Statystyka Opisowa z Demografią oraz Biostatystyka

Opisowa analiza struktury zjawisk statystycznych

Aleksander Denisiuk

denisjuk@euh-e.edu.pl

Elbląska Uczelnia Humanistyczno-Ekonomiczna

ul. Lotnicza 2

82-300 Elbląg

Opisowa analiza struktury zjawisk statystycznych

Najnowsza wersja tego dokumentu dostępna jest pod adresem

http://denisjuk.euh-e.edu.pl/

Rozkład empiryczny

- przyporządkowanie kolejnym wartościom zmiannej x_j odpowiadających im liczebnośi n_j
  - zamiast liczebności używane są także częstotliwości względne w_j, $w_j = \frac{n_j}{\sum n_k} \left(\frac{n_j}{\sum n_k} \cdot 100\% \right)$
- odzwierciadla strukturę badanej zbiorowości z punktu widzenia określonej cechy
- ustalany na podstawie konkretnych obserwacji

Rozkład empiryczny

- cechy skokowej, cechy ciągłej
  - jednomodalny
  - bimodalny
  - wielomodalny

Rozkład jednomodalny

- symetryczny
- normalny
- asymetryczny
  - prawostronny
  - lewostronny
- zbiorowości *jednorodne*

Rozkład empiryczny

- skrajnie asymetryczny
- siodłowy
- zbiorowości *skrajnie zróżnicowane*

Opisowe charakterystyki

- miary średnie
- miary rozproszenia
- miary asymetrii
- miary koncentracji

Opisowe charakterystyki

- są bardziej syntetycznymi sposobami opisu rozkładów, niż forma graficzna lub tabelaryjna
- pozwalają w sposób syntetyczny określić właściwości badanych rozkładów
- pozwalają porównać:
  - dwie różne zbiorowości pod względem tej samej cechy badania
  - różne cechy tej samej zbiorowości

Miary średnie

- klasyczne
  - średnia arytmetyczna
  - średnia harmoniczna
  - średnia geometryczna
- pozycyjne
  - dominanta (modalna, wartość najczęstsza)
  - kwantyle
    - kwartyle
    - kwintyle
    - decyle
    - centyle (percentyle)

Średnia arytmetyczna

- średnia nieważona (zwykła)
- $\bar{x} = \frac{x_1 + x_2 + \cdots + x_N}{N} = \frac{\sum_{i=1}^{N} x_i}{N}$
- średnia ważona
  - wagi — liczebności wariantów
  - $\bar{x} = \frac{x_1 n_1 + x_2 n_2 + \cdots + x_k n_k}{N} = \frac{\sum_{i=1}^{k} x_i n_i}{N}$

Średnia arytmetyczna. Przykład

- osoba przepracowała w pięciu kolejnych dniach liczbę godzin: 8, 3, 2, 10, 7.
  - średnio 6 godzin
- oblicz średnią arytmetyczną liczby dzieci na utrzymaniu zaobserwowanej w grupie liczącej 82 osób

liczba dzieci	ilość pracowników
0	34
1	26
2	11
3	10
4	1

Średnia arytmetyczna. Przykład, cd

- osoba przepracowała w pięciu kolejnych dniach liczbę godzin: 8, 3, 2, 10, 7.
  - średnio 6 godzin
- oblicz średnią arytmetyczną liczby dzieci na utrzymaniu zaobserwowanej w grupie liczącej 82 osób

liczba dzieci	ilość pracowników	$x_i n_i$
0	34	0
1	26	26
2	11	22
3	10	30
4	1	4

  - średnio 1 dziecko

Szeregi rozdzielcze przedziałowe

- środki przedziałów $\hat{x} = \frac{x_- + x_+}{2}$
- $\bar{x} = \frac{\hat{x}_1 n_1 + \hat{x}_2 n_2 + \cdots + \hat{x}_k n_k}{N} = \frac{\sum_{i=1}^{k} \hat{x}_i n_i}{N}$
- wskaźniki struktury $w_i = \frac{n_i}{N} \cdot 100$
- $\bar{x} = \frac{\sum_{i=1}^{k} \hat{x}_i w_i}{100}$

Szeregi rozdzielcze. Przykład

- średnia liczb podmiotów publicznych w gminach wiejsckich

liczba podmiotów	liczba gmin
5–9	22
10–14	37
15–19	17
20–24	3
25–29	2

Szeregi rozdzielcze. Przykład, cd

- średnia liczb podmiotów publicznych w gminach wiejsckich

x_D–x_G	n_i	$\hat{x}_i$	$\hat{x}_i n_i$
5–9	22	7	154
10–14	37	12	444
15–19	17	17	289
20–24	3	22	66
25–29	2	27	54

  - $\bar{x} = 12,4$

Średnia arytmetyczna

- $\bar{x}_i$ — średnia grupy i
- średnia dla wszystkich grup łącznie: $\bar{\bar{x}} = \frac{\sum_{i=1}^{k} \bar{x}_i n_i}{N}$

Średnia arytmetyczna. Właściwości

- jest wypadkową wszystkich wartości zmiennych, oraz $x_{\min} \leqslant \bar{x} \leqslant x_{\max}$
- suma odchyleń poszczególnych wartości od średniej arytmetycznej jest równa zeru
  - $\sum_{i=1}^{N}(x_i - \bar{x}) = 0$ (szereg wiliczający)
  - $\sum_{i=1}^{k}(x_i - \bar{x})n_i = 0$ (szereg rozdzielczy punktowy)
  - $\sum_{i=1}^{k}(\hat{x}_i - \bar{x})n_i = 0$ (szereg rozdzielczy przedziałowy)
- jeżeli wszystkie wartości pomniejszyć (powiększyć, pomnożyć, podzielić) przez stałą, to średnia arytmetyczna zostanie pomniejszona (powiększona, pomnożona, podzielona) przez tę stałą.

Średnia arytmetyczna. Właściwości, cd

- jeżeli liczebności poszczególnych wariantów cechy są jednakowe, to średnia arytmetyczna równa się ilorazowi sumy wartości wariantów i ich liczby
- suma wartości zmiennej jest równa iloczynowi średniej arytmetycznej i liczebności zbiorowej, $\sum_{i=1}^{N} x_i = N\bar{x}$ (szereg wiliczający)
- jeżeli wszystkie wartości pomniejszyć (powiększyć, pomnożyć, podzielić) przez stałą, to średnia arytmetyczna zostanie pomniejszona (powiększona, pomnożona, podzielona) przez tę stałą.
- na poziom śreniej arytmetycznej silny wpływ wywierają warości ekstremalne

Średnia arytmetyczna. Ograniczenia

- jest miarą prawidłową tylko w odniesieniu do zbiorowości jednorodnych
- w miarę wzrostu asymetrii i zróżnicowania, dla rozkładów bimodalnych i wielomodalnych średnia arytmetyczna traci poznawczą wartość
- nie można obliczyć dla szeregu o przedniałach otwartycch
  - można domykać przedziały otwarte, jeżeli liczba jednostek w nich nie przekracza 5%

Średnia harmoniczna

- jest odwrotnością średniej arytmetycznej odwrotności wartości zmiennych $H = \frac{N}{\sum_{i=1}^{N} \frac{1}{x_i}}$
- dla szeregów rozdzielczych punktowych $H = \frac{N}{\sum_{i=1}^{k} \frac{1}{x_i} n_i}$
- dla szeregów rozdzielczych przedziałowych $H = \frac{N}{\sum_{i=1}^{k} \frac{1}{\hat{x}_i} n_i}$

Średnia harmoniczna

- stosuje się, jeżeli wartości podane są w jednostkach względnych (km/h, kg/osobę), wagi — w jednostkach, występujących w licznikach
  - prędkość (km/h), wagi w km
  - gęstość zaludnienia (obob/km^2), wagi w osobach

Średnia harmoniczna. Przykład

- załóżmy, że gęstość zaludnienia w dwu 60-tysięcznych miastach wynosi odpowiednio 400 osób/km^2 oraz 600 osób/km^2
- jaka jest przeciętna gęstość zaludnienia?
- (odp.: 480 osób/km^2)

Średnia geometryczna

- $\bar{x}_g = \sqrt[n]{x_1 x_2 \dots x_N} = \sqrt[N]{\prod_{i=1}^{N} x_i}$
- $\bar{x}_g = \sqrt[N]{x_1^{n_1} x_2^{n_2} \dots x_k^{n_k}} = \sqrt[N]{\prod_{i=1}^{k} x_i^{n_i}}$
- stosuje się przy badaniu średniego tempa zmian zjawisk

Dominanta (modalna, wartość najczęstsza)

- taka wartość zmiennej, która w danym rozkładzie występuje najczęściej
  - tylko dla rozkładów jednomodalnych
- w szeregach wyliczalnych i rozdzielczych punktowych jest wartością cechy
- w szeregach rozdzielczych przedziałowych można określić tylko przedział
  - konkretna wartość dominanty oblicza się jako
    $D = x_D + \frac{n_D - n_{D-1}}{(n_D - n_{D-1}) + (n_D - n_{D+1})} i_D$
  - albo metodą graficzną
    - rozkład empiryczny jest jednomodalny
    - asymetria rozkładu jest umiarkowana
    - przedział w którym występuje dominanta oraz dwa sąsiadujące mają jednakowe rozpiętości

Dominanta. Przykład

- w przykładzie 12 dominantą jest 0 dzieci
- w przykładzie 14 dominantą jest 12 podmiotów publicznych

Kwantyle

- wartości, które dzielą zbiorowość na kokreślone części pod względem liczby jednostek
  - szewregi muszą być uporządkowane
- kwartyle
- decyle
- centyle (percentyle)

Kwartyle

- kwartyl pierwszy (dolny) —25%
- kwartyl drugi (mediana, wartość środkowa) —50%
- kwartyl trzeci (górny) —75%

Mediana

- szeregi wyliczalne:

$$Me = \begin{cases} x_{\frac{N+1}{2}}, & \text{gdy } N \text{ jest nieparzyste} \\ \frac{1}{2}\left(x_{\frac{N}{2}} + x_{\frac{N}{2}+1}\right), & \text{gdy } N \text{ jest parzyste} \end{cases}$$

- szeregi rozdzielcze punktowe: *kumulacja*

Mediana. Przykład

- czas dojazdu do pracy: 35, 5, 20, 15, 30, 10, 60, 20, 45, 60
  - mediana: 25 minut
- w przykładzie 12
  - mediana: 1 dziecko

Kwartyle. Szeregi rozdzielcze przedziałowe

- $Q_1 = x_{Q_1} + \frac{\frac{N}{4} - \sum_{i=1}^{k-1} n_i}{n_{Q_1}} i_{Q_1}$
- $Q_2 = Me = x_{Me} + \frac{\frac{N}{2} - \sum_{i=1}^{k-1} n_i}{n_{Me}} i_{Me}$
- $Q_3 = x_{Q_3} + \frac{\frac{3N}{4} - \sum_{i=1}^{k-1} n_i}{n_{Q_3}} i_{Q_3}$
- gdzie
  - Q_1, Q_2, Q_3 — odpowiednie kwartyle
  - x_{Q_1}, x_{Me}, x_{Q_3} — dolne granice przedziałów, w których znajdują się odpowiednie kwartyle
  - n_{Q_1}, n_{Me}, n_{Q_3} — liczebności tych przedziałów
  - i_{Q_1}, i_{Me}, i_{Q_3} — rozpiętości przedziałów
  - $\sum_{i=1}^{k-1} n_i$ — sumy liczebności do klasy, w której znajduje się kwartyl

Uwagi o średnich

- kwartyle mogą być wykorzystywane we wszystkich przypadkach
- decyle i centyle oreślane są w sposób podobny
- średnia arytmetyczna, dominanta i mediana są powiązane pewnymi zależnościami
  - w przypadku umiarkowanie asymetrycznego rozkładu $\bar{x} - D = 3(\bar{x} - Me)$

Miary zmienności

- *dyspesja (rozproszenie)* — zróżnicowanie jednostek ze względu na wartości badanej cechy
- miary pozycyjne
  - empiryczny obszar zmienności (rozstęp, amplituda wachań)
  - odchylenie ćwiartkowe
- miary klasyczne
  - odchylenie standardowe
  - wariancja
  - odchylenie przeciętne
- współczynnik zmienności

Miary zmienności

- bezwzględne (absolutne)
  - obszar zmienności
  - wariancja
  - odchylenie stadardowe
  - odchylenie przeciętne
  - odchylenie ćwiartkowe
- względne (relatywne)
  - współczynnik zmienności

Empiryczny obszar zmienności

- $R = x_{\max} - x_{\min}$
  - szereg wyliczalny
  - szereg rozdzilczy — tylko przybliżono
  - przedziały otwarte — niemożliwe
  - wstępna orientacja

Odchylenie przeciętne

- $d = \frac{1}{N} \sum_{i=1}^{N} |x_i - \bar{x}|$
- $d = \frac{1}{N} \sum_{i=1}^{k} |x_i - \bar{x}| n_i$
- $d = \frac{1}{N} \sum_{i=1}^{k} |\hat{x}_i - \bar{x}| n_i$

Odchylenie ćwiartkowe

- $Q = \frac{Q_3 - Q_1}{2}$
- typowy obszar zmienności
  - $Me - Q \leqslant x_{\text{typ}} \leqslant Me + Q$

Wariancja

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

Wariancja. Właściwości

- $s^2 = \overline{x_i^2} - \bar{x}^2$
- jeżeli zbiorowość podzielić na k grup, to

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

- nieujemna i mianowana
- wariancja obliczona na podstawie szeregów rozdzielczych przedziałowych jest *zawyżona*
  - poprawka Shepparda $s^2 = \frac{1}{N} \sum_{i=1}^{k} (\hat{x} - \bar{x})^2 n_i - \frac{i^2}{12}$

Odchylenie standardowe

- $s = \sqrt{s^2}$
- obszar typowy $\bar{x} - s < x_{\text{typ}} < \bar{x} + s$
- odchylenia standardowe, ćwiartkowe oraz przeciętne: $Q < d < s$

Odchylenie standardowe. Właściwości

- obliczane na podstawie wszystkich obserwacji w danym szeregu
- nie zmienia się, jeżeli liczebności szeregu wyrazić w liczbach względnych (procentach)
- nie zmienia się, jeżeli do wszystkich wartości zmiennej dodać pewną stałą
- jeżeli wszystkie wartości zmiennej pomnożyć przez pewną dodatnią stałą, to odchylenie standardowe pomnoży się przez tę samą stałą

Reguła trzech sigm

- wprzypadku rozkładu normalnego (zbliżonego do normalnego)
  - blisko trzecia część obserwacji różni się od średniej arytmetycznej o więcej niż $\pm s$
  - około jedna na 20 obserwacji przekracza tę średnią od wielkość $\pm 2s$
  - tylko jedna na 370 obserwacji przekracza średnią arytmetyczną o $\pm 3s$

Współczynnik zmienności

- miara bezwzględna
- jest ilorazem bezwzględnej miary dyspersji oraz odpowiednich średnich
  - klasyczne:
    - $V_s = \frac{s}{\bar{x}} \cdot 100\%$
    - $V_d = \frac{d}{\bar{x}} \cdot 100\%$
  - pozycyjne:
    - $V_Q = \frac{Q}{Me} \cdot 100\%$
    - $V_{Q_1 Q_3} = \frac{Q_3 - Q_1}{Q_3 + Q_1}$

Współczynnik zmienności. Przykład

- średnie miesięczne wpływy za świadczenie usług nolcegowych w trzech hotelach A, B i C były równe: $\bar{x}_A = 600\,000$ zł., $\bar{x}_B = 300\,000$ zł., $\bar{x}_C = 500\,000$ zł.
- odchylenia standardowe wynosiły $s_A = 110\,000$ zł., $s_B = 90\,000$ zł., $s_A = 120\,000$ zł.
- w którym hotelu występuje najmniejsza dyspersja?
  - $V_s(A) = \frac{110}{600} \cdot 100\% = 18{,}3\%$
  - $V_s(B) = \frac{90}{300} \cdot 100\% = 30{,}0\%$
  - $V_s(C) = \frac{120}{500} \cdot 100\% = 24{,}0\%$

Miary asymetrii

- w rozkładach symetrycznych trzy średnie są równe: $\bar{x} = D = Me$
- jeżeli $x > Me > D$, to rozkład charakteryzuje się *asymetrią prawostronną*
- jeżeli $x < Me < D$, to — *asymetrią lewostronną*

Wskaźnik skośności (asymetrii)

- $W_s = \bar{x} - D$
  - w przypadku symetrii $W_s = 0$
  - w przypadku asymetri lewostronnej $W_s < 0$
  - w przypadku asymetri prawostronnej $W_s > 0$

Wskaźnik skośności a kwartyle

- w przypadku symetrii $(Q_3 - Q_2) - (Q_2 - Q_1) = 0$
- w przypadku asymetri lewostronnej $(Q_3 - Q_2) - (Q_2 - Q_1) < 0$
- w przypadku asymetri prawostronnej $(Q_3 - Q_2) - (Q_2 - Q_1) > 0$

Wskaźnik skośności

- jest bezwzględną miarą aymetrii
- określa jedynie kirunek asymetrii

Współczynnik asymetrii (skośności)

- jest miarą niemieanowaną i unormowaną
  1. $As = \frac{\bar{x}-D}{s}$
  2. $As = \frac{\bar{x}-D}{d}$
  3. $As = \frac{(Q_3-Q_2)-(Q_2-Q_1)}{(Q_3-Q_2)+(Q_2-Q_1)} = \frac{Q_3+Q_1-2Me}{2Q}$
- współczynniki 1 i 2 są wzajemie zamienne
- (pozycyjny) współczynnik 3 jest stosowany, gdy nie można obliczyć dominanty czy średniej arytmetycznej

Współczynnik asymetrii. Przykład

Wiek w latach	Liczba zatrudnionych	
$x_{i-1} - x_i$	n_i	$\hat{x}_i$
15–25	14	20
25–35	32	30
35–45	26	40
45–55	7	50
55–65	3	60
Razem:	82	$\times$

- $D = 32{,}5$
- $As = 0{,}182$

Współczynnik asymetrii. Przedział otwarty

Miasta o liczbie ludności	Liczba miast	Skumulowana liczba miast
$x_{i-1} - x_i$	n_i	$n_s i$
<2 000	43	43
2 000–4 999	235	278
5 000–9 999	181	459
10 000–19 999	179	638
20 000–49 999	139	777
50 000–99 999	51	828
100 000–199 999	22	850
200 000 i więcej	20	870
Razem:	870	$\times$

- $A(Q) = 0{,}463$

Moment centralny rzędu trzeciego

- moment trzeci
- $m_3 = \frac{1}{N} \sum_{i=1}^{N} (x_i - \bar{x})^3 n_i$
  - dla szeregów symetrycznych $m_3 = 0$
  - dla lewostronnej asymetrii $m_3 < 0$
  - dla prawostronnej asymetrii $m_3 > 0$

Moment standardyzowany rzędu trzeciego

- *moment względny*
- $a_3 = \frac{m_3}{s^3}$

Moment trzeci. Przykład

- w przykładzie 12:
  - $\bar{x} = 1$
  - $s = 1{,}07$
  - $m_3 = 1{,}02$
  - $a_3 = 0{,}833$

Miary koncentracji

- nierównomierny podział zjawiska w zbiorowości
  - nierównomierny podział łącznego funduszu cechy pomiędzy poszczególne jednostki zbiorowości
- koncentracja zbiorowości wokół średniej (kurtoza)
- brak koncentracji
- zupełna koncentracja

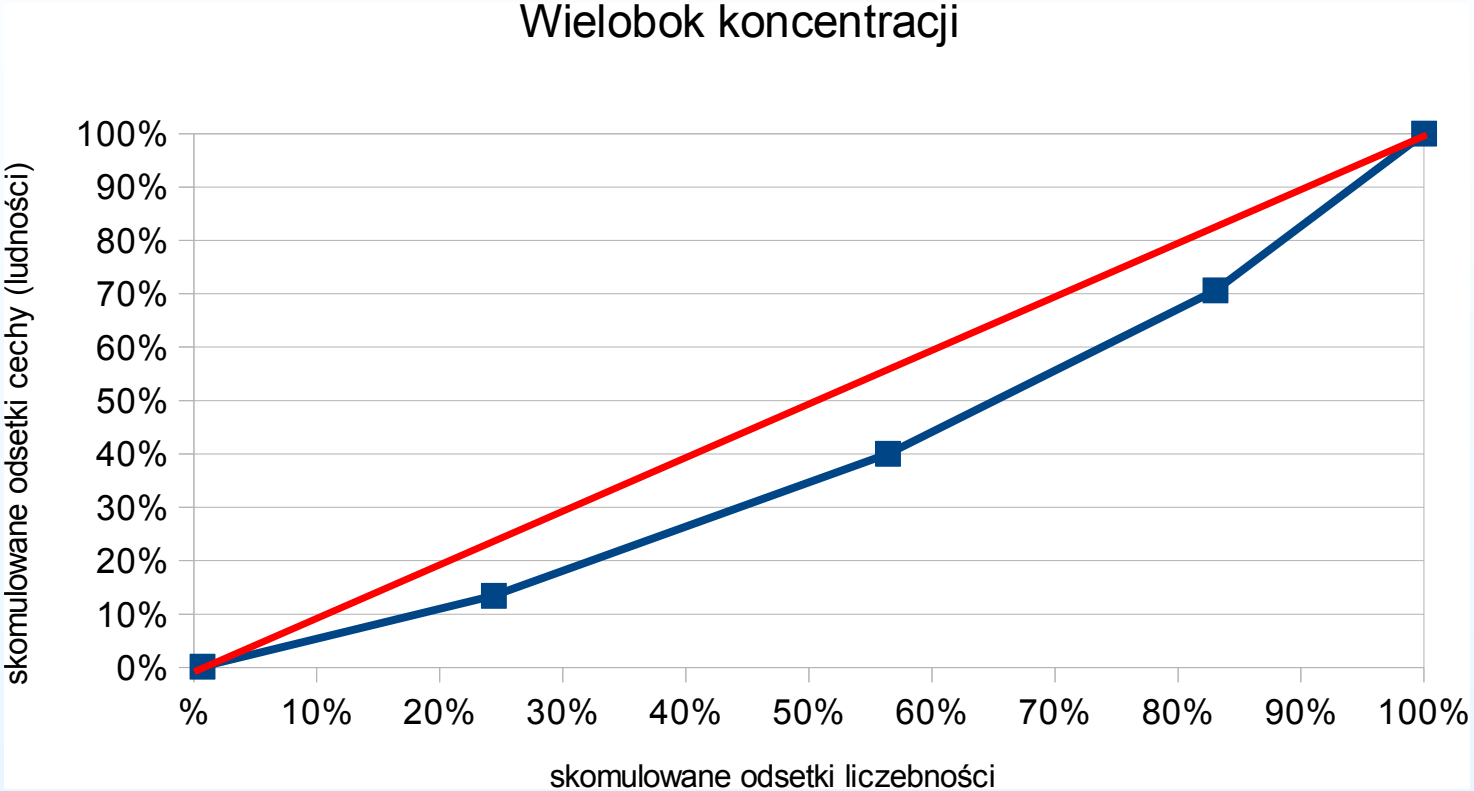
Wielobok koncentracji Lorenza

- na osi odciętych — skumulowane czestości względne (w %)
- na osi rzędnych — procentowe skumulowane częstości względne łącznego funduszu cechy
- krzywa Lorenza
- przekątna kwadratu: *linia równomiernego rozdziału*
- powierzchnia koncentracji
- współczynnik koncentracji Lorenza $k = \frac{a}{5000}$, gdzie a jest polem powierzchni koncentracji
  - jest miarą niemianowaną, $0 \leqslant k \leqslant 1$
  - jeżeli $k = 0$, brak koncentracji
  - jeżeli $k = 1$, to koncentracja zupełna

Wielobok koncentracji. Przykład

Gminy o liczbie ludności (w tys.)	Liczba gmin	Łączna liczba ludności
poniżej 2	15	23,4
2–5	490	1 972,5
5–7	663	3 951,3
7–10	551	4 551,0
powyżej 10	351	4 364,3

Wielobok koncentracji. Przykład



- $a = 1055,395$, $k = 0{,}21$
- koncentracja nie jest duża

Koncentracja obserwacji wokół średniej

- należy porównać rozkład z *normalnym*
- wykres bardziej wysmukły, niż krzywa normalna
  - większe skupienie wartości wokół średniej
  - *leptokurtyczny* rozkład
- wykres bardziej spłaszczony, niż krzywa normalna
  - mniejsza koncentracja wartości wokół średniej
  - *platokurtyczny* rozkład

Miara natężenia koncentracji wokół średniej

- moment centralny czwartego rzędu $m_4 = \frac{1}{N}\sum_{i=1}^{k}(x_i - \bar{x})^4 n_i$
- standardyzowany moment centralny czwartego rzędu $a_4 = \frac{m_4}{a_4}$
  - dla rozkładu normalnego $a_4 = 3$
  - dla rozkładu spłaszczonego $a_4 < 3$
  - dla rozkładu wysmukłego $a_4 > 3$
- dla rozkładów jednomodalnych określany jest *eksces*: $a_4 - 3$