

## ANALIZA WARIANCJI - KLASYFIKACJA JEDNOCZYNNIKOWA

Na poprzednich zajęciach omawialiśmy testy dla weryfikacji hipotez, że dwie populacje o rozkładach normalnych mają jednakowe wartości średnie. Co jednak zrobić, gdy średnich (grup porównywalnych) jest więcej? Możemy przykładowo badać czterema różnymi metodami czas krzepnięcia osocza krwi losowo wybranym pacjentom. Chcemy porównać średnie czasy krzepnięcia dla każdej metody. Wydawałoby się, że wystarczy przeprowadzić test t-Studenta dla każdej pary. Nie możemy tak jednak postąpić. Przy poziomie istotności 0,05 prawdopodobieństwo, że się nie pomylimy dla jednego porównania wynosi 0,95, a dla dwóch niezależnych porównań  $0,95^2 = 0,905$ . Dla sześciu niezależnych porównań wartość ta wynosi już  $0,95^6 = 0,7351$ . Prawdopodobieństwo, że pomylimy się co najmniej raz wynosi teraz  $1 - 0,7351 = 0,265$  a na tak duży błąd pierwszego rodzaju zgodzić się nie możemy. Do analizy takich problemów wykorzystujemy zespół metod statystycznych zwanych analizą wariancji.

Powstanie i rozwój analizy wariancji (ANOVA) zawdzięczamy angielskiemu biologowi R. A. Fisherowi (1925r.), który rozwinął ją dla rozwiązania problemów doświadczeń rolniczego. Nazwa analizy wzięła się stąd, że ocena istotności różnic między średnimi odbywa się przez porównanie (tj. analizę) wariancji. Dokonujemy tego poprzez podział całkowitej zmienności (czyli sumy kwadratów odchyleń wszystkich pomiarów od średniej) na różne źródła związane z efektami występującymi w badaniu.

Analiza wariancji (ANOVA) jest zespołem metod statystycznych wykorzystywanych do porównywania kilku populacji. Ogólnie mówiąc, jest to technika badania wyników (doświadczeń, obserwacji), które zależą od jednego lub kilku **czynników** działających równocześnie. Czynniki takimi mogą być leki, metody leczenia, diety, płeć itd. Czynniki te nazywane bywają też czynnikami grupującymi, klasyfikacyjnymi lub zmiennymi manipulacyjnymi. Najprostsza i zarazem najbardziej popularna to analiza wariancji dla klasyfikacji pojedynczej. Bada się w niej wpływ tylko jednego czynnika klasyfikującego (kontrolowanego na wielu poziomach) na wyniki przeprowadzanego badania..

Założmy, że czynnik grupujący ma k poziomów. Wówczas analiza wariancji weryfikuje hipotezę, że średnie w grupach są jednakowe:

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

wobec hipotezy alternatywnej:

**$H_1$  : co najmniej dwie średnie różnią się między sobą**

Podstawą tej analizy jest możliwość rozbicia sumy kwadratów wariancji całkowitej dla wszystkich wyników obserwacji na dwa składniki:

- sumę kwadratów opisującą zmienność wewnątrz prób.
- sumę kwadratów opisującą zmienność między grupami (populacjami).

Wówczas omawiane rozbicie przyjmuje postać:

Całkowita suma kwadratów (SS Całkowita) = wewnątrzgrupowa suma kwadratów (SS Błędu) + międzygrupowa suma kwadratów (SS Efektu)

Rozłożyliśmy więc całkowite zróżnicowanie na dwa składniki. Pierwszy jest miarą zmienności wewnątrz grup, a drugi miarą różnic pomiędzy grupami. Suma kwadratów powstać bowiem może zarówno z odchyleń spowodowanych wpływami przypadkowymi, jak i odchyleń spowodowanych systematycznymi różnicami między poszczególnymi grupami. Stosowany skrót SS pochodzi od angielskiej terminologii Sum of Squares (suma kwadratów). Analogicznie sumują się liczby stopni swobody (ang. df - degrees of freedom), czyli mamy:

$$n - 1 \text{ (df całkowite)} = k - 1 \text{ (df efektu)} + n - k \text{ (df błędu)}$$

Następnym etapem obliczeń są tzw. średnie kwadraty odchyłeń (oznaczane skrótem MS od angielskiego Mean squares). Obliczamy je następująco:

$$\text{ŚK Efektu (MS Effect)} = \text{SS efektu} / \text{df efektu} = \text{SS Efektu} / (k - 1)$$

$$\text{ŚK błędu (MS Error)} = \text{SS Błędu} / \text{df błędu} = \text{SS błędu} / n - k$$

Przy założeniu prawdziwości hipotezy zerowej zarówno średni kwadrat odchyłeń międzygrupowych (ŚK Efektu), jak i średni kwadrat odchyłeń wewnątrzgrupowych (ŚK błędu) są nieobciążonymi estymatorami wariancji. Przy braku istotnej różnicy pomiędzy średnimi (założenie prawdziwości hipotezy zerowej) oszacowanie wariancji powyższymi estymatorami powinno się, więc różnić w granicach losowych odchyłeń. Jeżeli hipoteza zerowa nie jest prawdziwa, to wówczas średni kwadrat między grupami (ŚK Efektu) rośnie i jest wyższy od średniego kwadratu odchyłeń wewnątrzgrupowych. Hipotezę zerową należałoby wtedy odrzucić. Do porównywania ŚK Efektu i ŚK błędu używamy statystyki F. Wartości F bliskie jedności „świadczą za” sprawdzaną hipotezą, natomiast wartości dużo większe od 1 przemawiają za jej odrzuceniem. Wszystko oczywiście zależy od przyjętego przez nas poziomu istotności. Jeżeli hipotezę zerową odrzucamy, to wówczas stwierdzamy, że wyróżniony czynnik wpływa istotnie na zmienną nazywaną zmienna objaśnianą.

Przyjęło się zapisywać wyniki analizy wariancji w postaci następującej tabeli:

Źródło zmienności	Suma kwadratów	Stopnie swobody	Średni kwadrat	Test F
Między grupami	SK Efektu	k - 1	ŚK Efektu	$F = \text{ŚK Efektu} / \text{ŚK błędu}$
Wewnątrz grup	SK błędu	n - k	ŚK błędu	
Całkowita	SK całkowita	n - 1		

## ANALIZA WARIANCJI W PAKIECIE STATISTICA

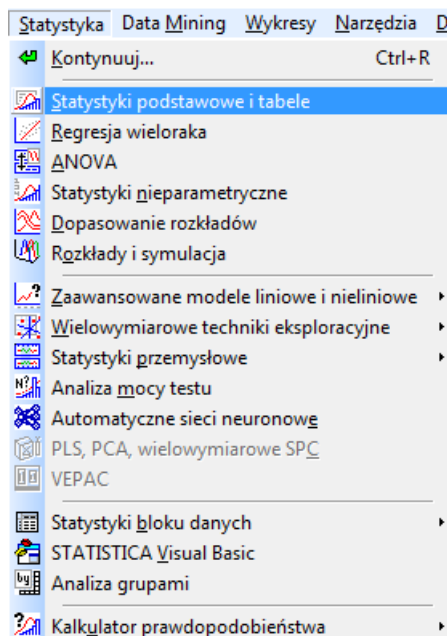
W praktyce przeprowadzamy analizę wariancji w oparciu o programy komputerowe. Pokażemy jak w tym celu wykorzystać pakiet *STATISTICA*. Korzystając z pakietów statystycznych musimy odpowiednio przygotować dane. Musimy je mieć wpisane w postaci dwóch zmiennych (kolumn). Pierwsza kolumna zawiera wyniki pomiarów, druga kolumna zawiera kody (wartości zmiennej grupującej), określające poziomy czynników (grupy), które chcemy porównać.

	Czas krzepliwości	
	1 CZAS_K	2 METODA
1	9,10	I
2	8,90	I
3	10,90	IV
4	10,40	IV
5	10,00	IV
6	9,20	I
7	9,40	II
8	8,50	II
9	10,00	III
10	9,90	III

Rys 1 Fragment przykładowego arkusza danych

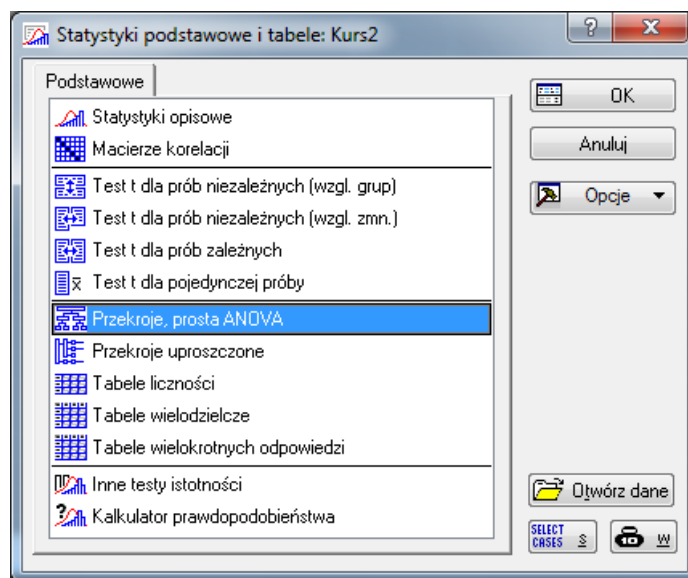
Przykładowo założmy, że chcemy porównać przeciętny czas krzepliwości badany czterema metodami. Fragment omawianych danych dla tego przykładu pokazany jest na rysunku 1.

Porównania dokonamy przeprowadzając, dla przykładowych danych, pełną analizę w pakiecie *STATISTICA*. W programie tym do przeprowadzenia jednoczynnikowej analizy wariancji służy opcja **Przekroje, prosta ANOVA** w module **Statystyki podstawowe i tabele** oraz bardzo rozbudowany moduł **ANOVA**. W celu przeprowadzenia analizy z menu **Statystyka** wybieramy opcję **Statystyki podstawowe i tabele**.



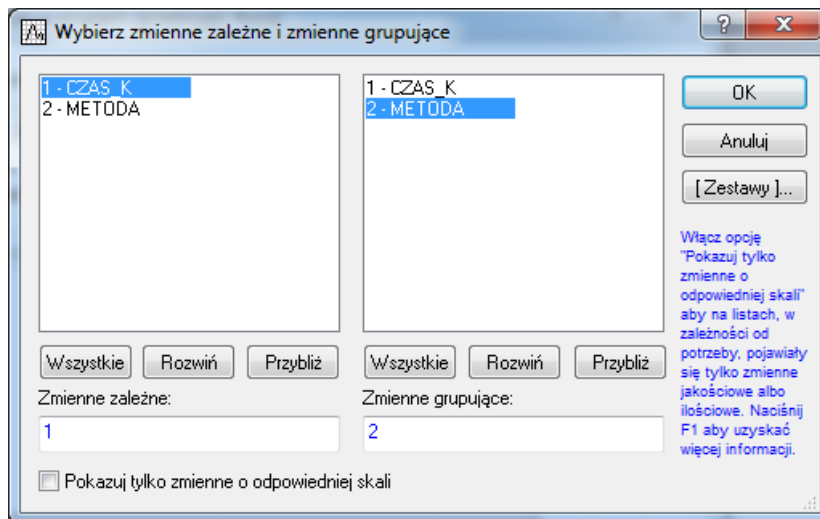
Rys 2 Rozwijalne menu Statystyka

Następnie w otwartym oknie **Statystyki podstawowe i tabele** wybieramy opcję **Przekroje, prosta ANOVA** (rysunek poniżej) a następnie klikamy **OK**.



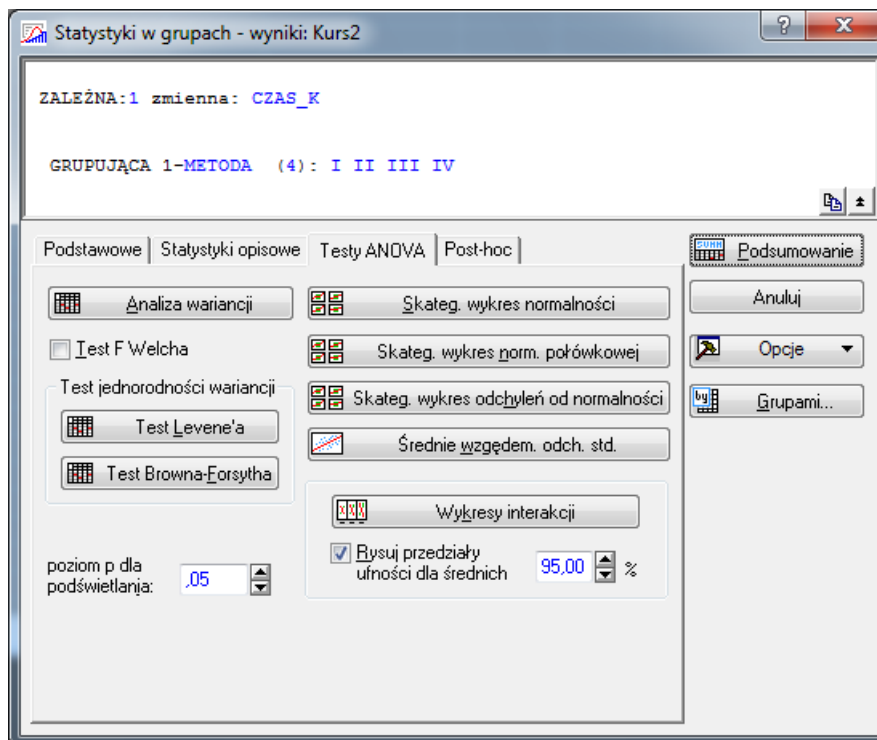
Rys 3 Okno wyboru podstawowych statystyk

W otwierającym się oknie **Statystyki w grupach (przekroje)** klikamy przycisk **Zmienne** dla określenia odpowiednich zmiennych. Jako zmienną zależną wybieramy zmienną **CZAS\_K**, a jako zmienną grupującą zmienną **METODA**. Klikając na przycisk **OK** kończymy wybór zmiennej. Opisana sytuacja pokazana jest na poniższym rysunku



Rys 4 Okno wyboru zmiennych do analizy wariancji

Kolejno klikamy dwa razy przycisk **OK**, aby rozpocząć analizę. W otwartym oknie **Statystyki w grupach** przechodzimy na kartę **Testy ANOVA** zawierająca interesujące nas analizy i opcje.



Rys 5 Okno wyboru opcji dla przeprowadzenia analizy wariancji

Za pomocą opcji **Test Levene'a** możemy zweryfikować założenie o jednorodności wariancji. Natomiast za pomocą opcji **Skateg. wykres normalności** możemy wzrokowo ocenić założenie normalności w każdej grupie.

Opcja **test F Welcha** umożliwia przeprowadzenie analizy wariancji, kiedy założenie o jednorodności wariancji jest naruszone.

Po kliknięciu przycisku **Analiza wariancji** otrzymamy następujący arkusz wyników:

Analiza wariancji (Arkusz1)								
Zaznaczone efekty są istotne z $p < ,05000$								
Zmienna	SS Efekt	df Efekt	MS Efekt	SS Błąd	df Błąd	MS Błąd	F	p
CZAS_K	24,13675	3	8,045583	54,47300	36	1,513139	5,317148	0,003883
[1]	[2]	[3]	[4]	[5]	[6]	[7]	[8]	[9]

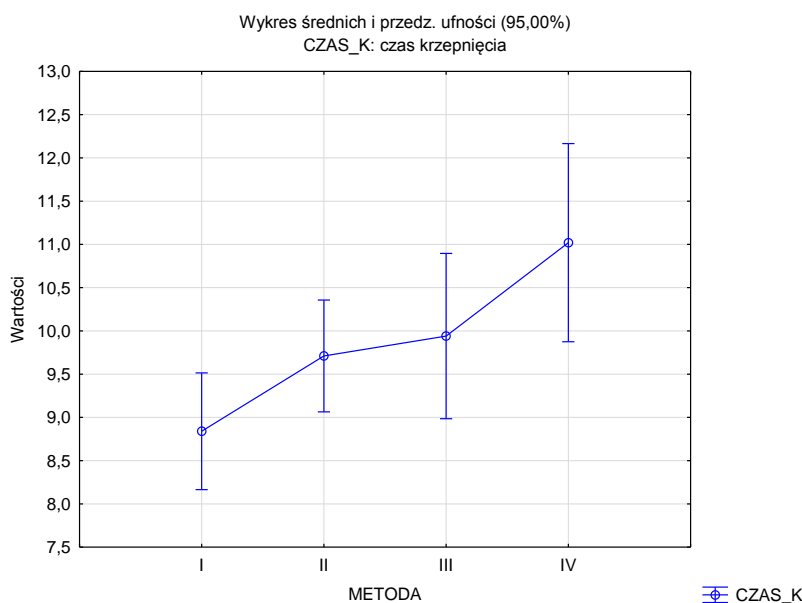
Rys. 6 Arkusz wyników analizy wariancji

W powyższym oknie wyników mamy następujące wartości liczbowe:

- [1] - nazwa zmiennej zależnej
- [2] - suma kwadratów efektu
- [3] - liczba stopni efektu
- [4] – średnia suma kwadratów efektu
- [5] - suma kwadratów błędu
- [6] - liczba stopni swobody błędu
- [7] - średnia suma kwadratów błędu
- [8] - wartość testu F
- [9] - poziom prawdopodobieństwa p

Jak wiemy istotą analizy wariancji jest rozbieżność na addytywne składniki (których liczba wynika z potrzeb eksperymentu). Porównanie poszczególnej wariancji wynikającej z działania danego czynnika oraz tzw. wariancji resztowej, czyli wariancji mierzącej losowy błąd daje odpowiedź, czy dany czynnik odgrywa istotną rolę w kształtowaniu się wyników eksperymentu. Porównując testem F wariancję między grupami z wariancją resztową (wewnątrz grupowa) rozstrzygamy, czy średnie grupowe różnią się istotnie od siebie czy nie. Jeśli podział na grupy np. przebiegał ze względu na różne poziomy badanego czynnika, to możemy w ten sposób wykryć istotny wpływ poziomu na efekt wartości badanej zmiennej. Jak widzimy, wyniki analizy ( $p=0,0039$ ) upoważniają nas do odrzucenia hipotezy zerowej, mówiącej o równości średnich.

Dla otrzymania interpretacji graficznej (rysunek poniżej) klikamy na przycisku **Wykres interakcji**. Otrzymany wykres potwierdza brak równości wszystkich średnich. Na tym wykresie mamy przedstawione punkty prezentujące wartości średnie w poszczególnych podgrupach



Rys. 7 Interpretacja graficzna wyników analizy wariancji

Jeżeli analiza wariancji nie pokaże istotności różnic między rozpatrywanymi średnimi, nie przeprowadza się już dalszych testów. Natomiast, kiedy hipoteza zerowa zostanie odrzucona powstaje pytanie, które z porównywanych populacji są odpowiedzialne za odrzucenie hipotezy zerowej. Chcemy wiedzieć, które z  $k$  średnich różnią się między sobą, a które nie. Musimy wtedy koniecznie przeprowadzić dokładniejsze badania różnic między średnimi z poszczególnych grup. Wykorzystujemy do tego celu specjalne testy post-hoc zwane też testami wielokrotnych porównań.

Dla danych z naszego przykładu okazało się, że nie wszystkie średnie są równe. Chcemy wiedzieć, które z analizowanych średnich różnią się istotnie między sobą. Do porównań szczegółowych wykorzystujemy, więc testy post-hoc.

Dla przeprowadzenia testów po fakcie przechodzimy na kartę **Post – hoc** i wybieramy **Test rozsądnej istotnej różnicy (RIR) Tukeya**. Otrzymany arkusz wyników pokazuje poniższy rysunek.

Test RIR Tukeya; zmienna: CZAS_K (Arkusz)				
Zaznaczone różnice są istotne z $p < ,05000$				
METODA	{1}	{2}	{3}	{4}
	M=8,8400	M=9,7100	M=9,9400	M=11,020
I {1}		0,401733	0,207153	0,001942
II {2}	0,401733		0,975086	0,099073
III {3}	0,207153	0,975086		0,220847
IV {4}	0,001942	0,099073	0,220847	

Rys. 8 Arkusz wyników testu Tukeya

Otrzymujemy następujące wyniki:

- Średnia dla metody I różni się istotnie od średniej dla skali IV ( $p = 0,0019$ )
- Średnie dla pozostałych metod nie różnią się istotnie (wartości  $p$  są większe od 0,05)

Prezentację metod analizy wariancji kończymy przedstawieniem podstawowych jej założeń:

1. **Analizowana zmienna zależna jest mierzalna**
2. **Rozważane niezależne populacje mają rozkłady normalne  $N(\mu_i, \sigma_i)$   $i = 1, 2, \dots, k$**
3. **Rozkłady te mają jednakową wariancję  $\sigma_1^2 = \sigma_2^2 = \dots = \sigma_k^2 = \sigma^2$**

Założenia te musimy sprawdzić przed rozpoczęciem analizy wariancji. Normalność rozkładu w każdej grupie sprawdzamy za pomocą testu Shapiro Wilka. Natomiast jednorodność wariancji sprawdzamy za pomocą testu Levene'a lub Browna – Forsytha. Oba znajdują się w oknie, gdzie przeprowadzamy analizę wariancji.

Jeżeli nie mamy spełnionego założenia o jednorodności wariancji stosujemy odpowiednią poprawkę **test F Welcha**. Otrzymamy wówczas następujący arkusz wyników:

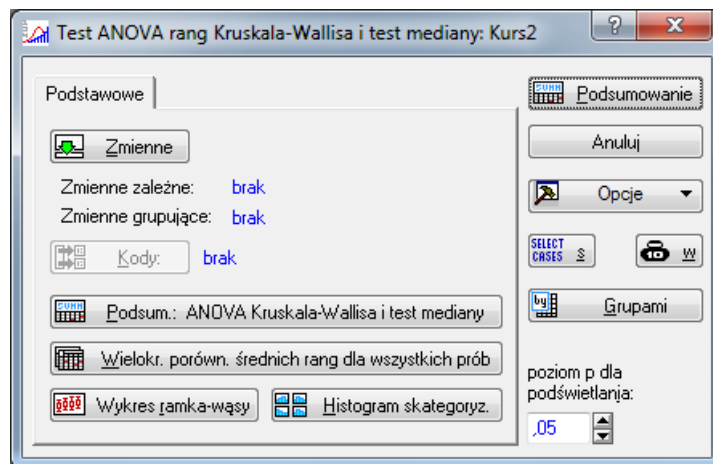
Analiza wariancji (Kurs2)												
Zaznaczone efekty są istotne z $p < ,05000$												
Zmienna	SS	df	MS	SS	df	MS	F	p	df Welch	df Welch	F Welcha	p Welcha
	Efekt	Efekt	Efekt	Błąd	Błąd	Błąd			Efekt	Błąd		
CZAS_K	24,13675	3	8,045583	54,47300	36	1,513139	5,317148	0,003883	3	19,59354	4,729440	0,012124

Rys. 9 Arkusz wyników analizy wariancji z poprawką Welcha

Tym razem interpretujemy cztery ostatnie kolumny (czcionka pogrubiona) związane z opisywaną poprawką.

Natomiast w przypadku braku normalności stosujemy nieparametryczny odpowiednik jednoczynnikowej analizy wariancji test Kruskala – Wallisa. Test ten weryfikujący hipotezę zerową o równości wszystkich median w analizowanych grupach znajdziemy w module testy nieparametryczne. Otrzymamy go wybierając z menu **Statystyka** opcję **Statystyki**

**nieparametryczne.** Następnie w otwierającym się oknie wybieramy opcję **Porównanie wielu prób niezależnych (grup)**. Po kliknięciu na przycisku **OK** otworzy się okno **Test ANOVA rang Kruskala-Wallisa i test mediany**.




Rys. 10 Okno z opcjami dla przeprowadzenia testu Kruskala - Wallisa

Klikając przycisk **Podsum. ANOVA Kruskala-Wallisa i test mediany** otrzymujemy dwa arkusze wyników. Nas interesuje ten dotyczący nieparametrycznej analizy wariancji. Pokazuje go poniższy rysunek.

ANOVA rang Kruskala-Wallisa; CZAS_K (Kurs2)					
Zmienna niezależna (grupująca): TEST					
Test Kruskala-Wallisa: H ( 3, N= 40) =12,46768 p =,0059					
Zależna: CZAS_K	Kod	N ważnych	Suma Rang	Średnia Ranga	
I	1	10	110,0000	11,00000	
II	2	10	204,0000	20,40000	
III	3	10	212,0000	21,20000	
IV	4	10	294,0000	29,40000	

Rys. 11 Arkusz wyników nieparametrycznej analizy wariancji.

Jak widzimy wartość  $p = 0,0059$ . Odrzucamy zatem hipotezę zerową o równości wszystkich median. Zatem istnieją, co najmniej, dwie różne mediany. Dla sprawdzenia, które mediany różnią się istotnie między sobą wykorzystujemy procedurę wielokrotnego porównywania średnich rang, wywołaną przyciskiem . Na ekranie pojawiają się dwa arkusze wynikowe. Nas interesuje ten zawierający wartości  $p$ .

Wartość p dla porównań wielokrotnych (d)				
Zmienna niezależna (grupująca): TEST				
Test Kruskala-Wallisa: H ( 3, N= 40) =12,				
Zależna: CZAS_K	I	II	III	IV
I	R:11,000	R:20,400	R:21,200	R:29,400
I		0,433094	0,306353	<b>0,002595</b>
II	0,433094		1,000000	0,511007
III	0,306353	1,000000		0,700673
IV	<b>0,002595</b>	0,511007	0,700673	

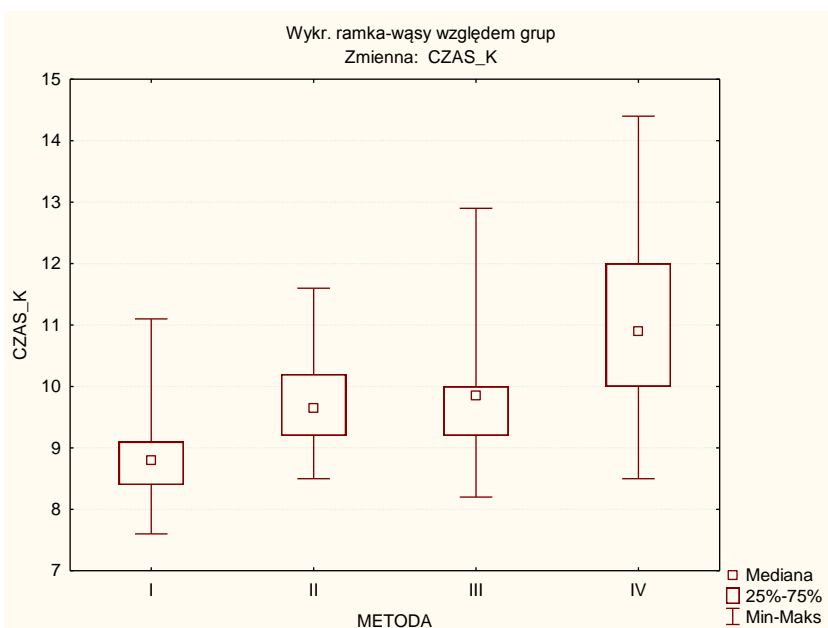
Rys. 12 Arkusz wyników testy post-hoc.

Możemy zatem powiedzieć, że:

- Mediana dla metody I różni się istotnie od średniej dla skali IV ( $p = 0,0026$ )
- Mediany dla pozostałych metod nie różnią się istotnie (wartości  $p$  są większe od 0,05)

Otrzymujemy zatem identyczne wyniki jak dla analizy wariancji.

Dla dopełnienia całości możemy, za pomocą opcji **Wykres ramka-wąsy** lub **Histogram skategoryzowany**, wywołać wykresy, stanowiące interpretację graficzną otrzymanych wyników.



Rys. 13 Wykres ramka wąsy przedstawiający położenie median

W programie *STATISTICA* do przeprowadzenia jednoczynnikowej analizy wariancji służy również bardzo rozbudowany moduł **ANOVA**.

### Dalsze możliwości

Sytuacje, w których pojedyncza zmienna pozwala wyjaśnić dane zjawisko medyczne należą do rzadkości. W eksperymentach medycznych branych jest pod uwagę wiele czynników. Przypuśćmy, że badając trzy leki w leczeniu pacjentów z depresją, rozważamy dodatkowy czynnik grupujący płeć. Dane do takiego dwuczynnikowego eksperymentu zapisujemy w postaci trzech kolumn. Pierwsza kolumna zawiera pomiary depresji pacjentów, a pozostałe dwie kolumny zawierają czynniki grupujące związane z rodzajem podawanego leku i płcią pacjenta. Wówczas różnice między średnimi wynikają zarówno z zastosowania różnych środków farmakologicznych jak i różnicy płci. Przy takim eksperymencie całkowitą zmienność (sumę kwadratów odchyłeń) możemy rozdzielić, na co najmniej trzy składniki:

- zmienność spowodowaną czynnikiem płeć
- zmienność spowodowaną zastosowaniem konkretnego leku
- zmienność spowodowaną błędem (zmienność wewnątrz grupowa)

Zauważmy, że istnieje jeszcze jedno dodatkowe źródło zmienności - **interakcja**, czyli oddziaływanie łączne. Mówi nam ona, w jakim stopniu wpływ jednego czynnika zależy od poziomu drugiego czynnika. Jeżeli pozostaje on taki sam, to nie ma żadnej interakcji. W przeciwnym wypadku między dwoma czynnikami zachodzi interakcja. Przyjmijmy, że jeden lek daje lepsze terapeutyczne wyniki niż pozostałe. Jeżeli tak jest równocześnie dla kobiet i mężczyzn (niezależnie od płci) to nie ma interakcji. Jeżeli jednak jeden lek daje lepsze wyniki dla kobiet a inny dla mężczyzn, to możemy mówić o wystąpieniu interakcji między tymi dwoma czynnikami (rodzajem leku a płcią). Możliwość wykrywania istotnych interakcji i w związku z tym bardziej złożonych hipotez na temat otaczającej nas rzeczywistości czyni z analizy wariancji bardzo uniwersalne narzędzie. Przy zastosowaniu testu t nie otrzymalibyśmy identycznych rezultatów.