Wykład dla studiów doktoranckich IMDiK PAN

Biostatystyka I

dr Anna Rajfura
Kat. Doświadczalnictwa i Bioinformatyki SGGW
anna_rajfura@sggw.pl

Program wykładu w skrócie

1. Wprowadzenie: rozkład empiryczny, parametry.
2. Rozkład normalny jako model cechy rzeczywistej.
3. Estymacja.
4. Hipotezy statystyczne. Porównanie dwóch populacji. testy parametryczne i nieparametryczne.
5. Testowanie zgodności rozkładu empirycznego z teoretycznym.
6. Porównania wielu średnich.Testy parametryczne i nieparametryczne.

W przykładach zastosowanie pakietu statystycznego STATISTICA, arkusza EXCEL lub CALC.

Literatura

1. Łomnicki A., *Wprowadzenie do statystyki dla przyrodników*, PWN Warszawa 2000, 2007
2. Meissner W., *Przewodnik do ćwiczeń z przedmiotu Metody statystyczne w biologii*, Wyd. Uniwersytetu Gdańskiego, Gdańsk 2010
3. Stanisz A. (red.), *Biostatystyka*, Wyd. Uniwersytetu Jagiellońskiego, Kraków 2005
4. Stanisz A., *Przystępny kurs statystyki z zastosowaniem STATISTICA PL na przykładach z medycyny* (tomy 1-2), Wyd. Statsoft Polska, Kraków 2006, 2007
5. Watała C., *Biostatystyka*, Wyd. α-medica press 2002

Statystyczny opis danych

Dane liczbowe z pomiaru: $x_1, x_2, \ldots, x_n$
w badaniu pełnym.

Pierwszym etapem analizy danych jest ich opis:

- rozkład wartości (w postaci tabel, wykresów)
- charakterystyka za pomocą parametrów (np. średnia arytmetyczna, mediana, odchylenie standardowe)

Przykład

Wiek chorych na pewną chorobę – **szereg statystyczny nieuporządkowany**.

30	80	92	65	90	19	21	81	76	24
80	38	31	38	38	20	92	28	50	55
73	47	30	62	41	48	69	61	44	57
49	56	43	56	45	27	45	70	89	44
52	56	61	49	53	66	81	56	34	88
76	49	51	45	48	48	60	57	43	59
41	51	85	69	17	36	76	59	52	43
77	54	51	47	48					

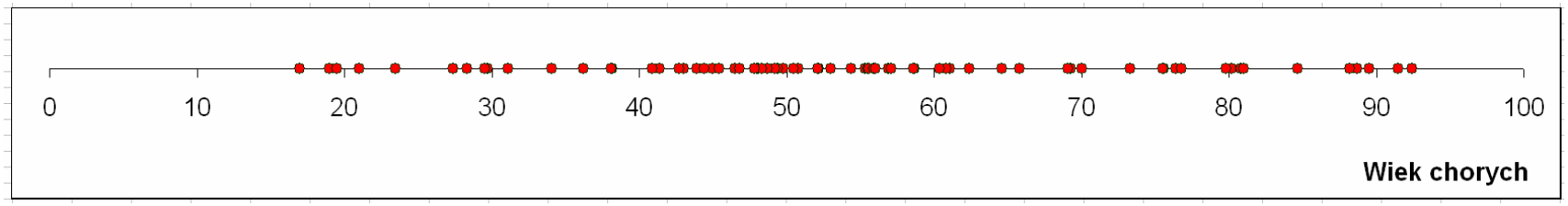
Przykład cd.

Wiek chorych na pewną chorobę – **szereg statystyczny uporządkowany** (posortowany rosnąco).

17	19	20	21	24	27	28	30	30	31
34	36	38	38	38	41	41	43	43	43
44	44	45	45	45	47	47	48	48	48
48	49	49	49	50	51	51	51	52	52
53	54	55	56	56	56	56	57	57	59
59	60	61	61	62	65	66	69	69	70
73	76	76	76	77	80	80	81	81	85
88	89	90	92	92					

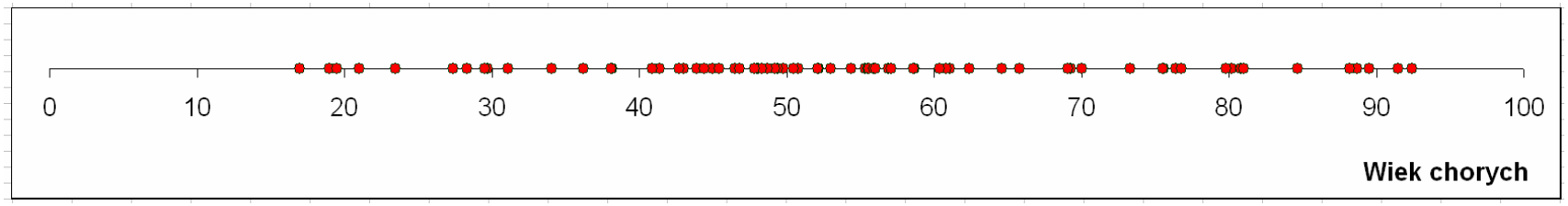
Przykład cd.

Wiek chorych na pewną chorobę – szereg statystyczny uporządkowany (posortowany rosnąco).

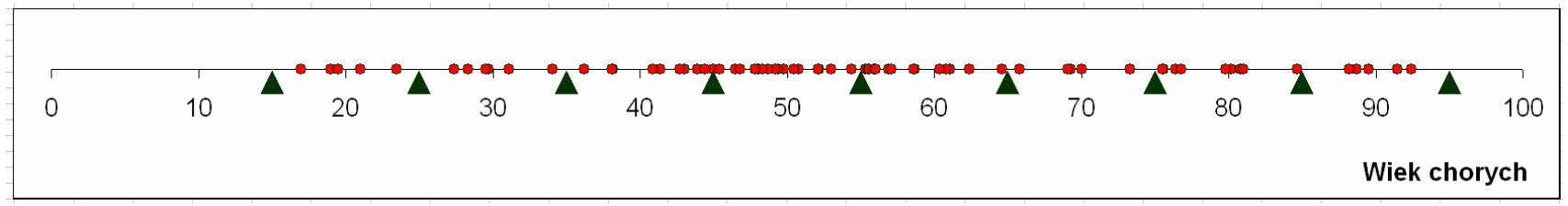


Przykład cd.

Wiek chorych na pewną chorobę – szereg statystyczny uporządkowany (posortowany rosnąco).



Cały zakres wartości dzielimy na klasy,



zliczamy wartości w każdej klasie. Rezultat można przedstawić w tabeli lub na wykresie.

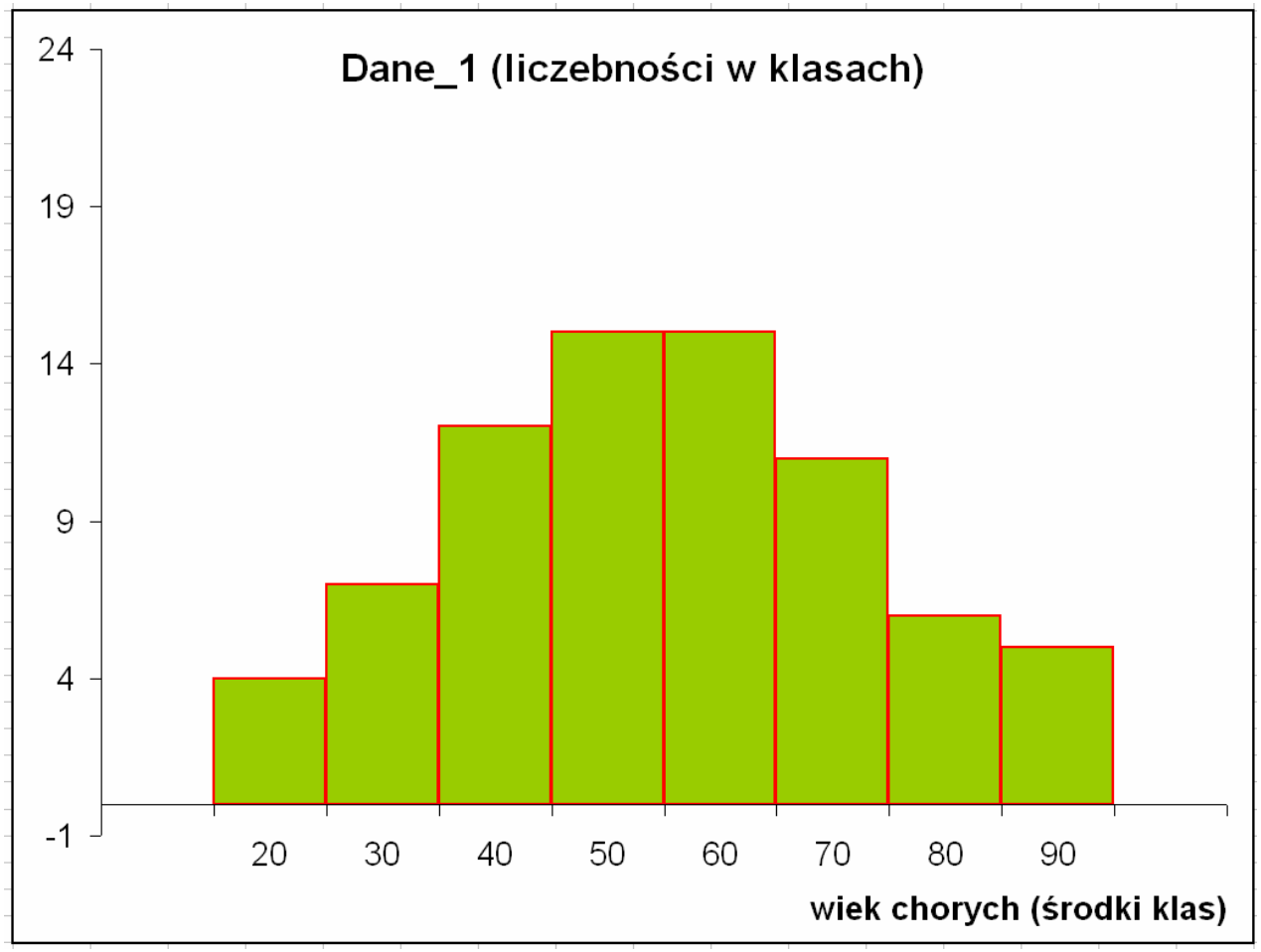
Empiryczny rozkład wartości

... przedstawiony w postaci **szeregu rozdzielczego**

Klasy wiekowe	Liczebność
<15; 25)	4
<25; 35)	7
<35; 45)	12
<45; 55)	15
<55; 65)	15
<65; 75)	11
<75; 85)	6
<85; 95>	5
Razem	75

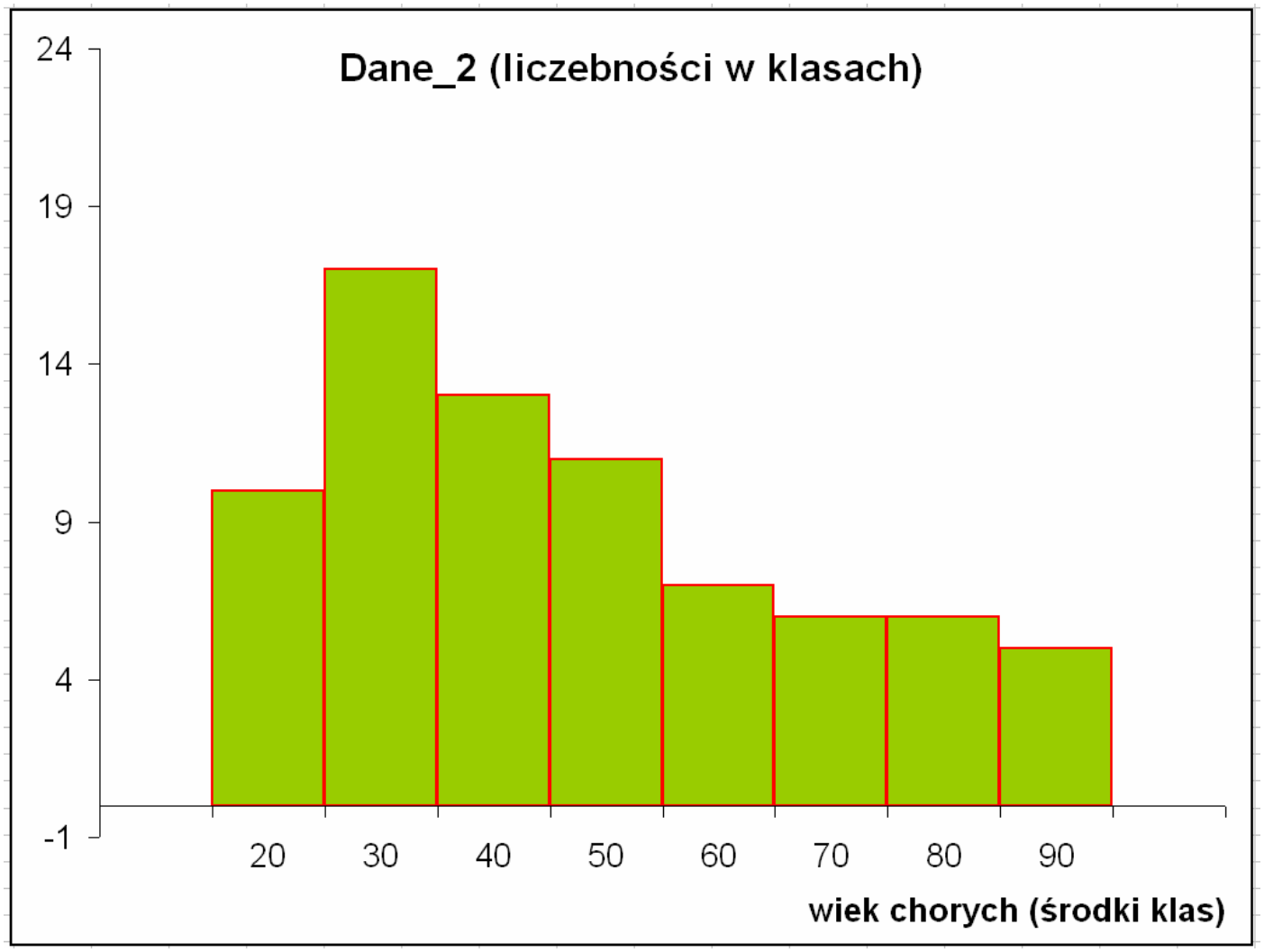
Empiryczny rozkład wartości

... przedstawiony w postaci **histogramu**



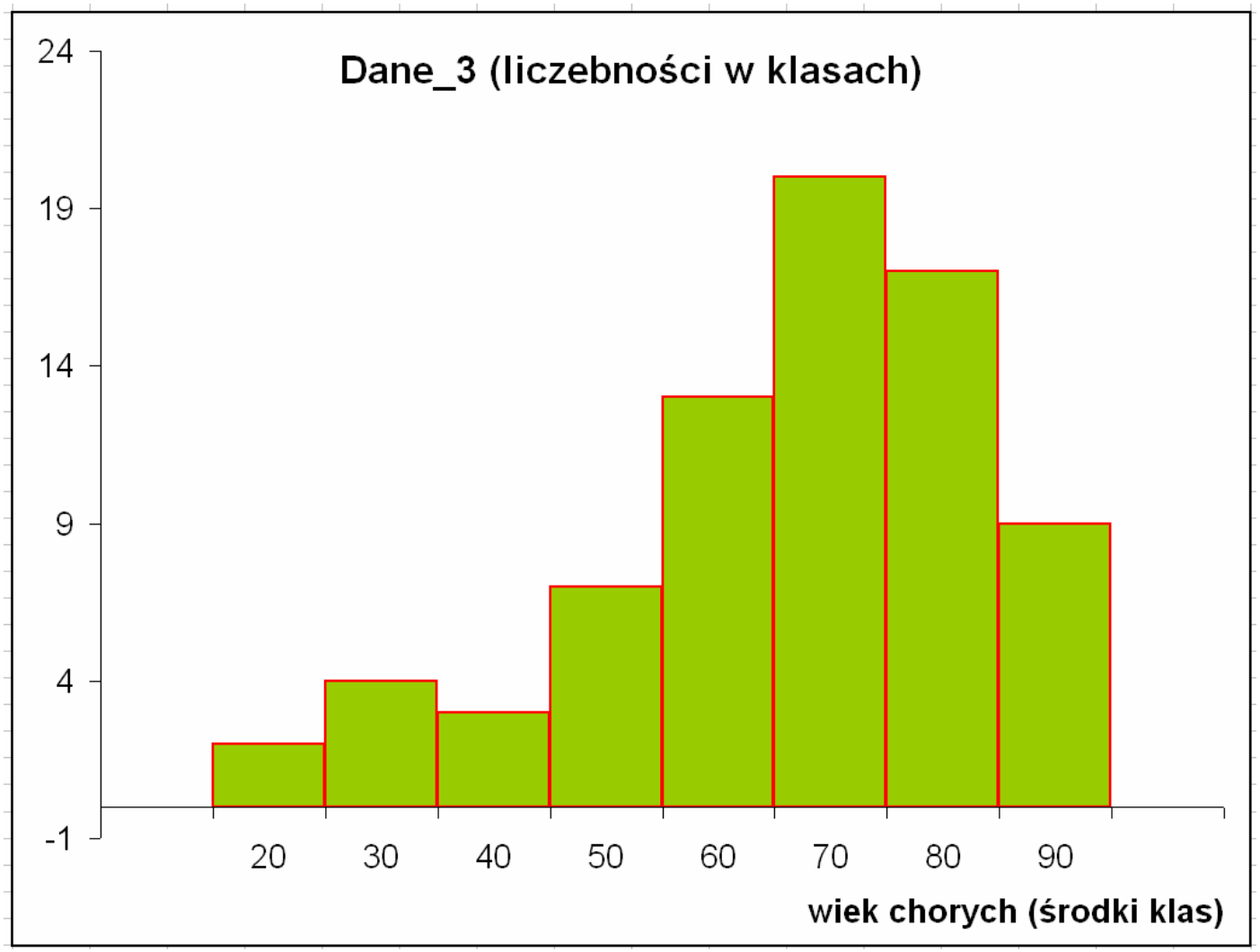
Rozkład symetryczny

Empiryczny rozkład wartości cd.



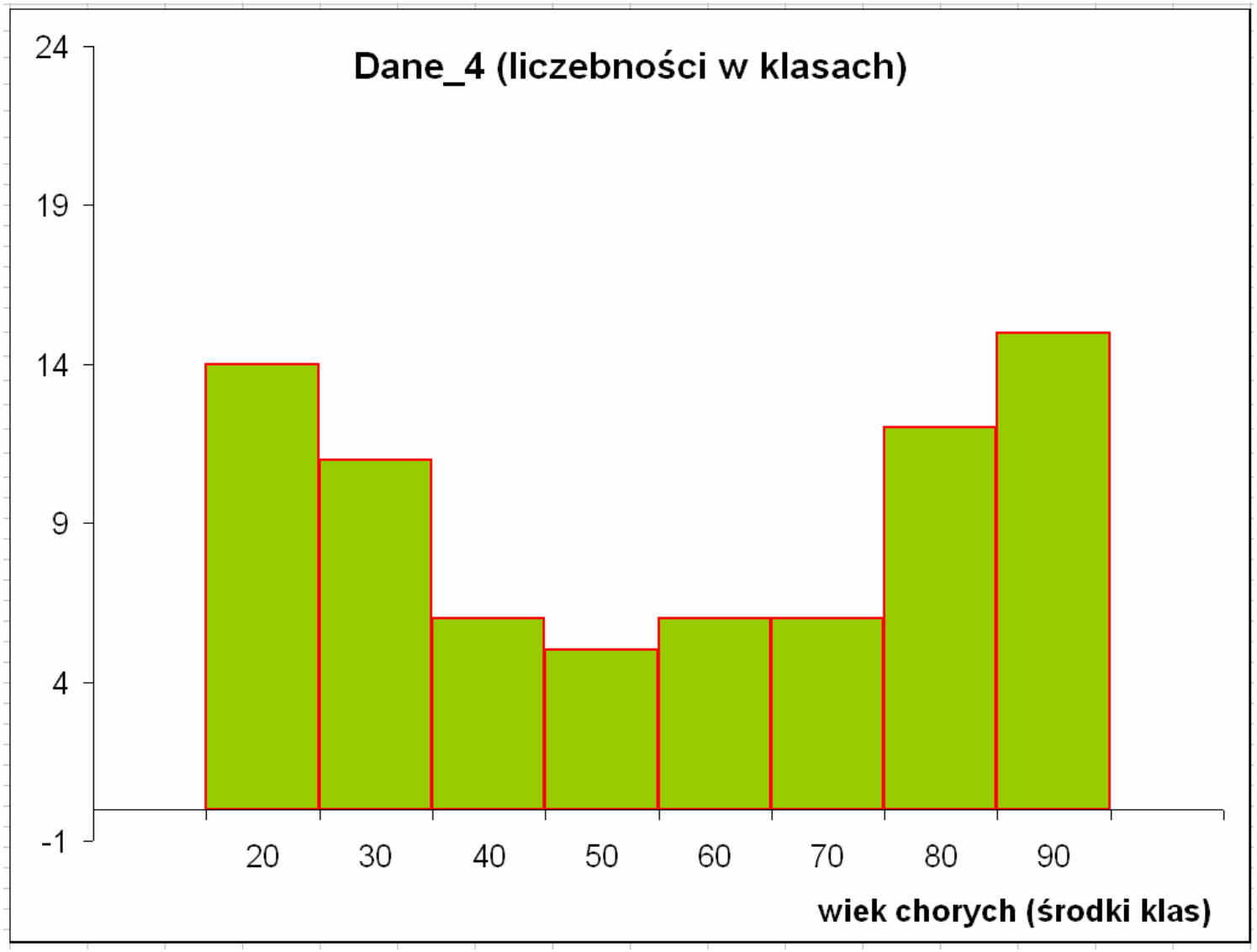
Rozkład asymetryczny
(prawoskośny, asymetria prawostronna)

Empiryczny rozkład wartości cd.



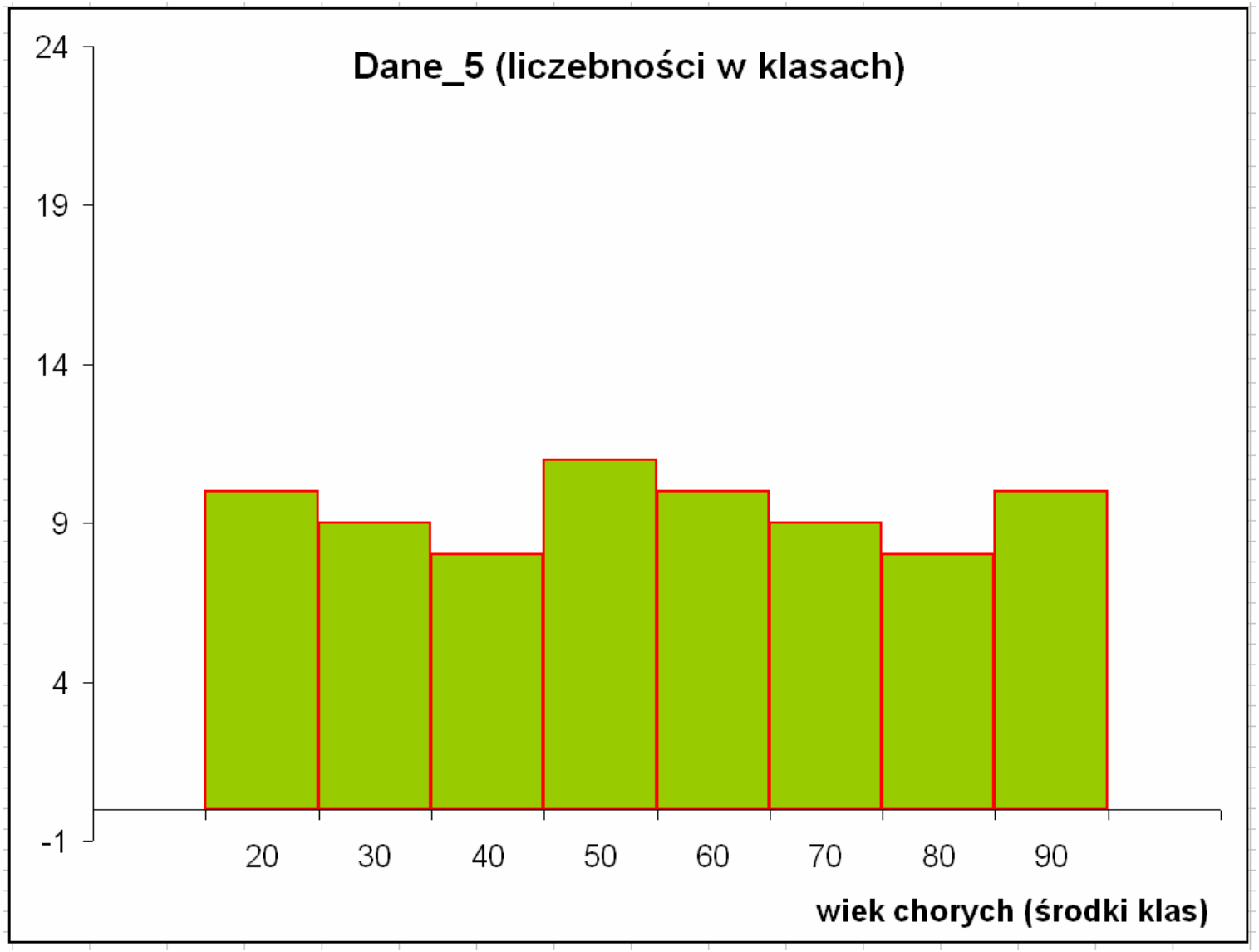
Rozkład asymetryczny
(lewoskośny, asymetria lewostronna)

Empiryczny rozkład wartości cd.



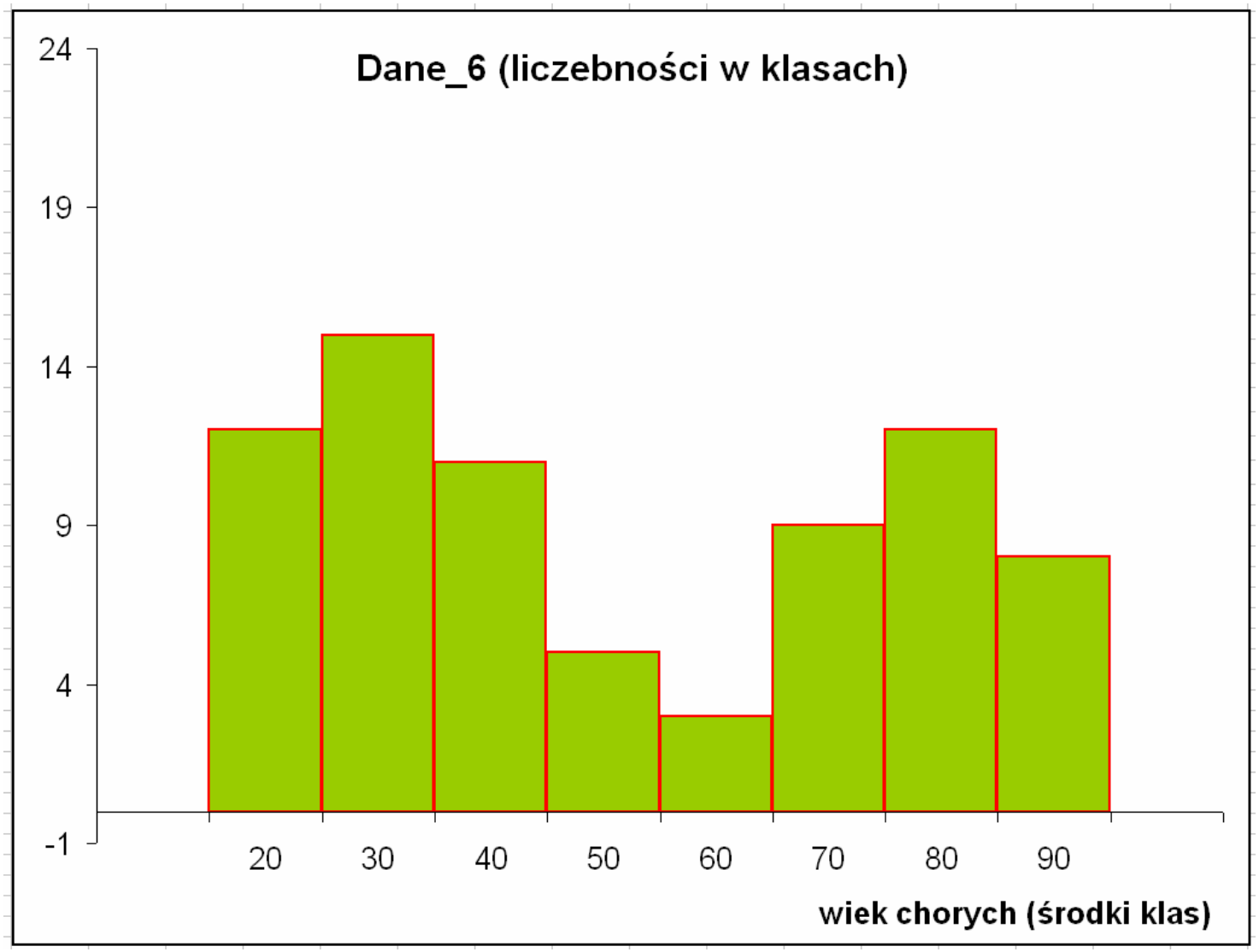
Rozkład typu U

Empiryczny rozkład wartości cd.



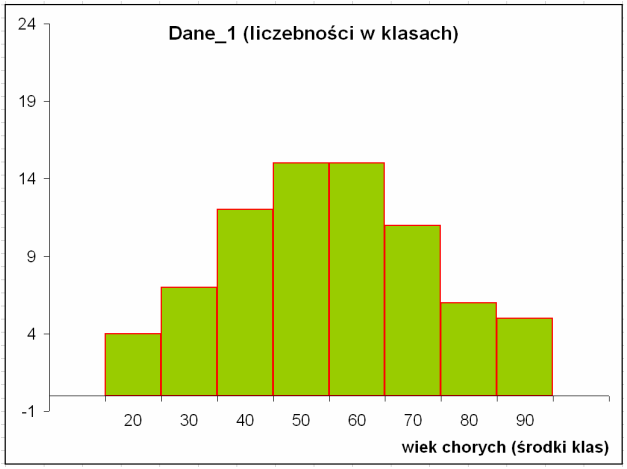
Rozkład równomierny

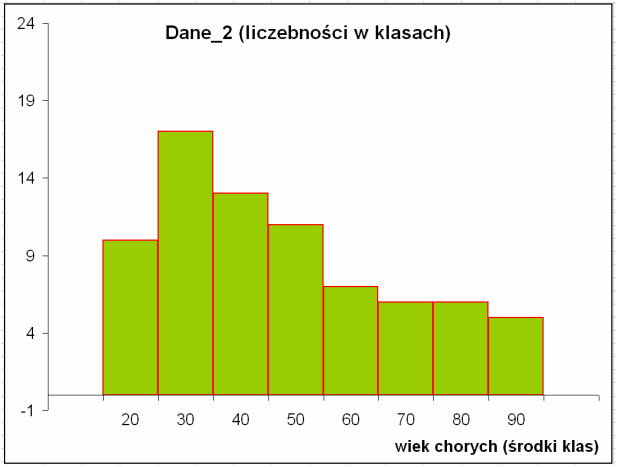
Empiryczny rozkład wartości cd.

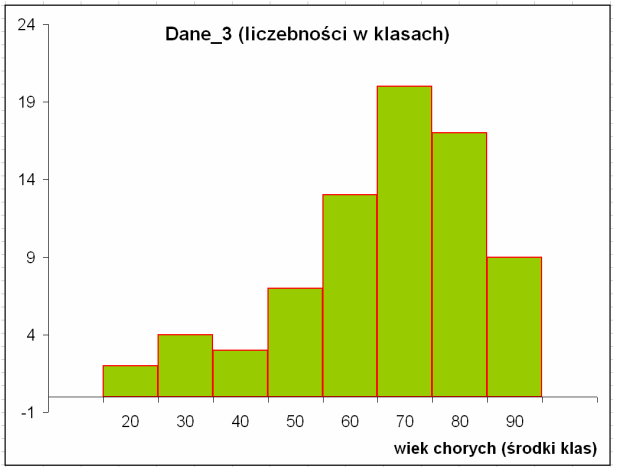


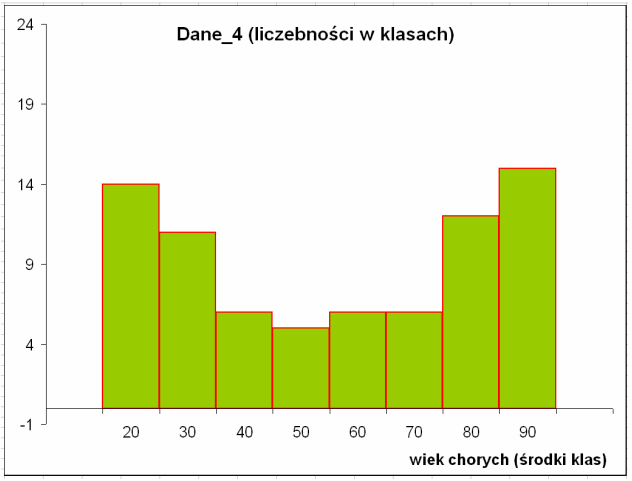
Rozkład dwumodalny

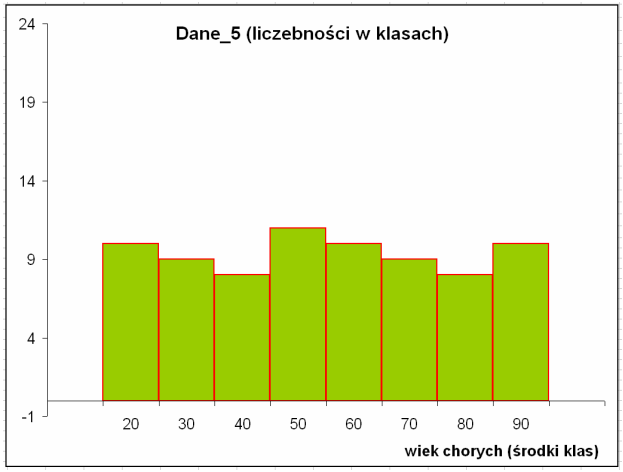
Typy rozkładów empirycznych

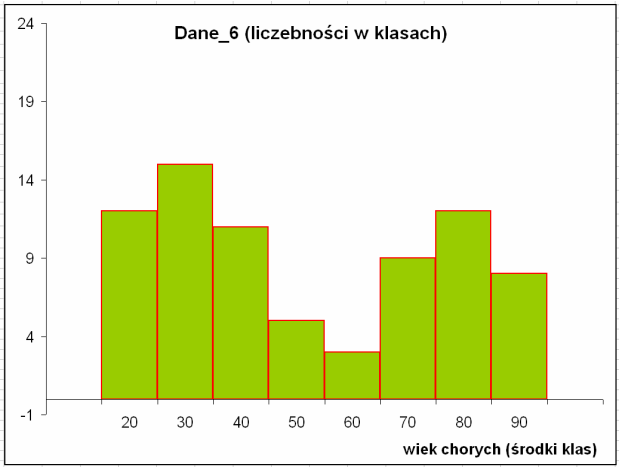












Statystyczny opis danych

Dane liczbowe z pomiaru: $x_1, x_2, \ldots, x_n$
w badaniu pełnym.

Pierwszym etapem analizy danych jest ich opis:

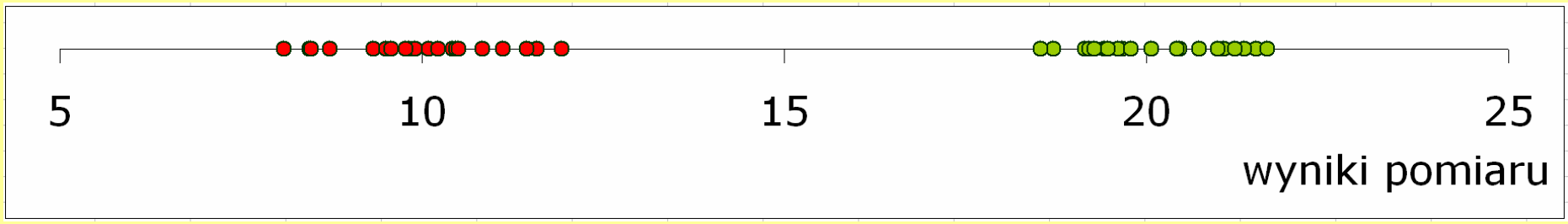
- rozkład wartości (w postaci tabeli, wykresu)
- charakterystyka za pomocą parametrów (np. średnia arytmetyczna, mediana, odchylenie standardowe)

Do czego potrzebujemy parametrów?

Do czego potrzebujemy parametrów?

Parametry opisują własności całego zespołu danych liczbowych:

Rys. 1

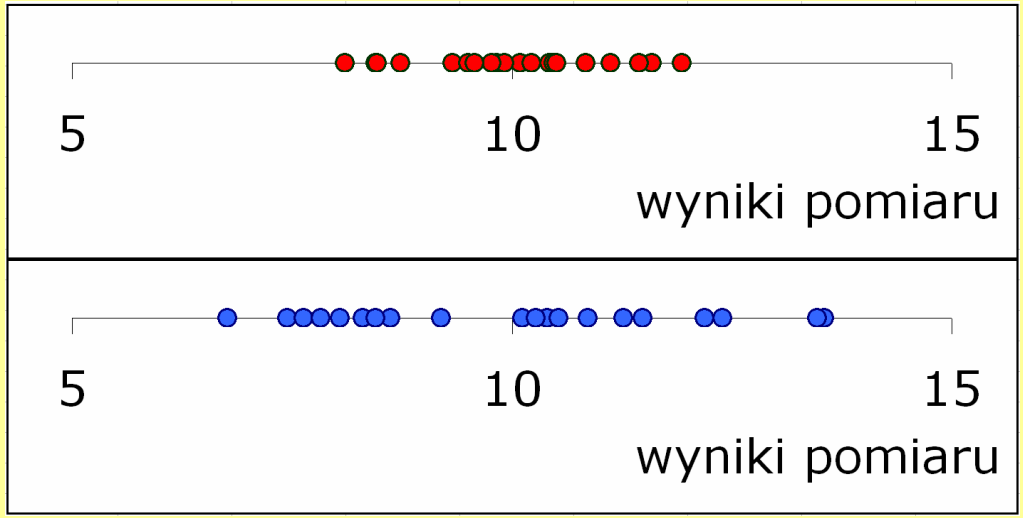


położenie (na osi liczbowej)

Do czego potrzebujemy parametrów?

Parametry opisują własności całego zespołu danych liczbowych:

Rys. 2

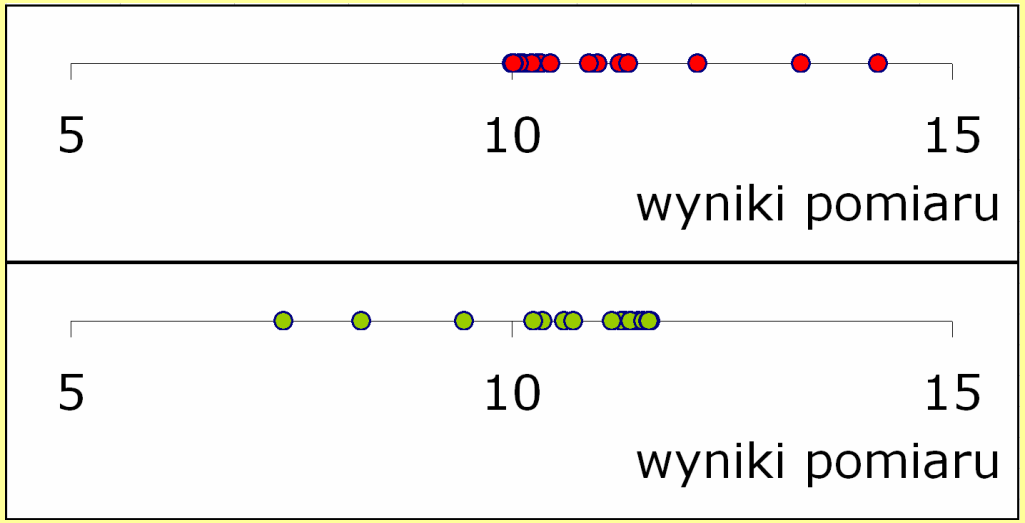


rozrzut (względem średniej)

Do czego potrzebujemy parametrów?

Parametry opisują własności całego zespołu danych liczbowych:

Rys. 3



asymetrię (względem średniej)

Parametry

Parametry opisują własności całego zespołu danych liczbowych:

a. położenie (przeciętny poziom, tendencja centralna)
b. rozrzut (rozproszenie, dyspersja) i zmienność
c. asymetrię (skośność)
d. spłaszczenie

Parametry klasyczne – obliczane na podstawie wszystkich wyników.

Parametry pozycyjne - wyznaczane na podstawie pozycji wyników w szeregu statystycznym lub częstości występowania.

Parametry położenia

Parametry położenia charakteryzują średni lub typowy poziom, wokół którego skupiają się wartości ze zbioru danych:

- wartość średnia (np. arytmetyczna, geometryczna, harmoniczna)
- wartość o ustalonej pozycji w rozkładzie (np. mediana i inne kwantyle)
- wartość najczęściej występująca (dominanta, inaczej: modalna, moda)

Średnia arytmetyczna

Dla danych $x_1, x_2, \ldots, x_n$:

$$\overline{x} = \frac{x_1 + x_2 + \ldots + x_n}{n} = \frac{1}{n}\sum_{i=1}^{n} x_i$$

*Dla szeregu rozdzielczego; k - liczba klas, n_i - liczebność w i-tej klasie:

$$\overline{x}_{sz} = \frac{\dot{x}_1 n_1 + \dot{x}_2 n_2 + \ldots + \dot{x}_k n_k}{n_1 + n_2 + \ldots + n_k} = \frac{\sum_{i=1}^{k} \dot{x}_i n_i}{\sum_{i=1}^{k} n_i}$$

Przykład

W pewnym doświadczeniu medycznym badano czas snu pacjentów leczonych na pewną chorobę. Dla 12 pacjentów otrzymano wyniki (w min.):

435, 389, 533, 324, 561, 395, 416, 500, 499, 397, 356, 398.

Średni czas snu w badanej grupie pacjentów wynosi

$$\overline{x} = \frac{435 + 389 + \ldots + 398}{12} = \frac{5203}{12} \approx 433{,}6 \quad \text{minuty}$$

Modyfikacje średniej arytmetycznej

Na średnią arytmetyczną duży wpływ mają skrajne wartości cechy, zwłaszcza przy małej liczebie danych. Aby go wyeliminować, można stosować inne średnie, które nie uwzględniają wartości skrajnych:

- **średnia przycięta** – liczona jak arytmetyczna, ale po usunięciu określonego procenta górnych i dolnych wartości (np. po 5%) ze zbioru danych

- **średnia Winsora** obliczana po zastąpieniu określonego procenta górnych i dolnych wartości w zbiorze danych wartością minimalną i maksymalną z pozostałej części

Przykład

Dane uporządkowane rosnąco $x_{(i)}$:

24, 35, 189, 195, 197, 198, 216, 235, 533, 561

Winsoryzacja:

189, 189, 189, 195, 197, 198, 216, 235, 235, 235

$$W = \frac{189 \cdot 3 + 195 + 197 + 198 + 216 + 235 \cdot 3}{10} = 207,8$$

Średnia typu Winsor stosowana jest w zagadnieniach medycznych, gdy występują obserwacje odstające.

Wzór na średnią typu Winsor

Dane uporządkowane rosnąco:

$$x_{(1)},\ x_{(2)},\ x_{(3)},\ \ldots,\ x_{(n-2)},\ x_{(n-1)},\ x_{(n)}$$

$$W_{\alpha,\beta} = \frac{1}{n}\left(\sum_{i=m+1}^{p} x_{(i)} + m x_{(m+1)} + (n-p)\, x_{(p)} \right)$$

gdzie: $\alpha \geq 0, \quad \beta \leq 0{,}5, \quad m = [\alpha n], \quad p = n - [\beta n]$

Symbol $[x]$ oznacza część całkowitą liczby x (np. [2,6] = 2).

Średnia typu Winsor jest obliczana z danych, w których (100α)% najmniejszych liczb zostało zastąpione przez $(m+1)$-szą, a (100β)% największych przez p-tą.

Średnia arytmetyczna ważona

Średnia arytmetyczna ważona stosowana jest, gdy pewnym obserwacjom w zbiorze wyników chcemy nadać większe znaczenie (aby miały większy wpływ na obliczaną wartość średniej).

$$\overline{x}_w = \frac{\sum_{i=1}^{n} w_i \cdot x_i}{\sum_{i=1}^{n} w_i}$$

gdzie w_i oznacza wagę wyniku x_i, $w_i > 0$.

Przykład

W grupie pacjentów przeprowadzono dwa testy psychomotoryczne, oba oceniane w skali 0-100. Psycholog uznał, że wagi wyników testów powinny być w stosunku jak 2:3.

Jeśli x_1 , x_2 oznaczają wyniki pacjenta z obu testów, to

$$\overline{x}_w = \frac{2 \cdot x_1 + 3 \cdot x_2}{2+3} = 0{,}4 \cdot x_1 + 0{,}6 \cdot x_2$$

Id_pacjenta	Test 1	Test 2	Ocena końcowa
1.	20	85	(20*2+85*3)/(2+3)=59
2.	85	20	(85*2+20*3)/(2+3)=46

Zwykła średnia arytmetyczna wynosi (20+85)/2=52,5.

Średnia arytmetyczna ważona cd.

Przykład zastosowania:

Średnia ze wcześniej obliczonych średnich, gdy nie są one równocenne (bo pochodzą ze zbiorów danych o różnej liczebności lub powierzchni badawczych o różnych rozmiarach lub gdy przy ich obliczaniu stosowano różną dokładność pomiaru). Wówczas wagą może być liczebność zbioru danych, powierzchnia badawcza, odwrotności dokładności pomiarów.

Przykład

Trzy osoby mierzyły pierśnicę drzew w jednym oddziale leśnym.

	Liczba zmierzonych drzew	Obliczona średnia [cm]
Osoba 1	5	120,0
Osoba 2	20	70,0
Osoba 3	50	40,0

Błędne podejście: średnia ze średnic wynosi 76,7 cm – nadmierną wagę przyjmujemy dla niewielu pomiarów pierwszej osoby, a zbyt małą do najliczniejszych pomiarów trzeciej osoby.

Poprawne podejście:

$$\frac{5\times 120+20\times 70+50\times 40}{5+20+50}=53{,}3$$

Średnia geometryczna

Średnia geometryczna jest stosowana do danych, które stanowią ilorazy pewnych wielkości. Oblicza się ją tylko dla liczb nieujemnych.

$$\overline{x}_g = \sqrt[n]{x_1 \cdot x_2 \cdot x_3 \cdot \ldots \cdot x_n}$$

Przykład zastosowania

W biologii oblicza się średnią ze współczynników śmiertelności lub przyrostu badanej cechy w kolejnych momentach czasu (w szeregach czasowych).

Przykład

Przez cztery lata badano rozrodczość i śmiertelność pewnego owada w wybranym siedlisku.

	Liczebność zaobserwowana	Współczynnik reprodukcji netto
Rok 1	10	
Rok 2	40	R1=40/10=4
Rok 3	40	R2=40/40=1
Rok 4	80	R3=80/40=2

Średni wsp. reprodukcji netto ze wszystkich lat

$$\overline{R} = \sqrt[3]{4 \cdot 1 \cdot 2} = 2$$

Sprawdzenie: 10·2·2·2=80

Przykład cd.

Zauważmy, że średnia arytmetyczna wynosi tu 2,33, a więc w ciągu trzech lat populacja motyli wzrosłaby $(2,33)^3 \sim 12,6$ raza, co nie miało miejsca.

Średnia harmoniczna

Średnia harmoniczna jest stosowana do danych wyrażonych w jednostkach względnych, np.: prędkość w km/h, zagęszczenie/zaludnienie w liczbie osób/km^2.

$$\overline{x}_h = \frac{n}{\sum_{i=1}^{n} \frac{1}{x_i}}$$

Średnią harmoniczną stosujemy tylko do liczb dodatnich. Pozwala nadać większe znaczenie mniejszym wartościom w zbiorze danych.

Średnia harmoniczna ważona

Średnia harmoniczna ważona pozwala nadać wagi poszczególnym wynikom ze zbioru danych:

$$\overline{x}_{hw} = \frac{\sum_{i=1}^{n} w_i}{\sum_{i=1}^{n} \frac{w_i}{x_i}}$$

Zależności między średnimi

Między średnią arytmetyczną, geometryczną i harmoniczną dla dowolnego zbioru danych występuje zależność:

$$\overline{x}_h \leq \overline{x}_g \leq \overline{x}$$

Dominanta

Dominanta to wartość występująca najczęściej w zbiorze danych (modalna, moda).

Dla danych zestawionych w szeregu rozdzielczym wskazuje na nią najliczniejsza klasa, na histogramie – najwyższy szczyt. W zależności od liczby szczytów widocznych na histogramie, mamy rozklad jednomodalny (unimodalny), dwumodalny (bimodalny), trójmodalny, itd. Wielomodalność rozkładu świadczy o niejednorodności zbioru danych.

Przykład

Pacjenci pewnej kliniki pogrupowani według czasu działania leku przeciwbólowego.

Czas działania leku [min.]	Liczba pacjentów
<7,5 ; 12,5)	4
<12,5 ; 17,5)	29
<17,5 ; 22,5)	39
<22,5 ; 27,5)	**81**
<27,5 ; 32,5)	35
<32,5 ; 37,5)	9
<37,5 ; 42,5>	4

Dominanta należy do przedziału od 22,5 do 27,5.

Dominanta w szeregu rozdzielczym

Rachunkowe wyznaczenie dominanty:

$$D(x) \approx {}_{D}x_0 + \frac{n_D - n_{D-1}}{\left(n_D - n_{D-1}\right) + \left(n_D - n_{D+1}\right)} \cdot h_D$$

n_D – liczebność w przedziale dominanty

n_{D-1} – liczebność w przedziale poprzedzającym przedział dominanty

n_{D+1} – liczebność w przedziale następujacym po przedziale dominanty

h_D – długość przedziału dominanty

${}_{D}x_0$ – początek przedziału dominanty

Przykład cd.

Dla wartości z przykładu:

$$D(x) \approx 22{,}5 + \frac{81-39}{(81-39)+(81-35)} \cdot 5 = 25{,}4$$

Najczęściej spotykanym czasem działania leku przeciwbólowego jest 25,4 min. (przy średniej arytmetycznej 23,9 min.).

Mediana

Mediana Me to wartość środkowa w zbiorze danych (parametr pozycyjny). Gdy liczba danych jest nieparzysta, medianę stanowi wartość środkowego elementu. W przypadku parzystej liczby pomiarów w próbie, medianę stanowi średnia wartość z dwóch sąsiadujących środkowych elementów uporządkowanego zbioru.

$$Me = \frac{x_{n/2} + n_{n/2+1}}{2} \qquad dla\ n\ parzystych$$

$$Me = x_{(n+1)/2} \qquad dla\ n\ nieparzystych$$

Przykłady

Przykład 1. Obliczanie mediany dla n = 9 danych:

35, 40, 36, 35, 39, 37, 38, 36, 38

Po uporządkowaniu:

35, 35, 36, 36, 37, 38, 38, 39, 40; $Me=37$

Przykład 2. Obliczanie mediany dla n = 8 danych:
35, 35, 36, 36, 37, 38, 38, 39;

$Me=(36+37)/2=36,5$

Mediana

Dla danych zagregowanych w szereg rozdzielczy przy wyznaczaniu mediany można wykorzystać szereg liczebności skumulowanych. Mediana znajduje się w klasie, w której skumulowane liczebności przekraczają lub co najmniej osiągają numer kolejnej jednostki środkowej.

Jeśli dane o charakterze skokowym pogrupowane są w szereg rozdzielczy, to medianę stanowi środek przedziału, w którym się ona znajduje.

Przykład

Pacjenci pewnej kliniki pogrupowani według czasu działania leku przeciwbólowego.

Czas działania leku	**Liczba pacjentów**	**L. pacjentów skumulowana**
<7,5; 12,5)	4	4
<12,5;17,5)	29	33
<17,5;22,5)	39	72
<22,5;27,5)	**81**	**153**
<27,5;32,5)	35	188
<32,5;37,5)	9	197
<37,5;42,5>	4	201
Razem	201	

Mediana należy do przedziału od 22,5 do 27,5.

Mediana w szeregu rozdzielczym

Jeśli dane o charakterze ciągłym pogrupowane są w szereg rozdzielczy (także, gdy występują rangi wiązane), to do wyznaczania mediany stosuje się wzór interpolacyjny:

$$Me = {}_{M}x_0 + \frac{h_M}{n_M} \cdot \left(\frac{n}{2} - F_0 \right)$$

gdzie:

${}_{M}x_0$ – dolna granica przedziału mediany,

h_M – długość przedziału mediany,

n_M – liczebność w przedziale mediany,

n – liczba wszystkich danych,

F_0 – liczebność skumul. przedz. przed medianą

Przykład cd.

Pacjenci pewnej kliniki pogrupowani według czasu działania leku przeciwbólowego.

$$Me = 22{,}5 + \frac{5}{81} \cdot \left(\frac{201}{2} - 72 \right) = 24{,}2$$

Mediana wynosi 24,2 min. Oznacza to, że dla połowy pacjentów czas działania leku nie przekracza 24,2 min. i dla takiej samej liczby pacjentów jest nie mniejszy od tej wartości.

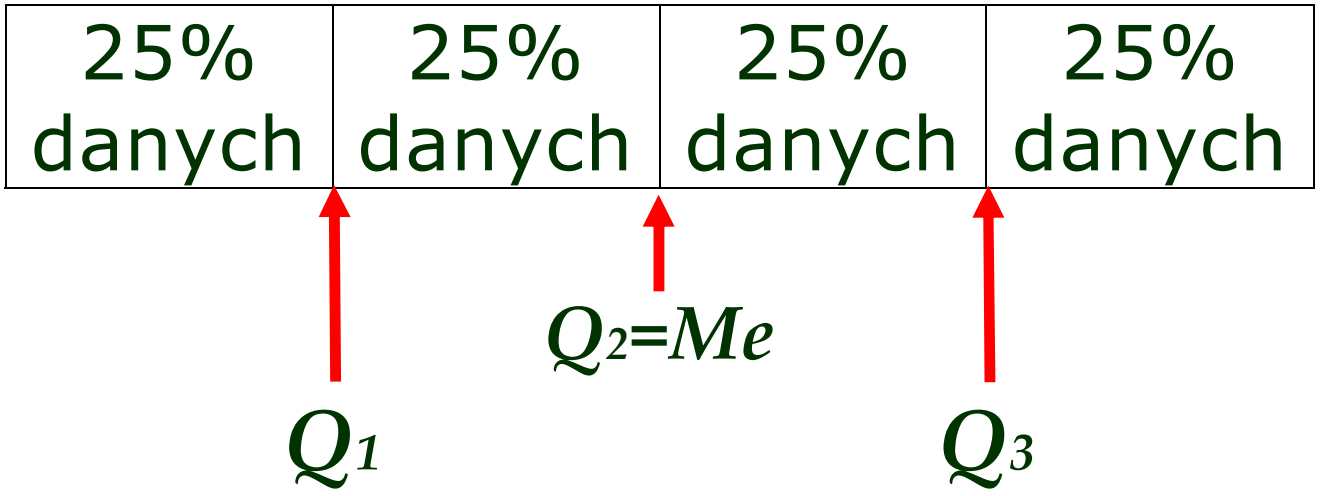
Kwantyle

Mediana jest jednym z kwantyli.

Kwantyle to wartości, które dzielą uporządkowany zbiór danych na części o jednakowej liczbie elementów.

Kwartyle dzielą zbiór danych na cztery części, **decyle** na dziesięć i **percentyle** (centyle) na sto części.

Kwartyle



Kwartyle

Pierwszy kwartyl Q_1 dzieli uporządkowany zbiór danych w taki sposób, że 25% elementów zbioru ma wartości nie większe, a 75% nie mniejsze od tego elementu.

Drugi kwartyl (mediana) $Q_2 = Me$ dzieli uporządkowany zbiór danych na dwie równe części. Wartości skrajne nie mają wpływu na jej wielkość.

Trzeci kwartyl Q_3 dzieli uporządkowany zbiór danych w taki sposób, że 75% elementów zbioru ma wartości nie większe, a 25% nie mniejsze od tego elementu.

Kwartyle

Przy wyznaczaniu kwartyli pierwszego i trzeciego stosuje się takie same zasady, jak dla mediany.

Dla danych ciągłych wzory interpolacyjne są następujące:

$$Q_1 = x_0 + \frac{h}{n_0} \cdot \left(\frac{N}{4} - F_0 \right), \qquad Q_3 = x_0 + \frac{h}{n_0} \cdot \left(\frac{3N}{4} - F_0 \right)$$

gdzie:

x_0 – dolna granica przedziału pierwszego lub trzeciego kwartyla, h – szerokość przedziału pierwszego lub trzeciego kwartyla, n_0 – liczebność w przedziale pierwszego lub trzeciego kwartyla, N – liczba wszystkich danych, F_0 – liczebność skumul. przedziału poprzedniego

Kwartyle w programie STATISTICA

Program STATISTICA daje mozliwość wyznaczania kwantyli, ale nie ma wśród nich metody opisanej powyżej. Dlatego w celu wyznaczenia mediany dla danych ciągłych pogrupowanych w przedziały trzeba korzystać z arkusza kalkulacyjnego.

Parametry rozrzutu i zmienności

Parametry rozrzutu opisują zróżnicowanie w zbiorze danych:

- wariancja i odchylenie standardowe
- współczynnik zmienności
- rozstęp
- rozstęp kwartylowy
- odchylenie ćwiartkowe

Wariancja

Wariancja s^2 pokazuje rozproszenie danych wokół średniej arytmetycznej.

$$s^2 = \frac{(x_1 - \bar{x})^2 + (x_2 - \bar{x})^2 + \ldots + (x_n - \bar{x})^2}{n} = \frac{\sum_{i=1}^{n} (x_i - \bar{x})^2}{n}$$

Najmniejsza wartość wariancji wynosi zero, im większe zróżnicowanie, tym większa wartość wariancji.

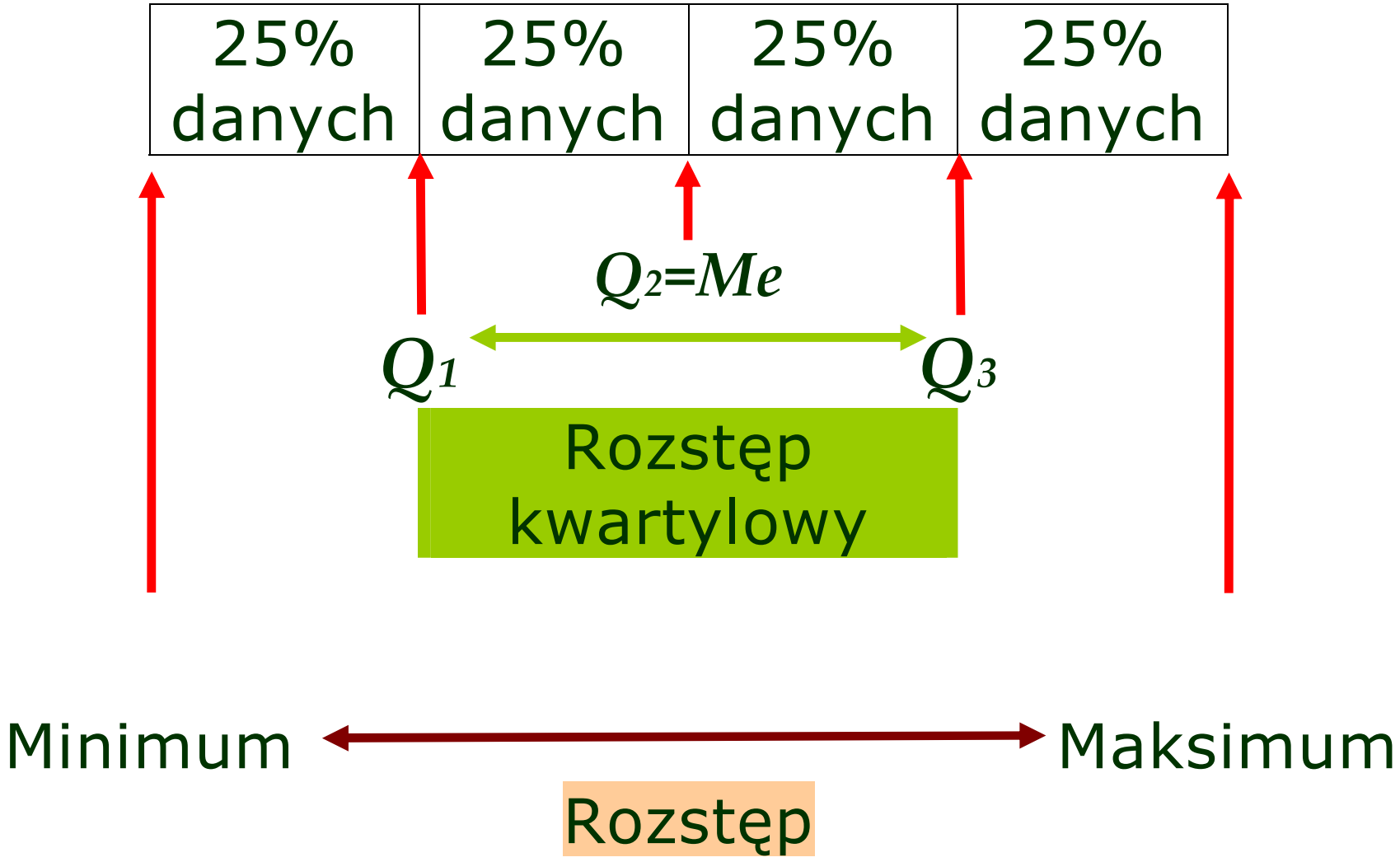
Odchylenie standardowe

Odchylenie standardowe s, SD (*standard deviation*)

$$SD = \sqrt{s^2}$$

Odchylenie standardowe pokazuje, o ile przeciętnie poszczególne wyniki różnią się od średniej, czyli pokazuje wielkość błędu pojedynczego pomiaru. Im mniejsza wartość odchylenia, tym obserwacje są bardziej skupione wokół średniej.

Parametry pozycyjne



Rozstęp

Rozstęp wskazuje na całkowity obszar zmienności badanej cechy. Jest obliczany jako różnica pomiędzy największą i najmniejszą wartością w zbiorze danych.

$$R = x_{max} - x_{min}$$

Określają go tylko dwie skrajne wartości, pozostałe nie mają żadnego wpływu na wielkość rozstępu. Dlatego zamiast rozstępu czasem podaje się zakres od 5% do 95% wartości, pomijając w ten sposób po 5% wartości skrajnych z każdego końca.

Odchylenie ćwiartkowe

Rozstęp kwartylowy obliczamy jako różnicę pomiędzy trzecim i pierwszym kwartylem. Pomiędzy tymi kwartylami znajduje się 50% wszystkich obserwacji.

Odchylenie ćwiartkowe Q obliczamy jako połowę różnicy pomiędzy trzecim i pierwszym kwartylem.

$$Q = \frac{Q_3 - Q_1}{2}$$

Rozstęp kwartylowy i odchylenie ćwiartkowe są obliczane, gdy miarą położenia jest mediana. Wskazują one na rozrzut wyników wokół mediany.

Współczynnik zmienności

Współczynnik zmienności v, cv (*coefficient of variation*) oblicza się dzieląc odchylenie standardowe przez średnią arytmetyczną - jest parametrem bezwymiarowym:

$$v = \frac{SD}{\overline{x}} \qquad \text{lub} \qquad v = \frac{SD}{\overline{x}} \cdot 100\%$$

Im mniejszą przyjmuje wartość, tym mniejsza zmienność zbioru danych względem średniej.

Współczynnik zmienności - pozycyjny

Współczynnik zmienności pozycyjny V oblicza się dzieląc odchylenie ćwiartkowe przez medianę:

$$V = \frac{Q}{Me} \qquad \text{lub} \qquad V = \frac{Q}{Me} \cdot 100\%$$

Przykład

					średnia arytmet.	odch. stand	wsp. zmienności
Dane1	6	10	12	12	10,0	2,4	24%
Dane2	56	60	62	62	60,0	2,4	4%

Współczynnik zmienności umożliwia porównanie zmienności zbiorów danych różniących się znacznie wartością średniej lub zawierających pomiary wykonane w różnych jednostkach.

Współczynnik zmienności stosujemy do porównania, która cecha jest bardziej zróżnicowana relatywnie względem swojej średniej. Jeśli oceniamy zróżnicowanie względem mediany, to stosujemy pozycyjny współczynnik zmienności.

Parametry asymetrii (skośności)

Miarę asymetrii można oprzeć na spostrzeżeniu, że w szeregu statystycznym o rozkładzie symetrycznym średnia arytmetyczna, mediana oraz dominanta są równe

$$\overline{x} = Me = D$$

W szeregach asymetrycznych wartości tych parametrów odbiegają od siebie tym bardziej, im asymetria jest większa. **Wskaźnikiem asymetrii** może być różnica

$$\overline{x} - D$$

Gdy

$$\overline{x} - D > 0 \quad asymetria \quad prawostronna$$

$$\overline{x} - D < 0 \quad asymetria \quad lewostronna$$

Współczynnik asymetrii

Współczynnik asymetrii (jeden z wielu) jest liczbą niemianowaną, a jego znak mówi o kierunku asymetrii.

$$A_3 = \frac{\overline{x} - D}{s}$$

- A_3=0 rozkład symetryczny
- A_3>0 rozkład prawoskośny, asymetria prawostronna
- A_3<0 rozkład lewoskośny, asymetria lewostronna

Parametr spłaszczenia

Kurtoza (wskaźnik spłaszczenia, wskaźnik smukłości) wskazuje na koncentrację danych wokół średniej. Wartości kurtozy mniejsze od zera świadczą o rozkładzie spłaszczonym (platykurtycznym), zaś wartości większe od zera o rozkładzie wysmukłym (leptokurtycznym) w stosunku do rozkładu normalnego.

Podsumowanie

Parametry tak charakteryzują zbiór danych, że porównanie różnych zbiorów danych można sprowadzić do porównań parametrów.

Przykład

W dwóch grupach chorych zmierzono skurczowe ciśnienie krwi. Otrzymano wyniki:

grupa I: 145, 125, 130, 155, 140, 150, 135

grupa II: 115, 150, 100, 180, 140, 165, 130

Obliczenia w arkuszu kalkulacyjnym

	A	B	C	D	E	F	G	H	I	J
1		**grupa I**	**grupa II**							
2		145	115							
3		125	150							
4		130	100							
5		155	180							
6		140	140							
7		150	165							
8		135	130							
9	średnia arytmetyczna	140,0	140,0							
10	odchylenie standardowe	10,0	25,8							
11	mediana	140,0	140,0							
12	odchylenie ćwiartkowe	7,5	17,5							

	A	B	C
1		**grupa I**	**grupa II**
2		145	115
3		125	150
4		130	100
5		155	180
6		140	140
7		150	165
8		135	130
9	średnia arytmetyczna	=ŚREDNIA(B2:B8)	=ŚREDNIA(C2:C8)
10	odchylenie standardowe	=ODCH.STANDARD.POPUL(B2:B8)	=ODCH.STANDARD.POPUL(C2:C8)
11	mediana	=MEDIANA(B2:B8)	=MEDIANA(C2:C8)
12	odchylenie ćwiartkowe	=(KWARTYL(B2:B8;3)-KWARTYL(B2:B8;1))/2	=(KWARTYL(C2:C8;3)-KWARTYL(C2:C8;1))/2

Przykład cd.

Jeśli położenie zbioru danych określamy używając średniej arytmetycznej (miara klasyczna), to do opisu rozrzutu także użyjemy miary klasycznej (odchylenie standardowe). Natomiast jeśli położenie rozkładu opisujemy używając miar pozycyjnych (mediana, dominanta), to rozrzut opiszemy za pomocą odchylenia ćwiartkowego (miara pozycyjna).

Obliczenia w pakiecie STATISTICA

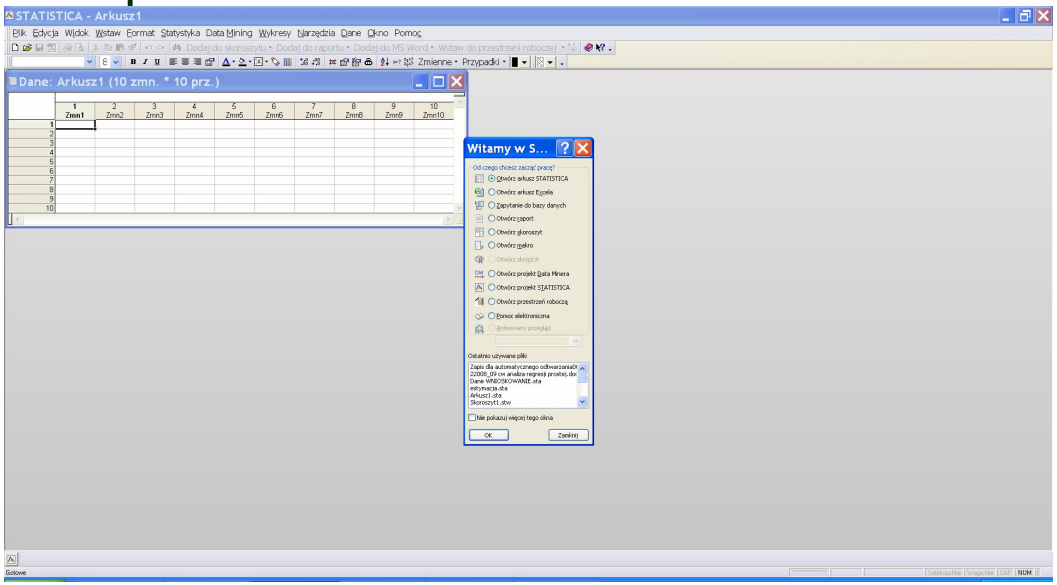
Przykład. (plik biegusy.xls) Zebrano dane o pomiarach 25 dorosłych i 38 młodych biegusów płaskodziobych. Należy przedstawić statystyki opisowe dla pomiaru skrzydła (kolumna SKRZ) oddzielnie dla obu grup wiekowych.

	A	B	C	D	E	F	G	H	I	J	K
1	WIEK	DCG	DZ	SKOK	SKRZ	CIEZ					
2	1	50	28,7	21,5	107	30					
3	1	55,3	32,6	21,4	108	31					
4	1	53	31	22,1	108	32					
5	1	55,1	33,7	23	105	31					
6	1	54,2	33	109	115	33					
7	1	53,3	30,3	22,8	108	33					
8	1	51	28,5	22,8	107	42					
9	1	48,5	27,5	21,9	105	26					
10	1	54,3	32,6	21,2	108	33					
11	1	54,1	32,7	22	113	32					
12	1	53,8	32,2	21,2	110	44					
13	1	50,7	30	22,2	107	28					
14	1	53,2	31,7	22,7	111	40					

Obliczenia w pakiecie STATISTICA

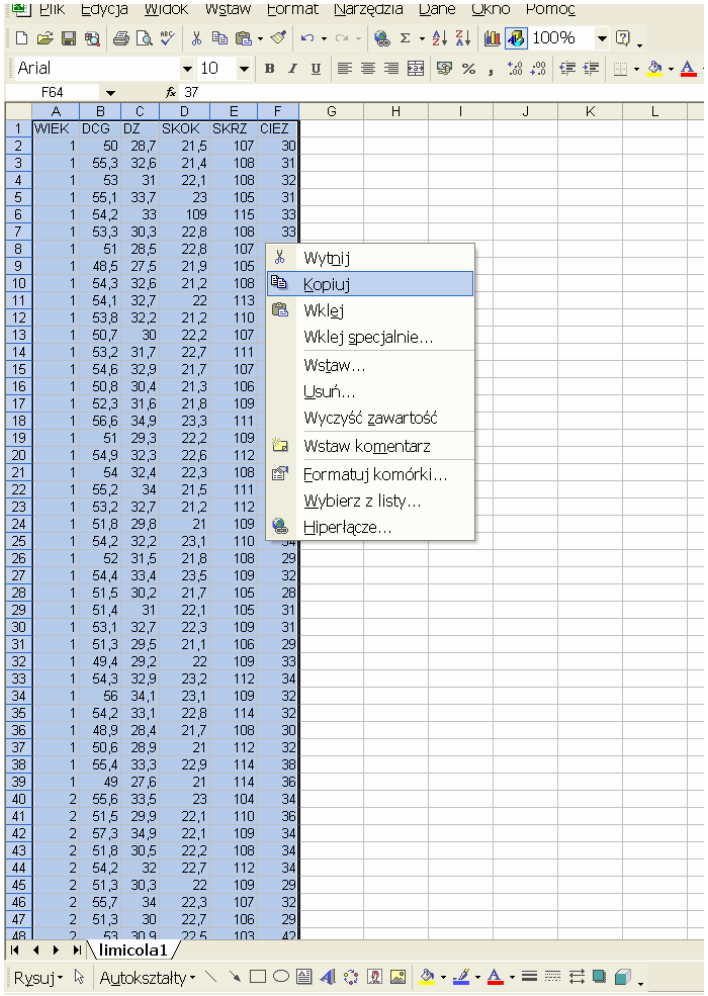
Uruchamiamy pakiet:

Zamykamy okienko powitalne:



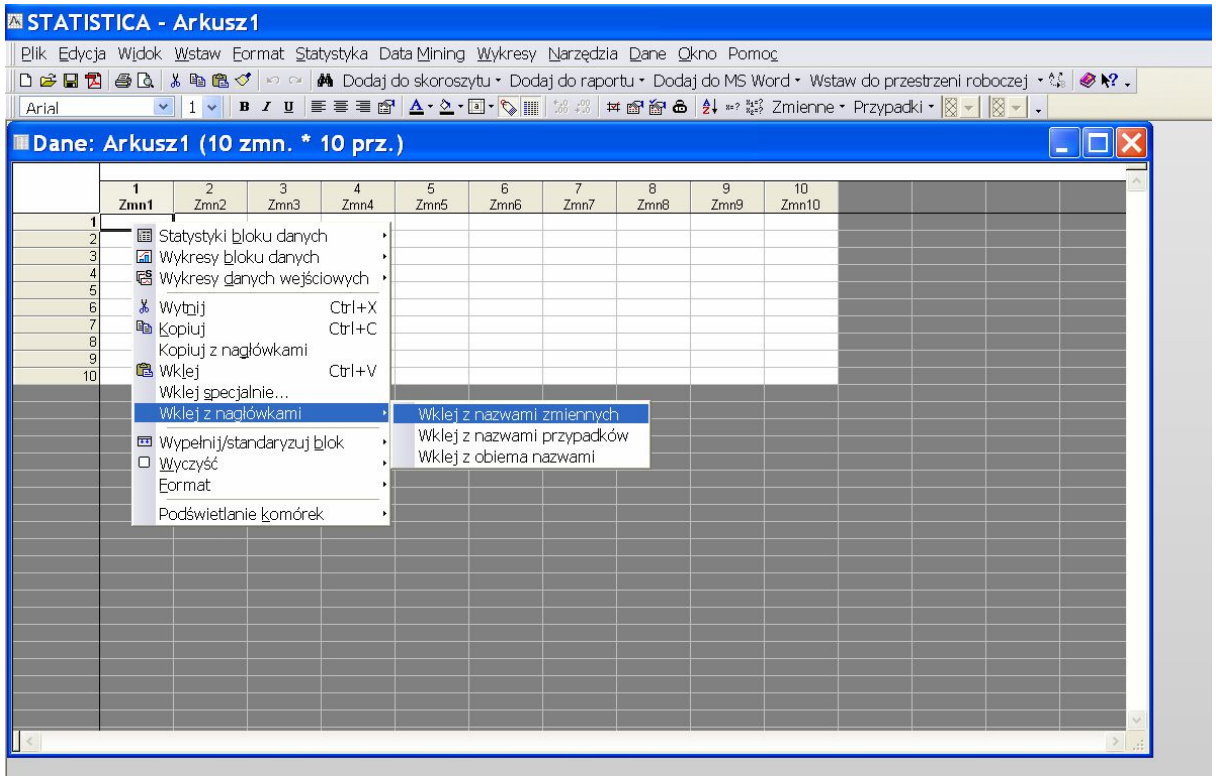
Obliczenia w pakiecie STATISTICA

Zaznaczamy dane w arkuszu, wybieramy **Kopiuj** z menu podręcznego:

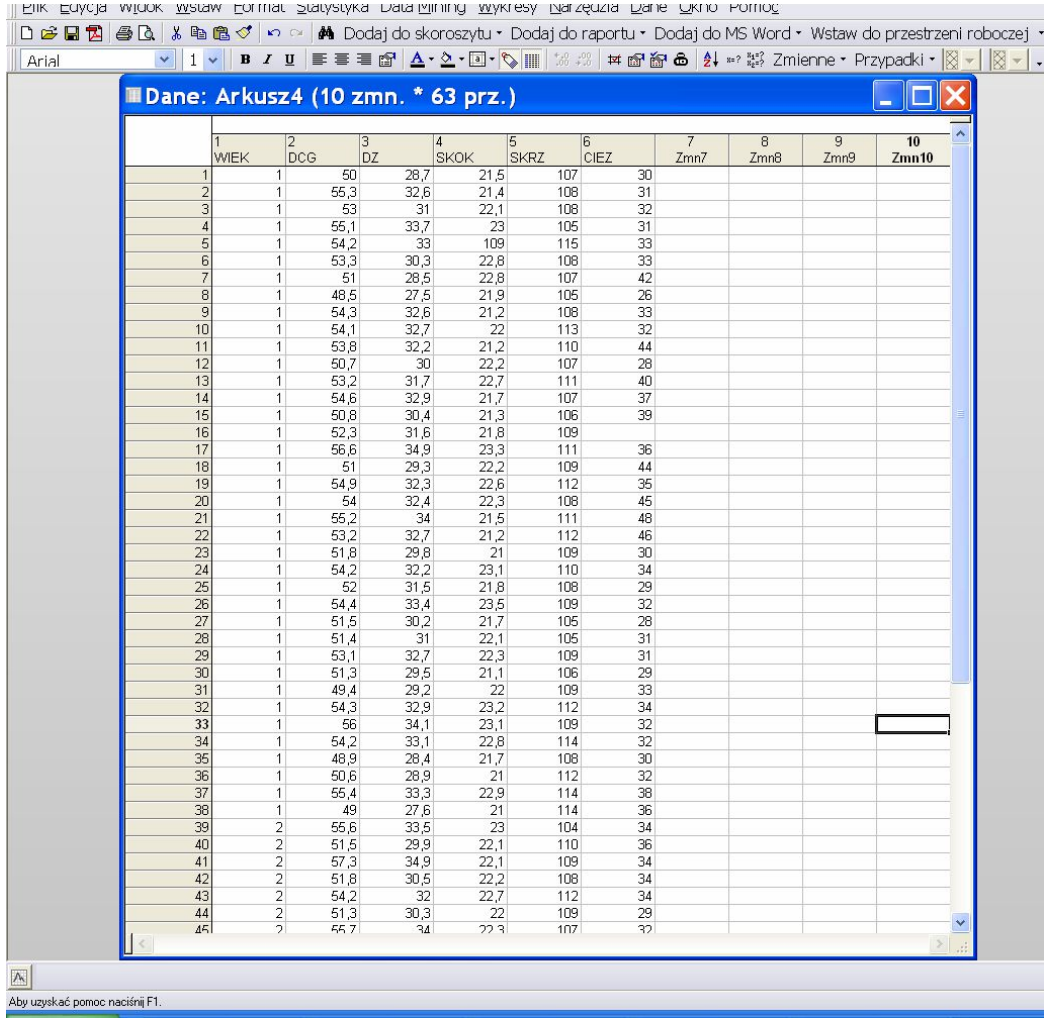


Obliczenia w pakiecie STATISTICA

Przechodzimy do pakietu, w pierwszej komórce pierwszej kolumny (zmiennej) klikamy prawym przyciskiem myszy, wybieramy **Wklej z nagłówkami/Wklej z nazwami zmiennych**



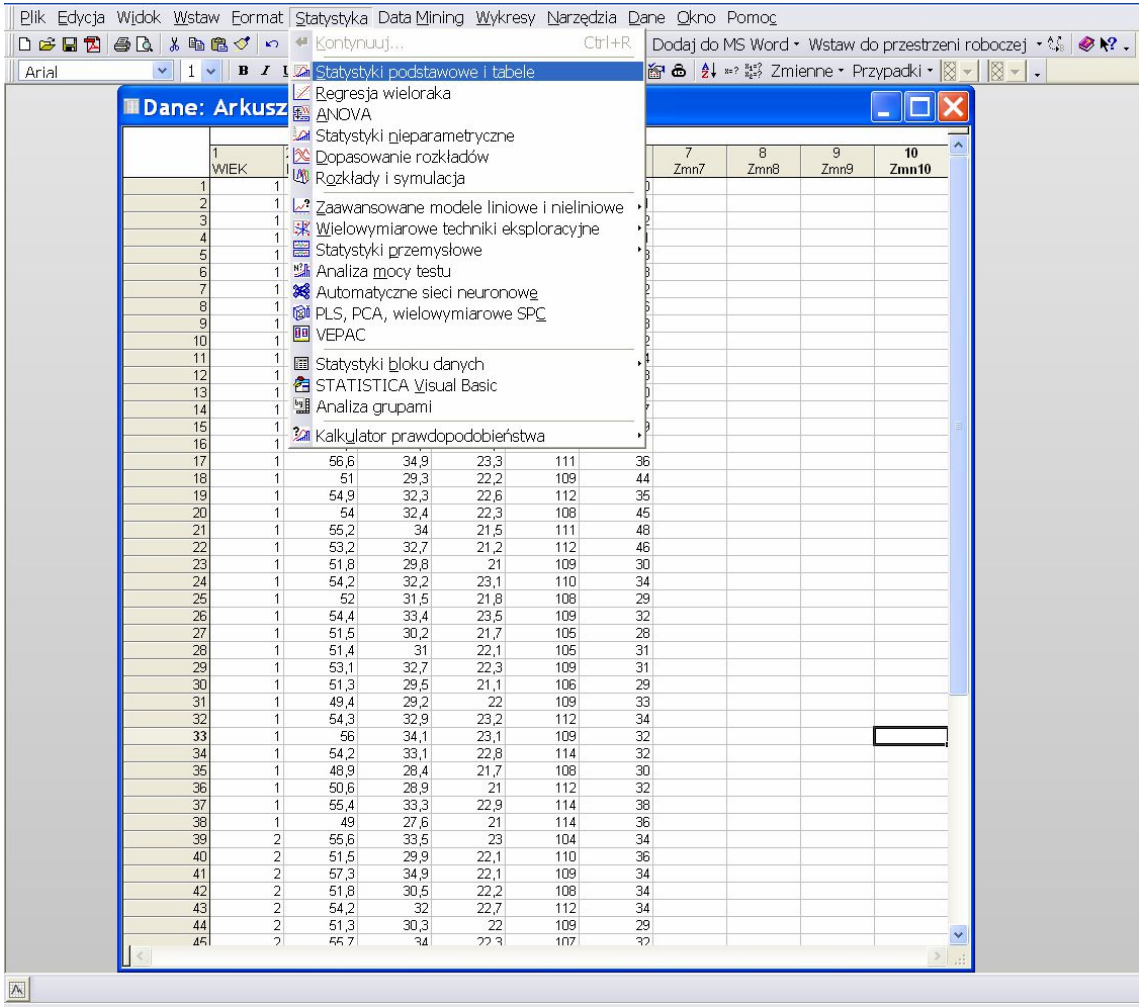
Obliczenia w pakiecie STATISTICA



	1 WIEK	2 DCG	3 DZ	4 SKOK	5 SKRZ	6 CIEZ	7 Zmn7	8 Zmn8	9 Zmn9	10 Zmn10
1	1	50	28,7	21,5	107	30				
2	1	55,3	32,6	21,4	108	31				
3	1	53	31	22,1	108	32				
4	1	55,1	33,7	23	105	31				
5	1	54,2	33	109	115	33				
6	1	53,3	30,3	22,8	108	33				
7	1	51	28,5	22,8	107	42				
8	1	48,5	27,5	21,9	105	26				
9	1	54,3	32,6	21,2	108	33				
10	1	54,1	32,7	22	113	32				
11	1	53,8	32,2	21,2	110	44				
12	1	50,7	30	22,2	107	28				
13	1	53,2	31,7	22,7	111	40				
14	1	54,6	32,9	21,7	107	37				
15	1	50,8	30,4	21,3	106	39				
16	1	52,3	31,6	21,8	109					
17	1	56,6	34,9	23,3	111	36				
18	1	51	29,3	22,2	109	44				
19	1	54,9	32,3	22,6	112	35				
20	1	54	32,4	22,3	108	45				
21	1	55,2	34	21,5	111	48				
22	1	53,2	32,7	21,2	112	46				
23	1	51,8	29,8	21	109	30				
24	1	54,2	32,2	23,1	110	34				
25	1	52	31,5	21,8	108	29				
26	1	54,4	33,4	23,5	109	32				
27	1	51,5	30,2	21,7	105	28				
28	1	51,4	31	22,1	105	31				
29	1	53,1	32,7	22,3	109	31				
30	1	51,3	29,5	21,1	106	29				
31	1	49,4	29,2	22	109	33				
32	1	54,3	32,9	23,2	112	34				
33	1	56	34,1	23,1	109	32				
34	1	54,2	33,1	22,8	114	32				
35	1	48,9	28,4	21,7	108	30				
36	1	50,6	28,9	21	112	32				
37	1	55,4	33,3	22,9	114	38				
38	1	49	27,6	21	114	36				
39	2	55,6	33,5	23	104	34				
40	2	51,5	29,9	22,1	110	36				
41	2	57,3	34,9	22,1	109	34				
42	2	51,8	30,5	22,2	108	34				
43	2	54,2	32	22,7	112	34				
44	2	51,3	30,3	22	109	29				
45	2	55,7	34	22,3	107	32				

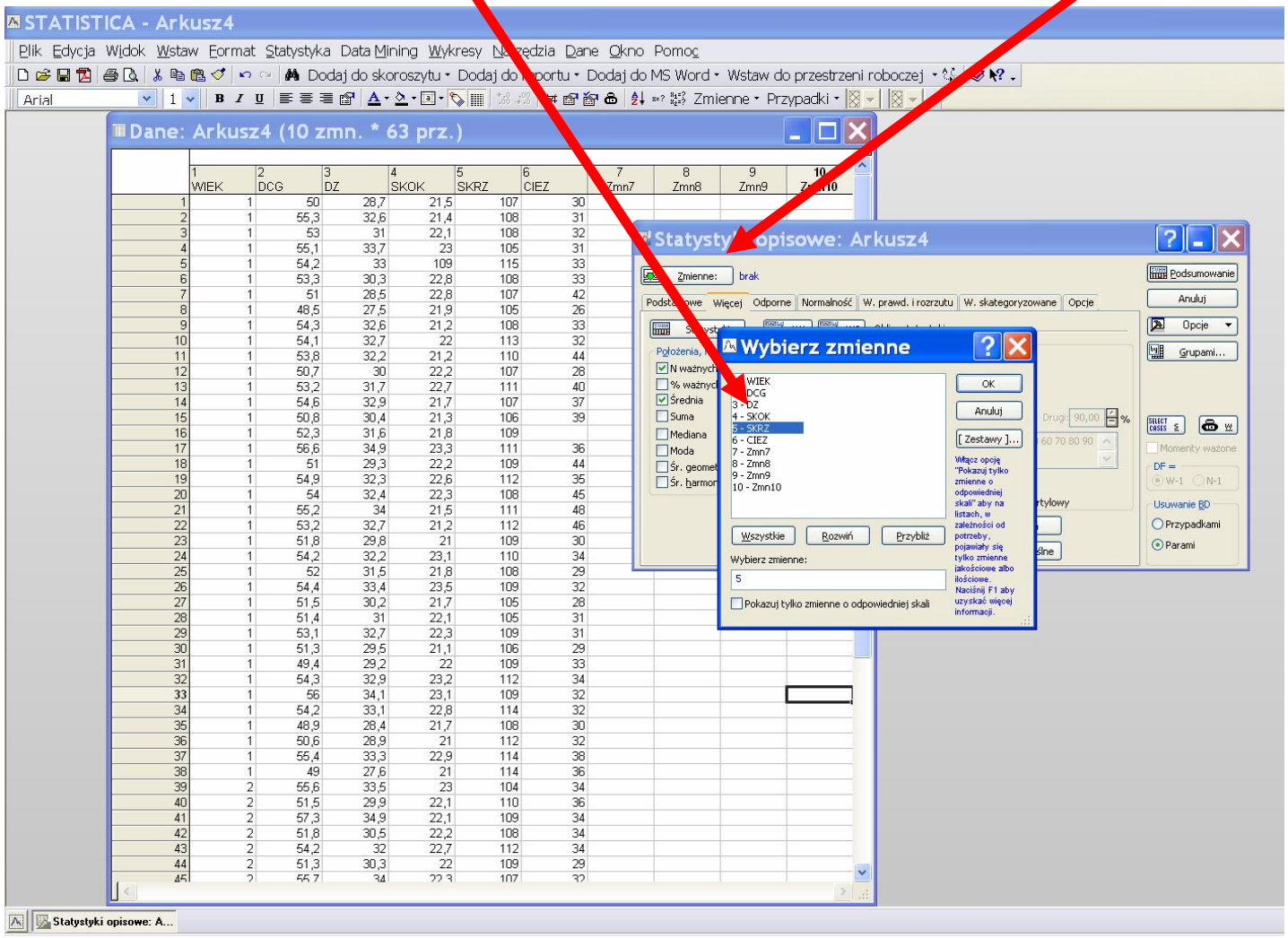
Obliczenia w pakiecie STATISTICA

Wybieramy z menu opcję **Statystyka/Statystyki podstawowe i tabele/Statystyki opisowe/OK**

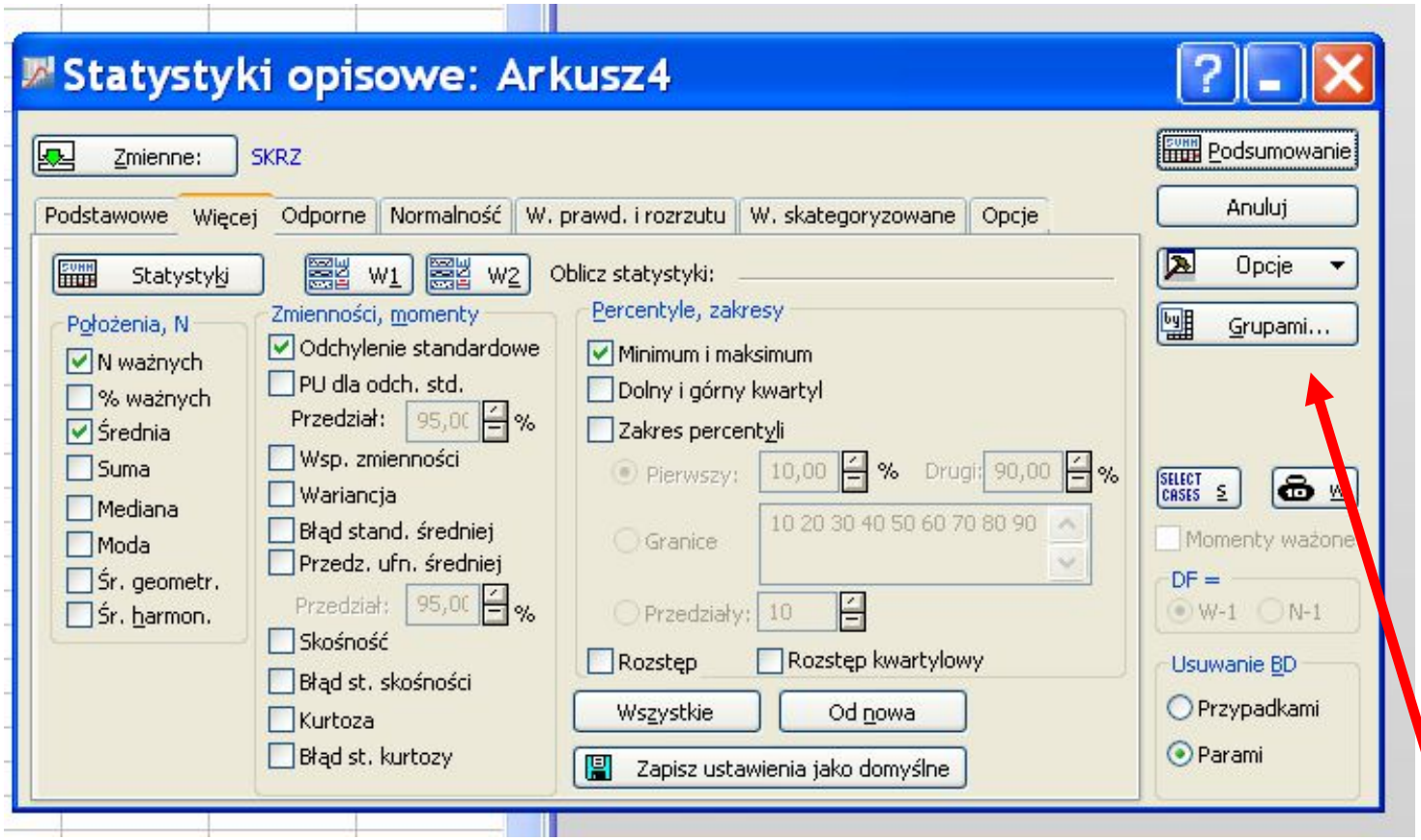


Obliczenia w pakiecie STATISTICA

W okienku **Statystyki opisowe** wybieramy przycisk **Zmienna** i zaznaczamy zmienną **SKRZ, przyciskamy OK.**

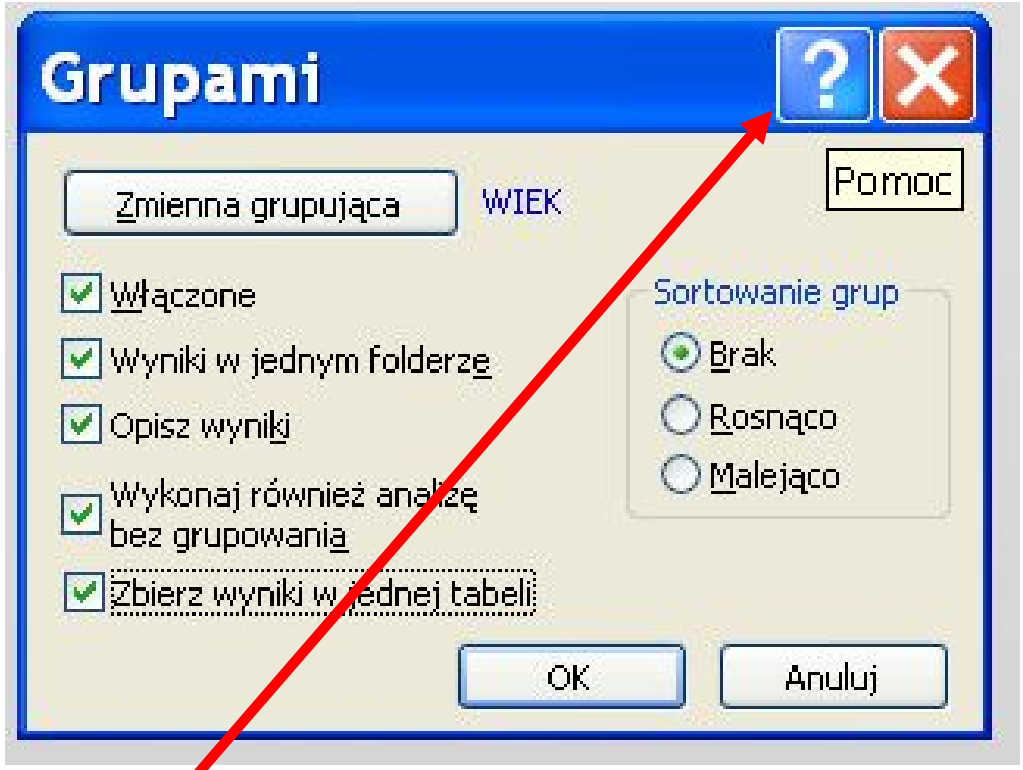


Obliczenia w pakiecie STATISTICA



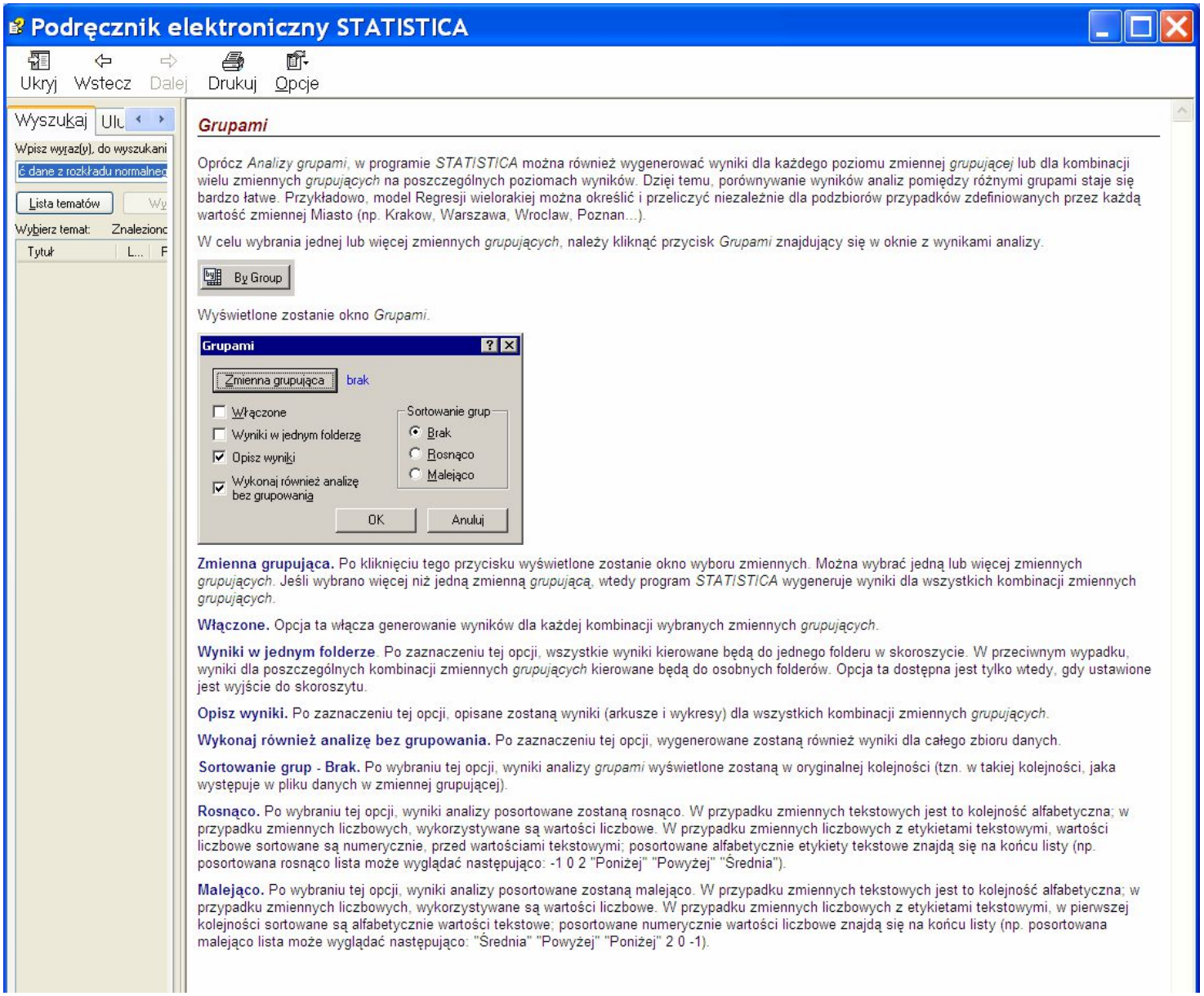
W oknie Statystyki opisowe, po prawej stronie przycisk **Grupami**... umożliwia zdefiniowanie podziału na grupy. Jako zmienną grupującą zaznaczamy WIEK. Zaznaczamy opcję **Wyniki w jednym folderze**.

Obliczenia w pakiecie STATISTICA



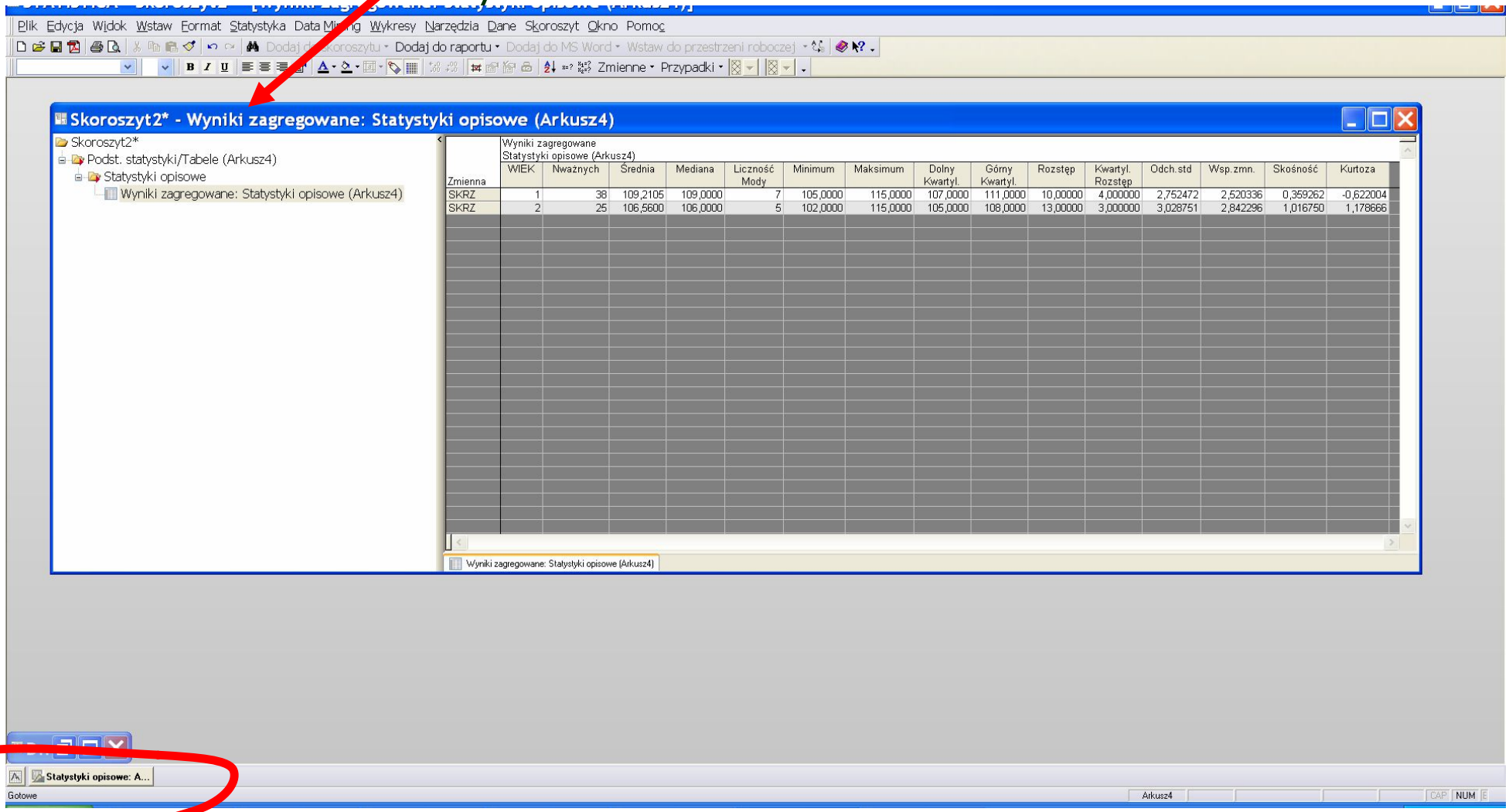
W razie niejasności, można kliknąć pytajnik, który otwiera pomoc ekranową.

Obliczenia w pakiecie STATISTICA



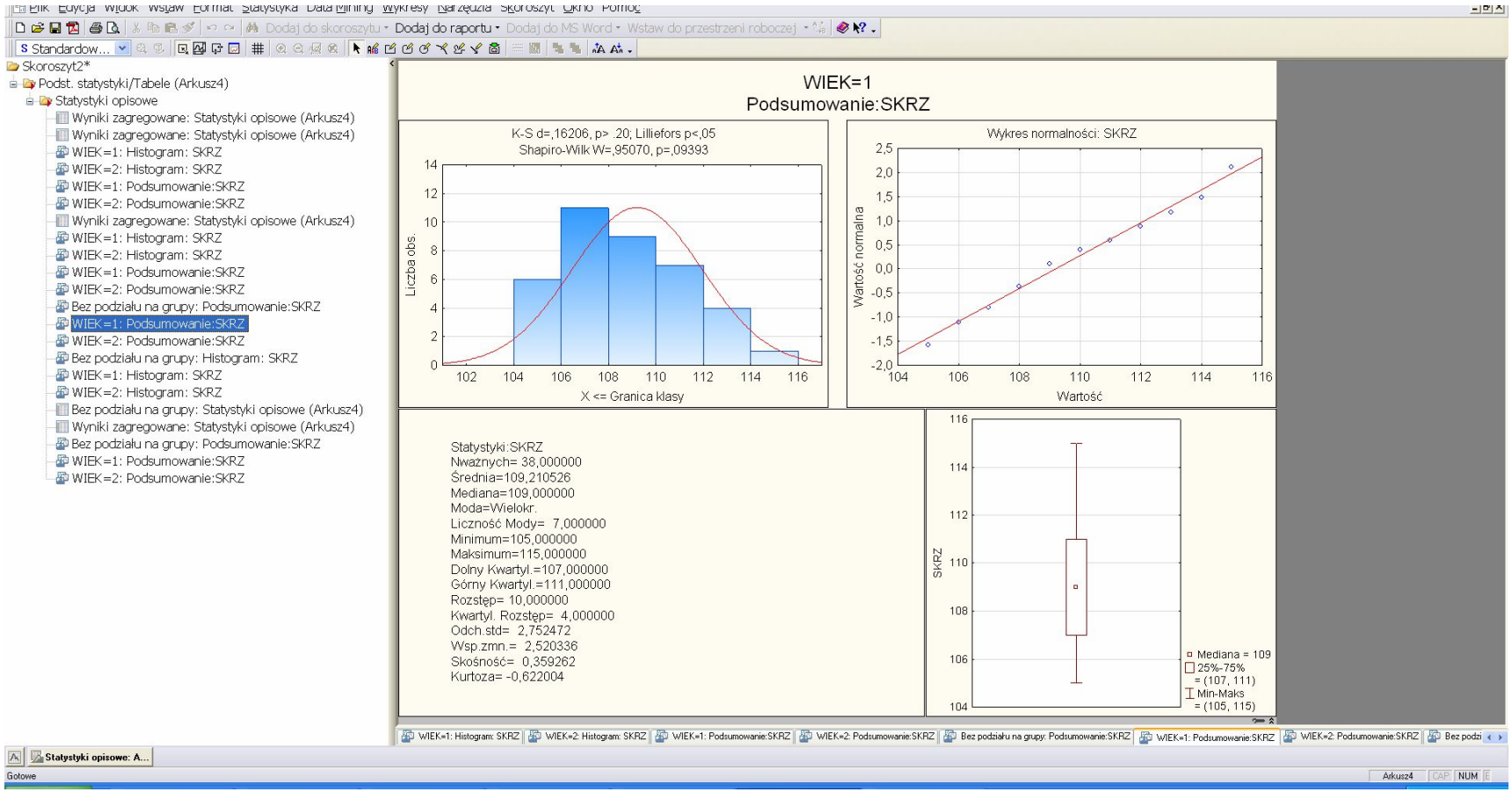
Obliczenia w pakiecie STATISTICA

W zakładce **Więcej** zaznaczamy parametry, które mają być obliczone. Po kliknięciu przycisku **Podsumowanie** wyniki są zestawione w skoroszycie.

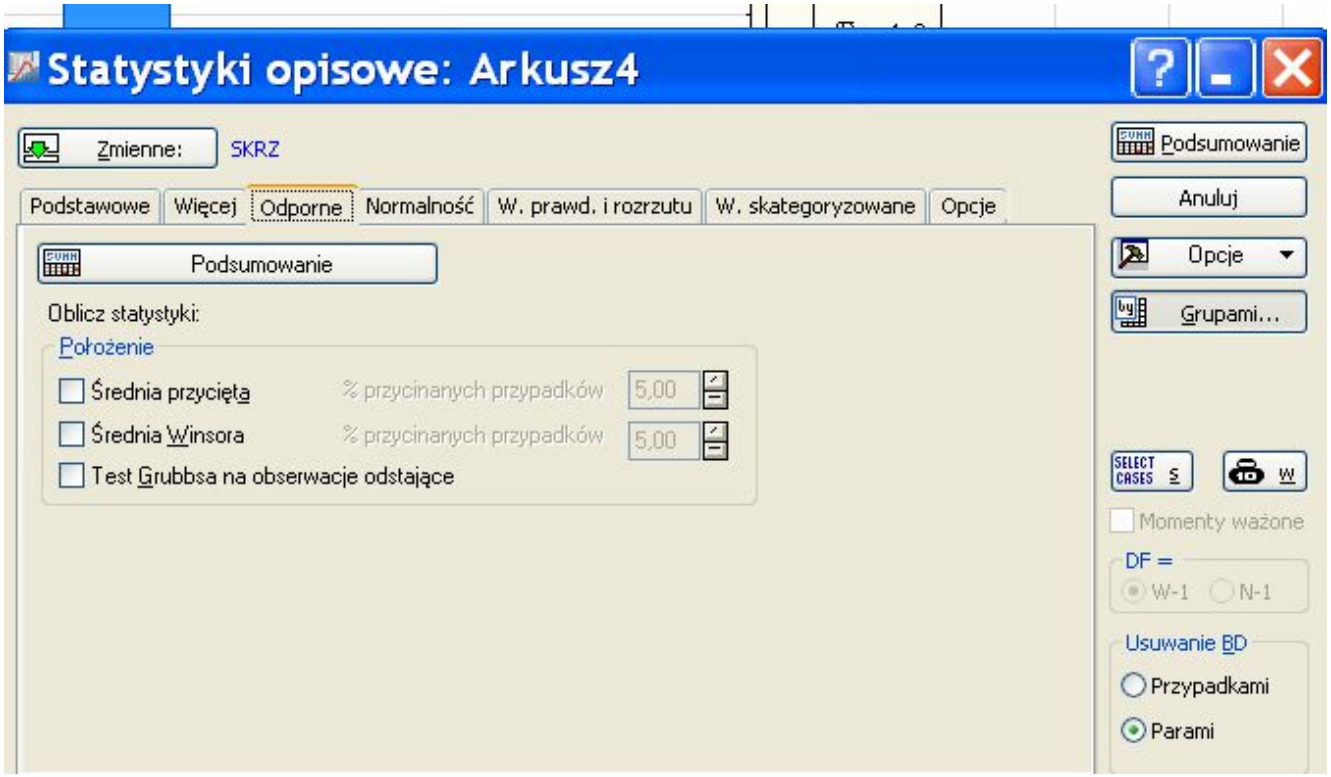


Obliczenia w pakiecie STATISTICA

Można obejrzeć histogramy rozkładu analizowanej zmiennej po kliknięciu na przycisk **Wykresy** na karcie **Więcej** lub **Histogramy** na karcie **Normalność**. W tym drugim przypadku możemy ustalić liczbę przedziałów.



Obliczenia w pakiecie STATISTICA



Obliczenia w pakiecie STATISTICA

Wykres ramka-wąsy, pudełko z wąsami (box-and-whisker plot)