Statystyka opisowa

Wykład 2:
Grupowanie danych (szeregi statystyczne)

+ porady dotyczące analizy danych w programie *STATISTICA*

Dobór metody prezentacji danych

Wybór sposobu prezentacji danych zależy od:

- **charakteru danych statystycznych (inne metody wybierzemy dla *danych przekrojowych*, inne dla *czasowych*, inne dla *danych ankietowych*);**
- **liczby analizowanych przypadków – innych metod prezentacji danych wymaga zbiór *danych przekrojowych* o powiatach w Polsce (kilkaset przypadków) a innych o państwach UE (27 przypadków);**
- **charakteru cechy statystycznej – inne narzędzia można zastosować do pogrupowania danych o charakterze liczbowych, a inne do danych o charakterze nominalnym („tekstowym”).**

Rodzaje szeregów statystycznych

(dane ankietowe, cecha nominalna)

SZEREG SZCZEGÓŁOWY (NIEUPORZĄDKOWANY)	*SZEREG SZCZEGÓŁOWY (UPORZĄDKOWANY)*	*SZEREG ROZDZIELCZY (PUNKTOWY)*

kobieta
kobieta
kobieta
mężczyzna
kobieta
mężczyzna
kobieta
kobieta
mężczyzna
kobieta
kobieta
mężczyzna
kobieta
mężczyzna
kobieta
...

→

kobieta
kobieta
kobieta
kobieta
kobieta
kobieta
kobieta
kobieta
kobieta
kobieta
mężczyzna
mężczyzna
mężczyzna
mężczyzna
mężczyzna
...

→

Płeć (x_i)	n_i	$\%_i$
kobieta	10	67%
mężczyzna	5	33%

$n = 15$ (wielkość próby)

x_i – wartości cechy, n_i – liczność, $\%_i$ – udział procentowy

Rodzaje szeregów statystycznych

(dane ankietowe, cecha mierzalna)

SZEREGI SZCZEGÓŁOWE | **SZEREG ROZDZIELCZY (PUNKTOWY)** | **SZEREG ROZDZIELCZY (PRZEDZIAŁOWY)**

WIEK
18
23
25
31
24
18
34
44
51
27
38
19
45
61
…

→

WIEK
18
18
18
18
18
…
19
19
19
19
…
20
20
20
…

→

Wiek (x_i)	n_i	$\%_i$
18	15	5,0%
19	10	3,3%
…		
….		
64	5	1,7%
65	3	1,0%

→

Wiek (x_i)	n_i	$\%_i$
18-24	100	33,3%
25-34	60	20,0%
35-44	50	16,7%
45-54	50	16,7%
55-65	40	13,3%

$n = 300$ (wielkość próby)

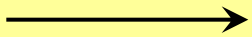
x_i – wartości cechy, n_i – liczność, $\%_i$ – udział procentowy

Rodzaje szeregów statystycznych

(dane przekrojowe, cecha mierzalna)

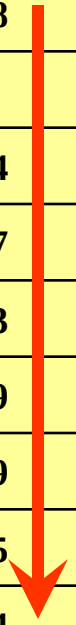
SZEREG SZCZEGÓŁOWY (NIEUPORZĄDKOWANY)

Państwo	Stopa bezrobocia 2007
Austria	4,4
Belgia	7,5
Bułgaria	6,9
Cypr	4
Czechy	5,3
Dania	3,8
Estonia	4,7
Finlandia	6,9
Francja	8,4



SZEREG SZCZEGÓŁOWY (UPORZĄDKOWANY)

Państwo	Stopa bezrobocia 2007
Dania	3,8
Cypr	4
Austria	4,4
Estonia	4,7
Czechy	5,3
Bułgaria	6,9
Finlandia	6,9
Belgia	7,5
Francja	8,4



$n = 27$ (wielkość próby)

x_i – wartości cechy

Rodzaje szeregów statystycznych

(dane czasowe, cecha mierzalna)

SZEREG SZCZEGÓŁOWY (NIEUPORZĄDKOWANY)

Rok	Wskaźnik rodności w Polsce
1998	10,23
1999	9,88
2000	9,79
2001	9,53
2002	9,16
2003	9,36
2004	9,55
2005	9,72
2006	9,85

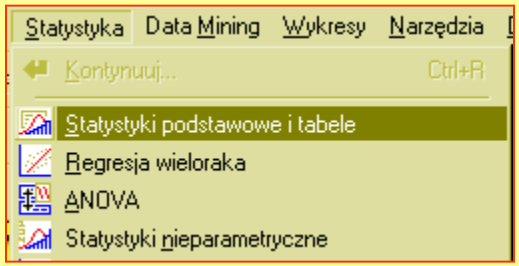
Prezentacja danych czasowych nie wymaga zwykle żadnych przekształceń wyjściowego ciągu danych (o ile oczywiście jest on uporządkowany chronologicznie).
W żadnym razie nie sortujemy danych czasowych według wartości poszczególnych cech!

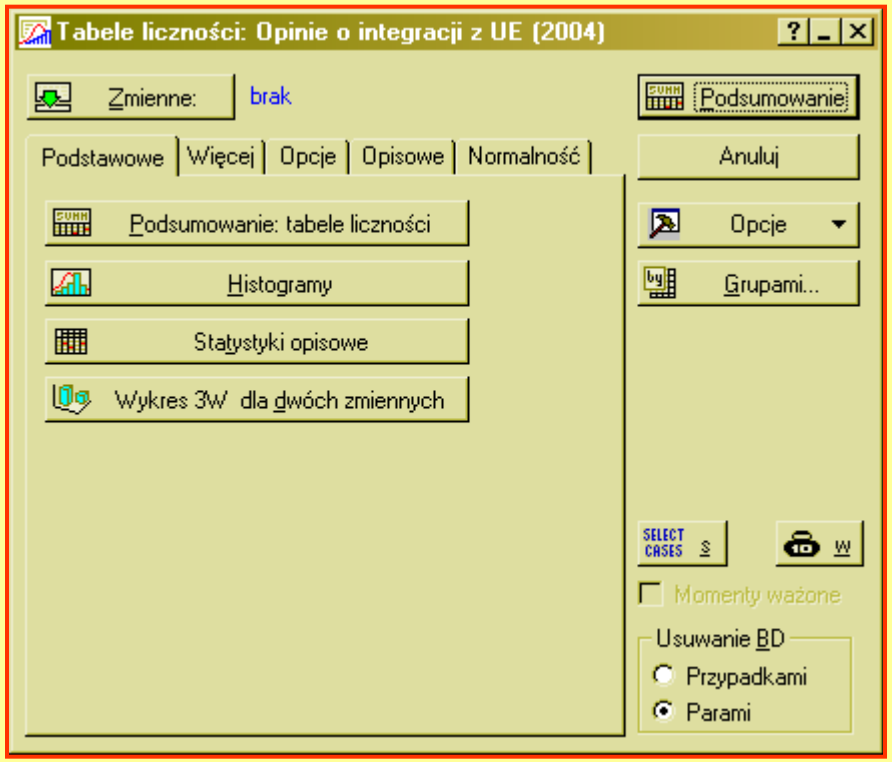
$n = 9$ **(wielkość próby)**

x_i **– wartości cechy**

Grupowanie danych – tabele liczności

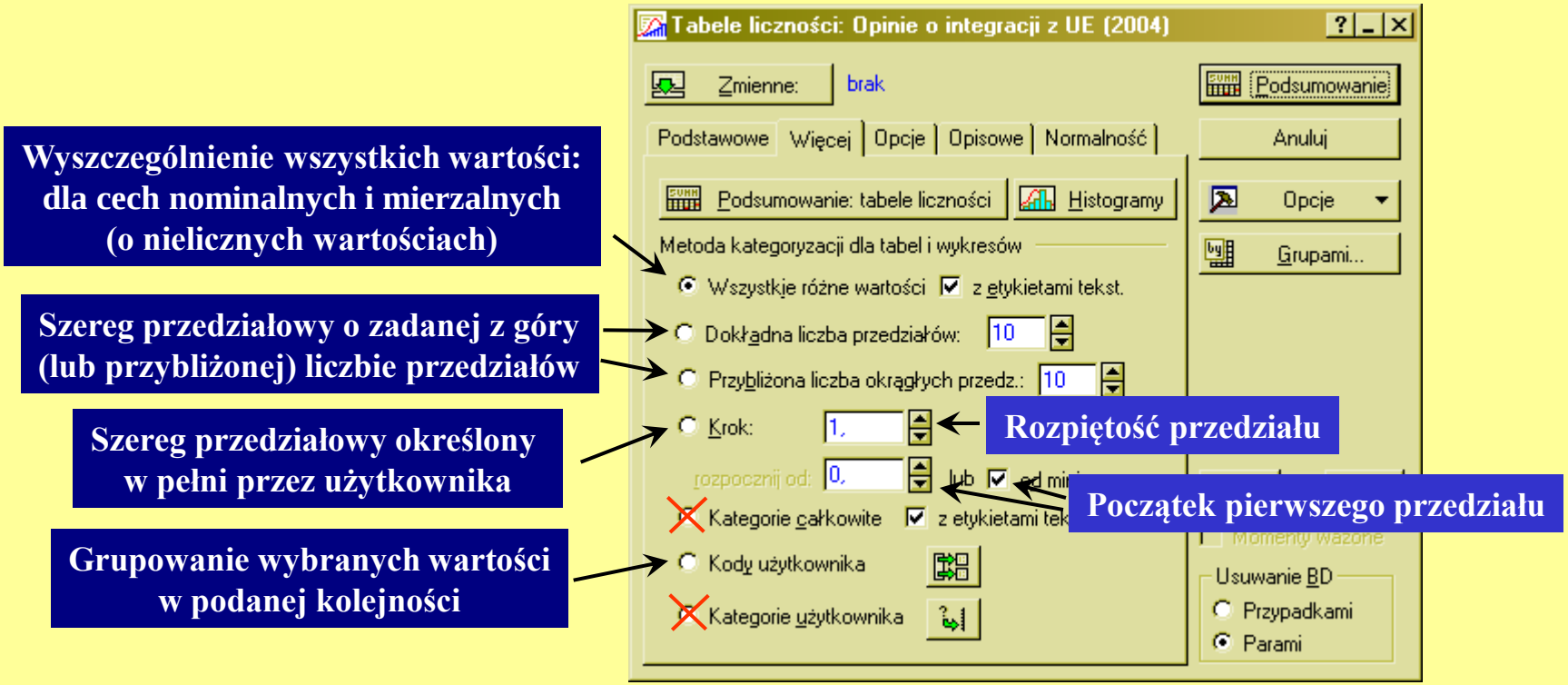
Narzędzia analiz statystycznych w programie *STATISTICA* są dostępne za pomocą dwóch poleceń *STATYSTYKA* i *WYKRESY*. Aby dokonać grupowania danych należy posłużyć się analizą *TABELE LICZNOŚCI*.





Rodzaje grupowania

W oknie *TABELE LICZNOŚCI* istnieje możliwość ustawienia różnych sposobów grupowania – w zakładce *WIĘCEJ*. Poniżej opisano najważniejsze sposoby grupowania (kategoryzacji) wartości zmiennych.



Przykłady grupowania

Przykład dotyczy pliku *Opinie o integracji z UE (2004).* Celem analizy jest przedstawienie odpowiedzi na pytania dotyczące skutku integracji dla Polski (zmienna 7) i sposoby głosowania respondentów w referendum akcesyjnym (zmienna 6).

Ponieważ obie zmienne mają ten sam charakter (nominalny) grupowanie możemy przeprowadzić jednocześnie, wybierając za pomocą przycisku ZMIENNE obie cechy i ustalając odpowiednio sposób grupowania.

Po naciśnięciu przycisku *PODSUMOWANIE* otrzymujemy dwie tabele – oddzielne wyniki grupowania dla obu cech. Wszystkie wyniki kolejnych analiz będę dodawane do otworzonego właśnie skoroszytu wyników.

Opis wyników grupowania

Tabele z wynikami grupowania zawierają następujące informacje:

- **warianty badanej cechy;**
- **liczbę przypadków dla każdego wariantu;**
- **skumulowaną liczbę przypadków (opis na rysunku)**
- **procentowy udział danego wariantu cechy;**
- **skumulowane procenty.**

UWAGA!!! Wartości skumulowane mają sens tylko wtedy, gdy grupowane warianty są w logiczny sposób uporządkowane (a więc dla cech porządkowych lub liczbowych).

W „roboczej” tabeli wyników należy pozostawić tylko te wartości, które się da zinterpretować. Należy także dokonać formatowania wartości.

Klasa	Tabela liczności: Czy integracja będzie dla Pol			
	Liczba	Skumulow. Liczba	Procent	Skumulow. Procent
trudno powiedzieć	49	49	16,33333	16,3333
niekorzystna	29	78	9,66667	26,0000
raczej niekorzystna	42	120	14,00000	40,0000
nie będzie mieć wpływu	11	131	3,66667	43,6667
raczej korzystna	114	245	38,00000	81,6667
korzystna	55	300	18,33333	100,0000
Braki	0	300	0,00000	100,0000

Klasa	Tabela liczności: C	
	Liczba	Procent
trudno powiedzieć	49	16,3
niekorzystna	29	9,7
raczej niekorzystna	42	14,0
nie będzie mieć wpływu	11	3,7
raczej korzystna	114	38,0
korzystna	55	18,3
Braki	0	0,0

Występowanie braków danych

Przy domyślnych ustawieniach opcji grupowania, w tabelach wyszczególniona jest także informacja o brakach danych. W rozpatrywanym przykładzie są one zapewne równoważne stwierdzeniu faktu, iż ktoś nie wziął udziału w referendum akcesyjnym.

Brak udziału w referendum →

Klasa	Tabela liczności: S	
	Liczba	Procent
tak	145	48,3
nie	64	21,3
Braki	91	30,3

W pewnych sytuacjach chcemy poznać strukturę danych po wykluczeniu z rozważań braków odpowiedzi. W omawianym przykładzie ma to sens, gdyż w ten sposób dowiadujemy się informacji o wynikach referendum w badanej zbiorowości. W oknie *TABELE LICZNOŚCI* w zakładce *OPCJE* wyłączamy ☐ Policz i podaj brakujące dane (BD)

Otrzymujemy informacje o strukturze procentowej tylko w grupie osób, które wzięły udział w głosowaniu.

Klasa	Tabela liczności: S	
	Liczba	Procent
tak	145	69,4
nie	64	30,6

Grupowanie danych liczbowych

Kontynuując analizę danych ankietowych z pliku *Opinie o integracji z UE (2004)* zbadamy strukturę wieku respondentów. Tego typu zestawienia umieszcza się w części *Charakterystyka badanej zbiorowości* – kwestia ta jest o tyle ważna, że poglądy na pewne zjawiska społeczne i polityczne są zwykle odmienne dla różnych grup wiekowych.

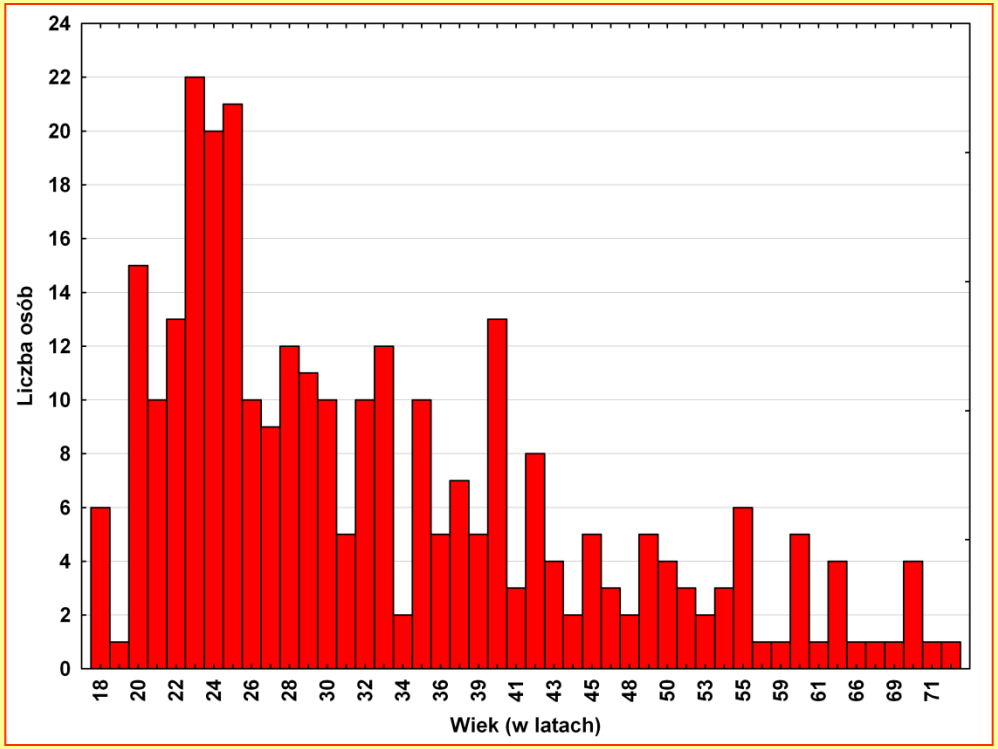
Należy więc stwierdzić, czy badana grupa jest reprezentatywną próbką z populacji dorosłych mieszkańców woj. podkarpackiego, czy też może odzwierciedla przede wszystkim poglądy osób młodszych.

Prezentacja rozkładu wieku respondentów

Klasa	Tabela liczności: Wiek	
	Liczba	Skumulow. Liczba
18	6	6
19	1	7
20	15	22
21	10	32
22	13	45
23	22	67
24	20	87
25	21	108
26	10	118
27	9	127
28	12	139
29	11	150
30	10	160
31	5	165
32	10	175
33	12	187
34	2	189
35	10	199
36	5	204
38	7	211
39	5	216
40	13	229
41	3	232
42	8	240
43	4	244
44	2	246
45	5	251
46	3	254
48	2	256

Po wybraniu analizy *TABELE LICZNOŚCI* i sporządzeniu (bez zmiany ustawień) szeregu rozdzielczego, okazuje się, że wyniki nie są zbyt czytelne…

Widać to zarówno podczas próby analizy informacji zawartych w tabeli liczności jak i na przykładzie graficznej prezentacji w postaci *HISTOGRAMU*.



Prezentacja rozkładu wieku respondentów

Przy tak dużej złożoności danych, należy je przedstawić w postaci szeregu przedziałowego, na przykład w następującej postaci…

Wiek (x_i)	n_i	$\%_i$
18-24	87	29,0%
25-34	102	34,0%
35-44	47	19,0%
45-54	27	9,0%
55-64	14	4,7%
65-75	13	4,3%

W tym celu wykorzystana zostanie opcja *KROK* umożliwiająca sporządzenie szeregu o jednakowej rozpiętości przedziałów.
Zostaną one następnie „ręcznie” skorygowane tak, by odpowiadały wzorcowi tabeli.

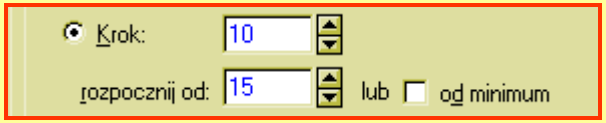


Tabela liczności: Wiek (Opinie o integracji z U

Od Do	Liczba	Skumulow. Liczba	Procent	Skumulow. Procent
15,00000<=x<25,00000	87	87	29,0	29,0
25,00000<=x<35,00000	102	189	34,0	63,0
35,00000<=x<45,00000	57	246	19,0	82,0
45,00000<=x<55,00000	27	273	9,0	91,0
55,00000<=x<65,00000	14	287	4,7	95,7
65,00000<=x<75,00000	12	299	4,0	99,7
75,00000<=x<85,00000	1	300	0,3	100,0
Braki	0	300	0,0	100,0

Uwagi i uzupełnienia dotyczące pracy z arkuszem danych programu *STATISTICA* (zmienne)

1) **Podczas nadawania nazw zmiennym w arkuszu programu *STATISTICA* można przejść do następnej (lub poprzedniej) zmiennej bez zamykania okna informacji o zmiennej.**
2) **Za pomocą przycisku *Wszystkie specyfikacje* można wywołać listę wszystkich zmiennych, gdzie również można edytować ich nazwy, także za pomocą poleceń *kopiuj* i *wklej*.**
3) **Aby dopasować rozmiar arkusza do długości zmiennych należy posłużyć się poleceniem *Format / Zmienne / Autodopasowanie* (zaznaczywszy uprzednio całość lub część arkusza).**
4) **W programie *STATISTICA* można także wprowadzić dodatkowe nazwy zmiennych (tzw. długie), które mogą być wyświetlane w arkuszu „na życzenie” użytkownika (*Widok / Nagłówki zmiennych / Pokaż długie nazwy*).**

Uwagi i uzupełnienia dotyczące pracy z arkuszem danych programu *STATISTICA* (etykiety tekstowe i wpisywanie danych)

1) **Wpisując etykiety tekstowe warto dopasować kolumny pomocniczego arkusza do ich długości.**
2) **Aby numerowanie etykiet zaczynało się od wartości 1 (a nie 101), należy zmienić domyślne ustawienia programu (*Narzędzia / Opcje – zakładka Arkusze*).**
3) **Każda nowa wartość tekstowa, która pojawi się w arkuszu podczas wpisywania danych zostanie zakodowana, jako kolejna etykieta tekstowa.**
4) **Podczas wpisywania danych do właściwie przygotowanego arkusza <u>korzystamy tylko</u> z klawiszy odpowiadających cyfrom oraz klawisza TAB lub strzałki w prawo.**
5) **Aby zastąpić wartość w komórce arkusza (to dotyczy także programu *EXCEL!!!*) nie trzeba jej edytować, ale wystarczy zaznaczyć (czyli nie trzeba klikać dwa razy i kasować!!!).**

Uwagi i uzupełnienia dotyczące pracy z arkuszem danych programu *STATISTICA* (skróty klawiaturowe)

CTRL + A – zaznaczenie całego arkusza
CTRL + C – kopiowanie zaznaczonego obszaru
CTRL + X – wycinanie zaznaczonego obszaru
CTRL + V – wklejanie zawartości schowka w pozycji kursora

CTRL + F – wyszukiwanie wartości w arkuszu (lub nazwach zmiennych i przypadków)
CTRL + H – wyszukiwanie i zamienianie wartości w arkuszu (lub nazwach zmiennych i przypadków)

CTRL + R – wznowienie ostatnio wykonywanej analizy statystycznej

Uwagi i uzupełnienia dotyczące pracy z arkuszem danych programu *STATISTICA* (narzędzia analiz)

1) **W programie *STATISTICA* można wyróżnić następujące podstawowe elementy: *ARKUSZ DANYCH*, *SKOROSZYT Z WYNIKAMI*, *OKNA ANALIZ*.**
2) **Nie ma potrzeby (ani sensu) zamykanie *SKOROSZYTU Z WYNIKAMI* po każdej analizie – wystarczy go zminimalizować.**
3) **Okna *ANALIZY* po wywołaniu wyników automatycznie się minimalizuje – analizę można w każdej chwili wznowić (CTRL + R).**
4) **Jednocześnie można korzystać z wielu *ANALIZ*.**
5) **Wyniki w postaci tabel i wykresów muszą być odpowiednio sformatowane – jeżeli nie dzieje się to automatycznie, należy takie czynności wykonać samodzielnie.**

Przekodowanie danych

Opcja przekodowania pozwala na zmianę wartości zmiennych nie w oparciu o formuły matematyczne, lecz na podstawie pewnych kryteriów logicznych.

Okno przekodowanie zmiennych można wywołać za pomocą przycisku 

i polecenia *PRZEKODUJ*.

Przykład stanowi kontynuację poprzednich rozważań (dotyczących formuły na BMI).

WHO dokonuje następującej klasyfikacji stanu zdrowia <u>osób dorosłych</u> według wartości BMI:

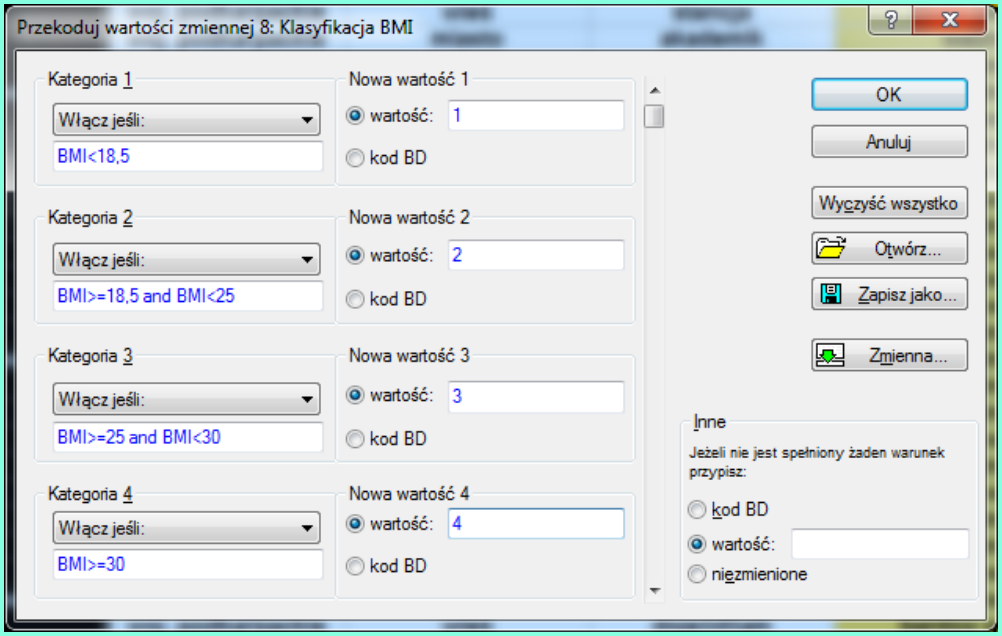
- **BMI < 18,5 – niedowaga;**
- **18,5 ≤ BMI < 25 – norma;**
- **25 ≤ BMI < 30 – nadwaga;**
- **BMI ≥ 30 – otyłość.**

Dokonamy takiej klasyfikacji osób z pliku *Ankieta studencka 2013-2016*.

E

W arkuszu danych wstawiamy nową zmienną i nazywamy ją Klasyfikacja wg BMI

Zaznaczamy nową zmienną i wybieramy za poleceniem przycisku ZMIENNE opcję PRZEKODUJ i wprowadzamy poniższe warunki. UWAGA!!! Zamiast BMI można też pisać v7, ale użycie BMI jest bardziej uniwersalne (dlaczego?)



7 BMI =v6/v5^2*10^4	8 Klasyfikacja BMI
21,9	norma
19,4	norma
24,2	norma
19,4	norma
18,2	niedowaga
20,8	norma
21,5	norma
23,5	norma
21,8	norma
22,0	norma
21,9	norma
19,2	norma
20,8	norma
27,3	nadwaga
18,4	niedowaga

Nadajemy wartościom 1-4 odpowiednie etykiety tekstowe: 1 – niedowaga, 2 – norma, 3 – nadwaga, 4 – otyłość