SIGMA KWADRAT

CZWARTY LUBELSKI KONKURS STATYSTYCZNO-DEMOGRAFICZNY

Opisowa analiza struktury zjawisk masowych

Demografia i statystyka

NBP Narodowy Bank Polski

PROJEKT DOFINANSOWANY ZE ŚRODKÓW NARODOWEGO BANKU POLSKIEGO

URZĄD STATYSTYCZNY W LUBLINIE

WYŻSZA SZKOŁA ZARZĄDZANIA I ADMINISTRACJI W ZAMOŚCIU

POLSKIE TOWARZYSTWO STATYSTYCZNE

Rozkład empiryczny zmiennej

Rozkładem empirycznym zmiennej nazywamy przyporządkowanie kolejnym wartościom zmiennej odpowiadających im liczebności.

Rozkłady empiryczne ustalane są na podstawie konkretnych danych statystycznych

Rodzaje rozkładów empirycznych jednowymiarowej zmiennej

Cecha	Skokowa			Ciągła		
Dominanta	Jednomodalne		Wielomodalne	Jednomodalne		Wielomodalne
Symetria	Symetryczne		Asymetryczne (prawo i lewoskośne)	Symetryczne		Asymetryczne (prawo i lewoskośne)
Spłaszczenie	Normalne	Lepto-kurtyczne	Plato-kurtyczne	Normalne	Lepto-kurtyczne	Plato-kurtyczne

Opisowe parametry struktury rozkładów empirycznych
Parametry klasyczne i pozycyjne

Do sumarycznej charakterystyki struktury rozkładów empirycznych służą parametry opisowe. Wyróżnia się **parametry klasyczne** (obliczane na podstawie wszystkich obserwacji) oraz **pozycyjne** (przy ich wyznaczaniu brane są pod uwagę tylko niektóre wartości zmiennej, stojące na określonej pozycji). **Parametry klasyczne** stosuje się przede wszystkim do analizy rozkładów symetrycznych lub umiarkowanie asymetrycznych. **Parametry pozycyjne** są wykorzystywane do badań każdego typu rozkładu, ale zazwyczaj stosowane są w analizie rozkładów silnie asymetrycznych oraz takich, w których występują otwarte przedziały klasowe.

Parametry opisowe rozkładu mogą być **wielkościami absolutnymi** (wyrażonymi w takich jednostkach, jak badana zmienna, np. w kg, godzinach, latach) lub **mieć postać liczb względnych** (ułamkowych lub procentowych). Parametry względne są szczególnie przydatne przy porównywaniu dwóch lub więcej struktur.

Najczęściej wykorzystywane parametry w opisie struktury zbiorowości masowych

- **miary przeciętne** (zwane też miarami poziomu wartości zmiennej, położenia lub średnimi). Służą one do określania tej wartości zmiennej opisanej przez rozkład, wokół której skupiają się wszystkie wartości zmiennej;
- **miary rozproszenia** (zmienności, zróżnicowania, dyspersji), służące do badania stopnia zróżnicowania wartości zmiennej;
- **miary asymetrii** (skośności), informujące o kierunku zróżnicowania wartości zmiennej;
- **miary koncentracji i spłaszczenia**. Miary koncentracji służą do ba dania stopnia nierównomierności rozkładu ogólnej sumy wartości zmiennej między poszczególne jednostki badanej zbiorowości. Miary spłaszczenia informują natomiast o tym, czy skupienie wartości badanej zmiennej wokół średniej w danym rozkładzie jest mniejsze czy większe niż w zbiorowości o rozkładzie normalnym.

Miary średnie

Najczęściej wykorzystywanymi w analizie średnimi są:

- **Miary klasyczne:**
  - średnia arytmetyczna
  - średnia harmoniczna
  - średnia geometryczna
- **Miary pozycyjne:**
  - dominanta (modalna, wartość najczęstsza)
  - kwantyle (kwartale – dzielące zbiorowość na cztery części, kwintale – dzielące zbiorowość na pięć części, decyle – dzielące zbiorowość na dziesięć części oraz percentyle – dzielące zbiorowość na sto części)

Obydwie grupy miar nie tylko nie wykluczają się ale uzupełniają. Każdy z nich opisuje bowiem poziom wartości cechy z innego punktu widzenia.

Prosta (zwykła) średnia arytmetyczna

Jest ilorazem sumy wartości zmiennej i liczebności badanej zbiorowości:

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

gdzie:

$\bar{X}$ — symbol średniej arytmetycznej

x_i — wariant cech mierzalnej lub wartość przyjęta przez *i*–tą obserwację (jednostkę, obiekt)

N — liczebność badanej zbiorowości

Ważona średnia arytmetyczna

Ważona średnia arytmetyczna obliczana jest na podstawie szeregów rozdzielczych punktowych i przedziałowych. Wagami są liczebności (częstości) odpowiadające poszczególnym wariantom zmiennej:

$$\bar{x} = \frac{x_1 n_1 + x_2 n_2 + ... + x_k n_k}{N} = \frac{\sum_{i=1}^{n} x_i n_k}{N}$$

gdzie:

$n_i (i = 1,2,....., k) -$ liczebności jednostek odpowiadających poszczególnym wartościom zmiennej

$N = n_1 + n_2 + ...n_k -$ ogólna liczebność badanej zbiorowości

Ważona średnia arytmetyczna w szeregach rozdzielczych przedziałowych

W szeregach rozdzielczych przedziałowych wartości zmiennej w każdej klasie nie są jednoznacznie określone, ale zawarte w przedziale od ... do. Dolną granicę przedziału klasowego będziemy oznaczać x_{0i} , górną zaś x_{1i} . W celu obliczenia średniej arytmetycznej z szeregu rozdzielczego przedziałowego należy uprzednio wyznaczyć środki przedziałów klasowych, które oznaczymy symbolem $\tilde{x}$ i obliczamy ze wzoru:

$$\tilde{x} = \frac{x_{0i} + x_{1i}}{2}$$

Wzór na średnią arytmetyczną z szeregu rozdzielczego przedziałowego jest więc następujący:

$$\bar{x} = \frac{\tilde{x}_1 n_1 + \tilde{x}_2 n_2 + ... + \tilde{x}_k n_k}{N} = \frac{\sum_{i=1}^{n} \tilde{x}_i n_k}{N}$$

Rozkład czasu trwania obsługi w banku

x_{0i} - x_{1i}	n_i	Obliczenia pomocnicze	
		$\tilde{x}_i$	$\tilde{x}_i n_i$
od 0 do 5	9	2,5	22,5
od 5 do 10	10	7,5	75
od 10 do 15	16	12,5	200
od 15 do 20	5	17,5	87,5
Ogółem	40	X	385

$$\bar{x} = \frac{385}{40} = 9{,}625$$

Średnia średnich

Jeżeli znane są średnie arytmetyczne dla pewnych grup i na tej podstawie chcemy policzyć średnią arytmetyczną dla wszystkich grup łącznie to korzystamy z formuły:

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

gdzie:

$\bar{x}_i$ – średnia arytmetyczna *i*-tej grupy

n_i – liczebność *i*-tej grupy

$N = n_1 + n_2 + ...n_k$ – ogólna liczebność badanej zbiorowości

Najważniejsze własności średniej arytmetycznej (1)

Jako miara klasyczna jest wypadkową działania wszystkich wartości badanej cechy i spełnia nierówność:

$$x_{\min} < \bar{x} < x_{\max}$$

Suma odchyleń poszczególnych wartości zmiennej od średniej arytmetycznej wynosi 0

$\sum_{i=1}^{k}(x_i - \bar{x}) = 0$ w przypadku szeregu wyliczającego

$\sum_{i=1}^{k}(x_i - \bar{x}) \cdot n_i = 0$ w przypadku szeregu rozdzielczego punktowego

$\sum_{i=1}^{k}(\tilde{x}_i - \bar{x}) \cdot n_i = 0$ w przypadku szeregu rozdzielczego przedziałowego

Najważniejsze własności średniej arytmetycznej (2)

Jeśli pomnożymy średnią przez ogólną liczebność badanej zbiorowości to otrzymamy sumę wartości wszystkich jednostek:

$$N \cdot \bar{x} = \sum_{i=1}^{N} x_i$$

Średnia arytmetyczna sumy (różnicy) zmiennych równa się sumie (różnicy) zmiennych

$$\frac{1}{N} \sum_{i=1}^{N} (x_i + c) = \bar{x} + c$$

Jeżeli wszystkie wartości zmiennej powiększymy (pomniejszymy, podzielimy lub pomnożymy) o pewną stałą c, to średnia arytmetyczna będzie równa sumie (różnicy, ilorazowi lub iloczynowi) średniej arytmetycznej stałej c:

Najważniejsze własności średniej arytmetycznej (3)

Na poziom średniej arytmetycznej silny wpływ wywierają wartości ekstremalne (skrajne), przy czym wpływ ten jest silniejszy w przypadku wysokich wartości zmiennej.

Średnia arytmetyczna – jako wypadkowa wszystkich zaobserwowanych wartości cechy – jest wielkością abstrakcyjną. Oznacza to, że w niektórych przypadkach może przyjmować wartości w ogóle nie występujące w zbiorowości, np. pół samochodu.

Średnia arytmetyczna jest miarą prawidłową tylko do zbiorowości jednorodnych, o niewielkim zróżnicowaniu wartości zmiennej u poszczególnych jednostek. W miarę wzrostu zróżnicowania wartości zmiennych (asymetrii i dyspersji rozkładu), a także w rozkładach Bi- i wielomodalnych należy do opisu stosować przeciętne pozycyjne.

Średnia harmoniczna

Jest odwrotnością średniej arytmetycznej z odwrotności wartości zmiennych:

$$H = \frac{N}{\sum_{i=1}^{N} \frac{1}{x_i}}$$

Średnią harmoniczną stosuje się wówczas, gdy wartości zmiennej podane są w jednostkach względnych („łamanych”), np. km/godz, kg/osobę. Przykładowo można tutaj wymienić:

- prędkość pojazdu
- gęstość zaludnienia
- spożycie artykułu X na głowę ludności.

Na przykład jeżeli turysta jechał rowerem przez 2 godziny z prędkością 15 km/godz., a przez następne 4 godziny z prędkością 9 km/godz. to średnią prędkość jazdy obliczamy za pomocą średniej harmonicznej następująco:

$$H = \frac{N}{\sum_{i=1}^{N} \frac{1}{x_i}} = \frac{2 \cdot 15 + 4 \cdot 9}{\frac{1}{15} \cdot 30 + \frac{1}{9} \cdot 36} = \frac{66}{2+4} = \frac{66}{6} = 11 \text{ km/godz.}$$

Średnia geometryczna

Jest pierwiastkiem *k*–tego stopnia z iloczynu *k* wartości zmiennej, czyli:

$$G = \sqrt[k]{x_1 \cdot x_2 \cdot \ldots\ldots x_k} = \sqrt[k]{\prod_{i=1}^{k} x_i}$$

Średnia geometryczna znajduje zastosowanie przy badaniu średniego tempa zmian zjawisk, których rozwój przedstawiony jest w postaci szeregów dynamicznych.

Dominanta

Dominanta jest to najczęściej powtarzająca się wartość zmiennej w szeregu statystycznym. Określa ona najbardziej typową wartość zmiennej w badanej zbiorowości. Charakterystyczną cechą dominanty jest możliwość jej wyznaczenia zarówno z szeregów dotyczących cechy mierzalnej, jak i nie mierzalnej. Wartość dominanty można jedynie ustalić z rozkładów jednomodalnych.

W szeregach wyliczających i rozdzielczych punktowych dominanta jest tą wartością cech, której odpowiada największa liczebność. W szeregach rozdzielczych przedziałowych bezpośrednio można określić tylko przedział, w którym znajduje się dominanta. Jest to przedział o największej liczebności. Konkretną wartość oblicza się za pomocą wzoru interpolacyjnego.

Kwantyle

Do obliczania kwantyli zbiorowość winna zostać uporządkowana niemalejąco.

Kwartyle – miary dzielące zbiorowość na cztery części.

- **Kwartyl pierwszy** (dolny) dzieli zbiorowość na dwie części w ten sposób, że 25% jednostek zbiorowości ma wartości zmiennej mniejsze lub równe kwartylowi pierwszemu, a 75% - równe lub większe od tego kwartyla
- **Mediana** (kwartl drugi) dzieli zbiorowość na dwie części w ten sposób, że 50% jednostek ma wartości mniejsze lub równe medianie a 50% - równe lub większe od mediany
- **Kwartyl trzeci** (górny) dzieli zbiorowość na dwie części w ten sposób,
  że 75% jednostek zbiorowości ma wartości zmiennej mniejsze lub równe kwartylowi trzeciemu, a 25% - równe lub większe od tego kwartyla

Z szeregów wyliczających (składających się zazwyczaj z niewielkiej liczby jednostek) najczęściej wyznacza się medianę. W przypadku gdy liczba obserwacji jest nieparzysta, mediana jest środkową. Jeśli natomiast liczba jednostek zbiorowości jest parzysta – mediana jest średnią arytmetyczną dwóch środkowych wartości zmiennej.

Kwantyle

Nieparzysta liczba obserwacji

	Wynagrodzenie	
1	1200	
2	**1250**	**I kwintyl**
3	**1300**	**I kwartyl**
4	1340	
5	**1365**	**Mediana**
6	1410	
7	**1700**	**III kwartyl**
8	**2500**	**IV kwintyl**
9	5000	

Liczebność kwartyli = 9*0,25 = 2,25
Liczebność kwintyli = 9*0,2 = 1,8

Parzysta liczba obserwacji

	Wynagrodzenie		
1	1200		
2	1250		**I kwintyl = 1275**
3	**1300**	**I kwartyl**	
4	1340		
5	1365	**Mediana = 1387,5**	
6	1410		
7	1700		
8	**2500**	**III kwartyl**	**IV kwintyl = 3750**
9	5000		
10	12000		

Liczebność kwartyli = 10*0,25 = 2,5
Liczebność kwintyli = 10*0,2 = 2

Miary zmienności

Wartości średnie nie wystarczają do scharakteryzowania struktury zbiorowości. Badana zbiorowość może charakteryzować się różnym stopniem zmienności (rozproszenia, dyspersji).

Dyspersją nazywamy zróżnicowanie jednostek zbiorowości ze względu na wartość badanej cechy.

Podział ze względu na liczbę obserwacji potrzebną do obliczeń:

- Klasyczne miary zmienności oblicza się na podstawie wszystkich wartości badanej cechy:
  - odchylenie standardowe
  - wariancja
  - współczynnik zmienności

Miary zmienności (2)

- Pozycyjne miary zmienności obliczane są na podstawie niektórych (stojących w określonej pozycji) wartości:
  - empiryczny obszar zmienności (zwany też rozstępem)
  - odchylenie ćwiartkowe
  - pozycyjny współczynnik zmienności

Podział ze względu na miano:

- Bezwzględne miary zmienności
  - rozstęp, odchylenie ćwiartkowe, wariancja, odchylenie standardowe
- Względne miary zmienności
  - współczynnik zmienności wyrażony w procentach

Rozstęp

Jest to różnica pomiędzy największą a najmniejszą wartością cechy:

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

Rozstęp kwartylny:

$$R_{kw} = \mathrm{Q}_3 - \mathrm{Q}_1$$

Rozstęp jest miarą pozycyjną i zależy tylko od dwóch wartości. Brakuje zatem informacji o zróżnicowaniu pozostałych jednostek zbiorowości pod względem badanej cechy. Dlatego też rozstęp stosowany jest głównie, gdy potrzebna jest wstępna orientacja o obszarze zmienności cechy.

Odchylenie ćwiartkowe

Oblicza się na podstawie różnicy pomiędzy trzecim i pierwszym kwartylem:

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

Mierzy poziom zróżnicowania jedynie połowy jednostek, pozostałych po odrzuceniu 25% jednostek o wartościach mniejszych od pierwszego kwartyla i większych od trzeciego kwartyla. Miara ta nie jest więc wrażliwa na skrajne wartości zbioru.

Odchylenie ćwiartkowe (2)

Jeżeli w danej zbiorowości do opisu tendencji centralnej użyto mediany, a do opisu zmienności odchylenia ćwiartkowego – to możliwe jest określenie typowego obszaru zmienności badanej cechy:

$$Me - Q < x_{typ} < Me + Q$$

Nietypowe w danej zbiorowości są jednostki o wartości niższej od różnicy $Me - Q$ oraz wyższe od sumy $Me + Q$

Odchylenie ćwiartkowe jest szczególnie przydatne w analizie statystycznej szeregów rozdzielczych przedziałowych o klasach otwartych. Interpretuje się je jako przeciętne zróżnicowanie badanych jednostek wokół mediany.

Wariancja

Wariancja to średnia arytmetyczna kwadratów odchyleń poszczególnych wartości cechy od ich średniej arytmetycznej.

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

dla szeregów wyliczających,

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

dla szeregów rozdzielczych punktowych,

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

dla szeregów rozdzielczych przedziałowych

Cechy wariancji

Jest różnicą między średnią arytmetyczną kwadratów wartości zmiennej i kwadratem jej średniej arytmetycznej czyli:

$$s^2 = \bar{x}^2 - (\bar{x})^2$$

Jeżeli badaną zbiorowość podzielimy na *k* grup, to wariancja ogólna (całej zbiorowości) jest sumą dwóch składników: wariancji wewnątrzgrupowej
i międzygrupowej. Własność ta jest nazywana równością wariancyjną

Wariancja jest wielkością nieujemną $(s^2 \geq 0)$ i mianowaną. Jej mianem jest kwadrat jednostki fizycznej w jakiej mierzona jest badana cecha. Stąd też wariancja jest trudna do merytorycznej interpretacji.

Odchylenie standardowe

W celu otrzymania miary zmienności o mianie zgodnym z mianem badanej cechy, oblicza się dodatni pierwiastek z wariancji. Otrzymana w ten sposób miara nazywa się ***odchyleniem standardowym***:

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

Odchylenie standardowe określa, o ile – średnio biorąc – jednostki zbiorowości różnią się od średniej arytmetycznej badanej zmiennej. Im zbiorowość jest bardziej zróżnicowana, tym wariancja (a więc i odchylenie standardowe) jest większe.

Typowy obszar zmienności

Odchylenie standardowe można wykorzystać do budowy ***typowego obszaru zmienności*** badanej cechy:

$$\bar{x} - s \leq x_{typ} \leq \bar{x} + s$$

Z odchyleniem standardowym wiąże się tzw. reguła trzech sigm. Zgodnie z nią. Wystąpienie obserwacji o wartości cechy spoza przedziału: $(\bar{x} - 3s; \bar{x} + 3s)$ jest mało prawdopodobne.
W przypadku rozkładów o małej asymetrii tylko 0,3% obserwacji wykracza poza ten przedział.
W rozkładach regularnych (symetrycznych, jednomodalnych) ok. 68% obserwacji odchyla się od średniej arytmetycznej o mniej niż jeno odchylenie standardowe, ok. 95% obserwacji odchyla się od średniej arytmetycznej o mniej niż dwa odchylenia standardowe i niemal wszystkie o mniej niż trzy odchylenia standardowe.

Standaryzacja wartości cechy

Jest to przekształcenie pierwotnych wartości cechy w wartości według wzoru:

$$z_i = \frac{x_i - \bar{x}}{s}$$

Wartości standaryzowane informują o tym, o ile odchyleń standardowych pierwotna wartość cechy jest większa lub mniejsza od średniej arytmetycznej. Wartości cechy większe od średniej odpowiada dodatnia wartość zmiennej standaryzowanej, a wartościom niższym – ujemna wartość standaryzowana.
Średnia arytmetyczna zbioru danych standaryzowanych wynosi zero a odchylenie standardowe jeden.
Dane standaryzowane pochodzące z różnych rozkładów mogą być ze sobą porównywalne.

Współczynnik zmienności

Pozwala porównywać zmienność tej samej cechy w różnych zbiorowościach. Jest ilorazem absolutnej miary zróżnicowania i przeciętnego poziomu wartości cechy. Z uwagi, że przy analizie rozkładu posługujemy się różnymi wartościami dyspersji i przeciętnymi, współczynnik zmienności można liczyć:

Klasyczny współczynnik zmienności:

$$V_s = \frac{s}{\bar{x}} \cdot 100$$

Pozycyjne współczynniki zmienności:

$$V_Q = \frac{Q}{Me} \cdot 100 \qquad V_Q = \frac{Q_3 - Q_1}{Q_3 + Q_1} \cdot 100$$

Jeżeli współczynnik zmienności przyjmuje wysokie wartości liczbowe, to fakt ten świadczy o niejednorodności badanej zbiorowości.

Umownie przyjmuje się, że jeżeli $V_s \leq 10\%$ to zbiorowość nie wykazuje dużego zróżnicowania i uznaje się ją za jednorodną.

Miary asymetrii

Badanie asymetrii polega na odpowiedzi na pytanie czy przeważająca liczba jednostek tworzących badaną zbiorowość ma wartości cechy wyższe czy niższe od przeciętnego poziomu. Problem ten wiążę się z oceną kierunku asymetrii (skośności) rozkładu.
Asymetrię rozkładu najłatwiej jest określić przez porównanie takich charakterystyk jak, średnia arytmetyczna, mediana oraz dominanta. W rozkładach symetrycznych średnie te są sobie równe.
Jeśli spełniona jest nierówność:

$\overline{x} > Me > D$ to rozkład charakteryzuje się asymetrią prawostronną (dodatnią)

$\overline{x} < Me < D$ to rozkład charakteryzuje się asymetrią lewostronną (ujemną)

Wskaźnik asymetrii

Służy do określenia kierunku asymetrii a więc stwierdzenia czy jest prawostronna czy lewostronna:

$$W_s = \bar{x} - D$$

Jeśli W_s jest dodatnia, mamy do czynienia z asymetrią prawostronną. W przeciwnym przypadku jest to asymetria lewostronna. W rozkładzie symetrycznym zachodzi: $W_s = 0$, a więc: $\bar{x} = D$

Moment standaryzowany trzeciego rzędu

Miarą określającą kierunek, jak i siłę asymetrii jest współczynnik definiowany za pomocą momentu standaryzowanego trzeciego rzędu:

$$A_s = \frac{m_3}{s^3}$$

Licznik wyraża przeciętną wielkość trzecich potęg odchyleń od średniej arytmetycznej:

$$m_3 = \frac{1}{N}\sum_{i=1}^{N}(x_i - \bar{x})^3$$ dla szeregu wyliczającego

$$m_3 = \frac{1}{N}\sum_{i=1}^{k}(x_i - \bar{x})^3 \cdot n_i$$ dla szeregu rozdzielczego punktowego

$$m_3 = \frac{1}{N}\sum_{i=1}^{k}(\tilde{x}_i - \bar{x})^3 \cdot n_i$$ dla szeregu rozdzielczego przedziałowego

Moment standaryzowany trzeciego rzędu (2)

Mianownik jest trzecią potęgą odchylenia standardowego
W przypadku rozkładów o asymetrii prawostronnej $A_s > 0$,
a $A_s < 0$ lewostronnej. W rozkładach symetrycznych $A_s = 0$.

Im większa jest wartość bezwzględna współczynnika tym silniejsza jest asymetria rozkładu.

Jeżeli asymetria nie jest zbyt silna, to wartość standaryzowanego momentu trzeciego rzędu zawiera się w granicach:

$$-1 \leq A_s \leq +1$$

Jedynie przy ekstremalnie silnej asymetrii, bezwzględna wartość współczynnika asymetrii przekracza 2.

Miary spłaszczenia i koncentracji

Zbiorowość statystyczną analizuje się również ze względu na stopień skupienia poszczególnych wartości cechy wokół średniej arytmetycznej. Skupienie to jest - w dużym stopniu - uzależnione od poziomu dyspersji. Im większe jest zróżnicowanie, tym mniejsze skupienie i odwrotnie. Miarą skupienia poszczególnych wartości cechy wokół jej średniej arytmetycznej jest współczynnik skupienia (kurtoza).

Kurtoza

Współczynnik skupienia (kurtoza) jest standaryzowanym momentem centralnym czwartego rzędu, czyli:

$$K = \frac{m_4}{s^4}$$

gdzie:

m_4 – moment centralny czwartego rzędu, określający przeciętną wielkość czwartych potęg odchyleń wartości cechy od średniej arytmetycznej:

$$m_4 = \frac{1}{N}\sum_{i=1}^{N}(x_i - \bar{x})^4$$ dla szeregu wyliczającego

$$m_4 = \frac{1}{N}\sum_{i=1}^{k}(x_i - \bar{x})^4 \cdot n_i$$ dla szeregu rozdzielczego punktowego

$$m_4 = \frac{1}{N}\sum_{i=1}^{k}(\tilde{x}_i - \bar{x})^4 \cdot n_i$$ dla szeregu rozdzielczego przedziałowego

Kurtoza

Im wyższa wartość współczynnika skupienia, tym krzywa liczebności jest bardziej wysmukła. Oznacza to większe skupienie wartości cechy wokół średniej. Małe wartości współczynnika skupienia wskazują na spłaszczenie rozkładu, a więc mniejsze skupienie wartości cechy wokół średniej arytmetycznej.

Przyjmuje się, że jeśli zbiorowość ma rozkład normalny, to $K = 3$. Jeśli natomiast $K < 3$, to rozkład jest bardziej spłaszczony niż normalny. Taki rozkład nosi nazwę platokurtycznego.
W przypadku, gdy $K > 3$, rozkład empiryczny badanej cechy jest bardziej wysmukły, a skupie nie jest silniejsze od normalnego. Mówimy wówczas o rozkładach leptokurtycznych.

Pojęcie koncentracji

W przypadku cech o charakterze zasobów (powierzchnia ziemi, dochody, kapitał, produkcja itp.) ważne znaczenie ma analiza rozkładu ogólnej sumy wartości badanej cechy (łącznego funduszu cechy) pomiędzy poszczególne jednostki zbiorowości statystycznej. Mówimy wówczas o koncentracji badanego zjawiska. Koncentracja jest bezpośrednio związana z asymetrią i dyspersją badanej cechy. Im silniejsza asymetria i większe zróżnicowanie wartości zmiennej - tym koncentracja jest większa.

Zupełna (całkowita) koncentracja występuje wtedy, gdy łączny fundusz cechy przypada na jedną jednostkę zbiorowości (np. łączny areał powierzchni ziemi w województwie pozostaje w posiadaniu jednego gospodarstwa rolnego).

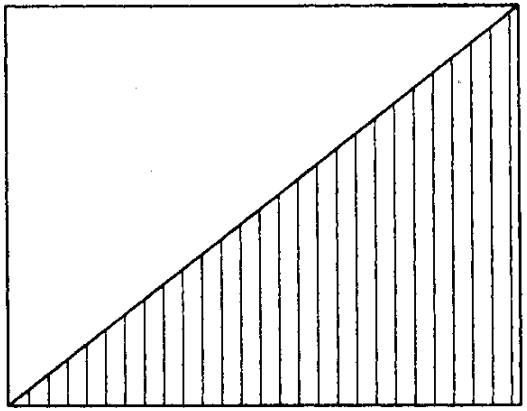
Z brakiem koncentracji mamy do czynienia wówczas, gdy na każdą jednostkę zbiorowości przypada taka sama część ogólnej sumy wartości cechy (np. każdy pracownik w przedsiębiorstwie otrzymuje taką samą część łącznego funduszu płac). W badaniach statystycznych zjawiska braku koncentracji i koncentracji zupełnej raczej nie występują. Najczęściej mamy do czynienia z różnym natężeniem koncentracji.

Ocena koncentracji metodą graficzną

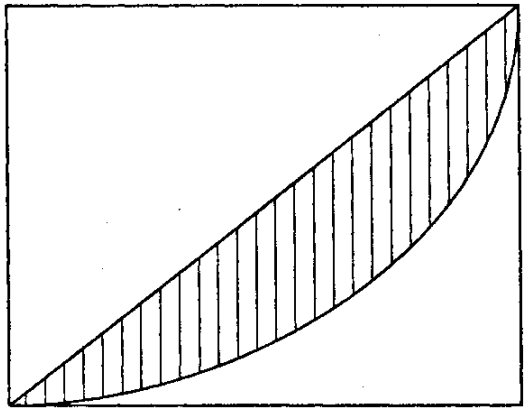
Metoda graficzna polega na wykreśleniu wieloboku koncentracji Lorenza. W tym celu na osi odciętych odmierza się skumulowane części względne liczebności (w %), natomiast na osi rzędnych – procentowe skumulowane częstości względne łącznego funduszu cechy. Łącząc punkty o tych współrzędnych otrzymujemy ***krzywą koncentracji*** (nazywaną też ***krzywą Lorenza***). W przypadku nierównomiernego rozdziału łącznego funduszu cechy pomiędzy jednostki zbiorowości wszystkie punkty leżałyby na przekątnej kwadratu o boku 100. Przekątna tego kwadratu nosi nazwę linii równomiernego rozdziału. Powierzchnia zawarta między linią równomiernego rozdziału a krzywą koncentracji Lorenza jest powierzchnią koncentracji. Im większy jest stopień koncentracji, tym bardziej krzywa Lorenza odchyla się od linii równomiernego rozdziału, a tym samym większa jest powierzchnia koncentracji.

Różne przypadki koncentracji

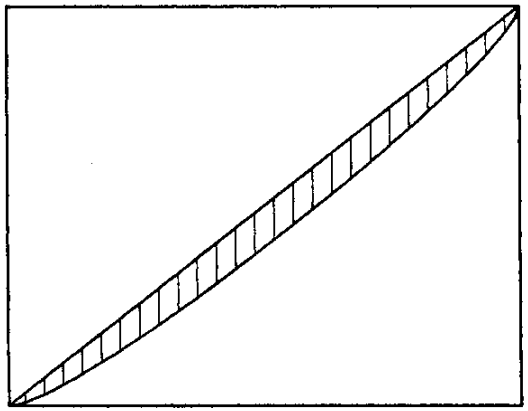
A. Koncentracja całkowita



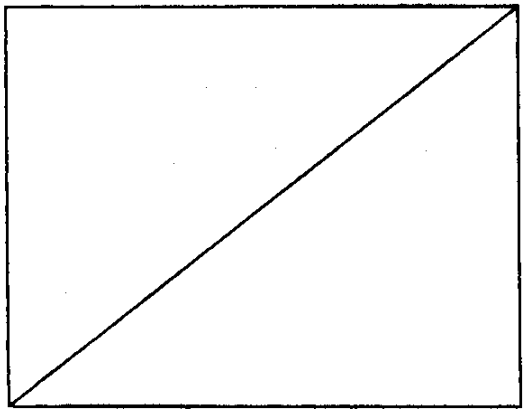
B. Koncentracja duża



C. Koncentracja słaba



D. Brak koncentracji



Źródło: A. Zeliaś, Metody statystyczne, PWE, Warszawa 2000, s. 73.

Współczynnika koncentracji Lorenza

Maksymalna wartość powierzchni koncentracji pozostaje równa połowie kwadratu, tj. 5000, gdyż dwa boki prostokątnego trójkąta równoramiennego mają długość 100, stąd jego pole jest równe 5000.
Stosunek pola zawartego między linią równomiernego rozdziału a krzywą koncentracji do pola połowy kwadratu (pola trójkąta) nosi nazwę współczynnika koncentracji Lorenza. Współczynnik ten ma następującą postać:

$$k = \frac{a}{5000}$$

Współczynnik jest miarą niemianowaną, przyjmującą wartości liczbowe z przedziału: $0 < k < 1$. Przy braku koncentracji $k = 0$, natomiast przy $k = 1$ występuje koncentracja zupełna (całkowita).

Przykład

Osoba	Wynagrodzenie
1	1200
2	1300
3	1400
4	1600
5	1600
6	1600
7	1700
8	1800
9	1800
10	2200
11	2500
12	2700
13	2800
14	2800

Statystyki podstawowe

	Wynagrodzenia
N	14
Średnia	1928,57
Mediana	1750
Dominanta (moda)	1600
Liczność mody	3
Minimum	1200
Maksimum	2800
Dolny	1600
Górny	2500
Rozstęp	1600
Rozstęp kwartylny	900
Odch.Std.	563,54
Skośność	0,54
Kurtoza	-1,20

Wykres Ramka – wąsy