon/2$.
4. Liczebność próby nie musi być większa niż wymiar obserwacji.
5. Nie wymaga się jednorodności wariancji błędów w populacjach.
6. Metoda nie wymaga określenia rozkładu badanych cech.

DISCO w R

DISCO analysis of multivariate iris data:

Distance Components : index 1.00

Source	Df	Sum Dist	Mean Dist	F-ratio	p-value
Between:					
Species	2	119.23731	59.61865	124.597	0.001
Within	147	70.33848	0.47849		

MANOVA analysis of multivariate iris data:

	Df	Pillai	approx F	num Df	den Df	Pr(>F)	
Species	2	1.192	53.466	8	290	<2.2e-16	***
Residuals	147						

[permutation test p =0.001]

DISCO analysis of residuals of linear model for iris data:

Distance Components : index 1.00

Source	Df	Sum Dist	Mean Dist	F-ratio	p-value
Between:					
Species	2	1.69845	0.84923	0.775	0.039
Within	147	70.33848	0.47849		