Statystyka i modelowanie w naukach o środowisku

Wykład 11

Wieloczynnikowa analiza wariancji ANOVA

Analiza skupień

Analiza składowych głównych – PCA (ang. PCA – principal component analysis)

Wieloczynnikowa analiza wariancji

- Układ losowanych bloków

- Dwuczynnikowy układ całkowicie losowy

- Układ z powtarzanymi pomiarami

Wieloczynnikowa analiza wariancji - układ całkowicie losowy

Układ całkowicie losowy stosuje się w przypadku, kiedy jednostki doświadczalne są jednorodne, czyli nie podlegają zmienności systematycznej warunków badanego zjawiska.

W układzie całkowicie losowym wszystkie kombinacje czynników A i B rozlosowujemy na powierzchni doświadczalnej w n powtórzeniach.

W dwuczynnikowej analizie wariancji postępowanie jest zbliżone do jednoczynnikowej analizy wariancji, jednak oceniamy jednocześnie wpływ dwóch czynników doświadczalnych.

Jedną z ważnych różnic w dwuczynnikowej analizie wariancji w stosunku do jednoczynnikowej analizy wariancji jest ocena współdziałania dwóch czynników. Jeśli występuję współdziałanie oznacza to, że jeden czynnik modyfikuje wpływ drugiego na badaną zmienną.

Wieloczynnikowa analiza wariancji - układ całkowicie losowy

Przykłady zastosowań 2-czynnikowej analizy wariancji:

- Porównanie plonowania kilku odmian pszenicy przy trzech różnych dawkach nawożenia azotem (zmienna zależna: plon ziarna, czynnik A: odmiana, czynnik B: dawka nawożenia)

- Porównanie wpływu kilku różnych temperatur przechowywania i kilku różnych wilgotności powietrza na trwałość bulw ziemniaka (zmienna zależna: czas przechowywania, czynnik A: temperatura przechowywania, czynnik B: wilgotność przechowywania)

Wieloczynnikowa analiza wariancji - układ całkowicie losowy

Cel: ocena wpływu dwóch czynników doświadczalnych na cechę (zmienną) o rozkładzie normalnym.

Założenia:
zmienne mają rozkład normalny $X_i \sim N(\mu,\sigma^2)$
wariancje (a tym samym odchylenia standardowe) dla badanych populacji są równe $\sigma_1 = \sigma_2 = \ldots = \sigma_i$

(powinny być one takie same dla poziomów czynnika A i poziomów czynnika B)

Wieloczynnikowa analiza wariancji - układ całkowicie losowy

Hipotezy zerowe:

- H_{01}: średnie dla poziomów pierwszego czynnika (A) nie różnią się istotnie
- H_{02}: średnie dla poziomów drugiego czynnika (B) nie różnią się istotnie
- H_{03}: nie występuje istotne współdziałanie czynników (A i B)

każda hipoteza zerowa ma odpowiadająca jej hipotezę alternatywną

Hipotezy alternatywne:

- H_{11}: co najmniej 2 średnie dla poziomów pierwszego czynnika (A) różnią się istotnie
- H_{12}: co najmniej 2 średnie dla poziomów drugiego czynnika (B) różnią się istotnie
- H_{13}: występuje istotne współdziałanie czynników (A i B)

Wieloczynnikowa analiza wariancji - układ całkowicie losowy

Jeśli odrzucimy co najmniej jedną hipotezę zerową, a więc przyjmiemy co najmniej jedną hipotezę alternatywną to kontynuujemy analizę wykonując porównania wielokrotne średnich.

Porównania wielokrotne, mają na celu porównanie:

Średnich dla poziomów czynnika A (łącznie dla wszystkich poziomów czynnika B)

Średnich dla poziomów czynnika B (łącznie dla wszystkich poziomów czynnika A)

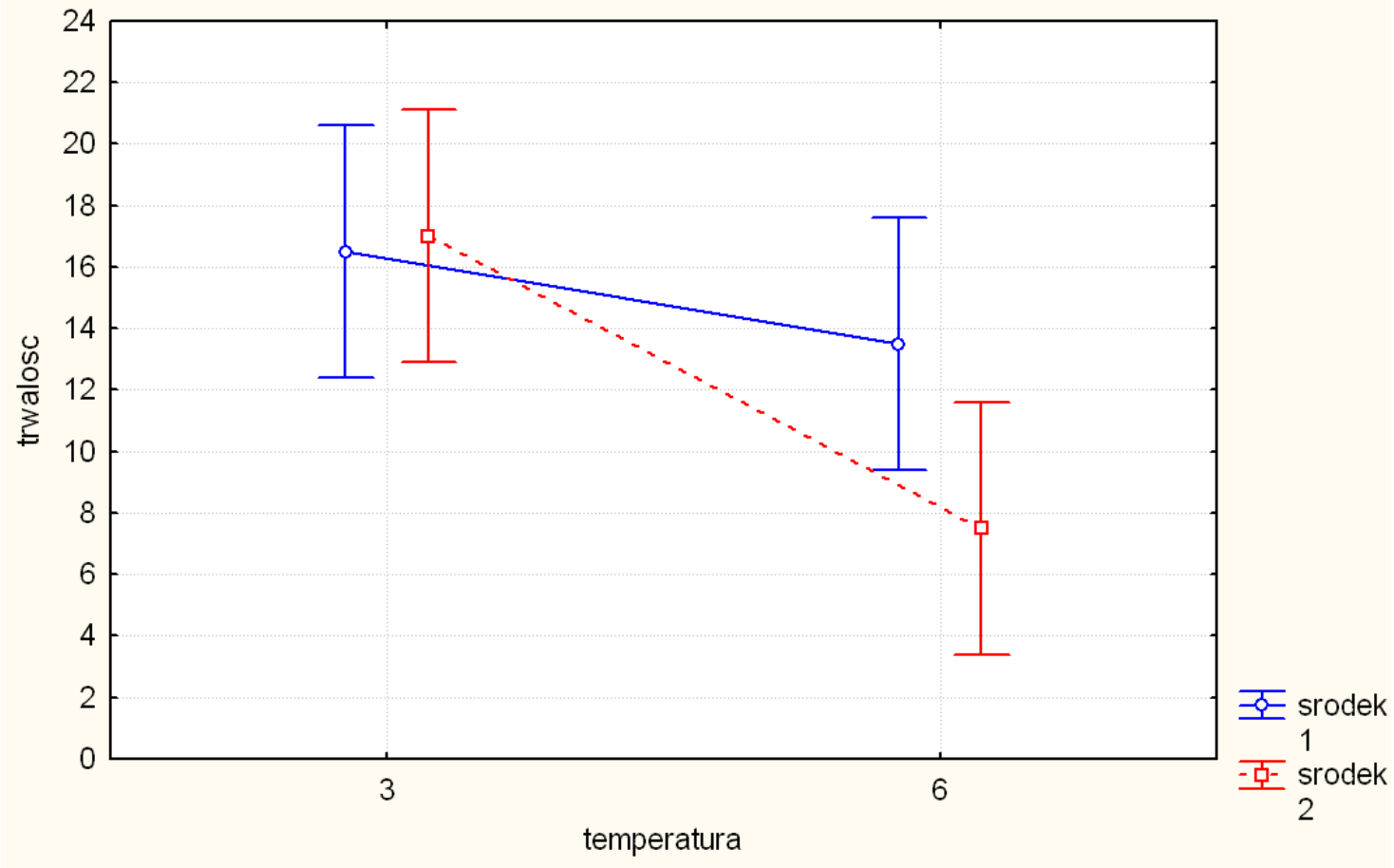
Średnich dla poziomów czynnika A (przy danym poziomie czynnika B)

Średnich dla poziomów czynnika B (przy danym poziomie czynnika A)

Średnich dla kombinacji czynników A i B

Porównania wielokrotne są wykonywane przy użyciu tych samych procedur jak w przypadku jednoczynnikowej analizy wariancji.

Wieloczynnikowa analiza wariancji - układ całkowicie losowy (interakcja czynników)



Wieloczynnikowa analiza wariancji –

układ całkowicie losowy vs układ losowanych bloków

3 poziomy czynnika A (A1, A2, A3), 2 poziomy czynnika B (B1, B2) oraz 3 powtórzenia

Układ całkowicie losowy

A1 B1	A2 B2	A2 B1
A2 B2	A3 B2	A1 B2
A3 B1	A1 B1	A3 B1
A2 B1	A3 B2	A1 B2
A3 B2	A3 B1	A2 B1
A1 B2	A1 B1	A2 B2

Obiekty rozlosowane na obszarze całego doświadczenia w sposób losowy

Układ losowanych bloków

A1 B1	A2 B2	A3 B1
A2 B2	A2 B1	A3 B2
A3 B1	A1 B1	A2 B1
A2 B1	A3 B2	A1 B2
A3 B2	A1 B2	A1 B1
A1 B2	A3 B1	A2 B2
blok 1	**blok 2**	**blok 3**

Kierunek zmienności glebowej

Obiekty rozlosowane w obrębie bloków – w jednym bloku tylko raz występuje każdy obiekt

Wieloczynnikowa analiza wariancji - ANOVA powtarzanych pomiarów

Jednoczynnikowa analiza wariancji dla powtarzanych pomiarów, czyli ANOVA dla grup zależnych (ang. single-factor repeated-measures analysis of variance).

Analiza ta jest metodą pozwalającą na porównanie ze sobą kilku pomiarów przeprowadzonych na tej samej grupie np. osób – tzw. grupy zależne.

W takim badaniu ta sama grupa np. osób jest badana kilkukrotnie, a poszczególne powtórzone jej pomiary stanowią oddzielne grupy, grupy pomiarów.

Układ losowanych bloków – przykład

Oceniano zawartość białka w ziarnie w czterech odmianach pszenicy na podstawie analiz chemicznych ziarna pobranego z losowo wybranych 9 regionów Polski. Na podstawie uzyskanych wyników określić, czy występuje zróżnicowanie pod względem zawartości białka między tymi odmianami pszenicy (przyjąć poziom istotności α=0,05, potraktować region jako blok).

odmiana	zawartość białka (%)	region
A	12,4	1
A	14,0	2
A	13,6	3
A	13,6	4
A	14,4	5
A	14,0	6
A	15,2	7
A	15,2	8
A	15,6	9
B	16,8	1
B	18,0	2
B	17,4	3
B	17,2	4
B	18,6	5
B	18,2	6
B	19,6	7
B	19,4	8
B	20,8	9
C	11,2	1
C	13,6	2
C	12,4	3
C	11,6	4
C	14,0	5
C	13,6	6
C	14,4	7
C	14,0	8
C	14,4	9
D	13,2	1
D	14,8	2
D	14,4	3
D	14,0	4
D	16,0	5
D	15,6	6
D	17,2	7
D	16,8	8
D	18,0	9

Układ losowanych bloków – przykład

Jeśli czynnik grupujący ma 4 poziomy, hipoteza zerowa w analizie wariancji ma postać:

H_{0A}: $\mu_A = \mu_B = \mu_C = \mu_D$ (odmiana nie ma wpływu na plon)

H_{0b}: $\mu_{b1} = \mu_{b2} = \mu_{b3} = \cdots = \mu_{b8}$ (blok nie ma wpływu na plon)

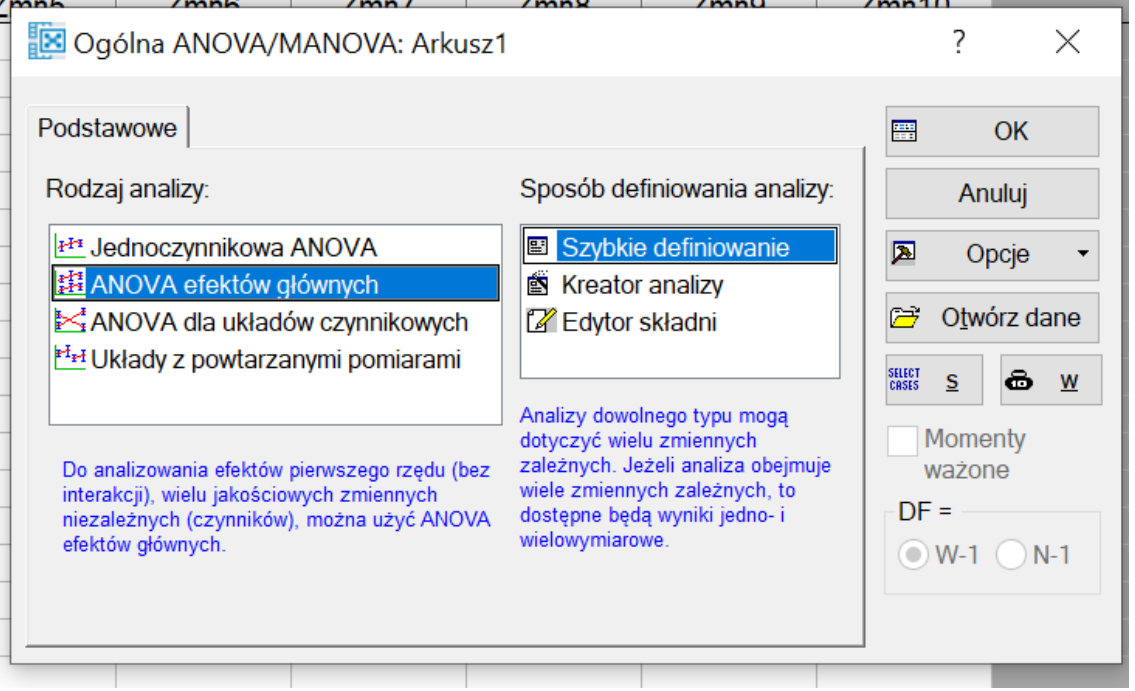
H_1: co najmniej dwie średnie różnią się między sobą ze względu na odmianę

H_1: co najmniej dwie średnie różnią się między sobą ze względu na blok

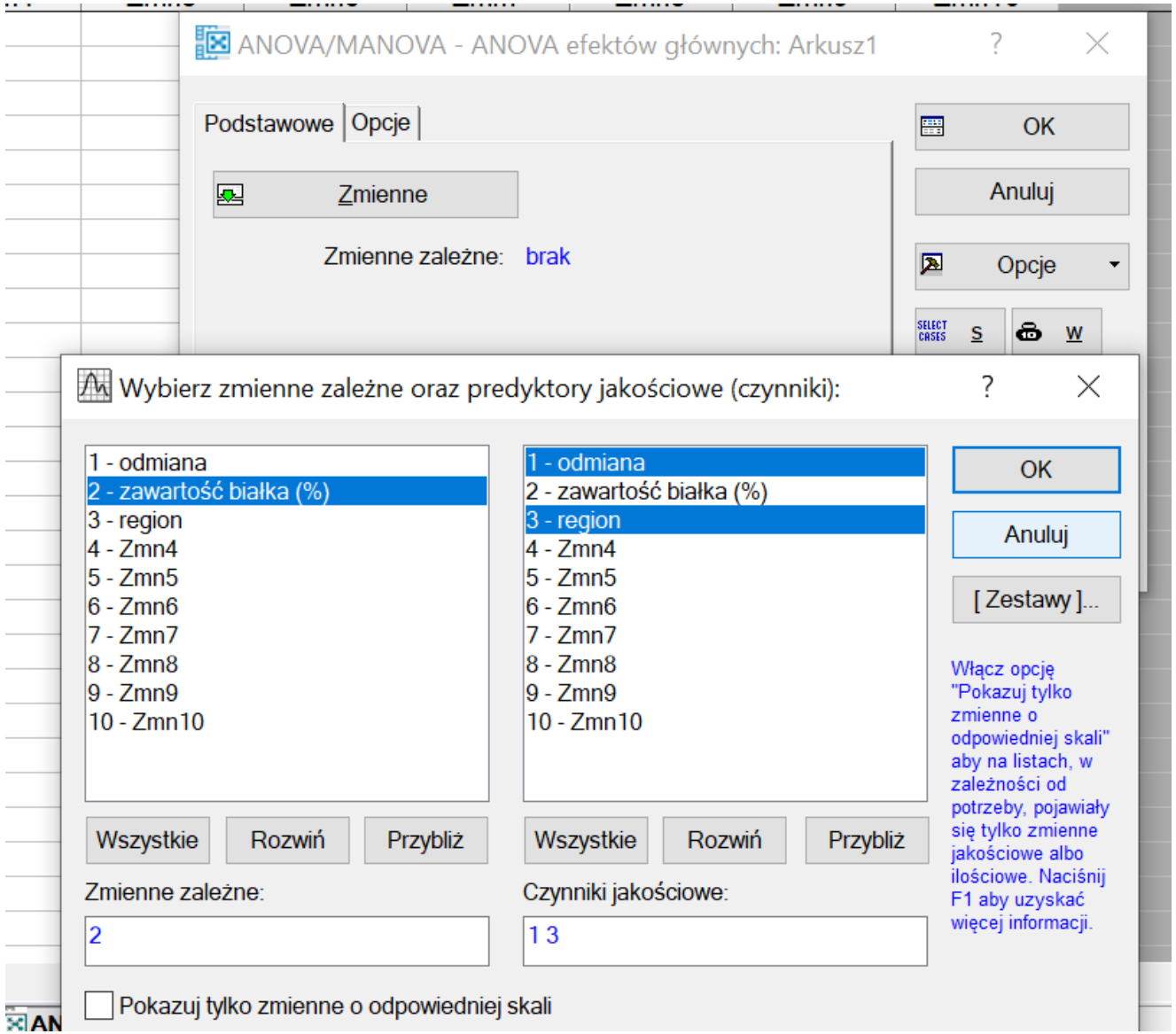
Układ losowanych bloków – przykład

Statistica→ ANOVA→ANOVA efektów głównych

	odmiana	zawartość	region	Zmn4
1	A	12.4	1	
2	A	14	2	
3	A	13.6	3	
4	A	13.6	4	
5	A	14.4	5	
6	A	14	6	
7	A	15.2	7	
8	A	15.2	8	
9	A	15.6	9	
10	B	16.8	1	
11	B	18	2	
12	B	17.4	3	
13	B	17.2	4	
14	B	18.6	5	
15	B	18.2	6	
16	B	19.6	7	
17	B	19.4	8	
18	B	20.8	9	



Układ losowanych bloków – przykład



Układ losowanych bloków – przykład

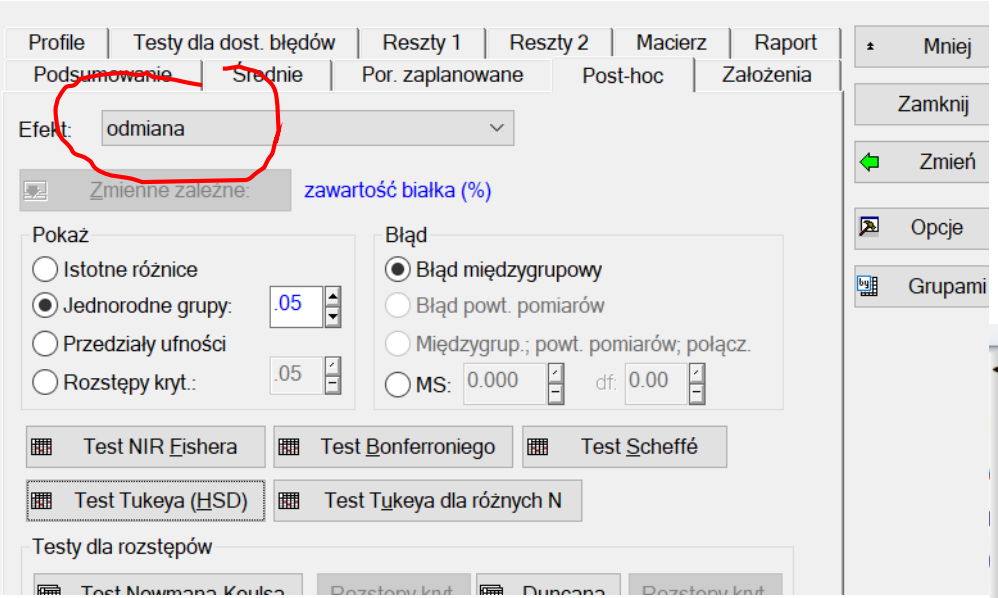
Efekt	Jednowymiarowe testy istotności dla zawartość białka (%) (Arkusz1 Parametryzacja z sigma-ograniczeniami Dekompozycja efektywnych hipotez				
	SS	Stopnie swobody	MS	**F**	p
Wyraz wolny	8500.840	1	8500.840	52283.07	0.000000
odmiana	137.898	3	45.966	282.71	0.000000
region	49.520	8	6.190	38.07	0.000000
Błąd	3.902	24	0.163		

Odrzucamy hipotezy zerowe, gdyż dla odmiany i dla bloku $p<\alpha$

H_1: co najmniej dwie średnie różnią się między sobą ze względu na odmianę
H_1: co najmniej dwie średnie różnią się między sobą ze względu na region

Układ losowanych bloków – przykład

Wydzielenie grup jednorodnych ze względu na odmianę:



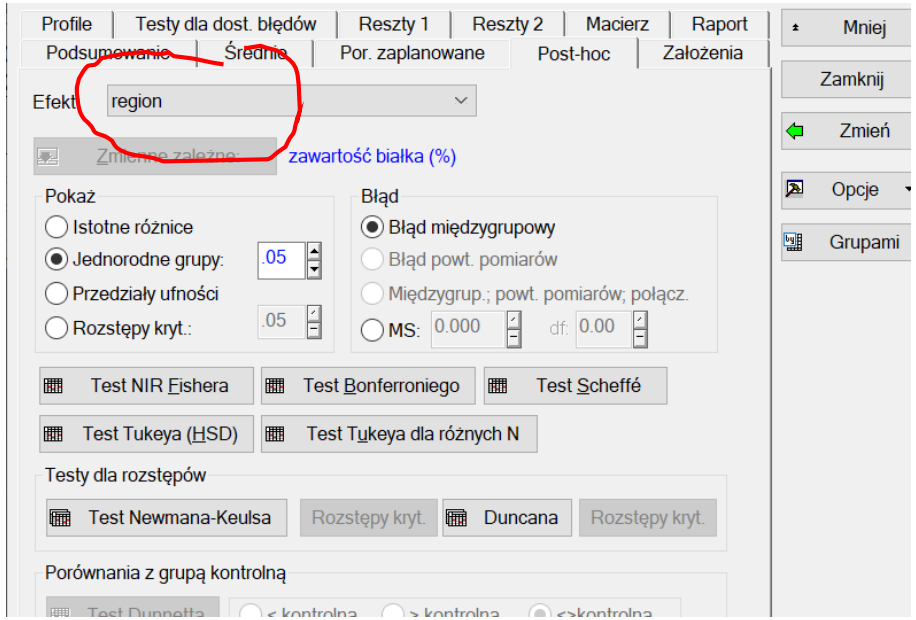
Test HSD Tukeya; zmienna zawartość białka (%) (Arkusz1)
Grupy jednorodne, alfa = .05000
Błąd: MS międzygrupowe = .16259, df = 24.000

Nr podkl.	odmiana	zawartość białka (%) Średnie	1	2	3	4
3	C	13.24444	****			
1	A	14.22222		****		
4	D	15.55556			****	
2	B	18.44444				****

Wszystkie odmiany znacząco różnią się od siebie. Wydzielono cztery grupy jednorodne.

Układ losowanych bloków – przykład

Wydzielenie grup jednorodnych ze względu na region:



Test HSD Tukeya; zmienna zawartość białka (%) (Arkusz1)
Grupy jednorodne, alfa = .05000
Błąd: MS międzygrupowe = .16259, df = 24.000

Nr podkl.	region	zawartość białka (%) Średnie	1	2	3	4	5	6
1	1	13.40000					****	
4	4	14.10000					****	****
3	3	14.45000	****					****
2	2	15.10000	****	****				
6	6	15.35000	****	****				
5	5	15.75000		****	****			
8	8	16.35000			****	****		
7	7	16.60000			****	****		
9	9	17.20000				****		

Wydzielono sześć grup jednorodnych.

Dwuczynnikowa ANOVA układ całkowicie losowy – przykład

Badano zgodność 3 metod określenia łącznej zawartości tłuszczu w mleku UHT. Wybrano losowo po 12 kartonów mleka tłustego, średnio-tłustego i chudego. Zastosowano 3 metody pomiaru (na 4 losowych kartonach każdego rodzaju mleka). Czy metody te działają tak samo dla każdego rodzaju mleka?

	mleko 0,5%	mleko 2%	mleko 3,2%
metoda 1	0.51%	2.02%	3.18%
metoda 1	0.52%	2.04%	3.24%
metoda 1	0.51%	2.02%	3.19%
metoda 1	0.51%	1.95%	3.21%
metoda 2	0.49%	1.90%	3.13%
metoda 2	0.48%	1.98%	3.17%
metoda 2	0.47%	1.92%	3.14%
metoda 2	0.49%	1.92%	3.14%
metoda 3	0.53%	2.07%	3.21%
metoda 3	0.53%	2.09%	3.22%
metoda 3	0.54%	2.04%	3.23%
metoda 3	0.55%	2.01%	3.27%

Układ losowanych bloków – przykład

Jeśli czynnik grupujący A ma 3 poziomy, czynnik grupujący B ma także 3 poziomy hipoteza zerowa w analizie wariancji ma postać:

H_{0A}: $\mu_{0,5\%} = \mu_{2,0\%} = \mu_{3,2\%}$ (rodzaj mleka nie ma wpływu na pomiar)

H_{0B}: $\mu_I = \mu_{II} = \mu_{III}$ (metoda nie ma wpływu na pomiar)

H_{0AxB}: nie występuje istotne współdziałanie czynników (A i B)

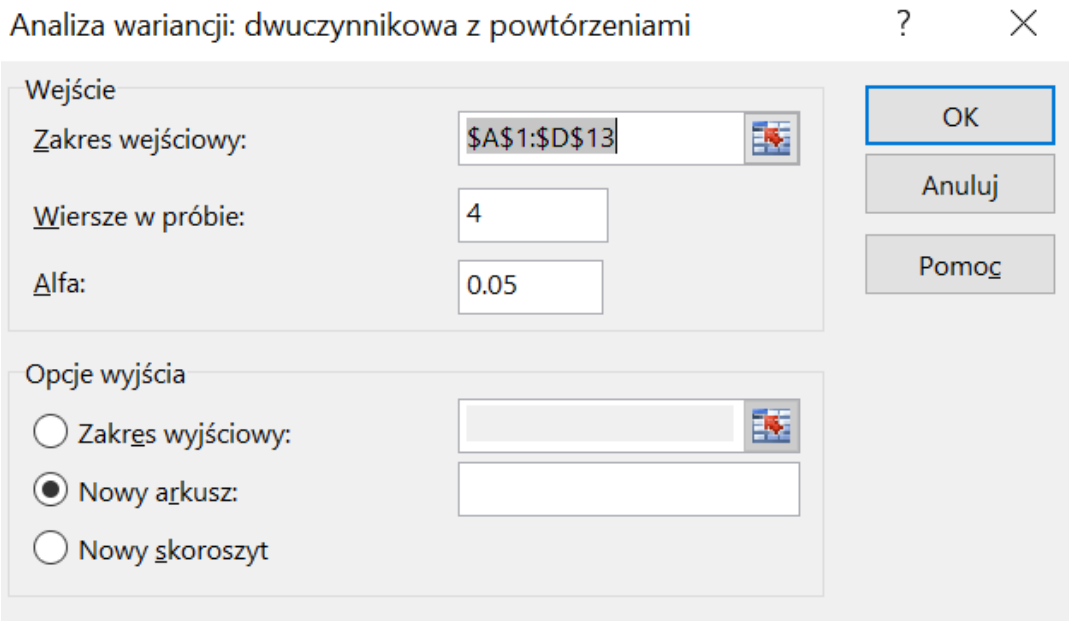
H_1: co najmniej dwie średnie różnią się między sobą ze względu na rodzaj mleka

H_1: co najmniej dwie średnie różnią się między sobą ze względu na metodę pomiaru

Dwuczynnikowa ANOVA układ całkowicie losowy – przykład

Excel: Dane→Analiza wariancji: dwuczynnikowa z powtórzeniami
Dane w postaci jak poniżej, zaznacza się całość (łącznie z nazwami kolumn i wierszy) i są 4 powtórzenia

	mleko 0,5%	mleko 2%	mleko 3,2%
metoda 1	0.51%	2.02%	3.18%
metoda 1	0.52%	2.04%	3.24%
metoda 1	0.51%	2.02%	3.19%
metoda 1	0.51%	1.95%	3.21%
metoda 2	0.49%	1.90%	3.13%
metoda 2	0.48%	1.98%	3.17%
metoda 2	0.47%	1.92%	3.14%
metoda 2	0.49%	1.92%	3.14%
metoda 3	0.53%	2.07%	3.21%
metoda 3	0.53%	2.09%	3.22%
metoda 3	0.54%	2.04%	3.23%
metoda 3	0.55%	2.01%	3.27%



Dwuczynnikowa ANOVA układ całkowicie losowy – przykład

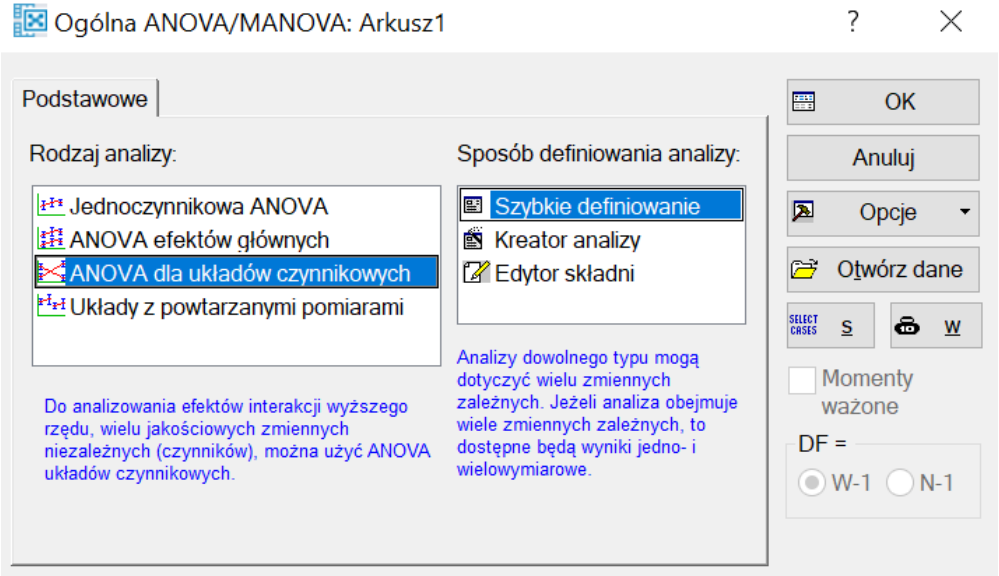
Excel: tabela analizy wariancji
Kolumny to rodzaj mleka (mleko 0,5%, mleko 2%, mleko 3,2%), próbka to metody I, II, III
Są też interakcje tłustości mleka i metody

ANALIZA WARIANCJI						
Źródło wariancji	SS	df	MS	F	Wartość-p	Test F
Próbka	4.79056E-06	2	2.39528E-06	36.64164306	2.03E-08	3.354131
Kolumny	0.004336794	2	0.002168397	33170.94476	1.69E-46	3.354131
Interakcja	4.89444E-07	4	1.22361E-07	1.871813031	0.144381	2.727765
W obrębie	0.000001765	27	6.53704E-08			
Razem	0.004343839	35				

W zadaniu jest pytanie o to, czy metody te działają tak samo dla każdego rodzaju mleka. Czyli pytanie jest o interakcje tych metod. Z analizy wynika, że interakcje są nieistotne, p=0,144 czyli jest większe od α=0,05.
Odpowiedź: Metody te działają tak samo dla każdego rodzaju mleka.
Widać także istotny wpływ każdego czynnika.

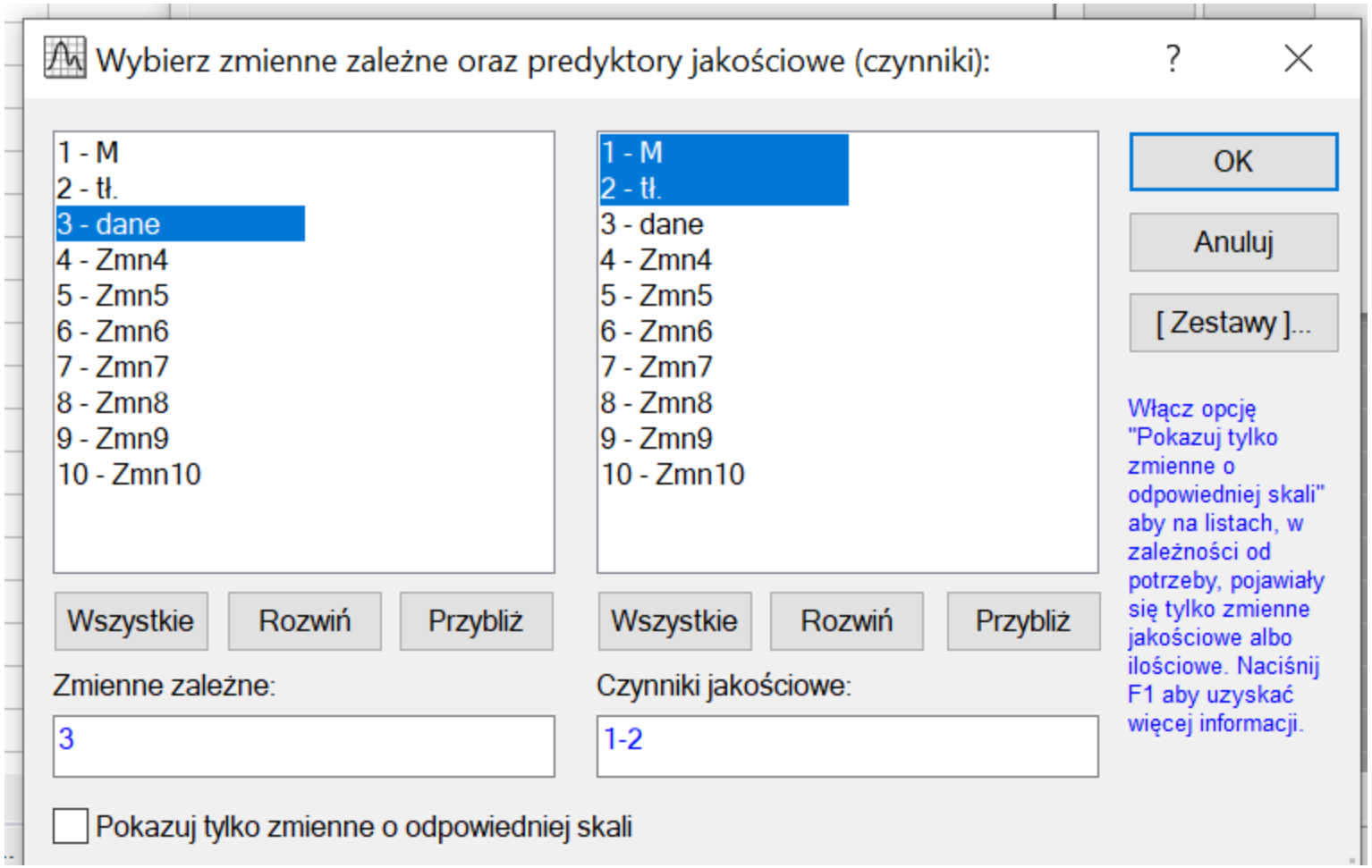
Dwuczynnikowa ANOVA układ całkowicie losowy – przykład

Statistica: ANOVA→ANOVA dla układów czynnikowych
Dane w postaci jak poniżej, każdy czynnik w osobnej kolumnie, dane zależne także w osobnej kolumnie



M	tł.	dane
metoda 1	mleko 0,5%	0.51%
metoda 1	mleko 0,5%	0.52%
metoda 1	mleko 0,5%	0.51%
metoda 1	mleko 0,5%	0.51%
metoda 2	mleko 0,5%	0.49%
metoda 2	mleko 0,5%	0.48%
metoda 2	mleko 0,5%	0.47%
metoda 2	mleko 0,5%	0.49%
metoda 3	mleko 0,5%	0.53%
metoda 3	mleko 0,5%	0.53%
metoda 3	mleko 0,5%	0.54%
metoda 3	mleko 0,5%	0.55%
metoda 1	mleko 2%	2.02%
metoda 1	mleko 2%	2.04%
metoda 1	mleko 2%	2.02%
metoda 1	mleko 2%	1.95%
metoda 2	mleko 2%	1.90%
metoda 2	mleko 2%	1.98%
metoda 2	mleko 2%	1.92%
metoda 2	mleko 2%	1.92%
metoda 3	mleko 2%	2.07%
metoda 3	mleko 2%	2.09%
metoda 3	mleko 2%	2.04%
metoda 3	mleko 2%	2.01%
metoda 1	mleko 3,2%	3.18%
metoda 1	mleko 3,2%	3.24%
metoda 1	mleko 3,2%	3.19%
metoda 1	mleko 3,2%	3.21%
metoda 2	mleko 3,2%	3.13%
metoda 2	mleko 3,2%	3.17%
metoda 2	mleko 3,2%	3.14%
metoda 2	mleko 3,2%	3.14%
metoda 3	mleko 3,2%	3.21%
metoda 3	mleko 3,2%	3.22%
metoda 3	mleko 3,2%	3.23%
metoda 3	mleko 3,2%	3.27%

Dwuczynnikowa ANOVA układ całkowicie losowy – przykład



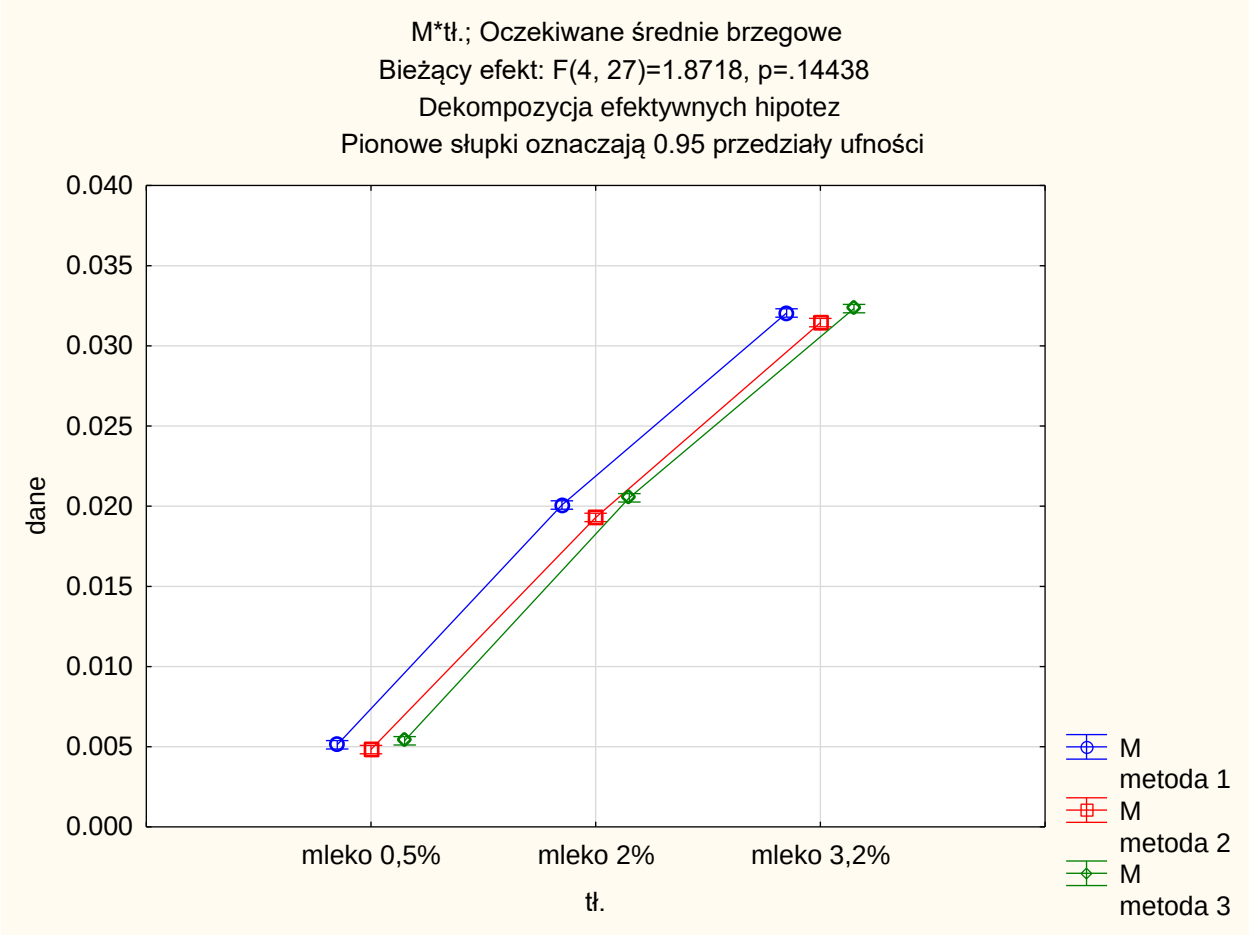
Dwuczynnikowa ANOVA układ całkowicie losowy – przykład

Efekt	Jednowymiarowe testy istotności dla dane (Arkusz1) Parametryzacja z sigma-ograniczeniami Dekompozycja efektywnych hipotez				
	SS	Stopnie swobody	MS	F	p
Wyraz wolny	0.013004	1	0.013004	198921.9	0.000000
M	0.000005	2	0.000002	36.6	0.000000
tł.	0.004337	2	0.002168	33170.9	0.000000
M*tł.	0.000000	4	0.000000	1.9	0.144381
Błąd	0.000002	27	0.000000		

W zadaniu jest pytanie o to, czy metody te działają tak samo dla każdego rodzaju mleka. Czyli pytanie jest o interakcje tych metod. Z analizy wynika, że interakcje są nieistotne, p=0,144 czyli jest większe od α=0,05.
Odpowiedź: Metody te działają tak samo dla każdego rodzaju mleka.
Widać także istotny wpływ każdego czynnika ($p<\alpha$).

Dwuczynnikowa ANOVA układ całkowicie losowy – przykład

W zadaniu jest pytanie o to, czy metody te działają tak samo dla każdego rodzaju mleka. Czyli pytanie jest o interakcje tych metod.
Z analizy wynika, że interakcje są nieistotne, $p>\alpha$.
Z rysunku widać, że krzywe dla metod I, II, III są równoległe – nie ma interakcji.



ANOVA dla powtarzanych pomiarów – przykład

Badacz testował efektywność działania czterech detergentów na rozwój bakterii w kontenerach na mleko po umyciu przez pięć kolejnych dni. Sprawdź, czy między detergentami występują statystycznie znaczące różnice.

			Dzień		
rozpuszczalnik	1	2	3	4	5
A	12	21	17	38	29
A	11	20	16	37	28
B	15	23	16	43	35
B	14	22	17	44	35
C	6	11	7	32	28
C	5	10	10	33	29
D	18	27	23	43	35
D	19	25	20	40	40

ANOVA dla powtarzanych pomiarów – przykład

Detergent to czynnik A. Ma on 4 poziomy.
Każda kombinacja jest powtórzona 2 razy (2 powtórzenia).
Pomiar jest powtarzany przez 5 kolejnych dni (1, 2, 3, 4, 5)
Hipoteza zerowa w analizie wariancji ma postać:
H_{0A}: $\mu_A = \mu_B = \mu_C = \mu_D$ (detergenty nie mają wpływu na rozwój bakterii)
H_{0B}: $\mu_1 = \mu_2 = \mu_3 = \mu_4 = \mu_5$ (czas badania nie ma wpływu na rozwój bakterii)

H_{0AxB}: nie występuje istotne współdziałanie czynników (A i B)

H_1: co najmniej dwie średnie różnią się między sobą ze względu na detergenty
H_1: co najmniej dwie średnie różnią się między sobą ze względu na czas badania

ANOVA dla powtarzanych pomiarów – przykład

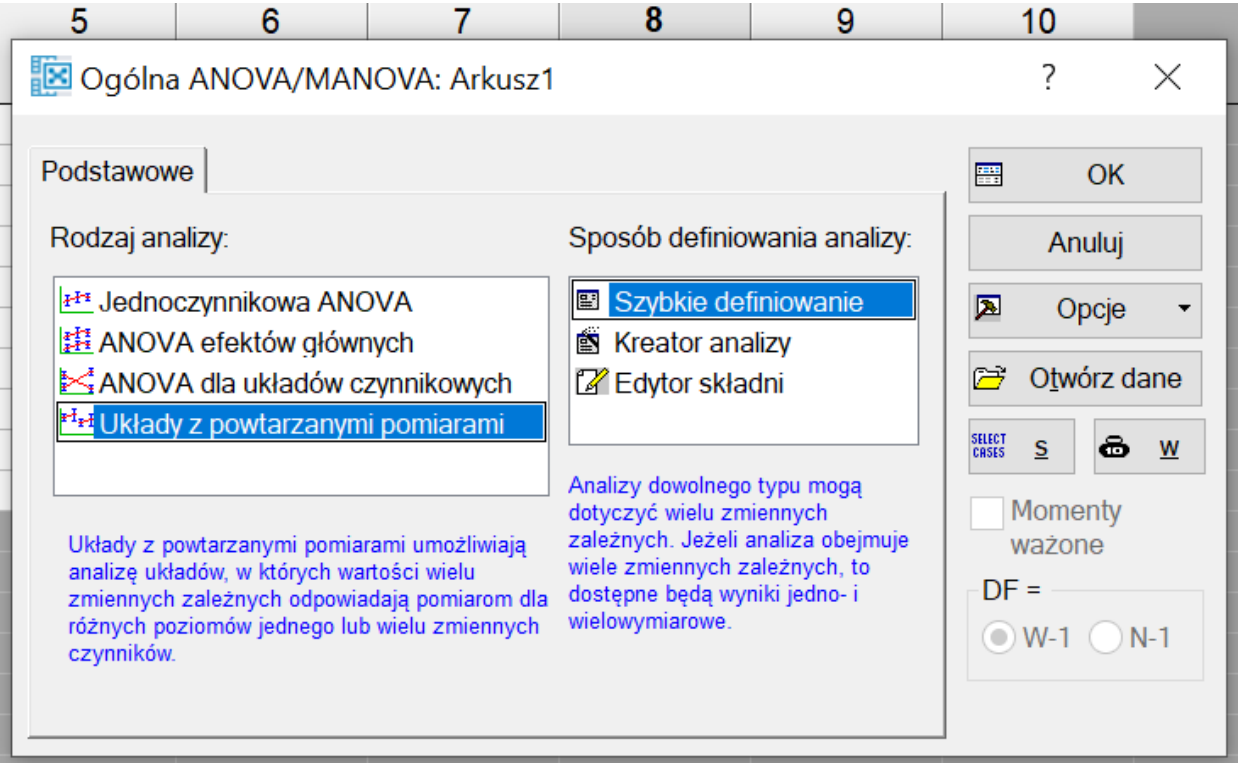
Statistica: dane wklejamy z nagłówkami jako 4 kolumny.

	1 rozpuszczalnik	2 1	3 2	4 3	5 4	6 5	7 Zmn7	**8 Zmn8**	9 Zmn9	10 Zmn10
1	A	12	21	17	38	29				
2	A	11	20	16	37	28				
3	B	15	23	16	43	35				
4	B	14	22	17	44	35				
5	C	6	11	7	32	28				
6	C	5	10	10	33	29				
7	D	18	27	23	43	35				
8	D	19	25	20	40	40				
9										
10										

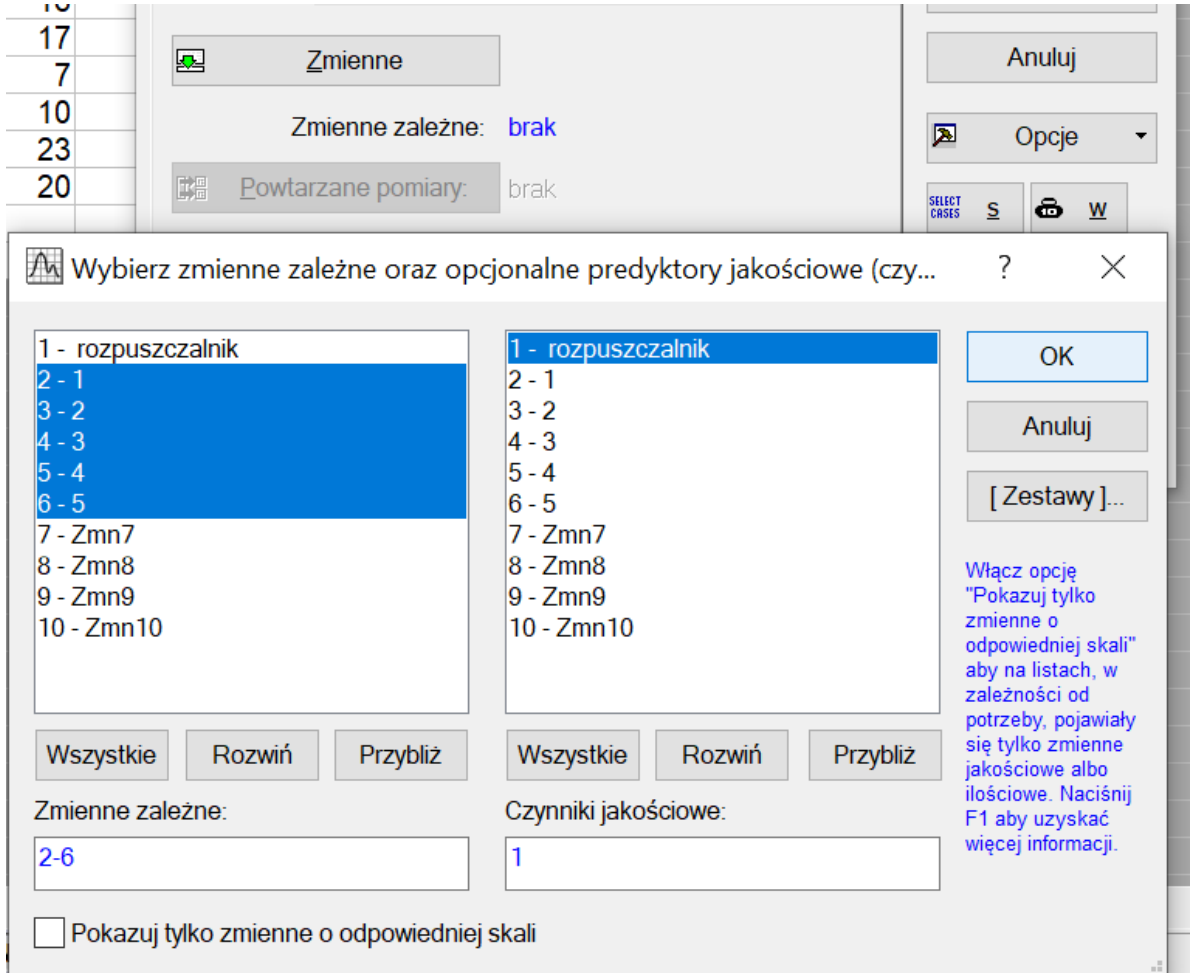
ANOVA dla powtarzanych pomiarów – przykład

Statistica: ANOVA →Układy z powtarzanymi pomiarami

	1 rozpuszczalnik	2 1	3 2	4 3
1	A	12	21	17
2	A	11	20	16
3	B	15	23	16
4	B	14	22	17
5	C	6	11	7
6	C	5	10	10
7	D	18	27	23
8	D	19	25	20
9				
10				

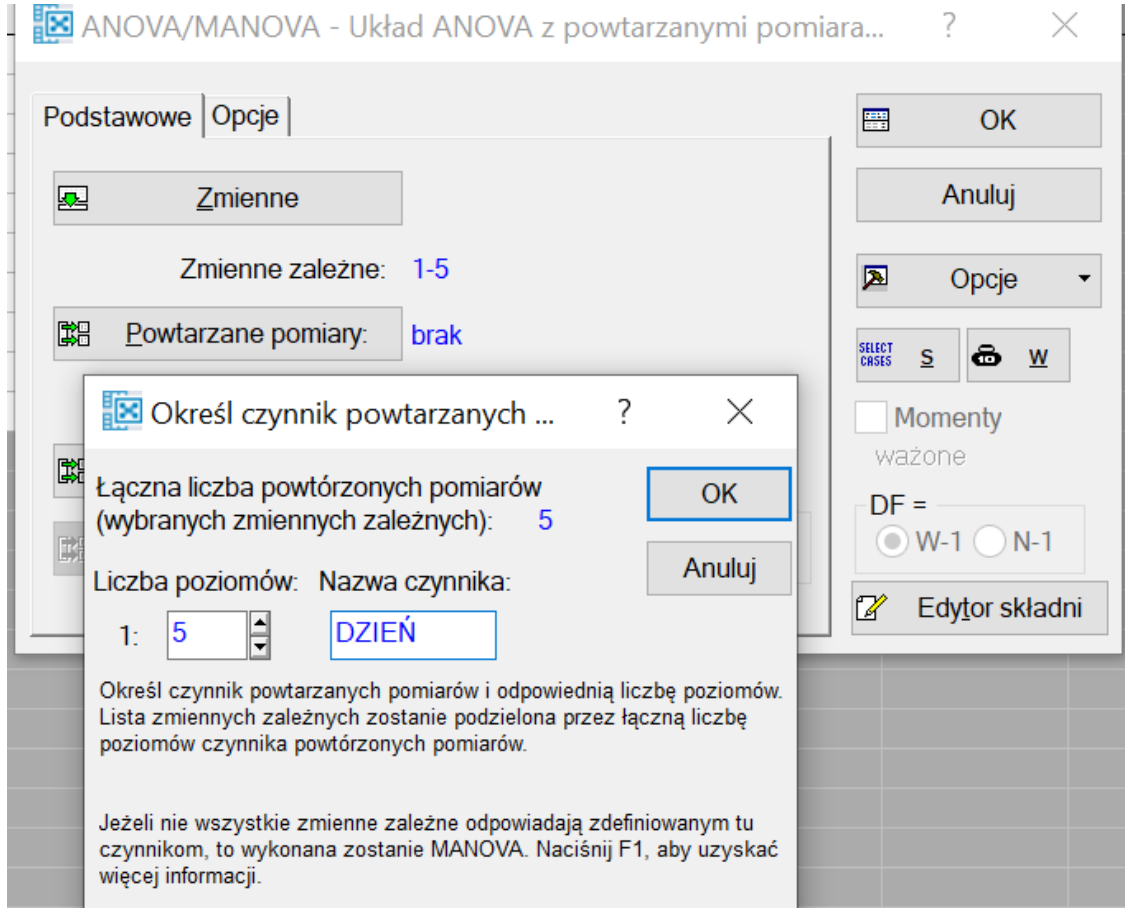


ANOVA dla powtarzanych pomiarów – przykład



ANOVA dla powtarzanych pomiarów – przykład

Teraz należy wcisnąć „powtarzane pomiary” i wpisać liczbę poziomów i nazwę czynnika.



ANOVA dla powtarzanych pomiarów – przykład

Analiza wariancji dla powtarzanych pomiarów (Arkusz1)
Parametryzacja z sigma-ograniczeniami
Dekompozycja efektywnych hipotez

Efekt	SS	Stopnie swobody	MS	F	p
Wyraz wolny	22752.90	1	22752.90	23950.42	0.000000
rozpuszczalnik	794.90	3	264.97	278.91	0.000042
Błąd	3.80	4	0.95		
DZIEŃ	4039.35	4	1009.84	517.87	0.000000
DZIEŃ* rozpuszczalnik	91.85	12	7.65	3.93	0.006200
Błąd	31.20	16	1.95		

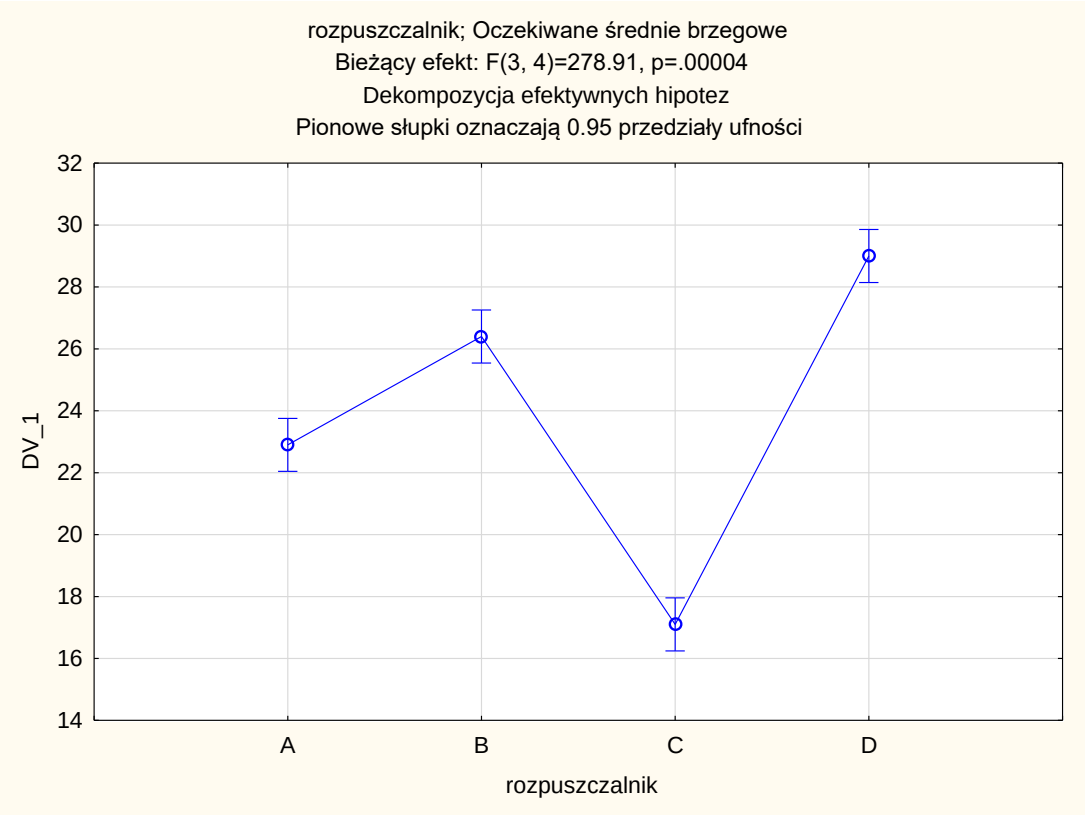
Rozpuszczalnik ma wpływ na liczbę bakterii. Dzień pomiaru także wpływa na liczbę bakterii. Obserwuje się także istotną interakcję.

ANOVA dla powtarzanych pomiarów – przykład

W zadaniu jest pytanie o to, czy rozpuszczalnik ma wpływ na liczbę bakterii. Rozpuszczalnik ma istotny wpływ na liczbę bakterii. Wyodrębnia się 4 grupy jednorodne.

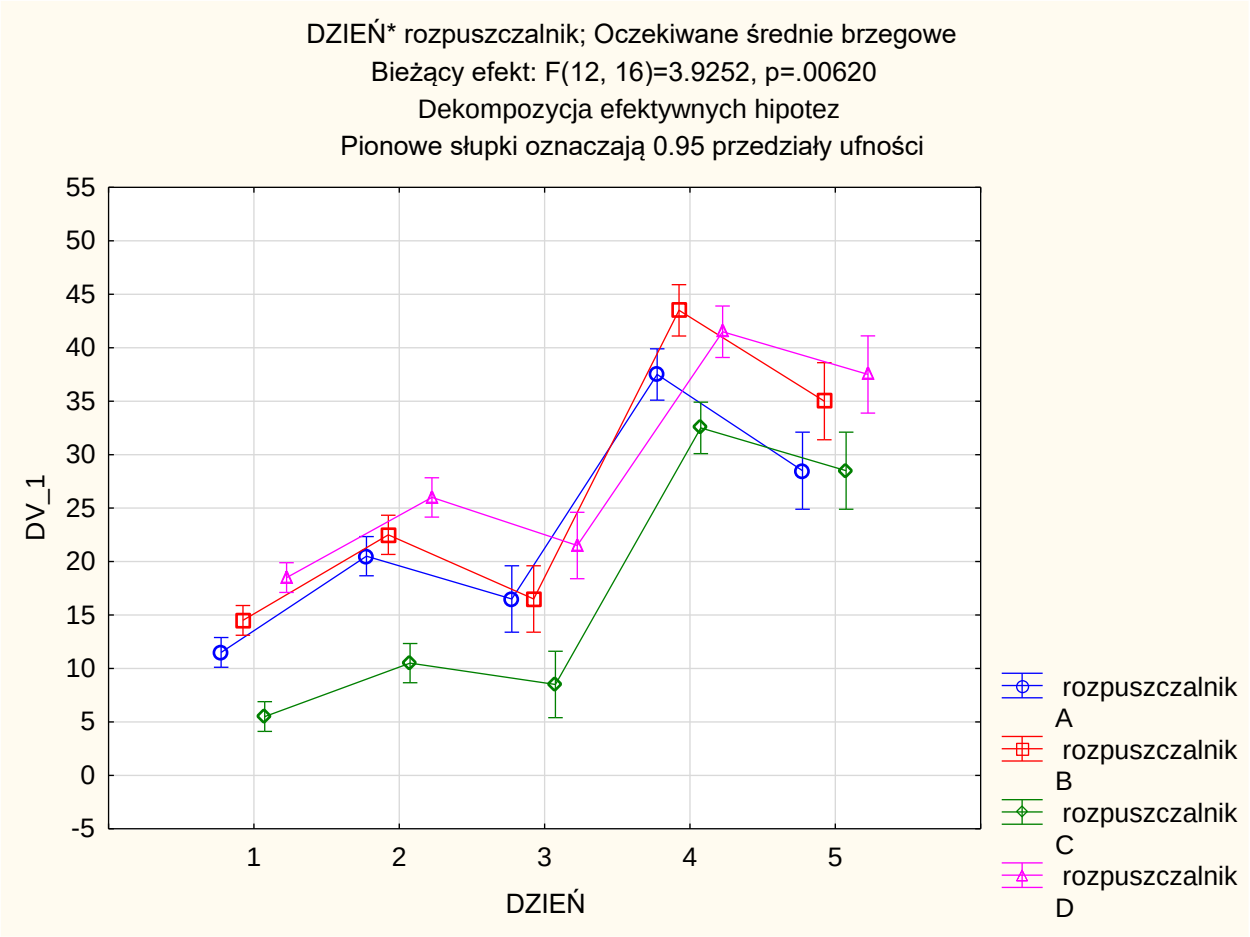
Test HSD Tukeya; zmienna DV_1 (Arkusz1)
Grupy jednorodne, alfa = .05000
Błąd: MS międzygrupowe = .95000, df = 4.0000

Nr podkl.	rozpuszczalnik	DV_1 Średnie	1	2	3	4
3	C	17.10000	****			
1	A	22.90000		****		
2	B	26.40000			****	
4	D	29.00000				****



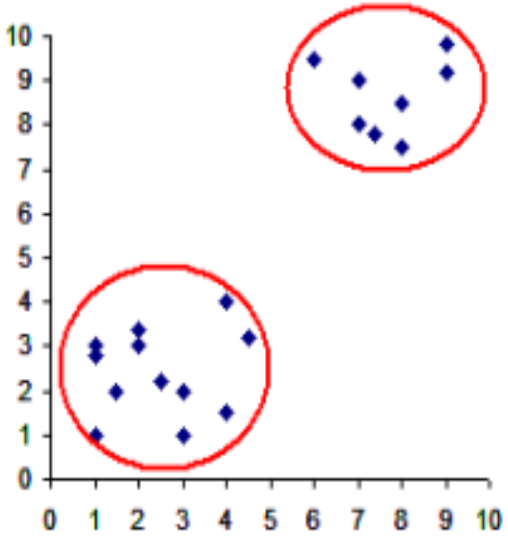
ANOVA dla powtarzanych pomiarów – przykład

Interakcje



Analiza skupień – wielowymiarowa klasyfikacja obiektów

Metoda, pozwalająca na grupowanie obiektów pod względem wielu cech jednocześnie. W przypadku dwóch lub trzech cech możliwe jest wskazanie obiektów podobnych do siebie na podstawie wartości tych cech (X, Y ewentualnie Z) na wykresie punktowym. W przypadku grupowania cech pod względem więcej niż 3 cech (zmiennych) nie ma możliwości graficznego przedstawienia na wykresie wartości wszystkich cech. Możliwe jest natomiast określenie odległości między obiektami w przestrzeni wielowymiarowej



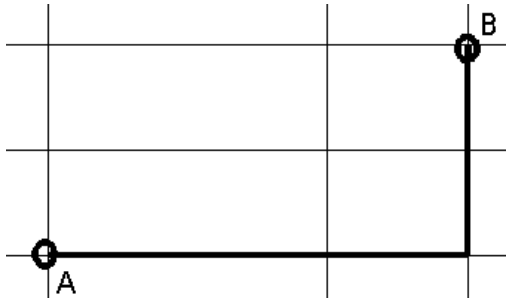
Analiza skupień – wielowymiarowa klasyfikacja obiektów

Odległość Euklidesowa w przestrzeni o *p* wymiarach między dwoma obiektami:

$X_{ij;}$ X_{kj} – wartości j-tej cechy dla obiektów *i* oraz *k*
p – liczba cech/zmiennych

$$d(x_i, x_k) = d_{ik} = \sqrt{\sum_{j=1}^{p} (x_{ij} - x_{kj})^2}$$

Istnieją również inne miary określania odległości, np. odległość miejska (tzw. city block lub typu Manhattan)



Analiza skupień – wielowymiarowa klasyfikacja obiektów

Ze względu na stosowanie różnych jednostek poszczególnych cech oraz różnych skal wartości, zwykle odległość między obiektami jest określana na podstawie zmiennych standaryzowanych

$$z_{ij} = \frac{x_{ij} - \bar{x}_j}{S_j}$$

z_{ij} – *wartość po standaryzacji*

x_{ij} – *wartość przed standaryzacją*

x_j – *wartość średnia cechy*

S_j – *odchylenie standardowe*

Standaryzacja zmiennych pozwala w analizie skupień na zachowanie podobnej wagi każdej zmiennej w klasyfikacji obiektów.

Analiza skupień – metody grupowania

<u>Hierarchiczne</u>

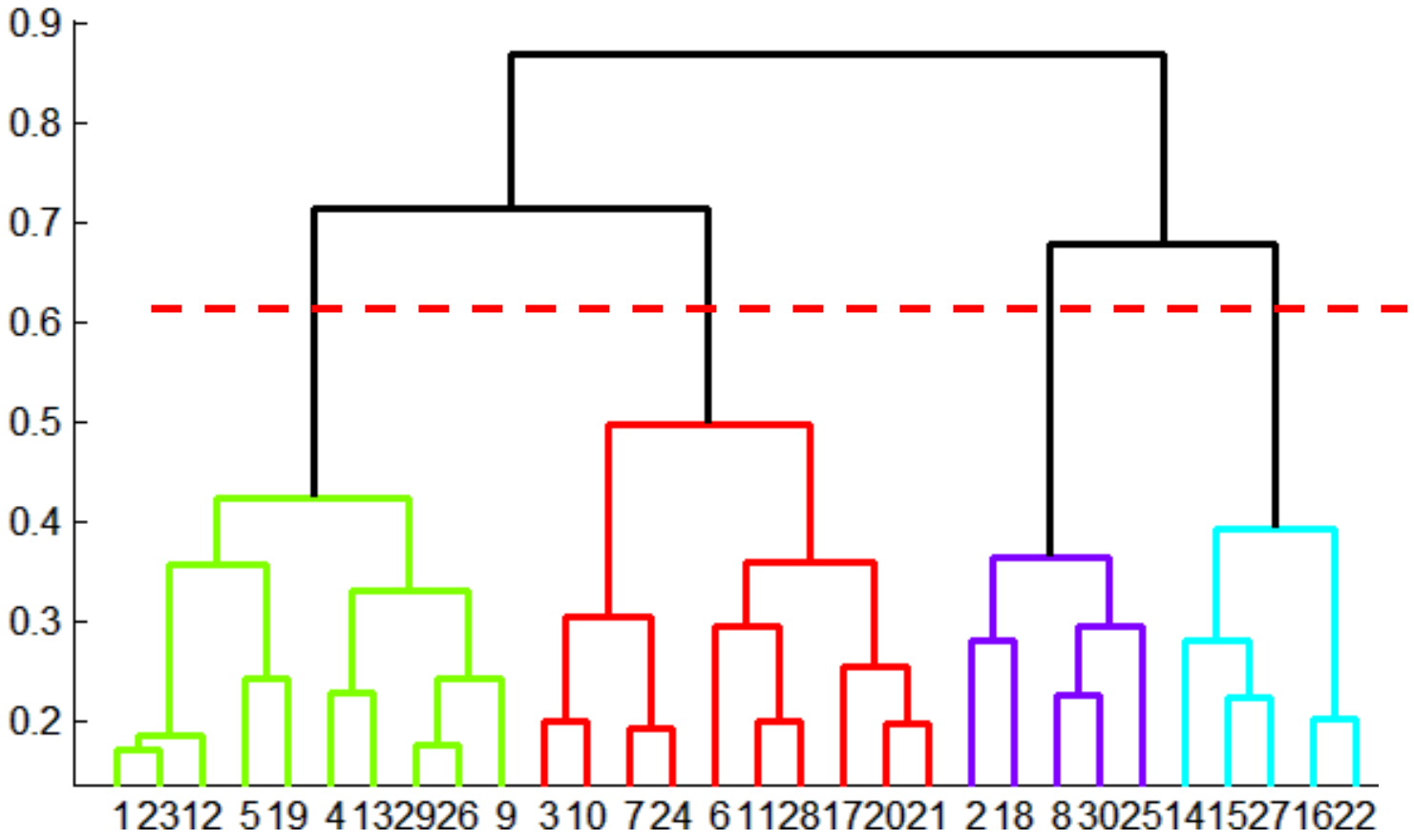
pozwalają na łączenie obiektów w grupy z zachowaniem hierarchii, tzn. możemy określić na podstawie dendrogramu które obiekty w obrębie wydzielonych grup są podobne, a które bardziej odległe

Wybrane metody aglomeracji (łączenia obiektów) w grupowaniu hierarchicznym:
- metoda najbliższego sąsiada
- metoda najdalszego sąsiada
- metoda Warda
- metoda centroidów

<u>Niehierarchiczne</u>

po zaliczeniu obiektu do danej grupy nie możemy powiedzieć, które z obiektów w obrębie jednej grupy są bardziej podobne.

Metodą niehierarchiczną jest metoda *k*-średnich.



Przykładowy dendrogram, powstały jako wynik analizy skupień. Kreskowana linia czerwona przedstawia podział obiektów na 4 grupy. Istnieje dowolność w ustalaniu liczby grup, tak więc można obiekty podzielić na większą lub tez mniejszą liczbę grup w zależności, od stawianych celów analizy

Analiza skupień – przykłady zastosowań

1) Wydzielenie grup odmian jabłoni podobnych pod względem wielu cech np.:

- koloru owoców (kolor musi być wyrażony ilościowo tzn. w postaci liczby np.

w skali 5 stopniowej 1- zielony…. 5- czerwony)

- wielkości owoców

- szybkości wzrostu

- wrażliwości na choroby

itp.

2) Wydzielenie grup gmin podobnych pod względem wielu cech np.:

- liczba mieszkańców

- dochody w przeliczeniu na mieszkańca

- stopa bezrobocia

- udział powierzchni użytków rolnych, lasów, sadów

-itp.

Analiza skupień – przykłady zastosowań

Dla scharakteryzowania poszczególnych wydzielonych grup obiektów można wykorzystać wartości średnich poszczególnych cech (zmiennych) na podstawie których była wykonana klasyfikacja.

Średnie wartości cech dla obiektów z poszczególnych grup nazywamy centroidami.

Nie ma jednej obiektywnej metody ustalania liczba wydzielonych grup obiektów. Liczba ta może być ustalana bądź przed wykonaniem analiz, lub też po wykonaniu analiz, np. na podstawie dendrogramu.

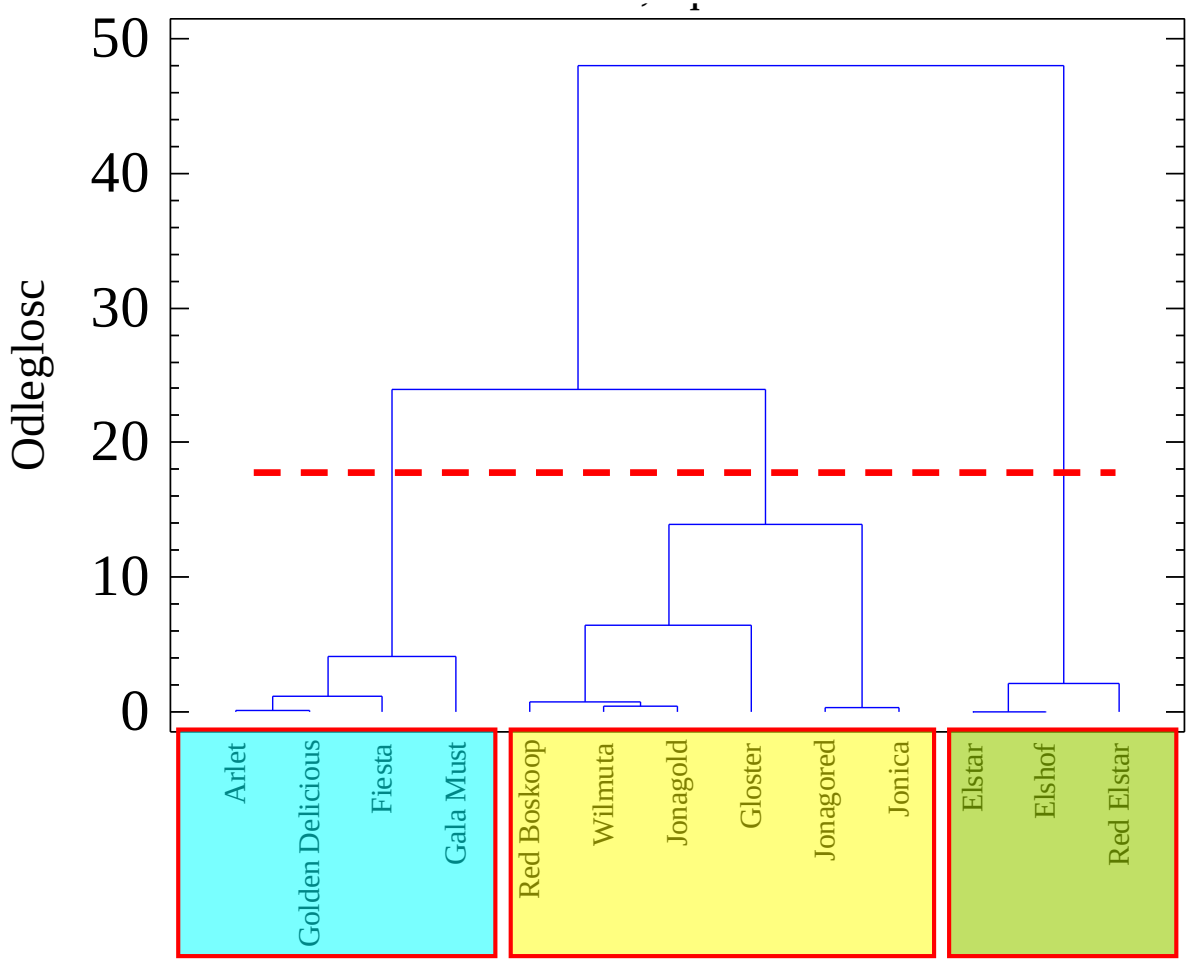
Analiza skupień – przykład

Podział odmian jabłoni podstawie kilku cech

Wartości średnie wybranych cech kilkunastu odmian jabłoni

odmiana	powierzchnia przekroju pnia (cm^2)	średnia produktywnośc drzew (kg/cm^2)	średni plon (t/ha)	średnia masa owocu (g)
Arlet	35,2	0,48	29,3	144
Red Boskoop	34	0,44	24,3	179
Fiesta	36,3	0,51	33,2	148
Gala Must	24,5	0,52	23,6	141
Golden Delicious Reinders	33,5	0,52	29	147
Jonagored	31,2	0,32	20,2	164
Jonica	28,6	0,34	18,2	167
Wilmuta	34,5	0,5	25,8	187
Jonagold	33,1	0,45	26,8	186
Elstar	62	0,3	33	135
Red Elstar	66,2	0,27	27	138
Elshof	60,2	0,3	32	139
Gloster	35,2	0,45	34,2	177

Dane na podstawie:
http://www.up.poznan.pl/ogrodnictwo/Ogrodnictwo%2041/67%20Wocir.pdf



Wartości średnie (centroidy) ocenianych cech dla wydzielonych grup w analizie skupień

Grupa	powierzchnia przekroju pnia (cm^2)	średnia produktywnośc drzew (kg/cm^2)	średni plon (t/ha)	średnia masa owocu (g)
1	32,4	0,51	28,8	145,0
2	32,8	0,42	24,9	176,7
3	62,8	0,29	30,7	137,3

Analiza skupień – przykład

Statistica – przygotowanie danych

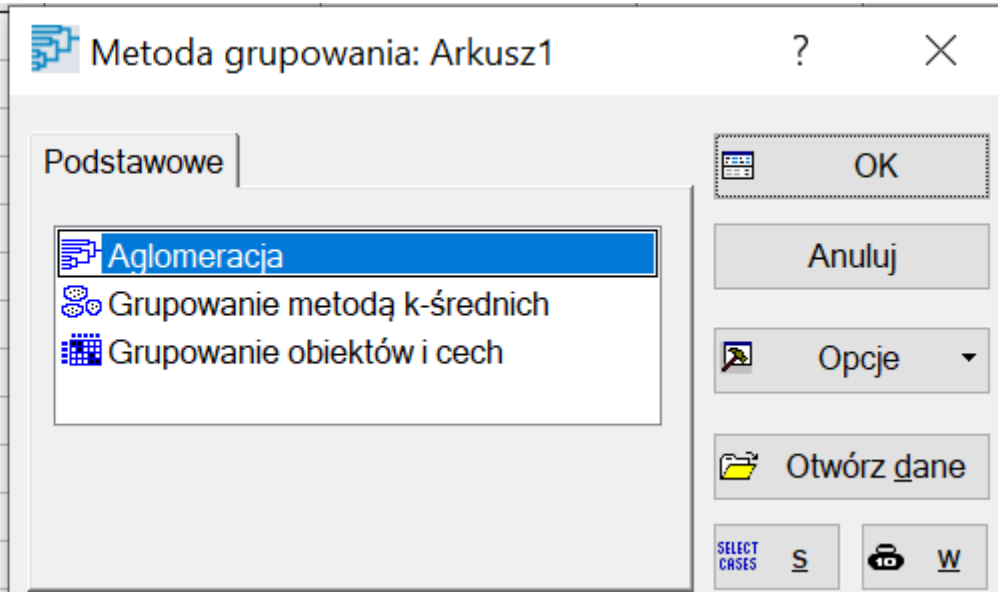
	1 powierzchnia	2 średnia produktywność	3 średni plon	4 średnia masa
Arlet	35.2	0.48	29.3	144
Red Boskoop	34	0.44	24.3	179
Fiesta	36.3	0.51	33.2	148
Gala Must	24.5	0.52	23.6	141
Golden Delicious Reinders	33.5	0.52	29	147
Jonagored	31.2	0.32	20.2	164
Jonica	28.6	0.34	18.2	167
Wilmuta	34.5	0.5	25.8	187
Jonagold	33.1	0.45	26.8	186
Elstar	62	0.3	33	135
Red Elstar	66.2	0.27	27	138
Elshof	60.2	0.3	32	139
Gloster	35.2	0.45	34.2	177

Analiza skupień – przykład

Statistica – uruchomienie analizy

Statystyka→Wielowymiarowe→Analiza skupień→Aglomeracja

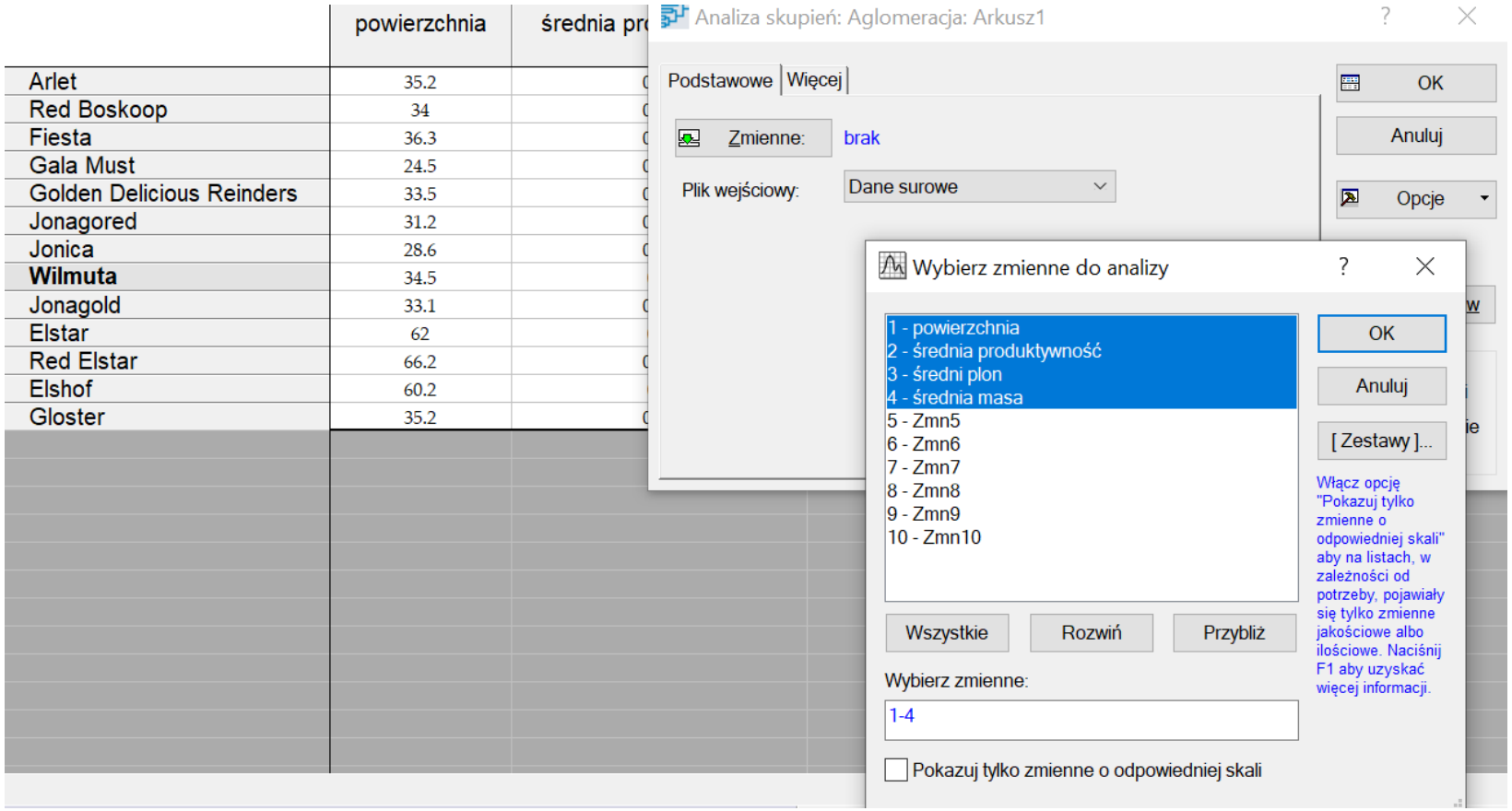
Arlet	35.2	0.48
Red Boskoop	34	0.44
Fiesta	36.3	0.51
Gala Must	24.5	0.52
Golden Delicious Reinders	33.5	0.52
Jonagored	31.2	0.32
Jonica	28.6	0.34
Wilmuta	34.5	0.5
Jonagold	33.1	0.45
Elstar	62	0.3
Red Elstar	66.2	0.27
Elshof	60.2	0.3



Analiza skupień – przykład

Statistica

Wybór zmiennych

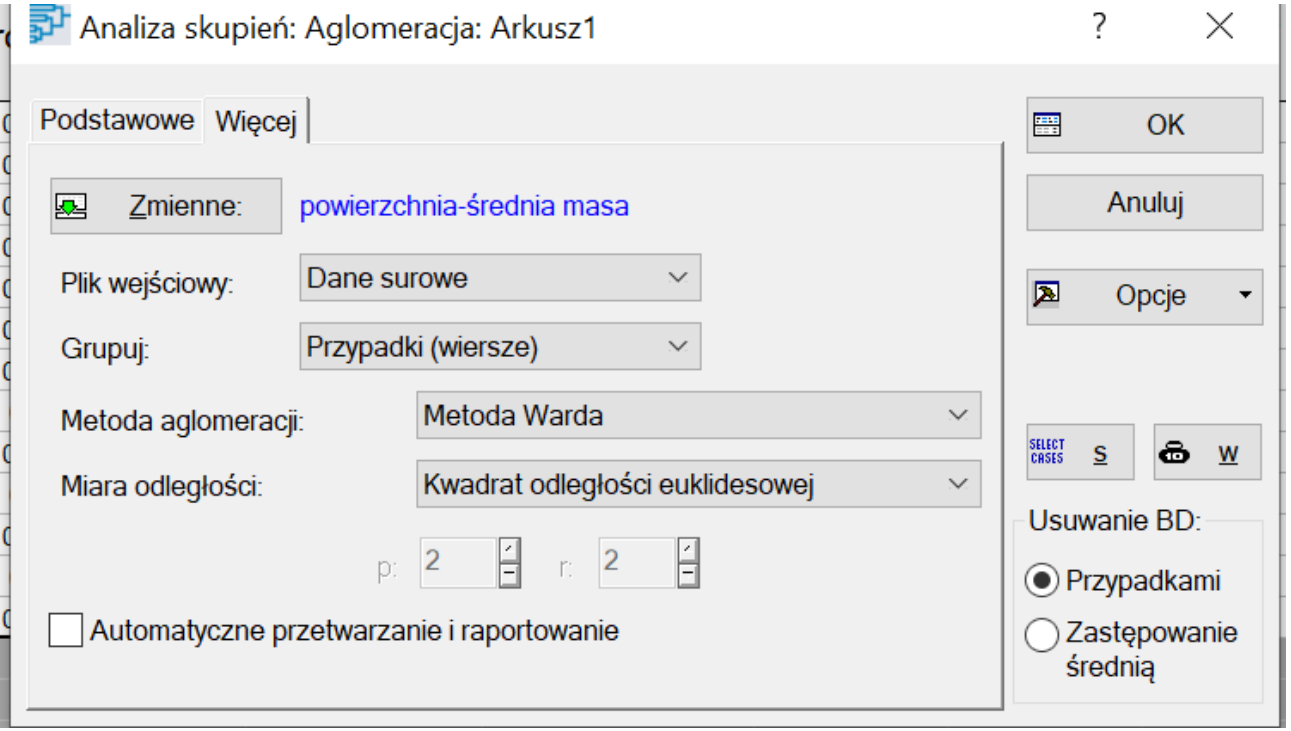


Analiza skupień – przykład

Statistica

Ustawienia do analizy

	powierzchnia	średnia pr
Arlet	35.2	0
Red Boskoop	34	0
Fiesta	36.3	0
Gala Must	24.5	0
Golden Delicious Reinders	33.5	0
Jonagored	31.2	0
Jonica	28.6	0
Wilmuta	34.5	
Jonagold	33.1	0
Elstar	62	
Red Elstar	66.2	0
Elshof	60.2	
Gloster	35.2	0



W następnym kroku wybieramy „Pionowy wykres sopelkowy”

Analiza skupień – przykład

Statistica

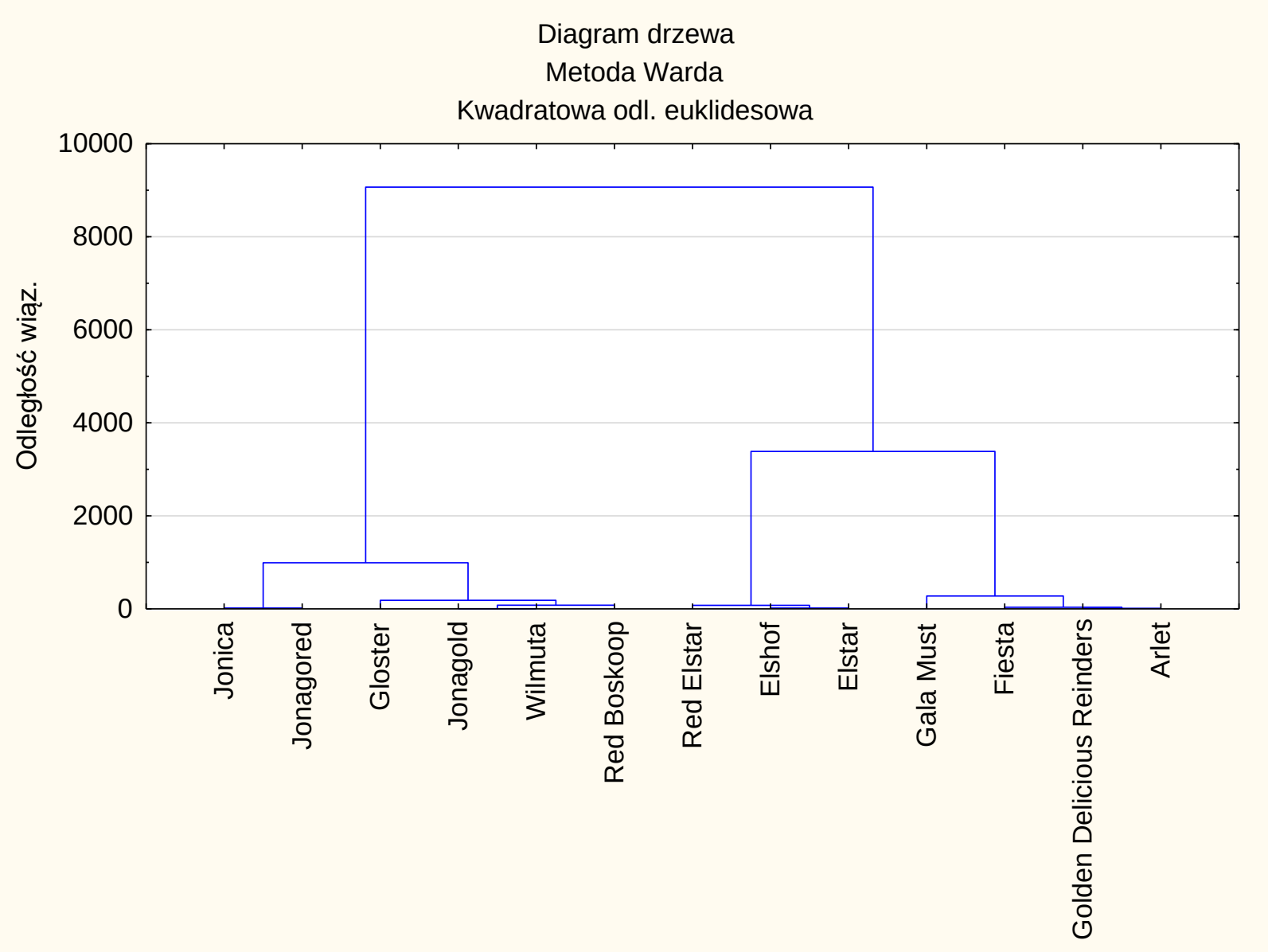
Wynik:

Na wykresie widać, które obiekty, czyli odmiany, są do siebie podobne ze względu na oceniane cechy.

Red Elstar, Elshof i Elstar są do siebie podobne.

Gloster, Jonagold, Wilmuta i Red Boskoop są do siebie podobne.

Gala Must, Fiesta, Golden Delicious Reinders i Arlet są do siebie podobne.



PCA

PCA jest często używana do redukcji liczby zmiennych zbioru danych statystycznych, poprzez wyznaczenie nowych zmiennych (składowych głównych) skorelowanych z zmiennymi źródłowymi.

Dzięki temu możemy zredukować liczbę np. do dwóch lub trzech zmiennych, co umożliwia uproszczone wielowymiarowe porządkowanie obiektów i ocenę wizualną zróżnicowania obiektów w układzie współrzędnych o dwóch lub trzech wymiarach.

Każda ze składowych głównych (czyli nowych zmiennych) wyjaśnia określony % zmienności całkowitej wszystkich zmiennych. Jeśli % wyjaśnianej zmienności przez pierwszą i drugą składową główną jest duży (bliski 100%), to możemy z wykorzystaniem dwuwymiarowego układu współrzędnych właściwie ocenić zróżnicowanie wielowymiarowe badanych obiektów.

Taka redukcja liczby wymiarów jest możliwa tylko w przypadku występowania dość silnych korelacji między poszczególnymi cechami

PCA – przykład

Podział odmian jabłoni podstawie kilku cech

Wartości średnie wybranych cech kilkunastu odmian jabłoni

odmiana	powierzchnia przekroju pnia (cm^2)	średnia produktywnośc drzew (kg/cm^2)	średni plon (t/ha)	średnia masa owocu (g)
Arlet	35,2	0,48	29,3	144
Red Boskoop	34	0,44	24,3	179
Fiesta	36,3	0,51	33,2	148
Gala Must	24,5	0,52	23,6	141
Golden Delicious Reinders	33,5	0,52	29	147
Jonagored	31,2	0,32	20,2	164
Jonica	28,6	0,34	18,2	167
Wilmuta	34,5	0,5	25,8	187
Jonagold	33,1	0,45	26,8	186
Elstar	62	0,3	33	135
Red Elstar	66,2	0,27	27	138
Elshof	60,2	0,3	32	139
Gloster	35,2	0,45	34,2	177

Dane na podstawie:
http://www.up.poznan.pl/ogrodnictwo/Ogrodnictwo%2041/67%20Wocir.pdf

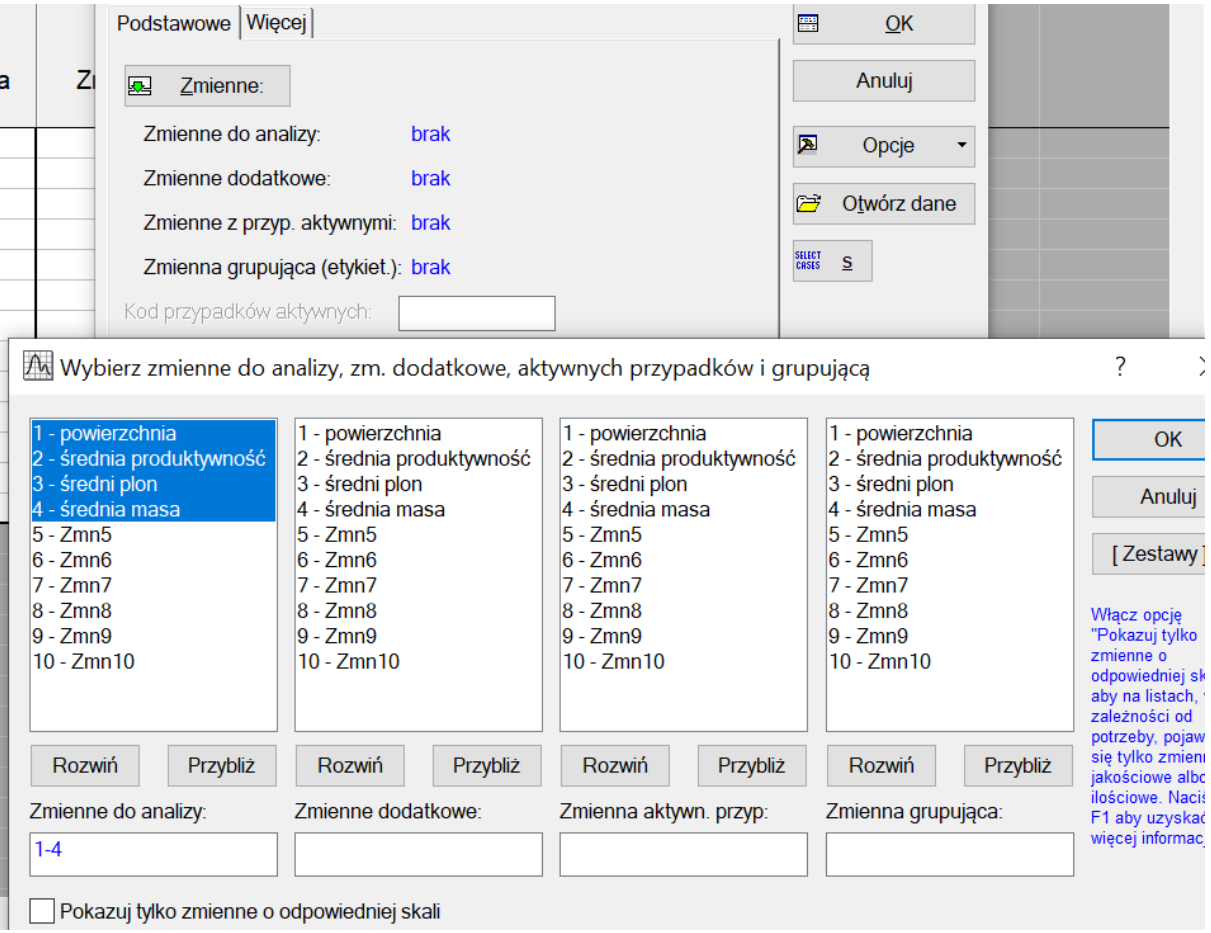
PCA – przykład

Statistica (dane wklejone z obiema nazwami: kolumn i wierszy)

Statystyka→Analiza składowych głównych i klasyfikacja

Następnie należy wybrać zmienne do analizy.

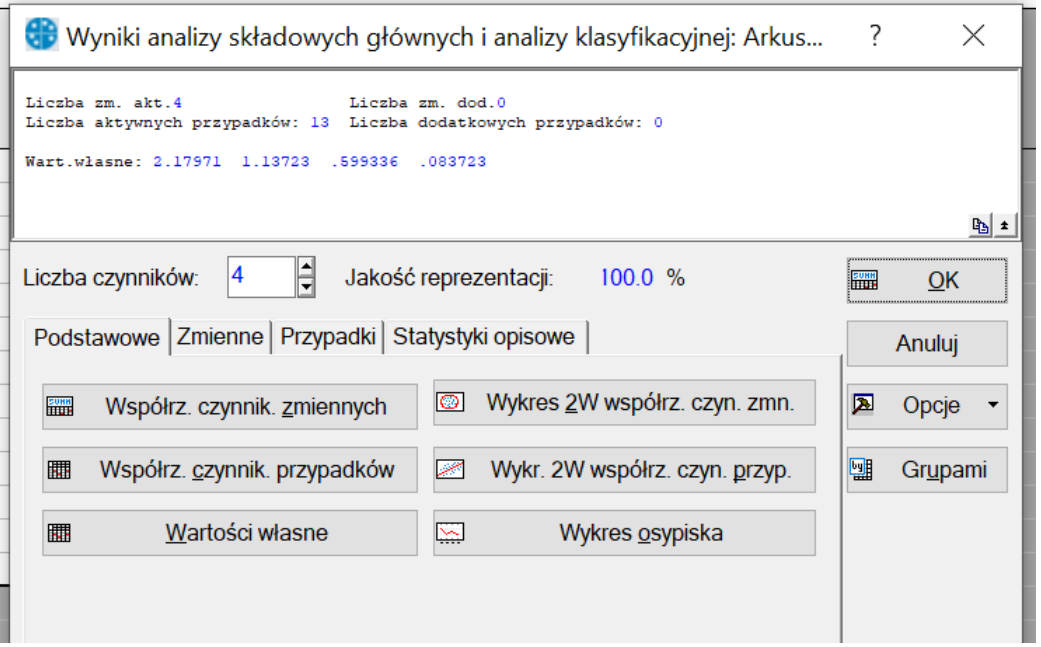
	1 powierzchnia	2 średnia produktywność	3 średni plon	4 średnia masa
Arlet	35.2	0.48	29.3	144
Red Boskoop	34	0.44	24.3	179
Fiesta	36.3	0.51	33.2	148
Gala Must	24.5	0.52	23.6	141
Golden Delicious Reinders	33.5	0.52	29	147
Jonagored	31.2	0.32	20.2	164
Jonica	28.6	0.34	18.2	167
Wilmuta	34.5	0.5	25.8	187
Jonagold	33.1	0.45	26.8	186
Elstar	62	0.3	33	135
Red Elstar	66.2	0.27	27	138
Elshof	60.2	0.3	32	139
Gloster	35.2	0.45	34.2	177



PCA – przykład

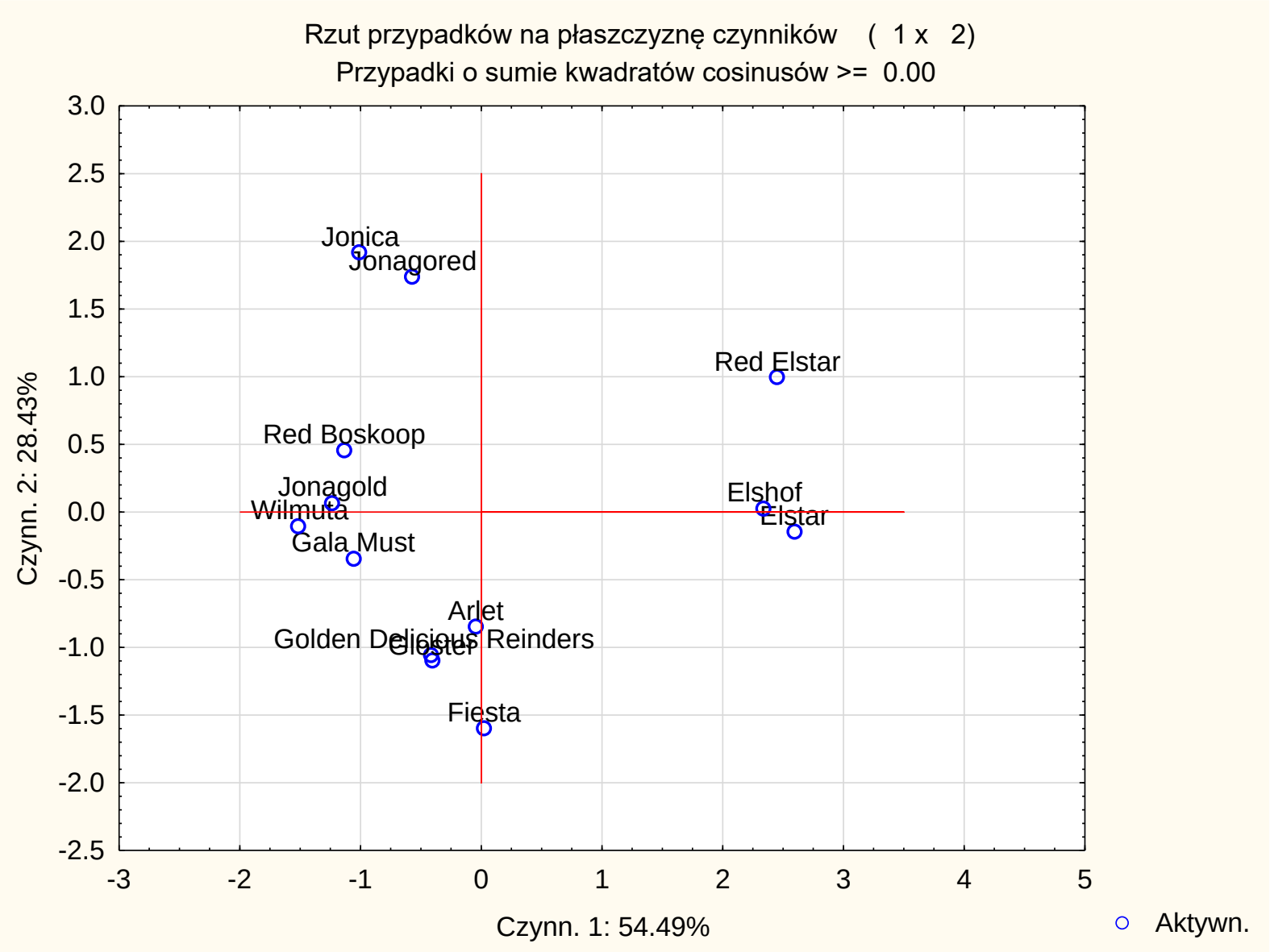
Podział odmian jabłoni podstawie kilku cech

	1 powierzchnia	2 średnia produktywność	3 średni plon	4 średnia masa
Arlet	35.2	0.48	29.3	144
Red Boskoop	34	0.44	24.3	179
Fiesta	36.3	0.51	33.2	148
Gala Must	24.5	0.52	23.6	141
Golden Delicious Reinders	33.5	0.52	29	147
Jonagored	31.2	0.32	20.2	164
Jonica	28.6	0.34	18.2	167
Wilmuta	34.5	0.5	25.8	187
Jonagold	33.1	0.45	26.8	186
Elstar	62	0.3	33	135
Red Elstar	66.2	0.27	27	138
Elshof	60.2	0.3	32	139
Gloster	35.2	0.45	34.2	177



PCA – przykład

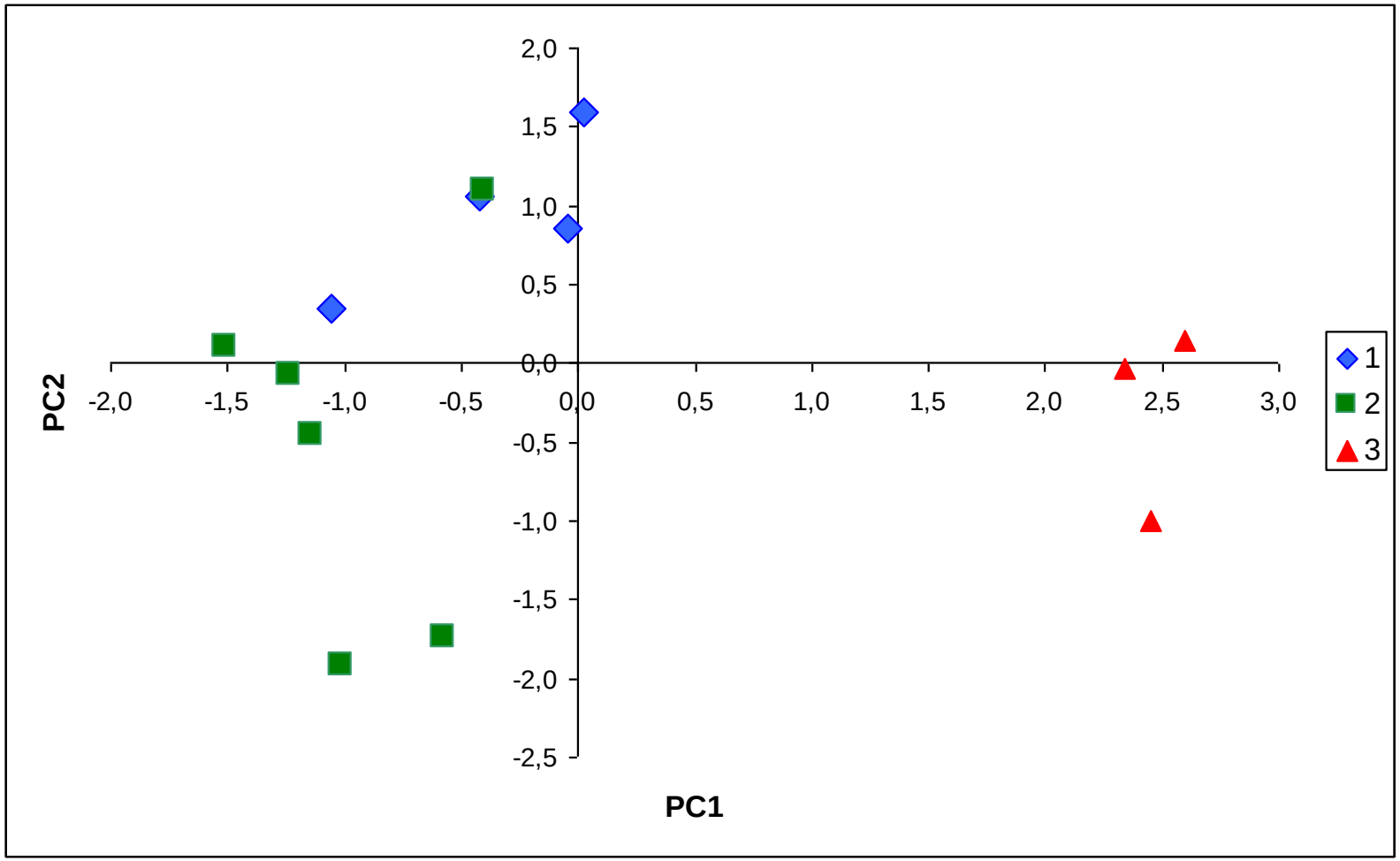
Wykres 2W współrz. czyn. przyp.



PCA – porównanie z analizą skupień

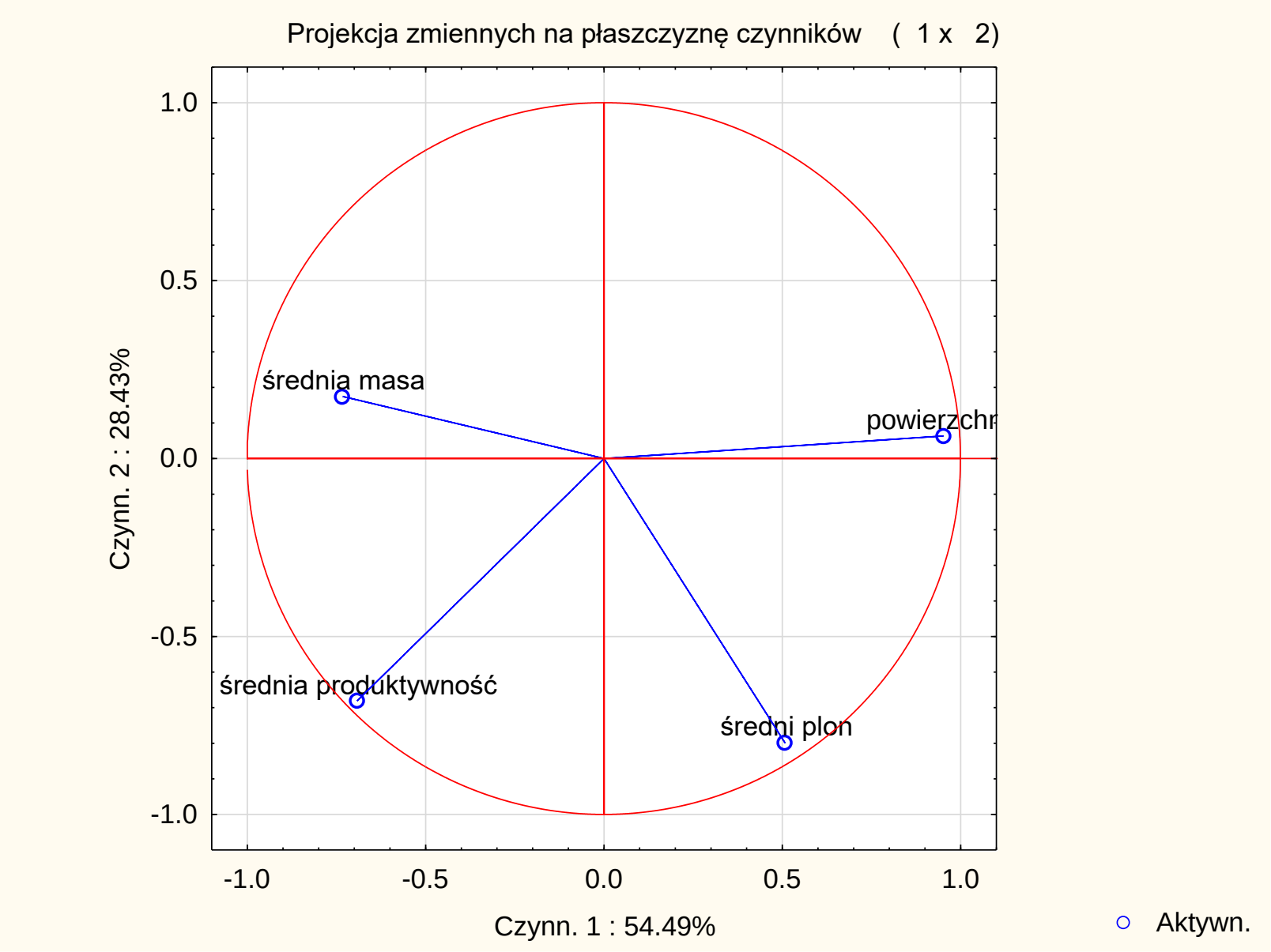
Wartości dwóch pierwszych składowych głównych dla poszczególnych odmian (numery oznaczają grupy odmian wydzielone w analizie skupień).

Analiza składowych głównych ułatwia scharakteryzowanie grup obiektów wydzielonych w analizie skupień, pozwala na graficzne przedstawienie zróżnicowania tych obiektów w układzie dwuwymiarowym.



PCA – przykład

Wykres 2W współrz. czyn. zmn.



PCA - przykład

Wartości pierwszych dwóch składowych (PC1 i PC2) dla poszczególnych odmian

odmiana	**PC1 (54,5%)**	**PC2 (28,4%)**
Arlet	-0,04	0,85
Red Boskoop	-1,14	-0,45
Fiesta	0,02	1,60
Gala Must	-1,06	0,34
Golden Delicious Reinders	-0,42	1,06
Jonagored	-0,57	-1,74
Jonica	-1,01	-1,92
Wilmuta	-1,52	0,10
Jonagold	-1,24	-0,07
Elstar	2,59	0,15
Red Elstar	2,45	-0,99
Elshof	2,34	-0,03
Gloster	-0,40	1,10

Wartości współczynników korelacji badanych cech z poszczególnymi składowymi głównymi

	powierzchnia przekroju pnia (cm^2)	**średnia produktywność drzew (kg/cm2)**	**średni plon (t/ha)**	**średnia masa owocu (g)**
PC1	**0,95**	**-0,69**	**0,51**	**-0,73**
PC2	**-0,06**	**0,68**	**0,80**	**-0,17**

PCA - przykład

Skoroszyt1*
- Analiza czynni
  - Wyniki anal
    - Ładunki c
    - Ładunki c
- Składowe głó
  - Wyniki zaav
    - Projekcja
    - Rzut przy
- Składowe głó
  - Wyniki zaav
    - Rzut przy
    - Projekcja
    - Współrzę
    - Współrzę

Przyp	Współrzędne czynnikowe przypadków, na podst. korelacji			
	Czynn. 1	Czynn. 2	Czynn. 3	Czynn. 4
Arlet	-0.04468	-0.85081	-0.69601	0.054554
Red Boskoop	-1.13969	0.45486	0.50409	-0.215322
Fiesta	0.02324	-1.59670	-0.26239	0.118358
Gala Must	-1.06066	-0.34231	-1.58752	-0.087354
Golden Delicious Reinders	-0.42020	-1.05821	-0.71459	-0.137030
Jonagored	-0.57435	1.73807	-0.25858	0.385122
Jonica	-1.00955	1.91540	-0.38364	0.221004
Wilmuta	-1.51520	-0.10440	0.84528	-0.516508
Jonagold	-1.23941	0.06942	0.96182	-0.056618
Elstar	2.59356	-0.14678	0.17420	0.100000
Red Elstar	2.45280	0.99449	-0.00489	-0.442081
Elshof	2.33806	0.02828	0.23236	0.092329
Gloster	-0.40392	-1.10131	1.18987	0.483546

Skoroszyt1*
- Analiza czynni
  - Wyniki anal
    - Ładunki c
    - Ładunki c
- Składowe głó
  - Wyniki zaav
    - Projekcja
    - Rzut przy
- Składowe głó
  - Wyniki zaav
    - Rzut przy
    - Projekcja
    - Współrze
    - Współrzę

Zmienna	Współrzędne czynnikowe zmiennych, na podst. korelacji (			
	Czynn. 1	Czynn. 2	Czynn. 3	Czynn. 4
powierzchnia	0.950580	0.063465	0.225253	-0.204038
średnia produktywność	-0.693142	-0.682087	-0.164549	-0.165032
średni plon	0.507113	-0.798387	0.302063	0.119048
średnia masa	-0.733825	0.174754	0.655956	-0.026154