

STATYSTYKA OPISOWA

WYKŁAD 3,4

Szereg rozdzielczy przedziałowy (dane pogrupowane) (stosujemy w przypadku dużej liczby niepowtarzających się danych)

Przedział ($w_i ; w_{i+1}$)	Środek $\dot{x}_i$	Liczebność n_i	Liczebność skumulowana s_i
($w_1 ; w_2$)	$\dot{x}_1$	n_1	$s_1 = n_1$
($w_2 ; w_3$)	$\dot{x}_2$	n_2	$s_2 = n_1 + n_2$
...	...	...	...
($w_r ; w_{r+1}$)	$\dot{x}_r$	n_r	$s_r = n$
razem	---	n	---

X – badana cecha,
 n – liczba danych statystycznych,
 r – liczba przedziałów (klas),
 $(w_i; w_{i+1})$ przedziały (klasy),
 $\dot{x}_i$ – środki przedziałów (klas),
 n_i – liczba danych należących do
 przedziału $(w_i; w_{i+1})$,
 s_i – liczebności skumulowane
 $(s_i = n_1 + n_2 + \dots + n_i)$.
 b – długość klasy

Liczbę klas r można ustalić orientacyjnie wg następującej tabeli:

Liczba danych n	Liczba klas r
$30 \div 60$	$6 \div 8$
$60 \div 100$	$7 \div 10$
$100 \div 200$	$9 \div 12$
$200 \div 500$	$11 \div 17$
> 500	$16 \div 25$

Będziemy przyjmować, że klasy są jednakowej długości.

Długość klas b

$$b = \frac{\text{rozstęp}}{r} = \frac{x_{\max} - x_{\min}}{r}$$

(wynik dzielenia **zawsze należy zaokrąglić w górę do klasy dokładności danych statystycznych**).

Jeśli stosujemy przedziały otwarte to
lewy koniec pierwszej klasy
 w_1 wyznaczamy ze wzoru

$$w_1 = x_{\min} - \frac{\alpha}{2}$$

α - klasa dokładności danych.
Kolejne końce klas wyznaczamy
dodając do poprzedniego końca długość
klasy b tzn. $w_2 = w_1 + b$, $w_3 = w_2 + b$, itd.

Średnia (arytmetyczna)

$$\bar{x} = \frac{1}{n} \sum_{i=1}^r n_i x_i$$

Dominanta

$$d = w_d + \frac{n_d - n_{d-1}}{(n_d - n_{d-1}) + (n_d - n_{d+1})} \cdot b$$

w_d – lewy koniec klasy zawierającej dominantę,

d – numer klasy zawierającej dominantę,

n_d – liczebność klasy zawierającej dominantę,

b – długość klas,

Uwaga

Wzoru nie można stosować gdy klasa najliczniejsza jest pierwsza lub ostatnia.

Mediana

$$m_e = w_m + \frac{b}{n_m} \left(\frac{n}{2} - s_{m-1} \right)$$

w_m – lewy koniec klasy zawierającej medianę

m – numer klasy zawierającej medianę

n_m – liczebność klasy zawierającej medianę

s_{m-1} – liczebność skumulowana klasy $m - 1$

Kwartyle

pierwszy kwartyl:

$$q_1 = w_{q_1} + \frac{b}{n_{q_1}} \left(\frac{n}{4} - s_{q_1-1} \right)$$

drugi kwartyl:

$$q_2 = m_e$$

trzeci kwartyl:

$$q_3 = w_{q_3} + \frac{b}{n_{q_3}} \left(\frac{n}{4} \cdot 3 - s_{q_3-1} \right)$$

oznaczenia analogiczne jak przy medianie.

Uogólnieniami mediany i kwartyli są **decyle** (podział na 10 równych części) i **percentyle** (podział na 100 równych części).

Wariancja

$$s^2 = \frac{1}{n} \sum_{i=1}^r n_i \left(x_i - \bar{x} \right)^2$$

Współczynnik asymetrii

$$a = \frac{\frac{1}{n} \sum_{i=1}^r n_i \left(x_i - \bar{x} \right)^3}{s^3}$$

$$\text{lub } a_1 = \frac{\bar{x} - d}{s} \quad (\text{wskaźnik asymetrii})$$

lub
$$a_2 = \frac{q_3 - 2m_e + q_1}{2q}$$

(pozycyjny wskaźnik asymetrii)

gdzie
$$q = \frac{q_3 - q_1}{2}$$

(odchylenie ćwiartkowe)

lub
$$a_3 = \frac{d_9 - 2m_e + d_1}{d_9 - d_1}$$

(decylowy wskaźnik asymetrii)

gdzie **d_1** pierwszy decyl
 d_9 dziewiąty decyl

Pozycyjne odpowiedniki
współczynnika zmienności:

$$v_p = \frac{q}{m_e}$$

i obszaru typowych wartości:

$$[m_e - q, m_e + q]$$

Współczynnik skupienia (kurtoza)

$$k = \frac{\frac{1}{n} \sum_{i=1}^r n_i \left(x_i - \bar{x} \right)^4}{s^4}$$

lub
$$k' = \frac{q_3 - q_1}{2(d_9 - d_1)}$$

(pozycyjny wskaźnik skupienia)

gdzie **d₁ pierwszy decyl**

d₉ dziewiąty decyl

q₁ pierwszy kwartyl

q₃ trzeci kwartyl

POMIAR KONCENTRACJI

Krzywa Lorenza. Współczynnik Giniego.

Najpierw wykreślamy tzw. krzywą Lorenza.

W tym celu na osi poziomej odkładamy

skumulowane częstości $w_i = \frac{n_i}{n}$

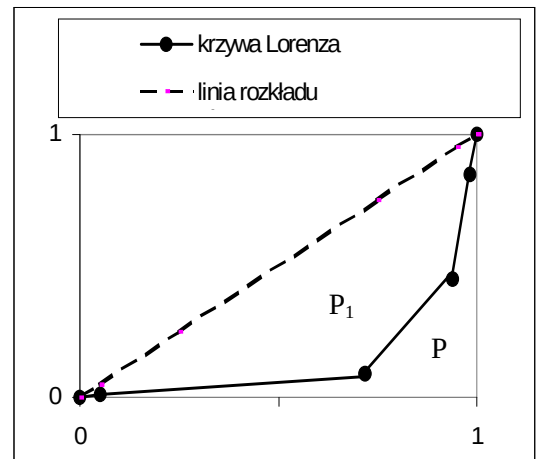
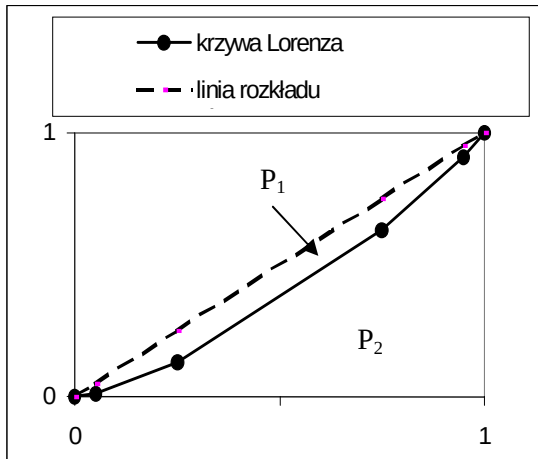
(wskaźniki struktury) a na osi pionowej skumulowane udziały wartości

$$z_i = \frac{\dot{x}_i n_i}{\sum_{i=1}^r \dot{x}_i n_i}$$

(są to udziały i – tego przedziału w wartości globalnej).

Wielkości te wyznaczają ciąg punktów, które łączymy łamaną i przedłużamy ją do początku układu współrzędnych, otrzymana łamana to **krzywa Lorenza**.

Odcinek o końcach $(0, 0)$ i $(1, 1)$ przedstawia **linię równomiernego rozkładu**.



Słaba koncentracja
koncentracja

Silna

Współczynnik Giniego definiujemy jako stosunek pola P_1 do $0,5$ (pole trójkąta pod linią równomiernego rozkładu) co jest równe podwojonemu polu P_1 , tzn.

$$K_G = \frac{P_1}{0,5} = 2P_1 = 1 - 2P_2$$

bo

$$2P_1 + 2P_2 = 1$$

gdzie P_2 jest polem obszaru pod krzywą Lorenza, jest to suma pól trapezów, które łatwo obliczyć.

Zauważmy, że

$$\begin{aligned} 2 * \text{Pole trapezu} &= \\ &= (\text{suma podstaw}) * \text{wysokość} = \\ &\quad (S_{z_{i-1}} + S_{z_{i+1}}) * w_i \end{aligned}$$

(pierwszy trapez jest trójkątem).

$$K_G \in [0,1]$$

Przykład

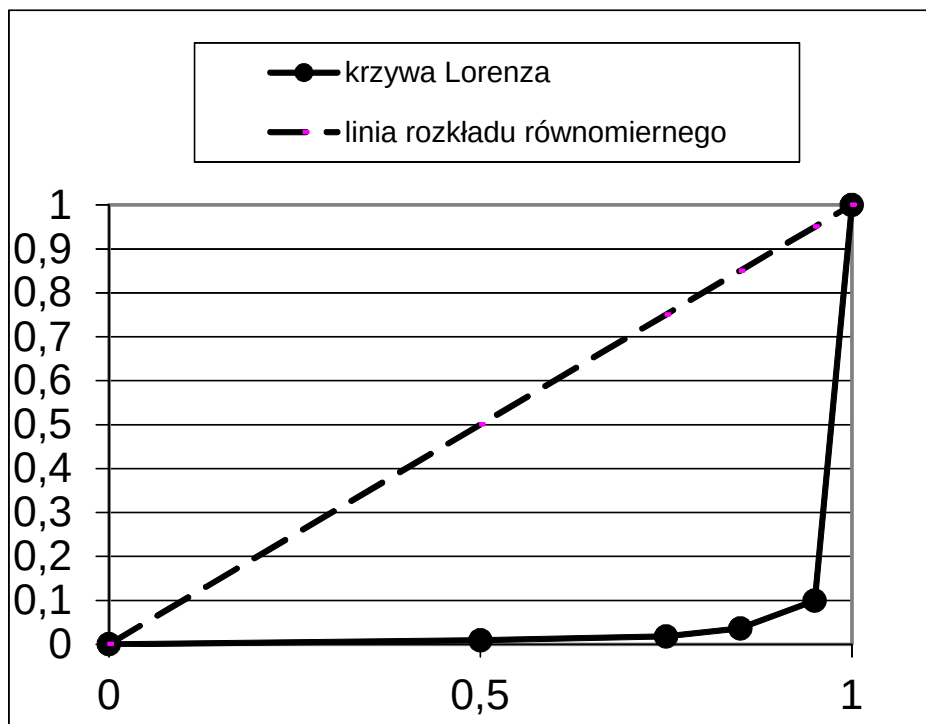
$\dot{x}_i$	n_i	w_i	S_{w_i}	$n_i \dot{x}_i$	$z_i = \frac{\dot{x}_i n_i}{\sum_{i=1}^r \dot{x}_i n_i}$	S_{z_i}	$(S_{z_{i-1}} + S_{z_{i+1}}) * w_i$
1	50	0,5	0,5	50	0,009	0,009	0,004504505
2	25	0,25	0,75	50	0,009	0,018	0,006756757
10	10	0,1	0,85	100	0,018	0,036	0,005405405
35	10	0,1	0,95	350	0,063	0,099	0,013513514
1000	5	0,05	1	5000	0,9009	1	0,054954955
	100	1		5550	1		0,085135135

W ostatniej kolumnie są wyliczone podwojone pola trapezów pod krzywą Lorenza, zatem $P_2 = 0,085135135$.

Stąd

$$K_G = 1 - 2P_2 \approx 1 - 0,085 = 0,915,$$

co świadczy o bardzo dużej koncentracji
(najwyższą wartość osiągają nieliczne
elementy próby).



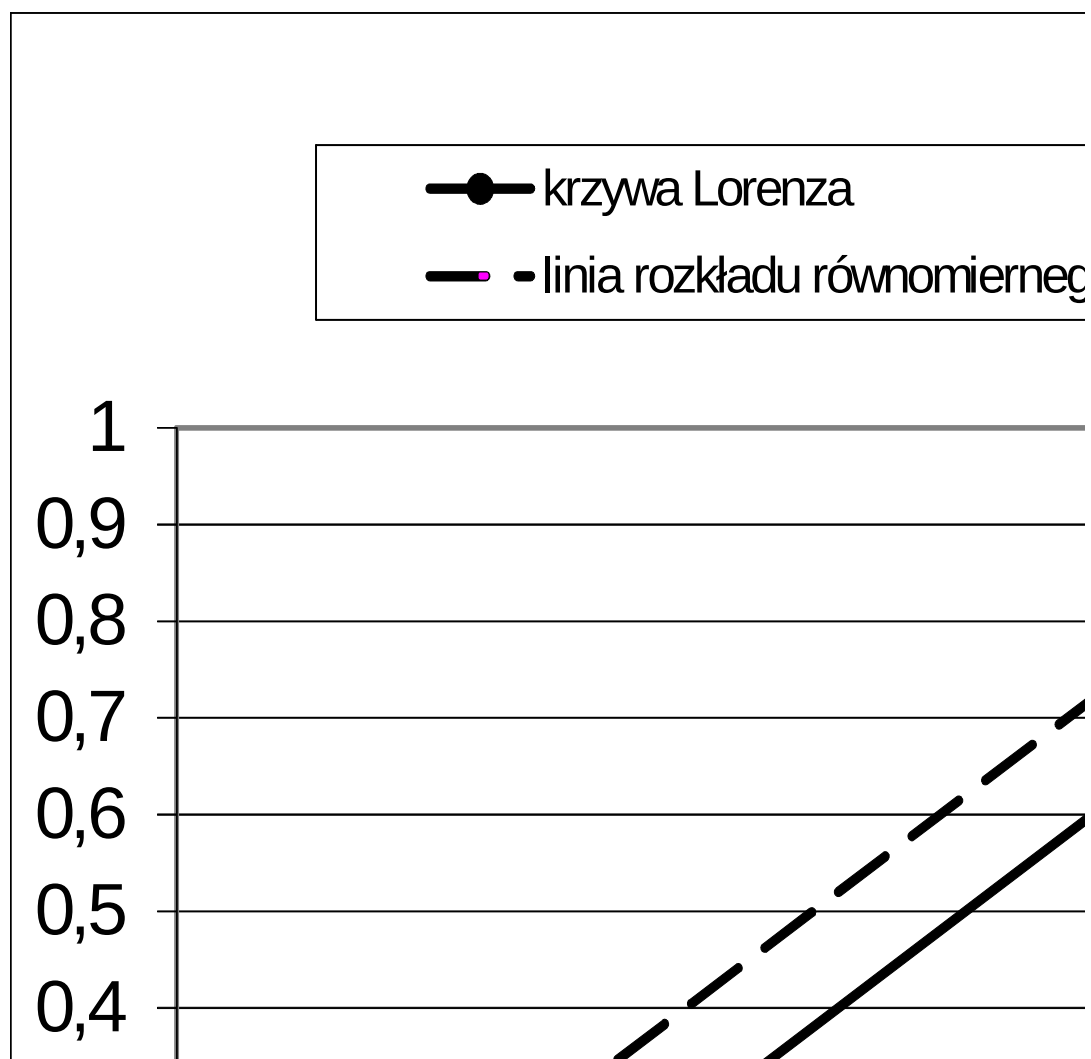
Przykład

$\dot{x}_i$	n_i	w_i	S_{w_i}	$n_i \dot{x}_i$	$z_i = \frac{\dot{x}_i n_i}{\sum_{i=1}^r \dot{x}_i n_i}$	S_{z_i}	$(S_{z_{i-1}} + S_{z_{i+1}}) * w_i$
5	5	0,05	0,05	25	0,01	0,01	0,0005
15	20	0,2	0,25	300	0,12	0,13	0,028
25	50	0,5	0,75	1250	0,5	0,63	0,38
35	20	0,2	0,95	700	0,28	0,91	0,308
45	5	0,05	1	225	0,09	1	0,0955
	100	1		2500	1		0,812

W ostatniej kolumnie są
wyliczone podwojone pola
trapezów pod krzywą Lorenza,
zatem $P_2 = 0,812$

Stąd

$K_G = 1 - 2P_2 \approx 1 - 0,812 = 0,188$,
co świadczy o słabej koncentracji.



Przykład

Dwa zakłady A i B wykonują ten sam detal. Średnia dzienna wydajność pracy na jednego pracownika w obu zakładach jest taka sama i wynosi $\bar{x}_A = \bar{x}_B = 40$ szt.

Wiadomo też, że

$$m_{eA} = 45, \quad d_A = 49 \quad m_{eB} = 36, \quad d_B = 34$$

Chcemy porównać wydajność w obu zakładach.

Rozwiązanie

Wydajność w zakładzie A ma rozkład o asymetrii ujemnej (bo $\bar{x}_A - d_A < 0$) zatem ponad połowa pracowników ma wydajność powyżej średniej 40 szt.

Wydajność w zakładzie B ma rozkład o asymetrii dodatniej (bo $\bar{x}_B - d_B > 0$) zatem mniej niż połowa pracowników ma wydajność powyżej średniej 40 szt.

Wniosek

Chociaż średnia wydajność w obu zakładach jest taka sama to bardziej korzystny rozkład wydajności jest w zakładzie A.

Charakterystyki połączonych populacji.

Niekiedy obserwujemy wartości badanej cechy i liczymy charakterystyki w populacji podzielonej na podzbiory a następnie chcemy wyznaczyć wartości tych charakterystyk w całej populacji.

Przyjmijmy, że populacja jest podzielona na k podzbiorów o liczebności równych odpowiednio N_i $i = 1, 2, \dots, k$; tzn. $\sum_{i=1}^k N_i = N$.

Jeśli $\bar{x}_i$ są średnimi w poszczególnych podzbiorach to średnia dla całej populacji jest równa

$$\bar{x} = \frac{N_1}{N} \bar{x}_1 + \frac{N_2}{N} \bar{x}_2 + \dots + \frac{N_k}{N} \bar{x}_k = \sum_{i=1}^k \frac{N_i}{N} \bar{x}_i$$

Jest to tzw. **średnia ważona**.

Jeśli s_i^2 są wariancjami w poszczególnych podzbiorach to wariancja dla całej populacji jest równa

$$s^2 = \sum_{i=1}^k \frac{N_i}{N} s_i^2 + \sum_{i=1}^k \frac{N_i}{N} (\bar{x}_i - \bar{x})^2$$

Pierwszy składnik to tzw. **wariancja wewnątrzgrupowa**, drugi składnik to tzw. **wariancja międzygrupowa**.

DODATEK-RODZAJE ŚREDNICH

Jedną z wielkości charakteryzujących dane liczbowe jest wartość średnia.

Rodzaje średnich:

- Arytmetyczna
- Geometryczna
- Harmoniczna
- Potęgowa

Wybór średniej zależy od rodzaju badanych wielkości i potrzeb analizy danych.

Najczęściej stosowaną średnią jest średnia arytmetyczna.

Średnią arytmetyczną liczb
rzeczywistych

$x_1, x_2, x_3, \dots, x_n$ nazywamy liczbę:

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

Przykład.

Pięciu studentów otrzymało na egzaminie z matematyki oceny: 3, 2, 5, 2, 3. Ile wynosi średnia ocena tych studentów?

(odp. 3)

Jeżeli wśród n danych występują wartości powtarzające się:

x_i występuje n_i razy, $i = 1, 2, \dots, r$

$$n_1 + n_2 + \dots + n_k = \sum_{i=1}^k n_i = n$$

to

$$\bar{x} = \frac{1}{n} (n_1 x_1 + n_2 x_2 + \dots + n_k x_k) = \frac{1}{n} \sum_{i=1}^k n_i x_i = \sum_{i=1}^k \frac{n_i}{n} x_i$$

Ten sposób liczenia średniej
 arytmetycznej nazywamy **średnią
 arytmetyczną ważoną.**

Przykład.

Dwudziestu pięciu studentów otrzymało
na egzaminie z matematyki oceny:

dziesięć ocen 3,

dziesięć ocen 2,

pięć ocen 5.

Ile wynosi średnia ocena tych
studentów?

$$\bar{x} = \frac{1}{25} (10 \cdot 2 + 10 \cdot 3 + 5 \cdot 5) = \frac{75}{25} = 3$$

Średnią geometryczną liczb

rzeczywistych dodatnich $x_1, x_2, x_3, \dots, x_n$

nazywamy pierwiastek n – tego stopnia

z ich iloczynu, tzn.

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

Średnia geometryczna znajduje

najczęściej zastosowanie przeciętnego

tempa zmian w czasie, np. do

uśredniania indeksów giełdowych.

Przykład.

Roczna stopa procentowa w czterech kolejnych latach wynosiła: 10%, 20%, 5%, 15%. Jaka była średnia stopa w tym okresie?

$$\bar{x}_g = \sqrt[4]{0,1 \cdot 0,2 \cdot 0,05 \cdot 0,15} = \sqrt[4]{0,00015} \approx 0,1107 \approx 11,07\%$$

Zauważmy, że średnia arytmetyczna tych danych wynosi 12,5%

Średnią harmoniczną liczb $x_1, x_2, x_3, \dots, x_n$ różnych od zera nazywamy odwrotność średniej arytmetycznej odwrotności liczb, tzn.

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

Średnią harmoniczną stosuje się przy uśrednianiu wielkości względnych, np. przy obliczaniu przeciętnej prędkości lub średniej gęstości zaludnienia.

Przykład.

Pan Kowalski codziennie dojeżdża do pracy samochodem z prędkością 40km/h. Pewnego dnia zaspał i wyjechał później niż zwykle. W połowie trasy zorientował się, że nie zdąży i zwiększył prędkość o 20km/h, dzięki czemu nie spóźnił się do pracy. Z jaką średnią prędkością jechał tego dnia pan Kowalski?

$$\bar{x}_h = \frac{2}{\frac{1}{40} + \frac{1}{60}} = \frac{240}{5} = 48$$

Zauważmy, że średnia arytmetyczna tych danych wynosi 50km/h

Średnią potęgową rzędu k liczb

rzeczywistych dodatnich $x_1, x_2, x_3, \dots, x_n$
nazywamy liczbę.

$$\bar{x}_{p(k)} = \sqrt[k]{\frac{x_1^k + x_2^k + \dots + x_n^k}{n}} = \sqrt[k]{\frac{1}{n} \sum_{i=1}^n x_i^k}$$

Uwaga:

Dla $k = 1$ jest to średnia
arytmetyczna,

Dla $k = -1$ jest to średnia
harmoniczna,

Dla $k = 2$ jest to **średnia
kwadratowa**,

Przykład.

Mamy 3 pojemniki sześciennie o krawędziach odpowiednio 1, 2 i 3. Chcemy znaleźć taką krawędź sześciennego pojemnika, aby trzy pojemniki o tej krawędzi zastąpiły dotychczas używane, to znaczy, aby łączna objętość poprzednich i nowych była taka sama.

$$\bar{x}_{p(3)} = \sqrt[3]{\frac{1^3 + 2^3 + 3^3}{3}} = \sqrt[3]{12} \approx 2,29$$

Zauważmy, że średnia arytmetyczna tych danych wynosi 2.

Twierdzenie

Dla dowolnych liczb rzeczywistych dodatnich $x_1, x_2, x_3, \dots, x_n$ zachodzą nierówności

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

przy czym równość zachodzi wtedy i tylko wtedy, gdy $x_1 = x_2 = x_3 = \dots = x_n$.