Statystyka opisowa

Prof. PK Dr Marek Malinowski

Udostępnione prezentacje z wykładu są wyłącznie do użytku osobistego z zakazem rozpowszechniania w jakikolwiek sposób przy użyciu jakiegokolwiek środku przekazu.

Charakterystyki liczbowe

Analiza danych statystycznych powinna prowadzić do zwięzłego przedstawienia wyników badań za pomocą odpowiednich charakterystyk liczbowych. Te charakterystyki dzielimy na:

1. miary położenia,
2. miary zmienności (rozrzutu, dyspersji),
3. miary asymetrii (skośności),
4. miary koncentracji (skupienia).

Miary położenia

Miary położenia charakteryzują średni poziom wartości zmiennej (badanej cechy), czyli są to takie wskaźniki liczbowe wokół których leżą pozostałe wartości badanej cechy. Dają odpowiedź na pytanie: Gdzie jest środek?

1. **Klasyczne miary położenia** (każda zmiana dowolnego elementu badanego zbioru pociąga za sobą zmianę wartości miary).

1.1. **Średnia arytmetyczna**:

a) dla szeregu szczegółowego $x_1, x_2, \ldots, x_n$

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

b) dla szeregu punktowego

x_i	x_1	$\cdots$	x_k
n_i	n_1	$\cdots$	n_k

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

c) dla szeregu klasowego wybieramy najpierw środki y_i klas $[x_{i-1}, x_i)$ jako reprezentantów i obliczamy

$$\bar{x} = \frac{1}{n} \sum_{i=1}^{k} y_i \cdot n_i$$

Np. dla szeregu klasowego

wartość	$[0,2)$	$[2,4)$	$[4,6)$	$[6,8)$	$[8,10)$	$[10,12]$
n_i	9	28	42	30	15	6
y_i	1	3	5	7	9	11

$\sum_{i=1}^{k} n_i = 130$, $\sum_{i=1}^{k} y_i \cdot n_i = 714$. Zatem $\bar{x} = \frac{714}{130} \approx 5,5$.

Średnia arytmetyczna jest wrażliwa na na skrajne wartości cechy, czyli na tzw. wartości odstające, wyraźnie oddalone od innych wartości i tym samym nietypowe, również na wartości przypadkowe, wynikające z błędnych pomiarów.

Średnia arytmetyczna z próby reprezentatywnej **jest dobrym przybliżeniem wartości przeciętnej w populacji generalnej**.

1.2. **Średnia harmoniczna**.

Średnią harmoniczną **stosujemy, gdy wartości cechy podane są w postaci wskaźników natężenia**, np. gęstość zaludnienia w os/km^2, spożycie w kg/os , wydajność pracy w szt/min.

a) dla szeregu szczegółowego $\bar{x}_H = \frac{n}{\sum_{i=1}^{n} \frac{1}{x_i}}$.

b) dla szeregu punktowego $\bar{x}_H = \frac{n}{\sum_{i=1}^{k} \frac{1}{x_i} \cdot n_i}$,

c) dla szeregu klasowego $\bar{x}_H = \frac{n}{\sum_{i=1}^{n} \frac{1}{y_i} \cdot n_i}$.

Np. Pierwszy pracownik wykonuje detal w ciągu 4 minut, drugi pracownik w ciągu 6 minut, a trzeci potrzebuje 12 minut. Ile czasu potrzebują średnio ci pracownicy, aby wykonać jeden detal? Wtedy $\bar{x}_H = \frac{3}{\frac{1}{4}+\frac{1}{6}+\frac{1}{12}} = 6$.

2. **Miary pozycyjne położenia**.

2.1. **Moda** (modalna, dominanta) to wartość cechy, która występuje najczęściej.

a) i b) dla szeregów szczegółowego i punktowego jest to wartość, która ma największą liczebność,

c) dla szeregu rozdzielczego klasowego można wskazać klasę, w której występuje moda, bo jest to klasa o największej liczebności. Przybliżoną wartość mody oblicza się ze wzoru

$$Mo = x_{m-1} + \frac{n_m - n_{m-1}}{(n_m - n_{m-1}) + (n_m - n_{m+1})} l_m,$$

gdzie m - numer klasy, w której występuje moda,
x_{m-1} - dolna granica klasy, w której występuje moda,
n_m - liczebność przedziału mody, n_{m-1} - liczebność przedział przed modą, n_{m+1} - liczebność przedziału po modzie,
l_m - długość przedziału, gdzie jest moda.

Np. odnosząc się do wcześniejszego szeregu przedziałowego widzimy, że $m = 3$, $x_{m-1} = 4$, $n_m = 42$, $n_{m-1} = 28$, $n_{m+1} = 30$, $l_m = 2$. Zatem

$$Mo = 4 + \frac{42 - 28}{(42 - 28) + (42 - 30)} \cdot 2 \approx 5,08.$$

Uwagi:

1. Wyznaczanie mody ma sens, gdy jest wyraźnie zaznaczone jedno maksimum,
2. Przedział mody i dwa sąsiednie powinny mieć tę samą długość,
3. Jeżeli klasą o największej liczebności jest klasa skrajna, to w zasadzie mody się nie wyznacza.

2.2. **Mediana** - wartość, która dzieli uporządkowany niemalejąco zbiór danych na dwie części tak, że co najmniej połowa jednostek ma wartość cechy nie większą od niej i równocześnie co najmniej połowa jednostek ma wartość cechy nie mniejszą od tej wartości.

a) i b) dla szeregu szczegółowego i punktowego

$$Me = \begin{cases} x_{\frac{n+1}{2}} & \text{, gdy} \quad n \text{ nieparzyste,} \\ \frac{x_{\frac{n}{2}} + x_{\frac{n}{2}+1}}{2} & \text{, gdy} \quad n \text{ parzyste.} \end{cases}$$

c) dla szeregu przedziałowego

$$Me = x_{m-1} + \frac{\frac{n}{2} - \sum_{i=1}^{m-1} n_i}{n_m} \cdot l_m,$$

gdzie m - numer przedziału mediany,

x_{m-1} - dolny kraniec klasy, w której znajduje się mediana,

n_m - liczebność przedziału mediany,

l_m - długość przedziału mediany.

Np. wracając do wcześniejszego szeregu przedziałowego mamy $n = 130$, $m = 3$, $x_2 = 4$, $n_3 = 42$, $\sum_{i=1}^{3-1} n_i = 37$, $l_3 = 2$. Zatem $Me = 4 + \frac{\frac{130}{2} - 37}{42} \cdot 2 \approx 5,3$.

Uwaga: Cechą mediany jest brak wrażliwości na skrajne wartości.

2.3. **Kwartyle**.

Kwartyl dolny (pierwszy) Q_1 dzieli zbiorowość na dwie części w ten sposób, że 25% jednostek ma wartości cechy co najwyżej równe Q_1, a 75% co najmniej równe temu kwartylowi.

Kwartyl górny (trzeci) Q_3 dzieli zbiorowość na dwie części w ten sposób, że 75% jednostek ma wartości cechy co najwyżej równe Q_3, a 25% co najmniej równe temu kwartylowi.

a) i b) dla szeregów szczegółowych i punktowych Q_1 to mediana dla podzbioru obserwacji, które są mniejsze, bądź równe medianie. Natomiast Q_3 to mediana dla podzbioru obserwacji, które są większe, bądź równe medianie.

c) dla szeregów przedziałowych kwartyle Q_1 i Q_3 wyznaczamy ze wzorów

$$Q_1 = x_{m-1} + \frac{\frac{n}{4} - \sum_{i=1}^{m-1} n_i}{n_m} \cdot l_m,$$

gdzie m - numer przedziału kwartyla pierwszego,
x_{m-1} - dolny kraniec klasy, w której znajduje się kwartyl pierwszy,
n_m - liczebność przedziału kwartyla pierwszego,
l_m - długość przedziału kwartyla pierwszego,

$$Q_3 = x_{m-1} + \frac{\frac{3n}{4} - \sum_{i=1}^{m-1} n_i}{n_m} \cdot l_m,$$

gdzie m - numer przedziału kwartyla trzeciego,
x_{m-1} - dolny kraniec klasy, w której znajduje się kwartyl trzeci,
n_m - liczebność przedziału kwartyla trzeciego,
l_m - długość przedziału kwartyla trzeciego.

Miary zmienności

Pozwalają one ocenić **poziom zróżnicowania jednostek badanej populacji ze względu na badaną cechę**. Ograniczenie się do miar położenia nie wystarcza, bo może się zdarzyć, że w dwóch zbiorowościach średnie arytmetyczne są takie same, ale rozproszenie danych jest inne w każdej z tych zbiorowości.

1. **Miary zmienności pozycyjne**.

1.1. **Rozstęp** R charakteryzuje empiryczny obszar zmienności badanej cechy

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

1.2. **Odchylenie ćwiartkowe** Q mierzy poziom zróżnicowania środkowych jednostek, tzn. pozostałych po odrzuceniu 25% jednostek o najmniejszych i największych wartościach

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

Typowy obszar zmienności cechy

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

2. **Miary zmienności klasyczne**.

2.1. **Wariancja s_n^2 z próby**.

a) dla szeregu szczegółowego

$$s_n^2 = \frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})^2 \text{ albo } s_{n-1}^2 = \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{x})^2,$$

b) dla szeregu punktowego

$$s_n^2 = \frac{1}{n}\sum_{i=1}^{k}(x_i - \bar{x})^2 n_i \text{ albo } s_{n-1}^2 = \frac{1}{n-1}\sum_{i=1}^{k}(x_i - \bar{x})^2 n_i,$$

c) dla szeregu przedziałowego

$$s_n^2 = \frac{1}{n}\sum_{i=1}^{k}(y_i - \bar{x})^2 n_i \text{ albo } s_{n-1}^2 = \frac{1}{n-1}\sum_{i=1}^{k}(y_i - \bar{x})^2 n_i.$$

Mianem wariancji jest kwadrat jednostki, w której mierzona jest badana cecha (np. kg^2). Zatem jest ona trudna w interpretacji. Aby uzyskać miano zgodne z jednostką mierzenia obliczamy odchylenie standardowe s.

2.2. **Odchylenie standardowe** s **z próby** określa średnie zróżnicowanie wartości badanej cechy wokół $\bar{x}$.

$$s = \sqrt{s_n^2} \text{ albo } s = \sqrt{s_{n-1}^2}.$$

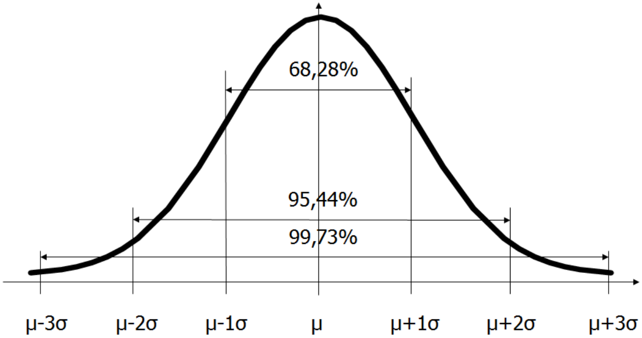
Typowy obszar zmienności

$$\bar{x} - s < x_{typ} < \bar{x} + s.$$

W tym obszarze mieszczą się wartości badanej cechy dla około $\frac{2}{3}$ wszystkich jednostek badanej zbiorowości.

Uwaga:
Z odchyleniem standardowym wiąże się **reguła trzech sigm**, na mocy której dla rozkładów o niewielkiej asymetrii

1. około 32% (około $\frac{1}{3}$) obserwacji jest poza przedziałem $(\bar{x} - s, \bar{x} + s)$,
2. około 5% obserwacji jest poza przedziałem $(\bar{x} - 2s, \bar{x} + 2s)$,
3. około 0,3% obserwacji jest poza przedziałem $(\bar{x} - 3s, \bar{x} + 3s)$.



2.3. **Współczynnik zmienności klasyczny**

$$V = \frac{s}{\bar{x}} \cdot 100\%.$$

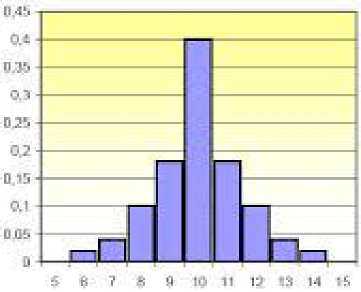
2.3. **Współczynnik zmienności pozycyjny**

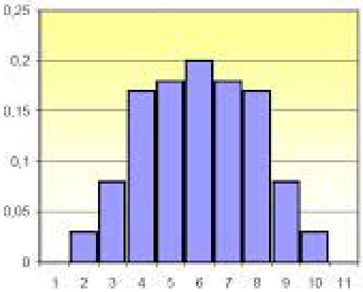
$$V = \frac{Q}{Me} \cdot 100\%.$$

Uwaga:
1. Przyjmuje się, że jeśli $V < 10\%$, to cecha wykazuje zróżnicowanie statystycznie nieistotne. Natomiast duże wartości V wskazują na istotne zróżnicowanie, czyli **niejednorodność** zbiorowości.
2. Ten współczynnik **stosuje się zwykle w porównaniach**, gdy chcemy ocenić zróżnicowanie:
a) kilku zbiorowości pod względem tej samej cechy,
b) tej samej zbiorowości pod względem kilku różnych cech.

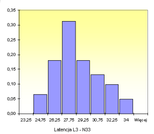
Do pełniejszej charakteryzacji danej zbiorowości używa się oprócz miar położenia i zmienności kolejnych miar zwanych miarami asymetrii.
Miary asymetrii pozwalają stwierdzić, **czy większa część populacji klasuje się powyżej, czy poniżej przeciętnego poziomu** badanej cechy. Asymetrię rozkładu można zbadać porównując modę Mo, medianę Me i średnią $\bar{x}$.

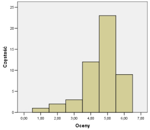
W przypadku rozkładu **symetrycznego** wszystkie te parametry są równe, tzn. $Mo = Me = \bar{x}$.



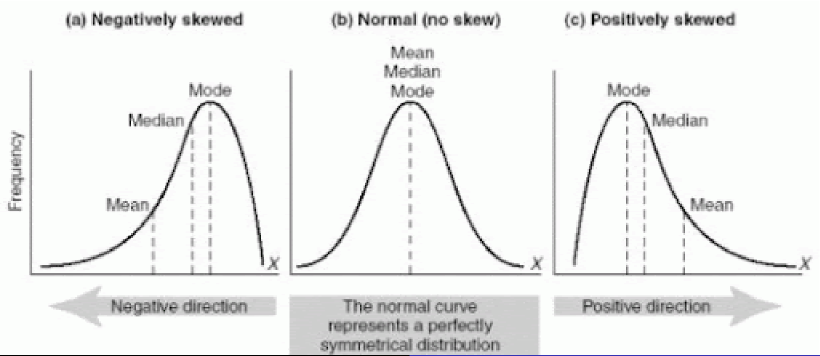


Jeśli zachodzi nierówność $Mo < Me < \bar{x}$ to rozkład jest asymetryczny prawostronnie (dodatnio). $\rightarrow$

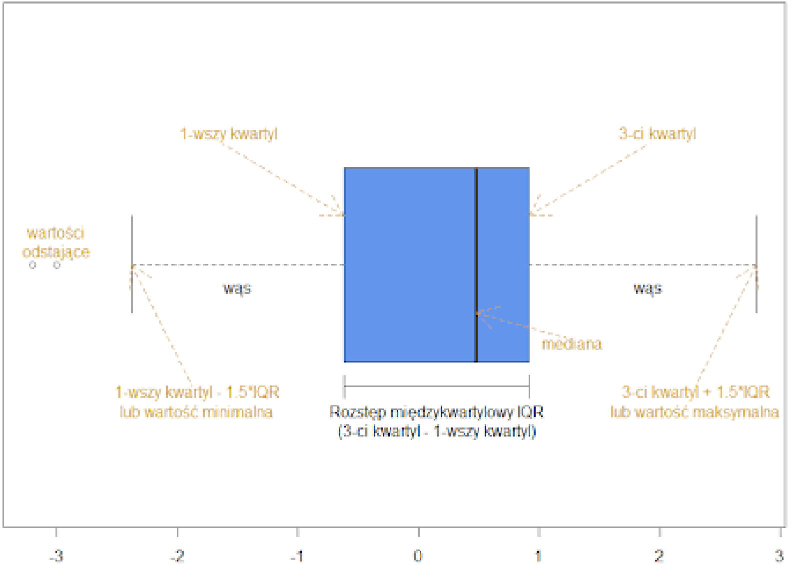


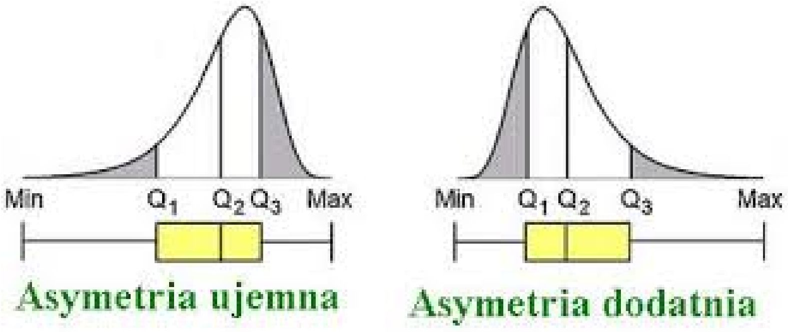


$\leftarrow$ Jeśli zaś zachodzi nierówność $\bar{x} < Me < Mo$ to rozkład jest asymetryczny lewostronnie (ujemnie).



Wykres pudełkowy. Z niego również można odczytać asymetrię. Na rysunku obok widoczna jest asymetria lewostronna (ujemna). Ilustracja pokazuje jak narysować wykres pudełkowy.





1. **Współczynniki asymetrii (skośności)** służą do określenia siły i kierunku asymetrii.

1.1. **Współczynnik asymetrii pozycyjny**.

$$A = \frac{Q_3 + Q_1 - 2Me}{Q_3 - Q_1} \text{ określa asymetrię jednostek środkowych,}$$

tzn. jednostek między Q_1 a Q_3. Na ogół $A \in (-1, 1)$. Jeśli asymetria nie jest silna, to $|A| \in (0, \frac{1}{2}]$.

1.2. **Współczynnik asymetrii klasyczny**.

$$A = \frac{m_3}{s^3} \text{ (standaryzowany moment centralny rzędu 3)}$$

a) dla szeregu szczegółowego $m_3 = \frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})^3$,

b) dla szeregu punktowego $m_3 = \frac{1}{n}\sum_{i=1}^{k}(x_i - \bar{x})^3 n_i$,

c) dla szeregu przedziałowego $m_3 = \frac{1}{n}\sum_{i=1}^{k}(y_i - \bar{x})^3 n_i$.

$A = 0$ - symetria rozkładu,
$A > 0$ - asymetria prawostronna (dodatnia),
$A < 0$ - asymetria lewostronna (ujemna).

Uwaga:
Jeśli asymetria nie jest zbyt silna, to wartość bezwzględna z A danego w punkcie 1.2 przyjmuje wartości z przedziału $(0, 2]$.

Miary koncentracji (skupienia)

Można wyróżnić 2 rodzaje koncentracji:

1. Koncentrację rozumianą jako **skupienie wartości poszczególnych jednostek wokół średniej**, czyli stopień spłaszczenia rozkładu.

2. Koncentracja rozumiana jako nierównomierny podział zjawiska w zbiorowości, a dokładniej **nierównomierny podział sumy wartości badanej cechy na poszczególne jednostki**, np. dużo dóbr luksusowych w niewielkiej liczbie gospodarstw domowych.

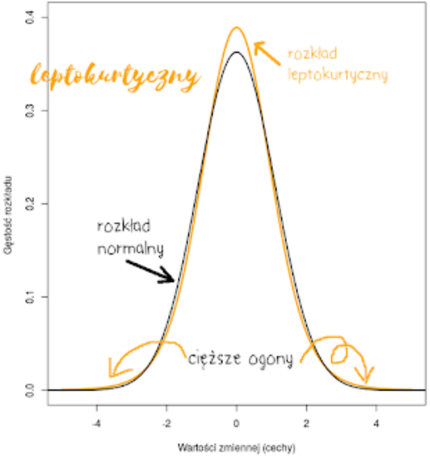
Ad. 1. Współczynnik koncentracji, standaryzowany moment centralny rzędu 4, tzw. **kurtoza** $K = \frac{m_4}{s^4}$

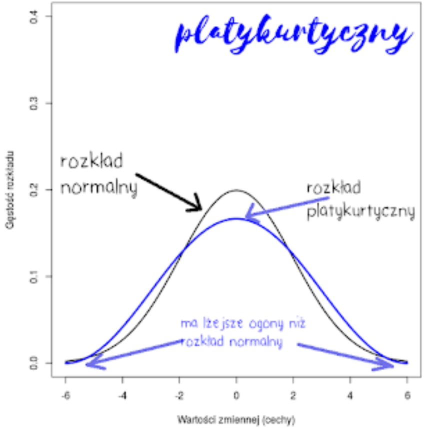
a) dla szeregu szczegółowego $m_4 = \frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})^4$,

b) dla szeregu punktowego $m_4 = \frac{1}{n}\sum_{i=1}^{k}(x_i - \bar{x})^4 n_i$,

c) dla szeregu przedziałowego $m_4 = \frac{1}{n}\sum_{i=1}^{k}(y_i - \bar{x})^4 n_i$.

Uwaga: Jeśli badana cecha ma tzw. **rozkład normalny**, to $K = 3$ (mezokurtyczny). Jeśli rozkład jest bardziej wysmukły, tzn. o skupieniu silniejszym niż w rozkładzie normalnym, to $K > 3$ i wartości mają tendencję do skupiania się wokół średniej (leptokurtyczny). Jeśli $K < 3$, to rozkład jest bardziej spłaszczony (platykurtyczny).





Ad. 2. (Informacyjnie) Analiza nierównomiernego podziału sumy wartości badanej cechy pomiędzy poszczególne jednostki polega na skonstruowaniu tzw. krzywej Lorenza i obliczeniu współczynnika Lorenza, zwanego również współczynnikiem Giniego G.

$G \in [0, 1]$, przy czym jeśli G jest liczbą z przedziału:
$(0, 0.3]$ - słaba koncentracja,
$(0.3, 0.6]$ - umiarkowana koncentracja,
$(0.6, 1)$ - silna koncentracja.

Przypadki skrajne:
$G = 0$ dla braku koncentracji, tzn. na każdą jednostkę zbiorowości przypada taka sama część ogólnej sumy wartości cechy,
$G = 1$ dla całkowitej koncentracji, tzn. ogólna suma wartości cechy przypada tylko jednej jednostce statystycznej.