Analiza struktury c.d.

Wykład 5

elzbieta.golata@ue.poznan.pl
dr hab. Elżbieta Gołata, prof. nadzw. UEP,
Katedra Statystyki
Wydział Informatyki i Gospodarki Elektronicznej
Uniwersytet Ekonomiczny w Poznaniu

Miary analizy struktury	

MIARY ANALIZY STRUKTURY

KLASYCZNE POZYCYJNE

1. CHARAKTERYSTYKI TENDENCJI CENTRALNEJ

- średnia arytmetyczna
- średnia geometryczna
- średnia harmoniczna
- średnia kwadratowa

- kwantyle (kwartyle, decyle, percentyle)
- dominanta (wartość najczęściej występująca, moda)

2. CHARAKTERYSTYKI ZRÓŻNICOWANIA - DYSPERSJI - ZMIENNOŚCI

- odchylenie przeciętne
- wariancja
- odchylenie standardowe
- klasyczny współ. zmienności

- rozstęp, obszar zmienności
- odchylenie ćwiartkowe
- odchylenie decylowe ...
- pozycyjny współ. Zmienności

3. CHARAKTERYSTYKI ASYMETRII - SKOŚNOŚCI

- moment trzeci centralny
- moment trzeci centralny stand.

- pozycyjny miernik asymetrii
- pozycyjny współ. asymetrii

klasyczno-pozycyjny miernik asymetrii

klasyczno-pozycyjny współczynnik asymetrii

4 A. CHARAKTERYSTYKI KONCENTRACJI WOKÓŁ ŚREDNIEJ

(kurtozy-ekscesu)

moment czwarty centralny

moment czwarty centralny standaryzowany

4 B. CHARAKTERYSTYKI KONCENTRACJI-RÓWNOMIERNOŚCI PODZIAŁU

współczynnik koncentracji K

Statystyka opisowa	Podstawowe rodzaje badań statystycznych

MIARY DYSPERSJI

Miary oparte na różnicy

ROZSTĘP (empiryczny obszar zmienności) (RANGE)

$$\mathbf{R = X_{max} - X_{min}}$$

- ✧ wstępna ocena zmienności
- ✧ miara łatwa do ustalenia
- ✧ mała wartość poznawcza
- ✧ uzależniona od wartości skrajnych (wynika to z definicji)
- ✧ przybliżona wartość R - szereg rozdzielczy przedziałowy

ODCHYLENIE ĆWIARTKOWE (INTER-QUARTILE RANGE)

1/2 obszaru zmienności 50 % środkowych jednostek zbiorowości

- ✧ wyeliminowanie wpływu jednostek skrajnych

$$Q(x) = \frac{Q_3 - Q_1}{2}$$

$$Q_2 - Q(x) < X_{typ} < Q_2 + Q(x)$$

Nietypowe dla danej zbiorowości są te jednostki, których wartości są niższe bądź wyższe od (*Me-Q(x); Me+Q(x)*)

- ✧ nazywane prawdopodobnym, bo z prawdopodobieństwem równym 0,5 "trafiamy" na jednostkę zawartą między Q_1 a Q_3
- ✧ nie nadaje się do działań algebraicznych
- ✧ stosowane wtedy, gdy nie można obliczyć średnich

MIARY OPARTE NA ODCHYLENIACH

WARIANCJA — VARIANCE

Wariancja w populacji generalnej $\sigma^2 = \frac{\sum_{i=1}^{N}(x_i - \mu)^2}{N}$

szereg szczegółowy $s^2(x) = \frac{\sum_{i=1}^{n}(x_i - \bar{x})^2}{n-1}$

szereg punktowy $s^2(x) = \frac{\sum_{i=1}^{k}(x_i - \bar{x})^2 n_i}{n-1}$

szereg z przedziałami $s^2(x) = \frac{\sum_{i=1}^{k}(x'_i - \bar{x})^2 n_i}{n-1}$

$$s^2(x) = \frac{\sum_{i=1}^{n} x_i^2 - \frac{(\sum_{i=1}^{n} x_i)^2}{n}}{n-1} = \frac{\sum_{i=1}^{n} x_i^2 - n \cdot (\bar{x})^2}{n-1}$$

$$s^2(x) = \frac{\sum_{i=1}^{k} x'^2_i\, n_i - \frac{(\sum_{i=1}^{k} x'_i\, n_i)^2}{n}}{n-1} = \frac{\sum_{i=1}^{k} x'^2_i\, n_i - n \cdot (\bar{x})^2}{n-1}$$

poprawka Sheppard'a $s^2(x) = \frac{\sum_{i=1}^{k}(x'_i - \bar{x})^2 n_i}{n-1} - \frac{i^2}{12}$

ODCHYLENIE STANDARDOWE

Twierdzenie Czebyszewa

Niech c będzie pewną wielkością spełniającą warunek $c \geq 1$. Proporcja obserwacji jakiejkolwiek populacji lub próby znajdujących się w przedziale c odchyleń standardowych od średniej arytmetycznej równa jest co najmniej $1-\frac{1}{c^2}$.

Przedział wartości jednostek populacji $(\mu - c \cdot \sigma\,;\ \mu + c \cdot \sigma)$ obejmuje wszystkie jednostki o wartościach średnia $\pm c$ odchyleń standardowych,

np. $c = 2$, wówczas $1-\frac{1}{2^2} = 0{,}75$

tzn. co najmniej 75% jednostek danej populacji przyjmie wartości z przedziału $\bar{x} \pm 2s(x)$

jeśli $c = 3$ wówczas $1-\frac{1}{3^2} = 0{,}89$

tzn. co najmniej 89% jednostek danej populacji przyjmie wartości z przedziału $\bar{x} \pm 3s(x)$

Reguła trzech sigm

Dla rozkładu normalnego mniej niż 0,026% jednostek przyjmuje wartości spoza przedziału $\mu \pm 3 \cdot \sigma(x)$

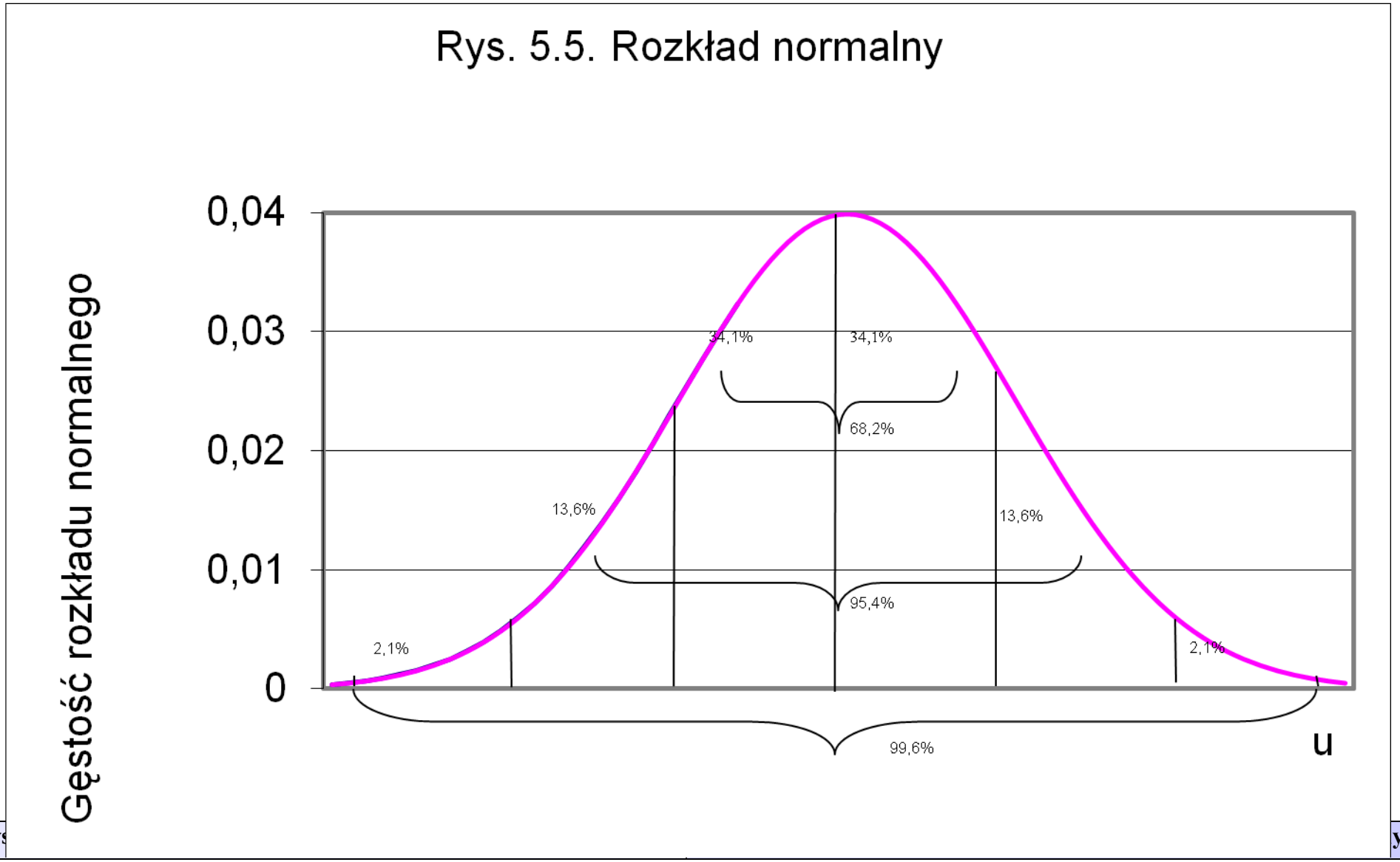
$\mu \pm \sigma$ zawiera 68% obserwacji

$\mu \pm 2\sigma$ zawiera 95% obserwacji

$\mu \pm 3\sigma$ zawiera ponad 99% obserwacji

Typowy przedział zmienności $\bar{x} - s(x) < x_{typ} < \bar{x} + s(x)$

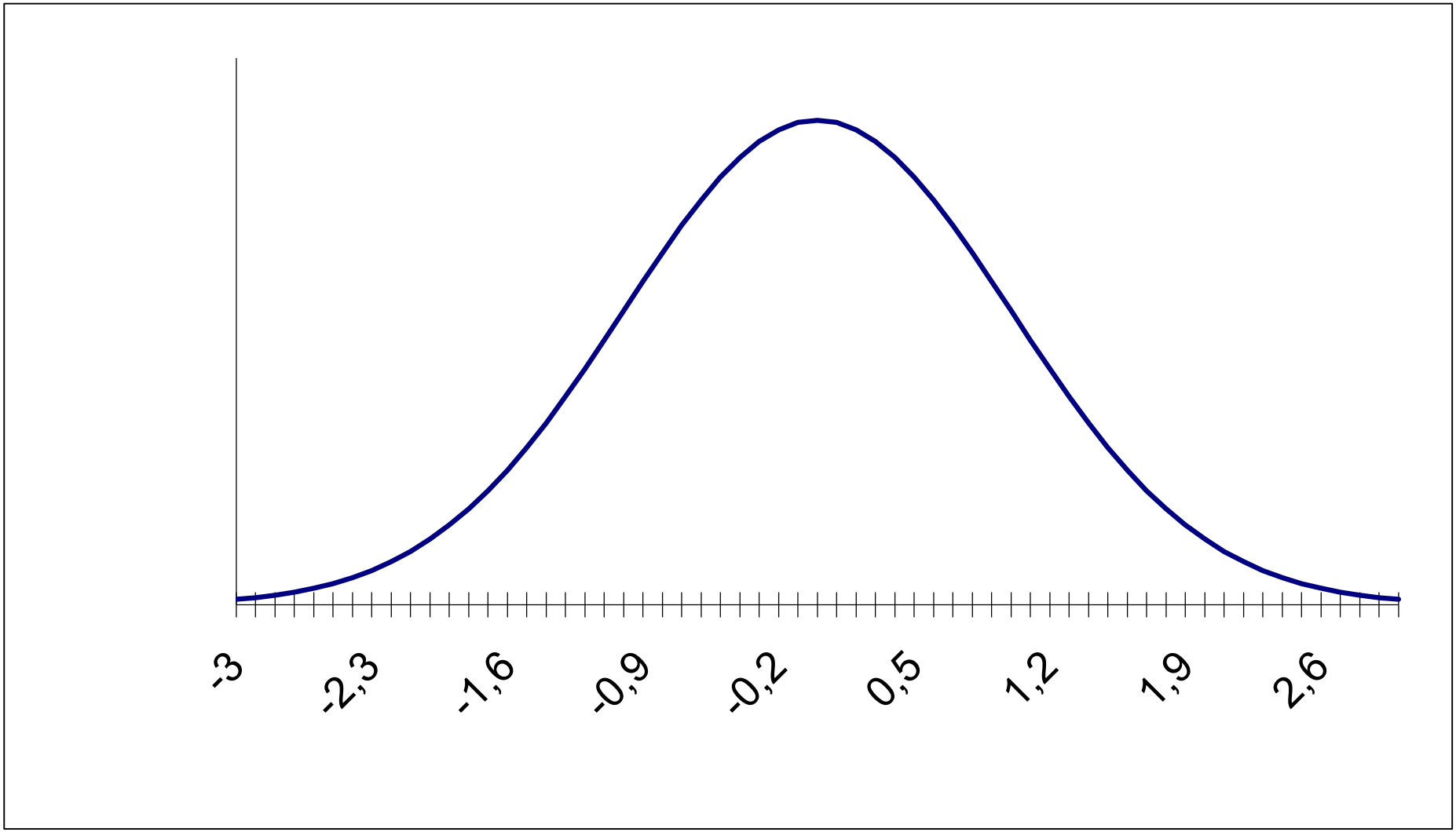
Rys. 5.5. Rozkład normalny



WŁASNOŚCI KRZYWEJ NORMALNEJ:

- jest symetryczna względem prostej $x=\mu$ *(symetryczność);*
- osiąga maksimum równe $\frac{1}{\sigma\sqrt{2\pi}}$ dla $x=\mu$ *(jednomodalność)*;
- jej ramiona mają punkty przegięcia dla $x=\mu-\sigma$ oraz $x=\mu+\sigma$ *(zmienność)*;
- pole powierzchni pod krzywą normalną równe jest jedności
- jest określona dla $x \in (-\infty;+\infty)$
- funkcja gęstości przyjmuje zawsze wartości dodatnie

Z własności pierwszej wynika, że μ decyduje o położeniu krzywej względem osi OX. Natomiast z własności drugiej i trzeciej wynika, że parametr σ określa „smukłość" krzywej. Dwa parametry : μ i σ całkowicie określają funkcję gęstości rozkładu normalnego *(określoność).*


-3
-2,3
-1,6
-0,9
-0,2
0,5
1,2
1,9
2,6

STANDARYZACJA

Znając wartości parametrów μ & σ rozkładu zmiennej losowej X można, stosując

przekształcenie standaryzacyjne

wyznaczyć wartość zmiennej Z dla dowolnej wartości badanego rozkładu

$$z = \frac{X - \mu}{\sigma}$$

- Zmienna Z przedstawia odchylenie od wartości oczekiwanej μ w jednostkach odchylenia standardowego s
- **W ten sposób pozwala na względna ocenę wartości oryginalnej zmiennej**
- Interpretacja zmiennej standaryzowanej ma 'dwie części':

a. **Znak (+) bądź (-)** informuje, czy zmienna oryginalna przyjmuje wartość **mniejszą czy większą od średniej** μ

b. **Wartość zmiennej standaryzowanej wskazuje na 'odległość' od średniej** w jednostkach odchylenia standardowego

Rozkład zmiennej standaryzowanej Z

Wszystkie rozkłady zmiennej losowej X o rozkładzie normalnym o dowolnych parametrach N(μ;σ) można transformować do rozkładu zmiennej Z ‘standaryzowanej’ o rozkładzie normalnym N(0;1)

Ważne cechy rozkładu standaryzowanego

1. Wartość oczekiwana jest równe zero - 0
2. Odchylenie standardowe jest równe jedności – 1
3. Kształt rozkładu po standaryzacji odpowiada rozkładowi zmiennej pierwotnej

Rozkład zmiennej standaryzowanej

	X	X - μ	(X - μ)/ σ	Z	$(z_i - 0)^2$
	26	26 – 19 = 7	1,40	+1.4	1,96
	18	18 - 19 = -1	-0,20	-0.2	0,04
	20	20 – 19 = 1	0,20	+0.2	0,04
	12	12 – 19 = -7	-1,40	-1.4	1,96
suma	**76**			**0,00**	**4,00**
średnia	**19**				**0**
std	**5**				**1**

$\mu = 19$ $\mu = 0$

$\sigma = 5$ $\sigma = 1$

$\mu = (1.4 + -0.2 + 0.2 + -1.4) / 4 = 0$ → **N(0;1)**

$$\sigma = \sqrt{\frac{(1.4 - 0)^2 + (-0.2 - 0)^2 + (0.2 - 0)^2 + (-1.4 - 0)^2}{4}} = 1$$

Porównanie wartości zmiennych z różnych rozkładów

Grzegorz uzyskał 64 pkt. z testu z Botaniki

Karol otrzymał 52 pkt. z testu ze Statystyki

Który z przyjaciół lepiej napisał test ?

Trudność w porównywaniu "surowych" wyników

Dlatego, stosując przekształcenie zwane 'standaryzacją', obydwa wyniki transformujemy w zmienną Z, która przedstawia je w porównywalnej skali, każdy z wyników w relacji do odpowiadającej rozkładowi pierwotnemu wartości oczekiwanej i odchyleniu standardowemu: μ & σ

Zmienna standaryzowana jest bezpośrednio porównywalna

Test z Botaniki (Grzegorz):

$\mu = 60$

$\sigma = 4.5$

$Z = (64 - 60) / 4.5 = +0.89$

Test ze Statystyki (Karol):

$\mu = 45$

$\sigma = 5$

$Z = (52 - 45) / 5 = +1.4$

Przykład

Dane są wyniki testu ze statystyki dla czterech studentów: *A, B, C, D.* Ponadto wiadomo, że: $\bar{x} = 58$ oraz $s(x) = 7$.Oceń, czy praca studenta *D* jest dobra, czy zła na tle pozostałych.

Rozwiązanie:

Zgodnie z regułą trzech sigm wyróżnić można następujące przedziały zmienności badanej cechy: typowy $\bar{x} \pm s(x)$, w którym mieści się około 2/3 obserwacji, zawierający około 75% jednostek ($\bar{x} \pm 2s(x)$) oraz obejmujący prawie wszystkie jednostki ($\bar{x} \pm 3s(x)$).

Dla zmiennej standaryzowanej przedziały te przyjmują następujące granice: (-1 ; 1), (-2 ; 2), (-3 ; 3).

$x_A = 72 \qquad u_a = \frac{72-58}{7} = 2 \qquad 1 - \frac{1}{2^2} = 0{,}75 \qquad x_B = 51p \quad u_B = \frac{51-58}{7} = -1$

$x_C = 58\,p \quad u_c = \frac{58-58}{7} = 0 \qquad x_D = 86 \quad u_D = \frac{86-58}{7} = 4 \quad 1 - \frac{1}{4^2} = 0{,}94$

Student *C* napisał test na poziomie przeciętnym (wartość zmiennej standaryzowanej równa zero).
Student B napisał pracę na poziomie niższym od przeciętnego (dolnej granicy typowego przedziału zmienności).
Student *A* napisał test bardzo dobrze. Jego praca należy do 12,5% najlepszych testów.
Praca studenta D jest jedną z najlepszych.

WSPÓŁCZYNNIK ZMIENNOŚCI

Klasyczny $V_{S(x)} = \frac{s(x)}{\bar{x}} \cdot 100\%$

COEFFICIENT OF VARIATION

Pozycyjny $V_{Q(x)} = \frac{Q(x)}{Me} \cdot 100\%$

Porównanie zmienności
- opisowa miara względna służąca do porównań
- cech jednoimiennych (takich samych) w różnych zbiorowościach
- cech różnorodnych w tej samej zbiorowości

RÓWNOŚĆ WARIANCYJNA

Dekompozycja wariancji na dwa addytywne składniki:

- **średnią arytmetyczną z wariancji warunkowych.**

  zmienność wewnątrzgrupowa: $\overline{s_j^2(x)} = \frac{1}{n}\sum_{j=1}^{l} s_j^2(x) n_j$

- **wariancja średnich warunkowych**

  zmienność międzygrupowa: $s^2(\bar{x}_j) = \frac{1}{n}\sum_{j}^{l} (\bar{x}_j - \bar{x})^2 n_j$

- **wariancja ogólna**: $s^2(x) = \overline{s_j^2(x)} + s^2(\bar{x}_j)$

! Równość wariancyjna nie jest przekształcalna na równość odchyleń standardowych,
stąd jej składniki nie są interpretowane merytorycznie,
mogą jednak odpowiadać na pytanie, jaki jest udział obu wariancji szczegółowych w wariancji ogólnej

PRZYKŁAD

W grupie 100 losowo wybranych klientów badano wydatki na zakup soków i napojów (cecha X) otrzymując następujące wyniki. Proszę wyznaczyć wariancję badanej cechy dla całej zbiorowości. Jak duże jest zróżnicowanie wysokości wydatków na napoje w zależności od miejsca zakupu?

n - 100 klientów, w tym: $n_1 = 40$ klientów małych sklepów

$n_2 = 60$ klientów supermarketów

małe sklepy: $\bar{x}_1 = 6$ PLN $s^2(x_1) = 2$ PLN^2

supermarkety: $\bar{x}_2 = 32$ PLN $s^2(x_2) = 10$ PLN^2

ogółem: $\bar{x} = ?$ PLN $s^2(x) = ?$ PLN^2

Rozwiązanie:

- Średnia ogólna średnia ważona ze średnich dla podgrup: $\bar{x} = \frac{1}{100}(6 \cdot 40 + 32 \cdot 60) = 21{,}6$
- Średnia arytmetyczna z dwóch wariancji: $\overline{s_j^2(x)} = \frac{1}{100}(2 \cdot 40 + 10 \cdot 60) = 6{,}8$
- Wariancja dwóch średnich warunkowych: $s^2(\bar{x}_j) = \frac{1}{100}[(6 - 21{,}6)^2 \cdot 40 + (32 - 21{,}6)^2 \cdot 60] = 162{,}24$
- Wariancja ogólna - równość wariancyjna: $s^2(x) = \overline{s_j^2(x)} + s^2(\bar{x}_j) = 6{,}8 + 162{,}24 = 169{,}04$

Jaką część zmienności ogólnej stanowi zmienność wewnątrzgrupowa ?

$$\frac{\overline{s_j^2(x)}}{s^2(x)} = \frac{6{,}8}{169{,}04} = 0{,}04$$

Jaką część zmienności ogólnej stanowi zmienność międzygrupowa?

$$\frac{s^2(\bar{x}_j)}{s^2(x)} = \frac{162{,}24}{169{,}04} = 0{,}96$$

WSPÓŁCZYNNIKI ASYMETRII

$A_{s(x)} = \frac{\bar{x} - D}{s(x)}$ klasyczno-pozycyjny współczynnik asymetrii $A_{s(x)} \in <-1;1>$

$A_{Q(x)} = \frac{Q_3 + Q_1 - 2Q_2}{2Q(x)}$ pozycyjny współczynnik asymetrii $A_{Q(x)} \in <-1;1>$

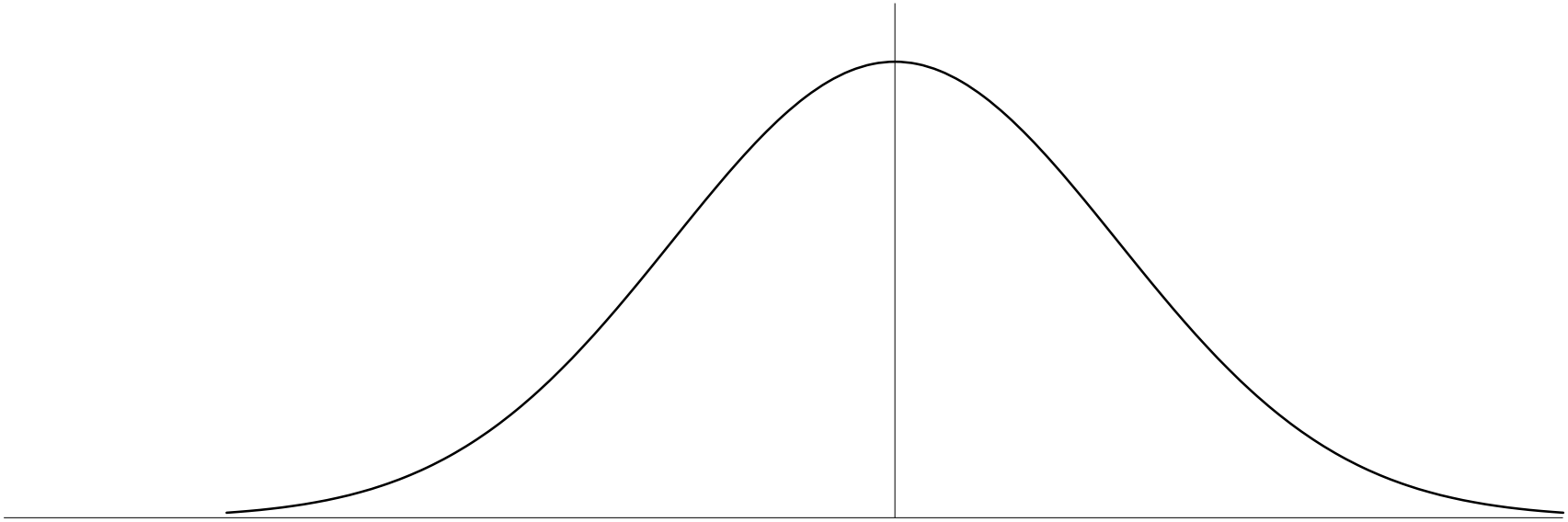
$A_{Q(x)} = \frac{(Q_3 - Q_2) - (Q_2 - Q_1)}{(Q_3 - Q_2) + (Q_2 - Q_1)}$

$Q_3 - Q_2 > Q_2 - Q_1$ $a+$

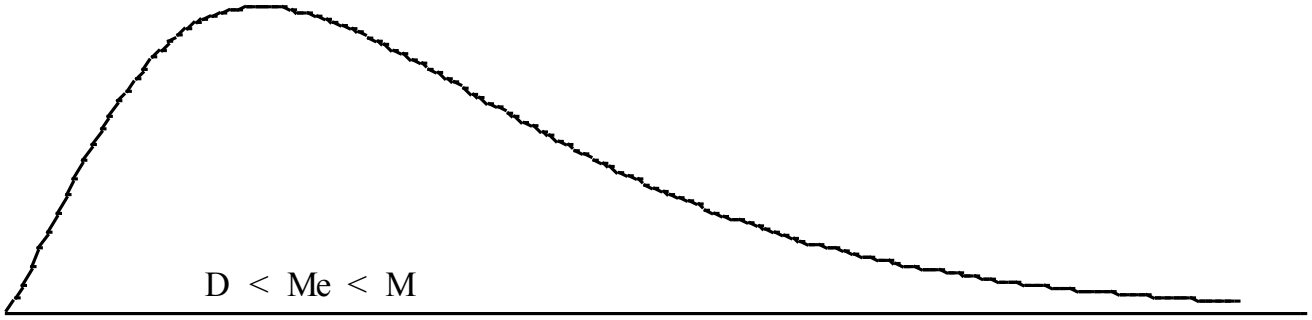
$Q_3 - Q_2 < Q_2 - Q_1$ $a-$

Miary asymetrii **Rachunek momentów**	

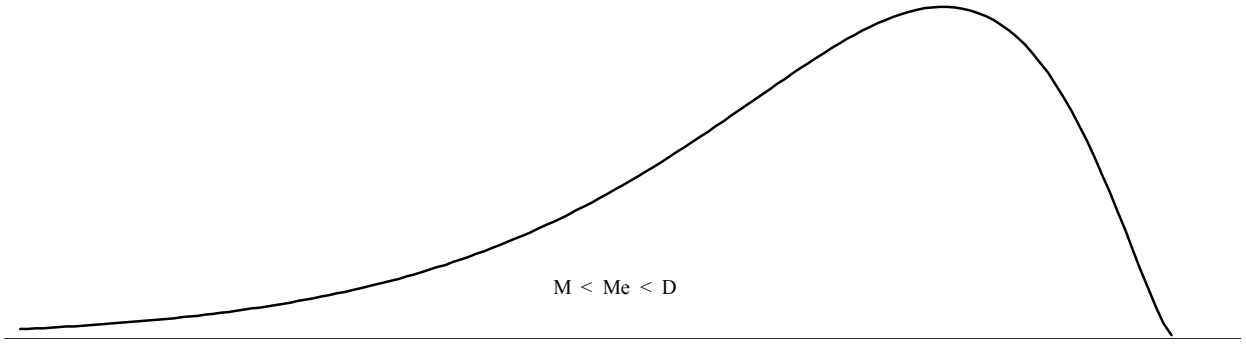
Rozkład symetryczny



Asymetria prawostronna



Asymetria lewostronna



RACHUNEK MOMENTÓW

Dowolnym k-tym momentem rozkładu nazywamy średnią arytmetyczną z odchyleń poszczególnych wartości zmiennej X od dowolnej liczby x_0 podniesionych do k-tej potęgi:

$$m_k = \frac{\sum (x_i - x_0)^k n_i}{n}$$

MOMENTY ZWYKŁE

$$m_1 = \frac{\sum (x_i - 0)}{n} = \frac{\sum x_i}{n} = \bar{x}$$

$$m_2 = \frac{\sum (x_i - 0)^2}{n} = \frac{\sum x_i^2}{n} = \overline{x^2}$$

MOMENTY CENTRALNE

$$\mu_1 = \frac{\sum (x_i - \bar{x}) n_i}{n} = 0$$ zawsze równy zero

$$\mu_2 = \frac{\sum (x_i - \bar{x})^2 n_i}{n} = s^2(x)$$ wariancja

$$\mu_3 = \frac{\sum (x_i - \bar{x})^3 n_i}{n}$$ miara asymetrii

$$\mu_4 = \frac{\sum (x_i - \bar{x})^4 n_i}{n}$$ miara ekscesu

moment trzeci centralny wyrażony w jednostkach odchylenia standardowego:

$$\alpha_3 = \frac{\frac{\sum (x_i - \bar{x})^3 n_i}{n}}{s^3(x)} \qquad \alpha_3 = \frac{\mu_3}{\sqrt{\mu_2{}^3}} \qquad \alpha'_3 = \frac{\mu_3}{s^3(x) + |\mu_3|} \qquad 1 < \alpha'_3 < 1$$

moment czwarty centralny wyrażony w jednostkach odchylenia standardowego:

$$\alpha_4 = \frac{\frac{\sum (x_i - \bar{x})^4 n_i}{n}}{s^4(x)} \qquad \alpha_4 = \frac{\mu_4}{\sqrt{\mu_2^2}}$$

$$\alpha_4 = 3 \qquad \alpha_4 > 3 \qquad \alpha_4 < 3$$

$$\alpha'_4 = \frac{\mu_4 - s^4(x)}{\mu_4} \qquad \alpha'_4 \in <0;1)$$

dla rozkładu normalnego $\alpha'_4 = 0{,}66(6)$

5 - LICZBOWA SYNTEZA FIVE NUMBER SUMMARY

SYNTETYCZNY OPIS ZBIOROWOŚCI PRZY POMOCY PIĘCIU LICZB

Przykład

W 2000 r. Pentor opublikował informacje dotyczące czasu oglądania telewizji przez dzieci w wieku szkolnym według różnych charakterystyk społecznych. W przykładowej próbie 20 dzieci w jednej z poznańskich szkół podstawowych otrzymano następujące informacje o przeciętnej liczbie godzin spędzanych przed telewizorem w ciągu tygodnia:

25 41 27 32 43 66 35 31 15 5
34 26 32 38 16 30 38 30 20 21

Proszę przedstawić i zinterpretować 5 – liczbową syntezę.

Rozwiązanie:

5 15 16 20 21 25 26 27 30 30 31 32 32 34 35 38 38 41 43 66

WARTOŚCI EKSTREMALNE: MIN=5 GODZIN MAX=66 GODZIN

Kwartyle:

- Ponieważ $n=20$ $n/4=5$ $Q_1= 21+0,25*(25-21)=22$
- Ponieważ $n=20$ $n/2=10$ $Q_2= 31+0,5*(31-30)=30,5$
- Ponieważ $n=20$ $3/4n=15$ $Q_3= 35+0,75*(38-35)=37,25$

5 - liczbowa synteza	

Łącznie z uzupełniającą informacją o zróżnicowaniu badanej zbiorowości, 5-liczbową syntezę można zapisać następująco:

Min = 5 ***$Q_1 = 22$*** ***$Q_2 = 30{,}5$*** ***$Q_3 = 37{,}25$*** ***Max = 66***

Q_1 - Min = 17 ***$Q_2 - Q_1 = 8{,}5$*** ***$Q_3 - Q_2 = 6{,}75$*** ***Max- Q_3=28,75***

JEDNOSTKI O WARTOŚCIACH SKRAJNYCH - OUTLIERS

<DG, GG> jednostki, które przyjmują wartość cechy z tego przedziału **nie** są traktowane jako odstające.

$$DG = Q_1 - 1{,}5 \cdot IQR$$

$$GG = Q_3 + 1{,}5 \cdot IQR$$

$$IQR = Q_3 - Q_1.$$

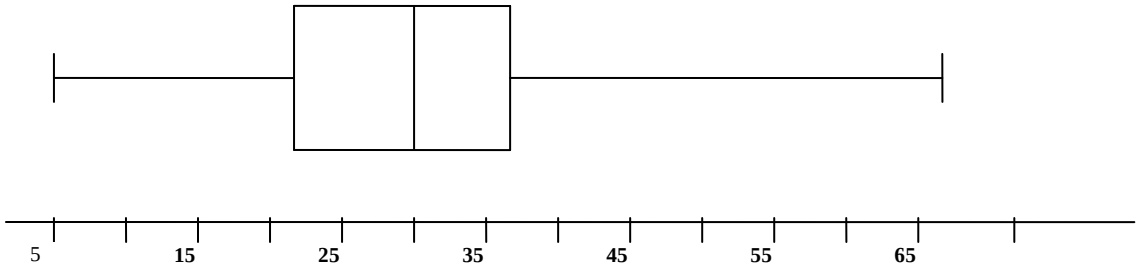
Rozstęp między-kwartylowy ***IQR = 37,25 – 22 = 15,25***

Dolna Granica ***$Q_1 - 1{,}5 \cdot IQR = 22 - 1{,}5 \cdot 15{,}25 = -0{,}875$***

Górna Granica ***$Q_3 + 1{,}5 \cdot IQR = 37{,}25 + 1{,}5 \cdot 15{,}25 = 60{,}125$***

WYKRES PUDEŁKOWY – BOXPLOT

- ✧ Przedstaw 5-liczbowa syntezę
- ✧ Narysuj oś liczbową i nanieś na nią wielkości obliczone w poprzednim kroku. Powyżej osi zaznacz krótkie odcinki pionowe w miejscach odpowiadających kwartylom i połącz je tworząc prostokąt podzielony na dwie części w miejscu odpowiadającym kwartylowi drugiemu.
- ✧ Zaznacz przy pomocy krótkich odcinków pionowych wartości ekstremalne. Połącz odcinkami (tzw. wąsami) boki prostokąta odpowiadające kwartylom z wartościami minimalna i maksymalną.



Min = 5 ***Q_1 = 22*** ***Q_2 = 30,5*** ***Q_3 = 37,25*** ***Max = 66***

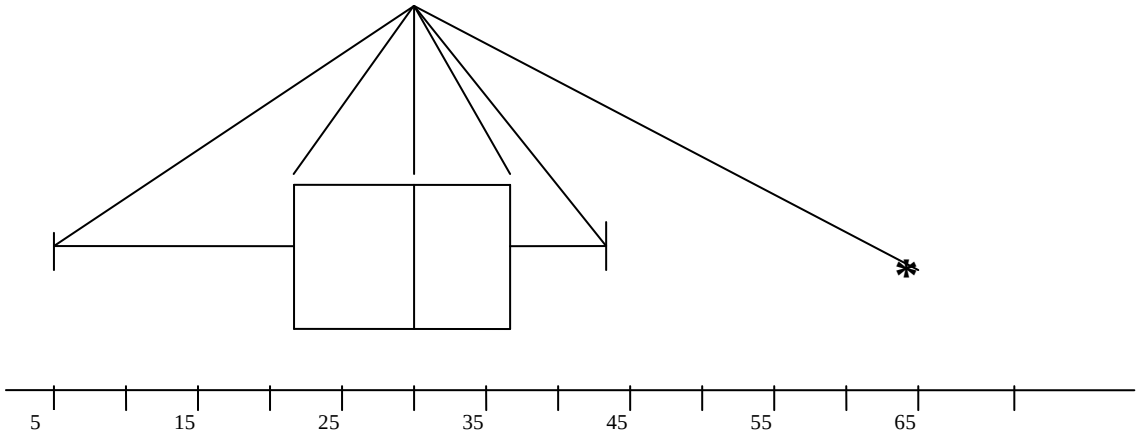
Q_1 - Min = 17 ***Q_2 -Q_1 = 8,5*** ***Q_3 -Q_2 = 6,75*** ***Max-Q_3 =28,75***

5 - liczbowa synteza	

ZMODYFIKOWANY WYKRES PUDEŁKOWY

- ✧ Wyznacz kwartyle
- ✧ Zidentyfikuj potencjalne wartości odstające oraz wartości *ekstremalne** nie będące wartościami odstającymi, tzn. zawarte w przedziale <*DG ; GG*>. Jeżeli w zbiorze nie występują wartości odstające, wówczas są to po prostu wartości ekstremalne tzn. *Min* i *Max*
- ✧ Narysuj oś liczbową i nanieś na nią wielkości obliczone w pierwszym kroku. Powyżej osi zaznacz krótkie odcinki pionowe w miejscach odpowiadających kwartylom i połącz je tworząc prostokąt podzielony na dwie części w miejscu odpowiadającym medianie. Zaznacz przy pomocy krótkich odcinków pionowych wartości *ekstremalne**. Połącz odcinkami (tzw. wąsami) boki prostokąta odpowiadające kwartylom z wartościami *ekstremalnymi**.
- ✧ Narysuj gwiazdkę odpowiadającą każdej potencjalnej wielkości odstającej

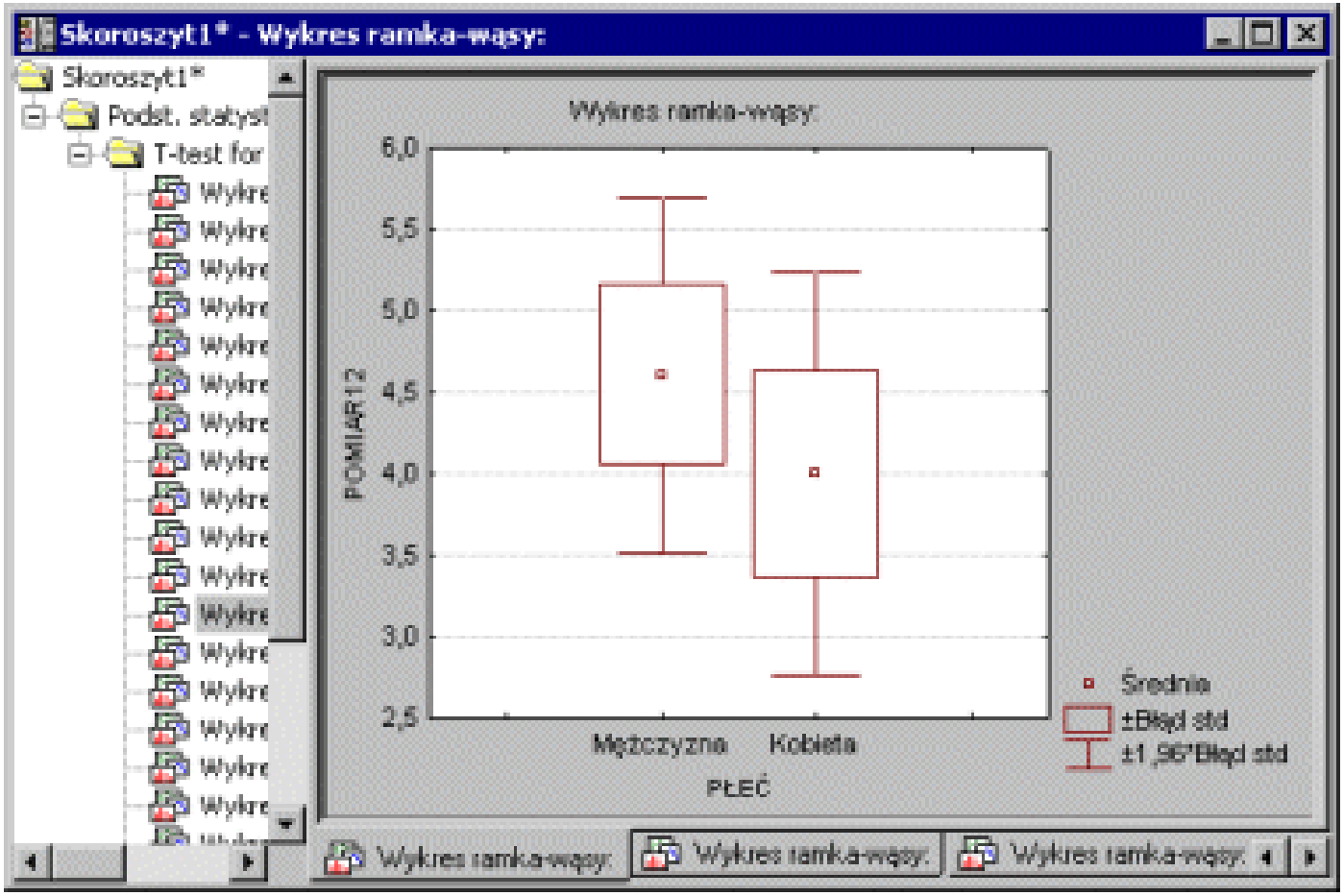
Ekstremalna D*= Min ***Ekstremalna G*= 43***

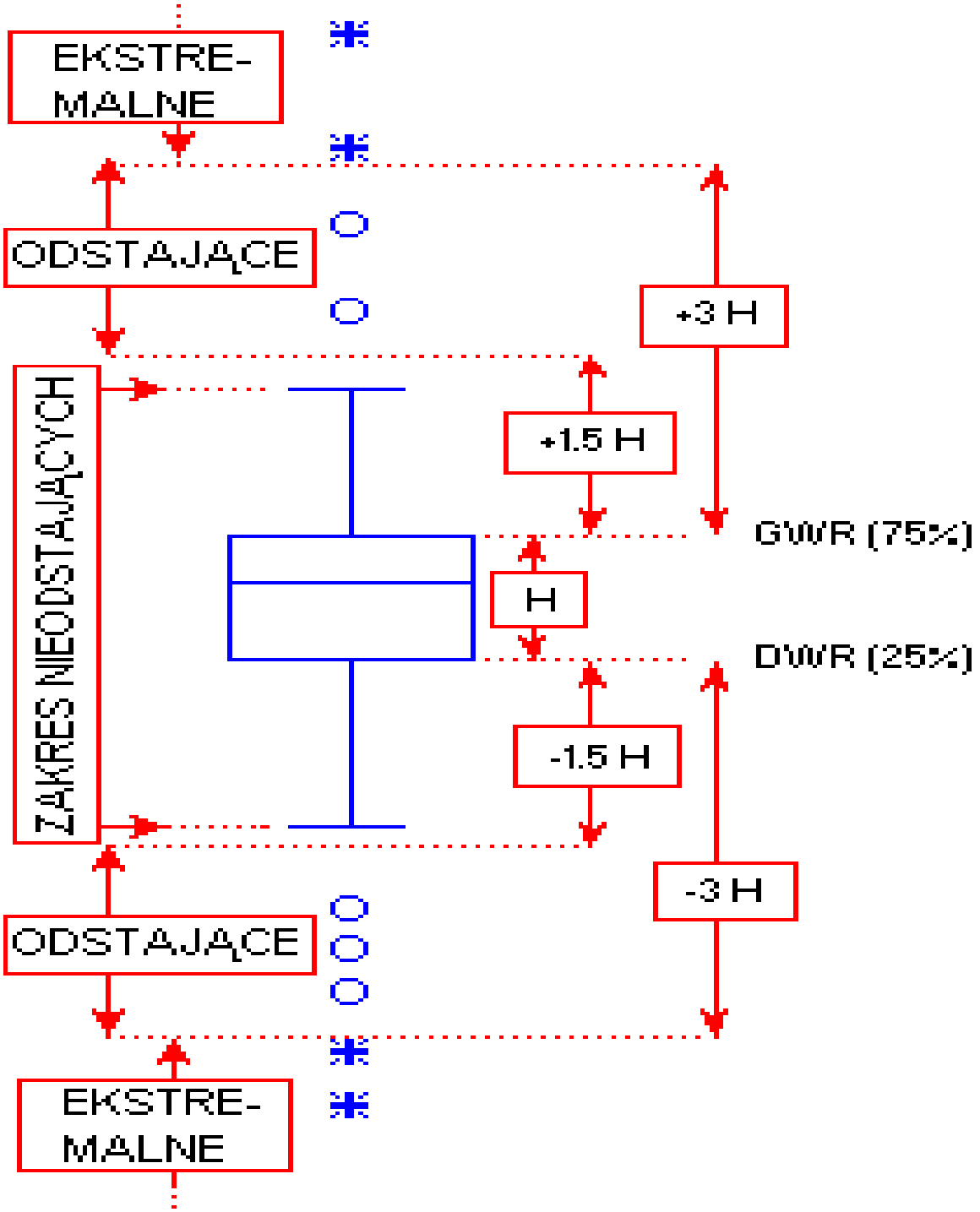


Min = 5 ***Q_1 = 22*** ***Q_2 = 30,5*** ***Q_3 = 37,25*** ***Max = 66***

Q_1 - Min = 17 ***Q_2 -Q_1 = 8,5*** ***Q_3 - Q_2 = 6,75*** ***Max-Q_3=28,75***

5 - liczbowa synteza	




EKSTRE-
MALNE
ODSTAJĄCE
ZAKRES NIEODSTAJĄCYCH
ODSTAJĄCE
EKSTRE-
MALNE
+3 H
+1.5 H
GWR (75%)
H
DWR (25%)
-1.5 H
-3 H

5 - liczbowa synteza	

Dziękuję za uwagę