AKADEMIA GÓRNICZO-HUTNICZA
IM. STANISŁAWA STASZICA W KRAKOWIE

Analiza wariancji ANOVA

Statystyka

Dr inż. Janusz Majewski
Katedra Informatyki

Literatura

- Prezentacja wykorzystuje fragmenty książki: Amir D. Aczel „Statystyka w zarządzaniu”, PWN, 2007

ANOVA – analiza wariancji

„Wszystkim ludziom żyjącym dziś, dnia 17 czerwca 1579, wiadomym czynimy, że z Bożej Łaski i w imieniu Jej Królewskiej Mości Królowej Anglii oraz Jej spadkobierców obejmuję w posiadanie na zawsze to królestwo, którego król oraz lud dobrowolnie zrzekają się swoich praw i tytułów do ziemi na rzecz Jej Królewskiej Mości, które to królestwo zostało przeze mnie nazwane i ma być przez wszystkich zwane pod nazwą **Nova Albion**”.

Francis Drake

W lecie 1936 roku na wzgórzach otaczających Point San Quentin i zatokę San Francisco natrafiono na mosiężną płytę z przytoczonym wyżej napisem…

W dzienniku okrętowym, który Sir Francis Drake prowadził w czasie swoich podróży dookoła świata, pisze on o dopłynięciu w roku 1579 do bezpiecznego miejsca na lądzie w celu poddania statku remontowi, którym to lądem było wybrzeże północnej Kalifornii. Wspomina też o pozostawieniu na nabrzeżu płyty upamiętniającej to wydarzenie…

Klasyfikacja pojedyncza – porównanie kilku średnich

Metoda analizy wariancji, w ogólnym przypadku, pozwala na sprawdzanie czy pewne czynniki wywierają wpływ na kształtowanie się średnich wartości badanych cech mierzalnych (o charakterze ilościowym). Jeśli uwzględnimy jeden czynnik – zadanie sprowadzi się do porównania kilku średnich.

Mamy k grup obserwacji o charakterze ilościowym. W każdej i-tej grupie dysponujemy próbką zawierającą n_i obserwacji. Zakładamy, że obserwacje w każdej grupie maja rozkład normalny lub zbliżony do normalnego, zaś wariancje we wszystkich grupach są równe i wynoszą σ^2 (σ^2 nie jest znane).

Klasyfikacja pojedyncza – porównanie kilku średnich

Grupa	1	2	…	i	…	k	Wszystkie
Liczba obserwacji	n_1	n_2	…	n_i	…	n_k	$N = \sum_{i=1}^{k} n_i$
Wartości obserwacji	y_{11} y_{12} ⋮ y_{1n_1}	y_{21} y_{22} ⋮ y_{2n_2}	…	y_{i1} y_{i2} ⋮ y_{in_k}	…	y_{k1} y_{k2} ⋮ y_{kn_k}	
Suma wartości	T_1	T_2	…	$T_i = \sum_{j=1}^{n_i} y_{ij}$	…	T_k	$T = \sum_{i=1}^{k} T_i$
Średnia wartości	$\overline{y_1}$	$\overline{y_2}$	…	$\overline{y_i} = T_i/n_i$	…	$\overline{y_k}$	$\overline{y} = T/N$
Suma kwadratów y^2	S_1	S_2	…	$S_i = \sum_{j=1}^{n_i} y_{ij}^2$	…	S_k	$S = \sum_{i=1}^{k} S_i$

Klasyfikacja pojedyncza – porównanie kilku średnich

y_{ij} – j-ta obserwacja w i-tej grupie

MODEL ADDYTYWNY: $y_{ij} = \mu_i + \varepsilon_{ij}$

μ_i – prawdziwa średnia w i-tej grupie

ε_{ij} – składnik losowy z zerową wartością średnią i stałą wariancją σ^2

Hipotezy w analizie wariancji w klasyfikacji pojedynczej

H_o: $\mu_1 = \mu_2 = \ldots = \mu_i = \ldots = \mu_k$

H_1: nie wszystkie średnie grupowe są równe

Klasyfikacja pojedyncza – porównanie kilku średnich

MODEL ADDYTYWNY: $y_{ij} = \mu + \alpha_i + \varepsilon_{ij}$

μ – wartość niezależna od grupy

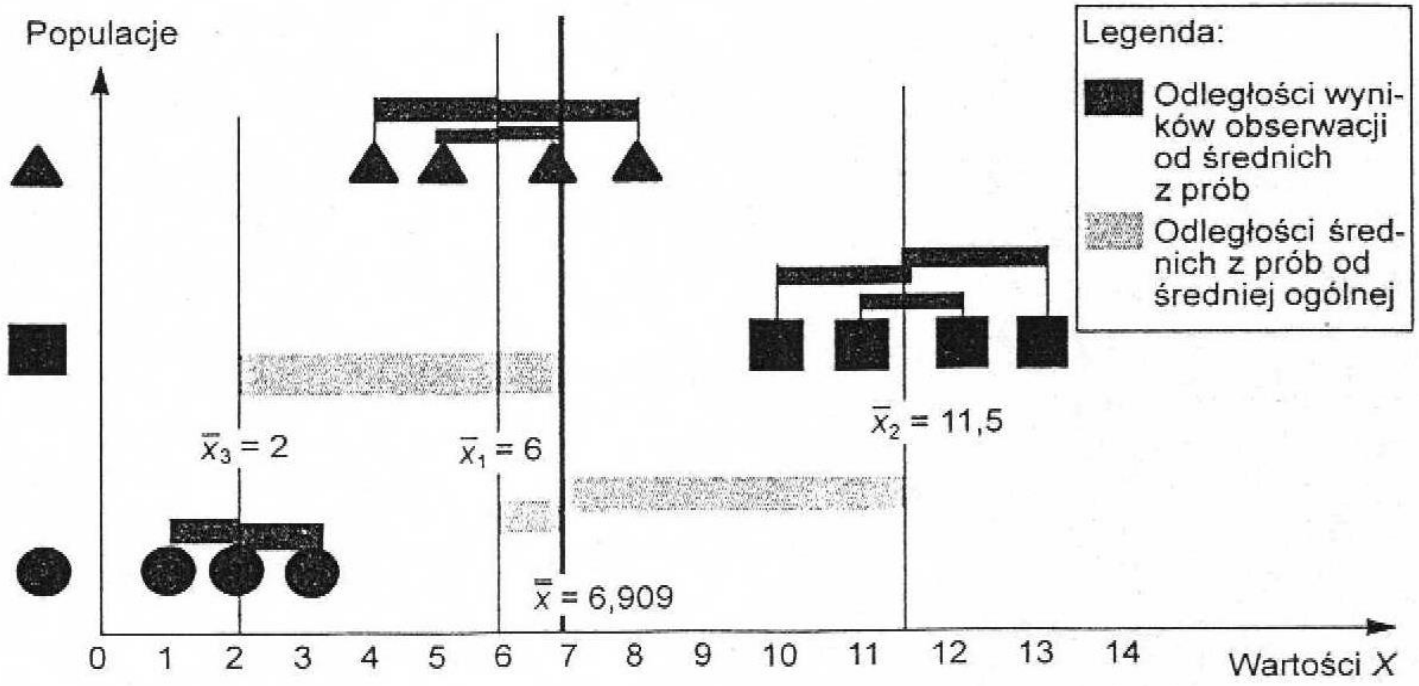
α_i – wartość „charakterystyczna” dla danej grupy, odpowiedzialna za różnice systematyczne pomiędzy grupami.

ε_{ij} – składnik losowy z zerową wartością średnią i stałą wariancją σ^2

Podział μ_i na $\mu + \alpha_i$ jest tak dokonany, aby

$$\sum_{i=1}^{k} \alpha_i = 0$$

Istota ANOVA – przykład



Rysunek 9.5. Odchylenia wyników obserwacji trójkątów, kwadratów i kółek od średnich z prób oraz odchylenia średnich z prób od średniej ogólnej

Podstawowa zasada ANOVA głosi, że **gdy średnie w populacjach nie są sobie równe, to „przeciętne" odchylenie losowe (błąd) jest stosunkowo mały w porównaniu z „przeciętnym" odchyleniem zabiegowym.**

Klasyfikacja pojedyncza – porównanie kilku średnich

$$\sum_{ij}(y_{ij} - \bar{y})^2 = \sum_{ij}(y_{ij} - \bar{y})^2 + \sum_{ij}(y_{ij} - \bar{y})^2$$

SK = SKWG + SKMG

Całkowita suma kwadratów

$$SK = \sum_{i_{ij}}(y_{ij} - \bar{y})^2 = S - \frac{T^2}{N}$$

Suma kwadratów wewnątrzgrupowa

$$SKWG = \sum_{i_{ij}}(y_{ij} - \bar{y}_i)^2 = S - \sum_{i}\frac{T_i^2}{n_i}$$

Suma kwadratów międzygrupowa

$$SKMG = \sum_{i_{ij}}(y_{ij} - \bar{y})^2 = \sum_{i}\frac{T_i^2}{n_i} - \frac{T^2}{N}$$

Klasyfikacja pojedyncza – porównanie kilku średnich

Jeśli hipoteza zerowa jest prawdziwa (czyli wszystkie średnie grupowe są równe), to istnieją trzy nieobciążone estymatory tej samej wartości σ^2 (σ^2 - wariancja identyczna we wszystkich grupach)

$$S_T^2 = \frac{SK}{N-1} \qquad S_W^2 = \frac{SKWG}{N-k} \qquad S_M^2 = \frac{SKMG}{k-1}$$

Jeśli hipoteza zerowa nie jest prawdziwa, to S_W^2 jest nadal nieobciążonym estymatorem σ^2, zaś S_M^2 wzrasta. Gdy S_M^2 znacznie przewyższa S_W^2, H_o trzeba odrzucić.

$$F = \frac{S_M^2}{S_w^2}$$

H_o trzeba odrzucić, gdy $F \geq F_{(N-k)}^{(k-1)}$

Klasyfikacja pojedyncza – porównanie kilku średnich

Źródło zmienności	Suma kwadratów	Liczba stopni swobody	Średni kwadrat	Stosunek wariancji
Miedzy grupami	*SKMG*	*k-1*	$S_M^2 = \frac{SKMG}{k-1}$	$F = \frac{S_M^2}{S_W^2}$
Wewnątrz grup	*SKWG*	*N-k*	$S_W^2 = \frac{SKWG}{N-k}$	
Całkowita	*SK*	*N-1*		

Klasyfikacja pojedyncza – porównanie kilku średnich

Gdy odrzucimy H_o można testować istotność różnicy miedzy dwiema wybranymi średnimi $\overline{y_a}$: $\overline{y_b}$ stosując test *t*:

$$t = \frac{\overline{y_a} - \overline{y_b}}{S_w \sqrt{\frac{1}{n_a} + \frac{1}{n_b}}}$$ (różnica jest istotna, gdy $|t| \geq t_{(N-k)}$)

lub dla przypadku równolicznych grup ($n_1 = n_2 = \ldots = n_k = n$) obliczyć najmniejsza istotna różnicę miedzy średnimi

$$D = \propto t_{(N-k)}\, s_w \sqrt{\frac{2}{n}}$$

Klasyfikacja pojedyncza – porównanie kilku średnich

Przykład: Czas krzepnięcia osocza krwi mierzono 4 metodami. Osocze pobrano od dziesięciu pacjentów i poddano czterem testom

Metoda	1	2	3	4
Ocena w minutach	9,1	10,0	10,0	10,9
	8,9	10,2	9,9	11,1
	8,4	9,8	9,8	12,2
	12,8	11,6	12,9	14,4
	8,7	9,5	11,2	9,8
	9,2	9,2	9,9	12,0
	7,6	8,6	8,5	8,5
	8,6	10,3	9,8	10,9
	8,9	9,4	9,2	10,4
	7,9	8,5	8,2	10,0
T_i	90,1	97,1	99,4	110,2
T_i^2	8118,01	9428,41	9880,36	12144,04
$\overline{y_i}$	9,01	9,71	9,94	11,02

Klasyfikacja pojedyncza – porównanie kilku średnich

N=40; T=396,0; S=4021,84

$$\frac{T^2}{N} = 3936{,}256$$

$$\sum_i \frac{T_i^2}{n_i} = 3957{,}082$$

Źródło zmienności	Suma kwadratów	Liczba stopni swobody	Średni kwadrat	Stosunek wariancji	Istotność
Miedzy grupami	20,826	3	6,924	3,85	TAK dla *α=0,025*
Wewnątrz grup	64,758	36	1,799		
Całkowita	85,584	39			

$${}_{0{,}025}F_{(36)}^{(3)} = 3{,}51$$

Klasyfikacja pojedyncza – porównanie kilku średnich

n=10

$$_{0,05}\, t_{(36)} = 2{,}029$$

$$D = 2{,}029 \cdot \sqrt{1{,}799} \cdot \sqrt{\frac{2}{10}} = 1{,}22$$

Dla *α=0,05* istotne są różnice miedzy

metodą 1 a metodą 4 oraz

metodą 2 a metodą 4 (ledwie, ledwie)

Dla uzyskania odpowiedzi na pytanie: czy wynik metody 4 istotnie odbiega od średniego wyniku metod 1, 2 i 3 należy zastosować badanie istotności kontrastu liniowego.

Klasyfikacja pojedyncza – porównanie kilku średnich

Dla uzyskania odpowiedzi na pytanie: czy wynik metody 4 istotnie odbiega od średniego wyniku metod 1, 2 i 3 należy zastosować badanie istotności kontrastu liniowego.

Kontrast liniowy określamy jako:

$$L = \sum \lambda_i \bar{y}_i$$

gdzie: $\sum \lambda_i = 0$ i testujemy używając zwykłego testu *t* przy *k(n-1)* stopniach swobody

$$t = \frac{L}{s_W \sqrt{\frac{\sum \lambda_i^2}{n}}}$$

Klasyfikacja pojedyncza – porównanie kilku średnich

Możemy też wykorzystać metodę analizy wariancji w schemacie:

Źródło zmienności	Suma kwadratów	Liczba stopni swobody	Średni kwadrat	Stosunek wariancji
Względem kontrastu *L*	$\frac{L^2}{\frac{1}{n}\sum \lambda_i^2}$	*1*	$S_1^2 = \frac{L^2}{\frac{1}{n}\sum \lambda_i^2}$	$F_1 = \frac{S_1^2}{S_W^2}$
Względem innych kontrastów	$SKMG - \frac{L^2}{\frac{1}{n}\sum \lambda_i^2}$	*k – 2*	$S_A^2 = \frac{SKMG - \frac{L^2}{\frac{1}{n}\sum \lambda_i^2}}{k-2}$	$F_2 = \frac{S_A^2}{S_W^2}$
Wewnątrz grup	*SKWG*	*k(n – 1)*	$S_W^2 = \frac{SKWG}{k(n-1)}$	
Całkowita	*SK*	*nk – 1*		

Klasyfikacja pojedyncza – porównanie kilku średnich

Przykład c.d.: Chcemy sprawdzić, czy średnia dla metody 4 odbiega istotnie od średnich dla pozostałych metod badania czasu krzepnięcia osocza krwi.

Konstruujemy kontrast liniowy:

$$L = 3\bar{y}_4 - \bar{y}_1 - \bar{y}_2 - \bar{y}_3$$

Mamy:

L = 4,40

Źródło zmienności	Suma kwadratów	Liczba stopni swobody	Średni kwadrat	Stosunek wariancji
Względem kontrastu *L*	3,667	1	3,667	2,04
Względem innych kontrastów	17,159	2	8,579	4,77
Wewnątrz grup	64,758	36	1,799	
Całkowita	85,584	39		

$_{0,05}F^{(1)}_{(36)} = 4,13$

$_{0,05}F^{(2)}_{(36)} = 3,29$

Kontrast nie jest istotny, inne kontrasty są istotne. Jakie? Dlaczego?

Porównanie kilku wariancji (test Bartletta)

Przyjmujemy oznaczanie takie same , jak przy porównywaniu średnich z *k* grup. Testujemy hipotezę zerową mówiąca że wariancje w każdej z grup są identyczne.

$$\boldsymbol{H_0}\colon\ \boldsymbol{\sigma_1^2 = \ldots = \sigma_i^2 = \ldots = \sigma_k^2}$$

$$\boldsymbol{H_1}\colon\ \textbf{nie wszystkie wariancje są identyczne}$$

Zakłada się, że próba została pobrana z populacji o rozkładzie normalnym.

Porównanie kilku wariancji (test Bartletta)

Zakłada się, że próba została pobrana z populacji o rozkładzie normalnym.

Obliczamy:

$$s_i^2 = \frac{\sum_{j=1}^{n_i} y_{ij}^2 - \frac{(\sum_{j=1}^{n_i} y_{ij})^2}{n_i}}{n_i - 1}$$

$$C = 1 + \frac{1}{3(k-1)}\left[\sum_{i=1}^{k}\frac{1}{n_i - 1} - \frac{1}{N-k}\right]$$

$$\widetilde{s^2} = \frac{\sum_{i=1}^{k}(n_i - 1)s_i^2}{N-k}$$

$$\chi^2 = \frac{M}{C}$$

$$M = (N-k)\ln\widetilde{s^2} - \sum_{i=1}^{k}(n_i - 1)\ln s_i^2$$

Jeśli $\chi^2 \geq {}_{\alpha}\chi^2_{(k-1)}$ to hipotezę o równości wszystkich wariancji odrzucamy.

Porównanie kilku wariancji (test Bartletta)

Przykład (c.d.) Badamy równość wariancji w grupach odpowiadających poszczególnym metodom oznaczania części krzepnięcia osocza. Mamy:

$$\widetilde{s^2} = 1{,}799$$

$$M = 2{,}9007$$

$$C = 1{,}0462$$

$$\chi^2 = 2{,}7726$$

$$_{0{,}05}\chi^2_3 = 7{,}815$$

$$\chi^2 < \chi^2_{kryt}$$

Wiec nie ma podstaw do odrzucenia hipotezy o równości wariancji w grupach i można było zastosować metodę analizy wariancji.

Analiza wariancji w klasyfikacji podwójnej

Brak efektu kolumn			
Brak efektu wierszy	5	5	5
	5	5	5
	5	5	5

Efekt kolumn			
Brak efektu wierszy	4	5	6
	4	5	6
	4	5	6

Brak efektu kolumn			
Efekt wierszy	4	4	4
	5	5	5
	6	6	6

Efekt wierszy	Efekt kolumn		
	4	5	6
	5	6	7
	6	7	8

Efekt wierszy	Efekt kolumn		
	4	5	6
	5	**16**	7
	6	7	8

Efekt interakcji

Analiza wariancji w klasyfikacji podwójnej

Wiersze\Kolumny	1	2	…	j	…	c	Suma
1							R_1
2							R_2
⋮			y_{ij1}	y_{ije},	…,		⋮
i			…,	y_{ijp},	y_{ijn}		R_i
⋮				T_{ij}, $S_{i,j}$,			⋮
r							R_r
Suma	C_1	C_2	…	C_j	…	C_c	T

$$T_{ij} = \sum_{p=1}^{w} y_{ijp} \qquad C_j = \sum_{i=1}^{r} T_{ij} \qquad R_i = \sum_{j=1}^{c} T_{ij}$$

$$T = \sum_{i=1}^{r} R_i = \sum_{j=1}^{c} C_j = \sum_{i,j} T_{ij} = \sum_{i,j,p} y_{ijp} \qquad S_{ij} = \sum_{p=1}^{n} y^{2}_{ijp}$$

$$N = r \cdot c \cdot n \qquad S = \sum_{i,j,p} y_{ijp}^2 \qquad \bar{y} = \frac{T}{N}$$

Analiza wariancji w klasyfikacji podwójnej

Przyjęty model addytywny:

$$y_{ijp} = \mu + \alpha_i + \beta_i + \gamma_{ij} + \varepsilon_{ijp}$$

y_{ijp} – p-ta obserwacja w i-tym wierszu i j-tej kolumnie
μ – wartość stała
α_i – odpowiedzialna za różnice pomiędzy wierszami
β_i – odpowiedzialna za równice pomiędzy kolumnami
γ_{ij} – odpowiedzialna za interakcję
ε_{ijp} – składnik losowy z zerową wartością średnią *0* i stałą wariancją σ^2

H_o : (1) brak efektu wierszy
(2) brak efektu kolumn
(3) brak efektu interakcji

Analiza wariancji w klasyfikacji podwójnej

$$SK = SKMW + SKMK + SKI + SKR$$

SK – Całkowita suma kwadratów odchyleń od średniej ogólnej

SKMW – Suma kwadratów odchyleń średnich wierszy od średniej ogólnej

SKMK – Suma kwadratów odchyleń średnich kolumn od średniej ogólnej

SKI – Suma kwadratów odchyleń średnich z kratek od wartości oczekiwanej wyjaśnionej efektami wierszy i kolumn

SKR – Resztowa suma kwadratów (odchyleń obserwacji wewnątrz kratki tabeli od średniej dla danej kratki)

Analiza wariancji w klasyfikacji podwójnej

$$SK = S - \frac{T^2}{N}$$

$$SKMW = \frac{\sum_i R_i^2}{nc} - \frac{T^2}{N}$$

$$SKMK = \frac{\sum_j C_j^2}{nr} - \frac{T^2}{N}$$

$$SKI = \frac{\sum_{ij} T_{ij}^2}{n} - \frac{T^2}{N} - (SKMW + SKMK)$$

$$SKR = SK - (SKMW + SKMK + SKI)$$

Źródło zmienności	Suma kwadratów	Liczba stopni swobody	Średni kwadrat	Stosunek wariancji
Miedzy wierszami	*SKMW*	*r-1*	$S_R^2 = \frac{SKMW}{r-1}$	$F_R = \frac{S_R^2}{S_0^2}$
Miedzy kolumnami	*SKMK*	*c-1*	$S_C^2 = \frac{SKMK}{c-1}$	$F_C = \frac{S_C^2}{S_0^2}$
Interakcja	*SKI*	*(r-1)(c-1)*	$S_I^2 = \frac{SKI}{(r-1)(c-1)}$	$F_I = \frac{S_I^2}{S_0^2}$
Reszta	*SKR*	*N-rc*	$S_0^2 = \frac{SKR}{N-rc}$	
Całkowita	*SK*	*N-1*		

Analiza wariancji w klasyfikacji podwójnej

Przykład: Fragment danych z poprzedniego przykładu. Rozważmy 3 metody i trzech pacjentów, zakładamy, że osocze pobrane od każdego z pacjentów badano każdą metoda trzykrotnie.

r=3 c=3 n=3 N=27

Analiza wariancji w klasyfikacji podwójnej

Pacjenci\Metody	*2*	*3*	*4*	*Sumy*
8	$T_{11} = 30{,}9$ *10,2* *10,5* *10,2* $S_{11} = 318{,}33$	*29,4* *9,9* *9,5* *10,0* *288,26*	*32,7* *11,3* *10,7* *10,7* *356,67*	$R_1 = 93{,}0$
9	*28,2* *9,6* *9,0* *9,6* *265,32*	*27,6* *9,1* *9,1* *9,4* *253,98*	*31,2* *10,3* *10,7* *10,2* *324,62*	$R_2 = 87{,}0$
10	*25,2* *9,0* *8,1* *8,4* *217,17*	*24,6* *8,6* *8,0* *8,4* *217,17*	*30,0* *9,8* *10,1* *10,1* *300,06*	$R_3 = 80{,}1$
Sumy	$C_1 = 93{,}0$	$C_2 = 81{,}6$	$C_3 = 93{,}9$	*T=260,1*

Analiza wariancji w klasyfikacji podwójnej

Zmienność	Suma kwadratów	Liczba stopni swobody	Średni kwadrat	Stosunek wariancji	Istotność
Miedzy pajentami	*9,26*	*2*	*4,63*	*52,08*	TAK α=0,001
Miedzy metodami	*9,14*	*2*	*4,57*	*51,41*	TAK α=0,001
Interakcja	*0,74*	*4*	*0,185*	*2,08*	NIE
Reszta	*1,60*	*18*	*0,0889*		
Całkowita	*20,74*	*26*			

Analiza wariancji w klasyfikacji podwójnej

Niektórzy autorzy [Blalock] polecają (ja też!) rozpocząć testowanie od ilorazu wariancji F_I. Przy braku podstaw do odrzucenia hipotezy o nieistotności interakcji zalecają oni sumę kwadratów interakcji dodać do składnika resztowego zmieniając odpowiednio liczbę stopni swobody

$$SKR' = SKI + SKR \qquad (= 0{,}74 + 1{,}60 = 2{,}34)$$

$$(S_o')^2 = \frac{SKR'}{N-r-c+1} \qquad (= 2{,}34/(18+4) = 0{,}1064)$$

i użyć tak zmodyfikowanego średniego kwadratu resztowego jako mianownika stosunków wariancji dla efektów głównych.

Np. Między pacjentami $F_R = \frac{S_R^2}{(S_0')^2} = \frac{4{,}63}{0{,}1064} = 43{,}53$ zamiast *52,08*

(co i tak daje w naszym przypadku bardzo wysoką istotność…)

Analiza wariancji w klasyfikacji podwójnej

Gdyby zaś interakcja była istotna, można obliczyć wartość:

$$d_{ij} = \overline{y_{ij}} - \overline{y_i} - \overline{y_j} + \overline{y},$$

która stanowi odchylenie średniej w polu tabeli od wartości spodziewanej dla braku interakcji i ten sposób zorientować się, gdzie jest „źródło” interakcji.