Statystyka - przypomnienie pojęć na przykładach

Statystyka - przypomnienie pojęć na przykładach

1. **Szeregi statystyczne**

- **Szereg szczegółowy**
  Badanie ilości usterek, próbę stanowi 50 losowo wybranych sztuk towaru, wartości cechy wypisujemy jeden obok drugiego

$$\underbrace{0, 0, \ldots, 0}_{\text{30 razy}}, 1, 1, 1, 1, 1, 1, 1, 1, 2, 2, 2, 2, 2, 2, 3, 3, 3, 3, 4, 4$$

- **Szereg rozdzielczy**
  Do określenia rozkładu stosuje się tzw. **wskaźnik struktury** ω_i czyli stosunek liczby jednostek o danej wartości cechy do liczebności całej próby

$$\omega_i = \frac{n_i}{n}$$

  Na jego podstawie grupujemy wartości z poprzedniego przykładu

Numer	Liczba usterek	Liczba wyrobów	Wskaźnik struktury
i	x_i	n_i	ω_i
1	0	30	0,60
2	1	8	0,16
3	2	6	0,12
4	3	4	0,08
5	4	2	0,04

- **Szereg rozdzielczy skumulowany**
  Budujemy go poprzez przyporządkowanie kolejnym wariantom cechy odpowiadających im liczebności skumulowanej, stosujemy w tym celu **skumulowany wskaźnik struktury** ω_{isk}

$$\omega_{isk} = \frac{n_{isk}}{n}$$

Numer	Liczba usterek	Liczebność skumulowana	Częstość skumulowana
i	x_i	n_{isk}	ω_{isk}
1	0	30	0,60
2	0-1	38	0,76
3	0-2	44	0,88
4	0-3	48	0,96
5	0-4	50	1

  Istotnym pojęciem jest również **dystrybuanta empiryczna** czyli przyporządkowywanie kolejnym wartościom cechy odpowiadającym im częstościom skumulowanym

- **Szereg rozdzielczy z przedziałami klasowymi**
  W celu pokazania konstrukcji tego typu szeregu użyjemy spisu województw w Polsce z roku 1993

Lp.	Województwo	Ilość gmin	Powierzchnia ogólna w km^2
1	Warszawskie	54	3788
2	Bialskopodlaskie	40	5348
3	Białostockie	55	10055
4	Bielskie	59	3704
5	Bydgoskie	59	10349
6	Chełmskie	30	3866
7	Ciechanowskie	51	6362
8	Częstochowskie	58	6182
9	Elbląskie	42	6103
10	Gdańskie	63	7394
11	Gorzowskie	40	8484
12	Jeleniogórskie	40	4379
13	Kaliskie	58	6512
14	Katowiskie	91	6650
15	Kieleckie	78	9211
16	Konińskie	47	5139
17	Koszalińskie	41	8470
18	Krakowskie	39	3254
19	Krośnieńskie	43	5702
20	Legnickie	37	4037
21	Leszczyńskie	32	4154
22	Lubelskie	69	6792
23	Łomżyńskie	46	6684
24	Łódzkie	17	1523
25	Nowosądeckie	55	5576
26	Olsztyńskie	58	12327
27	Opolskie	65	8535
28	Ostrołęckie	44	6498
29	Pilskie	43	8205
30	Piotrowskie	56	6266
31	Płockie	48	5117
32	Poznańskie	62	8151
33	Przemyskie	41	4437
34	Radomskie	63	7294
35	Rzeszowskie	47	4397
36	Siedleckie	74	8499
37	Sieradzkie	42	4869
38	Skierniewickie	43	3960
39	Słupskie	37	7453

Lp.	Województwo	Ilość gmin	Powierzchnia ogólna w km^2
40	Suwalskie	48	10490
41	Szczecińskie	54	9982
42	Tarnobrzeskie	55	6283
43	Tarnowskie	47	4151
44	Toruńskie	49	5348
45	Wałbrzyskie	45	4168
46	Włocławskie	46	4402
47	Wrocławskie	40	6287
48	Zamojskie	57	6960
49	Zielonogórskie	57	8868

Rozpoczynamy konstrukcję szeregu rozdzielczego z przedziałami klasowymi według powierzchnii województw.

Krok I - ustalamy liczbę **klas** na podstawie wzorów

$$\boxed{k \approx \sqrt{n}} \qquad \boxed{k \approx 1 + 3,322 \log n}$$

Dla naszego szeregu jest to $k = \sqrt{49} = 7$

Krok II - obliczamy **rozstęp**

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

U nas jest to $R = 12327 - 1523 = 10804$

Krok III - określamy **rozpiętość przedziału klasowego** (przybliżamy z nadmiarem)

$$\boxed{h \approx \frac{x_{\max} - x_{\min}}{k} \approx \frac{R}{k}}$$

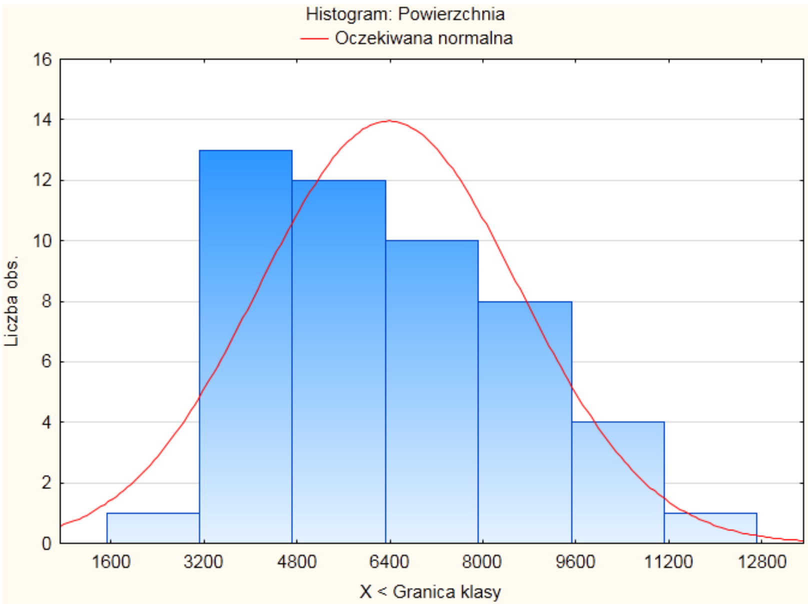
W przypadku powierzchni województw rozpiętość wyniesie $\frac{10804}{7} \approx 1600$ km

Krok IV - na podstawie wyliczonej ilości klas i rozpiętości budujemy **szereg**

Numer klasy	Powierzchnia w tys.km^2	Liczba województw	Wskaźnik struktury	Liczebność skumulowana	Częstość skumulowana
i	x_i	n_i	ω_i	n_{isk}	ω_{isk}
1	1,5-3,1	1	0,02	1	0,02
2	3,1-4,7	13	0,27	14	0,29
3	4,7-6,3	12	0,25	26	0,54
4	6,3-7,9	10	0,20	36	0,74
5	7,9-9,5	8	0,16	44	0,90
6	9,5-11,1	4	0,08	48	0,98
7	11,1-12,7	1	0,02	49	1

2. **Histogram**
W celu graficznego zaprezentowania danych konstruuje sie tak zwane histogramy, które składają się z wielu połączonych prostokątów. Każdy prostokąt symbolizuje jedną z klas, a jego wysokość oznacza liczebność względem badanej cechy.

Przykład:
Poniżej widzimy histogram obrazujący powierzchnię województw Polski zebraną w szeregu rozdzielczym powyżej.



3. **Klasyczne miary położenia**

- **Średnia arytmetyczna nieważona** - suma wartości cechy podzielona przez liczbę jednostek skończonej zbiorowości statycznej, wyliczajmy ją za pomocą wzoru

$$\boxed{\bar{x} = \frac{x_1 + x_2 + \cdots + x_n}{n} = \frac{1}{n}\sum_{i=1}^{n} x_i}$$

Przykład:
Mamy dwóch robotników (A i B) wykonujących te same produkty. Poniższe szeregi szczegółowe prezentują czasy wykonania pojedynczego produktu w minutach.
A: 12,15,15,18,20
B: 10,10,12,12,15,15,18,20,21,21
Wyliczmy średni czas produkcji jednej sztuki dla każdego robotnika:

$$\bar{x_A} = \frac{12 + 15 + 15 + 18 + 20}{5} = \frac{80}{5} = 16$$

$$\bar{x_B} = \frac{10 + 10 + 12 + 12 + 15 + 15 + 18 + 20 + 21 + 21}{10} = \frac{154}{10} = 15,4$$

Średnia arytmetyczna nieważona stosowana jest dla szeregów szczegółowych.

- **Średnia arytmetyczna ważona** - definicja jest analogiczna do średniej nieważonej z tym, że wyznaczamy ją dla szeregów rozdzielczych i rozdzielczych z przedziałami klasowymi, dana jest wzorem

$$\boxed{\bar{x} = \frac{1}{n}\sum_{i=1}^{k} x_i n_i} \qquad \boxed{\bar{x} = \frac{1}{n}\sum_{i=1}^{k} \dot{x}_i n_i}$$

gdzie k oznacza ilość klas, a $\dot{x}_i$ środek przedziału klasowego.

Przykład:
Weźmy szereg szczegółowy robotnika B z przykładu wyżej i zbudujemy na jego podstawie szereg rozdzielczy

Numer	Czas	Liczba sztuk
i	x_i	n_i
1	10	2
2	12	2
3	15	2
4	18	1
5	20	1
6	21	2

Wyliczmy średnią arytmetyczną ważoną

$$\bar{x_B} = \frac{10\cdot 2 + 12\cdot 2 + 15\cdot 2 + 18\cdot 1 + 20\cdot 1 + 21\cdot 2}{10} = \frac{154}{10} = 15,4$$

- **Średnia ważona** - średnia wyznaczana na podstawie średnich cząstkowych, dana wzorem

$$\boxed{\bar{\bar{x}} = \frac{1}{N}\sum_{i=1}^{r} \bar{x}_i n_i}$$

gdzie r oznacza ilość grup, n_i oznacza liczbność i-tej grupy, $\bar{x}_i$ oznacza średnią cząstkową w i-tej grupie, a N sumę liczebności we wszystkich grupach.

Przykład:
Wracamy do naszych robotników A i B, chcemy za pomocą ich średnich policzyć wspólną średnią czasu produkcji jednej sztuki towaru

$$\bar{\bar{x}} = \frac{16\cdot 5 + 15,4\cdot 10}{15} = 15,6$$

- **Średnia harmoniczna** - stosowana jest, gdy wartości cechy podane sa w przeliczeniu na stałą jednostkę innej zmiennej np. km/h, min/szt, zł/szt, dana jest następującymi wzorami

$$\boxed{\bar{x}_H = \frac{n}{\sum_{i=1}^{n}\frac{1}{x_i}} \text{ dla szeregu szczegółowego}}$$

$$\boxed{\bar{x}_H = \frac{\sum_{i=1}^{k} n_i}{\sum_{i=1}^{k}\frac{n_i}{x_i}} \text{ dla szeregu rozdzielczego}}$$

Przykład:
Mamy trójkę rzemieślników (A,B i C) pracujących po 8 godzin, Rzemieślnik A wykonuje jedną sztukę towaru w 4min, Rzemieślnik B w 6min, a C w 12min. Pytamy jaki jest średni czas produkcji jednej sztuki towaru.

$$\bar{x}_H = \frac{3}{\frac{1}{4} + \frac{1}{6} + \frac{1}{12}} = 6$$

- **Średnia geometryczna** - stosowana jest przy badaniu średniego tempa zmian zjawisk i dana wzorem

$$\boxed{\bar{x}_G = \sqrt[n]{x_1 \cdot x_2 \cdot \dots \cdot x_n} = \sqrt[n]{\prod_{i=1}^{n} x_i}}$$

Przykład:
W trzech okresach czasu mierzono zaludnienie miasta, wynosiło ono odpowiednio 5000, 7500, 8250. Mamy obliczyć średni przyrost względny ludności.

$$\bar{x}_G = \sqrt{\frac{7500}{5000} \cdot \frac{8250}{7500}} = \sqrt{1,5 \cdot 1,1} \approx 1,2845$$

4. **Pozycyjne miary położenia**

- **Modalna (dominanta)** - wartość cechy występująca najczęściej. W przypadku zarówno szeregów szczegółowych jak i rozdzielczych jej znalezienie nie jest trudne (w szeregach szczegółowych wystarczy znaleźć najczęściej powtarzająca się wartość, a w rozdzielczych jest to wartość dla której liczebność n_i jest największa). Z kolei w szeregach rozdzielczych z przedziałami klasowymi określić w tej sposób możemy tylko klasę, a nie dokładną wartość dominanty. Jej przybliżoną wartość wyznaczamy stosując następujący wzór:

$$\boxed{M_0 = x_{0m} + \frac{n_m - n_{m-1}}{(n_m - n_{m-1}) + (n_m - n_{m+1})} h_m}$$

gdzie m oznacza numer klasy zawierającej modalną, x_{0m} oznacza dolną granicę przedziału zawierającego modalną, n_m, n_{m-1}, n_{m+1} oznaczają liczebności klasy z modalną, poprzednią do niej klasę oraz następującą po niej, a h_m oznacza rozpiętość przedziału klasowego.

Przykład:
Wróćmy do szeregu z województwami. Wyczytujemy, że niego, że dominanta musi znajdowac się w drugim przedziale ($m = 2$). Teraz zgodnie z powyższym wzorem wyliczamy jej przybliżoną wartość.

$$M_0 = 3,1 + \frac{13 - 1}{(13 - 1) + (13 - 12)} \cdot 1,6 = 4,577 \text{ tys. km}^2$$

- **Kwantyle** - sa to wartości cechy dzielące zbiorowość na określone części pod wzaględem liczby jednostek. Części te są względem siebie w określonych proporcjach. Spośród kwantyli wyróżniamy **kwartyle** dzielące zbiorowość co 25% oraz **decyle** dzielące zbiorowość co 10%. Kwartyle dzielą się na

– **Kwartyl pierwszy** Q_1 – 25% zbiorowości ma wartości cechy mniejsze, bądź równe kwartylowi pierwszemu, a pozostałe 75% większe od kwartyla pierwszego, w celu jego wyliczenia musimy na początku wyznaczyć jego pozycję za pomocą wzoru

$$N_{Q_1} = \frac{n}{4}$$

a następnie wstawić odpowiednie dane do wzoru

$$\boxed{Q_1 = x_{0m} + \frac{N_{Q_1} - \sum_{i=1}^{m-1} n_i}{n_m} \cdot h_m}$$

– **Kwartyl drugi (mediana)** Me – 50% zbiorowości ma wartości cechy mniejsze, bądź równe kwartylowi pierwszemu, a pozostałe 50% większe od kwartyla pierwszego (czyli następuje podzielenie zbiorowości na połowy), medianę liczymy następująco dla szeregu szczegółowego

$$\boxed{Me = \begin{cases} x_{\frac{n+1}{2}} & \text{gdy } n \text{ jest nieparzyste} \\ \frac{1}{2}(x_{\frac{n}{2}} + x_{\frac{n}{2}+1}) & \text{gdy } n \text{ jest parzyste} \end{cases}}$$

Dla szeregu rozdzielczego działamy podobnie jak dla kwartyla pierwszego, na początku wyznaczamy pozycję mediany

$$N_{Me} = \frac{n}{2}$$

a później wstawiamy określone dane do wzoru

$$\boxed{Me = x_{0m} + \frac{N_{Me} - \sum_{i=1}^{m-1} n_i}{n_m} \cdot h_m}$$

– **Kwartyl trzeci** Q_3–75% zbiorowości ma wartości cechy mniejsze, bądź równe kwartylowi trzeciemu, a pozostałe 25% większe od kwartyla trzeciego, w celu jego wyliczenia musimy na początku wyznaczyć jego pozycję za pomocą wzoru

$$N_{Q_3} = \frac{3n}{4}$$

a następnie wstawić odpowiednie dane do wzoru

$$\boxed{Q_3 = x_{0m} + \frac{N_{Q_3} - \sum_{i=1}^{m-1} n_i}{n_m} \cdot h_m}$$

Oznaczenia we wszystkich wzorach na kwartyle są identyczne jak oznaczenia dla dominanty.

Przykład:
Wyliczmy wszystkie trzy kwartyle dla szeregu z województwami. Na początku wyliczamy ich pozycje

$$N_{Q_1} = \frac{49}{4} = 12,15 \qquad N_{Me} = \frac{49}{2} = 24,5 \qquad N_{Q_3} = \frac{147}{4} = 36,75$$

Teraz patrzymy w jakich klasach one leżą: pierwszy w $m = 2$, drugi (mediana) w $m = 3$ i trzeci $m = 5$. Mając te dane możemy przejść do dokładnego wyliczenia:

$$Q_1 = 3,1 + \frac{12,25 - 1}{13} \cdot 1,6 = 4,485 \text{ tys.km}^2$$

$$Me = 4,7 + \frac{24,5 - 14}{12} \cdot 1,6 = 6,1 \text{ tys.km}^2$$

$$Q_3 = 7,9 + \frac{36,75 - 36}{8} \cdot 1,6 = 8,05 \text{ tys.km}^2$$

5. **Klasyczne miary zmienności**

- **Wariancja** - średnia arytmetyczna kwadratów odchyleń wartości cechy od średniej arytmetycznej zbiorowości, dana wzorami

$$\boxed{s^2 = \frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})^2 \text{ dla szeregów szczegółowych}}$$

$$\boxed{s^2 = \frac{1}{n}\sum_{i=1}^{k}(x_i - \bar{x})^2 n_i \text{ dla szeregów rozdzielczych}}$$

$$\boxed{s^2 = \frac{1}{n}\sum_{i=1}^{k}(\dot{x}_i - \bar{x})^2 n_i \text{ dla szeregów rozdzielczych z przedziałami klasowymi}}$$

Istotnym parametrem jest również **odchylenie standardowe** s czyli pierwiastek z wariancji określający przeciętne zróżnicowanie piszczególnych wartości cechy od średniej arytmetycznej, dzięki niemu wyznaczyć możemy również **typowy obszar zmienności** dany wzorem

$$\bar{x} - s < x_{\text{typ}} < \bar{x} + s$$

w którym mieści się około $\frac{2}{3}$ wszystkich jednostek badanej zbiorowości.

Z pojęciem wariancji związana jest tak zwana **równość wariancyjna**, za pomocą której wyliczać możemy wariancję zbiorowości podzielonej na grupy. Wariancja ogólna jest sumą **wariancji wewnątrzgrupowej** czyli średniej arytmetycznej wewnątrzgrupowych wariancji wartości cechy oraz **wariancji międzygrupowej** stanowiącej wariancję średnich grupowych wartości tej zmiennej. Ogólnie równość wariancyjna wyraża się wzorem

$$\boxed{s^2 = \overline{s_i^2} + s^2(\bar{x}_i)}$$

gdzie

$$\boxed{\overline{s_i^2} = \frac{1}{N}\sum_{i=1}^{r} s_i^2 \cdot n_i \qquad s^2(x_i) = \frac{1}{N}\sum_{i=1}^{r}(\bar{x}_i - \bar{\bar{x}})^2 \cdot n_i}$$

Oznaczenia sa zgodne ze stosowanymi we wzorach powyżej.

Przykład:

Weźmy przykład ze strony 4 dotyczący robotników A i B, wyliczmy dla nich wariancję, odchylenie standardowe oraz zastosujmy równość wariancyjną.

$$s_A^2 = \frac{(12-16)^2 + (15-16)^2 + (15-16)^2 + (18-16)^2 + (20-16)^2}{5} = \frac{38}{5} = 7,6 \text{ min}$$

$$s_B^2 = 17,24 \text{ min} \ \ \text{(liczymy analogicznie jak wyżej)}$$

$$s_A = \sqrt{7,6} = 2,76 \text{ min}$$

$$s_B = \sqrt{17,24} = 4,1 \text{ min}$$

Równość wariancyjna:

$$\overline{s_i^2} = \frac{1}{15}(7,6 \cdot 5 + 17,24 \cdot 10) = 14,03$$

$$s^2(\bar{x}_i) = \frac{1}{15}((16-15,6)^2 \cdot 5 + (15,4-15,6)^2 \cdot 10) = 0,08$$

$$\overline{s_i^2} + s^2(\bar{x}_i) = 14,03 + 0,08 = 14,11$$

- **Odchylenie przeciętne** - średnia arytmetyczna bezwzględnych odchyleń wartości cechy od średniej arytmetycznej, dana za pomocą poniższych wzorów dla szeregu szczegółowego i rozdzielczego:

$$\boxed{d = \frac{1}{n}\sum_{i=1}^{n}|x_i - \bar{x}| \qquad\qquad d = \frac{1}{n}\sum_{i=1}^{k}|x_i - \bar{x}| \cdot n_i}$$

Przykład:
Wyliczmy odchylenie przeciętne dla robotnika A z poprzedniego zadania

$$d_A = \frac{|12-16| + |15-16| + |15-16| + |18-16| + |20-16|}{5} = \frac{12}{5} = 2,4 \text{ min}$$

6. **Pozycyjne miary zmienności**

- **Odchylenie ćwiartkowe** - połowa różnicy między trzecim, a pierwszym kwartylem

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

Również za pomocą tego parametru możemy wyliczać **typowy obszar zmienności**

$$\boxed{Me - Q < x_{\text{typ}} < Me + Q}$$

Przykład:
Wracamy do szeregu z województwami i wyliczonych dla niego kwartyli, wstawiamy odpowiednie dane do wzoru.

$$Q = \frac{8,05 - 4,485}{2} = 1,78 \text{ tys.km}^2$$

Zatem typowy obszar zmienności jest w zakresie od $4,32$ do $7,88$ tys.km^2.

- **Współczynniki zmienności** - ilorazy bezwględnej miary zmienności cechy i średniej jej wartości. Wyróżniamy klasyczne współczynniki zmienności

$$\boxed{V_s = \frac{s}{\bar{x}} \qquad\qquad V_d = \frac{d}{\bar{x}}}$$

oraz współczynniki pozycyjne

$$\boxed{V_Q = \frac{Q}{Me} \qquad\qquad V_{Q_1,Q_3} = \frac{Q_3 - Q_1}{Q_3 + Q_1}}$$

Przykład:
Wyliczmy podane powyżej współczynniki dla szeregu z województwami (obliczenia średniej, wariancji i odchylenia pomijamy).

$$\bar{x} = 6,38 \qquad s = 2,24 \qquad d = 1,91 \qquad V_s = \frac{2,24}{6,38} = 0,35 \qquad V_d = \frac{1,91}{6,38} = 0,30$$

$$Q = 1,78 \qquad Me = 6,1 \qquad Q_1 = 4,485 \qquad Q_3 = 8,05 \qquad V_Q = \frac{1,78}{6,1} = 0,29$$

$$V_{Q_1,Q_3} = \frac{8,05 - 4,485}{8,05 + 4,485} = \frac{3,565}{12,535} = 0,28$$

7. **Miary asymetrii**

- **Symetria i asymetria** - służy do opisywania położenia miar przeciętnych w szeregach, wyróżniamy:
  - **szeregi symetryczne**, gdy $\bar{x} = Me = M_0$
  - **asymetrię prawostronną**, gdy $\bar{x} > Me > M_0$
  - **asymetrię lewostronną**, gdy $\bar{x} < Me < M_0$
- **Wskaźnik skośności** - wyliczany jako różnica średniej i dominanty $\bar{x} - M_0$ lub za pomocą miar pozycyjnych, wtedy rodzaj asymetrii określamy następująco:
  - w rozkładzie symetrycznym mamy zależność $(Q_3 - Me) - (Me - Q_1) = 0$
  - przy asymetrii prawostronnej mamy zależność $(Q_3 - Me) - (Me - Q_1) > 0$
  - przy asymetrii lewostronnej mamy zależność $(Q_3 - Me) - (Me - Q_1) = 0$
- **Współczynniki skośności** - opisują kierunek i siłę asymetrii, dane są następująco:

$$\boxed{A_s = \frac{\bar{x} - M_0}{s} \qquad\qquad A_d = \frac{\bar{x} - M_0}{d} \qquad\qquad A_Q = \frac{Q_3 + Q_1 - 2Me}{2Q}}$$

Ponadto stosuje się (ale rzadziej) współczynnik asymetrii A dany wzorem

$$\boxed{A = \frac{m_3}{s^3} \text{ gdzie } m_3 = \begin{cases} \frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})^3 & \text{dla szeregu szczegółowego} \\ \frac{1}{n}\sum_{i=1}^{k}(x_i - \bar{x})^3 \cdot n_i & \text{dla szeregu rozdzielczego} \end{cases}}$$

Przykład:

Wyznaczmy wskaźnik skośności oraz współczynniki skośności A_s, A_d oraz A_Q dla szeregu z województwami.

$$Q = 1,78 \qquad Me = 6,1 \qquad Q_1 = 4,485 \qquad Q_3 = 8,05$$

Wskaźnik skośności: $(8,05-6,1)-(6,1-4,485) = 1,95-1,615 = 0,335 > 0-$ asymetria prawostronna

$$\bar{x} = 6,38 \qquad s = 2,24 \qquad d = 1,91 \qquad M_0 = 4,577$$

$$A_s = \frac{6,38 - 4,577}{2,24} = \frac{1,803}{2,24} = 0,80$$

$$A_d = \frac{6,38 - 4,577}{1,91} = \frac{1,803}{1,91} = 0,94$$

$$A_Q = \frac{8,05 + 4,485 - 2 \cdot 6,1}{2 \cdot 1,78} = \frac{0,335}{3,56} = 0,09$$

8. **Miary koncentracji**

- **Kurtoza** - opisuje koncentrację wartości cechy wokół średniej, dana jest wzorem

$$\boxed{K = \frac{m_4}{s^4} \text{ gdzie } m_4 = \begin{cases} \frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})^4 & \text{dla szeregu szczegółowego} \\ \frac{1}{n}\sum_{i=1}^{k}(x_i - \bar{x})^4 \cdot n_i & \text{dla szeregu rozdzielczego} \end{cases}}$$

Im wyższa jest wartość współczynnika K, tym bardziej wysmukła jest krzywa liczbności. Jeśli zbiorowość ma rozkład normalny to $K = 3$, jeśli $K < 3$ to rozkład jest bardziej spłaszczony od normalnego, a jesy $K > 3$ to krzywa jest jeszcze smuklejsza niż w rozkładzie normalnym. Z tego też powodu często wylicza się współczynnik K' dany następująco

$$\boxed{K' = \frac{m_4}{s^2} - 3}$$

Przykład:
Wyliczymy kurtozę dla szeregu z województwami, obliczenie m_4 pomijamy, skorzystamy z gotowego wyniku

$$K = \frac{58,846}{2,24^4} = \frac{58,846}{25,18} = 2,337 \Longrightarrow K' = -0,663$$

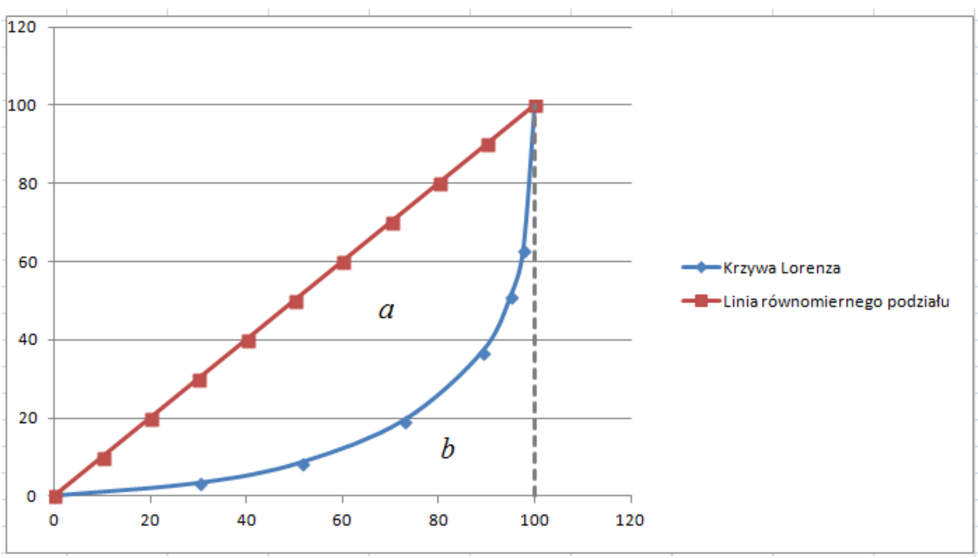
co informuje nas, że rozkład województw jest bardziej spłaszczony niż krzywa normalna.

- **Krzywa Lorenza**
Zjawisko koncentracji możemy rozważać również jako nierównomierny podział ogólnej wartości cech między poszczególne jednostki zbiorowości. W tym celu wyliczamy **współczynnik koncentracji Lorenza** i konstruujemy **krzywą Lorenza**. Mechanizm ten najlepiej zoobrazuje poniższy przykład

Przykład:
W tabeli zebrano dane dotyczące liczby miast i mieszkających w nich ludzi.

Grupy miast	Liczba	Ludność	Ods. miast	Ods. ludności	Skum.odsetki		Pole
w tys.		w tys.	w %	w%	miast	ludn.	
x_i	n_i		ω_i	z_i	ω_{isk}	z_{isk}	$\frac{z_{isk}+z_{i-1sk}}{2}\cdot\omega_i$
do 5	253	788,1	30,30	3,33	30,30	3,33	50,38
5-10	176	1239,2	21,08	5,23	51,38	8,55	125,19
10-20	178	2543,7	21,32	10,736	72,69	19,29	296,74
20-50	136	4140,3	16,29	17,43	88,98	36,76	458,39
50-100	50	3389,9	5,99	14,30	94,97	51,06	262,91
100-200	22	2849,1	2,63	12,02	97,60	63,08	150,36
200+	20	8750,5	2,40	36,92	100	100	195,30
Razem	835	23700,8	100	100			1537,27

W tabeli łatwo zauważyć, że w polskich miast ludność nie jest rozmieszczona równomiernie. Przykładowo w ok. 95% miast mieszka tylko nieco ponad połowa ludności. W przypadku, gdyby we wszystkich miastach mieszkała identyczna ilość osób mówilibyśmy o braku koncentracji, a gdyby jedno miasto skupiało całą ludność to wtedy o całkowitej koncentracji. Do opisu stopnia koncentracji używa się krzywej Lorenza. Aby ją skonstruoować musimy na osi odciętych zaznaczyć skumulowany odsetek miast ω_{isk}, a na osi rzędnych skumulowany odsetek mieszkańców z_{isk}. Następnie zaznaczamy na niej punkty o współrzędnych (ω_{isk}, z_{isk}).



W przypadku równomiernego podziału wszystkie punkty leżały na przekątnej kwadratu o boku 100, przekątna ta nosi nazwę **linii równomiernego podziału**. Im bardziej nasza krzywa oddala się od tej linii to wzrasta pole figury oznaczonej na rysunku jako a. Pole to służy nam do wyznaczania **współczynnika koncentracji Lorenza** danego wzorem

$$\boxed{K_L = \frac{a}{5000} = \frac{5000 - b}{5000}}$$

Dla całkowitej koncentracji wartości K_L wyniesie 1, a dla równomiernego podziału 0. W praktyce zamiast liczyć pole a liczymy wartość $5000 - b$, gdzie b jest sumą pól trapezów i trójkątów pod krzywą. Wtedy wzór ma postać

$$\boxed{K_L \approx 1 - \frac{1}{5000} \cdot \sum_{i=1}^{k} \frac{z_{isk} + z_{i-1sk}}{2} \cdot \omega_i}$$

Dla miast i ludności współczynnik ten wyliczamy następująco

$$K_L = 1 - \frac{1537,27}{5000} = 0,692$$

co daje nam informacje, że występuje dość duża koncentracja ludności w miastach.

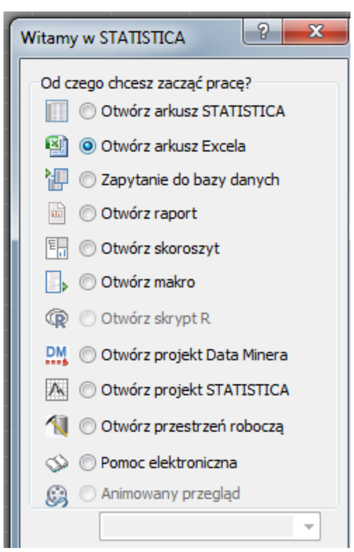
Tutorial - Statystyka opisowa w programie Statistica 12

- **Wczytywanie tabeli z Excela**

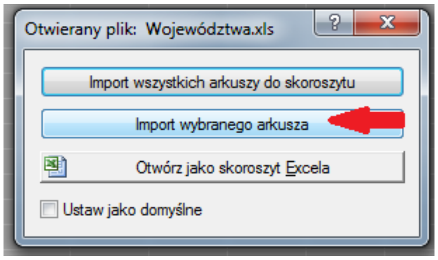
  Statistica 12 umożliwia wczytywanie plików z programu Excel, ale tylko tych z rozszerzeniem `.xls`, rozszerzenie `.xlsx` nie jest obsługiwane. Aby jak najlepiej zaimportować dane należy w Excelu posiadać tabele, w której kolumny posiadają nagłówki. Dzięki temu nazwy w nagłówkach w Statistice automatycznie staną się nazwami zmiennych. Poniżej prezentujemy taką oto tabelę.

	A	B	C	D
1	Lp.	Województwo	Liczba Gmin	Powierzchnia
2	1	Warszawskie	54	3788
3	2	Bialskopodlaskie	40	5348
4	3	Białostockie	55	10055
5	4	Bielskie	59	3704
6	5	Bydgoskie	59	10349
7	6	Chełmskie	30	3866
8	7	Ciechanowskie	51	6362
9	8	Czestochowskie	58	6182

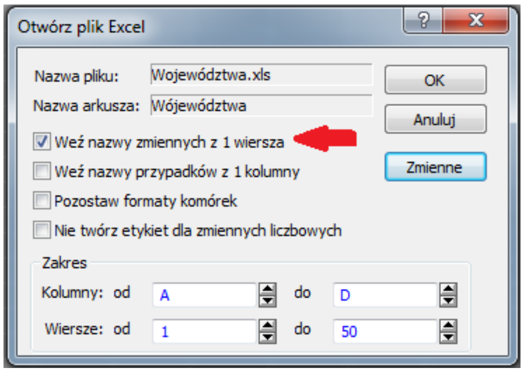
  Przejdźmy do importu tabeli. Włączamy Statisticę i zaznaczamy opcję **Otwórz arkusz Excela**.



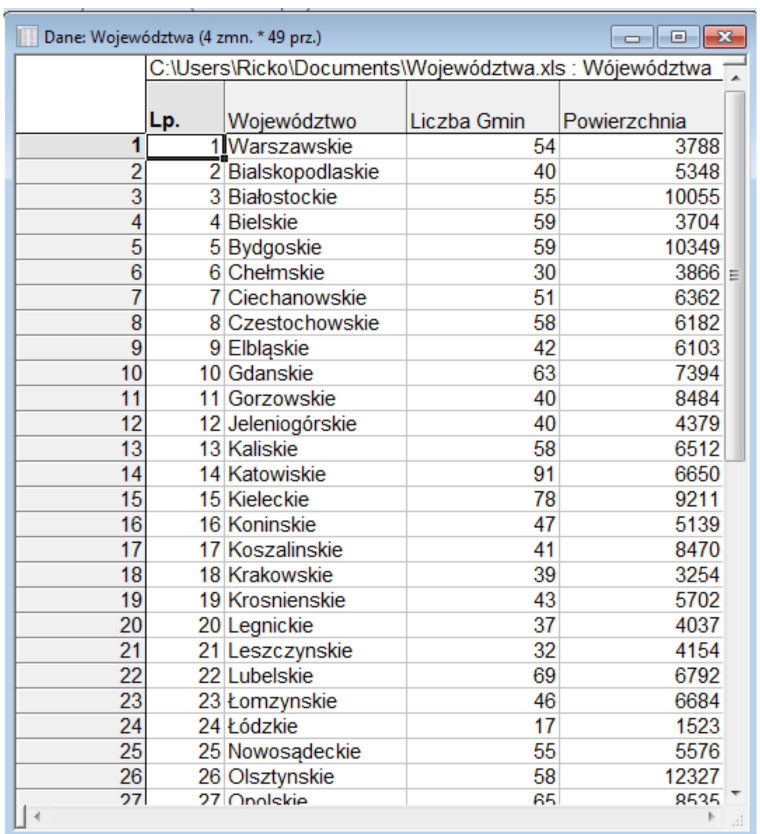
  Wybieramy plik. Następnie mamy możliwość pobrania jednego, wybranego arkusza lub wszystkich. My wybierzemy jeden.



Po wyborze arkusza zaznaczamy opcję **Weź nazwy zmiennych z 1 wiersza**



Klikamy **Ok** i kończymy import.

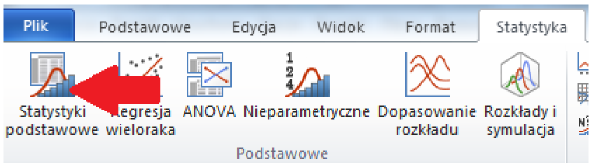

Dane: Województwa (4 zmn. * 49 prz.)

C:\Users\Ricko\Documents\Województwa.xls : Wójewództwa

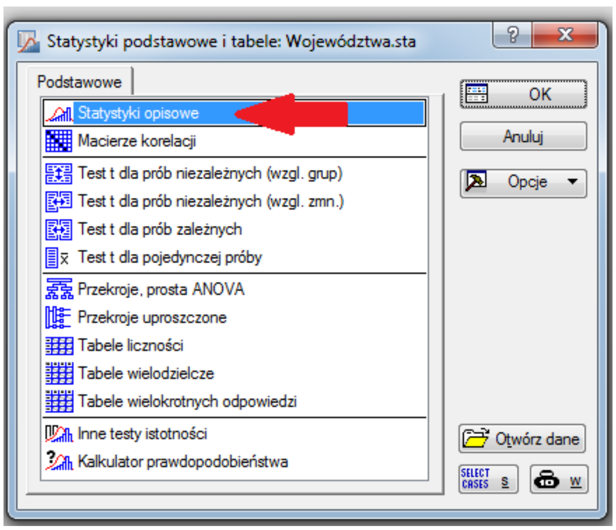
	Lp.	Województwo	Liczba Gmin	Powierzchnia
1	1	Warszawskie	54	3788
2	2	Bialskopodlaskie	40	5348
3	3	Białostockie	55	10055
4	4	Bielskie	59	3704
5	5	Bydgoskie	59	10349
6	6	Chełmskie	30	3866
7	7	Ciechanowskie	51	6362
8	8	Czestochowskie	58	6182
9	9	Elbląskie	42	6103
10	10	Gdanskie	63	7394
11	11	Gorzowskie	40	8484
12	12	Jeleniogórskie	40	4379
13	13	Kaliskie	58	6512
14	14	Katowiskie	91	6650
15	15	Kieleckie	78	9211
16	16	Koninskie	47	5139
17	17	Koszalinskie	41	8470
18	18	Krakowskie	39	3254
19	19	Krosnienskie	43	5702
20	20	Legnickie	37	4037
21	21	Leszczynskie	32	4154
22	22	Lubelskie	69	6792
23	23	Łomzynskie	46	6684
24	24	Łódzkie	17	1523
25	25	Nowosądeckie	55	5576
26	26	Olsztynskie	58	12327
27	27	Opolskie	65	8535

- **Szereg rozdzielczy z przedziałami klasowymi**
  Przedstawimy dwa sposoby konstrukcji szeregu rozdzielczego: podstawowy i zaawansowany z możliwością modyfikacji sposobu wyświetlania szeregu.

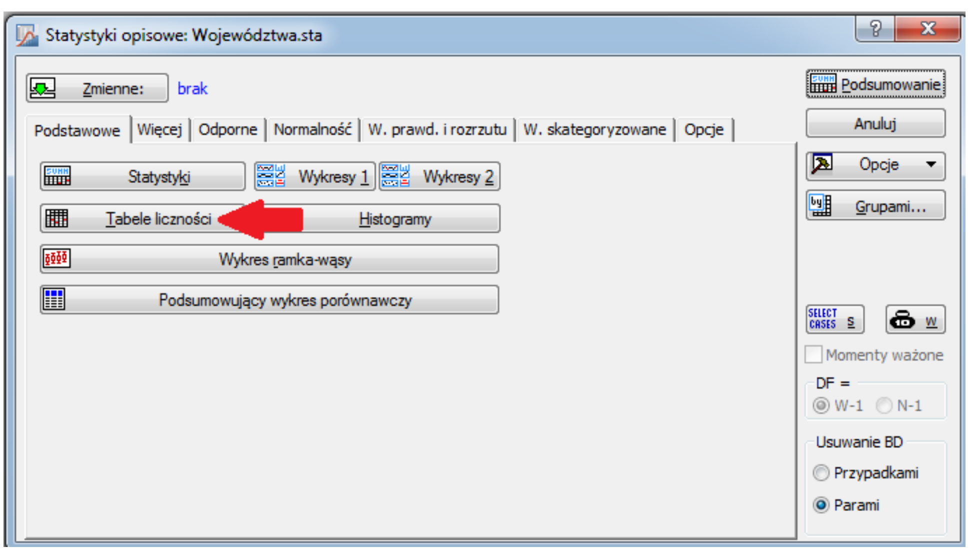
  1. Klikamy na menu **Statystyka** i wybieramy opcję najbardziej po lewej **Statystyki podstawowe**



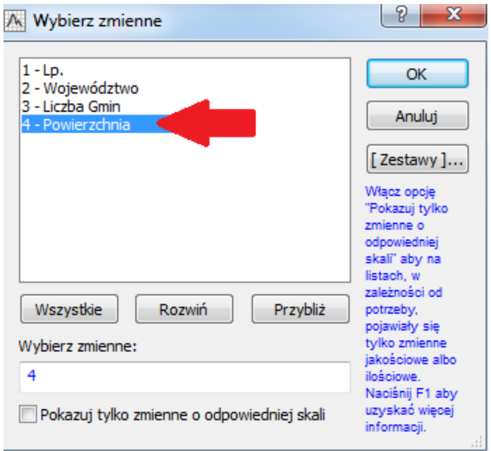
Następnie wybieramy **Statystyki opisowe**



Wybieramy **Tabele liczności**



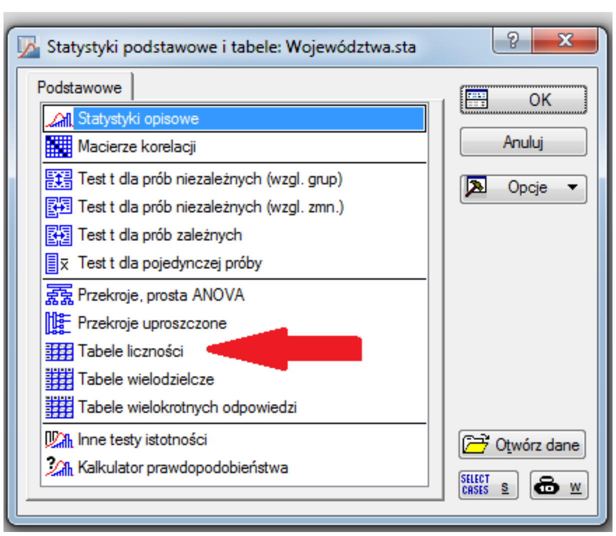
Zaznaczamy interesujące nas zmienne



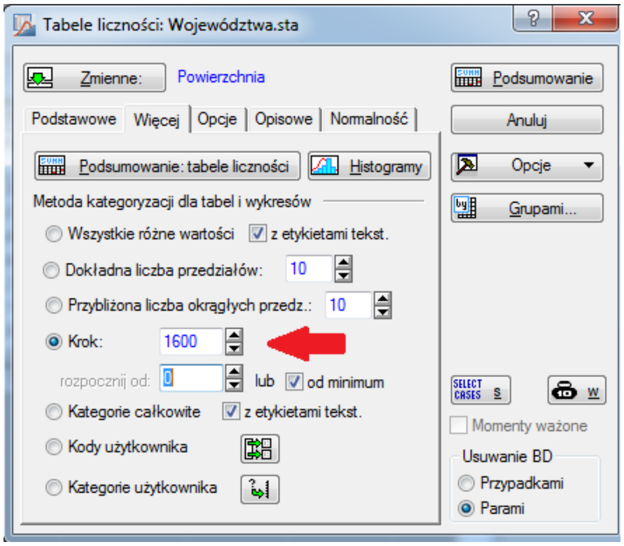
i klikamy **Ok**.

Klasa	Tabela liczności: Powierzchnia (Województwa) K-S d=,09291, p> .20; Lilliefors p> .20 Liczba	Skumulow. Liczba	Procent Ważnych	Skumul. % Ważnych	% ogółu Przypadki	Skumulow. % Ogółu
0,000000<x<=2000,000	1	1	2,04082	2,0408	2,04082	2,0408
2000,000<x<=4000,000	5	6	10,20408	12,2449	10,20408	12,2449
4000,000<x<=6000,000	15	21	30,61224	42,8571	30,61224	42,8571
6000,000<x<=8000,000	15	36	30,61224	73,4694	30,61224	73,4694
8000,000<x<=10000,00	9	45	18,36735	91,8367	18,36735	91,8367
10000,00<x<=12000,00	3	48	6,12245	97,9592	6,12245	97,9592
12000,00<x<=14000,00	1	49	2,04082	100,0000	2,04082	100,0000
Braki	0	49	0,00000		0,00000	100,0000

2. Podobnie jak wcześnie wybieramy **Statystyki podstawowe**, lecz tym razem klikamy na **Tabele liczności**



Wybieramy kartę **Więcej** i zaznaczamy kropkę obok opcji **Krok**, zmieniamy wartość kroku na długość przedziału klasowego, następnie klikamy **Podsumowanie**, zaznaczamy podobnie jak w poprzednim sposobie zmienną i klikamy **Ok**

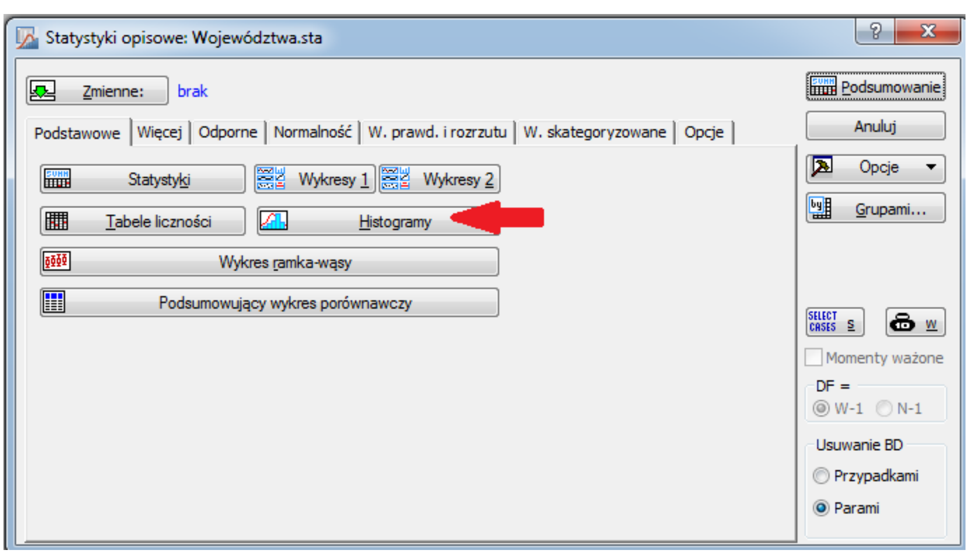


Od	Do	Liczba	Skumulow. Liczba	Procent	Skumulow. Procent
		Tabela liczności: Powierzchnia (Województwa)			
1523	<=x<3123	1	1	2,04082	2,0408
3123	<=x<4723	13	14	26,53061	28,5714
4723	<=x<6323	12	26	24,48980	53,0612
6323	<=x<7923	10	36	20,40816	73,4694
7923	<=x<9523	8	44	16,32653	89,7959
9523	<=x<11123	4	48	8,16327	97,9592
11123	<=x<12723	1	49	2,04082	100,0000
12723	<=x<14323	0	49	0,00000	100,0000
Braki		0	49	0,00000	100,0000

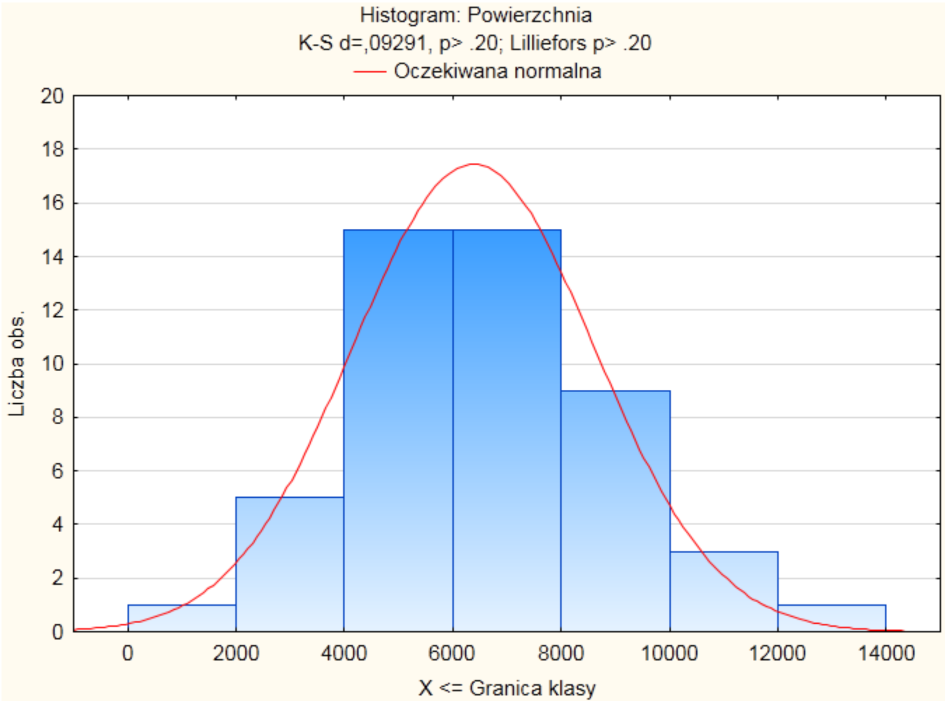
Możemy zamiast opcji **Krok** wybrac inne sąsiadujące opcje i zobaczyć w jaki sposób zmieni się nasz szereg.

- **Histogram**
  Podobnie jak z szeregami histogram budować możemy na dwa sposoby: podstawowy i zaawansowany z możliwością dostosowania przedziałów.

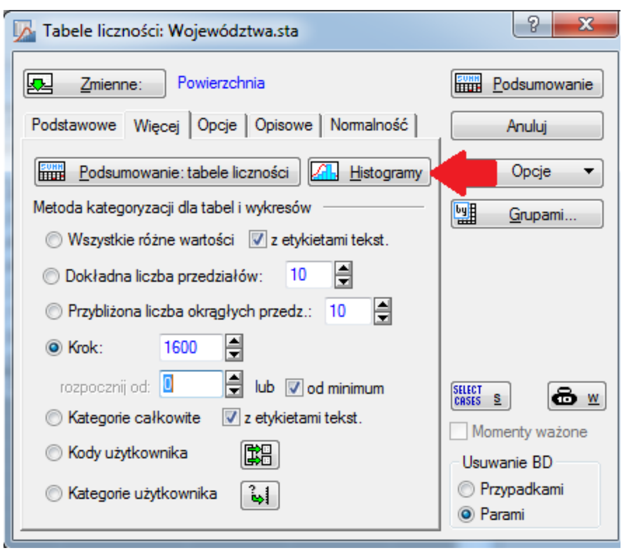
  1. Klikamy na menu **Statystyka** i wybieramy opcję **Statystyki podstawowe**. Następnie wybieramy **Statystyki opisowe** i klikamy przycisk **Histogramy**

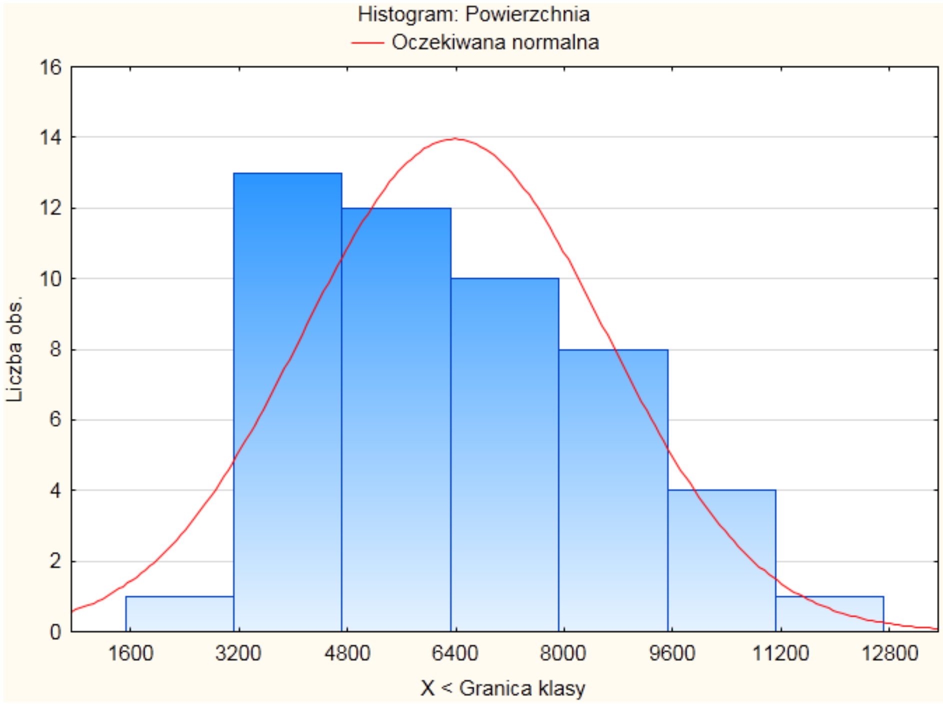


Teraz pozostaje nam wybrać zmienną i histogram gotowy.



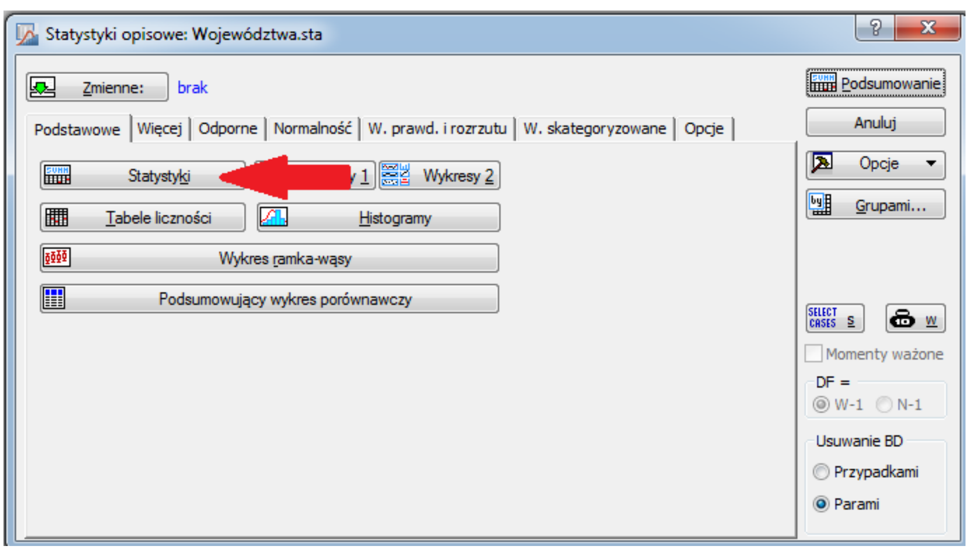
2. Klikamy na menu **Statystyka** i wybieramy opcję **Statystyki podstawowe**. Następnie wybieramy **Tabele liczności**, w niej kartę **Więcej** i zaznaczamy kropkę obok opcji **Krok** tak jak w przypadku szeregu, zmieniamy wartość kroku na długość przedziału klasowego, następnie klikamy **Histogramy**, zaznaczamy podobnie jak w poprzednim sposobie zmienną i klikamy **Ok**.





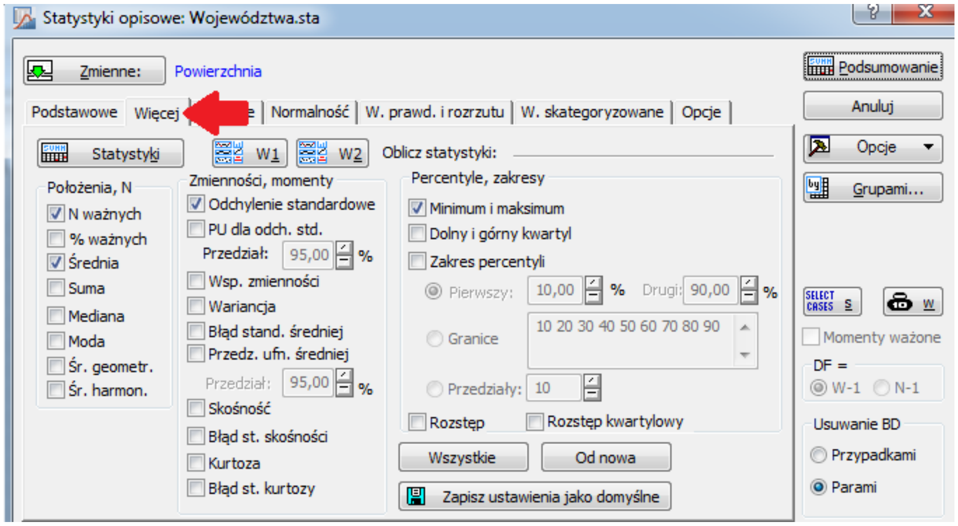
- **Średnia, moda, mediana itp.**
  Pierwsza, prostsza metoda pozwala wyliczyć takie parametry jak **ilość pomiarów, średnia, min, max, odchylenie standardowe**, druga, bardziej zaawansowana daje nam jednak większy wachlarz opcji do wyboru

  1. Klikamy na menu **Statystyka** i wybieramy opcję **Statystyki podstawowe**. Następnie wybieramy **Statystyki opisowe** i klikamy przycisk **Statystyki**

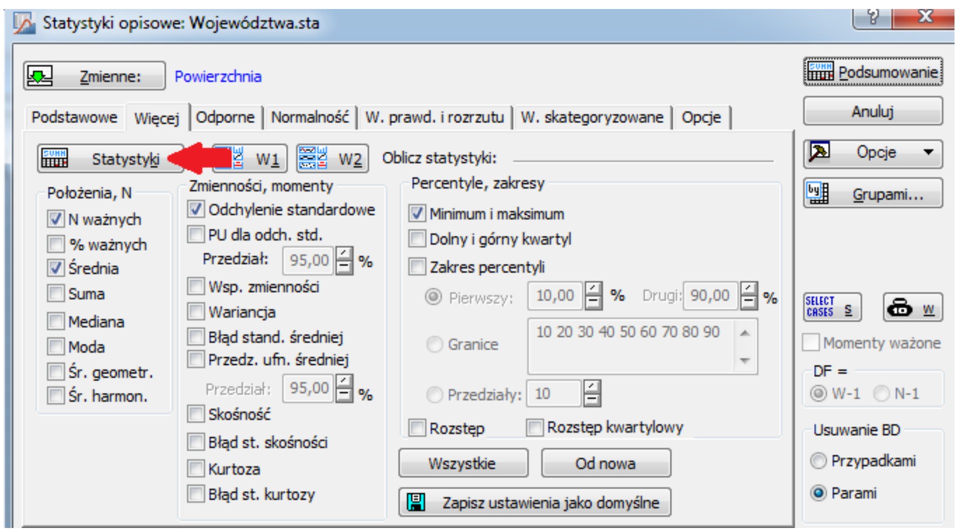
     

     Wybieramy zmienną i klikamy **Ok**

  2. Klikamy na menu **Statystyka** i wybieramy opcję **Statystyki podstawowe**. Następnie wybieramy **Statystyki opisowe** i zmieniamy kartę z **Podstawowe** na **Więcej**



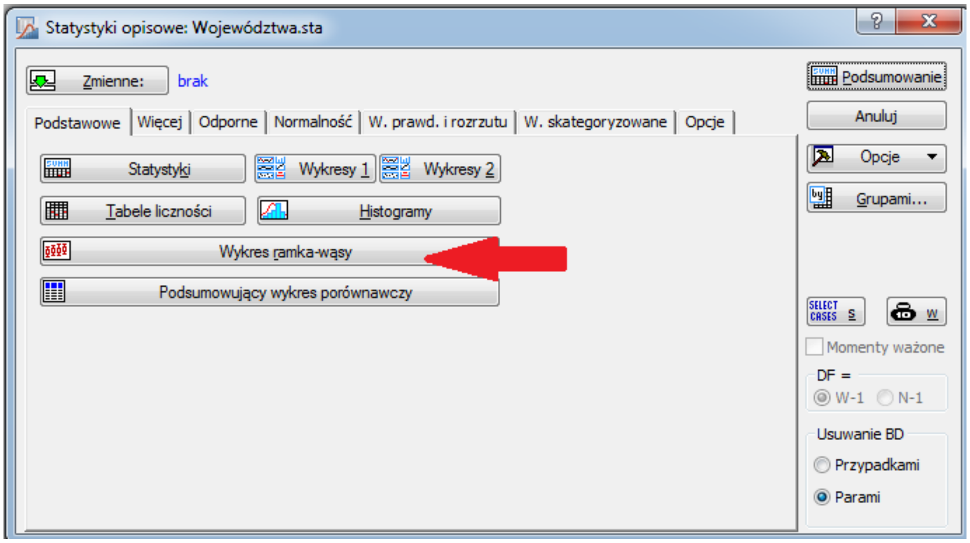
Następnie zaznaczamy interesujące nas wielkości, klikamy **Statystyki** i wybieramy zmienną.



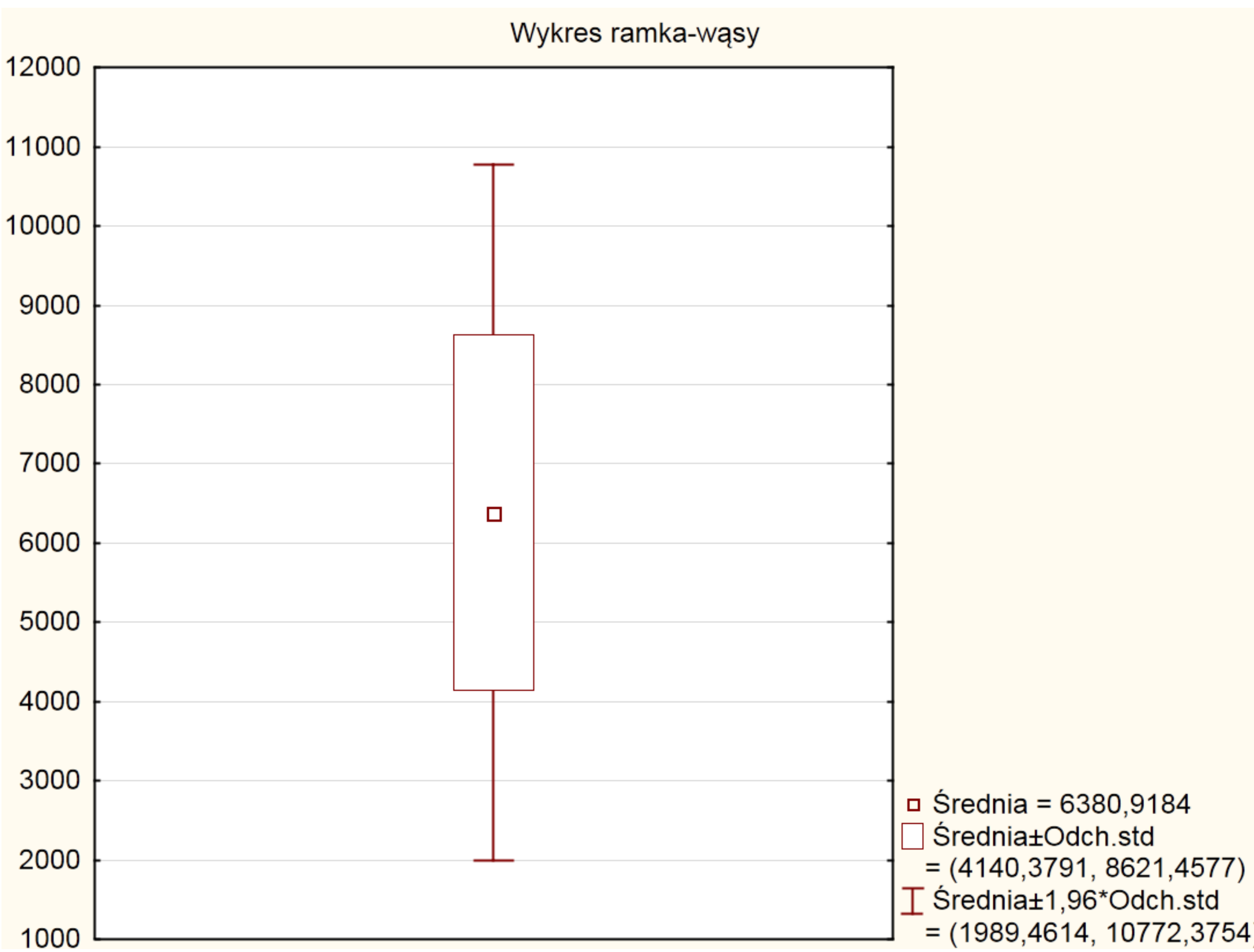
Otrzymujemy w wierszu policzone parametry.

	Statystyki opisowe (Województwa)											
Zmienna	**Nważnych**	Średnia	Mediana	Moda	Liczność Mody	Minimum	Maksimum	Wariancja	Odch.std	Wsp.zmn.	Skośność	Kurtoza
Powierzchnia	49	6380,918	6283,000	5348,000	2	1523,000	12327,00	5020016	2240,539	35,11312	0,444440	-0,066741

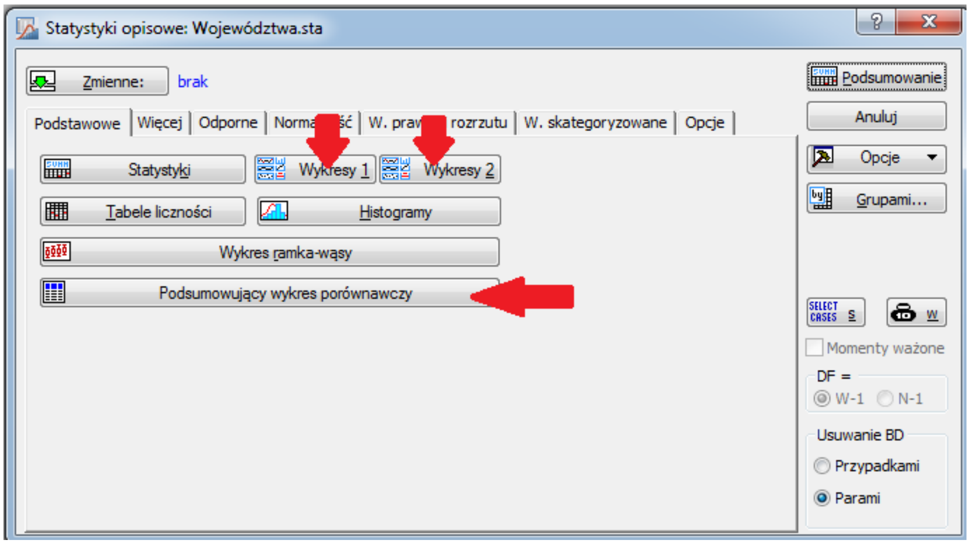
- **Wykres pudełkowy**
  Klikamy na menu **Statystyka** i wybieramy opcję **Statystyki podstawowe**. Następnie wybieramy **Statystyki opisowe** i klikamy przycisk **Wykres ramka-wąsy**



Wybieramy zmienną i klikamy **Ok**.



- **Prezentacja danych**
  Statistica oferuje prezentację danych takich jak podstawowe miary, histogram i wykres pudełkowy w jednym oknie. Do wyboru są trzy najbardziej podstawowe opcje. Dwie z nich to wykresy, aby wywołać te opcje klikamy na menu **Statystyka** i wybieramy opcję **Statystyki podstawowe**. Następnie wybieramy **Statystyki opisowe** i wybieramy albo przycisk **Wykresy 1** albo **Wykresy 2**. Trzecią możliwością jest kliknięcie przycisku **Podsumowujący wykres porównawczy**



Poniżej prezentujemy działanie ostatniej możliwości (wykresu porównawczego)