Statystyka opisowa

Statystyka opisowa różni się od statystyki matematycznej. W statystyce matematycznej wykonujemy badania dla próby statystycznej a wnioskujemy na temat całej populacji. W statystyce opisowej analizuje się tylko zależności w próbie.

Wyróżniamy trzy kategorie statystyk opisowych

1. miary położenia: np. średnia, dominanta (moda), kwantyle.

2. miary dyspersji (rozproszenia, zróżnicowania): np. wariancja, odchylenie standardowe,

odchylenie przeciętne.

3. miary kształtu rozkładu: np. skośność, kurtoza.

Wyróżniamy charakterystyki:

1. klasyczne - wyznaczone przez wszystkie wartości danych statystycznych, np. średnia, wariancja
2. pozycyjne - wyznaczone przez niektóre wartości danych, np. mediana, dominanta,

Dane można prezentować w postaci:

1. Szeregu prostego (stosujemy w przypadku małej liczby danych) – prezentowane są wszystkie wartości, np.: wyniki badania wzrostu w pewnej drużynie siatkówki:
   188, 175, 199, 202, 188, 194, 195, 195, 202, 193, 203, 188, 201, 193, 197, 198, 185, 195, 194, 202.

2. Szeregu rozdzielczego punktowego (stosujemy gdy dane się powtarzają), np.:

Wzrost	Liczba wystąpień
175	1
176	2
188	4
191	1
193	1
194	2
195	2
197	3
198	3
199	3
202	1
203	2
209	2

3. Szeregu rozdzielczego przedziałowego (stosujemy gdy danych jest dużo i się nie powtarzają), np.:

Wzrost	Liczba wystąpień
175-190	7
191-200	15
201-210	5

Liczba przedziałów i ich zakres jest zwykle ustalana zdroworozsądkowo w zależności od specyfiki danych.

Statystyki dla szeregu prostego (szczegółowego)

Miary położenia

Średnia arytmetyczna

$$\overline{X} = \frac{\sum_{i=1}^{n} x_i}{n}$$

W Excelu:

ŚREDNIA(liczba1;liczba2;...)

Liczba1; liczba2;... : od 1 do 255 argumentów liczbowych lub tablica

Dominanta - najczęściej występująca wartość

$$D(x) = x_{i\ (i:n_i=\max\limits_{i}(n_i))}$$

W Excelu:

WYST.NAJCZĘŚCIEJ(liczba1;liczba2;...)

Liczba1; liczba2;... : od 1 do 255 argumentów liczbowych lub tablica

Uwaga: dla danego zbioru danych dominanta może nie występować

Kwantyle - wartości cechy dzielące próbę w określonym stosunku

Mediana (wartość środkowa) - dzieli próbę w stosunku 1:1 – 50% obserwacji posiada wartość cechy nie większą niż mediana i 50% obserwacji posiada wartość nie mniejszą niż mediana.

Jeśli dane są uporządkowane czyli $x_1 \leq x_2 \leq \ldots \leq x_n$

$$m_e = \begin{cases} x_{\frac{n+1}{2}} & \text{dla nieparzystych} \\ \frac{1}{2}\left(x_{\frac{n}{2}} + x_{\frac{n+2}{2}} \right) & \text{dla parzystych} \end{cases}$$

W Excelu:

MEDIANA(liczba1;liczba2;...)

Liczba1; liczba2;... : od 1 do 255 argumentów liczbowych lub tablica

Kwartyl 1 – dzieli próbę na 25% obserwacji posiadających wartość cechy nie większą niż pierwszy kwartyl oraz 75% posiadających wartość nie mniejszą niż on.

$$q_1 = \begin{cases} x_{\frac{n+1}{4}} & \text{dla } n = 4k+3 \\ x_{\frac{n+2}{4}} & \text{dla } n = 4k+2 \\ \frac{1}{2}\left(x_{\frac{n+3}{4}-1} + x_{\frac{n+3}{4}} \right) & \text{dla } n = 4k+1 \\ \frac{1}{2}\left(x_{\frac{n}{4}} + x_{\frac{n}{4}+1} \right) & \text{dla } n = 4k \end{cases}$$

Kwartyl 3 – dzieli próbę na 75% obserwacji posiadających wartość cechy nie większą niż trzeci kwartyl oraz 25% posiadających wartość nie mniejszą niż on.

$$q_3 = \begin{cases} x_{\frac{3n+3}{4}} & \text{dla } n = 4k+3 \\ x_{\frac{3n+2}{4}} & \text{dla } n = 4k+2 \\ \frac{1}{2}\left(x_{\frac{3n+1}{4}} + x_{\frac{3n+5}{4}} \right) & \text{dla } n = 4k+1 \\ \frac{1}{2}\left(x_{\frac{3n}{4}} + x_{\frac{3n}{4}+1} \right) & \text{dla } n = 4k \end{cases}$$

W Excelu:

KWARTYL(tablica;kwartyl)
tablica: zakres danych dla których liczymy kwartyl

kwartyl: numer kwartyla

kwartyl	KWARTYL
0	Wartość minimalna
1	Pierwszy kwartyl
2	Wartość mediany
3	Trzeci kwartyl
4	Wartość maksymalna

Miary zróżnicowania

Wariancja - wskazuje jak bardzo wartości analizowanego zbioru rozrzucone są wokół średniej

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

Wariancja podawana jest w jednostkach kwadratowych dlatego jej interpretacja sprawia pewne kłopoty.

W Excelu:

WARIANCJA.POPUL(liczba1;liczba2;...)

Liczba1; liczba2;... : od 1 do 255 argumentów liczbowych lub tablica

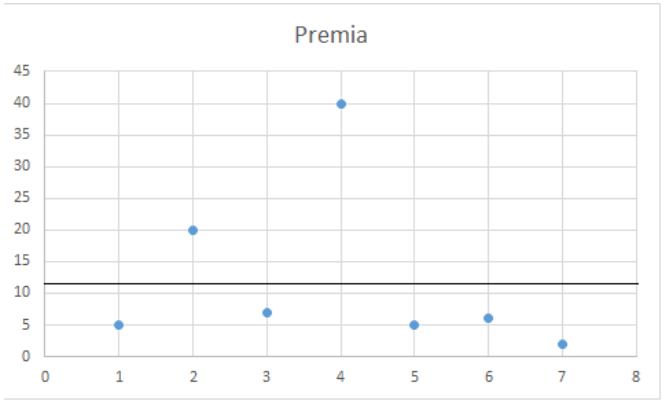
Przykład

Firma 1

Pracownik	Premia [%]
1	5
2	20
3	7
4	40
5	5
6	6
7	2

Średnia 12

$S^2 = [(5-12)^2+(20-12)^2+(7-12)^2+(40-12)^2+(5-12)^2+(6-12)^2+(2-12)^2]/7=158$

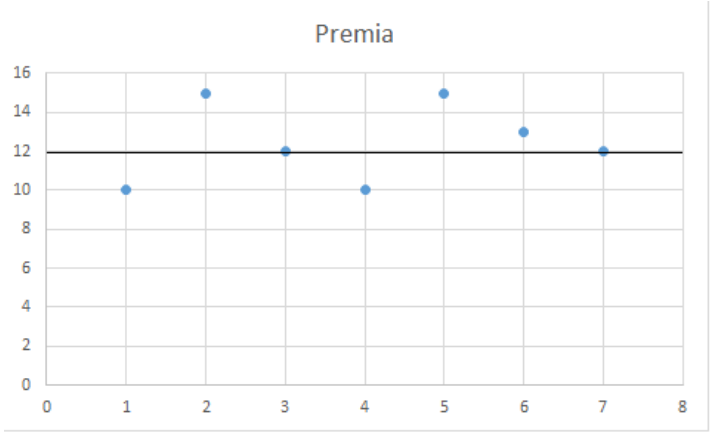


Firma 2

Pracownik	Premia [%]
1	10
2	15
3	12
4	10
5	15
6	13
7	12

Średnia 12

$S^2 = 3,7$



Odchylenie standardowe- wskazuje jak bardzo wartości analizowanego zbioru rozrzucone są wokół średniej

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

Odchylenie standardowe jest łatwiejsze w interpretacji niż wariancja

W Excelu:

ODCH.STANDARD.POPUL(liczba1;liczba2;...)
Liczba1; liczba2;... : od 1 do 255 argumentów liczbowych lub tablica

Przykład

Firma 1

S=12

Firma 2

S=1,92

Odchylenie przeciętne - wartość średnia odchyleń bezwzględnych punktów danych od ich wartości średniej.

$$s_p = \frac{1}{n}\sum_{i=1}^{n}\left|x_i - \bar{x}\right|$$

W Excelu:

ODCH.ŚREDNIE(liczba1;liczba2;...)
Liczba1; liczba2;... : od 1 do 255 argumentów liczbowych lub tablica

Przykład

Firma 1

S_p = [|5-12|+|20-12|+|7-12|+|40-12|+|5-12|+|6-12|+|2-12|]/7=10,2

Firma 2

S_p=1,63

Współczynnik zmienności - mierzy zróżnicowanie względne i określa jaką część (ile procent) przeciętnego poziomu badanej cechy stanowi odchylenie standardowe.

$$v = \frac{s}{\overline{x}}$$

Przykład:

Firma 1

v=1 (100%)

Firma 2

v=0,16 (16%)

Przedział typowych wartości – należy do niego większość danych, interpretacja ta jest uzasadniona wtedy gdy cecha ma rozkład zbliżony do rozkładu normalnego.

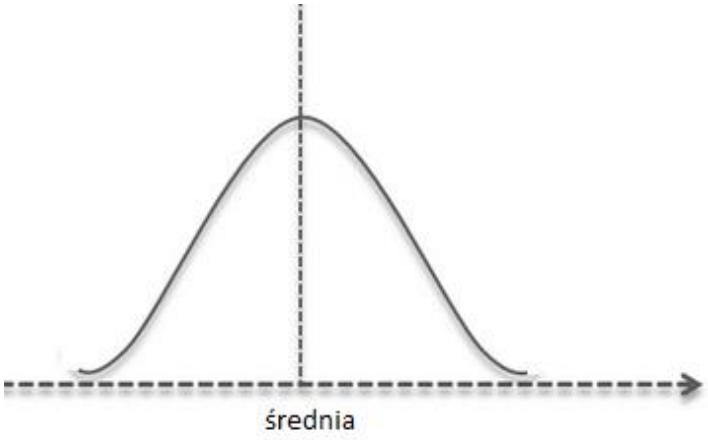
$$\left[\bar{x} - s, \bar{x} + s\right]$$

Rozstęp - najprostsza i najbardziej intuicyjna miara rozproszenia, różnica między największą i najmniejszą wartością występującą w analizowanym zbiorze danych.

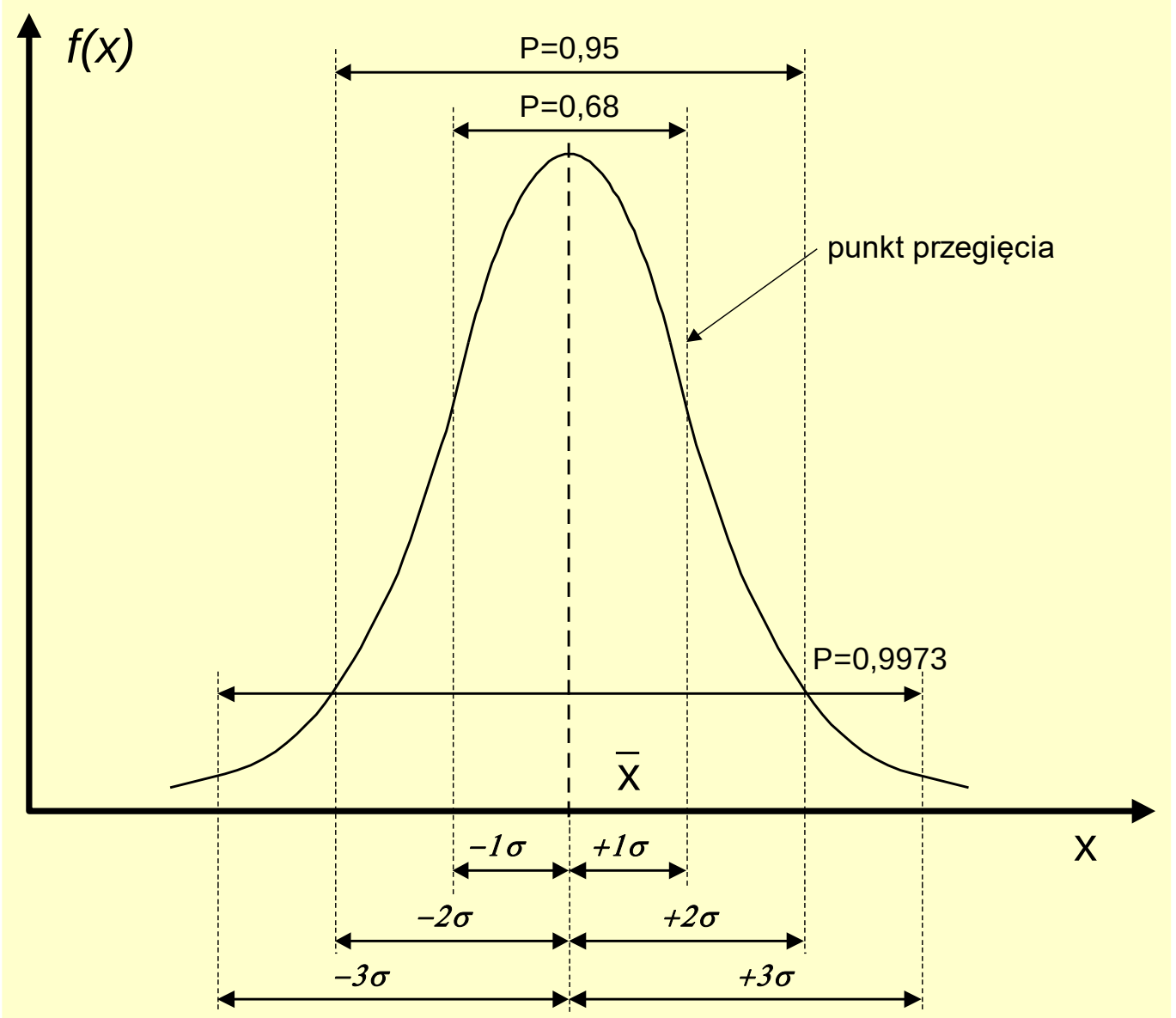
$$r_0 = x_{\max} - x_{\min}$$

Miary kształtu rozkładu

Rozkład normalny – wzorzec rozkładu

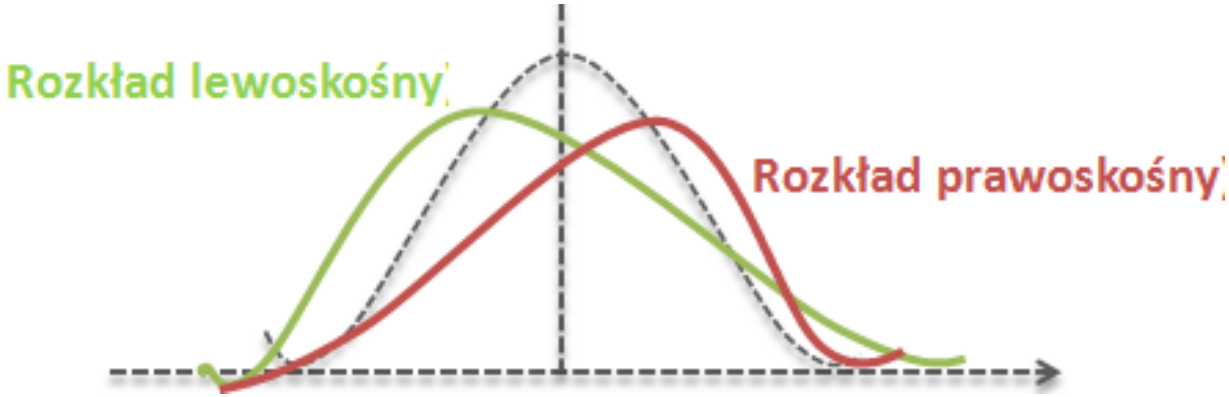


Liczba wystąpień danych, dla których badana cecha odbiega od średniej o pewną wartość „w górę” i „w dół” jest taka sama. Liczba wystąpień danych, dla których badana cecha odbiega od średniej maleje im większe jest odchylenie od średniej. Np. przeciętny noworodek waży 3 kg. Większość noworodków waży ok. 3 kg. Im większa różnica między wagą noworodka a średnią równą 3 kg, tym mniej takich dzieci się rodzi.


f(x)
P=0,95
P=0,68
punkt przegięcia
P=0,9973
$\overline{x}$
X
−1σ
+1σ
−2σ
+2σ
−3σ
+3σ

Skośność

Różnica między rozkładem normalnym a rozkładem uzyskanym empirycznie jeśli chodzi o symetrię względem średniej może przybierać 2 formy. Możliwe jest uzyskanie rozkładu lewoskośnego lub prawoskośnego.



Aby ocenić asymetrię oblicza się współczynnik asymetrii

$$a = \frac{\frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})^3}{s^3}$$

Znak współczynnika asymetrii określa kierunek asymetrii, wartość bezwzględna współczynnika asymetrii określa siłę asymetrii.

a=0 – rozkład jest symetryczny

a>0 – asymetria dodatnia (rozkład prawoskośny)

a<0 – asymetria ujemna (rozkład lewoskośny)

Dodatkowo uznaje się że:

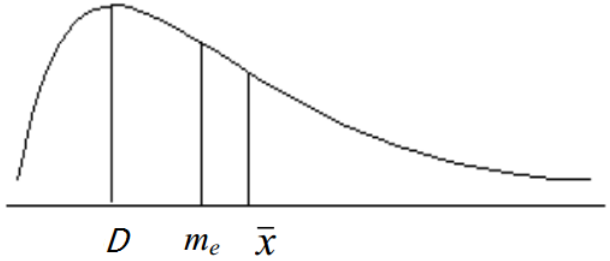
a≤ 0,7 - słaba asymetria,

a∈ (0,7;1, 4) – umiarkowana asymetria

a≥ 1,4 – silna asymetria

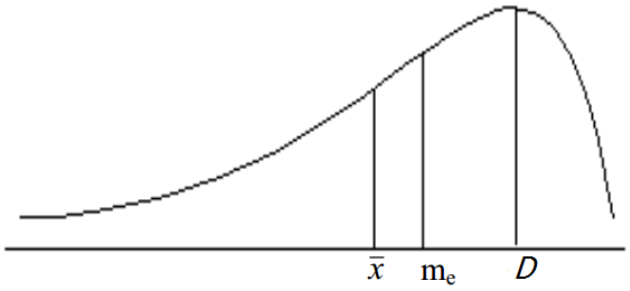
W jakim celu ocenia się asymetrię?

Asymetria dodatnia



Mediana dzieli dane na dwie równe części, średnia jest większa od mediany więc mniej niż połowa danych ma wartości większe od średniej.

Asymetria ujemna



Mediana dzieli dane na dwie równe części, średnia jest większa od mediany więc ponad połowa danych ma wartości większe od średniej.

W Excelu:

SKOŚNOŚĆ(liczba1;liczba2;...)

Liczba1; liczba2;... : od 1 do 255 argumentów liczbowych lub tablica

Przykład:

Firma 1

Pracownik	Premia [%]
1	5
2	20
3	7
4	40
5	5
6	6
7	2

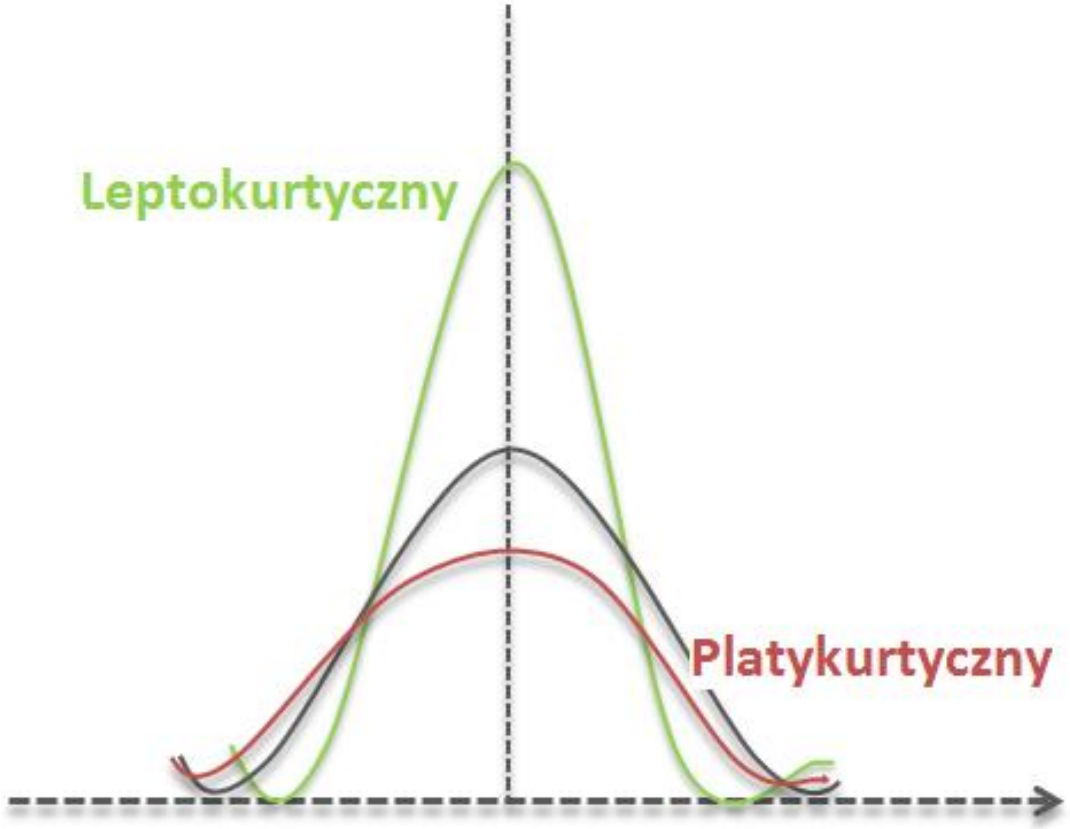
Średnia 12
S = 12

$a=[[(5-12)^3+(20-12)^3+(7-12)^3+(40-12)^3+(5-12)^3+(6-12)^3+(2-12)^3]/7]/(12)^3]=1,85$

Silna asymetria dodatnia

Kurtoza - miara zagęszczenia (koncentracji) wyników wokół wartości centralnej.

Różnica między rozkładem normalnym a rozkładem uzyskanym empirycznie jeśli chodzi o wysokość „garba” czyli liczbę przypadków, dla których wartość badanej cechy osiąga wartość średnią może przybierać 2 formy. Możliwe jest uzyskanie rozkładu spłaszczonego (platokurtycznego) lub wysmukłego (leptokurtycznego).



Aby ocenić kurtozę oblicza się wskaźnik kurtozy

$$k' = k - 3$$

Gdzie k to współczynnik skupienia (kurtoza)

$$k = \frac{\frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})^4}{s^4}$$

k'=0 – rozkład normalny

k'>0 – rozkład wysmukły

k'<0 – rozkład spłaszczony

W Excelu:

KURTOZA(liczba1;liczba2;...)

Liczba1; liczba2;... : od 1 do 255 argumentów liczbowych lub tablica

Przykład:

Firma 1

Pracownik	Premia [%]
1	5
2	20
3	7
4	40
5	5
6	6
7	2

Średnia 12
S = 12

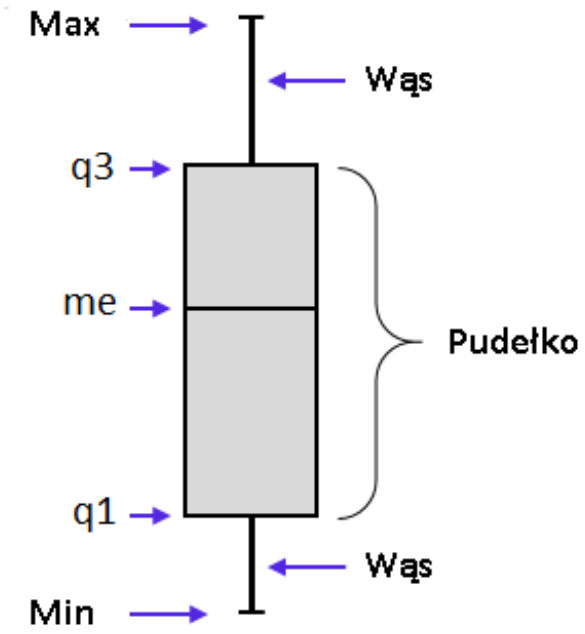
$k=[[(5-12)^4+(20-12)^4+(7-12)^4+(40-12)^4+(5-12)^4+(6-12)^4+(2-12)^4]/7]/(12)^4]=3,15$

$k'=3,15-3=0,15$

Rozkład wysmukły

Wykres pudełkowy (ramkowy) - służy do prezentacji wyników lub porównania danych. Zawiera informacje odnośnie położenia, rozproszenia i kształtu rozkładu danych.

Struktura wykresu



Gdzie:

Max - wartość maksymalna
q3 - trzeci kwartyl
me - mediana
q1 - pierwszy kwartyl
Min - wartość minimalna

Analiza wykresu pudełkowego:

1. **Położenie**: o położeniu świadczy cały wykres pudełkowy m.in możemy określić zakres danych (Min, Max).

2. **Rozproszenie**: im dłuższy wykres tym dane są bardziej rozproszone tzn. mogą przyjmować bardziej różniące się wartości. O rozproszeniu świadczą także długie wąsy - tzn. występują obserwacje skrajne (bardzo odbiegające od pudełka).

3. **Kształt:** jeżeli wykres wygląda symetrycznie względem kreski z medianą to możemy podejrzewać że wykres cechy jest **symetryczny**. Jeżeli pudełko nie jest równo podzielone albo/i wąsy są różnej długości to mamy do czynienia z **rozkładem asymetrycznym** - najczęściej to czy asymetria jest prawostronna czy lewostronna możemy odczytać po odległości Max i Min od Mediany: jeżeli jedna z tych odległości jest znacząco większa to mamy asymetrię prawostronną (jeżeli znacząco większa jest odległość Max od me) lub lewostronną (jeżeli odległość Min jest większa od me).

Przykład

Firma 1

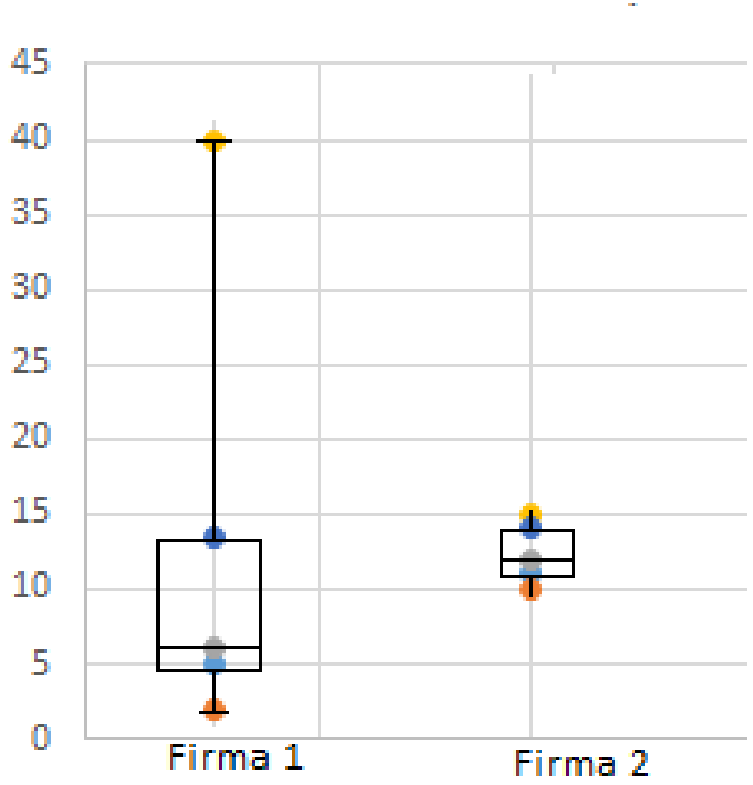
Pracownik	Premia [%]
1	5
2	20
3	7
4	40
5	5
6	6
7	2

q1	5
Min	2
Mediana	6
Max	40
q3	13,5

Firma 2

Pracownik	Premia [%]
1	10
2	15
3	12
4	10
5	15
6	13
7	12

q1	11
Min	10
Mediana	12
Max	15
q3	14


45
40
35
30
25
20
15
10
5
0
Firma 1
Firma 2

Zadanie do realizacji:

1. Przygotować dane do analizy (jedna zmienna niezależna i 2 zmienne zależne dla dwóch różnych przypadków, podmiotów itp. np. rok (zmienna niezależna), liczba mieszkańców z wykształceniem wyższym i liczba urodzeń w danym roku (zmienne zależne) dla województwa śląskiego i mazowieckiego)
   Dane można pobrać z dowolnego źródła np. http://stat.gov.pl/
2. Dla wszystkich zmiennych zależnych wyznaczyć: średnią arytmetyczną, dominantę, medianę, kwantyl pierwszy i trzeci, wariancję, odchylenie standardowe i przeciętne, współczynnik zmienności, przedział typowych wartości, rozstęp, skośność i kurtozę. Tam gdzie jest to możliwe wartości wyznaczyć za pomocą wzorów oraz za pomocą funkcji Excela.
3. Na dwóch oddzielnych wykresach przedstawić wykresy pudełkowe zmiennych zależnych (na jednym wykresie jedna zmienna dla dwóch przypadków np. liczba urodzeń w danym roku dla dwóch województw).
4. Zinterpretować uzyskane wyniki i porównać oba przypadki, podmioty odnośnie badanych cech (zmiennych zależnych).

Źródła:

Marek Cieciura, Janusz Zacharski PODSTAWY PROBABILISTYKI Z PRZYKŁADAMI ZASTOSOWAŃ W INFORMATYCE

Artur Zimny Statystyka opisowa Materiały pomocnicze do ćwiczeń