StatSoft® Polska

Zestaw do analiz marketingowych i rynkowych 4.0

instrukcja instalacji oraz podstawowe informacje o programie

1. Instrukcja instalacji i dezinstalacji programu

1.1. Instalacja wersji jednostanowiskowej i wersji sieciowej

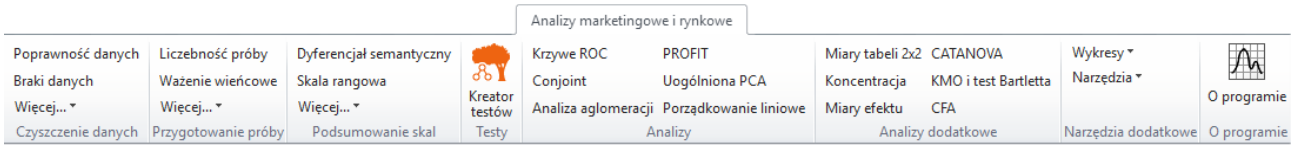
W celu zainstalowania programu należy:

a.) Zainstalować odpowiednią wersję *Statistica* (wymagana jest wersja 13.1 z odpowiednim zestawem modułów analitycznych).

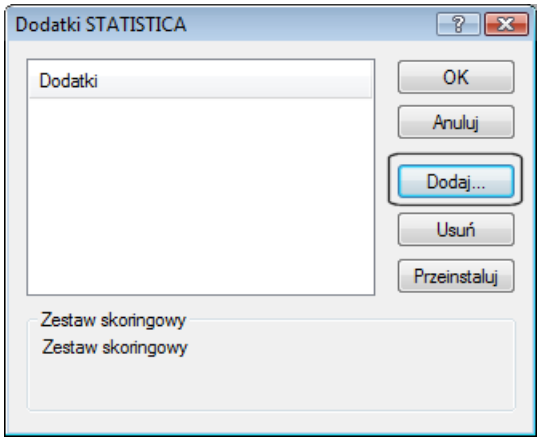
b.) Uruchomić instalator programu *Zestawy Analityczne.exe*, a następnie zatwierdzać kolejne kroki instalacji.

c.) W trakcie instalacji wskazać plik – *License.xml* – zostanie on skopiowany do katalogu z programem (domyślnie C:\Program Files\StatSoft\Zestawy Analityczne).

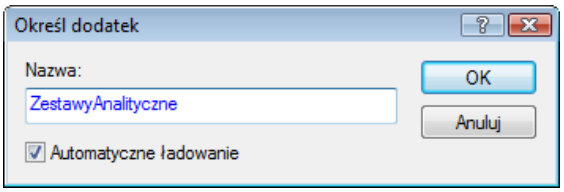
Po zainstalowaniu dodatku utworzone zostanie menu umożliwiające dostęp do modułów *Zestawu do analiz marketingowych i rynkowych*.



W przypadku niepojawienia się menu ***Analizy marketingowe i rynkowe*** po instalacji, należy w menu ***Narzędzia*** | ***Makro*** | ***Dodatki*** otworzyć okno ***Dodatki STATISTICA***.

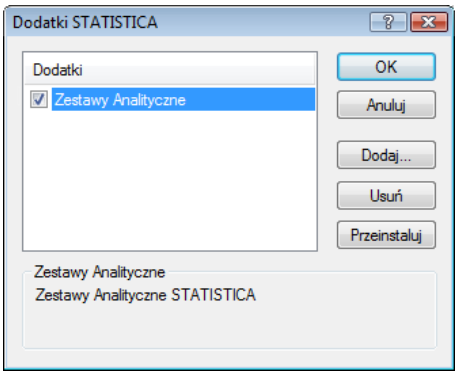


W wyświetlonym oknie kliknąć przycisk ***Dodaj…*** a następnie wprowadzić napis `ZestawyAnalityczne`.



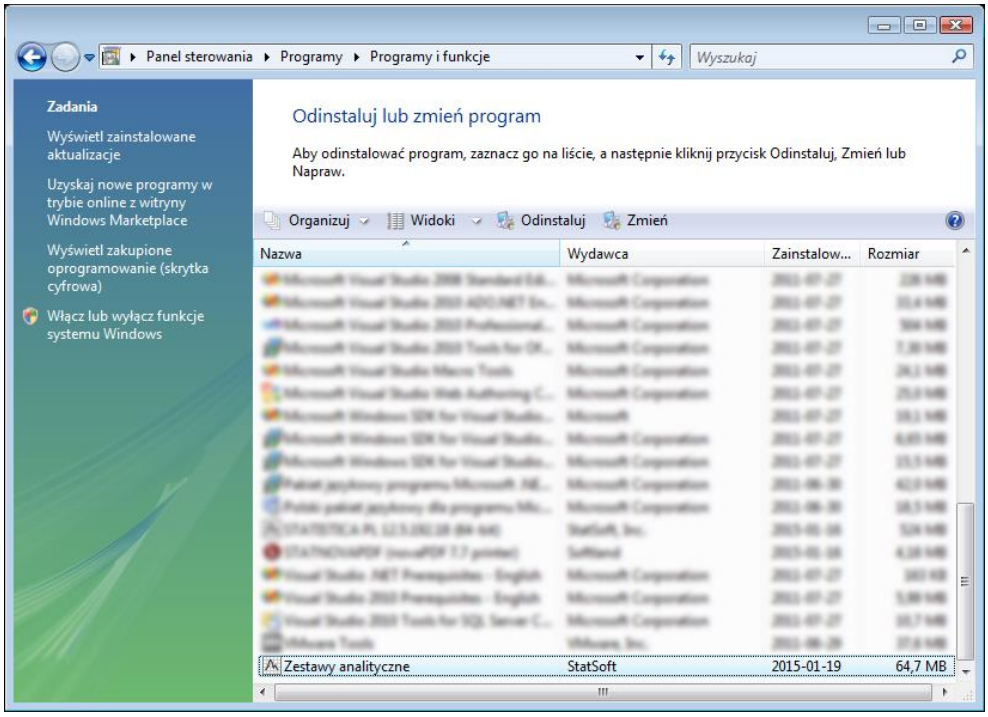
1.2. Odinstalowanie

W celu odinstalowania programu należy z menu ***Narzędzia*** wybrać ***Makro*** | ***Dodatki***



Pojawi się okno ***Dodatki STATISTICA***, należy zaznaczyć ***Zestawy analityczne***, a następnie kliknąć przycisk ***Usuń***.

Następnie należy wejść do ***Panelu sterowania*** systemu Windows, a następnie wybrać grupę ***Programy*** -> ***Programy i funkcje***.



W wyświetlonym oknie należy odnaleźć program ***Zestawy analityczne*** i kliknąć na niego dwukrotnie myszą lub wybrać opcję ***Odinstaluj*** w celu rozpoczęcia procesu dezinstalacji.

W przypadku trudności z odinstalowaniem programu ***Zestawy analityczne*** zalecane jest skorzystanie z narzędzia Microsoft Fix it dla problemów z instalacją i usuwaniem programów dostępnego na stronie firmy Microsoft.

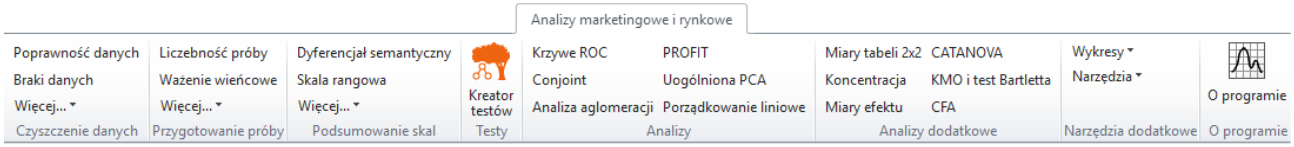
1.3. Wersja demonstracyjna

Wersja demonstracyjna programu ***Zestaw do analiz marketingowych i rynkowych*** udostępnia pełny zakres analiz z modułów ***Analiza ROC*** i ***Miary efektu dla tabel 2x2***.

2. Ogólne założenia programu

2.1. Przegląd modułów programu

Zestaw do analiz marketingowych i rynkowych jest zbiorem modułów analitycznych i kreatorów ułatwiających i przyspieszających proces opracowania wyników badań. Program rozszerza bogaty zestaw narzędzi do analizy zawarty w *STATISTICA* o dodatkowe funkcje, chętnie wykorzystywane przez badaczy podczas statystycznego opracowywania danych. Program składa się z siedmiu grup modułów:



- Czyszczenie danych
- Przygotowanie próby
- Podsumowanie skal
- Testy
- Analizy
- Narzędzia dodatkowe
    - Analizy dodatkowe
    - Wykresy
    - Narzędzia

Grupa ***Czyszczenie danych*** zawiera zestaw technik pozwalających na przygotowanie danych do analizy. Użytkownik ma możliwość sprawdzenia *Reguł poprawności danych*, *Uzupełnienia braków danych* czy *Zliczenia wystąpień* danych wartości we wskazanej grupie zmiennych. Możliwe są również przekształcenia i przekodowania zmiennych.

Grupa modułów ***Przygotowanie próby*** umożliwia użytkownikowi określenie liczebności próby a także wykonanie ważenia przypadków za pomocą metody *Ważenie wieńcowe*. Grupę zamyka moduł służący do korekty obciążenia próby *Propensity score matching* oraz moduł umożliwiający *Podział na podpróby*.

Podsumowanie skal zawiera moduły pozwalające w prosty sposób podsumować skale złożone takie jak skala *dyferencjału semantycznego*, *skala rangowa*, czy *skala Thurstone'a*. Grupa zawiera dodatkowo moduł oceniający *rzetelność skali Likerta* a także obliczający szereg wskaźników zgodności sędziów.

Testy zawiera *Kreator testów statystycznych* przeznaczony dla osób pragnących zweryfikować prawdziwość swojej hipotezy badawczej za pomocą testu statystycznego, mających jednocześnie trudności z określeniem testu, który byłby najbardziej odpowiedni w ich sytuacji. Kreator automatycznie sprawdza wszelkie założenia związane z danym typem problemu i w zależności od ich spełnienia proponuje poprawny test. Korzystając z tego narzędzia badacz musi jedynie określić kwestie merytoryczne prowadzonej analizy.

Grupa ***Analizy*** zawiera zestaw dziewięciu modułów analitycznych umożliwiających wykonanie bardziej zaawansowanych analiz. Użytkownik ma możliwość wykonania *Krzywych ROC* czy też metod: *Conjoint*, *Aglomeracji*, *PROFIT*, *Uogólnionej PCA* oraz *Porządkowania liniowego*. Dodatkowo grupa zawiera dwie podgrupy:

- ***ANOVA – układy niestandardowe*** zawierająca moduły pozwalające badaczowi na wygodne i intuicyjne zdefiniowanie mniej standardowych układów eksperymentów,
- ***Porównanie i ocena metod*** zawierająca szereg narzędzi umożliwiających sprawdzenie czy dwie metody pomiaru dają równoważne wyniki. Narzędzia ta pozwalają również na ocenę jakości wybranego sposobu pomiaru, poprzez wyznaczenie pewnych charakterystyk, świadczących o jego jakości.

Grupa ***Analizy dodatkowe*** zawiera szereg mniejszych modułów przydatnych w pracy analityków. Program pozwala obliczyć *Miary powiązania/efektu tabel 2x2*, wykonać *Analizę koncentracji* czy obliczyć *Standaryzowane miary powiązania/efektu*. Dodatkowo użytkownik ma możliwość wykonania analizy *CATANOVA*, kryterium *KMO* oraz *testu sferyczności Bartletta*. Grupa zawiera również metodę *CFA – konfiguracyjna analiza częstości*

Grupa ***Wykresy*** umożliwia przygotowanie szeregu dostosowanych wykresów. Dostępne są wykresy *słupkowy z kolorowymi słupkami*, wykres *sekwencyjny*, *radarowy*, *mozaikowy* oraz *kołowy (Spie plot)*.

Grupa ***Narzędzia*** pozwala na sformatowanie uzyskanych wyników a także na zapisanie ich w pliku Excela, bądź eksport grupy wykresów do wskazanego formatu.

2.2. Zbiory danych

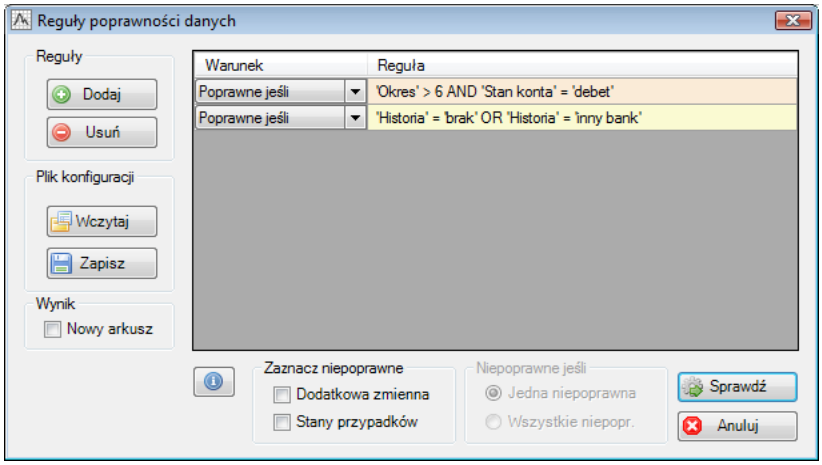
Wszystkie przykłady przedstawione w tej dokumentacji bazują na zestawie plików znajdujących się w katalogu *Zbiory danych*. W katalogu *Zbiory danych* znajdują się przykładowe arkusze *STATISTICA* używane podczas wykonywania przykładów przedstawiających funkcjonalność odpowiednich modułów ***Zestawu do analiz marketingowych i rynkowych***.

3. Czyszczenie danych

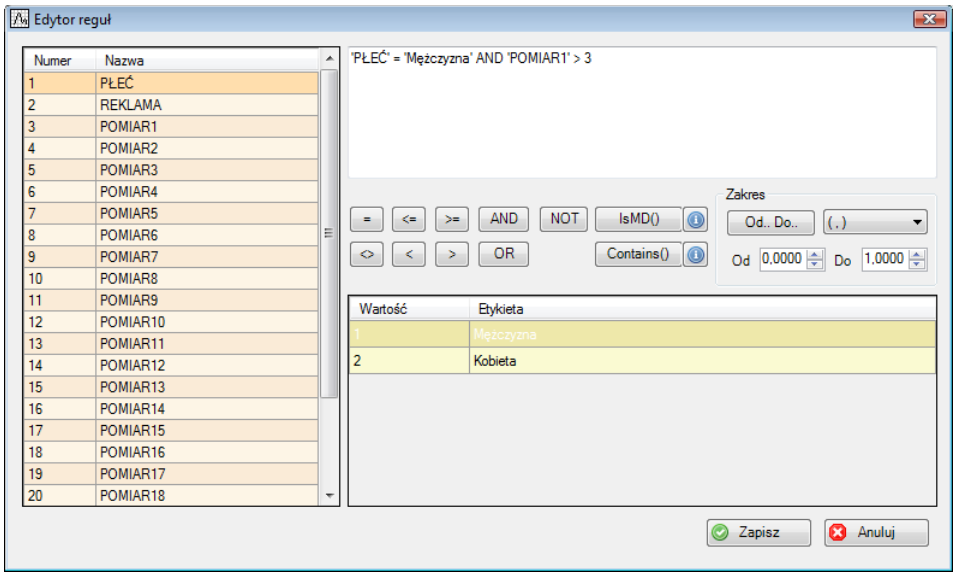
Pierwsza grupa modułów uzupełnia bogaty wybór narzędzi do czyszczenia danych zawarty w *STATISTICA* o:

3.1. Reguły poprawności danych

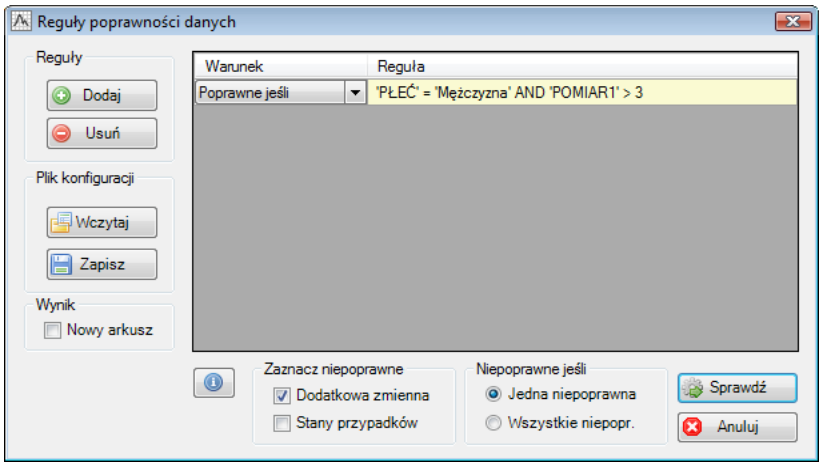
Dzięki tej opcji użytkownik ma możliwość łatwiejszego zdefiniowania reguł poprawności danych. W oknie definiowania reguł można wybrać nazwę zmiennej, wskazać żądane wartości lub odpowiadające im etykiety i połączyć je odpowiednim operatorem. Oprócz prostych reguł logicznych użytkownik ma możliwość wykorzystania zaawansowanej biblioteki funkcji zawartej w *STATISTICA*. Generowanie reguł ułatwia specjalnie przygotowany kreator. W arkuszu utworzone zostają zmienne wskazujące przypadki poprawne z punktu widzenia określonej reguły. Dodatkowo można utworzyć zmienną sprawdzającą poprawność względem wszystkich podanych reguł. Przygotowane reguły możemy zapisać do pliku konfiguracyjnego i wykorzystywać w innych modułach programu.



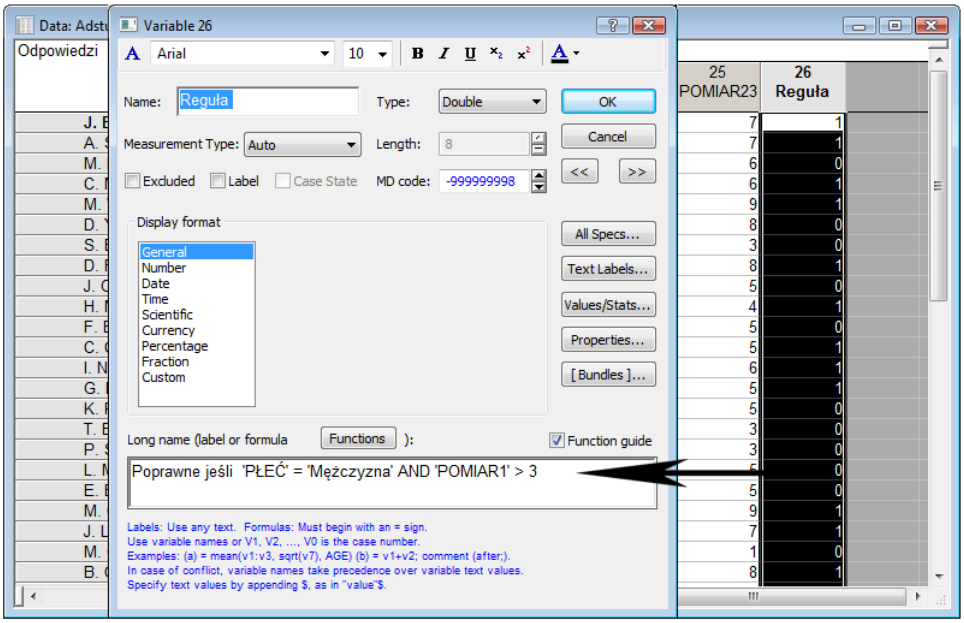
Przykład. Definiowanie reguł poprawności danych zaprezentowane zostanie na podstawie zbioru *Adstudy.sta*. Z menu ***Analizy marketingowe i rynkowe*** | ***Czyszczenie danych*** wybieramy opcję ***Poprawność danych***, przywołując okno ***Reguły poprawności danych***. Następnie klikamy przycisk ***Dodaj***, co spowoduje dodanie do listy reguł nowego wiersza. W polu ***Warunek*** możemy określić, czy definiowana reguła będzie regułą poprawności czy też błędu. Samą regułę określamy w polu ***Reguła***, wpisując ją ręcznie lub korzystając z pomocy kreatora (okno ***Edytor reguł***), który przywołujemy, klikając dwukrotnie na tym polu.



Edytor reguł umożliwia wygodne definiowanie reguł poprawności. Przykładowo załóżmy, że przypadki niepoprawne to takie, dla których zmienna *PŁEĆ* to *Mężczyzna* i równocześnie zmienna *POMIAR1* jest większa od 3. Aby wstawić do reguły nazwy zmiennych, klikamy dwukrotnie na wybranym polu listy znajdującej się po lewej stronie okna, natomiast nazwy klas zmiennych jakościowych dostępne są w dolnej części okna po kliknięciu na odpowiedniej zmiennej na liście. Operatory matematyczne oraz logiczne wstawiamy za pomocą odpowiednich przycisków (dodatkowo reguły mogą zawierać te same funkcje, jakie są dostępne w formułach zmiennych). Przygotowywana reguła wyświetla się w górnej części edytora. Po określeniu reguły klikamy przycisk ***Zapisz***, wracając do okna ***Reguły poprawności danych***.



Określona reguła poprawności została wstawiona do pola edycji. Klikając przycisk ***Dodaj***, możemy w sposób analogiczny definiować kolejne reguły. Sprawdzenie poprawności wykonujemy, klikając przycisk ***Sprawdź***, co spowoduje, że dla każdej zdefiniowanej reguły utworzona zostanie dodatkowa zmienna informująca, czy dany przypadek jest poprawny (wartość 1) czy też niepoprawny (wartość 0) względem zdefiniowanej reguły. Jeśli zdefiniowaliśmy więcej niż jedną regułę, wtedy zaznaczenie opcji ***Dodatkowa zmienna*** spowoduje utworzenie zmiennej informującej o poprawności względem wszystkich reguł równocześnie. Opcja ***Stany przypadków*** pozwala z kolei na wyróżnienie niepoprawnych przypadków za pomocą stanu ***Wyróżniony*** (w nazwie przypadku pojawi się czerwony wykrzyknik).

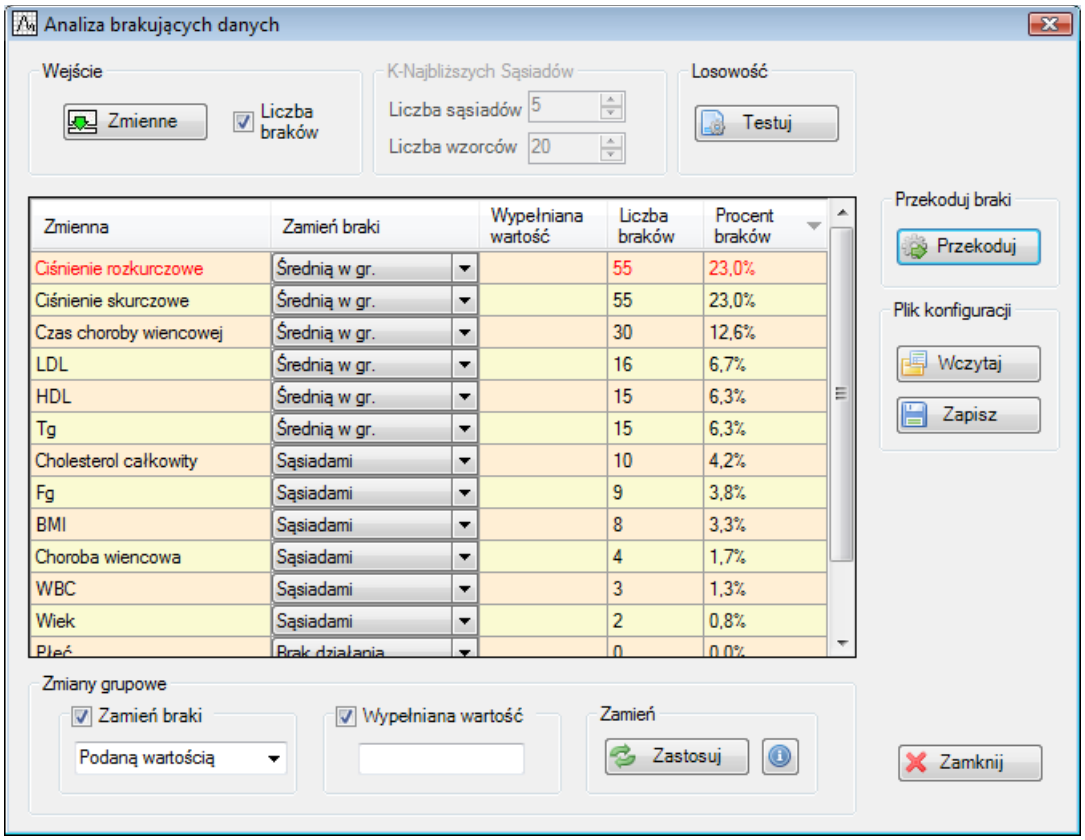


Po kliknięciu przycisku ***Sprawdź*** została dodana zmienna ***Reguła*** zawierająca informację o poprawności danego przypadku. Dodatkowo w tej zmiennej w polu *Długa nazwa* została wpisana treść zdefiniowanej reguły.

3.2. Analiza brakujących danych

Moduł umożliwia przekodowanie braków danych według wskazanego schematu. Moduł oferuje:

- Bogaty zestaw technik imputacji braków danych
  - Średnią, medianą, modalną
  - Średnią bądź medianą w grupach
  - Najbliższymi sąsiadami
  - Podaną wartością
- Łatwe określanie tej samej akcji dla wielu zmiennych
- Testowanie losowości braków danych
- Zapis określonych schematów kodowania do pliku konfiguracji



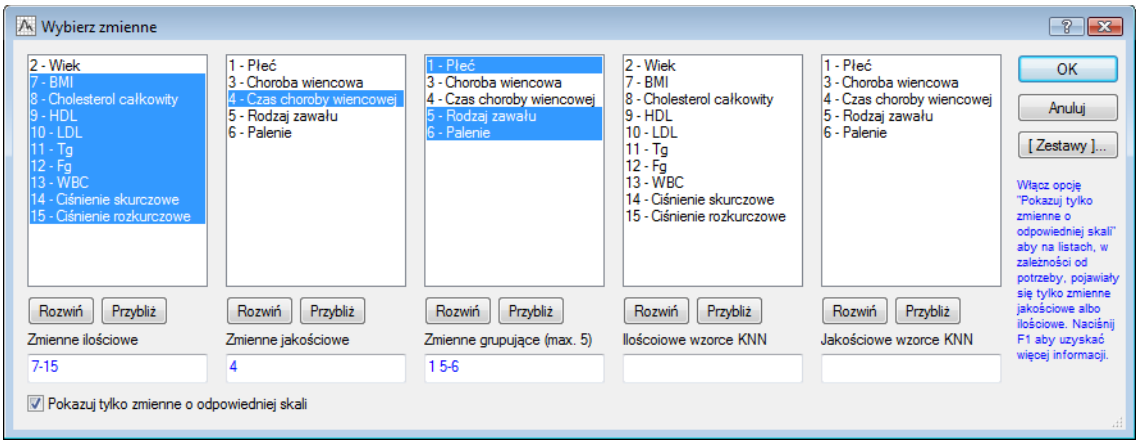
Przykład. Do analizy braków danych wykorzystamy plik *Zawały.sta*, który zawiera informacje o wybranych parametrach biochemicznych oraz klinicznych, zebranych wśród pacjentów z chorobą niedokrwienną serca.

Dane: Zawały (15 zmn. * 239 prz.)

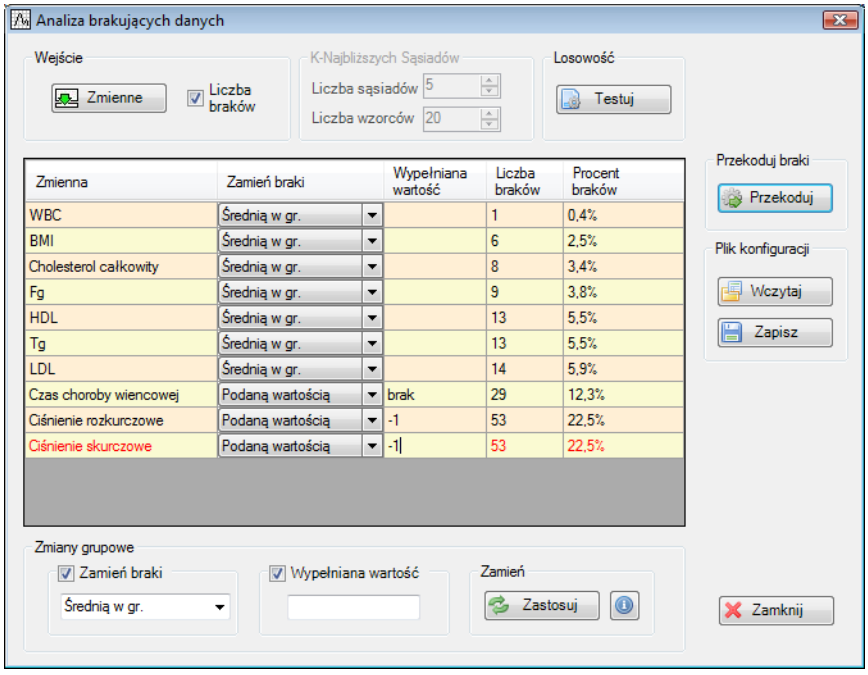
Dane dotyczą wybranych parametrów biochemicznych oraz klinicznych zebranych dla pacjentów z chorobą niedokrwienną serca.
Źródło: Watała C., Biostatystyka - wykorzystanie metod statystycznych w pracy badawczej w naukach biomedycznych, α-medica press 2002.

	1 Płeć	2 Wiek	3 Choroba wiencowa	4 Czas choroby wiencowej	5 Rodzaj zawału	6 Palenie	7 BMI	8 Cholesterol całkowity	9 HDL	10 LDL
KS/95/0004	Kobieta	32	tak	poniżej 2 m-cy	bez zawału	pali	18,65	205	27	100,8
KS/95/0007	Mężczyzna	32	tak	poniżej 2 m-cy	bez zawału	pali	18,65	205	27	100,8
KS/95/0014	Kobieta	49	tak	od 2 do 12 m-cy	bez zawału	nie pali	29,74	272	43	205,6
KS/95/0016	Mężczyzna	67	tak	od 2 do 12 m-cy	bez zawału	pali	28,84	208	46	138,2
KS/95/0018	Kobieta	63	tak	powyżej 12 m-cy	bez zawału	nie pali	33,20	205	41	121,0
KS/95/0021	Kobieta	64	tak	poniżej 2 m-cy	bez zawału	nie pali	18,62	209	66	120,6
KS/95/0023	Mężczyzna	47	tak	powyżej 12 m-cy	pełnościenny	pali	34,68	272	31	79,8
KS/95/0027	Mężczyzna	61	tak	poniżej 2 m-cy	bez zawału	pali	26,53	215	37	141,2
KS/95/0028	Mężczyzna	84	nie		bez zawału	nie pali		177	33	127,0
KS/95/0031	Kobieta	52	tak	powyżej 12 m-cy	bez zawału	nie pali	23,51	204	39	138,2
KS/95/0034	Kobieta	63	tak	powyżej 12 m-cy	bez zawału	nie pali	30,11	234	46	154,4
KS/95/0039	Kobieta	49	tak	od 2 do 12 m-cy	bez zawału	pali	26,30	212	48	141,0
KS/95/0040	Mężczyzna	68	nie		bez zawału	nie pali	22,55	156	38	93,0
KS/95/0046	Mężczyzna	32	tak	poniżej 2 m-cy	pełnościenny	pali	29,39	282	20	120,6

Aby uzupełnić braki danych z menu ***Analizy marketingowe i rynkowe*** – ***Czyszczenie danych*** wybieramy opcję ***Braki danych***. Następnie w oknie ***Analiza brakujących danych*** klikamy przycisk ***Zmienne*** i określamy wybór zgodnie z poniższą specyfikacją.



Dla zmiennych *Ciśnienie skurczowe* oraz *Ciśnienie rozkurczowe* wskazujemy zamianę braków wartością -1 (może to być dowolna wartość spoza zakresu poprawnych wartości ciśnienia). Pozostałe zmienne zastąpimy wartością średnią w grupach zdefiniowanych przez wskazane zmienne grupujące. W tym celu zaznaczamy wybrane zmienne w tabeli a następnie w obszarze ***Zmiany grupowe*** na liście wyboru wskazujemy opcję ***Średnią w gr.***, następnie w obszarze ***Zamień*** klikamy przycisk ***Zastosuj***. Spowoduje to wybór wskazanego schematu przekodowania braków we wszystkich zaznaczonych wierszach tabeli.

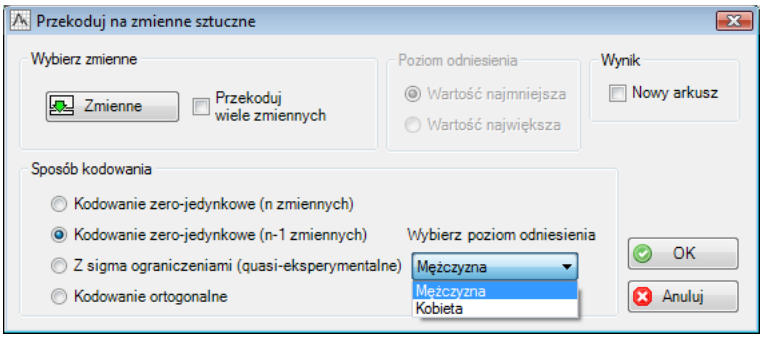


Klikamy ***Przekoduj***, aby zastąpić braki danych zgodnie z określonym schematem.

3.3. Zmienne sztuczne

Moduł umożliwia zamianę cech jakościowych na odpowiadające im zmienne sztuczne. W programie zaimplementowano cztery schematy kodowania:

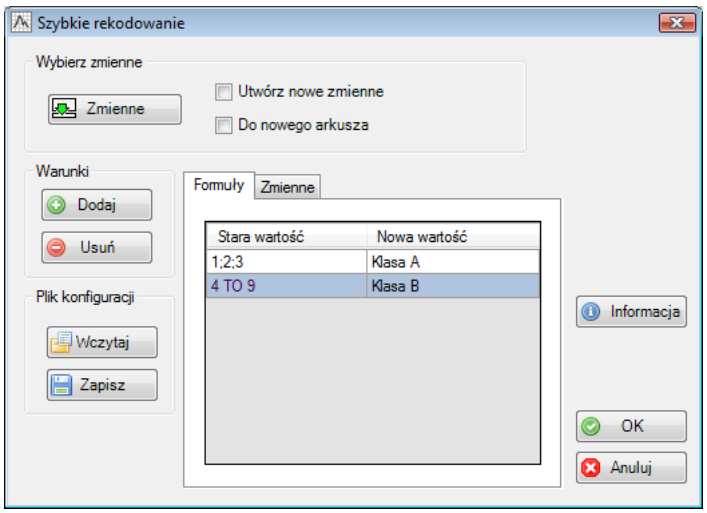
- Kodowanie zero-jedynkowe na n zmiennych (n to liczba poziomów kodowanej cechy),
- Kodowanie zero-jedynkowe na n-1 zmiennych,
- Kodowanie z sigma ograniczeniami (quasi-eksperymentalne),
- Kodowanie ortogonalne.



W przypadku wyboru trzech ostatnich schematów mamy możliwość ręcznego wyboru poziomu odniesienia (poziomu ukrytego, nieuwzględnionego w zmiennych sztucznych).

3.4. Szybkie rekodowanie

Moduł umożliwia przekodowanie wybranych wartości (dowolna wartość, braki danych, wartość z podanego zakresu) zmiennych do wskazanych nowych wartości. Użytkownik może przekodować jednocześnie wiele zmiennych, a nowe kody mogą zostać wprowadzone do tych samych lub nowych zmiennych.



Dopuszczalna składnia wartości, jakie możemy określić w kolumnie ***Stara wartość*** to:

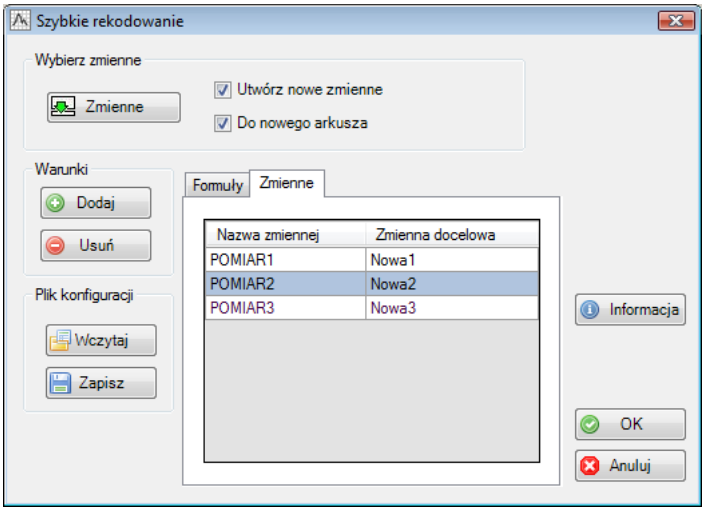
- *Wartość* - Dowolna wartość liczbowa
- *NULL* - Braki danych
- *NOT NULL* - Wartości inne niż braki danych
- *Wartość1 TO Wartość2* - zakres od Wartość1 do Wartość2 (włącznie)
- *Wartość1 TO* .. - Wartości większe bądź równe Wartość1
- .. *TO Wartość2* - Wartości mniejsze bądź równe Wartość2

Przykład. Przekodowanie danych zostanie zaprezentowane na podstawie zbioru *Adstudy.sta*. Z menu ***Analizy marketingowe i rynkowe*** | ***Czyszczenie danych*** wybieramy opcję ***Szybkie rekodowanie***, przywołując okno o tej samej nazwie. Następnie klikamy przycisk ***Zmienne***, aby wybrać zmienne, jakie będziemy chcieli przekodować, i zaznaczamy zmienne od 3 do 5.

Następnie klikamy przycisk ***Dodaj***, dodając pierwszy warunek kodowania do listy. W kolumnie ***Stara wartość*** wpisujemy wartości *1;2;3*, (symbol ; jest separatorem wartości) natomiast w kolumnie ***Nowa wartość*** wpisujemy *Klasa A*. Ponownie klikamy przycisk ***Dodaj***, by dodać kolejny warunek. Tym razem będzie on miał treść *4 TO 9*. W polu ***Nowa wartość*** wpiszemy *Klasa B* co spowoduje

przekodowanie wartości od 4 do 9 na wartość *Klasa B*. Szczegóły składni dostępne są po kliknięciu przycisku ***Informacja***.

Po wprowadzeniu formuł zaznaczamy opcję ***Utwórz nowe zmienne***, co spowoduje, że przekodowane wartości trafią do nowo utworzonych zmiennych. Zaznaczenie tej opcji uaktywni kartę ***Zmienne***, gdzie będziemy mogli określić nazwy nowych zmiennych.



Po przejściu na kartę zmienne określamy nazwy zmiennych docelowych jako *Nowa1*, *Nowa2* oraz *Nowa3*. Dodatkowo zaznaczamy opcję ***Do nowego arkusza***. Klikamy ***OK***., zatwierdzając analizę, i otrzymujemy nowy arkusz danych zawierający przekodowane wartości zmiennych. Reguły kodowania możemy zapisać do pliku konfiguracji za pomocą przycisku ***Zapisz***, a następnie wczytać je dla innego zbioru danych (przycisk ***Wczytaj***).

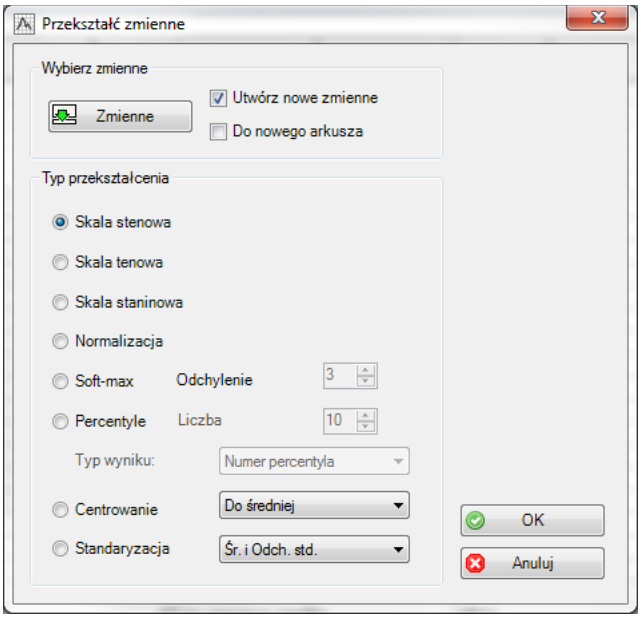
3.5. Przekształcenia zmiennych

Moduł umożliwia normalizację zmiennych zgodnie z jednym z ośmiu typów przekształceń:

- Skala stenowa
- Skala tenowa
- Skala staninowa
- Normalizacja
- Przekształcenie Soft-max
- Podział na percentyle
- Centrowanie
- Standaryzacja

Moduł umożliwia zapisanie przekształcanych zmiennych w nowym bądź bieżącym arkuszu danych.

Przykład Aby zademonstrować działanie modułu ***Przekształcenia zmiennych***, użyjemy pliku *Adstudy.sta* zawartego w zestawie plików przykładowych. Po jego otwarciu z menu ***Analizy marketingowe i rynkowe*** | ***Czyszczenie danych*** wybieramy odpowiednie polecenie, przywołując okno ***Przekształć zmienne***.



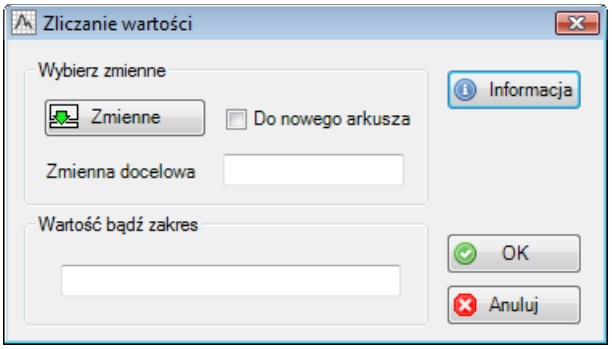
Za pomocą przycisku ***Zmienne*** wybieramy zmienne do analizy, *POMIAR1* oraz *POMIAR2*. Jeżeli chcemy aby przekształcone zostały oryginalne zmienne anulujemy opcję ***Utwórz nowe zmienne***. W obszarze ***Typ przekształcenia*** wybieramy opcję ***Normalizacja*** a następnie klikamy **OK.**, aby wykonać analizę. W wyniku analizy stare wartości zostały przekształcone zgodne z wybranym typem przekształcenia.

Dane: Adstudy (25 zmn. * 50 prz.)

Odpowiedzi	Badanie efektywności reklamy.				
	1 PŁEĆ	2 REKLAMA	3 POMIAR1	4 POMIAR2	5 POMIAR3
J. Baker	Mężczyzna	PEPSI	1	0,111111	6
A. Smith	Mężczyzna	COKE	0,666667	0,777778	1
M. Brown	Kobieta	COKE	1	0,888889	2
C. Mayer	Mężczyzna	PEPSI	0,777778	1	0
M. West	Mężczyzna	PEPSI	0,777778	0,111111	6
D. Young	Kobieta	COKE	0,666667	0	0
S. Bird	Kobieta	COKE	0,777778	0,444444	3
D. Flynd	Mężczyzna	PEPSI	1	1	2
J. Owen	Kobieta	PEPSI	0,777778	0,888889	2
H. Morrow	Mężczyzna	PEPSI	0,666667	0,666667	2
F. East	Kobieta	PEPSI	0,444444	0,666667	6
C. Clint	Mężczyzna	COKE	0,777778	0,333333	3
I. Neil	Mężczyzna	PEPSI	0,666667	0,222222	3
G. Boss	Mężczyzna	COKE	0,777778	0,222222	4
K. Record	Kobieta	PEPSI	0,666667	0,222222	7
T. Bush	Kobieta	PEPSI	0,333333	0,222222	5
P. Squire	Mężczyzna	COKE	0,222222	1	9
L. Mynard	Kobieta	PEPSI	0,111111	0	7
E. Bynum	Kobieta	COKE	0	0,666667	2
M. Quick	Mężczyzna	COKE	0,666667	0,888889	1
J. Liu	Mężczyzna	PEPSI	1	0,222222	7

3.6. Zliczanie wartości

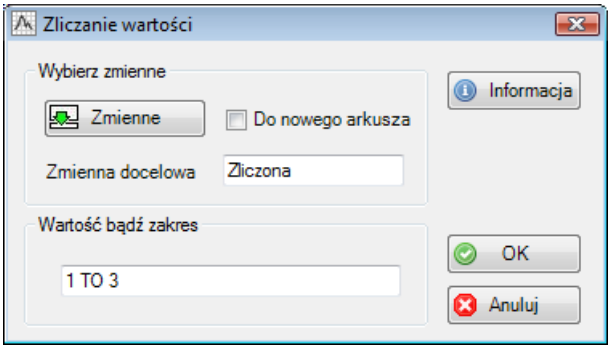
Moduł umożliwia utworzenie dodatkowej zmiennej zawierającej informację, ile razy w danym wierszu w określonej liście zmiennych wystąpiła wskazana przez użytkownika wartość. Oprócz pojedynczej wartości użytkownik może zliczyć wystąpienie wartości z dowolnego zakresu, a także braki danych oraz wartości nie będące brakami danych.



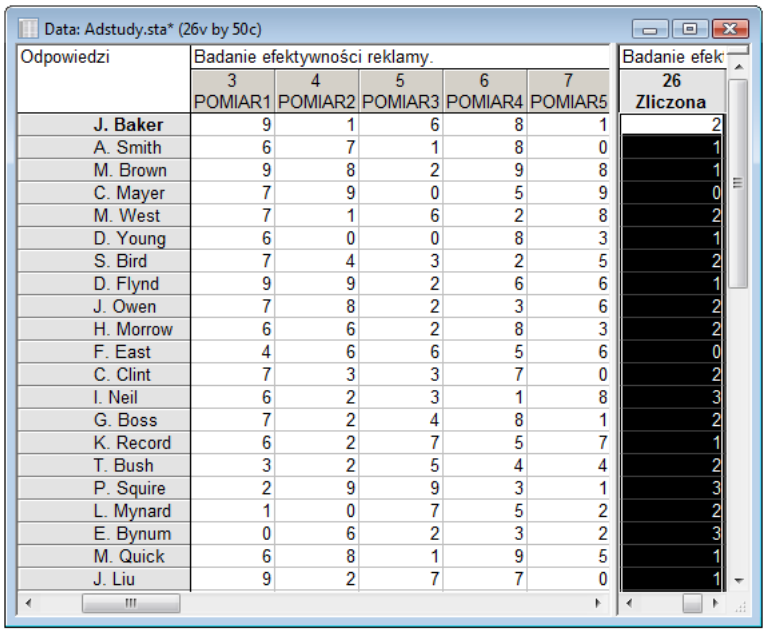
Dopuszczalna składnia wartości, jakie możemy określić w polu edycji ***Wartość bądź zakres*** to:

- *Wartość* - Dowolna wartość liczbowa
- *NULL* - Braki danych
- *NOT NULL* - Wartości inne niż braki danych
- *Wartość1 TO Wartość2* - zakres od Wartość1 do Wartość2 (włącznie)
- *Wartość1 TO* .. - Wartości większe bądź równe Wartość1
- .. *TO Wartość2* - Wartości mniejsze bądź równe Wartość2

Przykład. Zliczanie wartości zostanie zaprezentowane dla zbioru *Adstudy.sta* zawartego w zestawie plików przykładowych. Z menu ***Analizy marketingowe i rynkowe*** | ***Czyszczenie danych*** wybieramy opcję ***Zliczanie wartości*** przywołując okno o tej samej nazwie. Następnie klikamy przycisk ***Zmienne***, aby wybrać zmienne, jakie będziemy chcieli uwzględnić w analizie. W wyświetlonym oknie zaznaczamy zmienne od 3 do 7. Następnie określamy nazwę zmiennej, do której trafi wynik wykonywanej analizy. W polu ***Zmienna docelowa*** wpisujemy wartość ***Zliczona***.



Następnie w obszarze ***Wartość bądź zakres*** określamy, aby zliczane były wartości od jeden do trzech, wpisując formułę *1 To 3*. Po zatwierdzeniu analizy do analizowanego zbioru danych zostanie dodana zmienna *Zliczona,* zawierająca liczbę wystąpień podanego zakresu w wybranych zmiennych.



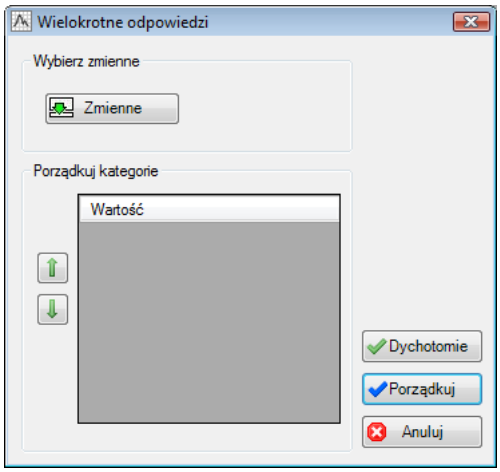
Data: Adstudy.sta* (26v by 50c)

Odpowiedzi	Badanie efektywności reklamy.					Badanie efek
	3 POMIAR1	4 POMIAR2	5 POMIAR3	6 POMIAR4	7 POMIAR5	26 Zliczona
J. Baker	9	1	6	8	1	2
A. Smith	6	7	1	8	0	1
M. Brown	9	8	2	9	8	1
C. Mayer	7	9	0	5	9	0
M. West	7	1	6	2	8	2
D. Young	6	0	0	8	3	1
S. Bird	7	4	3	2	5	2
D. Flynd	9	9	2	6	6	1
J. Owen	7	8	2	3	6	2
H. Morrow	6	6	2	8	3	2
F. East	4	6	6	5	6	0
C. Clint	7	3	3	7	0	2
I. Neil	6	2	3	1	8	3
G. Boss	7	2	4	8	1	2
K. Record	6	2	7	5	7	1
T. Bush	3	2	5	4	4	2
P. Squire	2	9	9	3	1	3
L. Mynard	1	0	7	5	2	2
E. Bynum	0	6	2	3	2	3
M. Quick	6	8	1	9	5	1
J. Liu	9	2	7	7	0	1

3.7. Zmienne wielokrotnych odpowiedzi

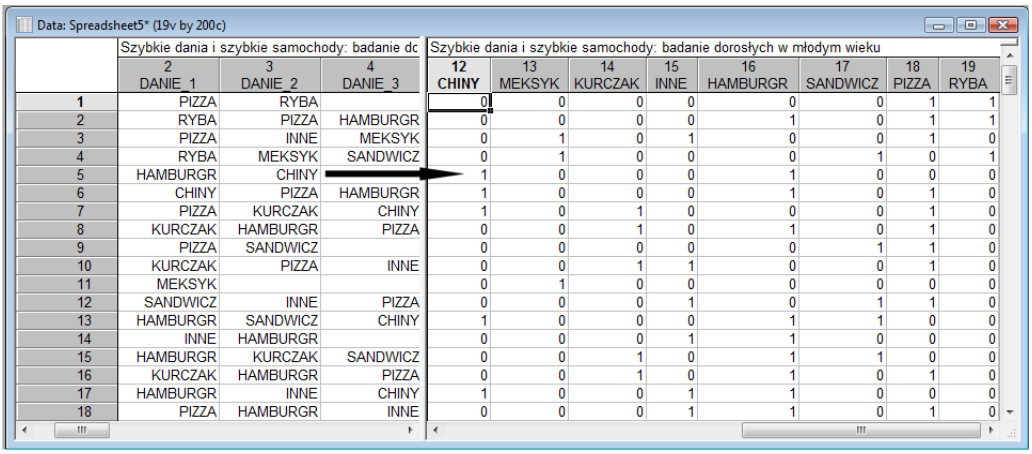
Moduł umożliwia przekodowanie zmiennych wielokrotnych odpowiedzi na wielokrotne dychotomie, bądź uporządkowanie wartości w zmiennych wielokrotnych odpowiedzi z zachowaniem sposobu kodowania. Odpowiedzi dla danego przypadku zostaną posortowane zgodnie z kolejnością podaną przez użytkownika.

Przykład. Aby zademonstrować działanie modułu ***Zmienne wielokrotnych odpowiedzi***, użyjemy pliku *Fastfood.sta* zawartego w zestawie plików przykładowych. Po jego otwarciu z menu ***Analizy marketingowe i rynkowe*** | ***Czyszczenie danych*** wybieramy odpowiednie polecenie, przywołując okno ***Wielokrotne odpowiedzi***.



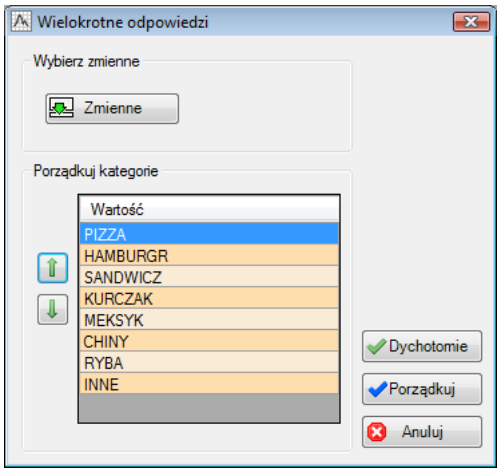
Następnie klikamy przycisk ***Zmienne*** i w oknie wyboru zmiennych wybieramy zmienne *DANIE_1*, *DANIE_2*, *DANIE_3*. Wybór tych zmiennych pozwoli na pobranie wszystkich występujących w nich unikalnych wartości i wyświetlenie ich w obszarze ***Porządkuj kategorie***.

Jeżeli chcemy przekodować wybrane zmienne na wielokrotne dychotomie, klikamy przycisk ***Dychotomie*** wyświetlając arkusz z nowoutworzonymi zmiennymi zerojedynkowymi.

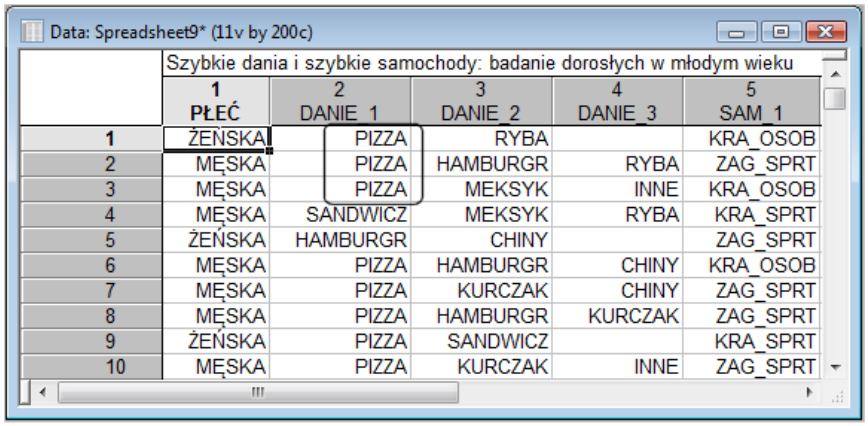


	2 DANIE_1	3 DANIE_2	4 DANIE_3	12 CHINY	13 MEKSYK	14 KURCZAK	15 INNE	16 HAMBURGR	17 SANDWICZ	18 PIZZA	19 RYBA
1	PIZZA	RYBA		0	0	0	0	0	0	1	1
2	RYBA	PIZZA	HAMBURGR	0	0	0	0	1	0	1	1
3	PIZZA	INNE	MEKSYK	0	1	0	1	0	0	1	0
4	RYBA	MEKSYK	SANDWICZ	0	1	0	0	0	1	0	1
5	HAMBURGR	CHINY		1	0	0	0	1	0	0	0
6	CHINY	PIZZA	HAMBURGR	1	0	0	0	1	0	1	0
7	PIZZA	KURCZAK	CHINY	1	0	1	0	0	0	1	0
8	KURCZAK	HAMBURGR	PIZZA	0	0	1	0	1	0	1	0
9	PIZZA	SANDWICZ		0	0	0	0	0	1	1	0
10	KURCZAK	PIZZA	INNE	0	0	1	1	0	0	1	0
11	MEKSYK			0	1	0	0	0	0	0	0
12	SANDWICZ	INNE	PIZZA	0	0	0	1	0	1	1	0
13	HAMBURGR	SANDWICZ	CHINY	1	0	0	0	1	1	0	0
14	INNE	HAMBURGR		0	0	0	1	1	0	0	0
15	HAMBURGR	KURCZAK	SANDWICZ	0	0	1	0	1	1	0	0
16	KURCZAK	HAMBURGR	PIZZA	0	0	1	0	1	0	1	0
17	HAMBURGR	INNE	CHINY	1	0	0	1	1	0	0	0
18	PIZZA	HAMBURGR	INNE	0	0	0	1	1	0	1	0

Jeżeli z kolei chcemy uporządkować kategorie w zbiorze nie zmieniając formatu kodowania, w obszarze ***Porządkuj kategorie*** określamy kolejność za pomocą strzałek. Przykładowo przesuniemy kategorię *PIZZA*, na pierwsze miejsce co spowoduje, że jeżeli wartość ta wystąpi w danym przypadku, to zawsze będzie ona zapisana w pierwszej zmiennej.



Po przygotowaniu odpowiedniej kolejności kategorii zatwierdzamy przygotowanie nowego zbioru danych, klikając ***Porządkuj***. Kategorie zmiennych *DANIE_1*, *DANIE_2* oraz *DANIE_3* w nowym zbiorze danych będą pojawiały się w określonej na liście kolejności.

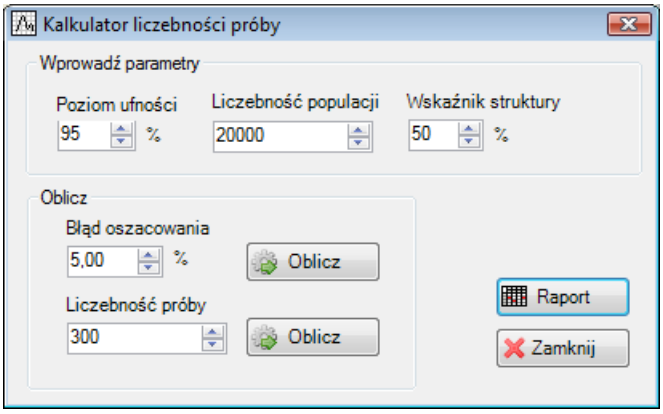


	1 PŁEĆ	2 DANIE_1	3 DANIE_2	4 DANIE_3	5 SAM_1
1	ŻEŃSKA	PIZZA	RYBA		KRA_OSOB
2	MĘSKA	PIZZA	HAMBURGR	RYBA	ZAG_SPRT
3	MĘSKA	PIZZA	MEKSYK	INNE	KRA_OSOB
4	MĘSKA	SANDWICZ	MEKSYK	RYBA	KRA_SPRT
5	ŻEŃSKA	HAMBURGR	CHINY		ZAG_SPRT
6	MĘSKA	PIZZA	HAMBURGR	CHINY	KRA_OSOB
7	MĘSKA	PIZZA	KURCZAK	CHINY	ZAG_SPRT
8	MĘSKA	PIZZA	HAMBURGR	KURCZAK	ZAG_SPRT
9	ŻEŃSKA	PIZZA	SANDWICZ		KRA_SPRT
10	MĘSKA	PIZZA	KURCZAK	INNE	ZAG_SPRT

4. Przygotowanie próby

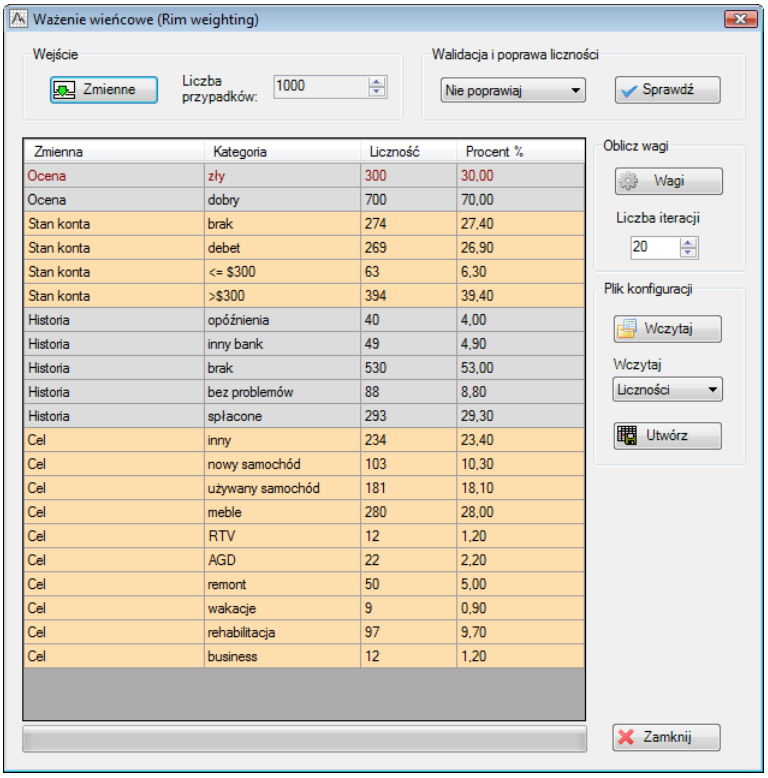
4.1. Liczebność próby

Moduł umożliwia z zadanym błędem oszacować liczebność próby dla frakcji lub dla zadanej liczebności próby obliczyć błąd oszacowania. Dodatkowymi parametrami uwzględnianymi w analizie są ***Poziom ufności***, ***Liczebność populacji*** oraz ***Wskaźnik struktury***.



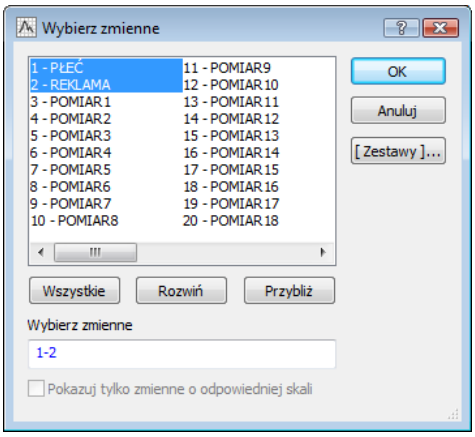
4.2. Ważenie wieńcowe

Moduł realizuje ważenie wieńcowe przypadków (*RIM weighting*). Moduł obsługuje ważenie względem maksymalnie sześciu wymiarów. Aby wygenerować zestaw wag dla przypadków wystarczy podanie rozkładów brzegowych dla poszczególnych wymiarów.

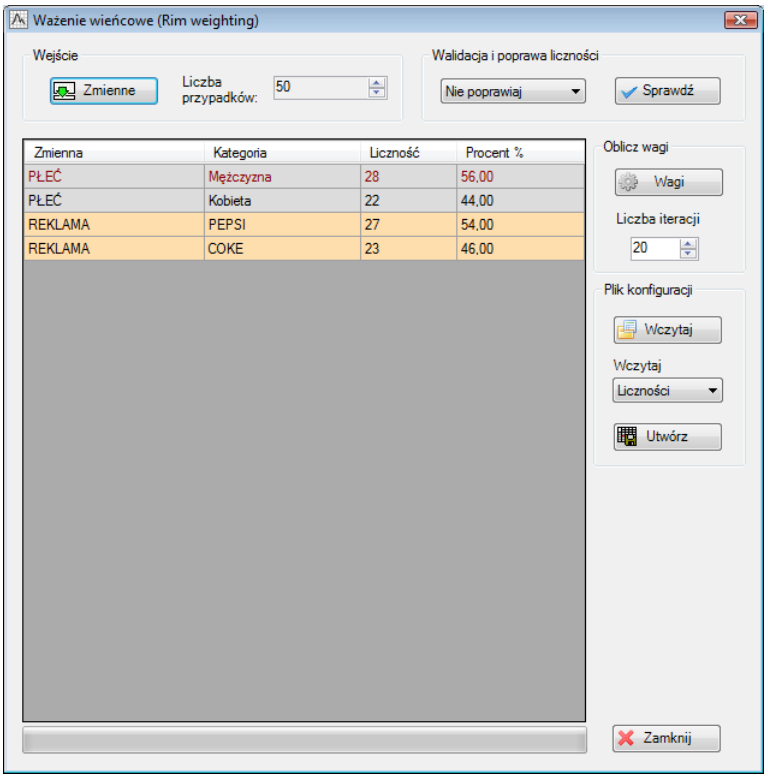


Przykład. Aby zademonstrować działanie modułu ***Ważenie wieńcowe***, użyjemy pliku *Adstudy.sta* zawartego w zestawie plików przykładowych. Po jego otwarciu z menu ***Analizy marketingowe i rynkowe*** | ***Przygotowanie próby*** wybieramy odpowiednie polecenie, przywołując okno ***Ważenie wieńcowe (Rim weighting)***. W oknie tym

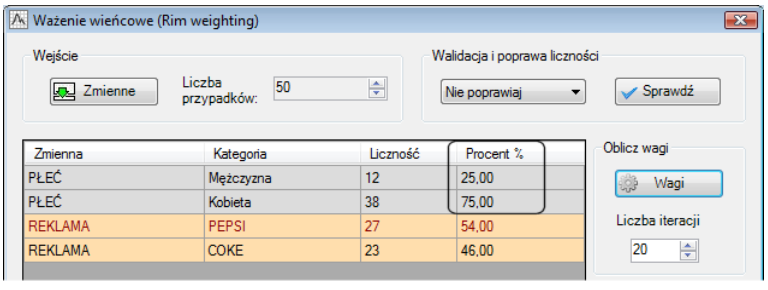
klikamy przycisk ***Zmienne***, aby wybrać zmienne, według których dobierane będą wagi. Wybieramy zmienne zgodnie z poniższym rysunkiem.



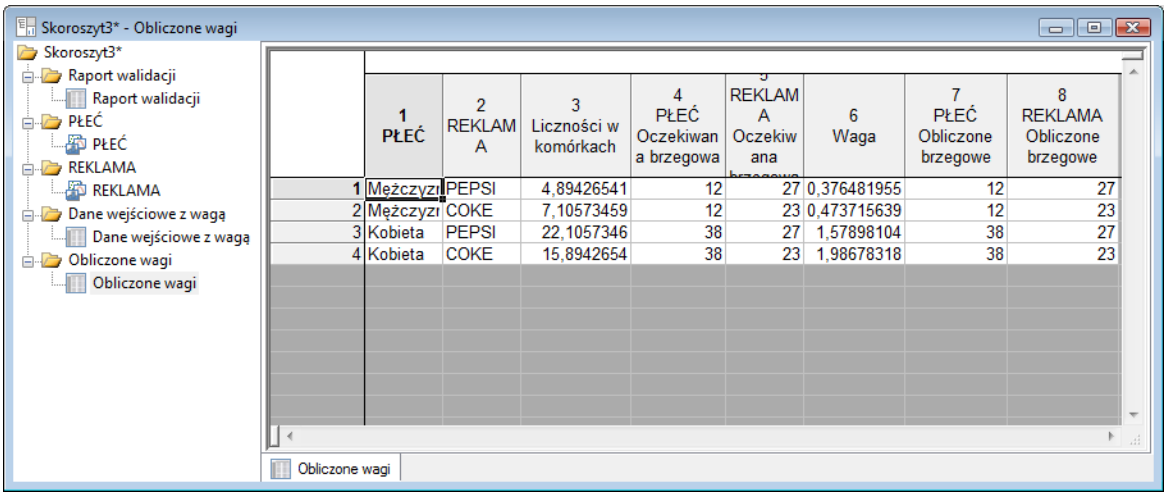
W oknie ***Ważenie wieńcowe (Rim weighting)*** pojawiają się wybrane zmienne wraz z przypisanymi im kategoriami oraz ich empirycznymi rozkładami.



Zgodnie z naszymi oczekiwaniami modyfikujemy rozkłady, wprowadzając nowe wartości w kolumnie ***Procent***. Załóżmy, że chcemy aby w zmiennej *Płeć* rozkład wartości klas był w stosunku 25:75.



Następnie klikamy przycisk ***Wagi***. W wyniku otrzymujemy rozbudowany raport zawierający wagi oraz rozkład po ich zastosowaniu:



Osobny arkusz zawiera dane wejściowe z nową zmienną *Waga*. Arkusz ten możemy następnie wykorzystać w dowolnych analizach wskazując zmienną Waga jako zmienną ważącą (dodatkowo zaznaczając opcję ***Momenty ważone***).

Uwaga. Podane wartości brzegowe możemy zapisać do pliku konfiguracyjnego (przycisk ***Utwórz***) i wykorzystać go ponownie w kolejnej analizie.

4.3. Propensity score matching

Obok badań eksperymentalnych, w których badacz ma pełną kontrolę nad procesem przydzielania obiektów do grup poddawanych określonym oddziaływaniom oraz do grupy kontrolnej drugim rodzajem badań są badania obserwacyjne. Są one powszechnie prowadzone w medycynie czy naukach społecznych. W tych badaniach z przyczyn prawnych, technicznych bądź etycznych badacz nie ma kontroli nad procesem przydzielania uczestników do badania, eksperymentu, bądź programu. Staje zatem przed problemem poprawnej oceny skuteczności danego oddziaływania, ponieważ nie dysponuje odpowiednią grupą porównawczą, która mogłaby być punktem odniesienia dla oceny mierzonego efektu.

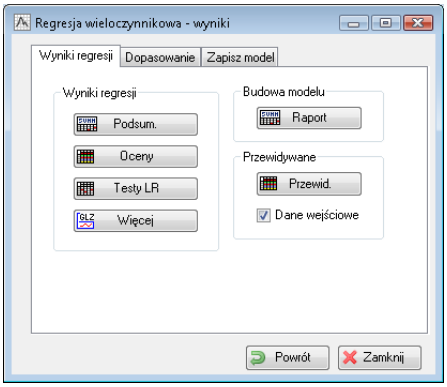
Przyjęcie za grupę kontrolną osób, które po prostu nie uczestniczyły w danym programie (a stanowiły dla niego grupę docelową) prowadzi do obciążonych ocen, ponieważ zakłada niezgodnie z rzeczywistością losowy (eksperymentalny) dobór do obydwóch grup. Celem metody *propensity score matching* jest korekta obciążenia oceny efektu spowodowanego nielosowym doborem do grupy interwencji i kontroli. Korekta ta odbywa się poprzez odpowiednie dopasowanie do każdego przypadku z grupy z interwencją przypadku bądź przypadków z grupy nie poddanej interwencji. Dopasowanie to przeprowadzamy w taki sposób aby rozkład cech analizowanych obiektów był maksymalnie zbliżony w obydwu grupach.

Ponieważ dopasowanie przypadków na podstawie wielu kryteriów jednocześnie wiąże się z wieloma trudnościami natury obliczeniowej, dlatego też zamiast tego dopasowujemy przypadki na podstawie syntetycznej miary *propensity score*, którą możemy zdefiniować jako skłonność do partycypacji w warunkach interwencji. Najczęściej miarę tą obliczamy za pomocą regresji logistycznej[1], zmienną zależną definiujemy jako fakt uczestnictwa\braku uczestnictwa w programie, predyktorami są zmienne, które w naszej ocenie determinują skłonność analizowanych obiektów do uczestnictwa w

[1] Oczywiście możemy użyć do tego celu dowolnej metody klasyfikacji np. sieci neuronowych czy losowego lasu.

programie. Prawdopodobieństwo uzyskane w wyniku zastosowania modelu na danych uczących traktujemy jako miarę skłonności do uczestnictwa w programie (*propensity score*).

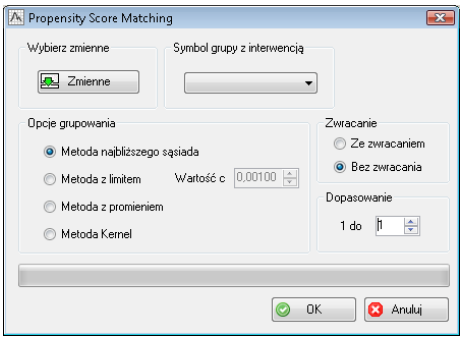
Model regresji logistycznej możemy zbudować za pomocą modułu do regresji logistycznej zawartego w programie *STATISTICA* lub za pomocą ***Kreatora regresji logistycznej*** wchodzącego w skład dodatku ***Zestaw Plus***. Po zbudowaniu modelu za pomocą *Kreatora* w oknie ***Regresja wieloczynnikowa*** – wyniki klikamy przycisk ***Przewid.*** zaznaczając opcję ***Dane wejściowe***.



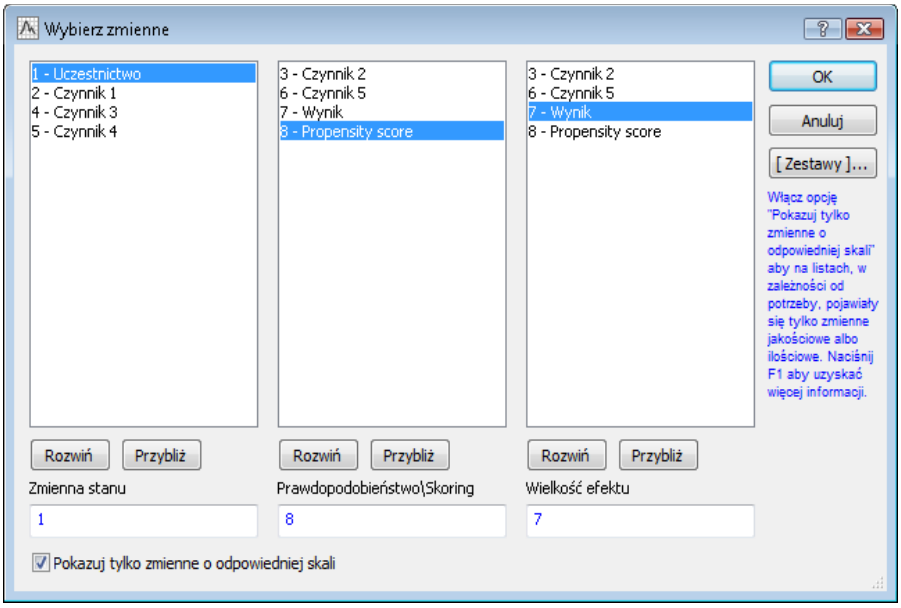
Wygenerowany arkusz zawiera oprócz danych wejściowych również prawdopodobieństwo obliczone na podstawie oszacowanego modelu, które traktujemy jako *propensity score*. Mając obliczoną wartość *propensity score* przechodzimy do opartej na niej selekcji przypadków.

Przykład wykorzystania techniki *propensity score matching* przedstawimy na wygenerowanym zbiorze *PSM.sta*, zawierającym informację o 1000 osobach, które znalazły się w grupie docelowej pewnego programu badawczego. Spośród tych osób 300 wzięło udział w programie, pozostałe 700 nie wzięło w nim udziału. Informacja o udziale w programie zawarta została w zmiennej *Uczestnictwo*. Zbiór zawiera dodatkowo cechy będące podstawą oceny skłonności uczestnictwa w programie, ocenę efektu programu oraz zmienną *Propensity score*.

Aby rozpocząć analizę z menu ***Analizy marketingowe i rynkowe*** z grupy ***Przygotowanie próby*** wybieramy opcję ***Propensity Score Matching*** otwierając okno o tej samej nazwie.

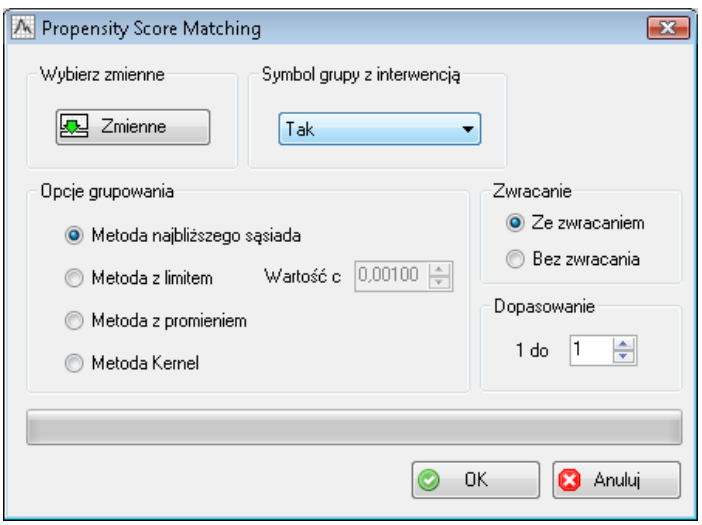


Klikamy przycisk ***Zmienne***, a następnie wskazujemy zmienne do analizy zgodnie z poniższą specyfikacją.

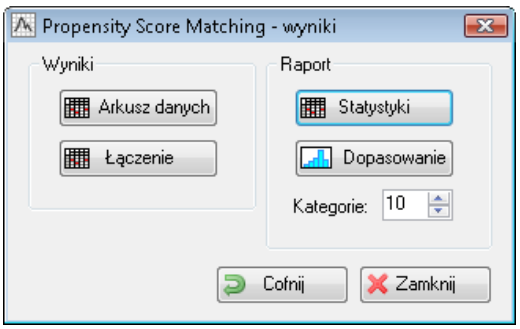


Po wybraniu zmiennych w obszarze ***Symbol grupy z interwencją*** określamy, jaka klasa zmiennej *Uczestnictwo* symbolizuje grupę z interwencją. Kolejnym zasadniczym krokiem analizy jest wybór metody grupowania. Do wyboru mamy cztery opcje:

- Metoda najbliższego sąsiada, która jest najprostszą a zarazem najbardziej popularną metodą polegającą na wyborze dla każdej osoby z grupy z interwencją osoby z grupy bez interwencji o najbliższej jej wartości *propensity score*. Algorytm ten możemy stosować zarówno ze zwracaniem jak i bez zwracania (opcja ***Zwracanie***) , dodatkowo wybierając za każdym razem jednego lub więcej sąsiadów (opcja ***Dopasowanie***).
- Metoda z limitem (*caliper*) jest modyfikacją metody najbliższego sąsiada, która powoduje, że w sytuacji, gdy nie uda się znaleźć sąsiada różniącego się nie więcej niż wskazana wartość *c*, dopasowanie nie dochodzi do skutku.
- Metoda z promieniem z kolei łączy z przypadkiem z grupy interwencji wszystkie przypadki z grupy bez interwencji różniące się od niego nie więcej niż wskazana wartość c.
- Metoda Kernel modyfikuje metodę z promieniem w ten sposób, że przypadkom z grupy bez interwencji, przypisanym do danego przypadku z grupy z interwencją, określa wagi odwrotnie proporcjonalne do odległości tych przypadków od przypadku z interwencją (w niniejszej implementacji ważenie odbywa się przy pomocy funkcji gęstości rozkładu normalnego ze średnią 0 i odchyleniem standardowym równym 1).

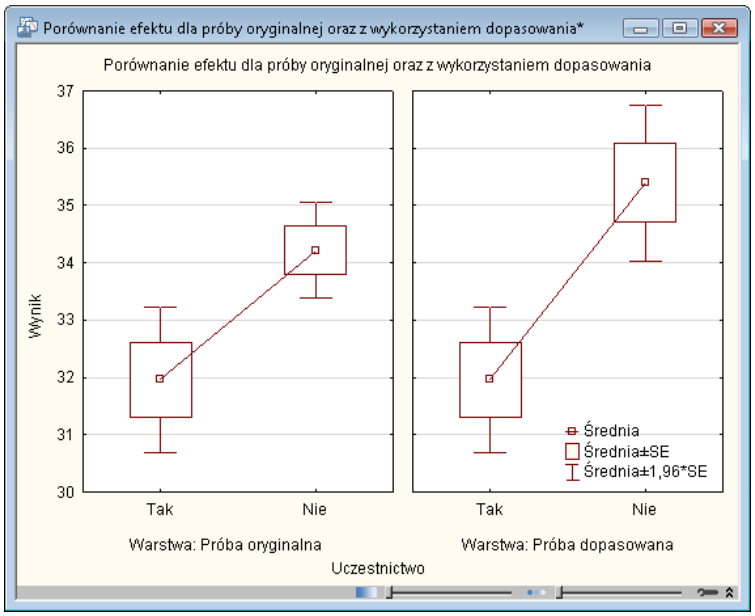


W naszym przykładzie wybierzemy najprostszą ***Metodę najbliższego sąsiada*** zaznaczając dodatkowo opcję ***Ze zwracaniem***. Klikamy ***OK***., aby rozpocząć procedurę łączenia. Po wykonaniu łączenia w oknie ***Propensity Score Matching – wyniki*** możemy wygenerować wyniki naszej analizy.

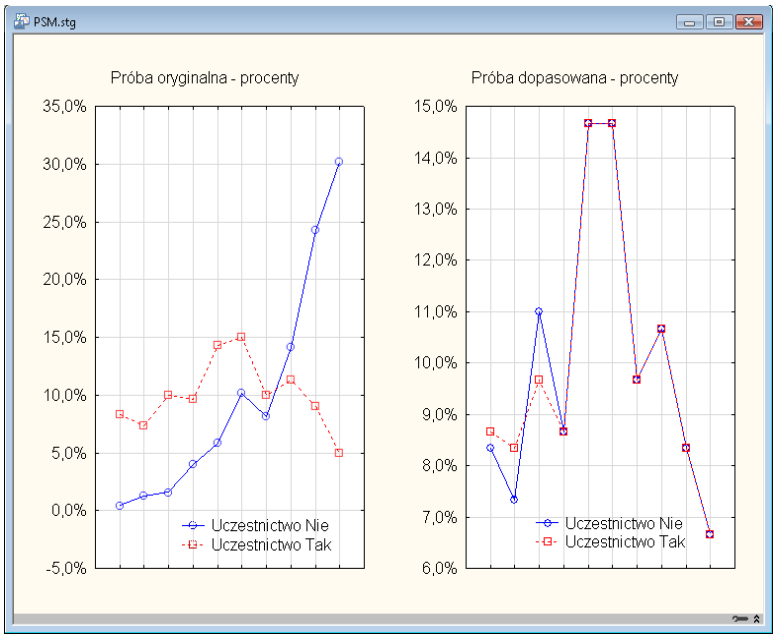


Za pomocą przycisku ***Arkusz danych*** znajdującego się w obszarze ***Wyniki*** możemy wygenerować arkusz zawierający przypadki z grupy interwencja oraz przypisane im przypadki z grupy brak interwencji. Jeżeli włączona została opcja ***Ze zwracaniem*** i w jej wyniku dany przypadek z grupy brak interwencji został wskazany wielokrotnie, to będzie on reprezentowany przez wiele wierszy w zbiorze danych. Arkusz ten może być podstawą do bardziej szczegółowych analiz za pomocą ogólnych narzędzi dostępnych w programie *STATISTICA*.

Za pomocą przycisku ***Łączenie*** otrzymujemy raport informujący, które przypadki z grupy *brak interwencji* zostały połączone z danymi przypadkami z grupy *interwencja*. Przycisk ***Statystyki*** pozwala nam ocenić i porównać wartość mierzonego efektu w zbiorze pierwotnym i po dopasowaniu. Na podstawie poniższego wykresu możemy stwierdzić, że po usunięciu obciążenia za pomocą metody *propensity score matching* oceniana wielkość efektu związana z uczestnictwem w programie wzrosła w porównaniu z próbą oryginalną.

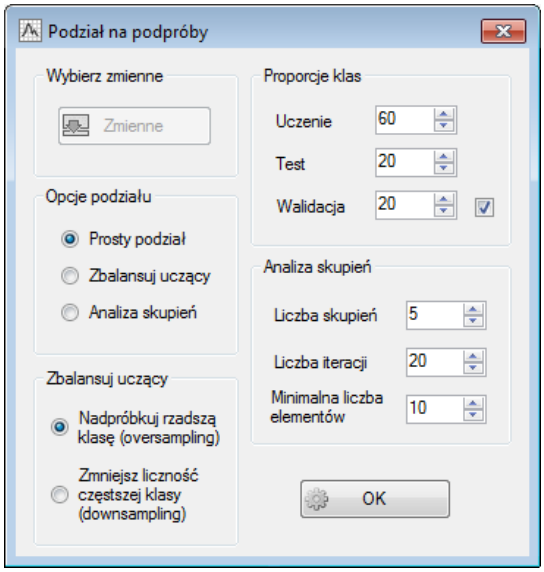


Dopasowanie z kolei pozwala ocenić za pomocą szeregu wykresów i raportów jakość dopasowania na podstawie rozkładu wartości *propensity score* w próbie oryginalnej i po procedurze łączenia.



4.4. Podział na podpróby

Moduł *Podział na podpróby* pozwala na przygotowanie prób uczącej, testowej oraz (opcjonalnie) walidacyjnej na potrzeby budowy modeli predykcyjnych (klasyfikacyjnych lub regresyjnych).



Moduł oferuje trzy główne opcje podziału zbioru na podpróby:

- *Prosty podział* umożliwia podzielenie zbioru danych w sposób losowy na klasy uczącą, testową oraz walidacyjną (opcjonalnie) zgodnie z proporcjami określonymi w grupie Proporcje klas.
- *Zbalansuj uczący* pozwala na przygotowanie próby uczącej zawierającej równe proporcje klas zmiennej zależnej. Opcja ta jest przydatna w sytuacji, gdy problem, jaki chcemy analizować jest zadaniem klasyfikacyjnym, z dwoma klasami zmiennej zależnej. W zależności od wyboru opcji w grupie Zbalansuj uczący, program dokona nadpróbkowania (*oversampling*) rzadszej klasy lub zmniejszy liczebność częstszej klasy (*downsampling*).
- *Analiza skupień* pozwala dokonać podziału na podpróby na podstawie losowania przypadków ze skupień utworzonych w wyniku analizy k-średnich dla wybranych predyktorów. Metoda

jest przydatna zwłaszcza w sytuacji mniejszych zbiorów danych, pozwala uniknąć nieproporcjonalnego rozłożenia się danej klasy przypadków w podpróbach.

W wyniku analizy badacz otrzymuje nowy zbiór danych, który oprócz pierwotnego zestawu zmiennych zawiera dodatkowo kolumnę informującą do jakiej klasy (uczenie, test, walidacja) trafił konkretny przypadek.

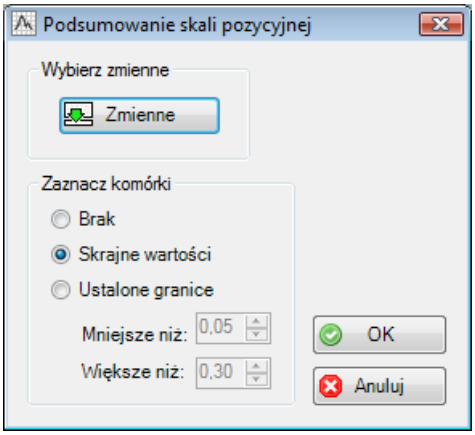
5. Podsumowanie skal

Kolejna grupa modułów – ***Podsumowanie skal*** umożliwia szybkie wygenerowanie raportu dla danych będących wynikiem pomiaru na skalach złożonych. Grupa zawiera dodatkowo moduł oceniający *rzetelność skali Likerta* a także obliczyć szereg wskaźników zgodności sędziów. Opis modułów znajduje się poniżej.

5.1. Podsumowanie skali pozycyjnej

Umożliwia utworzenie zestawienia rozkładów procentowych wartości wybranych zmiennych przedstawionych na skali pozycyjnej. W arkuszu wynikowym dodatkowo mogą zostać zaznaczone kolorami wartości najczęściej i najrzadziej występujące w każdej ze zmiennych wchodzących w skład skali pozycyjnej lub występujące częściej bądź rzadziej od podanych wartości granicznych.

Przykład. Przykładową analizę wykonamy, wykorzystując plik *Pozycyjna.sta* zawierający przykładowe oceny respondentów określające częstotliwość występowania pewnych cech na skali pozycyjnej. Po otwarciu pliku z menu ***Analizy marketingowe i rynkowe*** | ***Podsumowanie skal*** wybieramy pozycję ***Skala pozycyjna***, przywołując okno o tej samej nazwie. Następnie za pomocą przycisku ***Zmienne*** wybieramy zmienne do analizy – będą to wszystkie zmienne zawarte w zbiorze danych.



W obszarze ***Zaznacz komórki*** możemy wybrać opcję kolorowania skrajnych wartości lub wartości przekraczających ustalone granice. My wybierzemy opcję ***Skrajne wartości***, a następnie zatwierdzimy analizę za pomocą przycisku ***OK***.

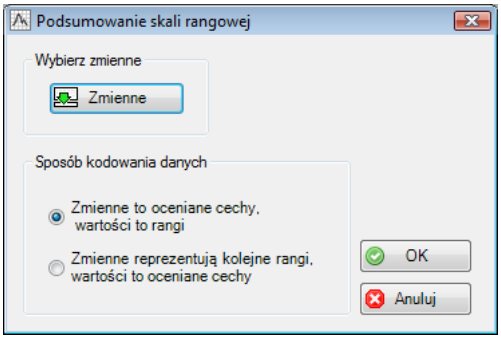
	Podsumowanie skali pozycyjnej					
	Stale	Bardzo często	Często	Czasami	Nigdy	Liczba odpowiedzi
Cecha A	25,0%	20,0%	20,0%	20,0%	15,0%	20
Cecha B	25,0%	30,0%	20,0%	10,0%	15,0%	20
Cecha C	30,0%	15,0%	35,0%	10,0%	10,0%	20
Cecha D	50,0%	5,0%	15,0%	15,0%	15,0%	20

Otrzymany arkusz pozwala ocenić częstość wyboru poszczególnych pozycji na skali oraz łatwo zidentyfikować wartości skrajne.

5.2. Podsumowanie skali rangowej

Umożliwia użytkownikowi wykonanie rankingu wartości wchodzących w skład skali rangowej oraz liczby wskazań danego obiektu na każdej pozycji. Analiza obsługuje dwa sposoby kodowania (zmienne oznaczają oceniane cechy, a wartości przypisane rangi, lub zmienne reprezentują kolejne rangi, a wartości oznaczają oceniane cechy).

Przykład. Podsumowanie skali rangowej wykonamy na podstawie zbioru *Napoje.sta* zawierającego oceny różnego rodzaju napojów. Najniższa wartość oznacza najwyższą ocenę respondenta. Aby wykonać analizę zgromadzonych danych z menu ***Analizy marketingowe i rynkowe*** | ***Podsumowanie skal***, wybieramy pozycję ***Skala rangowa***, przywołując okno o tej samej nazwie. Za pomocą przycisku ***Zmienne*** wybieramy wszystkie zmienne do analizy, a następnie zatwierdzamy wykonanie analizy, klikając przycisk ***OK***.



Uwaga! *Poza kodowaniem danych zapisanym w pliku Napoje.sta, gdzie zmienne to oceniane cechy, a wartości to rangi, istnieje możliwość innego sposobu kodowania, w którym zmienne reprezentują kolejne rangi, natomiast wartościami są oceniane cechy. Przykładowym zbiorem zawierającym takie kodowanie jest plik Napoje_2.sta.*

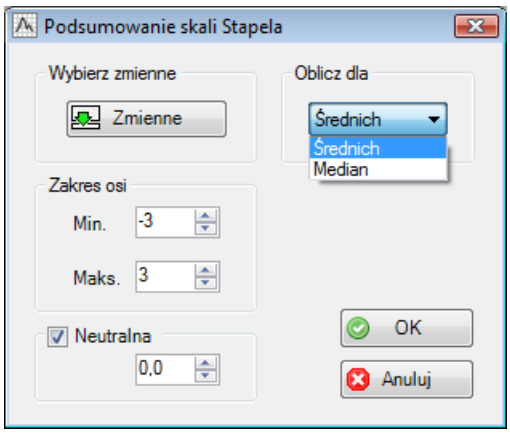
Po wykonaniu analizy uzyskujemy skoroszyt zawierający arkusz z podsumowaniem skali rangowej.

	Podsumowanie skali rangowej							
	Miejsce 1	Miejsce 2	Miejsce 3	Miejsce 4	Miejsce 5	Miejsce 6	Miejsce 7	Ranking
Napoje gazowane	4	5	0	0	0	1	0	60
Typu Cola	2	2	1	0	1	2	2	40
Kompot	0	2	2	3	1	1	1	40
Woda	0	0	3	5	1	1	0	40
Soki	3	0	0	1	4	0	2	39
Kefir	1	0	1	1	2	4	1	31
Herbata mrozona	0	1	3	0	1	1	4	30

Analizując uzyskane oceny możemy zauważyć, iż najwyżej ocenione zostały *Napoje gazowane*, natomiast najniżej *Herbata mrożona*.

5.3. Wykres dla skali Stapela

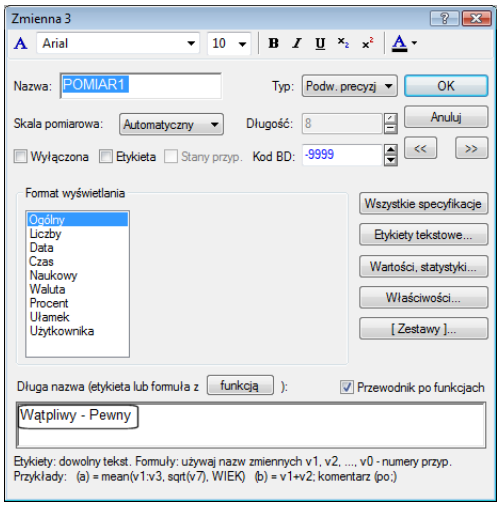
Umożliwia utworzenie wykresu podsumowującego średnie lub mediany wartości wskazanych zmiennych wyrażonych za pomocą skali Stapela. Dodatkowo istnieje możliwość wskazania zmiennej grupującej.



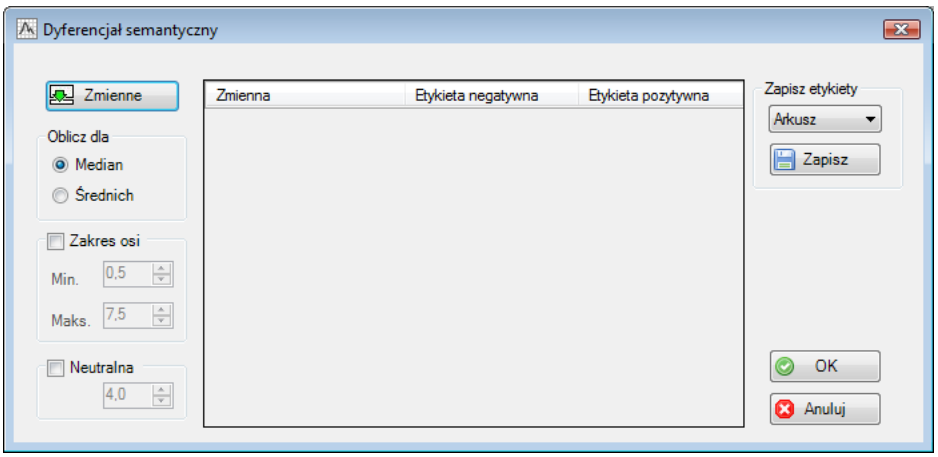
5.4. Wykres dyferencjału semantycznego

Skala dyferencjału semantycznego jest najczęściej zestawem szeregu prostych skal dwubiegunowych. W praktyce badań marketingowych technika ta stanowi zestaw skal szacunkowych do badań emocjonalnie nacechowanych ocen różnych obiektów (marek produktów, sloganów reklamowych, nazw firm). Przygotowanie skali wymaga opracowania par opozycyjnych terminów opisujących badane zjawisko, np. *dobry-zły*, *szybki-wolny*, *silny-słaby*. Następnie nasilenie każdej cechy jest oceniane na przykład od 1 do 7, gdzie 1 oznacza maksymalne nasilenie cechy pierwszej (np. negatywnej), 7 cechy przeciwnej (np. pozytywnej), a 4 stanowi punkt neutralny. Oceny respondentów są następnie uśredniane, bądź wyliczana jest mediana ocen, a wyniki nanoszone są na wykres dyferencjału semantycznego, zwany także profilem polaryzacji.

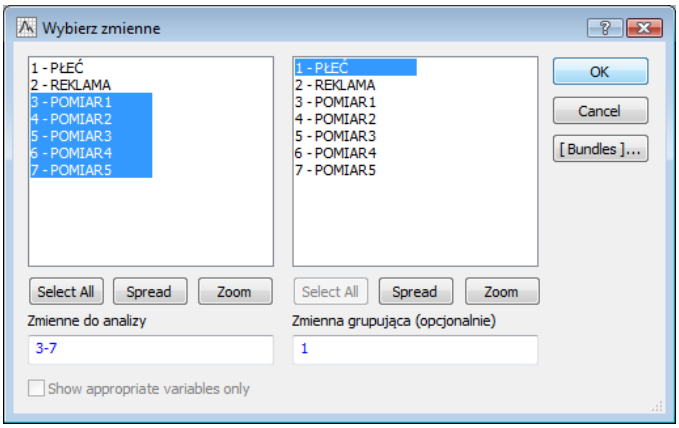
Przykład. Obliczenia wykresu dyferencjału semantycznego wykonamy na podstawie pliku *Dyferencjał.sta* zawierającego nieco zmodyfikowany podzbiór arkusza *Adstudy.sta*. Zbiór ten zawiera dwie zmienne grupujące *PŁEĆ* i *REKLAMA* oraz pięć zmiennych zawierających oceny nasilenia kolejno pięciu cech opisujących pewne zjawisko (*POMIAR1* do *POMIAR5*). Klikając dwukrotnie na nagłówku zmiennej *POMIAR1*, wchodzimy do okna specyfikacji zmiennej. W oknie specyfikacji w obszarze *Długa nazwa* widzimy parę opozycyjnych terminów (*Wątpliwy - Pewny*) oddzielonych myślnikiem.



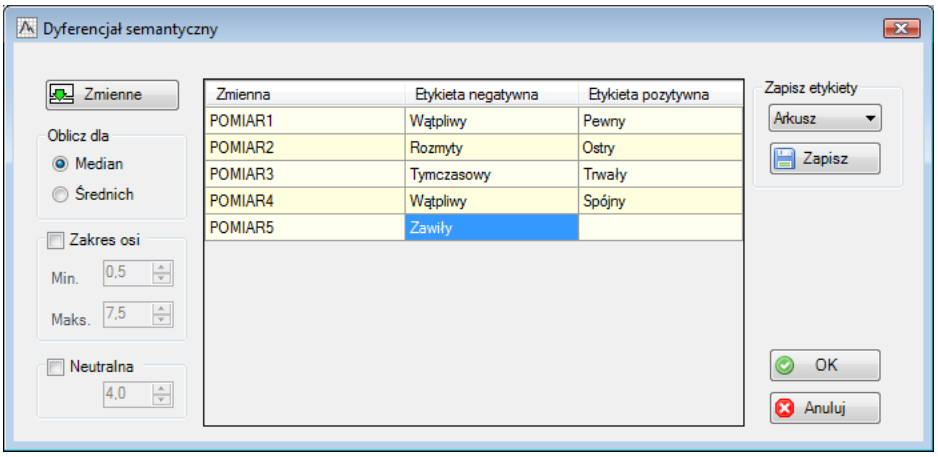
Podobna sytuacja ma miejsce w zmiennej *POMIAR2* oraz *POMIAR3*. Wprowadzimy kolejną parę terminów *Niespójny-Spójny* do zmiennej *POMIAR4*, pamiętając, by ocena negatywna znalazła się na pierwszym miejscu. Następnie z menu ***Analizy marketingowe i rynkowe*** | ***Podsumowanie skal*** wybieramy opcję ***Dyferencjał semantyczny***, przywołując okno ***Dyferencjał semantyczny***.



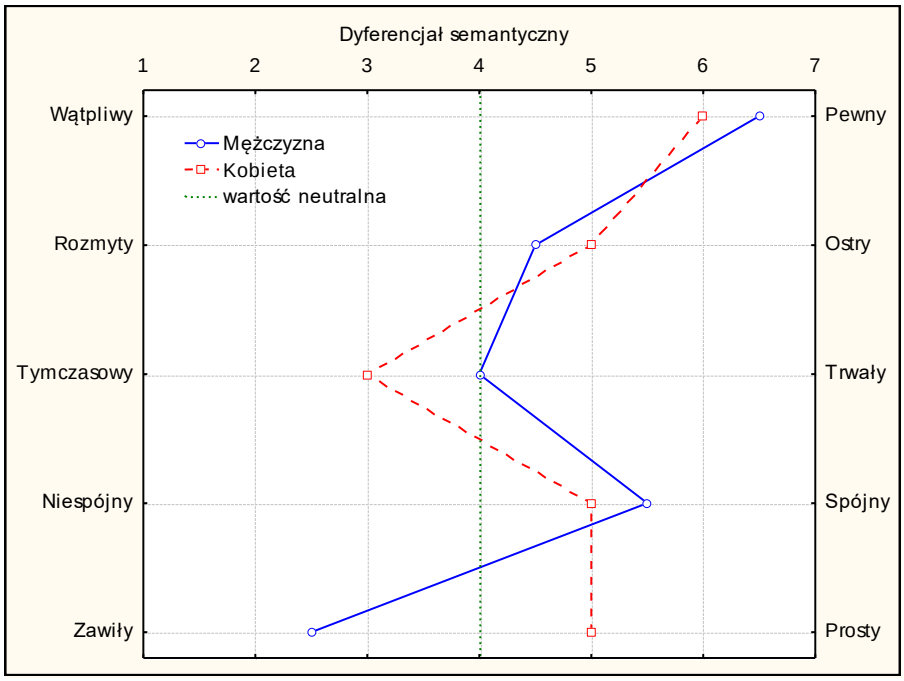
W oknie tym klikamy przycisk ***Zmienne***, aby wybrać zmienne do analizy. Analizować będziemy zmienne *POMIAR* (1 do 5) w podziale na *PŁEĆ* respondentów.



Po zatwierdzeniu wyboru zmiennych wracamy do okna ***Dyferencjał semantyczny***, w którym, w tabeli w środkowej części okna wyświetlone zostały oceny dla zmiennych 1 do 4. Ponieważ zmienna *POMIAR5* nie zawierała w swojej specyfikacji ocen, wprowadzimy je teraz bezpośrednio, w tabeli. Będą to oceny *Zawiły – Prosty*.

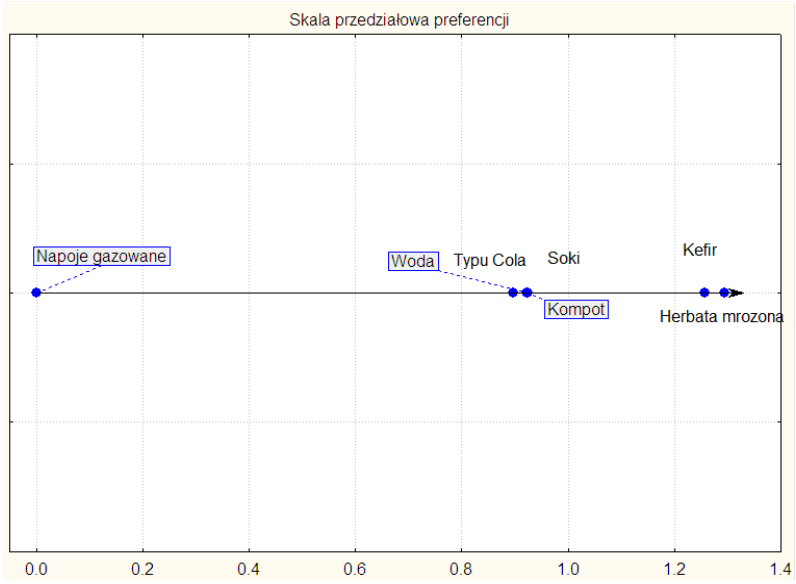


Po wprowadzeniu ocen możemy dopisać je do arkusza za pomocą przycisku ***Zapisz***. Możemy też utworzyć makro, które umożliwi dopisanie analogicznych etykiet do zbioru o tych samych nazwach zmiennych. W kolejnym kroku zaznaczymy obszar ***Zakres osi*** i wprowadzimy wartości ***Min.*** – 1 oraz ***Maks.*** – 7, dodatkowo podamy wartość neutralną równą 4. Następnie zatwierdzamy ustawienia analizy, klikając ***OK***, generując tym samym skoroszyt zawierający wykres dyferencjału semantycznego.



5.5. Metoda ocen porównawczych Thurstone'a

Moduł umożliwia zbudowanie metrycznej skali preferencji na podstawie danych o preferencjach uzyskanych z wykorzystaniem skali porównań parami bądź skali rangowej (jest ona przekształcana do skali porównań parami).

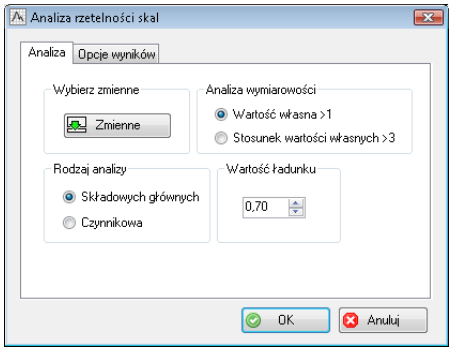


Dodatkowo uzyskane wyniki można zobrazować za pomocą mapy percepcji utworzonej metodą skalowania wielowymiarowego.

Szczegółowy opis tego modułu oraz przykłady jego wykorzystania dostępne są w artykule A. Sagana *Analiza preferencji konsumentów z wykorzystaniem programu STATISTICA – analiza conjoint i skalowanie wielowymiarowe*, dołączonym do niniejszej dokumentacji (zbiory danych użyte w artykule są dostępne w katalogu z pozostałymi plikami danych).

5.6. Analiza rzetelności skal

Moduł wykonuje analizę rzetelności dla danych mierzonych na skali *Likerta*. Moduł w pierwszej kolejności na podstawie analizy czynnikowej bądź głównych składowych (z rotacją *varimax znormalizowana*) dokonuje oceny wymiarowości analizowanego zbioru. Stosuje przy tym jedno z dostępnych kryteriów : ***Wartość własna >1*** lub ***Stosunek wartości własnych >3*** oraz wskazaną przez użytkownika wartość ładunku decydującą czy dana pozycja będzie wchodziła w skład skali.

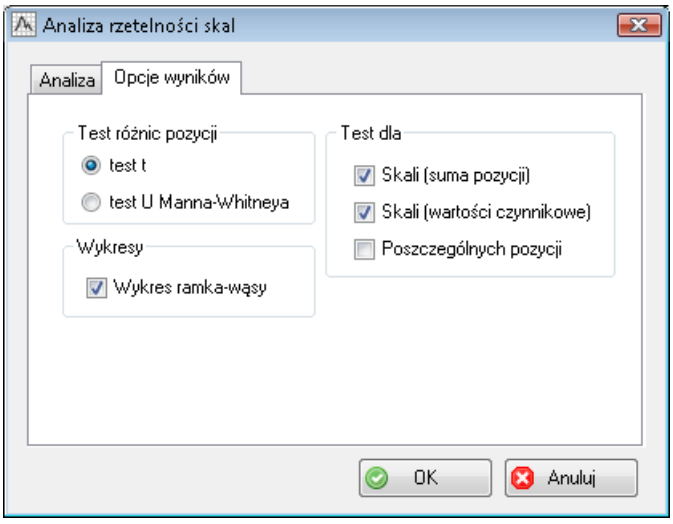


Następnie dla określonych w ten sposób wymiarów liczone są wskaźniki rzetelności *alfa Cronbacha*, *rho Joreskoga* oraz *theta Armora*. Na tym etapie następuje ewentualne usunięcie niektórych pozycji skali, jeżeli ich usunięcie zwiększa wskaźnik rzetelności *alfa Cronbacha*.

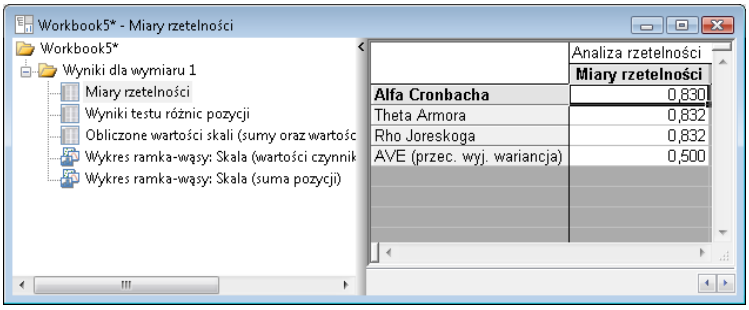
Po określeniu liczby wymiarów oraz pozycji, które wchodzą w ich skład oraz obliczeniu rzetelności, ostatnim etapem analizy jest ocena mocy dyskryminacyjnej dla wymiarów. Dla skali rozumianej jako wartości czynnikowe danego wymiaru lub jako suma pozycji ładujących dany wymiar dokonujemy podziału przypadków względem górnego oraz dolnego kwartyla a następnie za pomocą testu (parametrycznego, bądź nieparametrycznego) dokonujemy oceny istotności różnic wartości skali w uzyskanych grupach. Uzyskanie istotnych statystycznie różnic świadczy o dobrej mocy dyskryminacyjnej danego wymiaru.

Przykład analizy danych za pomocą tego modułu przedstawimy na podstawie zbioru danych *10Items.sta* dostępnego w zbiorze plików przykładowych programu *STATISTICA*.

Z menu ***Analizy marketingowe i rynkowe*** z grupy ***Podsumowanie skal*** wybieramy opcję ***Rzetelność skali*** wyświetlając okno ***Analiza rzetelności skal***. W oknie tym klikamy przycisk ***Zmienne***, a następnie w oknie ***Wybierz zmienne*** wybieramy wszystkie zmienne do analizy. Pozostawiamy wszystkie opcje programu na niezmienionym poziomie, jedynie na karcie ***Opcje wyników*** w obszarze ***Wykresy*** zaznaczamy opcję ***Wykres ramka-wąsy***.



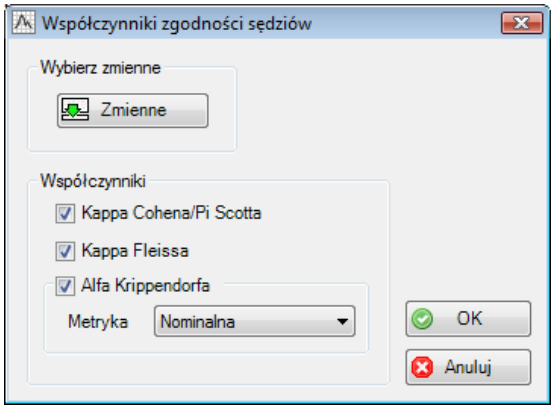
Po kliknięciu przycisku ***OK*** wykonana zostaje analiza wymiarowości, rzetelności skal oraz mocy dyskryminacyjnej zgodnie z opisanym powyżej scenariuszem. Na podstawie wprowadzonych danych program wyodrębnił jeden wymiar (skalę) składający się z pięciu pozycji. Poniżej zamieszczono przykładowe wyniki uzyskane na podstawie analizy.



5.7. Współczynniki zgodności sędziów

Umożliwiają określenie zgodności pomiędzy ocenami tych samych obiektów pochodzącymi z różnych źródeł. Pierwsze dwie miary *Kappa Cohena* oraz *Pi Scotta* umożliwiają obliczenie wskaźnika zgodności pomiędzy dokładnie dwoma sędziami przy założeniu, że oceny są wyrażane na skali nominalnej. *Kappa Fleissa* rozszerza możliwość obliczania wskaźnika dla więcej niż dwóch sędziów. *Alfa Krippendorfa* pozwala dodatkowo na uwzględnienie innych skal niż nominalna (przy dowolnej liczbie sędziów).

Przykład. Obliczenia współczynników zgodności sędziów wykonamy na podstawie zbioru *Sedziowie.sta*. Z menu ***Analizy marketingowe i rynkowe*** | ***Podsumowanie skal*** wybieramy polecenie ***Współczynniki zgodności sędziów***, przywołując okno o tej samej nazwie.



Klikamy przycisk ***Zmienne***, aby wybrać zmienne do analizy, a następnie wybieramy wszystkie zmienne. Ponieważ w zbiorze danych znajdują się oceny czterech sędziów, zaznaczamy jedynie opcje: *Kappa Fleissa* oraz *Alfa Krippendorfa* (dwa pierwsze wskaźniki umożliwiają obliczenie zgodności dla dokładnie dwóch sędziów). Klikamy ***OK***, aby zatwierdzić wykonanie analizy, otrzymując raport z wyliczonymi wskaźnikami zgodności.

	1 Kappa Fleissa	2 Alfa Krippendorfa
1	0,07	0,09

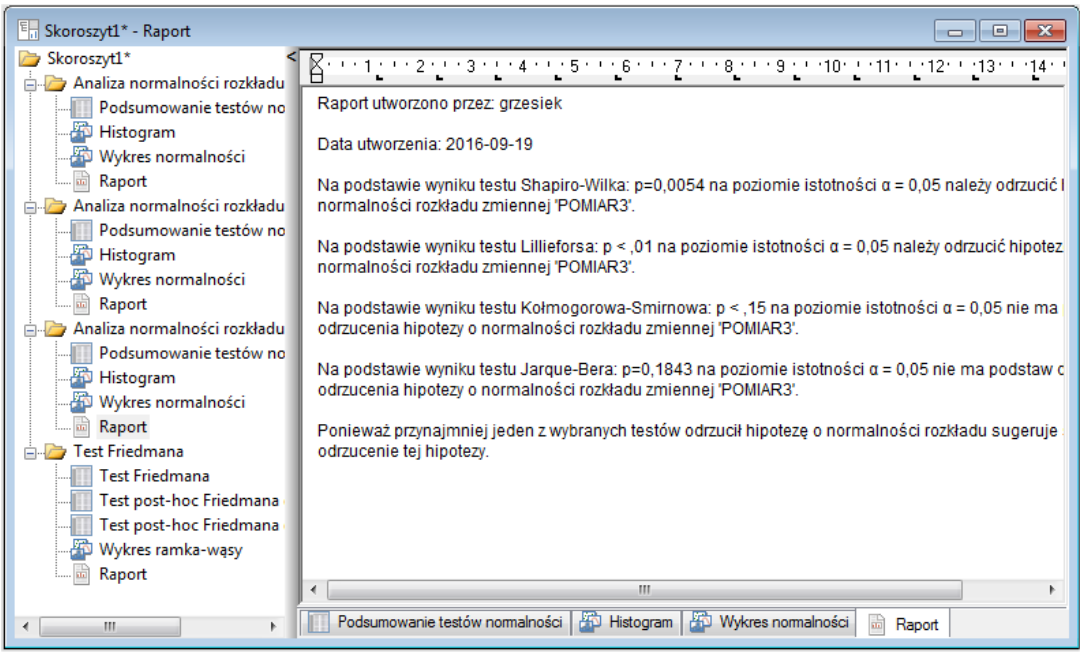
Uwaga! *Jeśli chcielibyśmy interpretować oceny sędziów na skalach mocniejszych niż nominalna, z listy rozwijalnej Metryka wybieramy odpowiednią dla naszych danych skalę pomiaru.*

6. Kreator testów statystycznych

Kreator testów statystycznych przeznaczony jest dla osób pragnących zweryfikować prawdziwość swojej hipotezy badawczej za pomocą testu statystycznego, mających jednocześnie trudności z określeniem testu, który byłby najbardziej odpowiedni w ich sytuacji. Kreator automatycznie sprawdza wszelkie założenia związane z danym typem problemu i w zależności od ich spełnienia proponuje poprawny test. Korzystając z tego narzędzia badacz musi jedynie określić kwestie merytoryczne prowadzonej analizy, takie jak:

- *Jaką analizę chcemy przeprowadzić?*
- *Czy badane próby są zależne/niezależne od siebie?*
- *Ile grup analizujemy?*
- *Na jakiej skali mierzone są badane cechy?*

Wynikiem działania programu jest skoroszyt zawierający wyniki poszczególnych testów (dot. założeń jak i głównego pytania) razem z interpretacją oraz wykresy i dodatkowe analizy generowane standardowo przy danym rodzaju badań.



Z powyższego opisu można odnieść wrażenie, że kreator testów jest swoistą „czarną skrzynką" do której możemy „wrzucić" dane, powiedzieć co chcemy uzyskać, kliknąć 'Uruchom analizę' i uzyskać wynik. Tak rzeczywiście jest, ale nie do końca. Głównym zadaniem badacza podczas przeprowadzania analizy jest wskazanie aspektów merytorycznych i wybór zmiennych. W tym miejscu rzeczywiście można po prostu uruchomić analizę i uzyskać pełen raport ze wszystkimi możliwymi wynikami, jednakże możemy również każdą analizę dostosować do naszych potrzeb. Możemy wybrać, które z możliwych testów chcemy zastosować, bądź też które wyniki nas interesują. Program automatycznie zapamiętuje wybrane opcje. Dodatkowym atutem kreatora testów jest możliwość wyeksportowania wyników do pliku MS Word.

W tym miejscu należałoby również zaznaczyć, że kreator testów ze względu na cel, który mu przyświeca, tj. pewna uniwersalność i prostota użytkowania, opiera się tylko na jednym z możliwych schematów testowania. Obsługuje on najczęściej występujące problemy analityczne. Bardziej złożone analizy czy też pewne dodatkowe opcje narzędzi tu użytych dostępne są w innych miejscach programu Statistica.

6.1. Możliwości programu

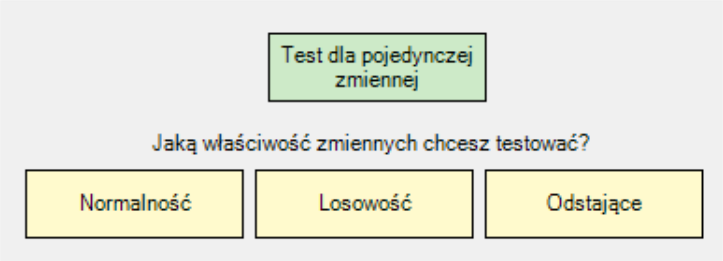
W obecnej wersji *Kreator testów statystycznych* umożliwia wykonanie dwóch rodzajów analiz:

- Testy dla pojedynczej zmiennej
- Badanie istotności różnic.

Wybranie jednego z opcji spowoduje wyświetlenie schematu w postaci drzewa, dzięki któremu badacz w prosty, intuicyjny sposób może określić jaki dokładnie typ analizy chce przeprowadzić.

6.2. Testy dla pojedynczej zmiennej

W celu przejścia do schematu dotyczącego analizy pojedynczej zmiennej wystarczy kliknąć w odpowiednie pole rodzaju analiz. Na ekranie pojawi się poniższy schemat.



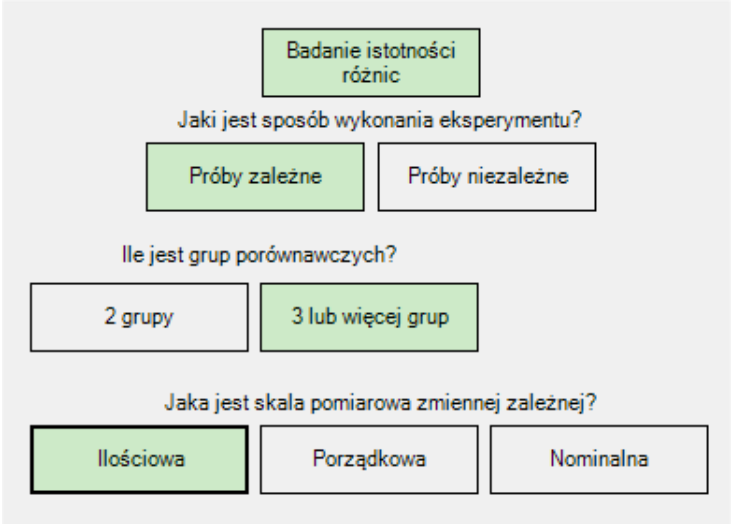
Jak można zauważyć, w chwili obecnej program oferuje trzy typy testów dla pojedynczej zmiennej:

- Normalność
- Losowość
- Obserwacje odstające.

W celu przeprowadzenia analizy wystarczy wybrać jeden z nich. Zostanie wtedy odblokowana możliwość wyboru zmiennych oraz wykonania analizy.

6.3. Badanie istotności różnic

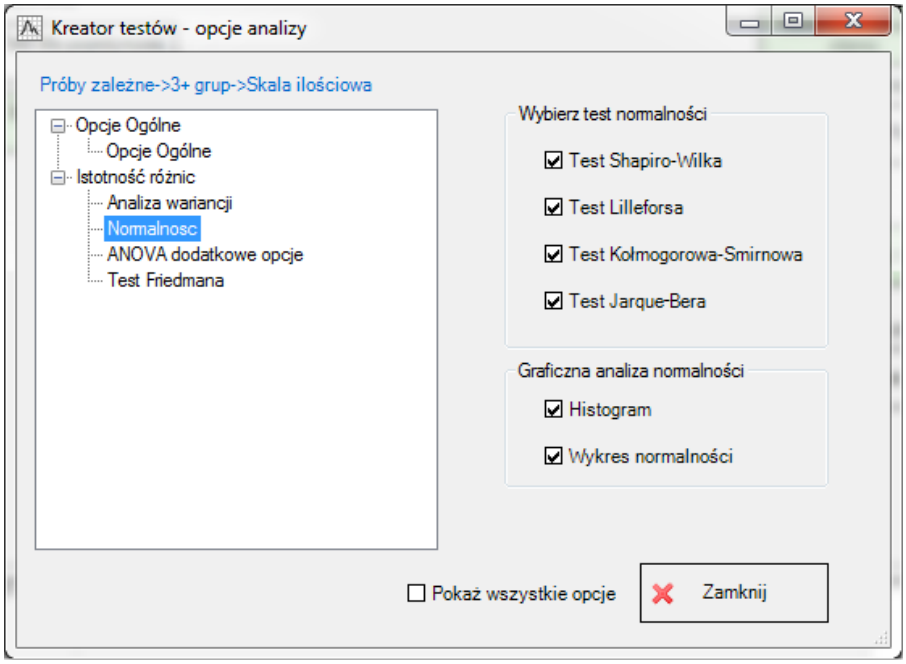
Po wybraniu tego rodzaju analizy, podobnie jak w przypadku analizy dla jednej zmiennej, pojawi się schemat ułatwiający dokładne określenie właściwej ścieżki postępowania. Kreator będzie prowadził użytkownika krok po kroku, zadając mu kolejne pytania pozwalające doprecyzować rodzaj analizowanego problemu. Przykładowy schemat decyzyjny zamieszczono poniżej.



Po udzieleniu odpowiedzi na powyższe pytania, użytkownik musi jedynie wybrać zmienne i uruchomić analizę. Program dokona sprawdzenia założeń związanych z daną klasą problemu, wybierze odpowiedni test i wygeneruje komplet wyników wraz z ich interpretacją.

6.4. Dodatkowe możliwości programu

Jak już wcześniej nadmieniono, program przeznaczony jest dla badaczy mających nieco mniejsze doświadczenie w analizie statystycznej. Parametry programu są zatem dostosowane do najbardziej typowych sytuacji. Bardziej doświadczeni analitycy mają jednakże możliwość określenia szeregu szczegółowych opcji dotyczących wyboru i konfiguracji testów.

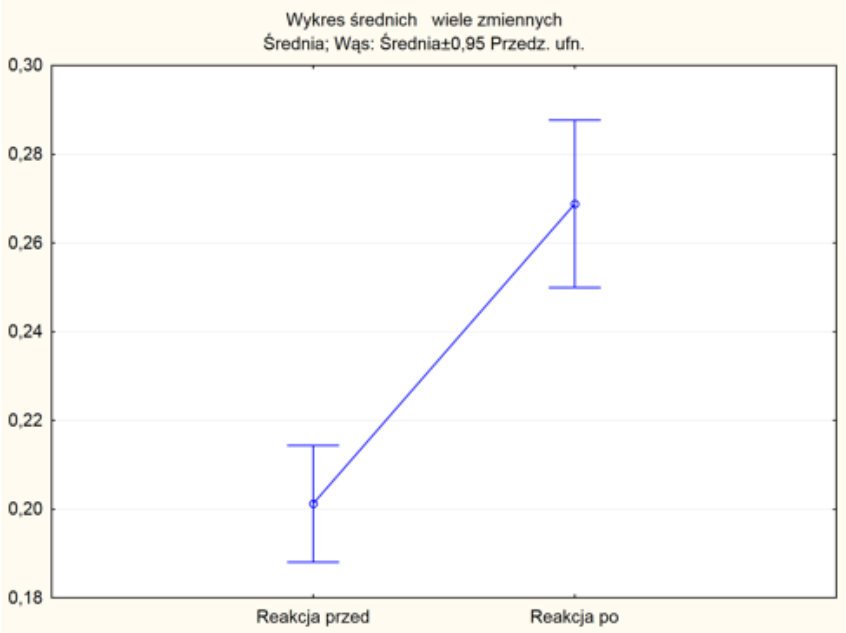


Ważnym atutem kreatora jest niewątpliwie możliwość zapisu raportu z analizy zarówno w formacie Statistica jak i MS Word. Poniżej fragment przykładowego raportu w formacie MS Word.

2.3. Statystyki opisowe

Zmienna	N ważnych	Średnia	Ufność - 95,000%	Ufność 95,000%	Odch.std
Reakcja przed	25,00	0,20	0,19	0,21	0,03
Reakcja po	25,00	0,27	0,25	0,29	0,05

2.4. Wykres interakcji



2.5. Interpretacja

Na podstawie wyniku testu t-Studenta dla prób zależnych: $p < 0,0001$ na poziomie istotności $\alpha = 0,05$ należy odrzucić hipotezę o równości średnich wartości zmiennej 'Reakcja przed' i zmiennej 'Reakcja po'.

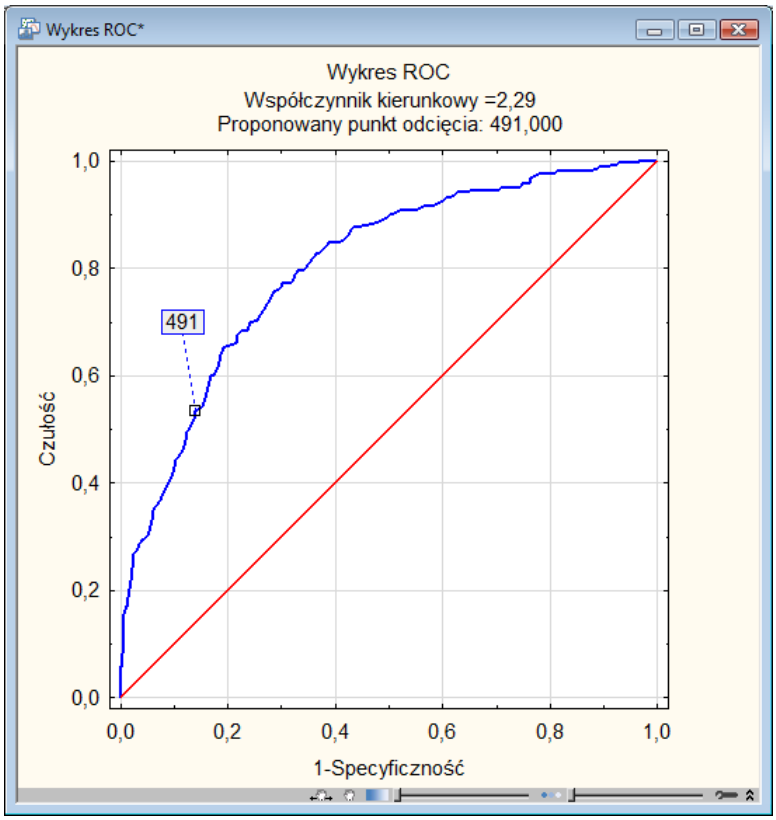
7. Analizy

Grupa ***Analizy*** zawiera zestaw sześciu modułów analitycznych umożliwiających wykonanie zaawansowanych analiz. Użytkownik ma możliwość wykonania *Krzywych ROC* czy też metod: *Conjoint*, *Aglomeracji*, *PROFIT*, *Uogólnionej PCA* oraz *Porządkowania liniowego*. Poniżej znajduje się opis poszczególnych modułów zawartych w tej grupie.

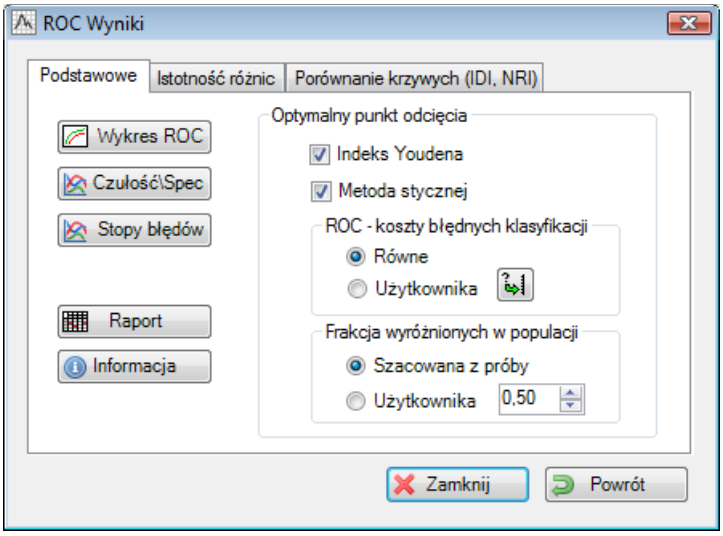
7.1. Krzywe ROC

Krzywa ROC (*Receiver Operating Characteristic*) jest narzędziem służącym do oceny poprawności klasyfikatora (pojedynczej zmiennej lub całego modelu), umożliwia ona łączny opis jego czułości i specyficzności. Ten sposób wspomagania procesu decyzyjnego jest szeroko stosowany w różnych obszarach analizy danych, m.in. w diagnostyce medycznej.
Moduł ***Krzywe ROC*** umożliwia:

- kreślenie krzywych ROC dla prób zależnych i niezależnych,
- obliczanie pola powierzchni pod krzywą,
- porównywanie istotności różnicy pól pomiędzy dwiema krzywymi,
- ustalanie optymalnego punktu odcięcia za pomocą indeksu Youdena oraz metodą stycznej dla podanych kosztów błędnej klasyfikacji i prawdopodobieństw a priori występowania badanego zjawiska,



- obliczanie miar FP,TP FN, FP, Sensitivity, Specificity, ACC, PPV, NPV, *False positive ratio*, *False negative ratio*, LR dla wszystkich możliwych punktów odcięcia,
- wykresy czułości i specyficzności,
- wskaźniki *IDI* oraz *NRI* służące do porównania krzywych.



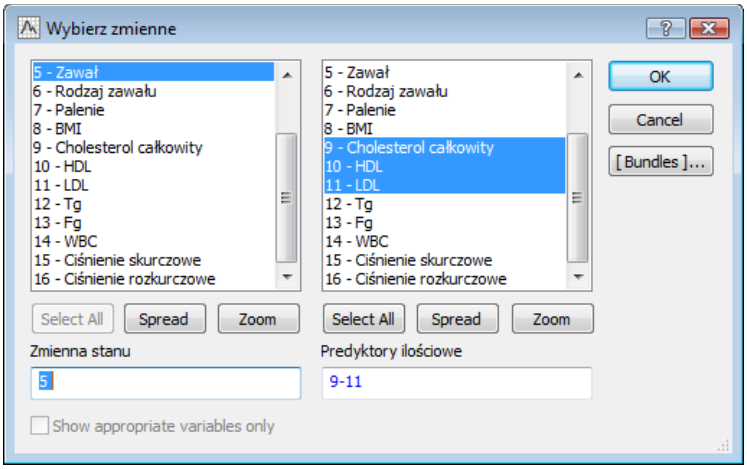
Przykład. Otwieramy plik zawierający dane do analizy (w naszym przypadku plik *Zawały.sta*), a następnie z menu ***Analizy marketingowe i rynkowe*** | ***Analizy*** wybieramy opcję ***Analiza ROC***.

Dane: Zawały (16 zmn. * 239 prz.)

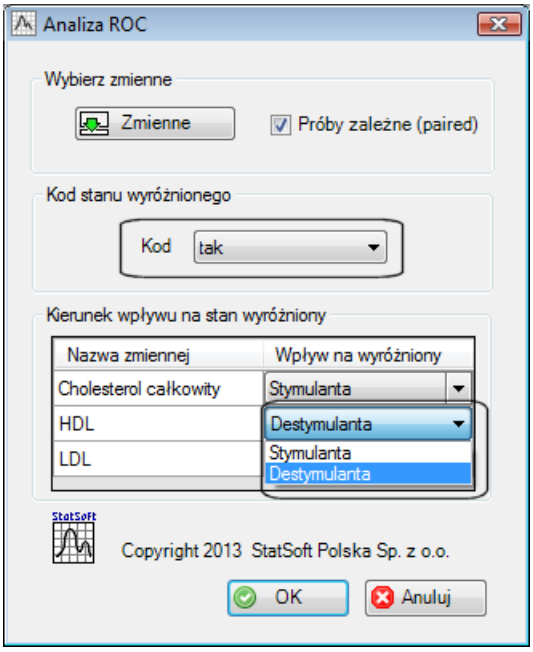
Dane dotyczą wybranych parametrów biochemicznych oraz klinicznych zebranych dla pacjentów z chorobą niedokrwienną serca.
Źródło: Watała C., Biostatystyka - wykorzystanie metod statystycznych w pracy badawczej w naukach biomedycznych, α-medica press 2002.

	1 Płeć	2 Wiek	3 Choroba wiencowa	4 Czas choroby wiencowej	5 Zawał	6 Rodzaj zawału	7 Palenie	8 BMI	9 Cholesterol całkowity
KS/95/0004	Kobieta	32	tak	poniżej 2 m-cy	nie	bez zawału	pali	18,65	205
KS/95/0007	Mężczyzna	32	tak	poniżej 2 m-cy	nie	bez zawału	pali	18,65	205
KS/95/0014	Kobieta	49	tak	od 2 do 12 m-cy	nie	bez zawału	nie pali	29,74	272
KS/95/0016	Mężczyzna	67	tak	od 2 do 12 m-cy	nie	bez zawału	pali	28,84	208
KS/95/0018	Kobieta	63	tak	powyżej 12 m-cy	nie	bez zawału	nie pali	33,20	205
KS/95/0021	Kobieta	64	tak	poniżej 2 m-cy	nie	bez zawału	nie pali	18,62	209
KS/95/0023	Mężczyzna	47	tak	powyżej 12 m-cy	tak	pełnościenny	pali	34,68	272
KS/95/0027	Mężczyzna	61	tak	poniżej 2 m-cy	nie	bez zawału	pali	26,53	215
KS/95/0028	Mężczyzna	84	nie		nie	bez zawału	nie pali		177
KS/95/0031	Kobieta	52	tak	powyżej 12 m-cy	nie	bez zawału	nie pali	23,51	204
KS/95/0034	Kobieta	63	tak	powyżej 12 m-cy	nie	bez zawału	nie pali	30,11	234
KS/95/0039	Kobieta	49	tak	od 2 do 12 m-cy	nie	bez zawału	pali	26,30	212
KS/95/0040	Mężczyzna	68	nie		nie	bez zawału	nie pali	22,55	156
KS/95/0046	Mężczyzna	32	tak	poniżej 2 m-cy	tak	pełnościenny	pali	29,39	282
KS/95/0048	Mężczyzna	64	tak	powyżej 12 m-cy	nie	bez zawału	pali	30,42	228
KS/95/0049	Mężczyzna	67	tak	powyżej 12 m-cy	nie	bez zawału	pali	22,60	200
KS/95/0050	Kobieta	51	nie		nie	bez zawału	pali	24,96	148
KS/95/0053	Mężczyzna	51	nie		nie	bez zawału	pali	37,74	216
KS/95/0055	Mężczyzna	60	nie		nie	bez zawału	pali	27,68	216

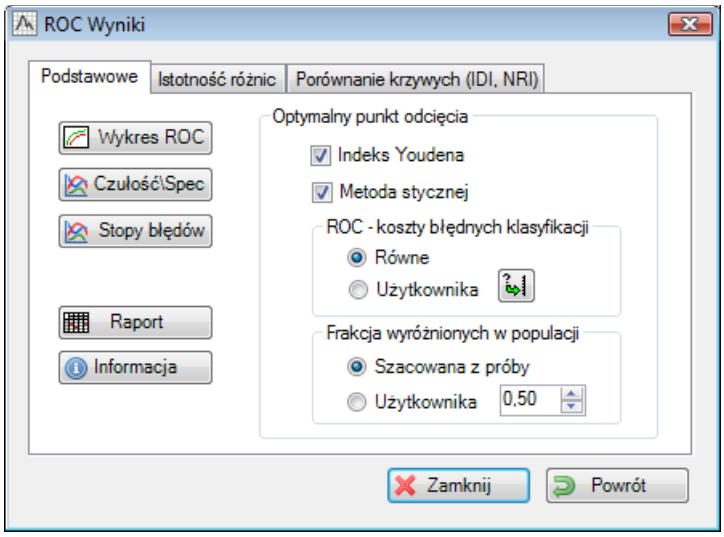
Kliknięcie opcji ***Analiza ROC*** spowoduje wyświetlenie okna o tej samej nazwie. Aby wybrać zmienne do analizy, klikamy przycisk ***Zmienne*** i określamy ***Zmienną stanu*** (w naszym przypadku *Zawał*) oraz ***Predyktory ilościowe*** (*Cholesterol całkowity*, *HDL* i *LDL*), po czym potwierdzamy wybór, klikając ***OK***.



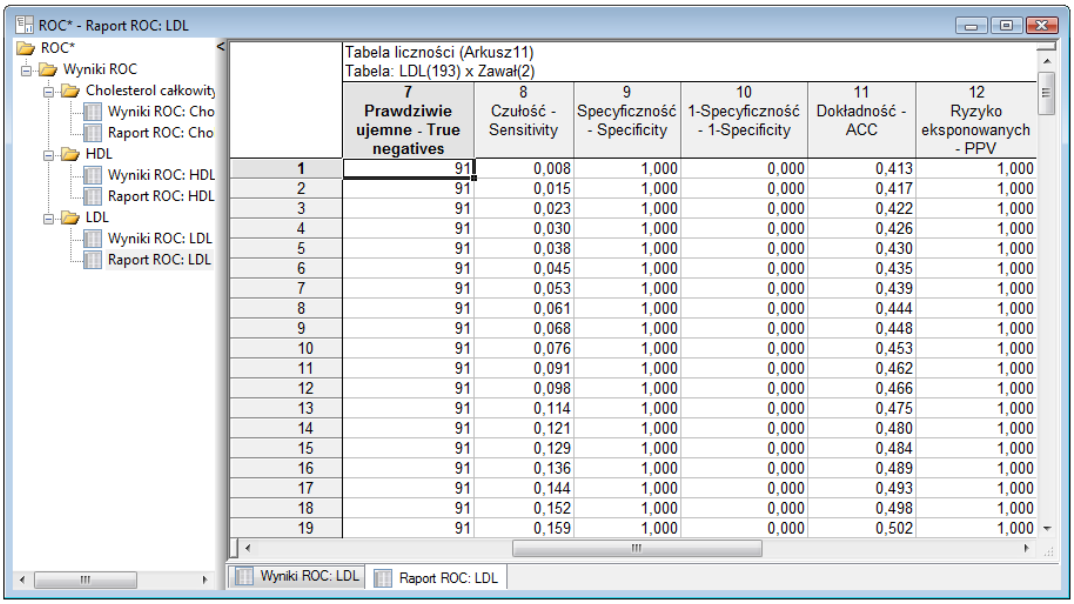
W obszarze ***Kod stanu wyróżnionego*** wskazujemy klasę *tak*, a następnie w obszarze ***Kierunek wpływu na stan wyróżniony*** określamy zmienną *HDL* jako destymulantę (domyślnie wszystkie predyktory są określone jako stymulanty).



Pozostałe ustawienia pozostawiamy bez zmian i potwierdzamy wybór, klikając przycisk ***OK***. Pojawia się okno ***ROC Wyniki***.

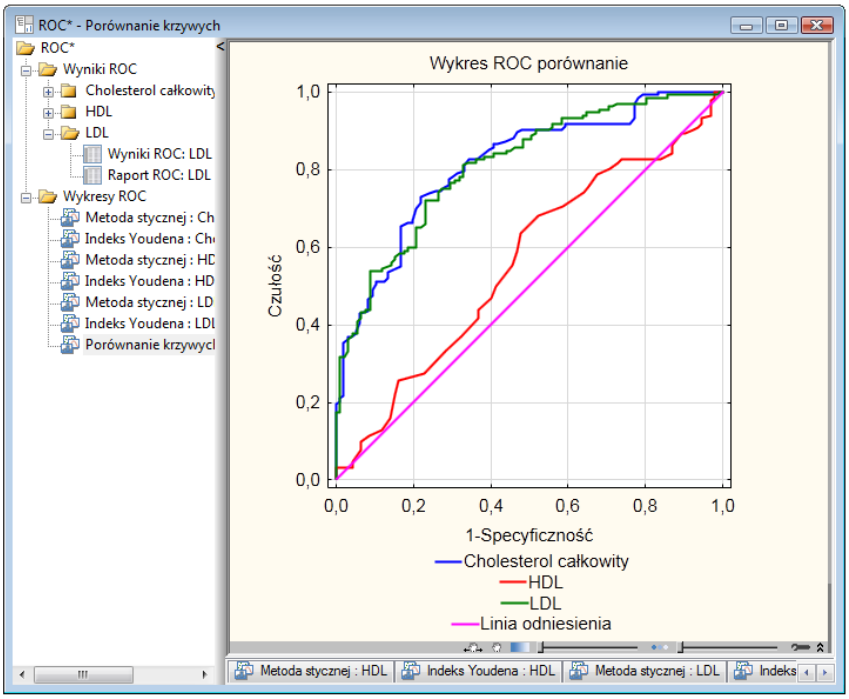


Klikając przycisk ***Raport*** na karcie ***Podstawowe***, otrzymujemy skoroszyt z arkuszami wyników dla każdego z wybranych predyktorów. Każdy *Raport ROC* zawiera między innymi wartości *Sensivity*, *Specificity*, *ACC*, *PPV*, *NPV*, *False positive ratio*, *False negative ratio* oraz *LR*. Wartości *AUC* (pole powierzchni pod krzywą) umieszczono w arkuszach *Wyniki ROC*.

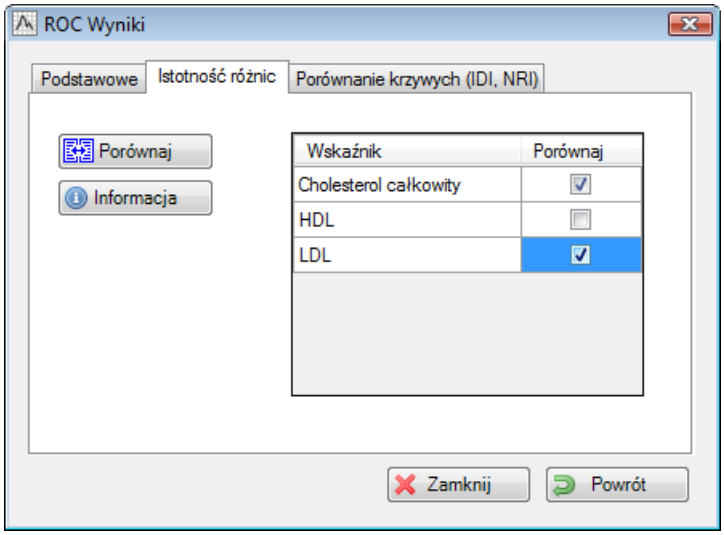


	7 Prawdziwie ujemne - True negatives	8 Czułość - Sensitivity	9 Specyficzność - Specificity	10 1-Specyficzność - 1-Specificity	11 Dokładność - ACC	12 Ryzyko eksponowanych - PPV
1	91	0,008	1,000	0,000	0,413	1,000
2	91	0,015	1,000	0,000	0,417	1,000
3	91	0,023	1,000	0,000	0,422	1,000
4	91	0,030	1,000	0,000	0,426	1,000
5	91	0,038	1,000	0,000	0,430	1,000
6	91	0,045	1,000	0,000	0,435	1,000
7	91	0,053	1,000	0,000	0,439	1,000
8	91	0,061	1,000	0,000	0,444	1,000
9	91	0,068	1,000	0,000	0,448	1,000
10	91	0,076	1,000	0,000	0,453	1,000
11	91	0,091	1,000	0,000	0,462	1,000
12	91	0,098	1,000	0,000	0,466	1,000
13	91	0,114	1,000	0,000	0,475	1,000
14	91	0,121	1,000	0,000	0,480	1,000
15	91	0,129	1,000	0,000	0,484	1,000
16	91	0,136	1,000	0,000	0,489	1,000
17	91	0,144	1,000	0,000	0,493	1,000
18	91	0,152	1,000	0,000	0,498	1,000
19	91	0,159	1,000	0,000	0,502	1,000

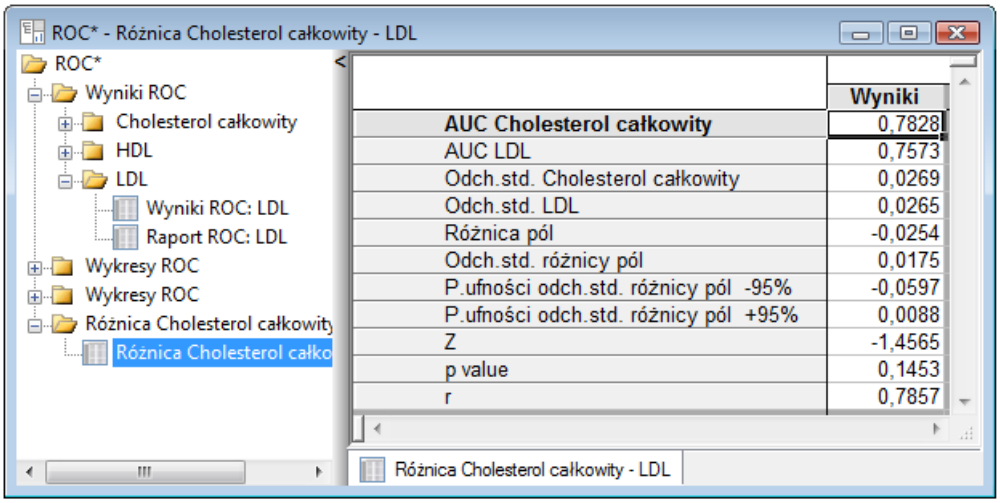
Aby uzyskać wykresy krzywych ROC dla każdego z wybranych predyktorów, należy kliknąć przycisk ***Wykres ROC*** w oknie ***ROC Wyniki***. Otrzymujemy wykresy krzywych ROC z zaznaczonymi proponowanymi punktami odcięcia oraz wykres porównujący wszystkie krzywe ROC.



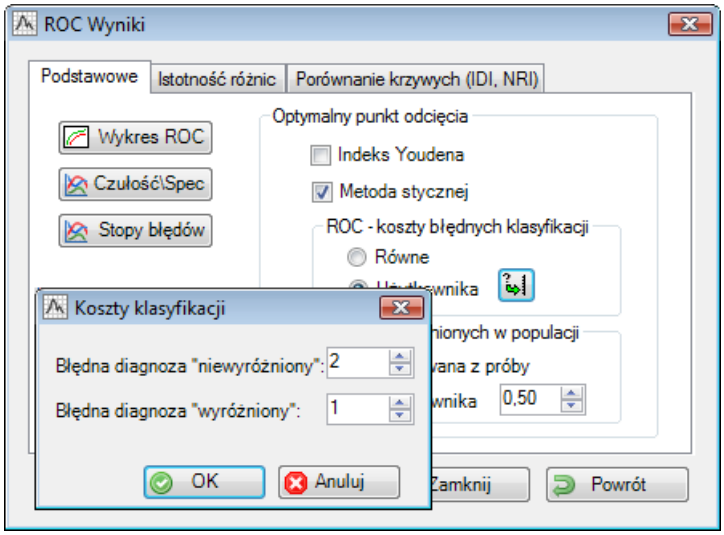
Moduł Krzywe *ROC* umożliwia przeprowadzenie analizy istotności różnic wartości AUC dla wybranych predyktorów. Dla przykładu porównamy predyktory: *Cholesterol całkowity* i *LDL*. Na karcie ***Istotność różnic*** zaznaczamy wybrane predyktory i klikamy przycisk ***Porównaj***.



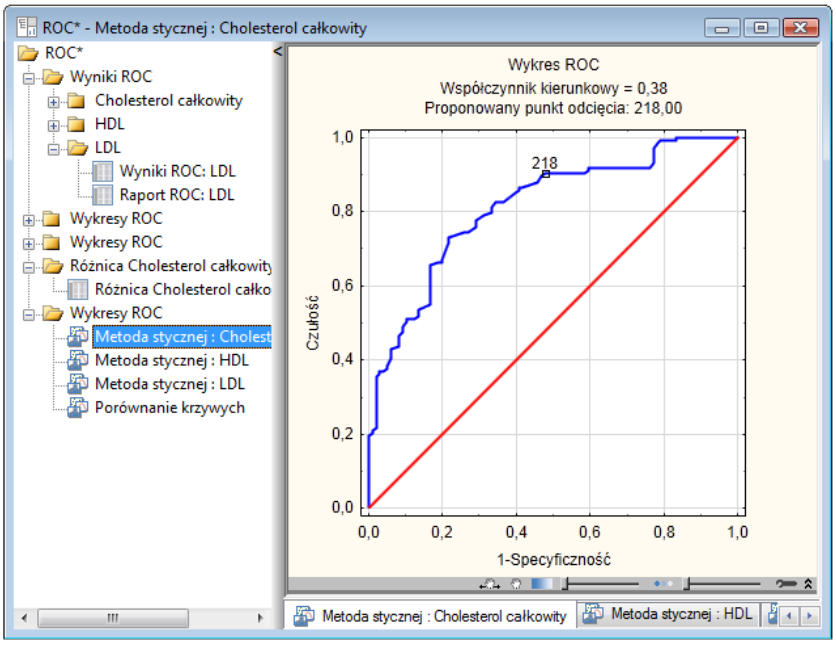
Do skoroszytu zostaje dodany arkusz wyników testu istotności różnic. W naszym przypadku *p value* wynosi 0,1453, czyli nie można wnioskować o występowaniu istotnych różnic w wartościach AUC dla wybranych predyktorów.



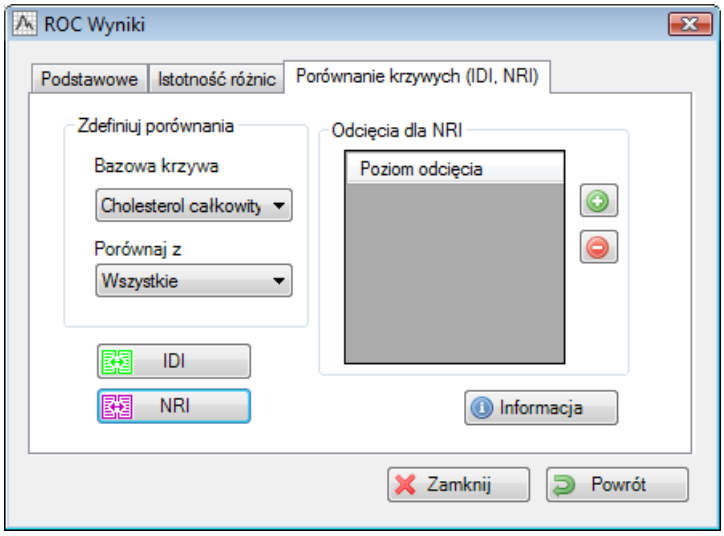
Kryterium doboru optymalnego punktu odcięcia można zmieniać w oknie ***ROC Wyniki*** na karcie ***Podstawowe*** za pomocą modyfikacji kosztów błędnych klasyfikacji lub podania frakcji wyróżnionych w populacji. Aby podać własne koszty błędnych klasyfikacji, zaznaczamy opcję ***Metoda stycznej*** a następnie w obszarze ***ROC – koszty błędnych klasyfikacji*** wybieramy opcję ***Użytkownika*** i klikamy przycisk po prawej stronie. W oknie ***Koszty klasyfikacji*** zmieniamy koszty i potwierdzamy wybór, klikając ***OK***.



Aby uzyskać punkt odcięcia dla zmodyfikowanych kosztów, klikamy przycisk ***Wykres ROC***.

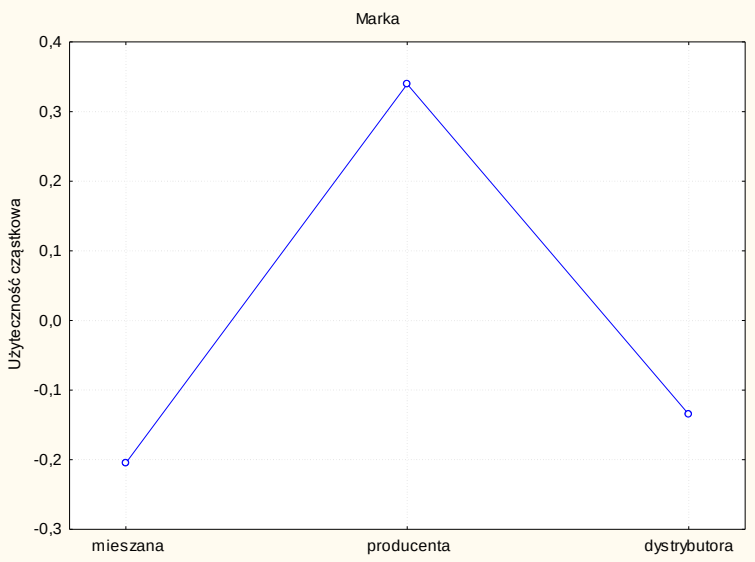


Dodatkową opcją jest możliwość porównania krzywych ROC za pomocą wskaźników IDI oraz NRI. Wskaźniki te przeznaczone są do porównania krzywych w sytuacji, gdy wartości predyktorów możemy interpretować jako prawdopodobieństwo (przyjmują wartości z zakresu od 0 do 1). Najczęściej wskaźniki te wykorzystuje się do porównywania modeli predykcyjnych.



7.2. Analiza conjoint

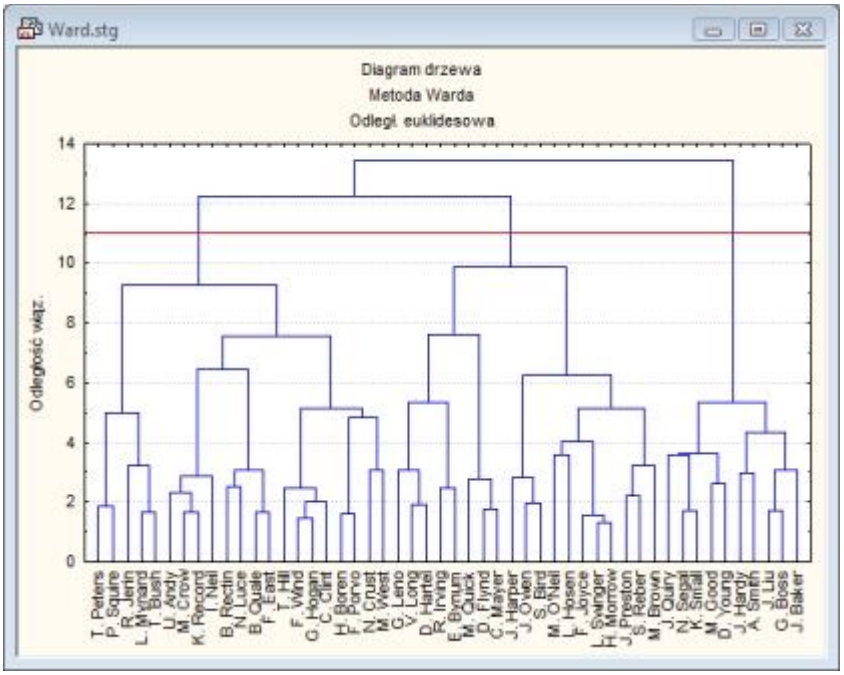
Moduł pozwala na wykonanie analizy dla zmiennych zależnych mierzonych na skali co najmniej przedziałowej. Program oblicza cząstkowe użyteczności poszczególnych poziomów cech (przedstawiane także w formie wykresów), a także użyteczności całkowite dla każdej kombinacji cech produktu i każdego respondenta oraz ranking profili. Dodatkowo obliczane są relatywne oraz sumaryczne ważności analizowanych zmiennych.



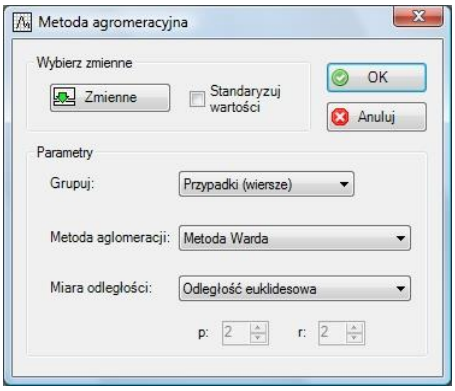
Szczegółowy opis tego modułu oraz przykłady jego wykorzystania dostępne są w artykule A. Sagana *Analiza preferencji konsumentów z wykorzystaniem programu STATISTICA – analiza conjoint i skalowanie wielowymiarowe*, dołączonym do niniejszej dokumentacji (zbiory danych użyte w artykule są dostępne w katalogu z pozostałymi plikami danych).

7.3. Aglomeracja z punktem odcięcia

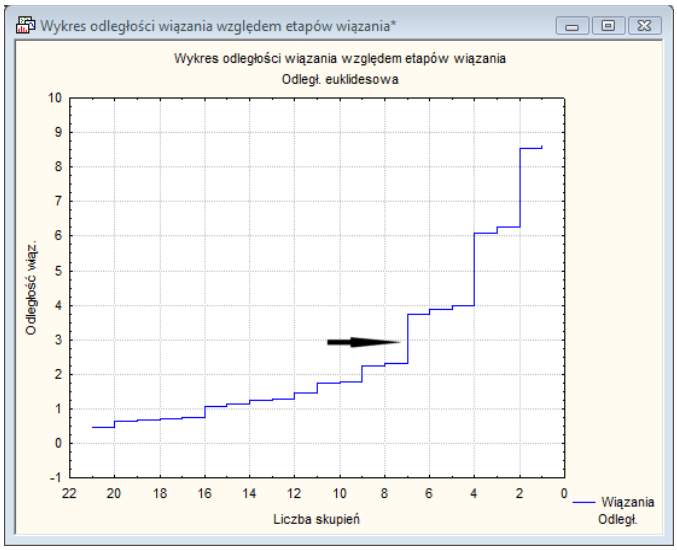
Aglomeracja z punktem odcięcia uzupełnia klasyczny moduł aglomeracyjnej analizy skupień o możliwość wskazania (w sposób ręczny bądź automatyczny) optymalnego punktu odcięcia dendrogramu oraz przygotowania na tej podstawie zbioru danych z przypisaniem analizowanych obiektów do poszczególnych grup.



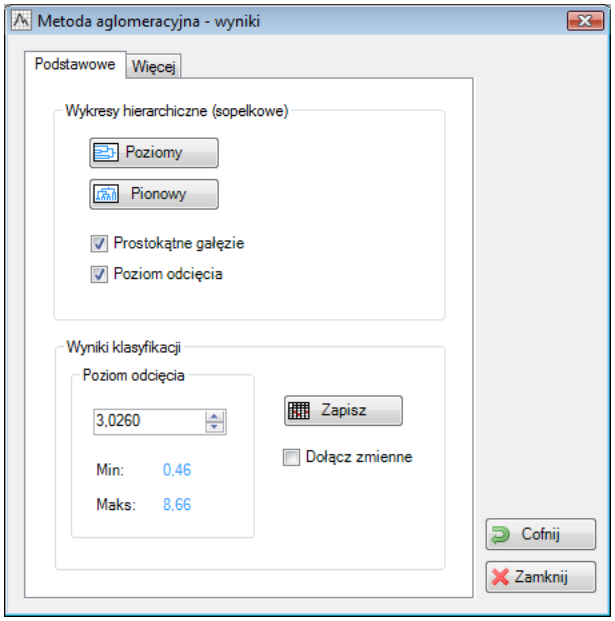
Przykład. Przedstawimy teraz przykład przygotowania modelu aglomeracyjnego na przykładzie pliku *Cars.sta*, zawierającego informacje o parametrach samochodów różnych marek (wybierano losowo jeden konkretny model spośród modeli oferowanych przez danego producenta). Z menu ***Analizy marketingowe i rynkowe*** | ***Analizy*** wybieramy polecenie ***Aglomeracja z punktem odcięcia*** przywołując okno ***Metoda aglomeracyjna***.



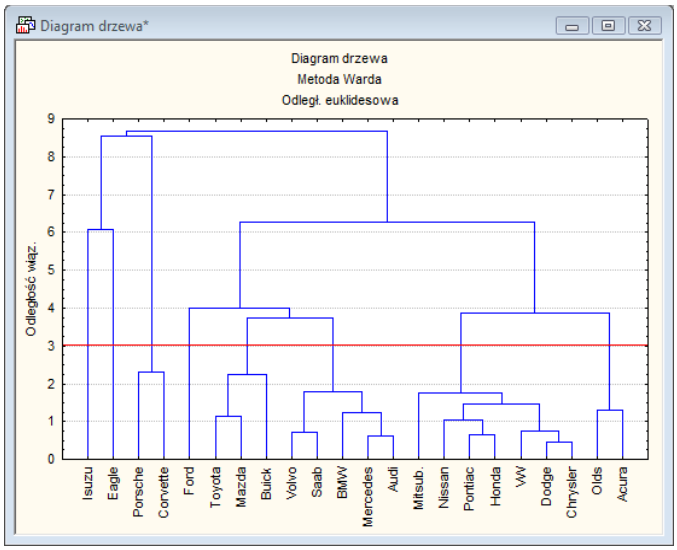
W pierwszej kolejności wskazujemy zmienne do analizy klikając przycisk ***Zmienne*** i wybierając do analizy wszystkie zmienne. Ponieważ dokonano już standaryzacji danych, nie zaznaczamy opcji ***Standaryzuj wartości***. Pozostałe opcje pozostawiamy na domyślnych ustawieniach i zatwierdzamy wykonanie analizy klikając ***OK.***, przywołując okno ***Metoda aglomeracyjna-wyniki***. Aby ocenić przebieg aglomeracji przechodzimy na kartę ***Więcej*** i klikamy ***Wykres***, aby wyświetlić przebieg procesu aglomeracji.



Zwykle bardzo dobrym punktem odcięcia jest pierwszy wyraźny przyrost odległości aglomeracyjnej. Możemy zauważyć taki przyrost dla odległości równej około 3 (rysunek powyżej). Przyjmijmy, że punkt ten będzie punktem odcięcia dendrogramu.



Przechodząc na kartę ***Podstawowe*** w obszarze ***Poziom odcięcia*** możemy zauważyć, że mechanizm automatycznego określania punktu odcięcia zaproponował podobne rozwiązanie (oczywiście my ręcznie możemy tę wartość zmienić kierując się na przykład kwestiami merytorycznymi). Klikamy ***Pionowy***, aby wyświetlić wykres hierarchiczny.



Analizując powyższy wykres możemy zauważyć, że efektem proponowanego rozwiązania będzie siedem skupień przy czym *Isuzu*, *Eagle* oraz *Ford* tworzą skupienia jednoelementowe. Jeśli uznalibyśmy proponowane rozwiązanie za satysfakcjonujące, na karcie ***Podstawowe*** klikamy przycisk ***Zapisz*** włączając wcześniej opcję ***Dołącz zmienne***. W wyniku analizy otrzymujemy arkusz danych z zestawem zmiennych wejściowych oraz dodatkową zmienną *Segment* zawierającą informację o segmencie, do którego trafiła analizowana obserwacja.

Dane: Klasyfikacja (7 zmn. * 22 prz.)

	Cena, wydajność, trzymanie się drogi różnych samochodów						
	1 CENA	2 PRZYSP	3 HAMOWAN	4 WSK_TRZY	5 ZUŻYCIE	6 Klasyfikowany element	7 Segment
Acura	-0,521	0,477	-0,007	0,382	2,079	1	5
Audi	0,866	0,208	0,319	-0,091	-0,677	2	3
BMW	0,496	-0,802	0,192	-0,091	-0,154	3	3
Buick	-0,614	1,689	0,933	-0,210	-0,154	4	2
Corvette	1,235	-1,811	-0,494	0,973	-0,677	5	1
Chrysler	-0,614	0,073	0,427	-0,210	-0,154	6	4
Dodge	-0,706	-0,196	0,481	0,145	-0,154	7	4
Eagle	-0,614	1,218	-4,199	-0,210	-0,677	8	7
Ford	-0,706	-1,542	0,987	0,145	-1,724	9	6
Honda	-0,429	0,410	-0,007	0,027	0,369	10	4
Isuzu	-0,798	0,410	-0,061	-4,230	1,067	11	8
Mazda	0,126	0,679	-0,133	0,500	-1,724	12	2
Mercedes	1,051	0,006	0,120	-0,091	-0,154	13	3
Mitsub.	-0,614	-1,003	0,084	0,382	0,718	14	4
Nissan	-0,429	0,073	-0,007	0,263	0,997	15	4
Olds	-0,614	-0,734	0,409	0,382	2,114	16	5
Pontiac	-0,614	0,679	0,536	0,145	0,195	17	4
Porsche	3,454	-2,215	-0,296	0,618	-1,026	18	1
Saab	0,588	0,679	0,246	0,263	0,021	19	3
Toyota	-0,059	1,218	0,228	0,736	-0,851	20	2
VW	-0,706	-0,128	0,102	0,382	0,195	21	4
Volvo	0,219	0,612	0,138	-0,210	0,369	22	3

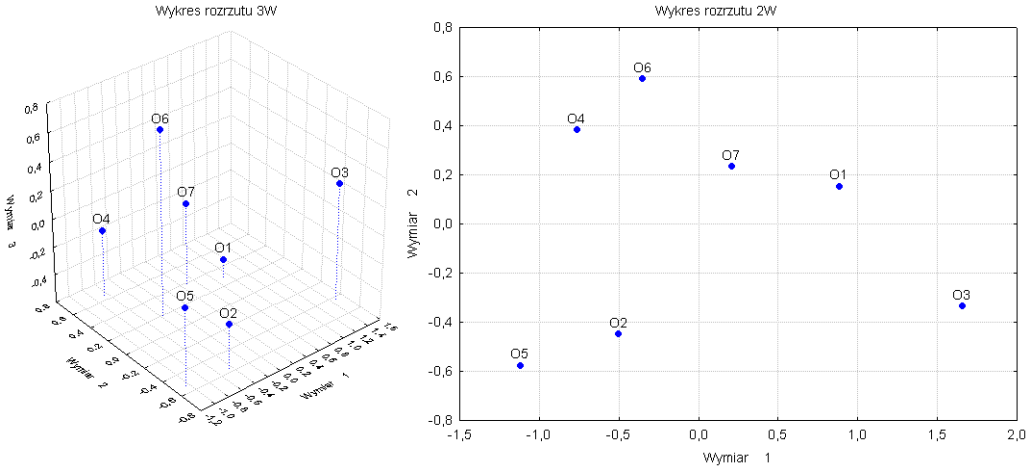
7.4. Analiza PROFIT

Analiza PROFIT jest procedurą łączącą dwie popularne techniki analityczne, skalowanie wielowymiarowe oraz regresję wieloraką. Celem skalowania wielowymiarowego jest graficzna prezentacja struktury podobieństwa pomiędzy analizowanymi obiektami względem wybranego zbioru zmiennych. Struktura ta jest zazwyczaj prezentowana za pomocą dwuwymiarowego (niekiedy trójwymiarowego) wykresu rozrzutu, często określanego mianem mapy percepcji bądź mapy podobieństw. Aby ułatwić interpretację uzyskanego wykresu rozrzutu obiektów oraz wyjaśnić wymiary mapy podobieństw, wykonujemy analizę regresji, w wyniku której na uzyskaną mapę nakładane są dodatkowe osie opisujące wymiary i ułatwiające interpretację zbudowanej mapy. Poniższy przykład analizy oraz opis metody został zaczerpnięty z opracowania P. Jabkowskiego[2].

[2] Jabkowski P., *O korzyściach wynikających z zastosowania analizy PROFIT*, Praktyczna analiza danych w marketingu i badaniach rynku, Materiały z seminariów, StatSoft Polska 2010

„Skalowanie wielowymiarowe jest w swojej naturze podobne do analizy czynnikowej z tą różnicą, że powiązania pomiędzy poszczególnymi obiektami mogą być wyrażone nie tylko poprzez macierze korelacji, ale także jako macierze dowolnych miar odległości, np. euklidesowej, kwadratu odległości euklidesowej, miejskiej Manhattan, Czebyszewa, niezgodności procentowej, potęgowej i innych. Jest to o tyle istotne, że skalowanie wielowymiarowe pozwala porównywać obiekty nie tylko względem cech ilościowych, ale także tych jakościowych. Zakres wykorzystania skalowania wielowymiarowego (analizy *PROFIT*) jest więc szerszy, niż zakres zastosowań analizy czynnikowej.

Głównym celem skalowania wielowymiarowego jest graficzna prezentacja struktury podobieństwa (lub odmienności) pomiędzy analizowanymi obiektami względem wybranego zbioru zmiennych (cech). Generalnie zatem rzecz ujmując, skalowanie wielowymiarowe dąży do takiego uporządkowania obiektów, by jednocześnie zredukować liczbę wymiarów i możliwie najlepiej odtworzyć pierwotne obserwowane odległości (różnice) pomiędzy obiektami. Graficzna prezentacja wyników przyjmuje dobrze znaną postać wykresu rozrzutu obiektów, nazywaną „mapą skalowania wielowymiarowego". Mapa taka, najczęściej 2- lub 3-wymiarowa, ma bardzo prostą interpretację. Przyjmuje się bowiem, że im mniejsza odległość pomiędzy badanymi obiektami, tym są one bardziej do siebie podobne. W ten sposób można wyznaczać grupy (klastry) obiektów, obiekty izolowane itd. Przykładowe mapy skalowania wielowymiarowego, dla dwóch oraz trzech wymiarów, zaprezentowane zostały na poniższym rysunku.

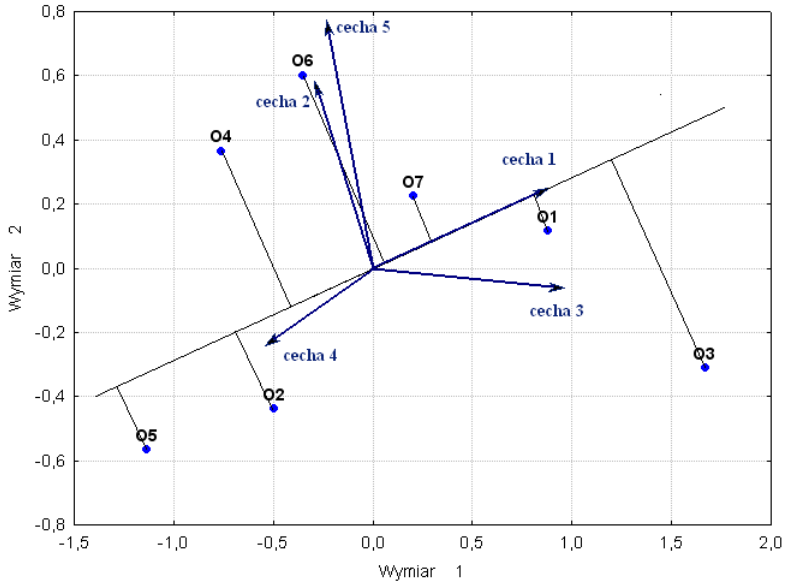


W wyniku skalowania wielowymiarowego otrzymujemy zatem przestrzeń, na której rozlokowane są interesujące nas obiekty. Po utworzeniu takiej mapy, każdemu obiektowi przyporządkować można wartości, które odpowiadają współrzędnym na mapie *MDS*. Ponieważ celem analizy *PROFIT* jest interpretacja tego, w jaki sposób wejściowy zbiór cech obiektów odwzorowany jest na osiach skalowania wielowymiarowego, to te współrzędne (koordynaty) przypisane obiektom, traktować będzie się jako zmienne wyjaśniające (niezależne), a wartości poszczególnych cech obiektów jako zmienne zależne (wyjaśniane). Potrzeba nam bowiem informacji o tym, w jaki sposób (w jakim kierunku) na płaszczyźnie (w przestrzeni) ułożone są obiekty, ze względu na natężenie każdej z tych wejściowych cech. Sposobem na to jest estymacja parametrów modelu, poprzez odniesienie każdej cechy do pozycji obiektów na mapie *MDS*.

Algorytm analizy *PROFIT* wykorzysta zatem informacje o współrzędnych (jako zmiennych niezależnych) oraz wartościach obiektów względem każdej z poszczególnych cech (jako zmiennych zależnych), przeprowadzając analizy regresji wielorakiej. Wykonanych będzie tyle analiz regresyjnych, ile cech (zmiennych) uwzględniono w skalowaniu wielowymiarowym. Dla przykładu, jeżeli marki pewnych produktów oceniane były pod względem pięciu cech, to dla każdej takiej cechy przeprowadzona będzie analiza regresji.

Standaryzowane współczynniki równań regresyjnych odpowiadające każdej z osi skalowania wielowymiarowego wyznaczają punkt na mapie *MDS* określający współrzędne danej cechy

(zmiennej). Innymi słowy, to właśnie one pozwalają ustalić, w jaki sposób ulokowane są interesujące nas obiekty ze względu na natężenie danej cechy. Należy mieć na uwadze fakt, że dla opisu wyników bez znaczenia pozostaje odległość danego obiektu od prostej, na której położony jest wektor, interpretuje się z kolei uszeregowanie rzutów obiektów na takie proste.

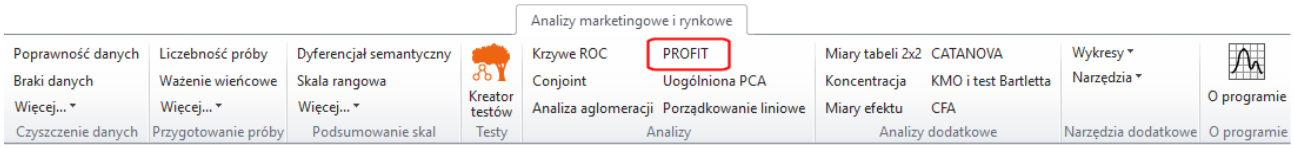


Z zaprezentowanych danych można odczytać, że względem cechy 1, uszeregowanie obiektów wygląda tak, że największym natężeniem wartości zmiennej (cechy) charakteryzuje się obiekt O3, natomiast najmniejszym obiekt O5. Dla przykładu w badaniach marketingowych można by dokonać skalowania wielowymiarowego pewnych marek produktów względem kilku cech, z których jedną stanowiłaby ocena prestiżu danej marki. Rzutując dane marki produktów na wektor prestiżu, otrzymano by informację o tym, która marka cieszy się największym, a która najmniejszym prestiżem wśród respondentów.

Dokonując interpretacji wyników analizy *PROFIT*, należy także rozpatrzyć współczynniki determinacji równań regresyjnych. Pokazują one, w jakim stopniu uszeregowanie obiektów względem wartości danej cechy wyjaśniane jest poprzez położenie tych obiektów na płaszczyźnie.

W badaniach marketingowych oraz badaniach rynku procedura skalowania wielowymiarowego, w tym także analiza *PROFIT*, zyskała swoje szczególne uznanie w zakresie budowy map percepcyjnych. Wykorzystuje się je np. wtedy, gdy głównym celem analiz jest ustalenie, w jakim zakresie (względem jakich wymiarów) porównywane między sobą towary, usługi, czy też produkty, uznawane są przez respondentów jako podobne. Podstawą skalowania wielowymiarowego jest w takich przypadkach macierz relacji podobieństwa pomiędzy analizowanymi markami produktów (obiektami). Następnie wyliczane są uśrednione oceny preferencji dla poszczególnych marek produktów, które wykorzystuje się w równaniach regresyjnych. Standaryzowane współczynniki równań regresyjnych wyznaczają zwrot i kierunek wektora danej cechy produktu i tym samym umożliwiają ustalenie preferencji produktów względem tej cechy.

Przejdźmy teraz do praktycznej implementacji procedury ***PROFIT*** w programie ***Analizy marketingowe i rynkowe***. Przykład wykonamy na podstawie pliku *PROFIT.sta*.



Dane wykorzystywane w przykładzie pochodzą z badań opartych na wywiadach kwestionariuszowych prowadzonych wśród odbiorców usług komunalnych w Poznaniu. Celem badań było ustalenie tego, w jaki sposób mieszkańcy miasta oceniają firmy świadczące usługi w zakresie: (1) dostarczania ciepła do mieszkań, (2) dostarczania gazu, (3) dostarczania wody, (4) dostarczania energii elektrycznej, (5) oczyszczania miasta oraz (6) transportu publicznego. Poszczególne firmy oceniane były pod względem (a) nowoczesności, (b) konkurencyjności, (c) jakości świadczonych usług, (d) dynamiczności rozwoju, (e) dbania o klienta oraz (f) wiarygodności. Wykorzystano w tym celu tzw. dyferencjał semantyczny, tzn. proszono respondentów, aby w ramach oceny każdej firmy dokonali ewaluacji tego, czy uważają ją za: (a) nowoczesną czy przestarzałą, (b) konkurencyjną czy niekonkurencyjną itd. dla każdej z takich cech. W ramach każdego wymiaru oceny odpowiedziom respondentów przypisano wagi liczbowe w ten sposób, że odpowiedzi skrajnie pozytywnej przypisano wartość +3, odpowiedzi neutralnej wartość 0, natomiast odpowiedzi skrajnie negatywnej wartość -3. Cała gama wartości obejmowała następujący zbiór wag: -3, -2, -1, 0, +1, +2 +3. Na podstawie jednostkowych odpowiedzi respondentów uśredniono wyniki dyferencjału semantycznego, uzyskując oceny poszczególnych firm z punktu widzenia wyróżnionych wymiarów oceny.

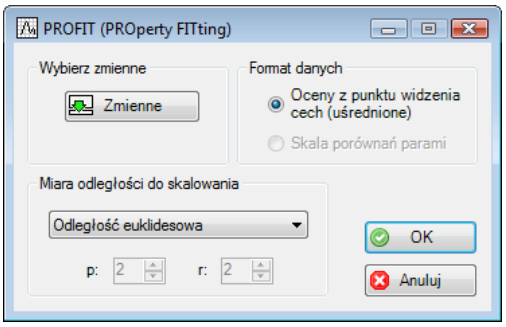
Pierwszym ułatwieniem, jakie daje moduł analizy *PROFIT* względem modułu analizy skalowania wielowymiarowego, jest sposób przygotowania pliku wejściowego do analizy. Prowadząc w programie *STATISTICA* skalowanie wielowymiarowe, należy przekształcić plik wejściowy do postaci macierzowej, z wyliczonymi odległościami pomiędzy rozpatrywanymi obiektami. Przeprowadzenie skalowania wielowymiarowego możliwe jest więc dopiero po odpowiednim zapisaniu danych w pliku macierzowym.

	Baza danych - przykład 1					
	1 **Firma 1**	2 Firma 2	3 Firma 3	4 Firma 4	5 Firma 5	6 Firma 6
Firma 1	0,00000	1,13371	1,69979	1,34696	2,39962	1,20632
Firma 2	1,13371	0,00000	1,40968	1,38470	2,28256	1,42327
Firma 3	1,69979	1,40968	0,00000	1,96209	1,18520	1,43878
Firma 4	1,34696	1,38470	1,96209	0,00000	2,13670	1,54606
Firma 5	2,39962	2,28256	1,18520	2,13670	0,00000	1,92338
Firma 6	1,20632	1,42327	1,43878	1,54606	1,92338	0,00000
Średnie	0,45833	0,04333	-0,00333	0,17333	-0,04167	0,25833
Odch.std	0,45849	0,52110	0,32303	0,59691	0,52674	0,40504
No.Cases	6,00000					
Matrix	3,00000					

W analizie *PROFIT* taka transformacja pliku danych do postaci macierzowej nie jest konieczna, wystarczy wyliczyć statystyki punktowe (np. średnią arytmetyczną, medianę czy też wskaźnik struktury) charakteryzujące obiekty względem interesujących nas cech. Poniższy rysunek zawiera zestawienie danych wejściowych wykorzystanych w opisywanym przykładzie.

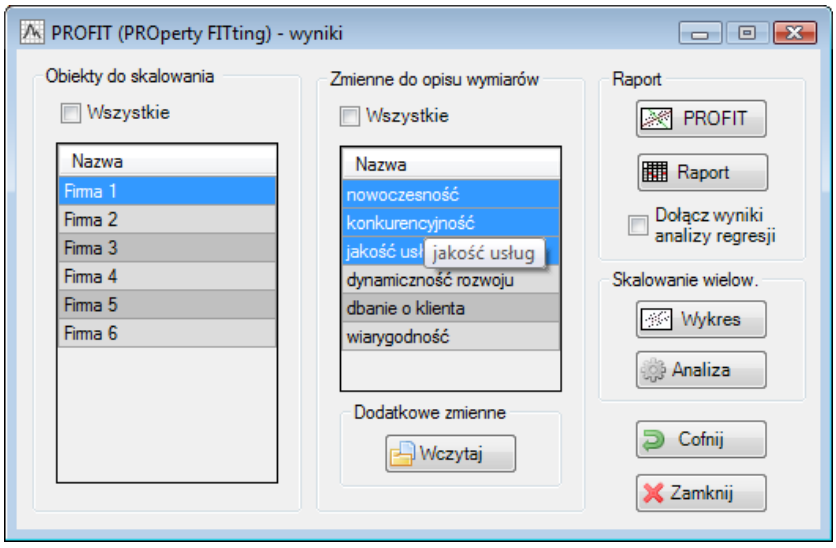
	Baza danych - analiza PROFIT - przykład 1					
	1 **nowoczesność**	2 konkurencyjność	3 jakość usług	4 dynamiczność rozwoju	5 dbanie o klienta	6 wiarygodność
Firma 1	0,29	0,22	0,37	-0,14	1,04	0,97
Firma 2	-0,46	-0,13	-0,06	-0,42	0,47	0,86
Firma 3	0,38	-0,08	-0,55	0,12	-0,1	0,21
Firma 4	-0,32	0,54	0,64	-0,66	0,85	-0,01
Firma 5	0,51	0,42	-0,37	0,29	-0,29	-0,81
Firma 6	0,21	-0,46	0,53	0,36	0,72	0,19

Na tak przygotowanych danych rozpocząć można procedurę analizy *PROFIT*. Po jej uruchomieniu okno programu wygląda następująco:



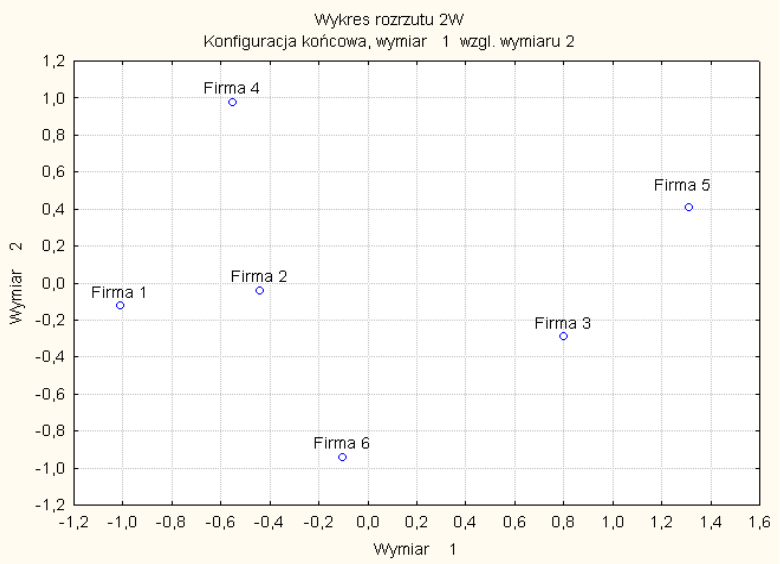
Konieczne jest ustalenie wejściowych warunków dla wykonania analizy *PROFIT*. W pierwszej kolejności określić należy format danych dla analizy. Ponieważ dane w przykładzie zapisane zostały jako uśrednione oceny poszczególnych firm w ramach każdego z sześciu wymiarów, to format danych ustalamy jako „uśrednione oceny obiektów z punktu widzenia cech". Do wyboru mamy także siedem miar odległości w tym: kwadrat odległości euklidesowej, odległość euklidesową, odległość miejską (tzw. Manhattan), odległość Czebyszewa, odległość potęgową, niezgodność procentową, a także odległość opartą na korelacjach liniowych Pearsona. Zauważmy, że w programie *STATISTICA* analiza *PROFIT* oferuje ten sam zestaw miar odległości, który dostępny jest również w procedurze skalowania wielowymiarowego (*PROFIT* jest przecież rozszerzeniem *MDS*), a także w module analizy skupień programu *STATISTICA*. Celem wszystkich tych procedur jest analiza podobieństw, stąd konsekwentnie dostępny jest ten sam zestaw miar odległości.

Po ustaleniu tych podstawowych warunków procedury *PROFIT* możemy przejść do wyboru zmiennych (cech), które będą podstawą skalowania obiektów. W naszym przykładzie do utworzenia mapy percepcji firm świadczących usługi komunalne wybierzemy wszystkie sześć cech, względem których firmy te oceniane były przez mieszkańców. Po zaakceptowaniu listy zmiennych program przechodzi do okna *PROFIT* – wyniki.



Na podstawie danych wejściowych algorytm analizy PROFIT wykonał już skalowanie wielowymiarowe obiektów na płaszczyźnie, chociaż na tym etapie program nie przedstawia jeszcze mapy podobieństw. Co ważne, w klasycznym skalowaniu wielowymiarowym możliwe jest utworzenie mapy o dowolnej liczbie wymiarów, natomiast algorytm procedury *PROFIT* w programie *STATISTICA* został zaprogramowany tak, by zawsze dać rozwiązanie dwuwymiarowe. Z technicznych przyczyn niemożliwe było wykreślenie rozwiązania w trzech wymiarach.

Załóżmy zatem, że po przeprowadzeniu klasycznego skalowania wielowymiarowego (analiza ta dostępna jest w module *wielowymiarowe techniki eksploracyjne* programu *STATISTICA*) także otrzymalibyśmy wynik w postaci mapy 2D. Takie rozwiązanie pozwoliłoby wprawdzie ustalić, które firmy są bardziej, a które mniej do siebie podobne, nie sposób jednak byłoby w prosty sposób stwierdzić, jak grupują się te obiekty względem interesujących nas cech. Innymi słowy nie bylibyśmy w stanie intuicyjnie ustalić, względem jakich cech rozpatrywane obiekty są do siebie podobne, a względem jakich cech odmienne.



Powróćmy zatem do analizy *PROFIT*, która na tym etapie dała „na razie" te same rezultaty co skalowanie wielowymiarowe. W oknie wyników analizy *PROFIT* teraz ponownie wybieramy obiekty oraz ustalamy, które zmienne wykorzystane będą do opisu wymiarów skalowania wielowymiarowego. W naszym przypadku porównywać będziemy wszystkie firmy (obiekty), a do opisu wymiarów wykorzystamy wszystkie zmienne wejściowe.

Zauważmy jednak, że lista zmiennych do opisu wymiarów może być inna, niż lista zmiennych na podstawie której dokonano skalowania wielowymiarowego obiektów. Bardzo często jednak do opisu wymiarów wykorzystujemy zarówno dane wejściowe, jak i dodatkowe informacje. Dla przykładu moglibyśmy dla analizowanych przez nas firm zebrać dodatkowe dane dotyczące wydatków na kampanie promocyjne i zdiagnozować, na ile ocena firm jest powiązana z działaniami *Public Relations*.

Jeżeli w oknie wyników analizy *PROFIT* zaznaczymy również opcje dołączenia wyników regresji, to program zamieści w skoroszycie wyniki tylu modeli regresyjnych, ile zmiennych wybrano do opisu wyników. Każda z tych zmiennych jest objaśniana poprzez położenie (współrzędne) obiektów na płaszczyźnie (rys. poniżej).

	Baza danych - analiza PROFIT - przykład 1							
	1 **nowoczesność**	2 konkurencyjność	3 jakość usług	4 dynamiczność rozwoju	5 dbanie o klienta	6 wiarygodność	7 WYM. 1	8 WYM. 2
Firma 1	0,29	0,22	0,37	-0,14	1,04	0,97	-1,01050	-0,121828
Firma 2	-0,46	-0,13	-0,06	-0,42	0,47	0,86	-0,44366	-0,037936
Firma 3	0,38	-0,08	-0,55	0,12	-0,1	0,21	0,79771	-0,287516
Firma 4	-0,32	0,54	0,64	-0,66	0,85	-0,01	-0,55312	0,979472
Firma 5	0,51	0,42	-0,37	0,29	-0,29	-0,81	1,31197	0,407734
Firma 6	0,21	-0,46	0,53	0,36	0,72	0,19	-0,10240	-0,939925

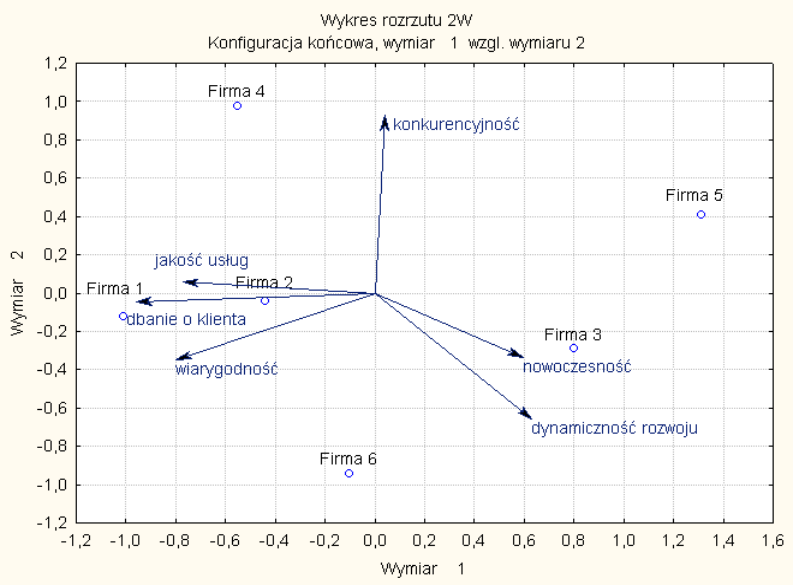
Na podstawie takich danych przeprowadzane są analizy regresyjne. Przykładowy wynik analizy regresji dla wymiaru oceny *konkurencyjności* rozpatrywanych firm przedstawiony został na poniższym rysunku.

N=6	Podsumowanie regresji zmiennej: konkurencyjność (Arkusz183) R= ,92897699 R^2= ,86299825 Skoryg. R2= ,77166374 F(2,3)=9,4488 p<,05071 Błąd std. estymacji: ,17981 BETA	Bł. std. BETA	B	Bł. std. B	t(3)	poziom p
W. wolny			0,085000	0,073405	1,157953	0,330698
WYM. 1	0,042770	0,213699	0,018241	0,091142	0,200142	0,854169
WYM. 2	0,927992	0,213699	0,537785	0,123842	4,342519	0,022541

Po przeprowadzeniu analiz regresyjnych dla wszystkich zmiennych, algorytm analizy *PROFIT* naniesie współczynniki kierunkowe określające zwrot i kierunek wektora, odpowiadającego każdej z cech wybranych do opisu wymiarów.

	1 **nowoczesność**	2 konkurencyjność	3 jakość usług	4 dynamiczność rozwoju	5 dbanie o klienta	6 wiarygodność
Wymiar 1	0,594409927	0,0427701069	-0,771437882	0,63148127	-0,958990711	-0,797056435
Wymiar 2	-0,334341627	0,927991899	0,0581444096	-0,657889285	-0,0450272172	-0,348072289

Ostatecznym wynikiem analizy *PROFIT* jest mapa percepcji z wektorami opisującymi poszczególne wymiary płaszczyzny.

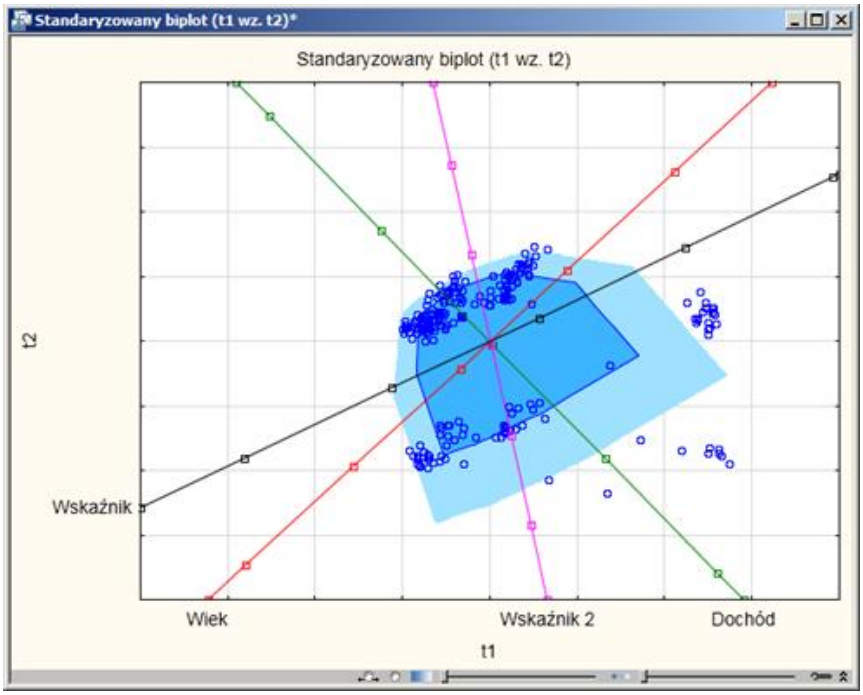


Wynik analizy *PROFIT* z naniesionymi współrzędnymi poszczególnych cech pozwala teraz na bardzo intuicyjną interpretację podobieństw i odmienności pomiędzy analizowanymi firmami. Dla przykładu z wykresu tego można odczytać, że Firma 4 jest oceniana jako najbardziej konkurencyjna i pod tym względem podobna jest do Firmy 5. To, co odróżnia obie firmy, to np. ocena ich nowoczesności i dynamiczności rozwoju; Firma 5 oceniana jest w obu wymiarach najwyżej, natomiast Firma 4 – najniżej. Podobnie można zauważyć, że Firma 1 ma najwyższe oceny w zakresie wiarygodności działań, dbania o klienta oraz jakości świadczenia usług, z kolei najgorsze oceny w tym względzie otrzymała Firma 3 oraz Firma 4. Analogiczne wnioskowanie można przeprowadzić dla dowolnego obiektu oraz dowolnego wymiaru osi skalowania wielowymiarowego".

7.5. Uogólniona metoda składowych głównych

Narzędzie umożliwiające wykonie analizy zarówno dla zmiennych ilościowych jak i jakościowych. Moduł umożliwia automatyczne określenie optymalnej liczby składowych za pomocą sprawdzianu krzyżowego, a także dynamiczne dodawanie\usuwanie kolejnych składowych.

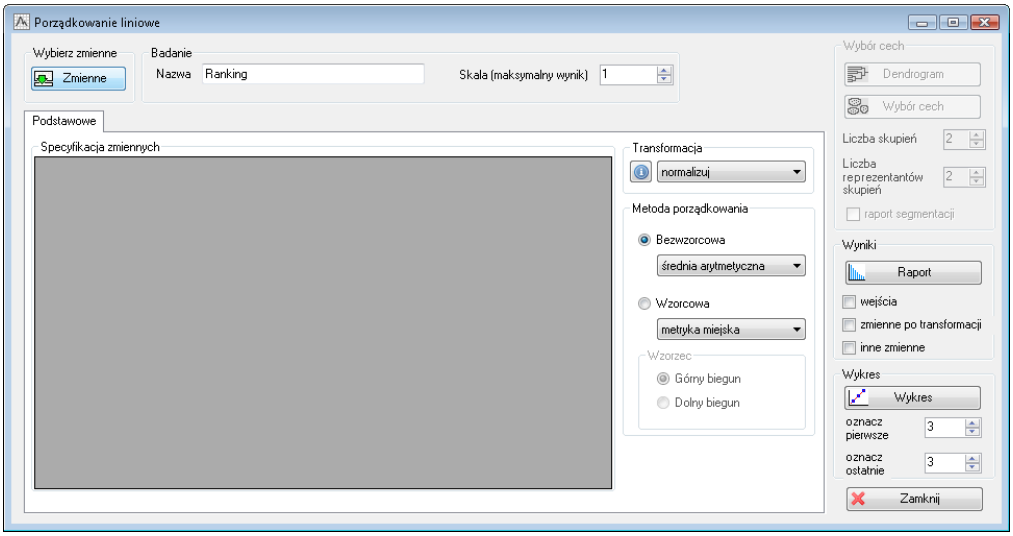
Dostępny jest kompletny zestaw wyników przeprowadzonych analiz w tym wykres osypiska oraz biplot.



7.6. Porządkowanie liniowe

Zadaniem metod porządkowania liniowego zbioru obiektów jest uszeregowanie, czyli ustalenie kolejności obiektów lub ich zbiorów według określonego kryterium. Naszym celem jest określenie syntetycznej miary agregującej cechy obiektu a następnie uporządkowanie na jej podstawie obiekty od „najlepszego" do „najgorszego".

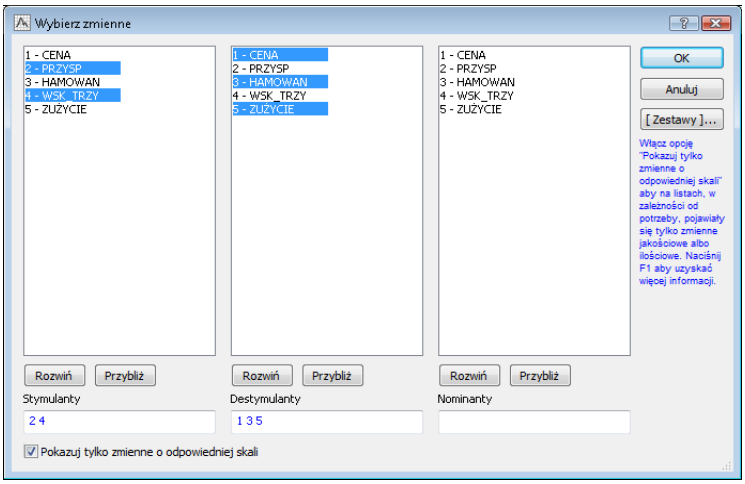
Przykład porządkowania liniowego zaprezentujemy na podstawie pliku *Cars.sta* zawierającego informacje o cenie, oraz parametrach eksploatacyjnych wybranych marek samochodów. Aby rozpocząć analizę, z menu ***Analizy marketingowe i rynkowe*** z grupy ***Analizy*** wybieramy opcję ***Porządkowanie liniowe***, wyświetlając okno o tej samej nazwie.



Moduł rozróżnia trzy typy zmiennych (cech obiektu):

- stymulanty,
- destymulanty,
- nominanty.

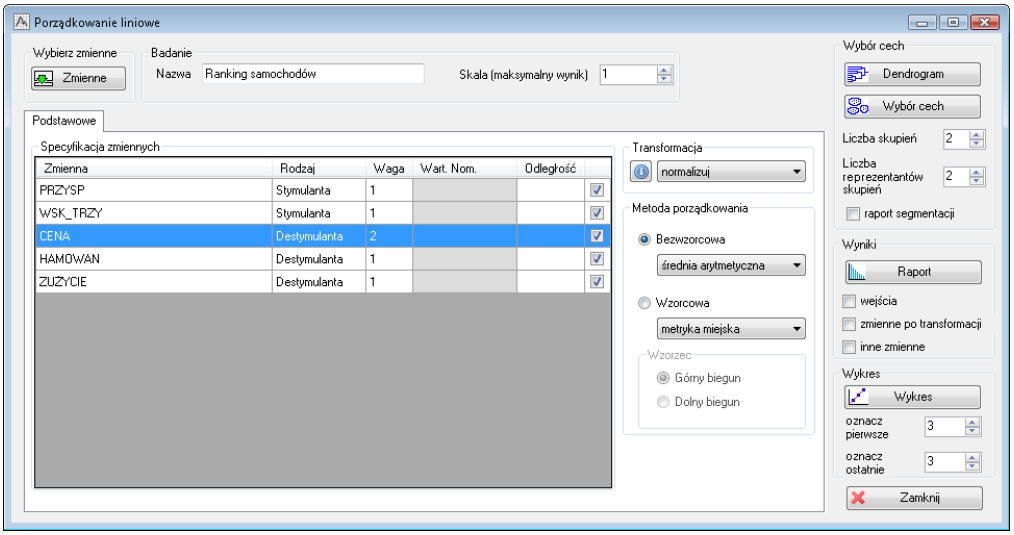
Ma to swoje odzwierciedlenie w oknie wyboru zmiennych, które wyświetlamy za pomocą przycisku ***Zmienne***.



W oknie tym przyspieszenie oraz wskaźnik trzymania się szosy określamy jako ***Stymulanty*** (im większe wartości tym lepiej), natomiast cenę, długość hamowania oraz zużycie paliwa określamy jako ***Destymulanty*** (im mniejsze wartości tym lepiej). Po zatwierdzeniu wyboru zmiennych wskazane predyktory pojawiają się w tabeli w obszarze ***Specyfikacja zmiennych***.

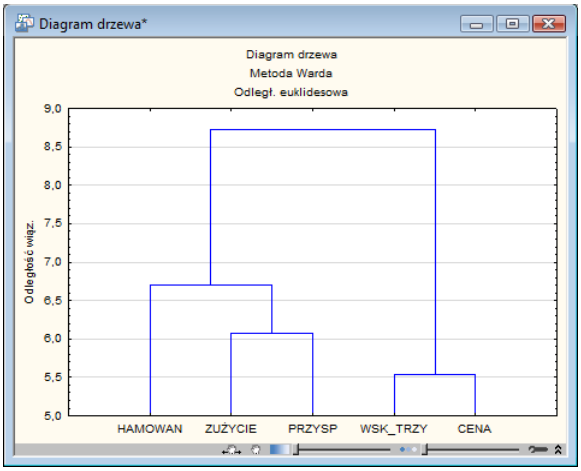
W tabeli tej obok nazwy zmiennej oraz jej polaryzacji zamieszczono kolumnę ***Waga***, w której możemy określić subiektywną ważność poszczególnych zmiennych w rankingu. Możemy na przykład założyć, że *CENA* jest dla nas dwa razy ważniejszym kryterium porządkowania niż pozostałe wymiary i określić jej wagę na poziomie 2. Jeśli wybrana przez nas zmienna jest nominantą, w kolumnie ***Wart. Nom***. podajemy optymalną dla tej zmiennej wartość.

W kolejnym kroku w obszarze ***Badanie*** wprowadzamy nazwę badania *Ranking samochodów*, wartość pola ***Skala (maksymalny wynik)*** ustawiamy na 1.



Ostatnia kolumna tabeli umożliwia ręczne włączanie\usuwanie zmiennych z rankingu. W sytuacji, gdy wskazaliśmy większą liczbę zmiennych do analizy możemy posłużyć się metodami automatycznego wyboru zmiennych opartymi na analizie skupień.

Kliknięcie przycisku ***Dendrogram*** wyświetli wykres sopelkowy będący wynikiem analizy aglomeracyjnej przeprowadzonej dla zmiennych. Na jego podstawie możemy ocenić podobieństwo (w sensie odległości a nie korelacji) pomiędzy poszczególnymi zmiennymi i określić liczbę skupień, jakie tworzą.



Analiza dendrogramu może być podstawą określenia liczby skupień, jakie tworzą wybrane przez nas cechy i ewentualnie eliminacji niektórych cech. Liczbę skupień wprowadzamy w polu ***Liczba skupień***, pole ***Liczba reprezentantów skupień*** określa ile zmiennych z danego skupienia chcielibyśmy uwzględnić w rankingu. Wybór zmiennych do analizy wykonujemy klikając przycisk ***Wybór cech*** co uruchomi procedurę grupowania metodą k-średnich. Wskazana liczba reprezentantów danego skupienia zostanie wybrana na podstawie ich odległości od środka swojego skupienia – odległość ta zostanie także wpisana do kolumny ***Odległość*** w tabeli. W naszej sytuacji nie będziemy eliminowali (ani ręcznie ani automatycznie) żadnych z wybranych zmiennych.
Przed obliczeniem syntetycznej miary dla każdego z obiektów, wartości cech je opisujące mogą zostać poddane jednej z następujących transformacji:

- normalizacja (przekształcenie zmiennej do przedziału od 0 do 1),
- standaryzacja (odjęcie średniej i podzielenie przez odchylenie standardowe),
- rangowanie,
- dane surowe.

W naszym przykładzie wybieramy opcję ***Normalizacja*** ponieważ jedynie ona pozwala uwzględnić różny charakter wpływu poszczególnych zmiennych na analizowane zjawisko. Kolejną decyzją, jaką badacz musi podjąć przed wykonaniem analizy jest określenie sposobu agregacji poszczególnych wymiarów do jednej syntetycznej miary. W obszarze ***Metoda porządkowania*** mamy możliwość wybrania opcji ***Bezwzorcowa***. W takiej sytuacji syntetyczna miara agregująca obiekty będzie obliczana na podstawie poniższych opcji:

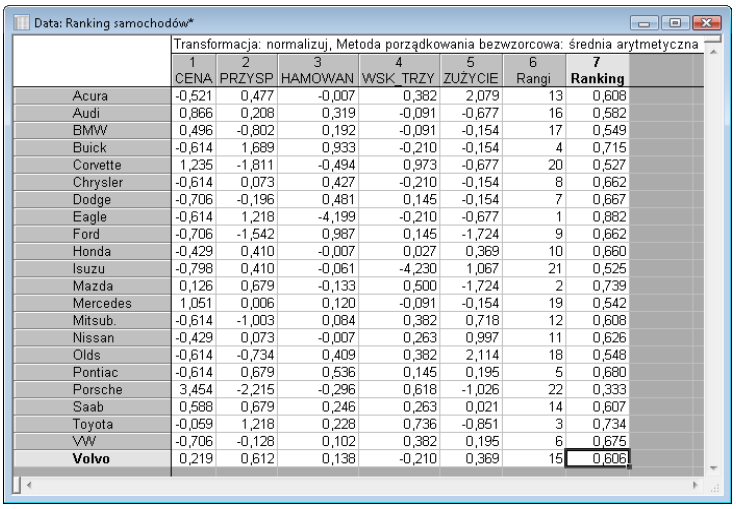
- średnia arytmetyczna
- średnia geometryczna
- średnia harmoniczna

Wybranie opcji ***Wzorcowa*** w zależności od wyboru w polu ***Wzorzec*** skutkuje w pierwszym kroku wyznaczeniem górnego bieguna – maksimum po współrzędnych dla każdej ze zmiennych (opcja ***Górny biegun***) lub dolnego bieguna – minimum po współrzędnych dla każdej ze zmiennych (opcja ***Dolny biegun***). Wartości te (maksymalne lub minimalne) utworzą „idealny obiekt" Następnie dla każdego obiektu obliczana jest odległość od „idealnego obiektu" czyli wzorca. Jako metrykę można zastosować:

- metrykę miejską
- metrykę euklidesową.

W naszym przykładzie wybieramy opcję ***Bezwzorcowa*** oraz średnią arytmetyczną jako sposób agregacji.
Wynik porządkowania liniowego prezentowany jest w postaci arkusza z przypisanymi do przypadków rangami oraz wykresu liniowego, który pozwala graficznie ocenić ranking przypadków. Aby uzyskać wyniki porządkowania, klikamy przycisk ***Raport*** uprzednio zaznaczając opcję ***wejścia***. W wyniku analizy otrzymujemy arkusz, który poza danymi wejściowymi zawiera dwie dodatkowe

zmienne: *Rangi* informującą o kolejności oraz *Ranking* informującą o uzyskanym wyniku w rankingu.

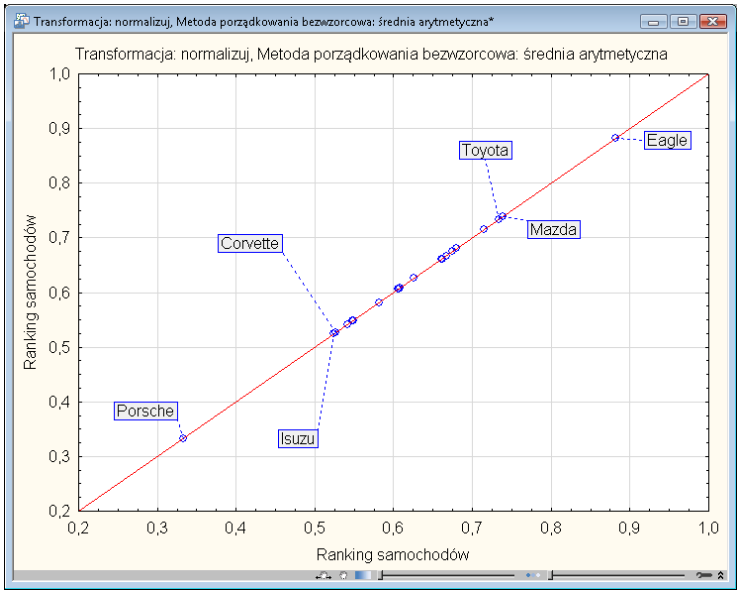


Data: Ranking samochodów*

Transformacja: normalizuj, Metoda porządkowania bezwzorcowa: średnia arytmetyczna

	1 CENA	2 PRZYSP	3 HAMOWAN	4 WSK_TRZY	5 ZUŻYCIE	6 Rangi	7 Ranking
Acura	-0,521	0,477	-0,007	0,382	2,079	13	0,608
Audi	0,866	0,208	0,319	-0,091	-0,677	16	0,582
BMW	0,496	-0,802	0,192	-0,091	-0,154	17	0,549
Buick	-0,614	1,689	0,933	-0,210	-0,154	4	0,715
Corvette	1,235	-1,811	-0,494	0,973	-0,677	20	0,527
Chrysler	-0,614	0,073	0,427	-0,210	-0,154	8	0,662
Dodge	-0,706	-0,196	0,481	0,145	-0,154	7	0,667
Eagle	-0,614	1,218	-4,199	-0,210	-0,677	1	0,882
Ford	-0,706	-1,542	0,987	0,145	-1,724	9	0,662
Honda	-0,429	0,410	-0,007	0,027	0,369	10	0,660
Isuzu	-0,798	0,410	-0,061	-4,230	1,067	21	0,525
Mazda	0,126	0,679	-0,133	0,500	-1,724	2	0,739
Mercedes	1,051	0,006	0,120	-0,091	-0,154	19	0,542
Mitsub.	-0,614	-1,003	0,084	0,382	0,718	12	0,608
Nissan	-0,429	0,073	-0,007	0,263	0,997	11	0,626
Olds	-0,614	-0,734	0,409	0,382	2,114	18	0,548
Pontiac	-0,614	0,679	0,536	0,145	0,195	5	0,680
Porsche	3,454	-2,215	-0,296	0,618	-1,026	22	0,333
Saab	0,588	0,679	0,246	0,263	0,021	14	0,607
Toyota	-0,059	1,218	0,228	0,736	-0,851	3	0,734
VW	-0,706	-0,128	0,102	0,382	0,195	6	0,675
Volvo	0,219	0,612	0,138	-0,210	0,369	15	0,606

Kliknięcie przycisku ***Wykres*** skutkuje wyświetleniem wykresu ilustrującego zbudowany ranking. W przypadku przyjętego przez nas sposobu agregacji, najwyżej w rankingu znajdują się obiekty przedstawione w prawym górnym rogu, najniżej w lewym dolnym rogu. Wybrana przez użytkownika liczba przodujących i ostatnich obiektów może zostać wyróżniona na wykresie. W naszym przypadku pozostawiamy domyślną opcję trzech najlepszych oraz trzech najgorszych.



Na podstawie powyższego rysunku możemy stwierdzić, że pierwsze miejsca w rankingu przypadły samochodom Eagle, Mazda oraz Toyota.

7.7 Bootstrap

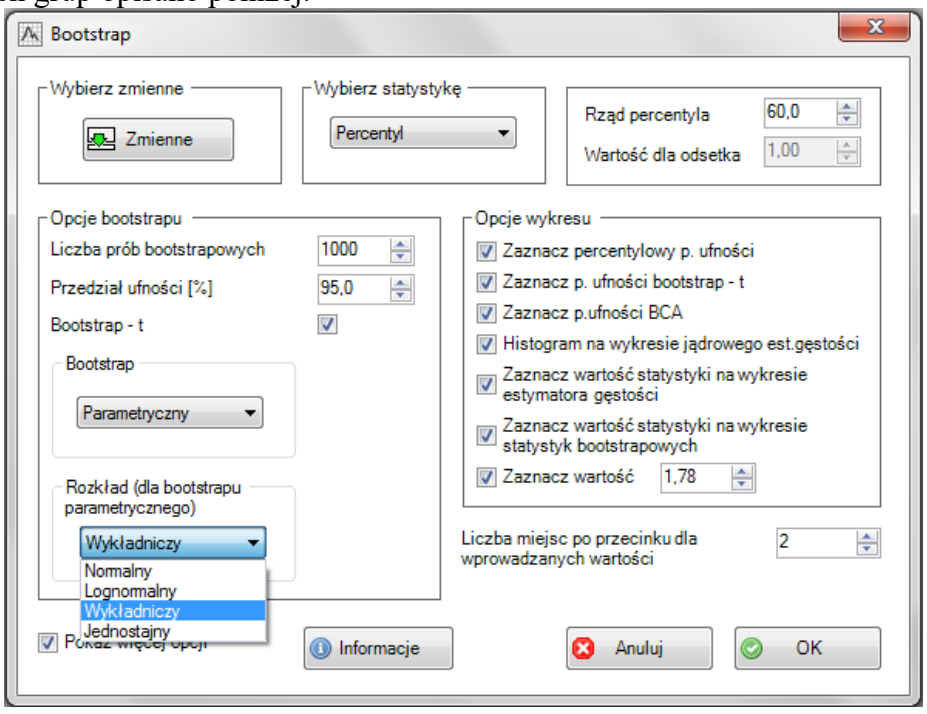
7.7.1 Wprowadzenie

Moduł *Bootstrap* pozwala na obliczenie bootstrapowych przedziałów ufności wybranych statystyk opisowych dla jednej lub dwóch prób, a także wartości błędu standardowego bootstrapu i obciążenia bootstrapu. Przedziały ufności są obliczane metodą percentylową i BCA (*bias corrected and accelerated*), a opcjonalnie także metodą *bootstrap-t*. Wyniki zostają zobrazowane dwoma wykresami - wykresem estymatora jądrowego gęstości rozkładu statystyk bootstrapowych i wykresem rozrzutu tych wartości z zaznaczonymi na nim przedziałami ufności.

Zaletą metod bootstrapowych jest elastyczność oraz mała liczba założeń. Bootstrap jest użyteczny zwłaszcza wtedy, kiedy nie znamy rozkładu, z którego pochodzi próba lub kiedy nie wiadomo jaki rozkład ma statystyka z próby natomiast jego idea polega na wielokrotnym losowaniu ze zwracaniem z próby w celu dobrego oszacowania błędu estymacji danej statystyki opisowej. Wyniki nie są idealnie powtarzalne - wartości liczbowe zazwyczaj odrobinę różnią dla każdej tury obliczeń o tych samych ustawieniach, gdyż za każdym razem obliczane są one poprzez próbkowanie losowe.

7.7.2 Opcje modułu

Wśród opcji wydzielono grupy ustawień dotyczące statystyk opisowych, bootstrapu oraz wykresów. Każdą z tych trzech grup opisano poniżej.



Statystyki

Wyniki, czyli przedziały ufności oraz wykresy, można uzyskać dla różnych statystyk opisowych. Dla pojedynczej próby dostępne opcje to:

- Średnia
- Odchylenie std.
- Percentyl
- Współczynnik skośności
- Współczynnik smukłości
- Odsetek (frakcja)

Natomiast dla dwóch prób można obliczyć następujące statystyki:

- Różnica średnich
- Różnica median
- Iloraz odchyleń std.
- Iloraz odsetków
- Współczynnik korelacji

W przypadku percentyla wskazujemy jego rząd, np. wartość 50 daje medianę a 25 - dolny kwartyl. Odsetek oznacza odsetek wystąpień wskazanej przez nas wartości, w związku z czym iloraz odsetków jest odpowiedni zwłaszcza do badania takich wielkości jak ryzyko względne. Natomiast współczynnik korelacji to klasyczny współczynnik korelacji liniowej Pearsona.

Bootstrap

W tej grupie dostępne są następujące opcje:

- ***Liczba prób bootstrapowych***. Domyślni ustawienie to N = 1000, a najmniejsza możliwa wartość to N = 40. Im większa wartość, tym lepiej dla dokładności wyników. Zaleca się wziąć N ≥ 5000, choć rzecz jasna powoduje to wydłużenie czasu obliczeń.

- ***Rodzaj bootstrapu***. Do wyboru są rodzje: nieparametryczny, półparametryczny, parametryczny. Klasyczna i podstawowa wersja to bootstrap nieparametryczny. Bootstrap półparametryczny różni się od niej tylko tym, że do każdej wartości w próbach bootstrapowych dodajemy zaburzenie o rozkładzie normalnym. Zalecany jest w przypadku szacowania percentyli. Bootstrap parametryczny zakłada, że próba pochodzi z ustalonej rodziny rozkładów i mamy wówczas do wyboru jedną z czterech rodzin: normalne, lognormalne, wykładnicze, jednostajne.

- ***Bootstrap-t.*** Zaznaczenie tej opcji oznacza obliczenie bootstrapowych przedziałów ufności również metodą bootstrap-t. Domyślnie opcja ta jest odznaczona, ze względu na czasochłonność obliczeń.

Wykresy

Jest kilka niezależnych od siebie opcji związanych z wykresami. Na wykresie statystyk bootstrapowych można zaznaczyć:

- Percentylowy przedział ufności
- Przedział ufności wyznaczony metodą BCA
- Przedział ufności wyznaczony metodą bootstrap-t
- Wartości statystyk z oryginalnej próby - czarnym pionowym odcinkiem

Na wykresie estymatora jądrowego można za to zaznaczyć:

- Wartości statystyki z oryginalnej próby - czarnym pionowym odcinkiem
- Gęstości słupków histogramu

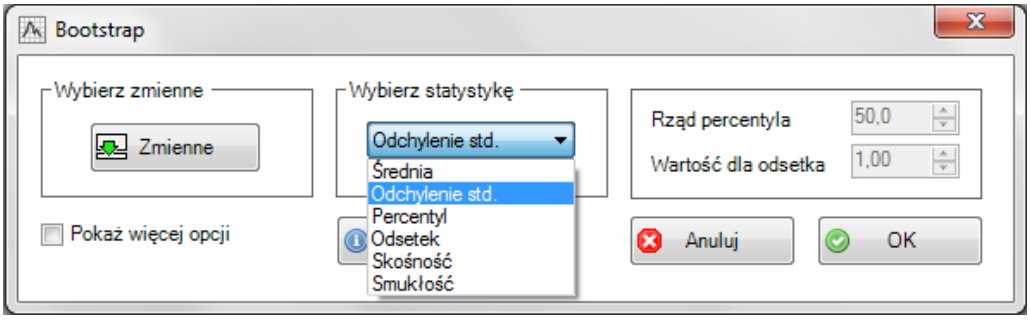
Możliwe jest też zaznaczenie na **obydwu** wykresach wskazanej przez nas wartości czerwonym pionowym odcinkiem

Na wykresie statystyk bootstrapowych przedziały ufności obliczone różnymi sposobami są zaznaczane w takiej samej kolejności (patrząc od góry do dołu), jak w arkuszu z wynikami liczbowymi.

7.12.2 Przykłady

W przykładach wykorzystamy arkusze dołączone do podstawowego pakietu Statistica, znajdujące się w folderze instalacyjnym Statistica w katalogu *Examples/Datasets.*

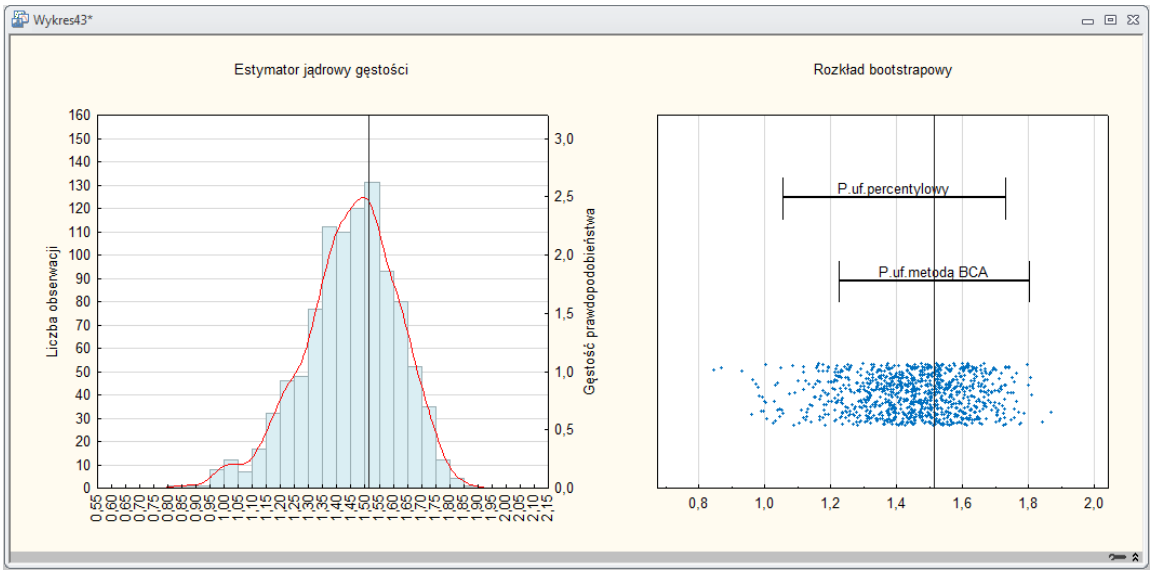
Przykład Otwieramy arkusz *CellCountA.sta*. Wybieramy z menu ***Zestaw Plus*** | ***Analizy***| ***Bootstrap*** w celu obliczenia przedziałów ufności wartości odchylenia standardowego pierwszej zmiennej. Za pomocą przycisku wyboru zmiennych wskazujemy zmienną nr 1 (*Temperature*) na pierwszej liście zmiennych (*Pierwsza próba*) a drugą listę pozostawiamy pustą. Następnie na liście rozwijalnej ***Wybierz statystykę*** wybieramy ***Odchylenie std***.



Zatwierdzamy przyciskiem ***OK***. Jako wynik otrzymujemy poniższy arkusz oraz dwa wykresy.

Dane: Wyniki liczbowe* (3 zmn. * 5 prz.)

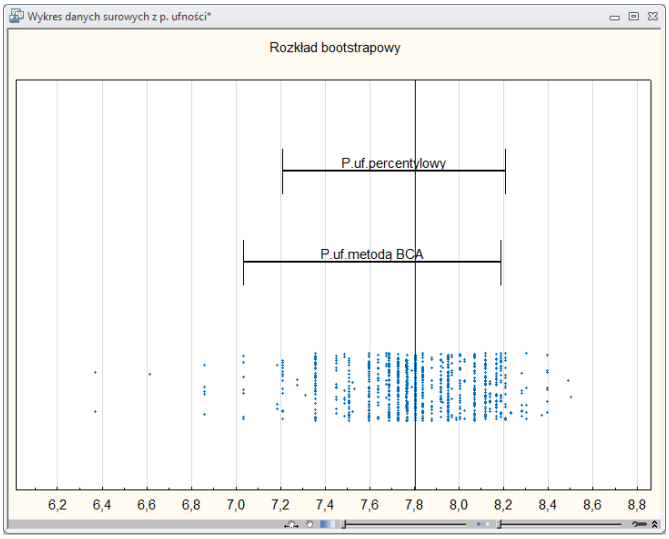
	Bootstrap nieparametryczny, N = 1000		
	Wartość	Dolna granica	Górna granica
Odchylenie std.	1,51273249		
95% p. ufności percentylowy		1,05457444	1,73124121
95% p. ufności metodą BCA		1,22546794	1,80177956
Błąd std. bootstrapu	0,165693766		
Obciążenie bootstrapu	-0,0644629434		



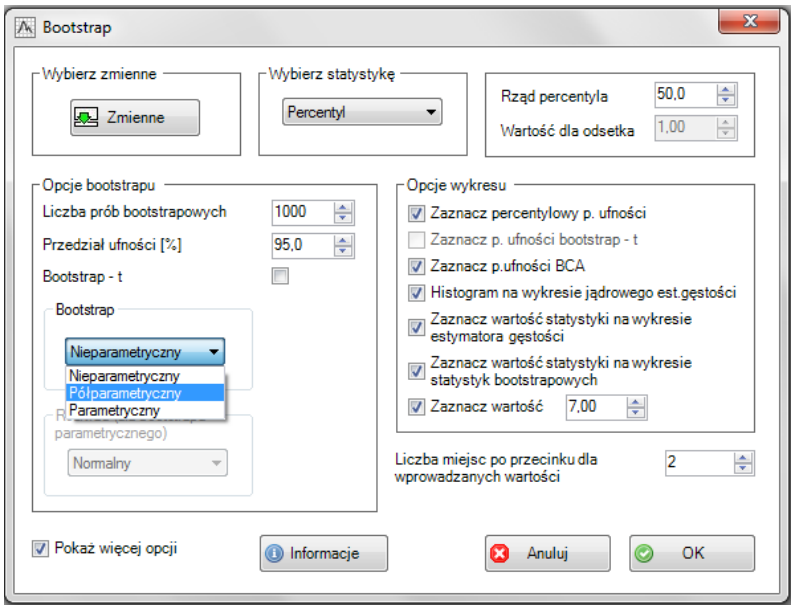
Wyniki mówią, że odchylenie standardowe z próby wynosi około 1,513, z kolei jego 95% percentylowy prz. ufności to [1,055; 1,731] a prz. ufności metodą BCA to [1,225; 1,802]. Błąd std. bootstrapu jest równy 0,166 a obciążenie bootstrapu równa się -0,064. Przypomnijmy, że za każdym razem otrzymamy nieco inne wyniki ze względu na procedurę próbkowania losowego stanowiącą istotę bootstrapu.

Wykres estymatora jądrowego gęstości(po lewej) ukazuje krzywą gęstości dopasowaną do histogramu wartości statystyk bootstrapowych.
Wykres rozkładu bootstrapowego (po prawej) oprócz obliczonych przedziałów ufności pokazuje niebieskimi znacznikami wartości statystyk bootstrapowych. Rzędne tych punktów nie mają znaczenia, różnią się między sobą wyłącznie po to by lepiej uwidocznić rozrzut odciętych.
Na obu wykresach czarna pionowa linia oznacza wartość statystyki z oryginalnej próby.

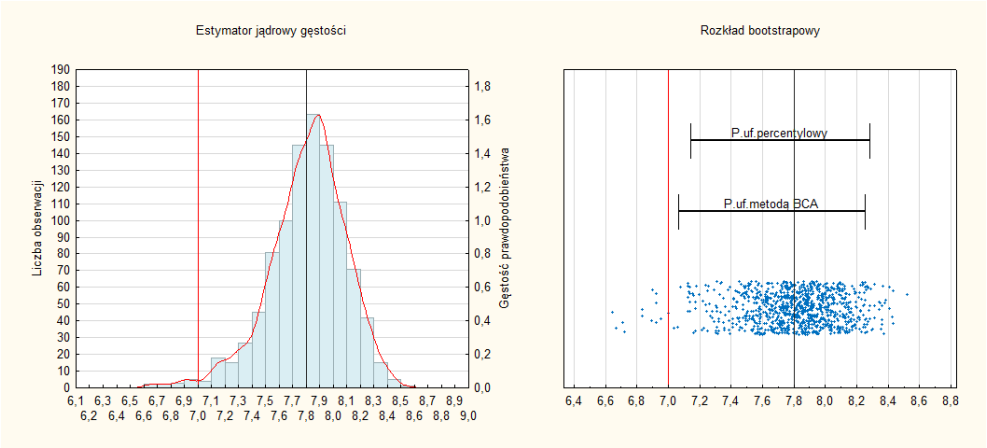
Przykład Otwieramy arkusz *CellCountA.sta*. Wybieramy z menu ***Zestaw Plus*** | ***Analizy***| ***Bootstrap*** w celu obliczenia przedziałów ufności wartości mediany drugiej zmiennej. Za pomocą przycisku wyboru zmiennych wskazujemy zmienną nr 2 (*pH*) na pierwszej liście zmiennych (*Pierwsza próba*) a drugą listę pozostawiamy pustą. Następnie na liście rozwijalnej ***Wybierz statystykę*** wybieramy ***Percentyl*** a w polu ***Rząd percentyla*** pozostawiamy domyślną wartość równą 50. Ze względu na naturę kwantyli, jeśli pozostawilibyśmy inne opcje jako domyślne, otrzymalibyśmy mało różnych wartości statystyk bootstrapowych, co widać na poniższym wykresie rozkładu bootstrapowego.



W związku z tym zaznaczamy pole ***Pokaż więcej opcji*** i w polu ***Bootstrap*** wybieramy ***Półparametryczny.*** Poza tym chcemy zaznaczyć na wykresach wartość 7, więc wprowadzamy ją w polu ***Zaznacz wartość***.



Zatwierdzamy przyciskiem ***OK*** i otrzymujemy w wyniku arkusz oraz dwa niżej ukazane wykresy.



Teraz już wartości statystyk bootstrapowych przyjmują znacznie bardziej zróżnicowane wartości.

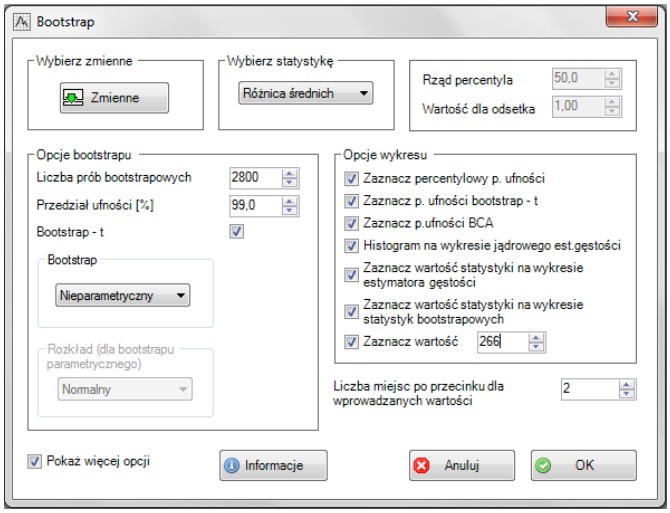
Przykład Otwieramy arkusz *Beverage.sta*. Wybieramy z menu ***Zestaw Plus | Analizy| Bootstrap*** w celu obliczenia przedziałów ufności odsetka występowania wartości 0 w drugiej zmiennej. Za pomocą przycisku wyboru zmiennych wskazujemy zmienną nr 2 (*COKE_N*) na pierwszej liście zmiennych (*Pierwsza próba*) a drugą listę pozostawiamy pustą. Następnie na liście rozwijalnej ***Wybierz statystykę*** wybieramy ***Odsetek*** a w polu ***Wartość dla odsetka*** wpisujemy wartość 0 i zatwierdzamy klikając ***OK***. Oprócz wykresów otrzymujemy arkusz z wynikami liczbowymi jak poniżej.

Dane: Wyniki liczbowe* (3 zmn. * 5 prz.)

	Bootstrap nieparametryczny, N = 1000		
	Wartość	Dolna granica	Górna granica
Odsetek	0,588235294		
95% p. ufności percentylowy		0,426470588	0,764705882
95% p. ufności metodą BCA		0,352941176	0,705882353
Błąd std. bootstrapu	0,0866087694		
Obciążenie bootstrapu	0,00388235294		

Warto porównać uzyskane właśnie granice przedziałów, a zwłaszcza przedziału uzyskanego czysto empirycznie metodą percentylową, z klasycznym przedziałem ufności - wyznaczonym wyłącznie za pomocą wzorów matematycznych. Próba liczy 34 elementy i 20 razy występuje w niej wartość 0 i dlatego po przejściu ścieżki ***Statystyka | Analiza mocy testu | Estymacja przedziałowa | Jedna frakcja*** wpisujemy wartości 34 i 0,5882 (=20/34) odp. w polach *Liczność próby N* i *Obserwowana frakcja p* i potwierdzamy przyciskiem *OK*. Dostajemy w wyniku (dokładny) 95% przedział ufności równy [0,4070; 0,7535].

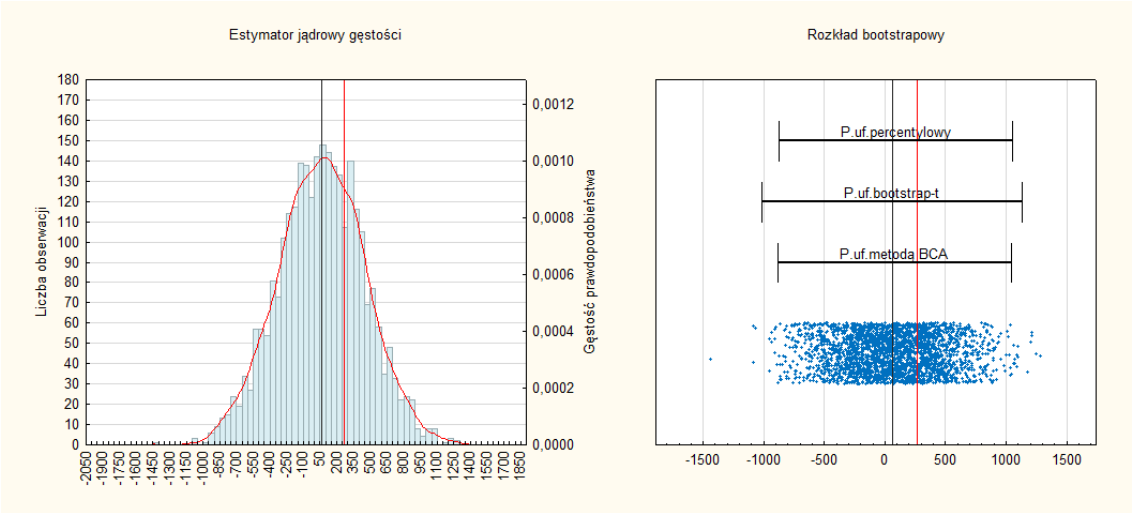
Przykład Otwieramy arkusz *Taguchi.sta* i wybieramy z menu ***Zestaw Plus | Analizy | Bootstrap*** w celu obliczenia przedziałów ufności różnicy średnich zmiennych nr 18 i 19. Za pomocą przycisku wyboru zmiennych wskazujemy te zmienne: na pierwszej liście zmiennych (*Pierwsza próba*) zmienną nr 18 a na drugiej liście (*Druga próba*) zmienną nr 19. Następnie na liście rozwijalnej ***Wybierz statystykę*** wybieramy ***Różnica średnich*** i zaznaczamy pole ***Pokaż więcej opcji***, a na nim pole ***Bootstrap-t*** po czym wpisujemy ***Liczbę prób bootstrapowych*** równą 2800, odsetek ***Przedziału ufności[%]*** równy 99 a w polu ***Zaznacz wartość*** liczbę 266.



Zatwierdzamy klikając ***OK*** i otrzymujemy następujące wyniki:

Dane: Wyniki liczbowe* (3 zmn. * 6 prz.)

	Bootstrap nieparametryczny, N = 2800		
	1 Wartość	2 Dolna granica	3 Górna granica
Różnica średnich	62,39		
99% p. ufności percentylowy		-875,28	1055,14
99% p. ufności bootstrap-t		-1012,72	1137,20
99% p. ufności metodą BCA		-886,33	1041,33
Błąd std. bootstrapu	382,65		
Obciążenie bootstrapu	1,47		



Obliczenie wyników metodą bootstrap-t zajęło więcej czasu, lecz mamy możliwość porównania przedziałów ufności uzyskanych trzema różnymi metodami. Przypominamy, że na wykresach czarna pionowa linia oznacza wartość statystyki(różnica średnich) a czerwona - wskazaną przez nas ręcznie wartość.

7.12.3 **Szczegóły obliczeniowe**

Poniższa sekcja ma zawiera bardziej zaawansowany matematycznie opis dla użytkowników zainteresowanych podejściem boostrapowym od tej strony. Jej znajomość nie jest wymagana do poprawnego korzystania z modułu. Obliczenia i opis są oparte na monografii B.Efron, R.Tibshirani, *An Introduction to the Bootstrap* (Chapman & Hall/CRC 1993).

Niech oryginalna próba będzie oznaczona jako $P = \{X_1, ..., X_k\}$ a θ niech będzie wartością wybranej statystyki obliczonej z niej.

Każda z *N* prób bootstrapowych w ***bootstrapie nieparametrycznym*** jest postaci $P^* = \{X_1^*, ..., X_k^*\}$, gdzie każda wartość X_i^* jest wybraną losowo i niezależnie od pozostałych jedną z wartości z próby *P*.

W ***bootstrapie półparametrycznym*** dodajemy do każdej z wartości X_i^* niezależne zaburzenia o rozkładzie normalnym o wartości oczekiwanej 0 i odchyleniu standardowym równym σ/k, gdzie σ jest odchyleniem standardowym próby *P*.

W ***bootstrapie parametrycznym***, inaczej niż w dwóch poprzednich rodzajach, nie odnosimy się bezpośrednio do pojedynczych wartości próby P, lecz zakładamy, że pochodzi ona z wybranej rodziny rozkładów(normalny/lognormalny/wykładniczy/jednostajny) i szacujemy z niej całej

parametry rozkładu metodą największej wiarygodności. Próby bootstrapowe składają się każda z niezależnych wartości wylosowanych z tego rozkładu.

Mając N prób bootstrapowych $P_1^*, ..., P_N^*$ obliczamy z nich kolejno tę samą co wybrana na początku statystyka, otrzymując ciąg statystyk bootstrapowych $\theta_1^*, ..., \theta_N^*$. Niech θ^* oznacza ich średnią arytmetyczną. Wówczas:

- ***Błąd standardowy bootstrapu*** określony jest wzorem

$$se(\theta): = (\Sigma_i(\theta_i^* - \theta^*)^2/(N - 1))^{1/2}$$

  czyli jest to wartość odchylenia standardowego tego ciągu wartości.

- ***Obciążenie bootstrapu*** to liczba równa różnicy $\theta^* - \theta$.

Mając wybrany poziom ufności α, $100(1 - \alpha)\%$ przedziały ufności są określone jak niżej:

- ***Percentylowy***: jego granice to $[\theta^*(\alpha/2), \theta^*(1-\alpha/2)]$, gdzie $\theta^*(c)$ jest kwantylem rzędu c zbioru statystyk bootstrapowych.

- ***Bootstrap-t***: jego granice to $[\theta - t^*(1-\alpha/2)se(\theta), \theta - t^*(\alpha/2)se(\theta)]$, gdzie $t^*(c)$ jest kwantylem rzędu c zbioru wartości $t_1^*, ..., t_N^*$ określonych wzorem $t_i^* = (\theta_i^* - \theta)/se(\theta_i^*)$. Błąd standardowy $se(\theta_i^*)$ jest obliczany poprzez wylosowanie z próby P_i^* kolejnych prób bootstrapowych, w liczbie $M = 100$, w ten sam sposób co dotychczas, czyli próbkowanie losowe ze zwracaniem. Mamy tu zatem do czynienia z *podwójnym* bootstrapem.

- ***BCA***: jego granice to $[\theta^*(\lambda_1), \theta^*(\lambda_2)]$, gdzie

$$\lambda_1 = \Phi(b+(b+z(\alpha/2))/(1-a(b+z(\alpha/2)))),$$

$$\lambda_2 = \Phi(b+(b+z(1-\alpha/2))/(1-a(b+z(1-\alpha/2))))$$

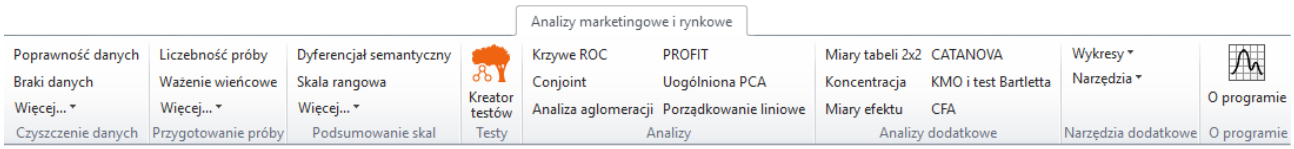
  a Φ i $z(c)$ są odp. dystrybuantą i kwantylem rzędu c standardowego rozkładu normalnego, $b = z(d)$, gdzie d jest proporcją tych statystyk bootstrapowych, które spełniają nierówność $\theta_i^* < \theta$, natomiast

$$a = (1/6)\cdot(\Sigma_i(\theta_{(\cdot)}-\theta_{(i)})^3)/(\Sigma_i(\theta_{(\cdot)}-\theta_{(i)})^2)^{3/2}$$

  gdzie z kolei $\theta_{(i)}$ to wartość statystyki obliczonej na P z pominiętym i-tym przypadkiem a $\theta_{(\cdot)}$ jest ich średnią.

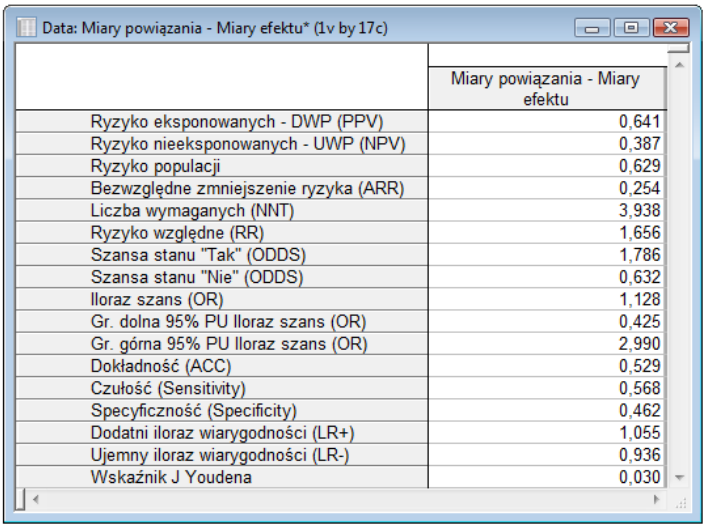
8. Analizy dodatkowe

Grupa **Analizy dodatkowe** zawiera moduły umożliwiające obliczenie miar powiązania\efektu dla tabel 2x2, wykonanie testu post hoc ANOVA Friedmana oraz kart kontrolnych CUSUM ważonych ryzykiem. Dodatkowo użytkownik ma możliwość obliczenia wskaźników koncentracji, analizy CATANOVA, standaryzowanych miar korelacji/efektu oraz CFA.



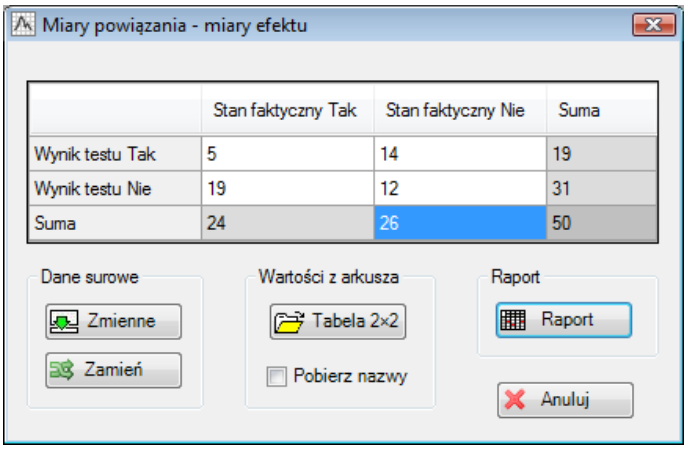
8.1. Miary powiązania/efektu dla tabel 2x2

Moduł przeznaczony jest do obliczania na podstawie tabeli 2x2 szeregu wskaźników powiązania lub efektu. Na przykład umożliwia on obliczenie efektu związanego z binarną zmienną zależną, spowodowanego manipulacją zmienną niezależną. Moduł umożliwia zarówno obliczenie wskaźników na podstawie danych surowych, jak również ręczne wprowadzenie lub korektę wartości w tabeli.



	Miary powiązania - Miary efektu
Ryzyko eksponowanych - DWP (PPV)	0,641
Ryzyko nieeksponowanych - UWP (NPV)	0,387
Ryzyko populacji	0,629
Bezwzględne zmniejszenie ryzyka (ARR)	0,254
Liczba wymaganych (NNT)	3,938
Ryzyko względne (RR)	1,656
Szansa stanu "Tak" (ODDS)	1,786
Szansa stanu "Nie" (ODDS)	0,632
Iloraz szans (OR)	1,128
Gr. dolna 95% PU Iloraz szans (OR)	0,425
Gr. górna 95% PU Iloraz szans (OR)	2,990
Dokładność (ACC)	0,529
Czułość (Sensitivity)	0,568
Specyficzność (Specificity)	0,462
Dodatni iloraz wiarygodności (LR+)	1,055
Ujemny iloraz wiarygodności (LR-)	0,936
Wskaźnik J Youdena	0,030

Przykład Dla zadanych liczności tabeli 2x2 możemy wyznaczyć miary powiązania/efektu. Aby uzyskać miary powiązania/efektu, wybieramy opcję ***Miary powiązania/efektu 2x2*** z paska narzędzi ***Analizy marketingowe i rynkowe | Analizy dodatkowe***. W oknie ***Miary powiązania – miary efektu*** wprowadzamy liczności opisujące wynik testu i stan faktyczny.

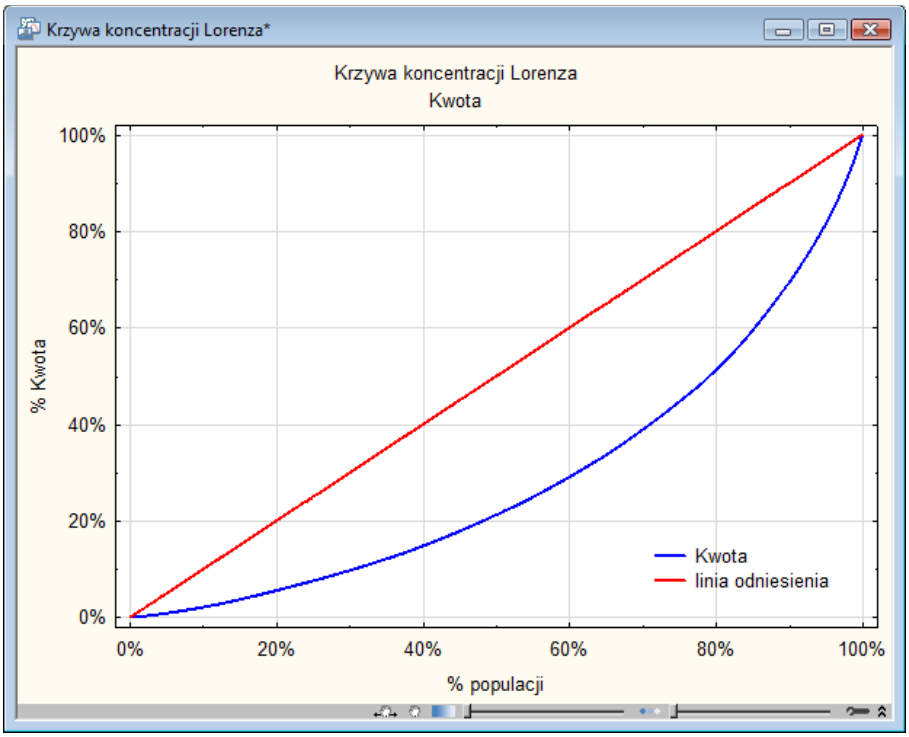


Po kliknięciu przycisku Raport otrzymujemy skoroszyt z wynikami, zawierający miary: Ryzyko eksponowanych, Ryzyko nieeksponowanych, Ryzyko populacji, ARR, NNT, Ryzyko względne, Szansę sukcesu, Iloraz szans, Przedział ufności dla ilorazu szans, Poprawność frakcji, Czułość, Specyficzność, Dodatni iloraz wiarygodności, Ujemny iloraz wiarygodności, Wskaźnik J Youdena obliczone na podstawie zadanych liczności.

Wskaźniki możemy policzyć również na podstawie danych surowych. W obszarze ***Dane surowe*** wskazujemy zmienne, na podstawie których chcielibyśmy obliczyć tabelę 2x2. Wyniki obliczeń zostaną automatycznie wprowadzone do tabelki. Za pomocą przycisku ***Zamień*** możemy zamienić klasy określające interpretację wyniku testu oraz stanu faktycznego.

8.2. Analiza koncentracji

Moduł ten pozwala obliczyć miary koncentracji dla pojedynczych cech. Koncentracja jest tutaj rozumiana jako nierównomierny podział ogólnej sumy wartości analizowanej zmiennej pomiędzy poszczególne przypadki należące do analizowanego zbioru. Bardzo często tego typu analizy wykonuje się przy badaniu dochodów, koncentracji produkcji, gęstości zaludnienia itp. Koncentracja może być mierzona za pomocą wskaźników *Giniego* bądź *Herfindahla* oraz przedstawiona za pomocą krzywej koncentracji *Lorenza*.



8.3. Standaryzowane miary efektu

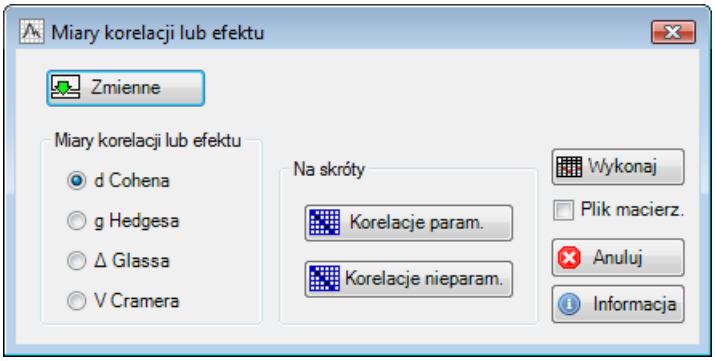
Moduł umożliwia obliczenie miar pozwalających na zbadanie siły związku pomiędzy dwiema zmiennymi w analizowanej zbiorowości. W module zaimplementowano trzy miary efektu –

- d Cohena,
- g Hedgesa,
- Δ Glassa

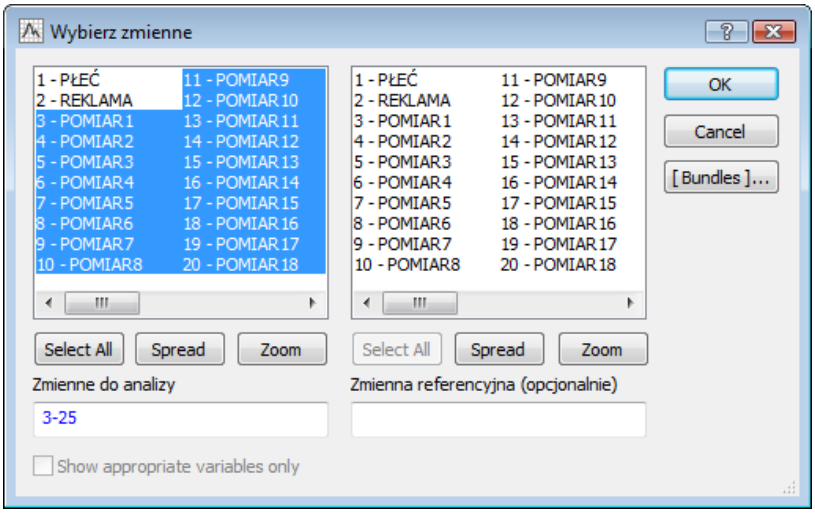
oraz współczynnik korelacji:

- V Cramera.

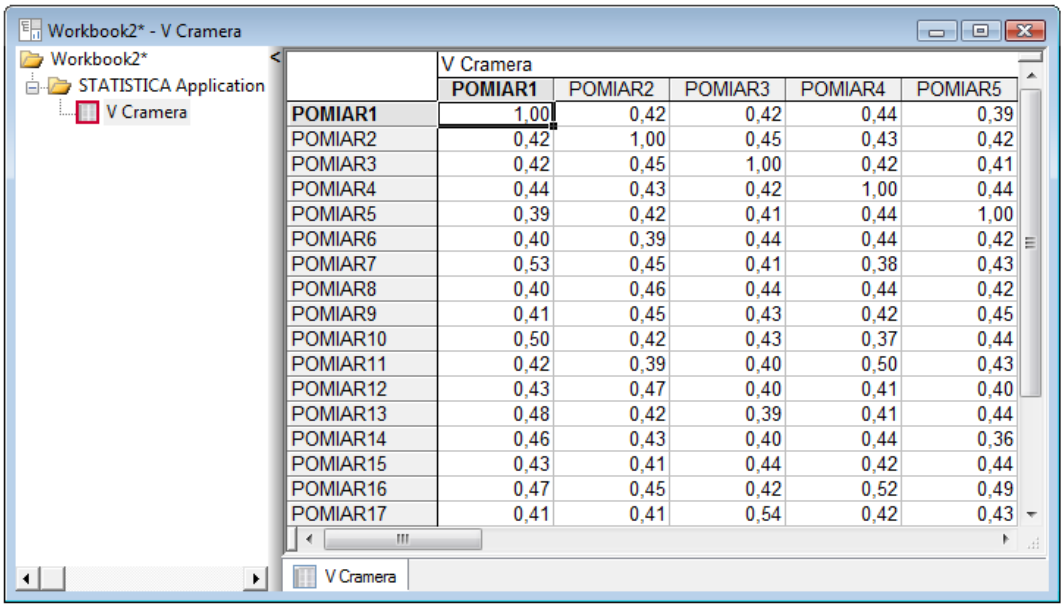
Przykład. Obliczenia wskaźników koncentracji wykonamy na podstawie zbioru *Adstudy.sta*. Z menu ***Analizy marketingowe i rynkowe*** | ***Analizy dodatkowe*** wybieramy polecenie ***Standaryzowane miary efektu*** przywołując okno o tej samej nazwie.



W oknie tym po kliknięciu na przycisk ***Zmienne*** wybieramy zmienne do analizy.



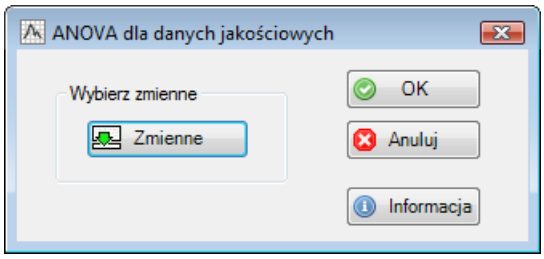
Na liście ***Zmienne do analizy*** wybieramy zmienne 3-25, natomiast nie wskazujemy zmiennej referencyjnej. Po określeniu zmiennych w oknie ***Miary efektu*** wybieramy opcję ***V Cramera*** i klikamy przycisk ***Wykonaj***, co spowoduje utworzenie skoroszytu zawierającego arkusz z obliczoną macierzą korelacji dla miary V Cramera.



	POMIAR1	POMIAR2	POMIAR3	POMIAR4	POMIAR5
POMIAR1	1,00	0,42	0,42	0,44	0,39
POMIAR2	0,42	1,00	0,45	0,43	0,42
POMIAR3	0,42	0,45	1,00	0,42	0,41
POMIAR4	0,44	0,43	0,42	1,00	0,44
POMIAR5	0,39	0,42	0,41	0,44	1,00
POMIAR6	0,40	0,39	0,44	0,44	0,42
POMIAR7	0,53	0,45	0,41	0,38	0,43
POMIAR8	0,40	0,46	0,44	0,44	0,42
POMIAR9	0,41	0,45	0,43	0,42	0,45
POMIAR10	0,50	0,42	0,43	0,37	0,44
POMIAR11	0,42	0,39	0,40	0,50	0,43
POMIAR12	0,43	0,47	0,40	0,41	0,40
POMIAR13	0,48	0,42	0,39	0,41	0,44
POMIAR14	0,46	0,43	0,40	0,44	0,36
POMIAR15	0,43	0,41	0,44	0,42	0,44
POMIAR16	0,47	0,45	0,42	0,52	0,49
POMIAR17	0,41	0,41	0,54	0,42	0,43

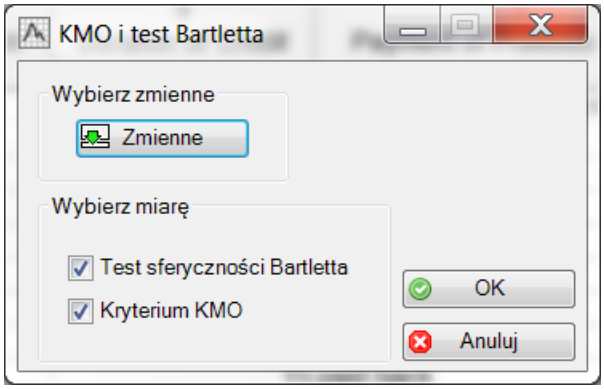
8.4. CATANOVA

Moduł umożliwia obliczenie analizy danych jakościowej analogicznej do analizy wariancji.



8.5. KMO i test Bartletta

Zaimplementowane miary umożliwiają sprawdzenie zasadności wykonywania analizy głównych składowych (PCA) na analizowanym zbiorze danych. Moduł umożliwia obliczenie kryterium Kaisera-Mayera-Olkina (KMO) zarówno na poziomie ogólnym, jak i dla poszczególnych zmiennych. Wykonuje również test sferyczności Bartletta.

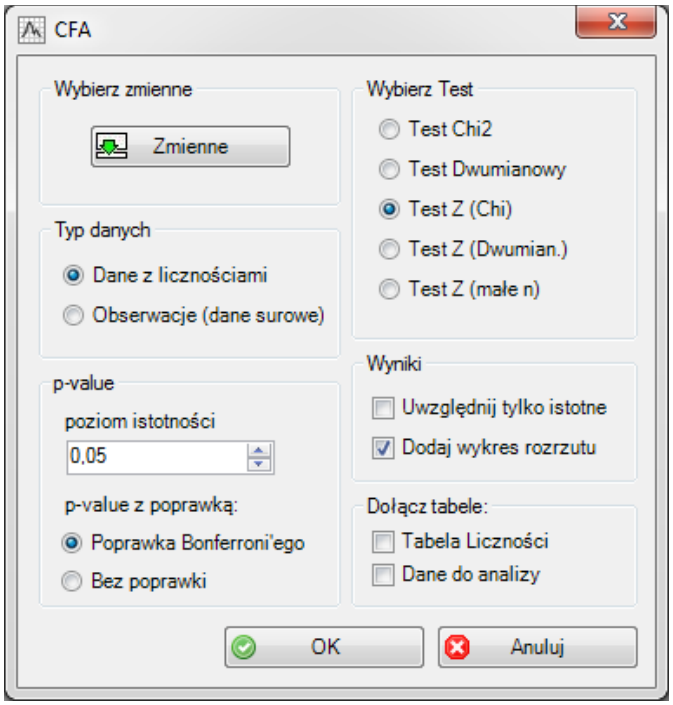


8.6. Konfiguracyjna analiza częstości (CFA)

CFA (*Configural Frequency Analysis*) jest narzędziem służącym do wyszukiwania wzorców i schematów w tabelach wielodzielczych. Pozwala odpowiedzieć na pytanie, czy wśród zgromadzonych danych występują pewne schematy pojawiające się częściej (typ) bądź rzadziej (anty-typ) niż byśmy się tego spodziewali. Wykorzystywana jest na przykład w naukach społecznych (określanie typów, wzorców zachowań klientów/pacjentów), czy badaniach skuteczności nowych programów nauczania (badania na dwóch grupach, w jednej wykorzystywano nowe metody, w drugiej stare; badanie porównuje postępy w obu grupach).

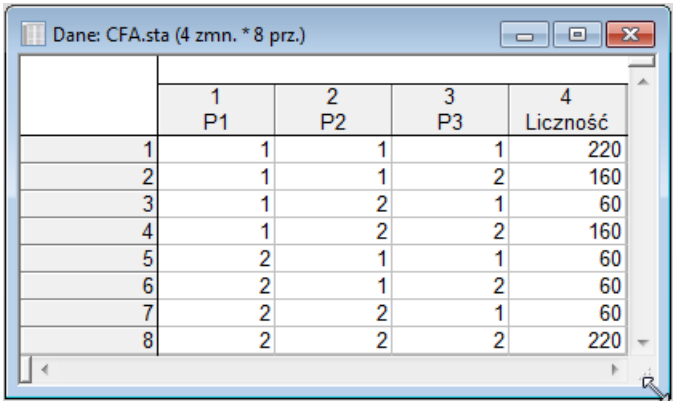
Moduł CFA umożliwia:

- budowę tabel liczności w przypadku danych surowych (zawierających pojedyncze obserwacje),
- obliczenie częstości brzegowych,
- obliczenie wartości oczekiwanych,
- przeprowadzenie wybranego testu sprawdzającego występowanie typów/anty-typów w tabeli liczności.



Dodatkowo moduł umożliwia badaczowi określenie zakresu wyników, jakie powinny trafić do raportu.

Przykład Ilustrację możliwości modułu oprzemy na danych przygotowanych przez *Lazarsfelda* i *Henriego* (1968). (plik *CFA.sta* znajdujący się w zestawie danych dołączonych do dokumentacji) W danych mamy 1000 przypadków dot. rozwiązania trzech problemów (zmienne od *P1* do *P3*). Wartość 1 oznacza problem rozwiązany; 2 odnosi się do problemu nierozwiązanego. Ostatnia kolumna (*Liczność*) informuje o liczbie osób, które uzyskały takie same rezultaty. Przykładowo pierwszy wiersz informuje, że było 220 osób, które rozwiązały wszystkie trzy problemy.


Dane: CFA.sta (4 zmn. * 8 prz.)

	1 P1	2 P2	3 P3	4 Liczność
1	1	1	1	220
2	1	1	2	160
3	1	2	1	60
4	1	2	2	160
5	2	1	1	60
6	2	1	2	60
7	2	2	1	60
8	2	2	2	220

Naszym zadaniem jest określenie czy pewna kombinacja występuje częściej/rzadziej od sytuacji, w której analizowane cechy byłyby od siebie niezależne. W tym celu użyjemy narzędzia CFA. W wyniku analizy otrzymujemy poniższą tabelę testującą wszystkie kombinacje pod tym kątem.

Dane: CFA (8 zmn. * 8 prz.)

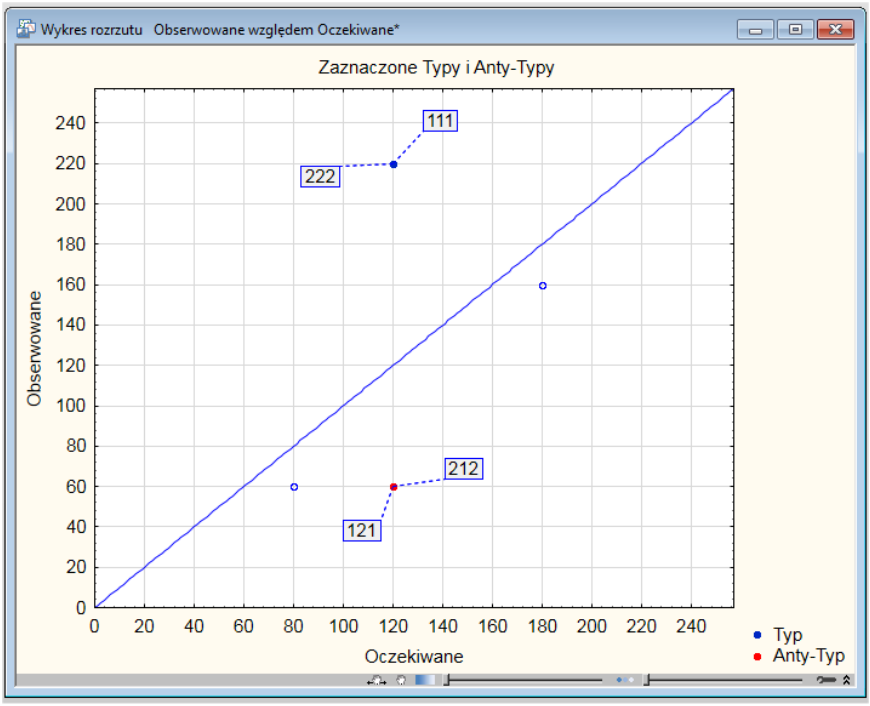
Tabela zwraca następujące wartości:
Obserwowane: Rzeczywiste zaobserwowane częstości występowania danej kombinacji zmiennych
Oczekiwane: Średnie oczekiwane częstości występowania danej kombinacji zmiennych
z: Wartość statystyki standaryzowanego rozkładu normalnego
p: Wartość p-value dla statystyki z
Typ/Antytyp: czy dana kombinacja zmiennych jest istotnie większa/mniejsza od wartości oczekiwanej

	V1	V2	V3	Obserwowane	Oczekiwane	z Chi	p	Typ/Antytyp
111	1	1	1	220	120	9,13	0,0000	T
112	1	1	2	160	180	1,49	0,0680	
121	1	2	1	60	120	5,48	0,0000	A
122	1	2	2	160	180	1,49	0,0680	
211	2	1	1	60	80	2,24	0,0127	
212	2	1	2	60	120	5,48	0,0000	A
221	2	2	1	60	80	2,24	0,0127	
222	2	2	2	220	120	9,13	0,0000	T

W nagłówku tabeli znajduje się opis poszczególnych kolumn wyników. Ostatnia kolumna informuje czy dana kombinacja to typ (wartość T), anty-typ (wartość A), czy też żadna z nich. Na czerwono zostały oznaczone kombinacje, których różnica między wartością obserwowaną a oczekiwaną jest istotna statystycznie.

W naszym przykładzie częściej (typ) występują sytuacje, gdy wszystkie problemy zostały rozwiązane lub nie został rozwiązany żaden. Rzadziej niż byśmy się spodziewali (anty-typ) występują sytuacje, gdy poprawnie rozwiązano jedynie problem 1 i 3 oraz gdy problemy 1 i 3 nie zostały rozwiązane, a problem 2 tak.

Uzyskane wyniki możemy również przedstawić na wykresie.



Linia na wykresie przedstawia sytuację, gdy wartości obserwowane są równe wartościom oczekiwanym. Punkty leżące nad linią oznaczają obserwacje będące typem, a zaetykietowane punkty pod linią to anty-typy. Etykiety punktów są tożsame z wynikami kolejnych problemów.

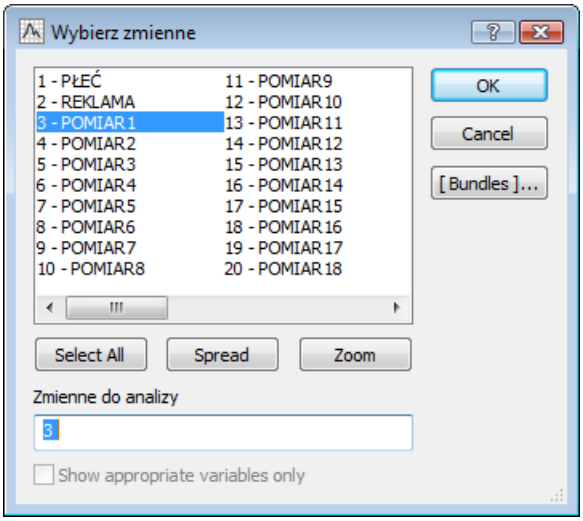
9. Wykresy

Kolejna grupa modułów to ***Wykresy*** umożliwiająca utworzenie szeregu dostosowanych wykresów. Dostępne są wykresy *słupkowy z kolorowymi słupkami*, wykres *sekwencyjny*, *radarowy*, *mozaikowy* oraz *kołowy (Spie plot)*.

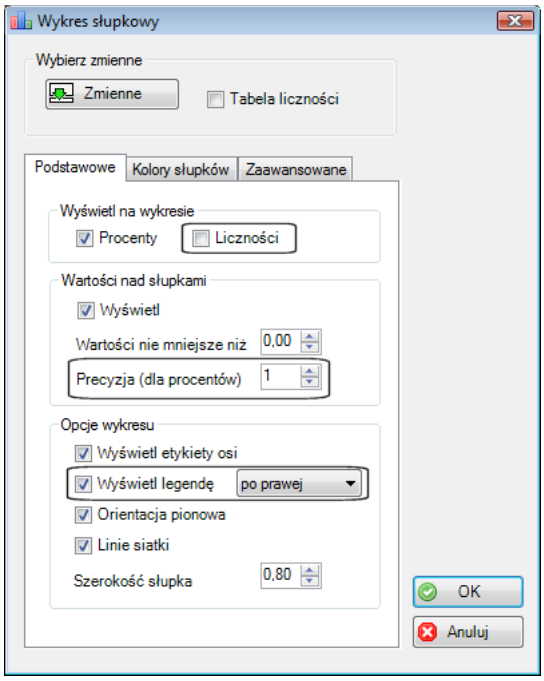
9.1. Wykres słupkowy (kolorowe słupki)

Wykres słupkowy jest typowym wykresem dostępnym w programie *Statistica*. Moduł zawarty w ***Zestawie do analiz marketingowych i rynkowych*** został dodatkowo zmodyfikowany, aby umożliwić (między innymi) niezależną zmianę kolorów wyświetlanych słupków.

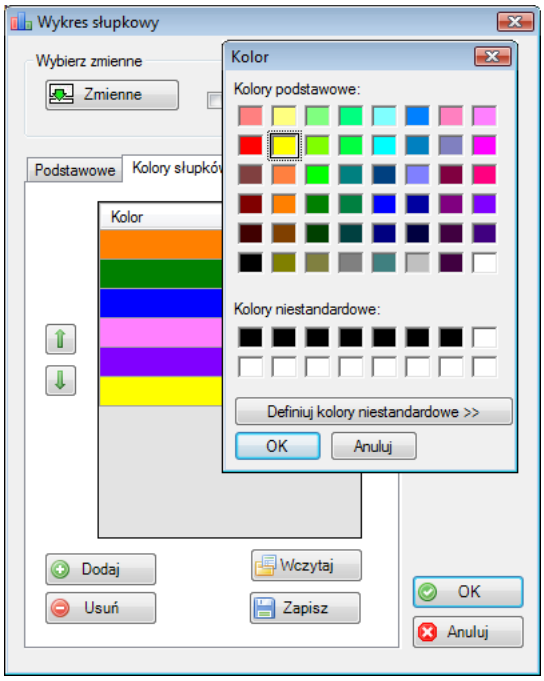
Przykład. Przedstawimy teraz przykład przygotowania dostosowanego wykresu słupkowego na przykładzie pliku *Adstudy.sta*. Z menu ***Analizy marketingowe i rynkowe*** | ***Wykresy*** wybieramy polecenie ***Wykres słupkowy (kolorowe słupki)*** przywołując okno ***Wykres słupkowy***. W pierwszej kolejności określamy zmienne, jakie będziemy analizować. W tym celu klikamy przycisk ***Zmienne***, a następnie na liście ***Zmienne do analizy*** wybieramy zmienną *POMIAR1*.



Następnie na karcie ***Podstawowe*** anulujemy zaznaczenie opcji ***Liczności*** w obszarze ***Wyświetl na wykresie***. W obszarze ***Wartości nad słupkami*** zmieniamy wartość pola ***Precyzja (dla procentów)*** na 1- wartości na wykresie zostaną wyświetlone z dokładnością jednego miejsca po przecinku. Dodatkowo w obszarze ***Opcje wykresu*** dla opcji ***Wyświetl legendę*** wybieramy wartość ***po prawej***, aby była ona wyświetlana po prawej stronie wykresu.

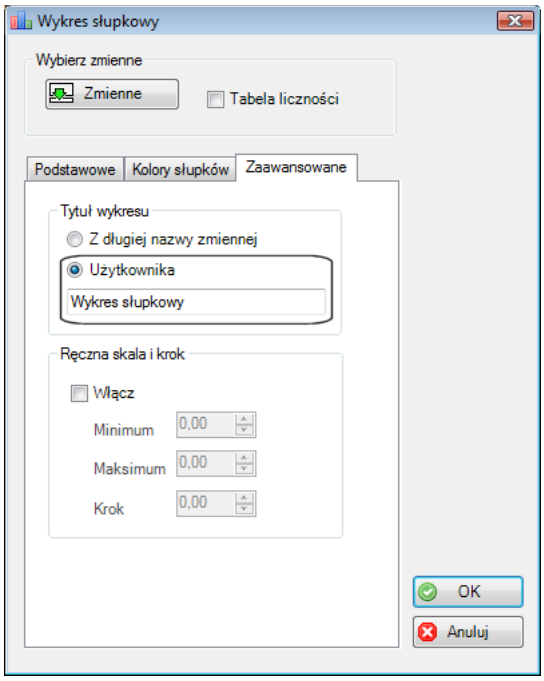


Następnie przechodzimy na kartę ***Kolory słupków***, aby określić kolory kategorii na wykresach.

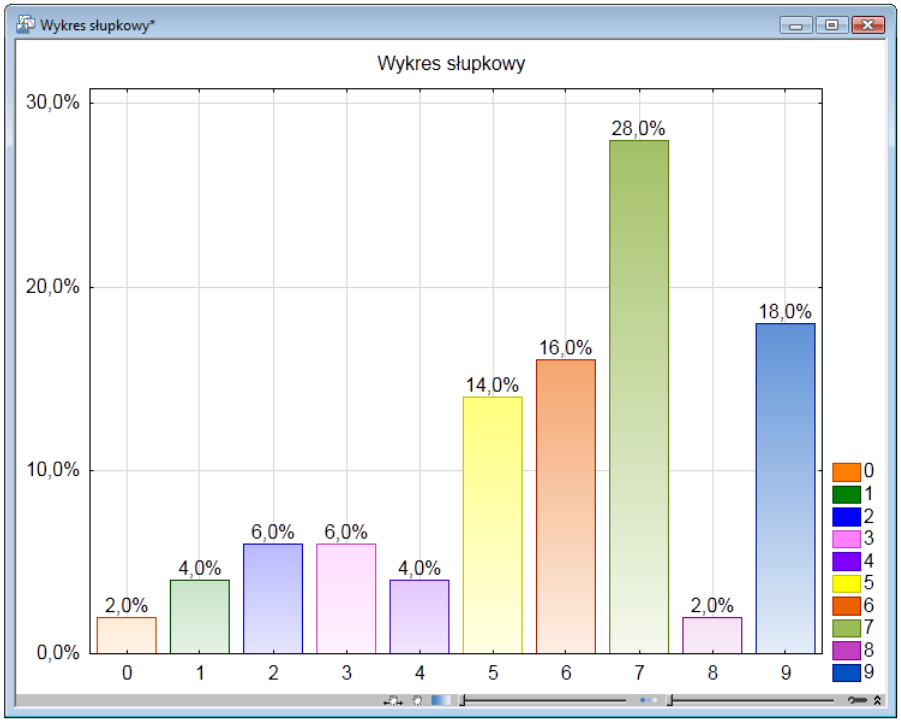


Kolejny kolor dodajemy za pomocą przycisku ***Dodaj*** co spowoduje pojawienie się nowego wiersza z kolorem w tabeli. Kolory zmieniamy za pomocą standardowego okna dialogowego ***Kolor***, które przywołujemy klikając dwukrotnie na wybranym wierszu na liście kolorów. Jeżeli na liście nie zdefiniowaliśmy kolorów lub jest ich mniej niż kategorii na wykresie, program uzupełni brakujące kolory z domyślnego zestawu kolorów. Po zdefiniowaniu kolorów możemy je następnie zapisać (przycisk ***Zapisz***) do pliku, aby użyć ich w innych wykresach (przycisk ***Wczytaj***). Przyciski ze strzałkami pozwalają zmienić kolejność kolorów na wykresie.

Po określeniu odpowiednich kolorów przechodzimy na kartę ***Zaawansowane***, aby określić nazwę wykresu. Domyślnie program czyta ją z długiej nazwy analizowanej zmiennej. My zmienimy tą domyślną opcję klikając ***Użytkownika*** a następnie w polu edycji wpisując *Wykres słupkowy*.



Klikamy ***OK*** otrzymując wykres słupkowy z procentami na wykresie o zadanej precyzji oraz wskazanymi kolorami słupków.

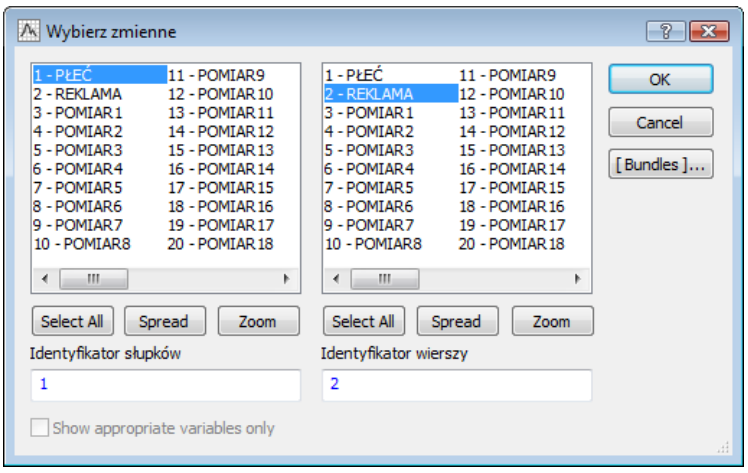


9.2. Wykres sekwencyjny

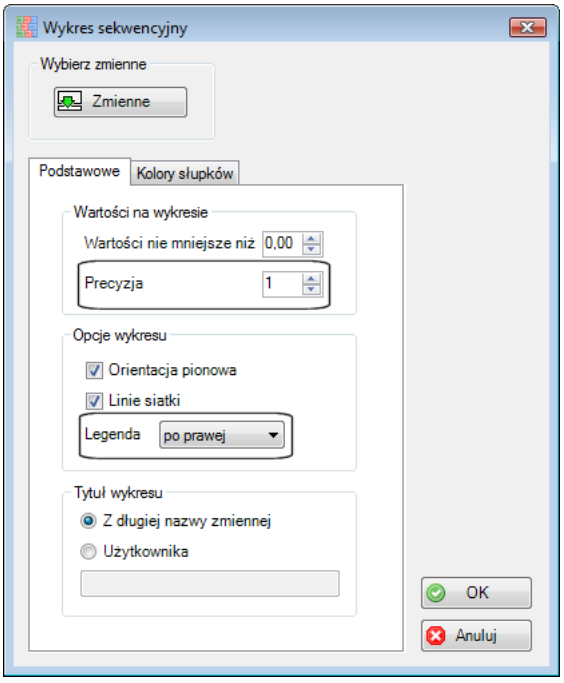
Wykresy **sekwencyjny** jest typowym wykresem dostępnym w programie *Statistica*. Moduł zawarty w ***Zestawie do analiz marketingowych i rynkowych*** został dodatkowo zmodyfikowany, aby umożliwić między innymi niezależną zmianę kolorów wyświetlanych słupków.

Przykład. Przedstawimy teraz przykład przygotowania dostosowanego wykresu sekwencyjnego na przykładzie pliku *Adstudy.sta*. Z menu ***Analizy marketingowe i rynkowe*** | ***Wykresy*** wybieramy polecenie ***Wykres sekwencyjny*** przywołując okno o tej samej nazwie. W pierwszej kolejności określamy zmienne, jakie będziemy analizować. W

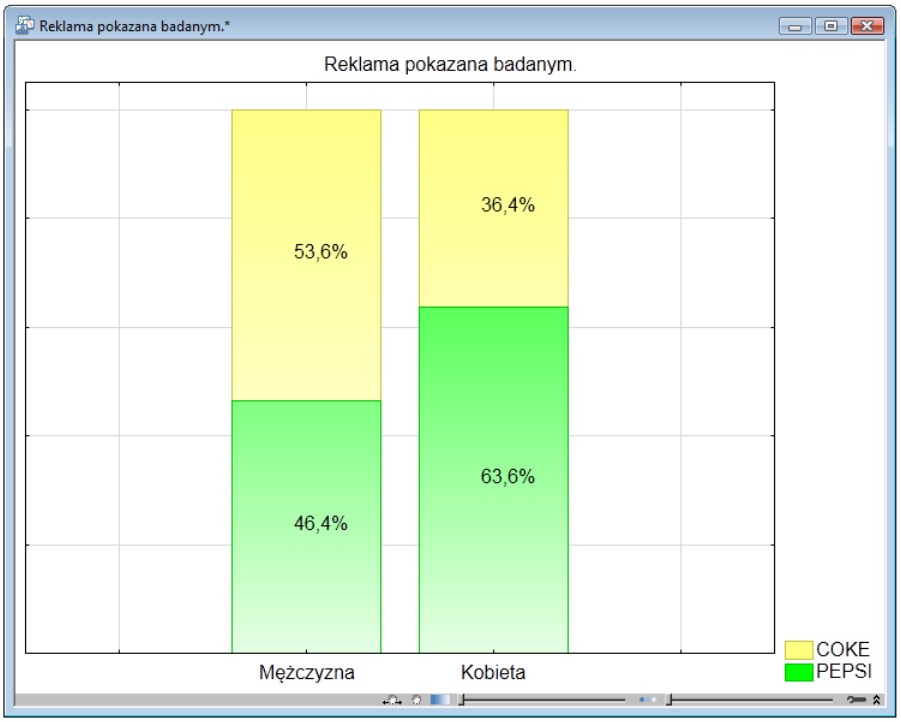
tym celu klikamy przycisk ***Zmienne***, a następnie na liście ***Zmienne do analizy*** wybieramy zmienną *PŁEĆ* (otrzymamy zatem dwa słupki reprezentujące kobiety i mężczyzn), na liście ***Identyfikator wierszy*** wybieramy zmienną *REKLAMA*.



Następnie na karcie ***Podstawowe*** zmieniamy wartość pola ***Precyzja*** na 1- wartości na wykresie zostaną wyświetlone z dokładnością jednego miejsca po przecinku. Dodatkowo na liście ***Legenda*** wskazujemy, aby była ona wyświetlana po prawej stronie wykresu.



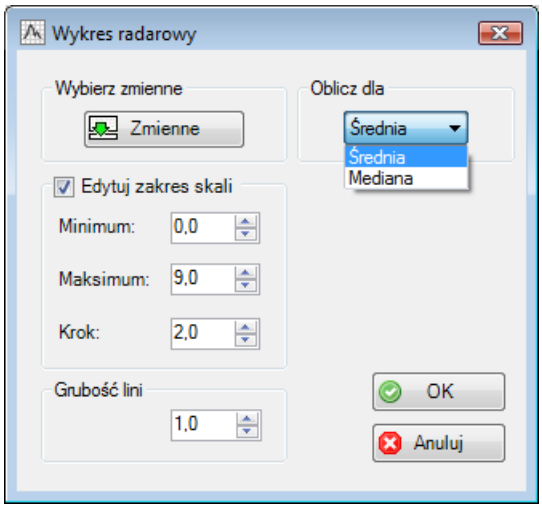
Następnie przechodzimy na kartę ***Kolory słupków***, aby wybrać kolory kategorii na wykresach. Opcje kolorów słupków są analogiczne do tych opisanych w przykładzie dla wykresu słupkowego. Po określeniu odpowiednich kolorów klikamy ***OK*** otrzymując wykres sekwencyjny z wartościami na wykresie o zadanej precyzji oraz wskazanymi kolorami słupków.



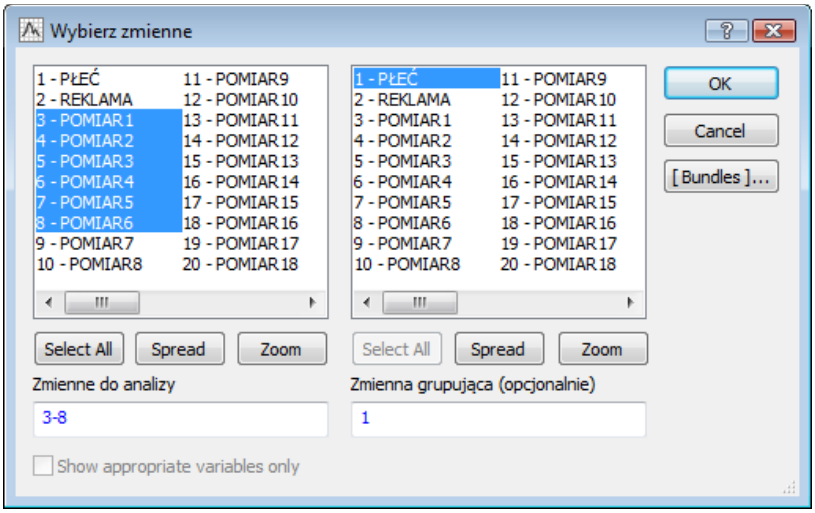
9.3. Wykres radarowy

Moduł umożliwia utworzenie wykresu radarowego dla wskazanej liczby wymiarów i obiektów. Obliczenia można wykonywać dla średnich oraz median.

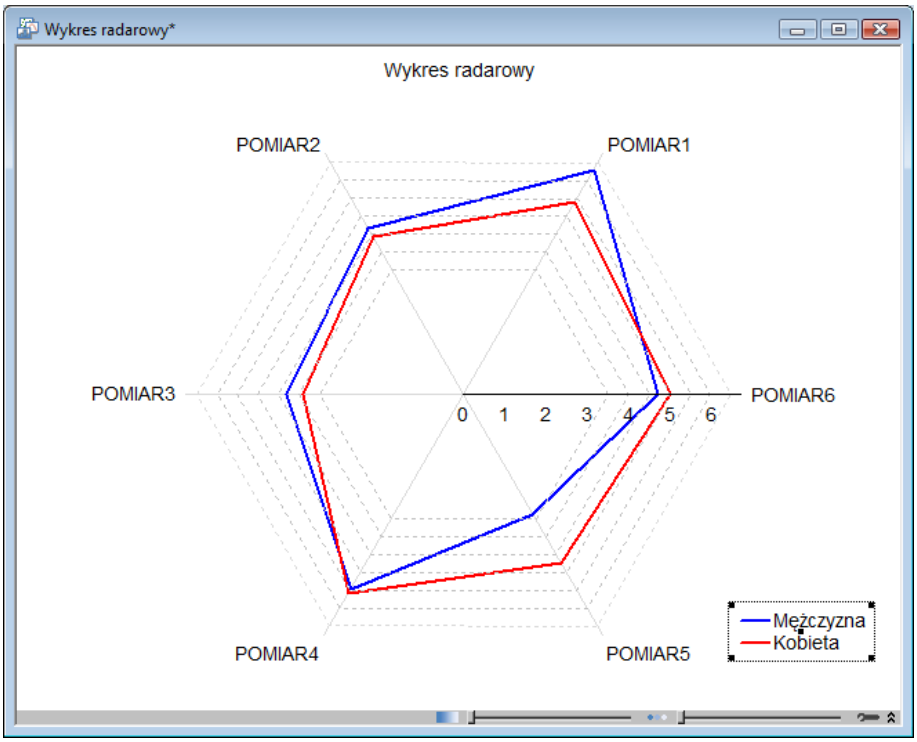
Przykład. Obliczenia wykresu radarowego wykonamy na podstawie zbioru *Adstudy.sta*. Z menu ***Analizy marketingowe i rynkowe*** | ***Wykresy*** wybieramy polecenie ***Wykres radarowy*** przywołując okno o tej samej nazwie.



Przyjmijmy, że interesują nas różnice w ocenach kobiet i mężczyzn dla pierwszych sześciu pomiarów. Aby wybrać interesujące nas cechy klikamy przycisk ***Zmienne*** a następnie na liście ***Zmienne do analizy*** wybieramy zmienne *POMIAR1* do *POMIAR6* (zmienne 3-8) na liście ***Zmienna grupująca*** wskazujemy zmienną *PŁEĆ*.



Po wybraniu zmiennych zaznaczamy opcję ***Edytuj zakres skali***, a następnie w polu ***Maksimum*** wpisujemy wartość 9, a w polu ***Krok*** wartość 2. W polu ***Oblicz dla*** wybieramy opcję ***Średnia***, a następnie zatwierdzamy analizę.



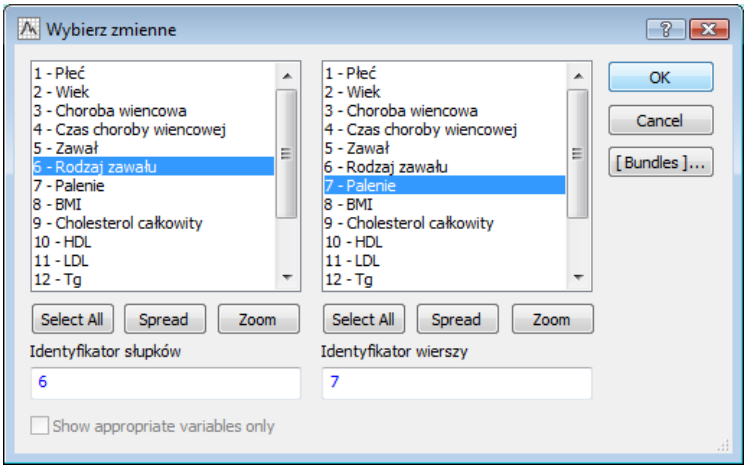
Na uzyskanym wykresie radarowym przedstawione zostały średnie wartości analizowanych cech dla mężczyzn oraz dla kobiet. Analizując uzyskane profile możemy zauważyć dosyć duże podobieństwo odpowiedzi w obydwu grupach. Jedynie dla zmiennej *POMIAR5* średnia odpowiedź w grupie kobiet jest wyraźnie większa od średniej odpowiedzi w grupie mężczyzn.

9.4. Wykres mozaikowy

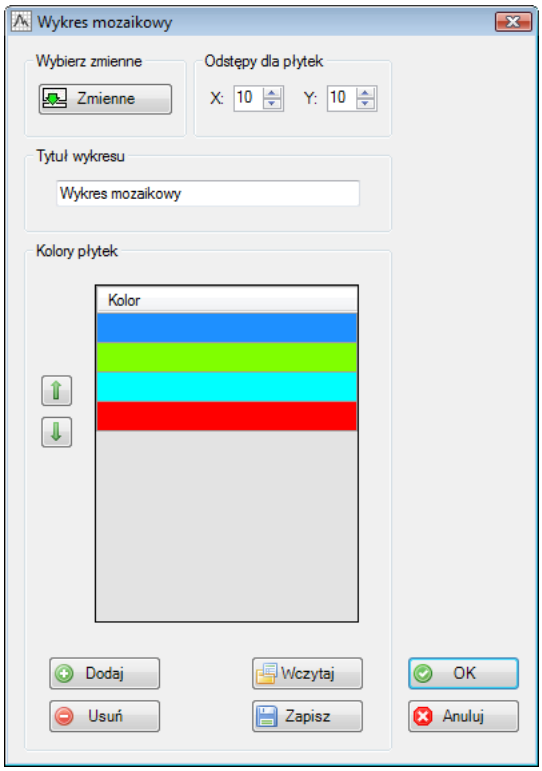
Wykres mozaikowy pozwala w prosty i intuicyjny sposób przedstawić wartości tabeli dwudzielczej (tabeli kontyngencji).

Przykład. Przedstawimy teraz przykład przygotowania wykresu mozaikowego na przykładzie pliku *Zawały.sta*. Z menu ***Analizy marketingowe i rynkowe*** | ***Wykresy*** wybieramy polecenie ***Wykres mozaikowy*** przywołując okno o tej samej nazwie.

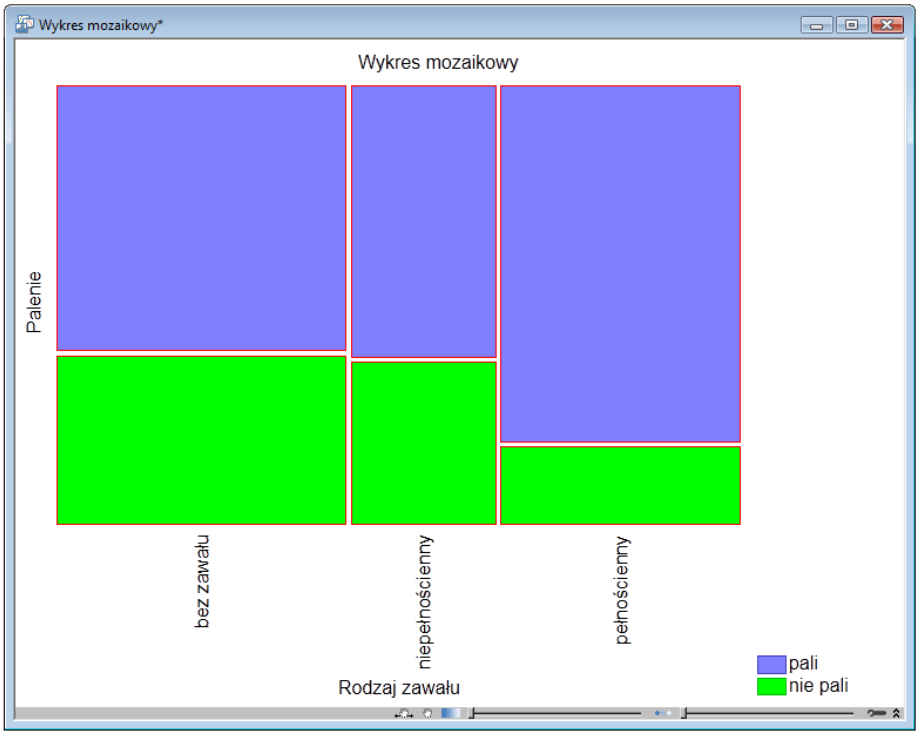
W pierwszej kolejności określamy zmienne, jakie będziemy analizować. W tym celu klikamy przycisk ***Zmienne***, a następnie na liście ***Identyfikator słupków*** wybieramy zmienną *Rodzaj zawału* (otrzymamy zatem trzy słupki reprezentujące występujące w tej zmiennej wartości), na liście ***Identyfikator wierszy*** wybieramy zmienną *Palenie*.



Następnie w polu ***Tytuł wykresu*** wprowadzamy tekst *Wykres mozaikowy*, a w obszarze ***Kolory płytek***, możemy określić kolory płytek na wykresach. Opcje kolorów są analogiczne do tych opisanych w przykładzie dla wykresu słupkowego.



Po określeniu odpowiednich kolorów klikamy ***OK*** otrzymując wykres mozaikowy ze wskazanymi kolorami płytek.

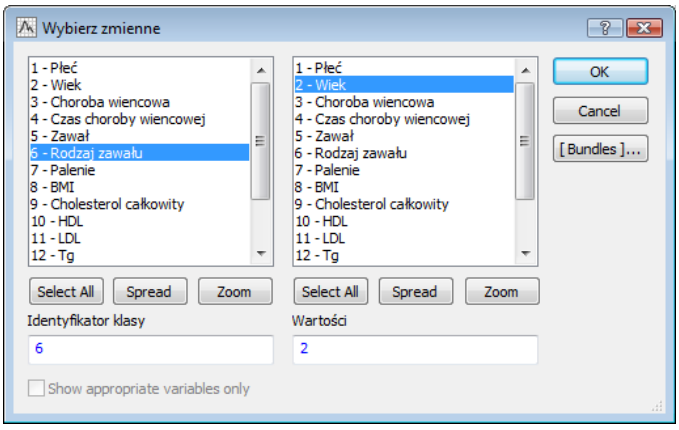


Na powyższym wykresie szerokość kolumn informuje o rozkładzie brzegowym zmiennej *Rodzaj zawału*. Wysokość poszczególnych słupków o rozkładzie warunkowym zmiennej *Palenie* w poszczególnych klasach zmiennej *Rodzaj zawału.*

9.5. Wykres kołowy (spie plot)

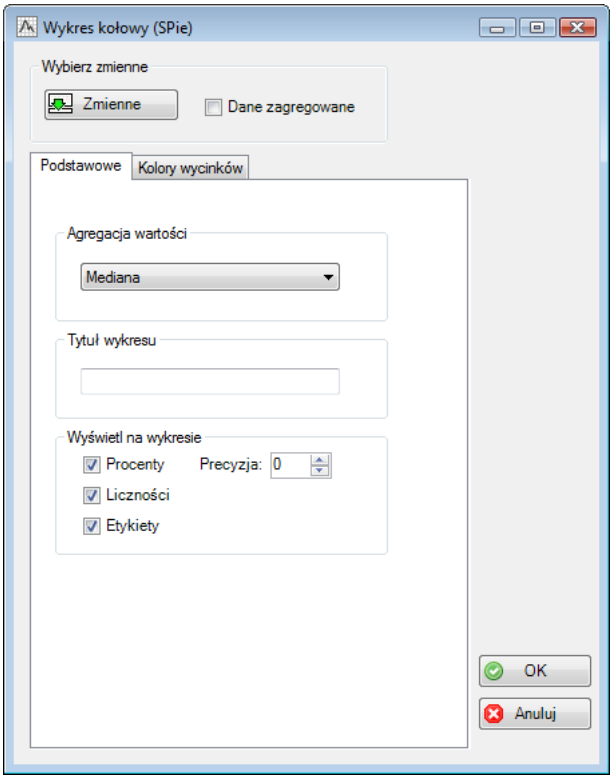
Wykres kołowy (spie plot) jest modyfikacją klasycznego wykresu kołowego. Na wykresie, oprócz szerokości wycinka reprezentującego procentowy udział danej kategorii zmiennej interpretujemy także wysokość promienia danego wycinka, który odpowiada wybranej statystyce pozycyjnej (np. średniej lub medianie) dodatkowej zmiennej.

Przykład. Przedstawimy teraz przykład przygotowania dostosowanego wykresu kołowego *Spie plot* na przykładzie pliku *Zawały.sta*. Z menu ***Analizy marketingowe i rynkowe*** | ***Wykresy*** wybieramy polecenie ***Wykres kołowy*** przywołując okno o tej samej nazwie. W pierwszej kolejności określamy zmienne, jakie będziemy analizować. W tym celu klikamy przycisk ***Zmienne***, a następnie na liście ***Identyfikator klasy*** wybieramy zmienną *Rodzaj zawału* (otrzymamy zatem trzy wycinki koła reprezentujące występujące w tej zmiennej wartości), na liście ***Wartości*** wybieramy zmienną *Wiek*.

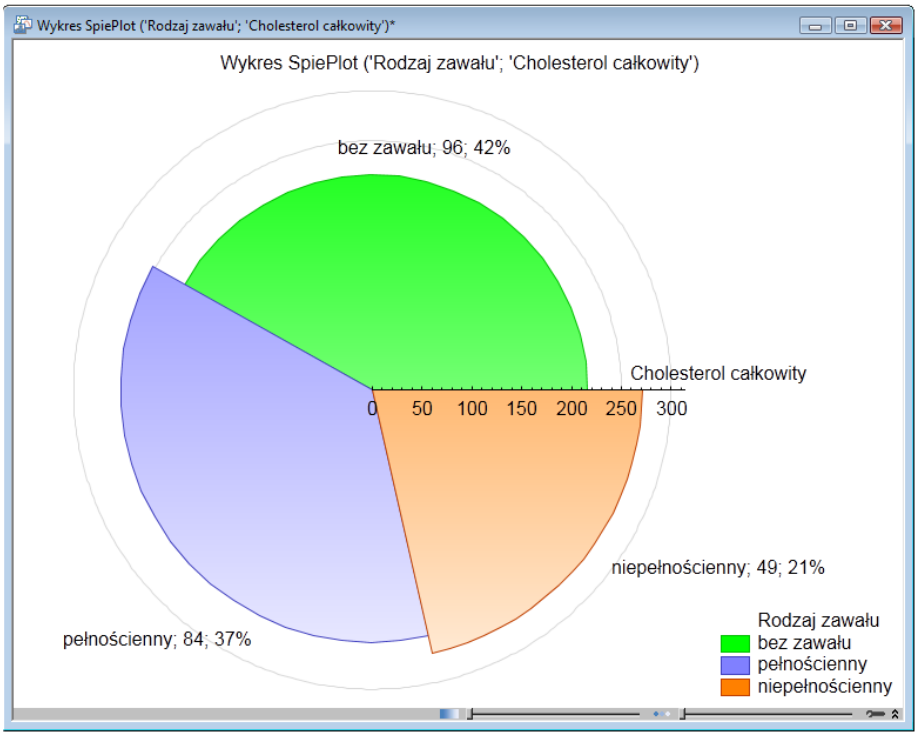


Następna na liście rozwijalnej ***Agregacja wartości*** wybieramy opcję ***Mediana***, aby wartości poszczególnych promieni wycinków koła odpowiadały wartościom median w grupach zmiennej

Rodzaj zawału. Opcje kolorów są analogiczne do tych opisanych w przykładzie dla wykresu słupkowego.



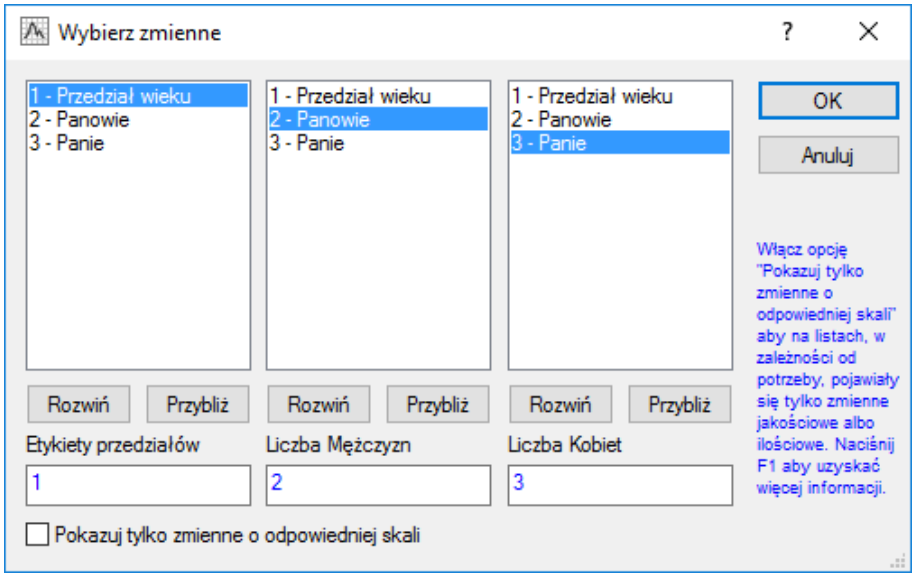
Po określeniu odpowiednich kolorów klikamy ***OK*** otrzymując wykres kołowy. Każdy z wycinków posiada etykietę z zawartością zgodną z wyborem opcji ***Wyświetl na wykresie***. Poszczególne wartości są odseparowane średnikami.



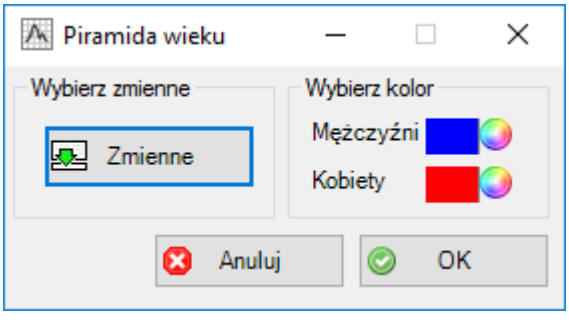
9.6 Piramida populacyjna

Piramida populacyjna pozwala przedstawić rozkład zmiennej w obrębie dwóch kategorii. Piramida składa się z dwóch poziomych wykresów słupkowych posiadających wspólna, pionową oś. Piramida populacyjna najczęściej wykorzystywana jest do w celu prezentacji danych demograficznych. Na wspólnej, pionowej osi zaznaczone są przedziały wiekowe (najmłodsze grupy na dole), słupki prezentują liczebność osób w danym przedziale w podziale na płeć.

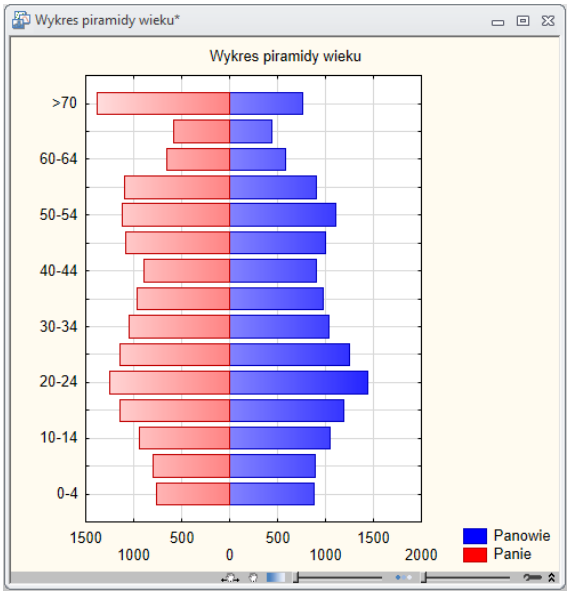
Przykład. Przedstawimy teraz przykład przygotowania ***Piramidy populacyjnej*** na podstawie pliku *PiramidaWieku.sta*. Z menu ***STATISTICA Zestaw Plus*** | ***Wykresy*** wybieramy polecenie ***Piramida populacyjna***. W pierwszej kolejności określamy zmienne, jakie będziemy analizować. W tym celu klikamy przycisk ***Zmienne***, a następnie na liście ***Etykiety przedziałów*** wybieramy zmienną *Przedział wieku* (zmienna określa dla jakich przedziałów wiekowych będzie budowana piramida), na liście ***Liczba Mężczyzn*** wybieramy zmienną *Panowie* (zmienna określa ilość mężczyzn, bądź dowolnej innej kategorii, którą chcemy zobrazować na wykresie) oraz analogicznie wybieramy zmienną *Panie* na liście ***Liczba Kobiet.***



Moduł ***Piramida populacyjna*** dodatkowo pozwala określić kolor słupków dla poszczególnych kategorii.



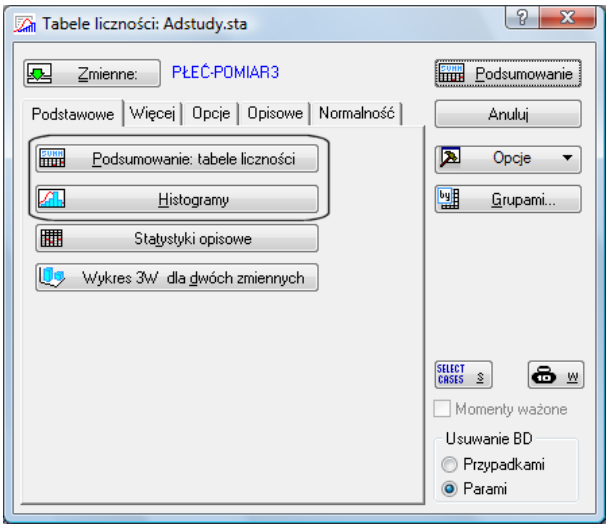
Po określeniu odpowiednich kolorów klikamy ***OK*** i otrzymujemy wykres piramidy populacyjnej.



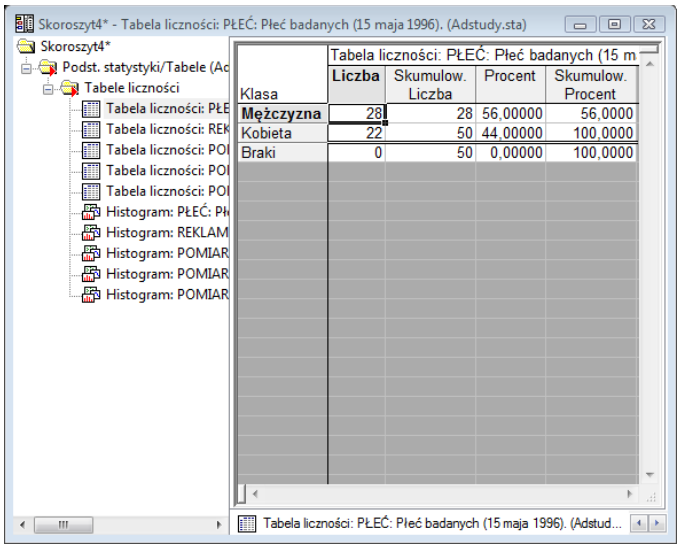
10. Narzędzia

Grupa ***Narzędzia*** umożliwia zapis wykresów zawartych w skoroszycie *STATISTICA* w postaci plików graficznych o wskazanym formacie i rozdzielczości – moduł ***Zapisz pliki graficzne***. Dodatkowo moduł ***Zapisz do MS Excel*** umożliwia zapisanie kolejnych obiektów skoroszytu *STATISTICA* w postaci osobnych zakładek arkusza *MS Excel*. Grupę uzupełnia moduł ***Formatuj arkusz lub skoroszyt***, który pozwala nadać wynikowym arkuszom *STATISTICA* pożądany przez użytkownika format – użytkownik określa sposób wyświetlania wartości komórek, typ i wielkość czcionki i inne parametry arkusza istotne podczas publikacji wyników.

Przykład. Aby zaprezentować możliwości modułów zawartych w grupie ***Narzędzia*** wygenerujemy na wstępie skoroszyt z przykładowymi wynikami. Analizę rozpoczniemy od otwarcia znanego nam pliku *Adstudy.sta*. Po jego otwarciu z menu ***Statystyka*** | ***Statystyki podstawowe i tabele*** wybieramy opcję ***Tabele liczności***.

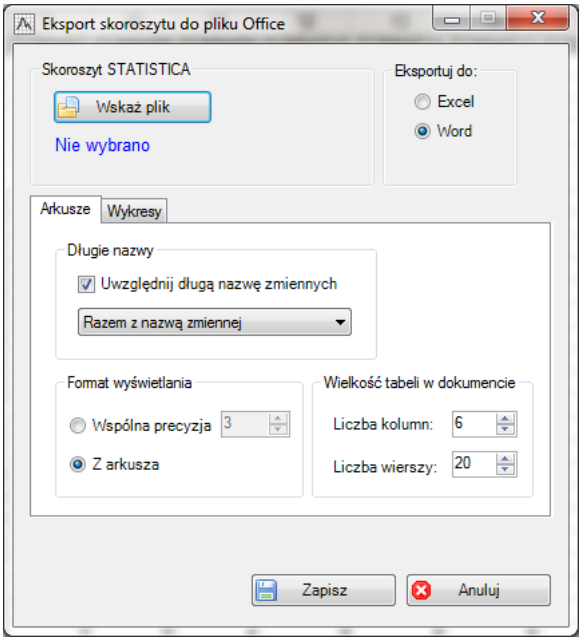


Klikamy przycisk ***Zmienne*** i wybieramy zmienne od 1 do 5 następnie klikamy ***Podsumowanie: tabele liczności*** generując zestaw tabel wynikowych oraz ***Histogramy*** generując zestaw histogramów. Wynikiem analizy powinien być skoroszyt zawierający wszystkie uzyskane raporty.



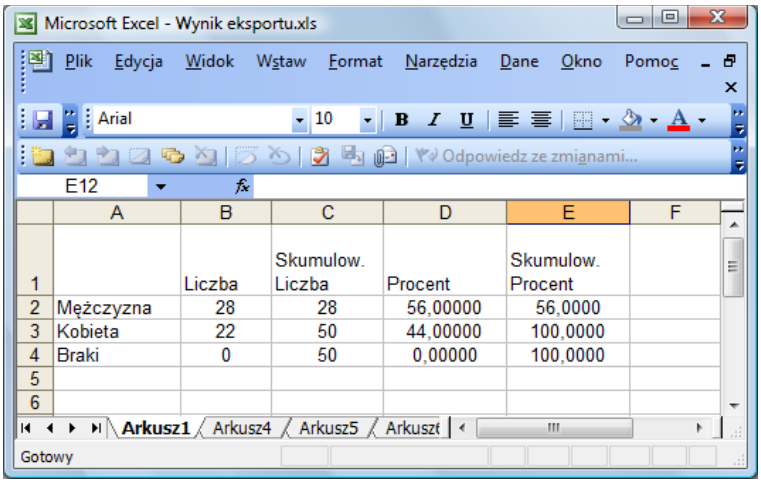
10.1. Zapisz do pliku Office

Umożliwia zapisanie utworzonego skoroszytu Statistica do arkusza MS Excel lub dokumentu MS Word. W przypadku wzbrania formatu MS Excel, każdy obiekt skoroszytu zapisywany jest na osobnym arkuszu. Aby zapisać utworzony skoroszyt do formatu MS Office z menu ***Zestaw do analiz marketingowych i rynkowych*** | ***Narzędzia*** wybieramy opcję ***Zapisz do pliku Office*** wyświetlając okno ***Eksport skoroszytu do pliku Excel***.



Jeżeli w programie Statistica otwarty był skoroszyt, zostanie on automatycznie wprowadzony do modułu. W sytuacji, gdy brak jest otwartego skoroszytu, za pomocą przycisku ***Wskaż plik*** możemy wskazać skoroszyt zapisany na dysku.

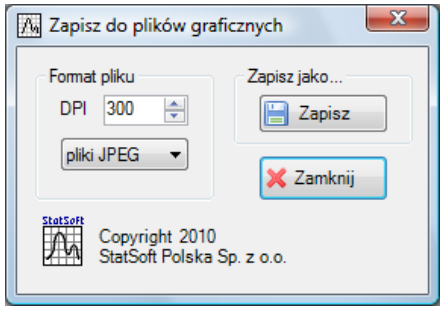
Ponieważ w utworzonym przez nas skoroszycie część nazw zmiennych zapisana jest w długiej nazwie zmiennej zaznaczamy opcję ***Uwzględnij długą nazwę zmiennych*** na liście rozwijalnej wybierając pozycję ***Razem z nazwą zmiennej***. Klikamy przycisk ***Zapisz***, aby zapisać dokument do formatu MS Excel. Otrzymany plik MS Excel zawierał będzie 8 arkuszy zawierających dokumenty ze skoroszytu Statistica.



Uwaga! Opcja zapisu do Excela lub Worda działa niezależnie od faktu zainstalowania bądź nie programu MS Office.

10.2. Zapisz do plików graficznych

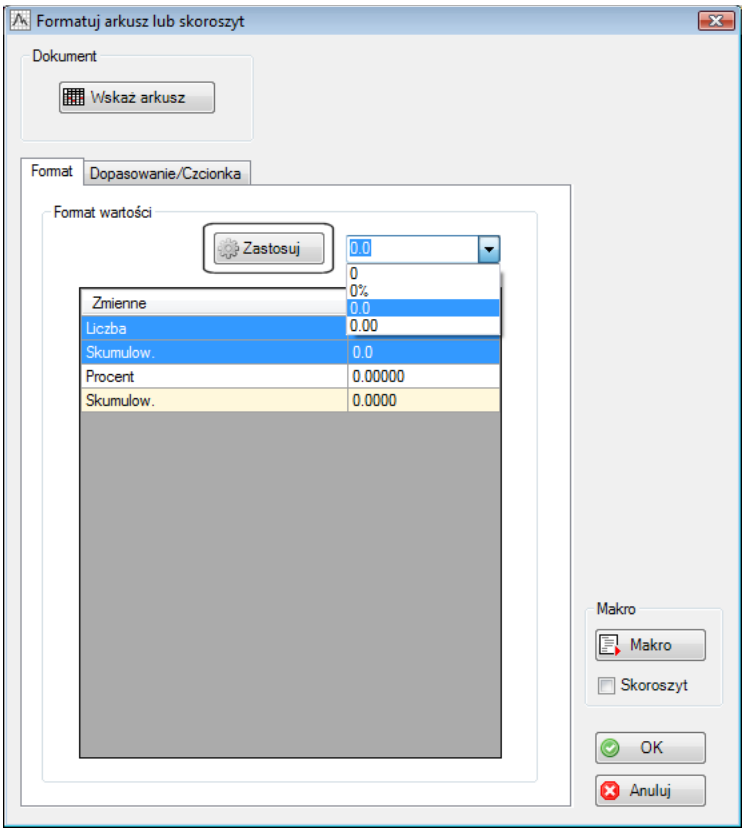
Aby zapisać wykresy znajdujące się w skoroszycie do plików graficznych o wskazanym formacie, z menu ***Analizy marketingowe i rynkowe*** | ***Narzędzia*** wybieramy opcję ***Zapisz do plików graficznych*** przywołując okno o tej samej nazwie.



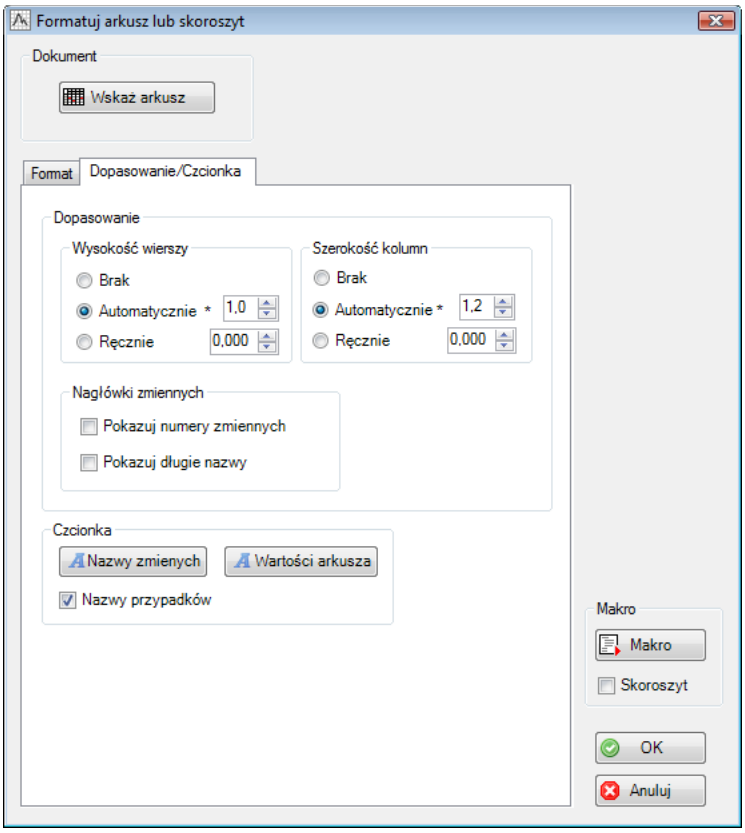
Po otwarciu okna określamy docelowy format zapisywanych plików w obszarze ***Format pliku***. Następnie zapisujemy obiekty graficzne do plików za pomocą przycisku ***Zapisz***.

10.3. Formatuj arkusz lub skoroszyt

Aby rozpocząć formatowanie arkusza, z menu ***Analizy marketingowe i rynkowe*** | ***Narzędzia*** wybieramy opcję ***Formatuj arkusz lub skoroszyt*** wyświetlając okno o tej samej nazwie.

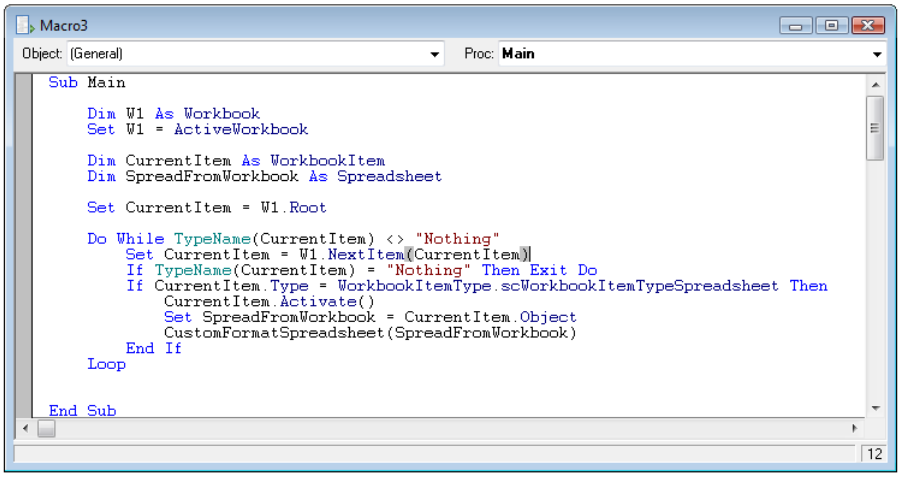


Następnie klikamy przycisk ***Wskaż arkusz*** i w oknie ***Wybierz arkusz*** wskazujemy interesujący nas arkusz (w naszym przypadku będzie to jeden z arkuszy zapisanych w skoroszycie). Po wskazaniu arkusza, w tabeli na karcie ***Format*** zostaną wyświetlone nazwy zmiennych występujących w arkuszu wraz z aktualnym formatem wyświetlania ich wartości. Możemy zauważyć, że format dwóch ostatnich zmiennych przewiduje zbyt dużą liczbę miejsc po przecinku. Aby zmienić ten format zaznaczamy interesujące nas wiersze, a następnie na liście rozwijalnej wybieramy format 0.0 (jedno miejsce po przecinku), a następnie klikamy ***Zastosuj***. Oczywiście poza zdefiniowanymi formatami możemy też określić własny format wyświetlania wpisując go jako element listy.



Następnie przechodzimy na kartę ***Dopasowanie/Czcionka*** i w obszarze ***Nagłówki zmiennych*** zaznaczamy opcję ***Pokazuj długie nazwy***. W obszarze ***Czcionka*** możemy zmienić typ i krój czcionki zarówno wartości komórek arkusza jak i nazw zmiennych i przypadków. Dodatkowo możemy także sterować szerokością kolumn i wysokością wierszy docelowego arkusza. Aby sformatować wskazany arkusz według wybranych opcji klikamy przycisk ***OK***.

Aby sformatować wszystkie elementy skoroszytu według ustalonego schematu, w obszarze ***Makro*** zaznaczamy opcję ***Skoroszyt***, a następnie klikamy przycisk ***Makro*** tworząc dokument makra wykonujący określone formatowanie.



Makro to uruchamiamy przyciskiem ***F5***. Jego uruchomienie spowoduje przekształcenie wszystkich arkuszy zgodnie z określonym schematem.