Analiza wariancji

Spis treści

Wiadomości ogólne

1. Wprowadzenie teoretyczne

1.1 Powtórzenie najważniejszych wiadomości o testach

Na poprzednich zajęciach zajmowaliśmy się między innymi taki testami istotności, które weryfikowały hipotezy, mówiące o równości średnich pomiędzy dwoma populacjami. Wykorzystywaliśmy testy t Studenta, przyjmując poziom istotności α. Weryfikowane hipotezy zerowe można było przedstawić jako:

zerową: H_0: $\mu = \mu_0$ wobec alternatywnej: H_1: $\mu \neq \mu_0$ lub H_1: $\mu > \mu_0$ lub H_1: $\mu < \mu_0$

Dla przypomnienia: wyznaczoną w programie w wyniku przeprowadzenia testu wartość p porównywaliśmy z ustalonym poziomem istotności α:

- jeżeli $p \leq \alpha \Rightarrow$ odrzucaliśmy H_0 przyjmując H_1,
- jeżeli $p > \alpha \Rightarrow$ nie było podstaw do odrzucenia H_0.

Weryfikacja hipotez powinna być prowadzona tak, aby zapewnić jak najmniejsze prawdopodobieństwo popełnienia pomyłki. Przy podejmowaniu decyzji o odrzuceniu lub nie odrzuceniu hipotezy zerowej istnieje możliwość popełnienia 2 rodzajów błędów:

Błąd I rodzaju – odrzucenie hipotezy zerowej, pomimo że jest prawdziwa. Prawdopodobieństwo popełnienia błędu I rodzaju to właśnie poziom istotności α. Czyli weryfikując hipotezę zerową na poziomie istotności α, zakładamy, że prawdopodobieństwo tego, że odrzucimy prawdziwą hipotezę zerową jest równe α.

Błąd II rodzaju – przyjęcie hipotezy zerowej, która w rzeczywistości jest fałszywa, oznaczane jako β.

Hipoteza zerowa	Podjęta decyzja	
	Przyjęcie H_0	Odrzucenie H_0
H_0prawdziwa	Decyzja prawidłowa	**Błąd I rodzaju**
H_0fałszywa	**Błąd II rodzaju**	Decyzja prawidłowa

Jak jednak należałoby postąpić, gdybyśmy mieli do porównania wartości średnie dla większej liczby grup (większej niż 2)?

Na pierwszy rzut oka, wydaje się, że wystarczyłoby dla każdej pary przeprowadzić test t-Studenta porównania średnich. Przy poziomie istotności $\alpha = 0{,}05$ prawdopodobieństwo, że się nie pomylimy dla jednego porównania wynosi $1 - \alpha = 0{,}95$, dla dwóch $0{,}95^2 = 0{,}9025$, natomiast dla pięciu porównań $0{,}95^5 = 0{,}7738$. Z tego wynika, że przy 5 porównaniach prawdopodobieństwo tego, że się pomylimy przynajmniej 1 raz wynosi: $1 - 0{,}7738 = 0{,}2262$, a to już bardzo dużo (na 100 testów, popełniamy błąd I rodzaju przynajmniej 22 razy).

Do analizy problemu takiego typu wykorzystujemy właśnie analizę wariancji.

1.2Analiza wariancji

Analiza wariancji to zespół metod statystycznych służących do porównywania wartości średnich w trzech lub więcej populacjach. W literaturze rozpowszechnione jest określenie ANOVA (ang. *ANalysis Of VAriance*). W najprostszej formie analiza wariancji to test statystyczny, którego wynik mówi o tym, czy wartości średnie w rozpatrywanych populacjach są sobie równe.

Badany może być wpływ pojedynczego czynnika klasyfikującego, mamy wtedy do czynienia z analizą wariancji dla klasyfikacji pojedynczej, lub z uwzględnieniem wielu czynników, wtedy mowa o analizie wieloczynnikowej.

Podstawowe założenia analizy wariancji:

1) analizowana zmienna jest zmienną ilościową;
2) każda z *k* niezależnych populacji ma rozkład normalny $N(\mu_i, \sigma_i)$, gdzie $i = 1, 2, \dots\ k$;
3) rozkłady te mają równe wariancje (założenie jednorodności wariancji):

$$\sigma_1^2 = \sigma_2^2 = \ \dots\ = \sigma_k^2 = \sigma^2.$$

Z każdej z tych populacji wylosowano próbę o liczebności n_i elementów. Biorąc pod uwagę wszystkie próby mamy łącznie $n = \sum_{i=1}^{k} n_i$ niezależnych obserwacji.

Gdy powyższe założenia są spełnione, można przystąpić do weryfikacji hipotezy zerowej mówiącej o równości średnich we wszystkich *k* grupach:

H_0: $\mu_1 = \ \mu_2 = \ \dots\ = \mu_k$ wobec alternatywnej:

H_1: co najmniej dwie średnie nie są sobie równe.

Do przeprowadzenia testu przyjmujemy poziom istotności α. Wprowadza się następujące oznaczenia:

x_{ij} *j*-ta wartość *i*-tej próby,

$\bar{x}$ średnia arytmetyczna wszystkich wartości,

$\bar{x}_i$ średnia arytmetyczna wartości w *i*-tej próbie.

SS Całkowita (ang. *Sum of Squares*) – tzw. całkowita suma kwadratów:

$$SS_{całkowita} = \sum_{i=1}^{k}\sum_{j=1}^{n_i}(x_{ij} - \bar{x})^2 \qquad (1)$$

SS Błąd – tzw. wewnętrzna suma kwadratów:

$$SS_{błąd} = \sum_{i=1}^{k}\sum_{j=1}^{n_i}(x_{ij} - \bar{x}_i)^2 \qquad (2)$$

SS Efekt – tzw. międzygrupowa suma kwadratów:

$$SS_{efekt} = \sum_{i=1}^{k}\sum_{j=1}^{n_i}(\bar{x}_i - \bar{x})^2 \quad (3)$$

Można wykazać, że:

$$SS_{całkowita} = SS_{błąd} + SS_{efekt}$$

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

Tym samym okazuje się, że całkowite zróżnicowanie można rozłożyć na dwa składniki, z których pierwszy jest miarą zmienności wewnątrz grup, a drugi miarą zmienności między grupami. Powyższa zależność jest podstawą metody ANOVA.

Liczba stopni swobody **df** (ang. *Degrees of Freedom*) może być rozłożona w podobny sposób, czyli:

$$df_{całkowita} = df_{błąd} + df_{efekt}$$

$$n - 1 = (n - k) + (k - 1)$$

Średnie kwadraty odchyleń MS (ang. *MeanSquares*) są obliczane następująco:

$$MS_{błąd} = \frac{SS_{błąd}}{n - k} = \frac{\sum_{i=1}^{k}\sum_{j=1}^{n_i}\left(x_{ij} - \bar{x}_i\right)^2}{n - k}$$

$$MS_{efekt} = \frac{SS_{efekt}}{k - 1} = \frac{\sum_{i=1}^{k}\sum_{j=1}^{n_i}(\bar{x}_i - \bar{x})^2}{k - 1}$$

Jeśli wartości średnie we wszystkich *k* populacjach są sobie równe (hipoteza zerowa jest prawdziwa), to:

1) średni kwadrat odchyleń MS_{efekt} (średni kwadrat odchyleń w grupach) jest nieobciążonym estymatorem wariancji;
2) średni kwadrat odchyleń $MS_{błąd}$ (średni kwadrat odchyleń między grupami) jest również nieobciążonym estymatorem wariancji.

Jeśli hipoteza zerowa nie jest prawdziwa, to średni kwadrat odchyleń między grupami $MS_{błąd}$ jest większy od średniego kwadratu odchyleń wewnątrz grup MS_{efekt}.

Statystyka F (Fischera-Snedecora) o $k - 1$ i $n - k$ stopniach swobody na przyjętym poziomie istotności służy do testowania hipotezy zerowej o równości wartości przeciętnych.

Statystykę F dla tego testu oblicza się ze wzoru:

$$F_{obl} = \frac{MS_{efekt}}{MS_{błąd}}$$

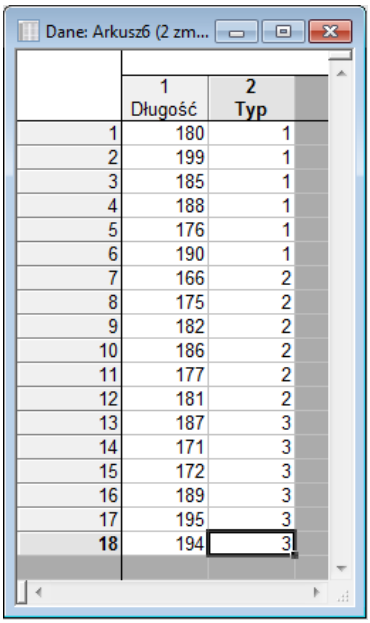
Wartości F bliskie jedności nie pozwalają nam na odrzucenie hipotezy zerowej, natomiast wartości F większe od 1 przemawiają za jej odrzuceniem. Wszystko zależy od przyjętego poziomu istotności, hipotezę zerową odrzuca się jeśli $F_{obl} \geq F_{n-k}^{k-1}$

Wartość F_{n-k}^{k-1} odczytuje się z tablic rozkładu F dla danego poziomu istotności.

2. Analiza wariancji w STATISTICE

2.1 Organizacja danych

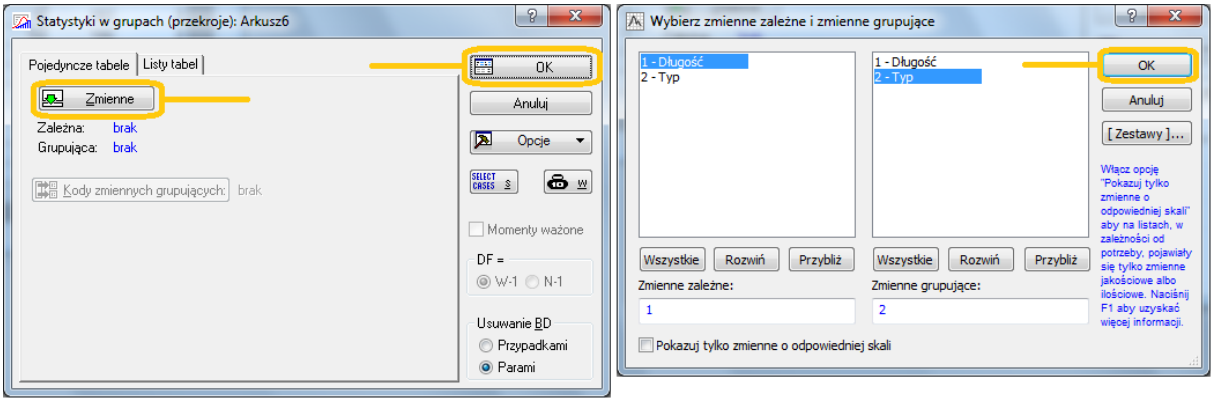
Do przeprowadzenia analizy ANOVA w klasyfikacji pojedynczej konieczne jest wprowadzenie danych w następującym formacie:



	1 Długość	2 Typ
1	180	1
2	199	1
3	185	1
4	188	1
5	176	1
6	190	1
7	166	2
8	175	2
9	182	2
10	186	2
11	177	2
12	181	2
13	187	3
14	171	3
15	172	3
16	189	3
17	195	3
18	194	3

Rys.1 Organizacja danych w arkuszu.

Pierwsza zmienna przyjmuje wartości próby losowej (zmienna zależna), druga określa przynależność do danej klasy (zmienna niezależna, grupująca). Aby rozpocząć analizę wybieramy z menu głównego ***Statystyka/Statystyki podstawowe i tabele/Przekroje prosta ANOVA***. Określenie zmiennych zależnych i grupujących odbywa się w sposób intuicyjny:

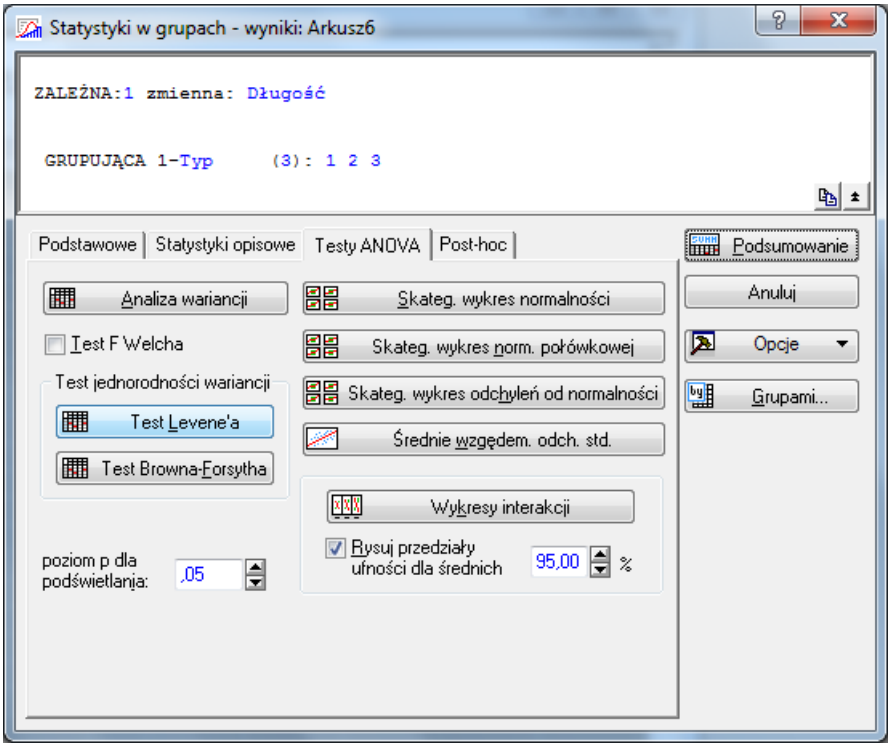


Rys.2 Okna po wejściu w Przekroje proste ANOVA.

Po kliknięciu OK, wyświetla się okno służące do analizy wariancji. Pierwszym krokiem powinno być sprawdzenie założenia o jednorodności wariancji.

2.3 Test jednorodności wariancji

Do sprawdzenia jednorodności wariancji można wykorzystać na przykład test Levene'a dostępny z menu ***Statystyka/Statystyki podstawowe i tabele/Przekroje prosta ANOVA*** w zakładce testy ANOVA:

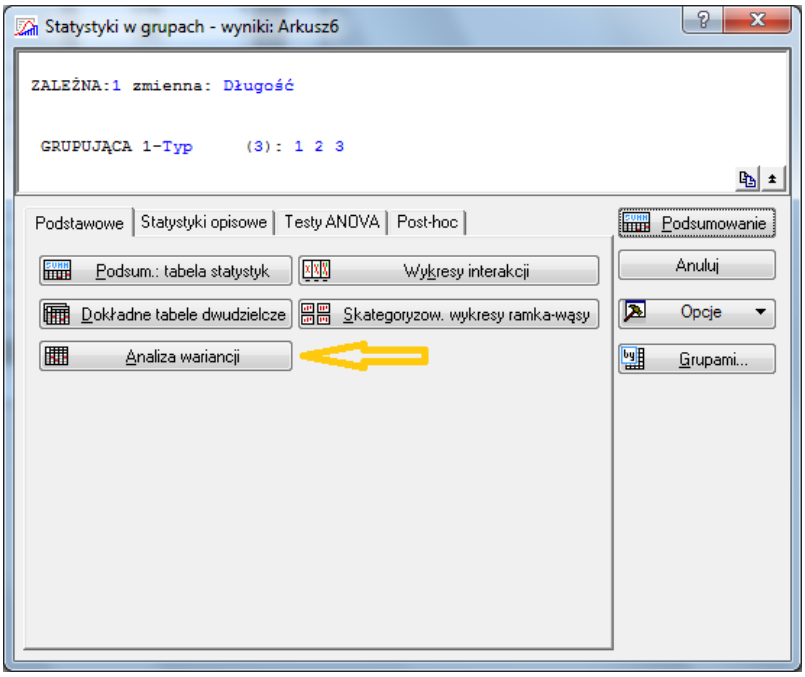


Rys. 3 Przycisk Test Levene'a do zweryfikowania hipotezy o jednorodności wariancji.

Hipoteza zerowa w teście Levene'a zakłada, że wariancje w różnych grupach są jednorodne (takie same). Jeśli test Levene'a okaże się istotny, wówczas należy odrzucić hipotezę o jednorodności wariancji.

2.3 ANOVA

Analizę wariancji dokonuje się w zakładce Podstawowe.



Rys.4 Okno analizy wariancji.

Po wciśnięciu przycisku w Skoroszycie wyświetla się tabelka z wynikiem testu. Decyzję o odrzuceniu lub nie odrzuceniu H_0podejmujemy porównując wartość p z ustalonym poziomem istotności α:

- jeżeli $p \leq \alpha \Rightarrow$ odrzucamy H_0 przyjmując H_1,
- jeżeli $p > \alpha \Rightarrow$ nie ma podstaw do odrzucenia H_0.

Analiza wariancji (Arkusz6)
Zaznaczone efekty są istotne z p < ,05000

Zmienna	SS Efekt	df Efekt	MS Efekt	SS Błąd	df Błąd	MS Błąd	F	p
Długość	243,4444	2	121,7222	1133,500	15	75,56667	1,610793	0,232428

Rys.5 Rezultaty analizy wariancji.

W tym przypadku $p > \alpha \Rightarrow$ nie ma podstaw do odrzucenia H_0.

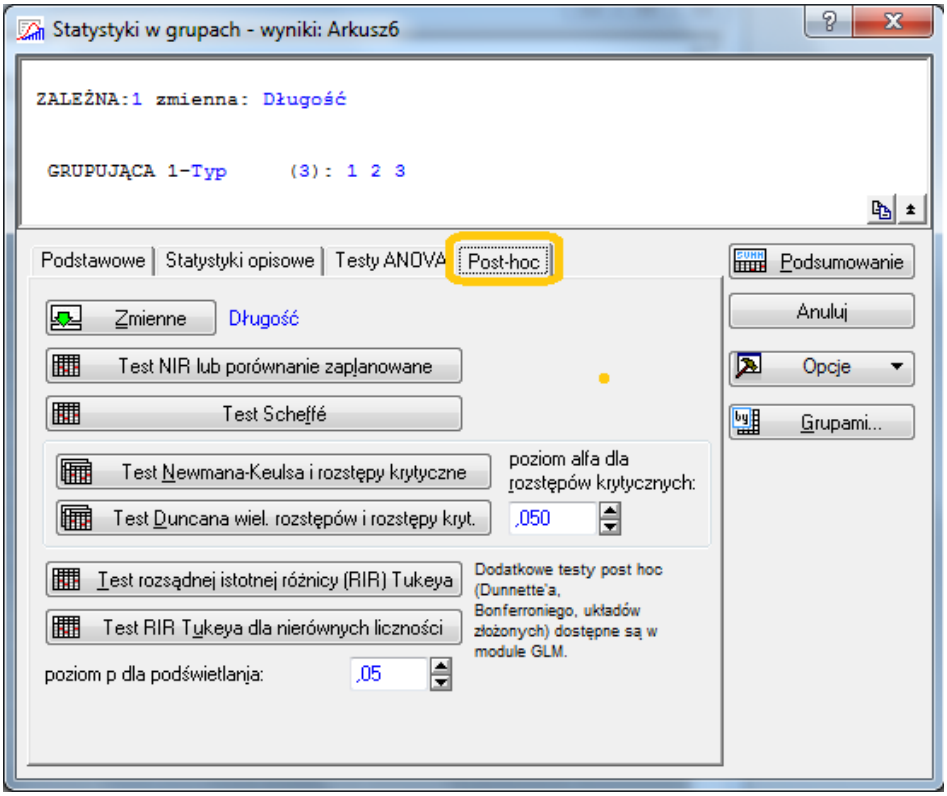
2.2 Testy Post-hoc

Czyli testy „po fakcie" przeprowadzane są w sytuacji, gdy wynik przeprowadzonej analizy metodą ANOVA jest negatywny, tj. nie wszystkie wartości średnie są sobie równe. Umożliwiają one znalezienie średnich, które są znacząco różne od pozostałych. W programie STATISTICA możliwe do przeprowadzenia są m.in. następujące testy:

1. NIR (najmniejszych istotnych różnic);

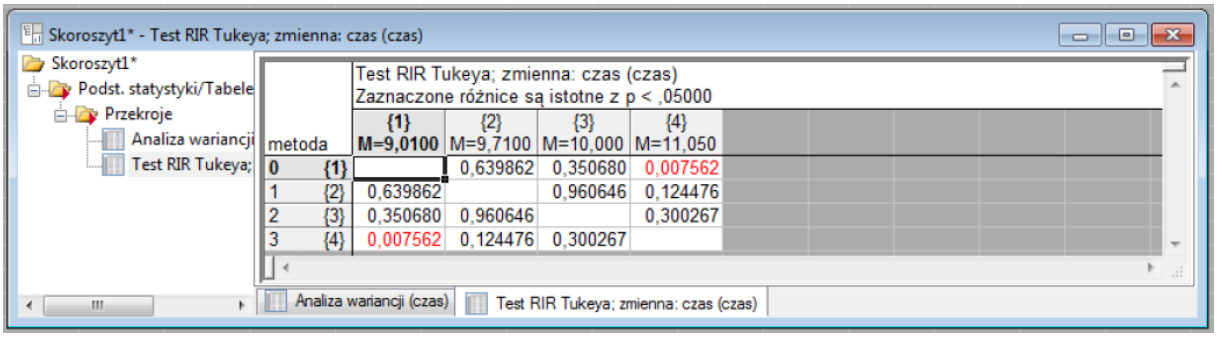
2. Scheffego;
3. Tukeya;
4. Neumana-Keulsa;
5. Duncana.

Testy te dostępne są w zakładce „Post-hoc” w oknie, w którym wykonywana była ANOVA:



Rys.6 Testy Post-hoc.

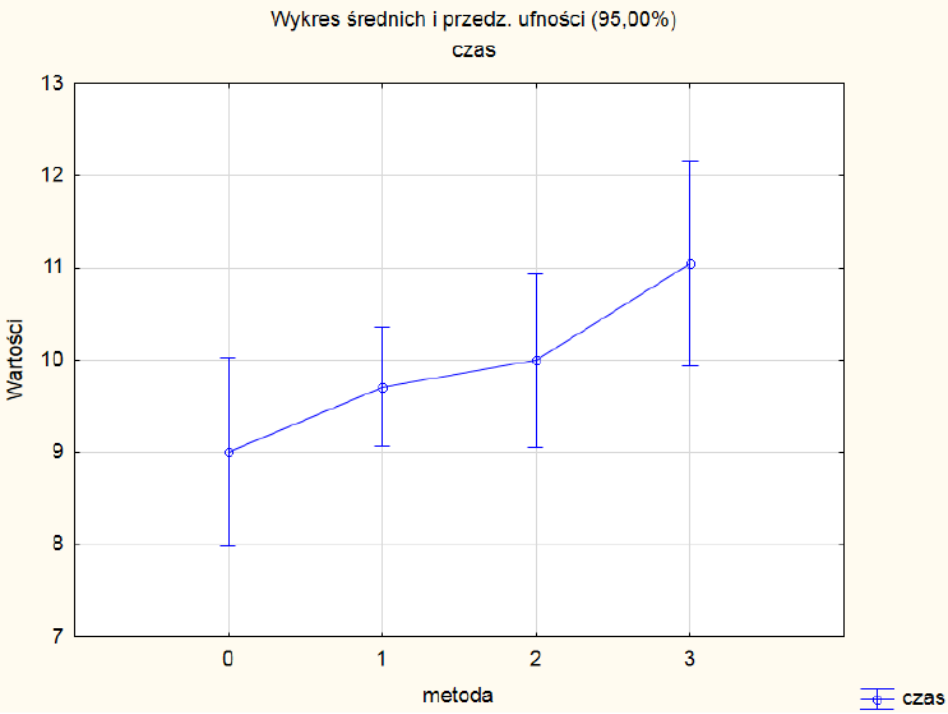
Przykładowy wynik testu Tukeya dla danych, dla których analiza ANOVA wskazała na odrzucenie hipotezy o równości wszystkich średnich przedstawiony jest poniżej:



Rys.7 Wyniki testu Tukeya.

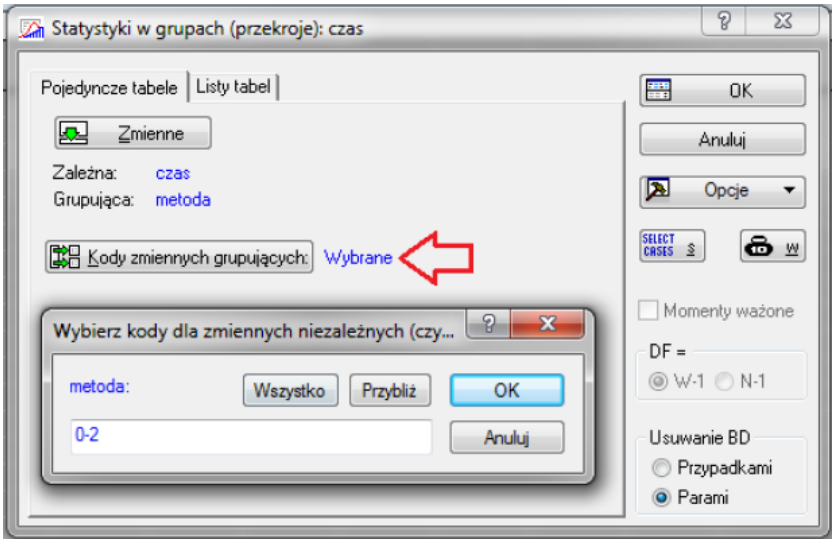
Wyniki wskazują na to, że średnie zbiorów 1 i 4 znacząco się różnią (na przyjęty poziomie istotności $\alpha = 0{,}05$), co ilustruje też poniższy wykres interakcji (przycisk ***Wykres interakcji*** dostępny z okna analizy ANOVA):

Instrukcja do laboratorium



Rys.8 Wykres interakcji.

Analiza ANOVA przeprowadzona ponownie dla tych samych danych z wykluczeniem zbioru {4} nie daje podstaw do odrzucenia hipotezy zerowej. Wykluczenie zbiorów z analizy odbywa się w prosty sposób poprzez użycie kodów zmiennych grupujących:



Rys.9 Wybór kodów zmiennych grupujących.

Ćwiczenie

Część I

Przeprowadź analizę wariancji krok po kroku, czytając poniższe instrukcje, wykorzystując ***Kalkulator prawdopodobieństwa*** (dostępny z menu ***Statystyka/Kalkulator prawdopodobieństwa/Rozkłady***). Wiedząc, że spełnione są założenia o normalności rozkładów i jednorodności wariancji, dla danych przedstawionych w poniższej tabeli, zweryfikuj hipotezę zerową na poziomie istotności $\alpha = 0{,}05$:

H_0: $\mu_1 = \mu_2 = \mu_3$ wobec alternatywnej:

H_1: co najmniej dwie średnie nie są sobie równe.

Tab. Średnica komórek hodowlanych (w mikrometrach) w zależności od podawanego czynnika wzrostu dla trzech wybranych próbek.

Czynnik A	Czynnik B	Czynnik C
30	16	37
49	25	21
35	32	22
38	36	39
26	27	45
40	31	44

a) Uzupełnij tabelę:

$\bar{x}_1$	$\bar{x}_2$	$\bar{x}_3$	$\bar{x}$	n	k	$df_{błąd}$	df_{efekt}

b) Po podstawieniu wyliczonych w podpunkcie a) parametrów do wzorów (2) i (3), uzyskano następujące wartości wewnętrznej sumy kwadratów: $SS_{błąd} = 1133{,}5$ oraz międzygrupowej sumy kwadratów: $SS_{efekt} = 40{,}6$. Wykorzystując te wartości wylicz średni kwadrat odchyleń w grupachMS_{efekt} oraz średni kwadrat odchyleń między grupami $MS_{błąd}$.

MS_{efekt}	$MS_{błąd}$

c) Oblicz wartość statystyki testowej F_{obl}.

d) Do wyznaczenia wartości krytycznej F_{n-k}^{k-1} dla danego poziomu istotności wykorzystaj ***Kalkulator prawdopodobieństwa***. Wybierz rozkład F, wprowadź odpowiednie wartości dla liczby stopni swobody $df\ 1 = k - 1$oraz $df\ 2 = n - k$. W miejsce p wpisz wartość poziomu istotności. Aby wyświetlił się odpowiedni obszar krytyczny zaznacz opcję (1-p).

e) Porównując wartość F_{obl} z wartością krytyczną F_{n-k}^{k-1} podejmij decyzję o odrzuceniu bądź nie odrzuceniu hipotezy zerowej.

f) Wykonaj skategoryzowany wykres ramka wąsy, przedstawiający średnie wartości dla każdej z grup, uwzględniający odchylenie standardowe i 1,96*odchylenie standardowe.
Wskazówka: w zależności od organizacji danych arkuszu skategoryzowany wykres można zrobić na 2 sposoby:

*1) w menu **Statystyka / Statystyki podstawowe i tabele / Statystyki opisowe** – zakładki **Opcje** i **W.skategoryzowane**), jeśli dane zorganizowane są w 2 zmiennych: zmiennej zależnej i grupującej;*
*2) w menu **Statystyka / Statystyki podstawowe i tabele / Statystyki opisowe** – zakładki **Opcje** i **Podstawowe**, jeśli pomiary dla każdej próby znajdują się w osobnej zmiennej.*

Część II

Dane do analizy znajdują się w pliku dane4.sta. Zawierają informacje o średnicy komórek hodowanych w obecności różnych czynników stymulujących wzrost. Czynniki oznaczono literami od A do J. Wykorzystując analizę ANOVA, sprawdź, czy można przyjąć, że komórki pochodzące ze wszystkich próbek mają taką samą średnicę? Przed przystąpieniem do analizy, sprawdź, czy spełnione są p**odstawowe założenia analizy wariancji:**

1) analizowana zmienna jest zmienną ilościową;
2) każda z *k* niezależnych populacji ma rozkład normalny $N(\mu_i, \sigma_i)$, gdzie $i = 1, 2, \dots\ k$;
3) rozkłady te mają równe wariancje (założenie jednorodności wariancji):

$$\sigma_1^2 = \sigma_2^2 = \ \dots\ = \sigma_k^2 = \sigma^2.$$

Do sprawdzenia normalności rozkładu wykorzystaj test Shapiro-Wilka, a do sprawdzenia jednorodności wariancji test Levena.

Wskazówka:.Przy wyborze zmiennych do testowania normalności można skorzystać z przycisku 

(dostępny w oknie Statystyki opisowe), wskazując odpowiednią zmienną grupującą. Wtedy jako rezultat wykonania testu, wyświetlą się wyniki dla całej zaznaczonej listy zmiennych jednocześnie. Jeśli któraś z próbek nie spełni któregoś z założeń, nie należy jej brać pod uwagę przy analizie wariancji.

1. Jeśli założenia są spełnione przejdź do analizy wariancji. Sformułuj hipotezę zerową oraz alternatywną.
2. Wykorzystując funkcje STATISTIKI opisane w rozdz. 2, zweryfikuj postawioną hipotezę zerową na poziomie istotności $\alpha = 0{,}05$.Wynik testu zilustruj wykresem interakcji.
3. Jeśli w wyniku analizy ANOVA otrzymasz negatywny wynik, przeprowadź test Post-hoc (wybierając test Tukeya), aby wskazać grupy, których wartość średnia istotnie różni się od pozostałych. Jeśli takie grupy wystąpią, to je wskaż.
4. Przeprowadź analizę wariancji ponownie, nie uwzględniając tych grup, których średnice istotnie różnią się od pozostałych. Jako ilustrację analizy wykonaj wykres interakcji.