Analiza struktury

Wykład 4

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

Analiza struktury	

ETAPY BADANIA STATYSTYCZNEGO

1. **PROGRAMOWANIE BADANIA,**
2. **OBSERWACJA STATYSTYCZNA,**
3. **OPRACOWANIE I PREZENTACJA MATERIAŁU STATYSTYCZNEGO**
4. **ANALIZA**

Analiza struktury	

1. ANALIZA

ANALIZA STRUKTURY

OPIS ZBIOROWOŚCI STATYSTYCZNEJ

ANALIZA WSPÓŁZALEŻNOŚCI

BADANIE ZALEŻNOŚCI MIĘDZY CECHAMI STATYSTYCZNYMI

ANALIZA DYNAMIKI

BADANIE ZMIENNOŚCI ZJAWISK W CZASIE

WNIOSKOWANIE STATYSTYCZNE

UOGÓLNIANIE OBSERWACJI DLA ZBIOROWOŚCI PRÓBNEJ NA CAŁĄ ZBIOROWOŚĆ GENERALNĄ

Analiza struktury	

ROZKŁAD EMPIRYCZNY - przyporządkowanie kolejnym uporządkowanym - niemalejącym wartościom zmiennej (x_i) odpowiadających im liczebności (n_i).

Rozkład odzwierciedla strukturę badanej zbiorowości z punktu widzenia określonej cechy.

Kształt rozkładu empirycznego:

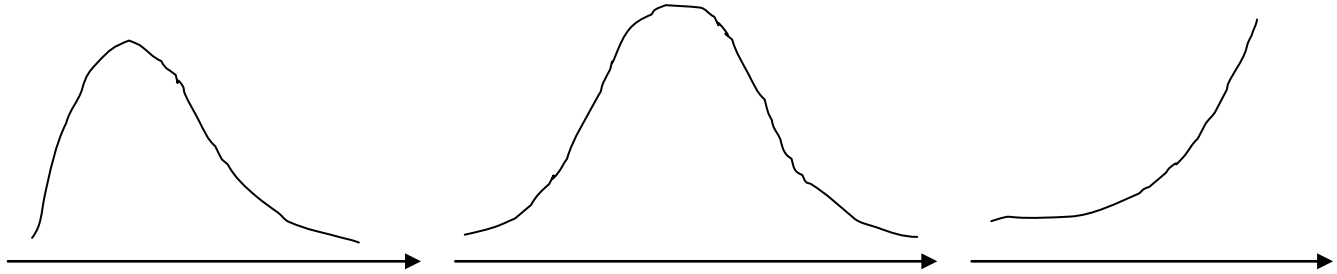
- świadczy o stopniu jednorodności badanego zbioru jednostek

- określa rodzaje miar statystycznych, które można zastosować do opisu struktury

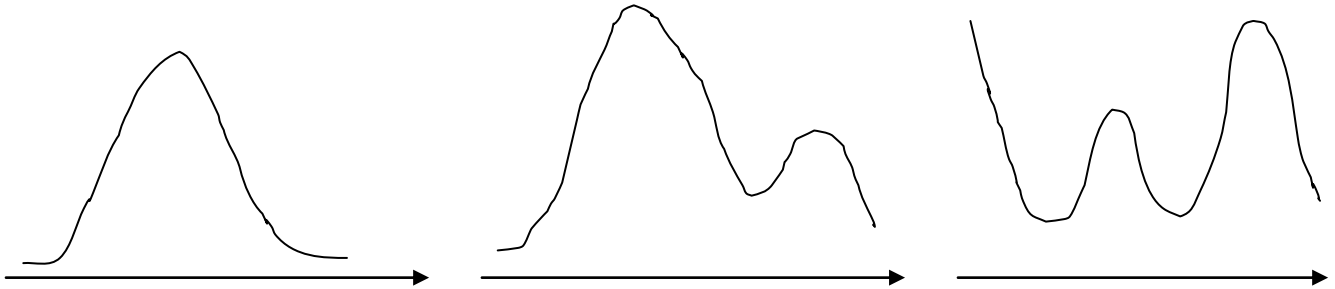
- rozkład jednorodny bądź względnie jednorodny charakteryzuje jedno wyraźnie zaznaczone maksimum, umiarkowane asymetria, zróżnicowanie i koncentracja

Analiza struktury

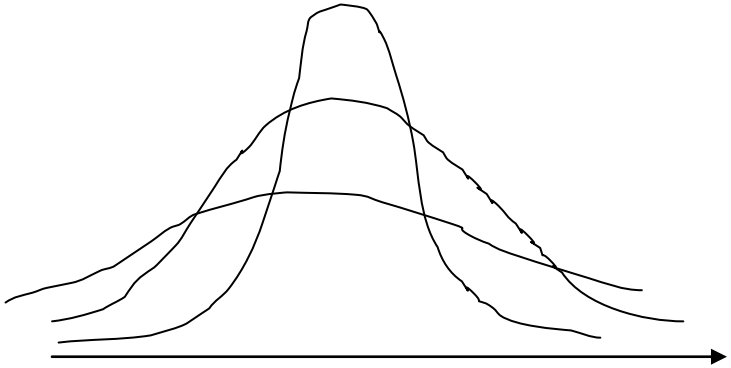
symetryczne i asymetryczne



- jednomodalne i wielomodalne

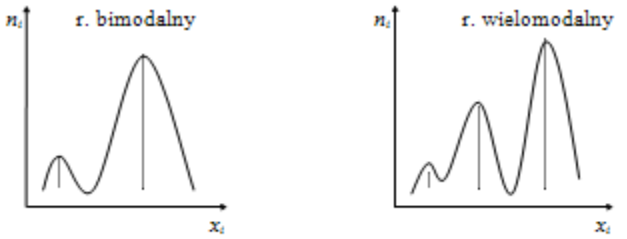


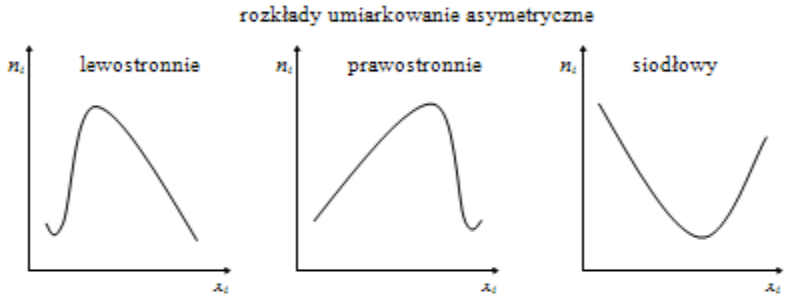
- spłaszczone i wysmukłe

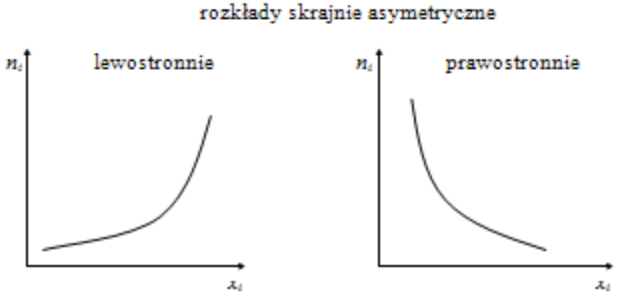


Analiza struktury

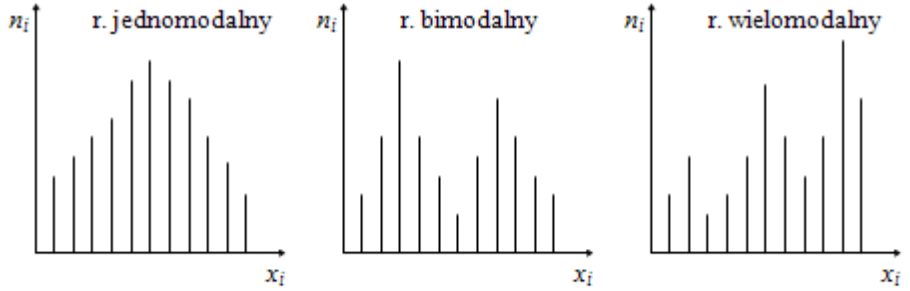
DLA CECHY CIAGŁEJ

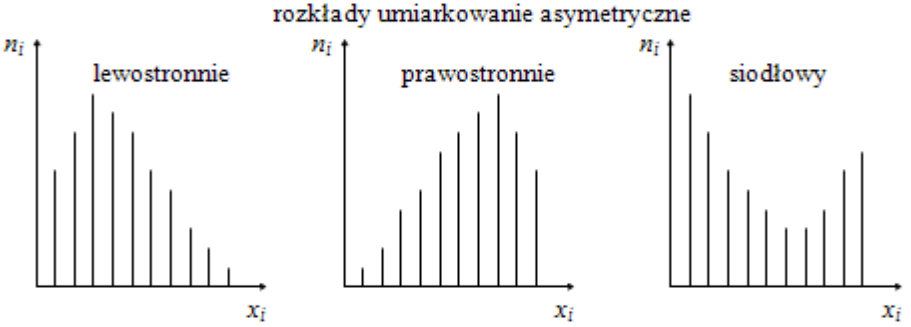


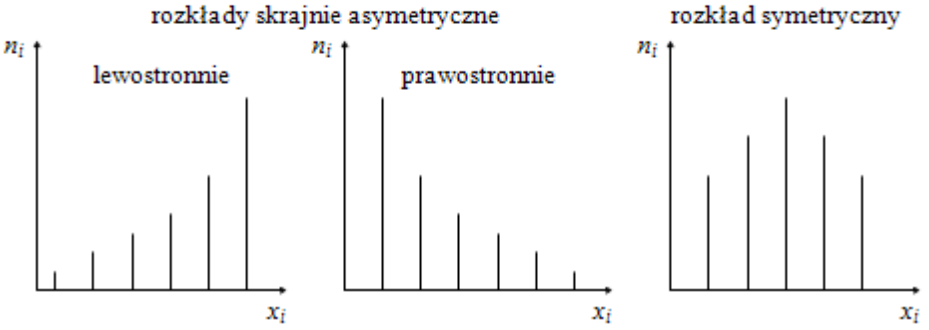




DLA CECHY SKOKOWEJ







POMIAR CECH - SKALE POMIAROWE

ILORAZOWA (ratio scale) - najmocniejsza ze skal pomiarowych, ma następujące 3 właściwości (np. odległość):

1. jakiekolwiek dwie wielkości mogą być wyrażone jako znaczący stosunek (ile razy większe)
2. może być określona różnica pomiędzy dwoma wielkościami (o ile większe)
3. jednostki można uporządkować od najmniejszej do największej (relacja większe lub mniejsze)

PRZEDZIAŁOWA (interval scale) - w odróżnieniu od skali ilorazowej nie posiada naturalnego początku (zera, np. temperatura). Ważność zachowują właściwości 2 i 3.

PORZĄDKOWA (ordinar scale) - posiada tylko własność 3 (relacja większe lub mniejsze, np. oceny wystawiane studentom na zaliczenie)

NOMINALNA (nominal scale) - stosowana dla cech jakościowych – pozwala na wyszczególnienie różnych kategorii - relacja równe lub różne (przypisanie etykiet dla grup jednostek, np. kolor samochodu)

Dystrybuanta empiryczna *(Skumulowane liczebności względne)*

- funkcja $G(x_i)$ ukazująca skumulowany rozkład cechy w *n*-elementowej próbie.

- funkcję $G(x_i)$ można zdefiniować jako skumulowaną częstość empiryczną, sumę częstości empirycznych od pierwszego do danego (*k*-tego) przedziału klasowego w rozkładzie empirycznym badanej cechy:

$$G(x_i) = \frac{n_i}{n}(X \leq x_i)$$ lub

$$G(x_i) = w_i(X \leq x_i)$$ gdzie $w_i = \frac{n_i}{n}$ - liczebności względne

Analiza struktury	

Liczebności absolutne
Absolutne skumulowane
Względne
Skumulowane liczebności względne

Wiek	Urodzenia żywe na 1000 kobiet w grupach wieku							
	1989		2003			2007		
	liczebność	częstości względne %	liczebność	częstości względne %	dystrybuanta empiryczna %	liczebność	częstości względne %	dystrybuanta empiryczna %
15-19	30,9	7,6	14,5	6,0	6,0	12,4	6	6
20-24	168,0	41,1	64,1	26,7	32,7	36	17	23
25-29	124,8	30,5	88,1	36,6	69,3	63	30	53
30-34	60,2	14,7	52,9	22,0	91,3	72	34	87
35-39	24,9	6,1	20,9	8,7	100,0	28	13	100

Konstrukcja szeregu strukturalnego

Wartości cechy x_i	Liczebność n_i	Częstość w_i
x_1	n_1	w_1
x_2	n_2	w_2
x_3	n_3	w_3
.	.	.
.	.	.
.	.	.
x_r	n_r	w_r
Ogółem	n	1

Przedziały klasowe x_{0i}-x_{1i}	Częstość absolutna n_i	Częstość względna w_i	Dystrybuanta empiryczna $F_n(x_i)$	Skumulowana liczebność $n(x_i)$
x_{01}-x_{11}	n_1	w_1	w_1	n_1
x_{02}-x_{12}	n_2	w_2	w_1+w_2	n_1+n_2
x_{03}-x_{13}	n_3	w_3	$w_1+w_2+w_3$	$n_1+n_2+n_3$
.	.	.	.	.
.	.	.	.	.
.	.	.	.	.
x_{0r}-x_{1r}	n_r	w_r	$w_1+w_2+...+w_r$	$n_1+n_2+...+n_r$
Ogółem	n	1	X	X

Analiza struktury	

Emeryci i renciści ZUS* według wysokości wypłacanych świadczeń, 2007, 2008

Wysokość wypłaty		Świadczeniobiorcy 2007 r.				Świadczeniobiorcy 2008 r.			
		w tys.	w odsetkach	w tys.	w odsetkach	w tys.	(%)	w tys.	(%)
		n_i	w_i	$_kn_i$	$_kw_i$	n_i	w_i	$_kn_i$	$_kw_i$
mniej niż	800	1851,9				1413,6			
800	1200	2352,8				2150,3			
1200	1600	1302,5				1850,7			
1600	2000	705				910,1			
2000	i więcej	773,4				1037,8			
Ogółem		**6985,6**				**7362,5**			

Uwaga:
Bez osób pobierających także świadczenia rolnicze
Źródło: Opracowanie własne na podstawie *Ważniejsze informacje z zakresu ubezpieczeń społecznych, 2008,* ZUS, Warszawa, 2009

$$\frac{y_{2008}}{y_{2007}} = \frac{7362,5}{6985,6} = 105,4 \quad (\%)$$

$$\frac{y_{mniej\ niż\ 800}}{y_{2000\ i\ wiecej}} = \frac{1413,6}{1037,8} = 136,21 \quad (\%)$$

Analiza struktury	

Emeryci i renciści ZUS* według wysokości wypłacanych świadczeń, 2007, 2008

Wysokość wypłaty		Świadczeniobiorcy 2007 r.				Świadczeniobiorcy 2008 r.			
		w tys.	w odsetkach	w tys.	w odsetkach	w tys.	(%)	w tys.	(%)
		n_i	w_i	$_kn_i$	$_kw_i$	n_i	w_i	$_kn_i$	$_kw_i$
mniej niż	800	1851,9	26,51			1413,6	19,20		
800	1200	2352,8	33,68			2150,3	29,21		
1200	1600	1302,5	18,65			1850,7	25,14		
1600	2000	705	10,09			910,1	12,36		
2000	i więcej	773,4	11,07			1037,8	14,10		
Ogółem		**6985,6**	100,00			**7362,5**	100,00		

Uwaga:
Bez osób pobierających także świadczenia rolnicze
Źródło: Opracowanie własne na podstawie *Ważniejsze informacje z zakresu ubezpieczeń społecznych, 2008,* ZUS, Warszawa, 2009

Analiza struktury	

Emeryci i renciści ZUS* według wysokości wypłacanych świadczeń, Polska 2007, 2008

Wysokość wypłaty		Świadczeniobiorcy 2007 r.				Świadczeniobiorcy 2008 r.			
		w tys.	w odsetkach	w tys.	w odsetkach	w tys.	(%)	w tys.	(%)
		n_i	w_i	$_kn_i$	$_kw_i$	n_i	w_i	$_kn_i$	$_kw_i$
mniej niż	800	1851,9	26,51	1851,90	26,51	1413,6	19,20	1413,60	19,20
800	1200	2352,8	33,68	4204,70	60,19	2150,3	29,21	3563,90	48,41
1200	1600	1302,5	18,65	5507,20	78,84	1850,7	25,14	5414,60	73,54
1600	2000	705	10,09	6212,20	88,93	910,1	12,36	6324,70	85,90
2000	i więcej	773,4	11,07	6985,60	100,00	1037,8	14,10	7362,50	100,00
Ogółem		**6985,6**	100,00			**7362,5**	100,00		

Uwaga:
Bez osób pobierających także świadczenia rolnicze
Źródło: Opracowanie własne na podstawie *Ważniejsze informacje z zakresu ubezpieczeń społecznych, 2008,* ZUS, Warszawa, 2009

Analiza 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

MIARY TENDENCJI CENTRALNEJ

Kiedy należy - nie należy liczyć średniej arytmetycznej

- średnia jest miarą prawidłową tylko w odniesieniu do zbiorowości w jednorodnych;

- szereg powinien mieć wszystkie przedziały jednakowej rozpiętości;

- powinny one być domknięte;

  - umowne zamykanie przedziałów otwartych o niewielkiej liczebności ($\frac{n_i}{N}$ <5%, $2 \cdot \frac{n_i}{N}$ <3%)

- nie należy liczyć, gdy:

  -bardzo silna asymetria;

  - rozkład bimodlany czy wielomodalny;

  - rozkład siodłowy, w kształcie litery U.

  średnia arytmetyczna zatraca swoją typowość i szansę pojawienia się w rzeczywistości

- szereg szczegółowy $\bar{x} = \frac{\sum_{i=1}^{N} x_i}{N}$ **- szereg punktowy** $\bar{x} = \frac{\sum_{i=1}^{k} x_i n_i}{N}$ **- szereg z przedziałami** $\bar{x} = \frac{\sum_{i=1}^{k} x'_i n_i}{N}$

Dla częstości względnych $w_i = \frac{n_i}{N}$ $\bar{x} = \sum_{i=1}^{k} w_i x'_i$

Gdy dostępne są tylko informacje o wartościach średnich dla grup, to średnia arytmetyczna całości jest średnią arytmetyczną ważoną ze średnich dla poszczególnych grup:

$$\bar{\bar{x}} = \frac{\sum_{i=1}^{k} \bar{x}_i n_i}{N} \qquad N = \sum_i n_i$$

PRZYKŁAD

Przebadano n - 35 studentów ze względu na tygodniowy czas poświęcany na naukę

studenci: $\bar{x}_1 = 12$ godz. $n_1 = 15$

studentki $\bar{x}_2 = 8{,}5$ godz. $n_2 = 20$

$$\bar{\bar{x}} = \frac{\sum \bar{x}_i n_i}{N} = \frac{12 \cdot 15 + 8{,}5 \cdot 20}{35} = 10$$

Własności średniej arytmetycznej

1. *Średnia arytmetyczna jest wypadkową wartości cechy dla wszystkich jednostek zbiorowości* $x_{\min} \langle \bar{x} \langle x_{\max}$

   - do obliczenia średniej nie trzeba znać poszczególnych obserwacji, ale tylko ich ogólną sumę i liczebność

   - jest najlepszą i najczęściej używaną charakterystyką przeciętnego poziomu

2. *Suma wartości zmiennej jest równa iloczynowi średniej arytmetycznej i liczebności zbiorowości:* $\sum_{i=1}^{N} x_i = N\bar{x}$

3. *Suma odchyleń wartości cechy dla poszczególnych jednostek od średniej arytmetycznej jest równa zeru, tzn.:*

$$\sum_{i=1}^{N}(x_i - \bar{x}) = 0$$

3. *Suma kwadratów odchyleń poszczególnych jednostek od wartości średniej równa się minimum, tzn. mniejsza niż od jakiejkolwiek innej dowolnej liczby.* *podstawa KMNK*

$$\sum_{i=1}^{N}(x_i - \bar{x})^2 = MIN \qquad tzn. \quad \sum_{i=1}^{N}(x_i - \bar{x})^2 \le \sum_{i=1}^{n}(x_i - z)^2$$

4. *Jeśli wszystkie wartości zmiennej powiększymy (pomniejszymy, podzielimy lub pomnożymy) o pewna stałą, to średnia arytmetyczna będzie równa sumie (różnicy, ilorazowi, iloczynowi) średniej arytmetycznej wyjściowych zmiennych i tej stałej.*

Analiza struktury	

5. *Na poziom średniej arytmetycznej silny wpływ wywierają wartości skrajne.*

***Wpływ błędu grupowania** na średnią arytmetyczną*

Błąd grupowania jest skutkiem rezygnacji z danych szczegółowych i zastąpienia ich informacjami ogólnymi

- przyjmujemy założenie, że środek przedziału jest rzeczywistą średnią z wartości cechy dla jednostek należących do tego przedziału
- im szerszy przedział tym większe są różnice między środkiem przedziału a średnią;

Różnice pomiędzy środkiem przedziału a średnią $x'_i - \bar{x}_i$ wpływają na wartość średniej:

- asymetria prawostronna - średnia zawyżona
- asymetria lewostronna - średnia zaniżona

Miary tendencji centralnej	

PRZYKŁAD:

nr jedn.	wynagrodzenie
	1 000
2	1 000
3	1 000
4	2 000
5	2 000
6	2 000
7	3 000
8	3 000
9	3 000
SUMA	**18 000**

$$\bar{x} = \frac{18000}{9} = 2000$$

wynagrodzenie x_i	l.pracowników n_i	x'_i	$x'_i\, n_i$
1000 - 1999	3	1500	4500
2000 - 2999	3	2500	7500
3000 - 3999	3	3500	10500
SUMA	9		22500

$$\bar{x} = \frac{22500}{9} = 2500$$

Średnia geometryczna

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

gdy różna częstotliwość występowania wartości zmiennych - wzór ważony:

$$\bar{x}_g = \sqrt[n]{x_2^{n_1} \cdot x_2^{n_2} \cdot \ldots \cdot x_k^{n_k}}$$

Stosowana jest przy badaniu średniego tempa zmian zjawisk.

Własności średniej geometrycznej

1. Jeśli choć jedna wartość jest równa zero, $\Rightarrow \bar{x}_g = 0$;

2. Jeśli choć jedna wartość w szeregu jest ujemna, to $\bar{x}_g$ może się stać liczbą urojoną

   - średnia geometryczna nadaje się więc jedynie do charakteryzowania wartości dodatnich.

3. Jest mniej wrażliwa na wartości skrajne aniżeli średnia arytmetyczna (zmniejsza wpływ różnic, które są często przypadkowe i nie mają większego znaczenia dla rozpatrywanego zjawiska).

4. Odchylenia względne wartości cechy od średniej geometrycznej znoszą się wzajemnie. Iloczyn odchyleń względnych od średniej geometrycznej równa się jedności: $\frac{x_1}{\bar{x}_g} \cdot \frac{x_2}{\bar{x}_g} \cdot \ldots \cdot \frac{x_n}{\bar{x}_g} = 1$

5. Średnia geometryczna iloczynów dwóch szeregów równa się iloczynowi średnich geometrycznych obu szeregów:

   $z = x\,y \qquad \bar{z}_g = \bar{x}_g \cdot \bar{y}_g$

Miary tendencji centralnej	

Dominanta – Moda $D=Mo$

- wartość modalna, typowa – wartość cechy, która powtarza się w szeregu największą ilość razy

- nadaje się do charakteryzowania cech jakościowych

- jest mniej abstrakcyjna niż średnia

Kiedy należy - nie należy liczyć Dominanty

- jednomodalność szeregu - tylko wtedy dominanta ma sens
- wielomodalność świadczy o niejednorodności zbiorowości

  ↓ wyznaczamy maksima lokalne krzywej liczebności
- szereg siodłowy posiada antymodalną
- jednakowa rozpiętość przedziału dominanty i dwóch sąsiadujących
- moda nie znajduje się w pierwszym, ani w ostatnim przedziale - szereg nie jest skrajnie asymetryczny

Miary tendencji centralnej	

*** w szeregu szczegółowym wartość pojawiająca się najczęściej**

*** w szeregu rozdzielczym punktowym wariant cechy najliczniej reprezentowany**

* **szereg rozdzielczy przedziałowy -** wskazanie przedziału

$$D = x_D + \frac{(n_D - n_{D-1})}{(n_D - n_{D-1}) + (n_D - n_{D+1})} c_D$$

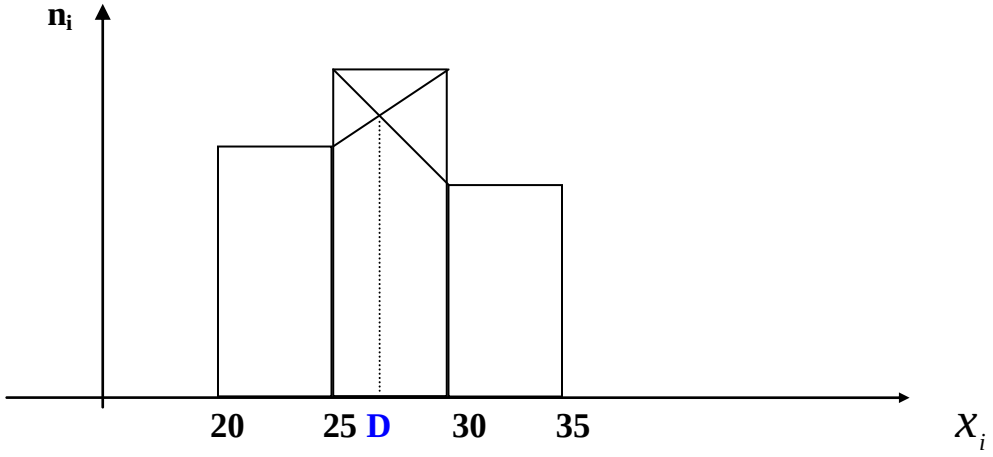
Miary tendencji centralnej

PRZYKŁAD

nowożeńcy wg wieku 2005 x_i	odsetek w_i	
	kobiety	mężczyźni
poniżej 19	6	1
20 - 24	44	29
25 - 29	34	44
30 - 34	8	14
35 – 39	2	4
40 – 49	3	4
50 i więcej	3	4
Suma	N= 100	N= 100

graficzny sposób wyznaczania dominanty

* forma graficzna histogramu

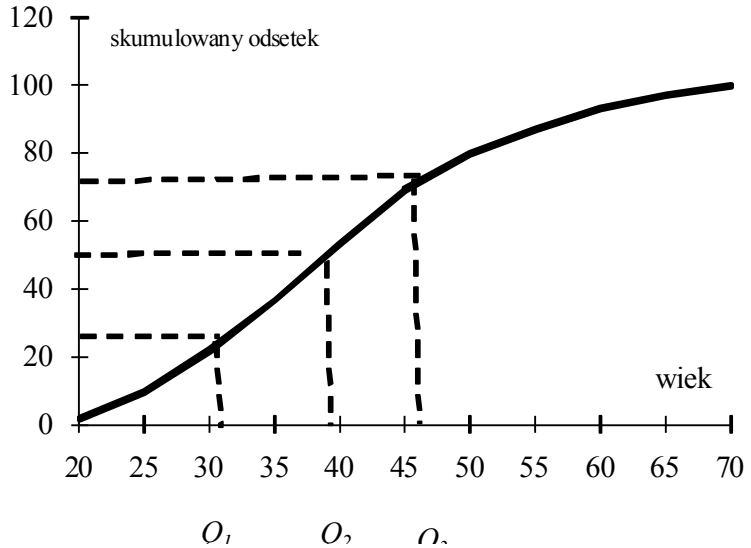


WŁAŚCIWOŚCI MEDIANY:

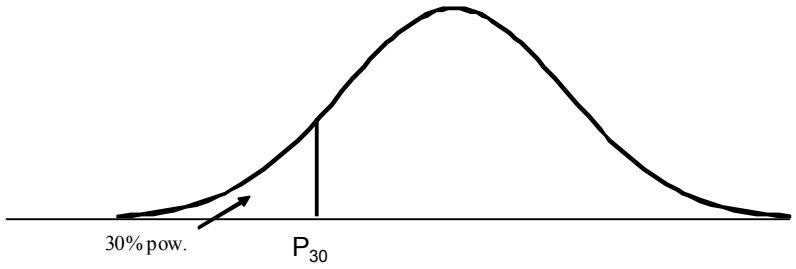
- nie zależy od wartości skrajnych
- można ją wyznaczać przy otwartych przedziałach klasowych i skrajnej asymetrii;
- dokładność przybliżeń Me zależy od rozpiętości przedziałów klasowych;
- nie nadaje się do dalszych przekształceń.

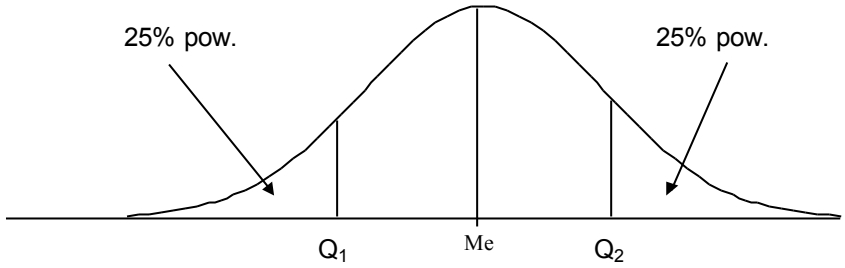
$$Q_2 = Me = x_{Me} + \frac{\frac{n}{2} - \sum_{i=1}^{k_{Me}-1} n_i}{n_{Me}} \cdot i_{Me}$$

Wykres dystrybuanty empirycznej - graficzna metoda wyznaczania kwartyli



Ilustracja graficzna kwantyli





Miary tendencji centralnej	

Dziękuję za uwagę