STATYSTYKA OPISOWA

WYKŁAD 1 i 2

Literatura:

Marek Cieciura, Janusz Zacharski, „Metody probabilistyczne w ujęciu praktycznym”,

L. Kowalski, „Statystyka”, 2005


CZY UWAŻASZ ŻE WYNIKI SONDAŻY SĄ FAŁSZOWANE?
NIE 92%
TAK 8%

Statystyka to dyscyplina naukowa, której zadaniem jest wykrywanie, analiza i opis prawidłowości występujących w procesach masowych.

Populacja to zbiorowość podlegająca badaniu statystycznemu.
Aby populację określić jednoznacznie charakteryzujemy ją pod względem:

–rzeczowym

–czasowym

–przestrzennym (terytorialnym).

Cecha to właściwość elementów populacji ze względu na którą prowadzimy badanie statystyczne.

Warianty to wartości cechy (cecha powinna mieć przynajmniej dwa warianty).

Przykład

Populacja:

Studenci II semestru Wydziału Elektroniki WAT, wg stanu na 1.10.2010.

Cechy:

- płeć,
- wzrost,
- kolor oczu,
- ocena na egzaminie z matematyki po I semestrze,
- ulubiony tygodnik,
- wysokość miesięcznych dochodów,
- czas poświęcony na naukę w tygodniu poprzedzającym ostatnią sesję egzaminacyjną.

Przykład

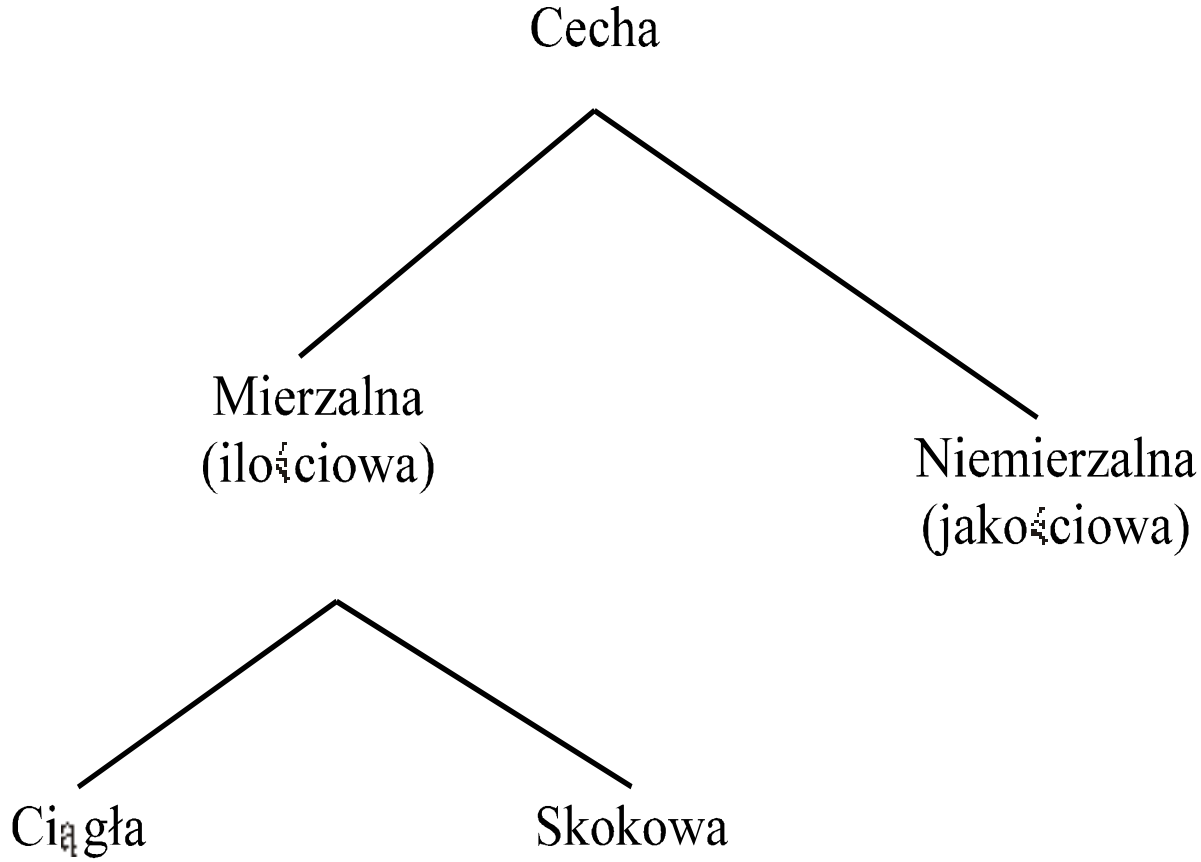
Populacja:

Samochody osobowe zarejestrowane w Warszawie, wg stanu na 1.09.2010.

Cechy:

- kolor karoserii,
- przebieg,
- średnie zużycie paliwa na 100 km,
- marka,
- czas osiągania prędkości 100 km/godz.

Uproszczona klasyfikacja cech:



Badanie statystyczne może być:

– **pełne** (obejmuje całą populację),
– **częściowe** (obejmuje część populacji – próbę).

Próba powinna być **reprezentatywna** tzn. rozkład wariantów badanej cechy w próbie powinien być zbliżony do rozkładu w całej populacji.

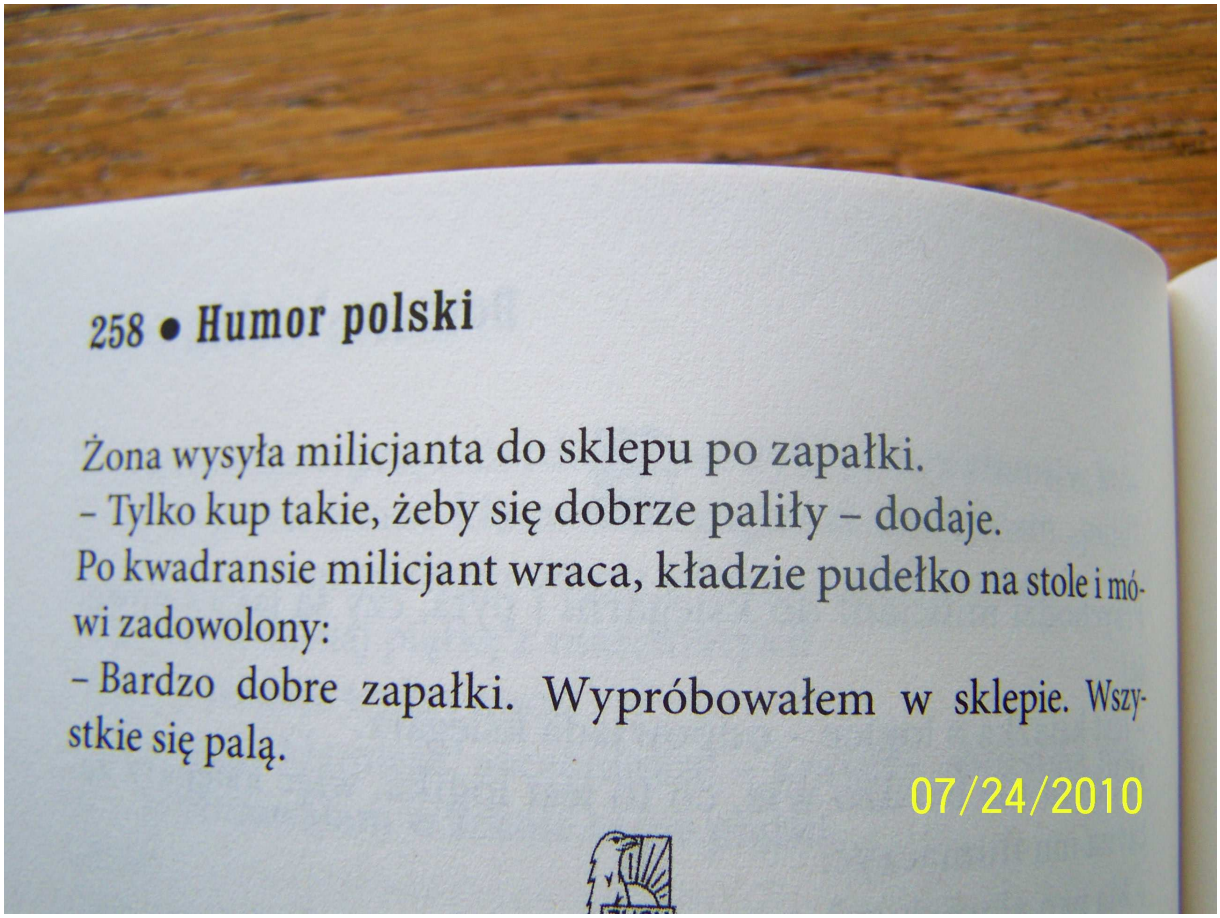
George Gallup 1901-1984

Pionier w dziedzinie badania opinii publicznej.

Rozwinął technikę doboru grupy reprezentatywnej

Uwaga

Badania pełne nie zawsze są możliwe lub celowe (badania niszczące, duża poplacja, wysokie koszty).



258 • Humor polski

Żona wysyła milicjanta do sklepu po zapałki.
– Tylko kup takie, żeby się dobrze paliły – dodaje.
Po kwadransie milicjant wraca, kładzie pudełko na stole i mówi zadowolony:
– Bardzo dobre zapałki. Wypróbowałem w sklepie. Wszystkie się palą.

07/24/2010

„Humor Polski” – lata 80-te

Liczebność próby.

Dla reprezentatywnej próby dorosłej liczebności Polski zwykle 1000 – 1300 osób.

Jerzy Spława-Neyman (1894 - 1981)
polski i amerykański matematyk i statystyk.

Wprowadził pojęcie przedziału ufności.

CHARAKTERYSTYKI LICZBOWE

Charakterystyki liczbowe to wielkości wyznaczone na podstawie danych statystycznych, **charakteryzujące** własności badanej cechy.

Zakładamy, że badana cecha jest **mierzalna**.

Klasyfikacja charakterystyk:

- charakterystyki **położenia** (np. średnia, mediana, dominanta),
- charakterystyki **rozproszenia** (np. wariancja, odchylenie standardowe, odchylenie ćwiartkowe, współczynnik zmienności),
- charakterystyki **asymetrii** (np. współczynnik asymetrii, wskaźnik asymetrii),
- charakterystyki **spłaszczenia** (np. kurtoza).

Charakterystyki mogą być:

- **klasyczne** (wyznaczone przez wszystkie wartości danych statystycznych, np. średnia, wariancja, odchylenie standardowe, współczynnik zmienności, współczynnik asymetrii),
- **pozycyjne** (wyznaczone przez niektóre (decyduje ich pozycja) wartości danych statystycznych, np. mediana, dominanta, kwartyle),
- **mieszane** (np. wskaźnik asymetrii).

Dane statystyczne prezentujemy zwykle w postaci

- **Szeregu prostego**

  (stosujemy w przypadku małej liczby danych),
- **Szeregu rozdzielczego punktowego** (stosujemy gdy dane się powtarzają),
- **Szeregu rozdzielczego przedziałowego** (stosujemy gdy danych jest dużo i się nie powtarzają),

Szereg prosty

Oznaczenia:
X – badana cecha,
n–liczba danych statystycznych,
x_i – dane statystyczne ($i = 1, 2, ..., n$),

Przykład
X – czas dojazdu do pracy (min),
Dane od 20 pracowników:
18, 26, 35, 12, 38, 45, 25, 54, 32, 15,
28, 22, 15, 18, 48, 42, 55, 14, 36, 16,

tzn. $x_1 = 18$, $x_2 = 26$, …., $x_{20} = 16$,

Średnia (arytmetyczna)

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

Np. dla danych 2, 5, 3, 4, 6, średnia wynosi 4 (sumujemy dane i sumę dzielimy przez liczbę danych).


Aksjomat Cole'a
Suma inteligencji na Ziemi jest stała. Liczba ludności rośnie

Uwaga

Własność (suma odchyleń od średniej jest równa zero)

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

EXCEL:

ŚREDNIA

Zwraca wartość średnią (średnią arytmetyczną) argumentów.

Składnia

ŚREDNIA(**liczba1**;liczba2;...)

Liczba1; liczba2;... to od 1 do 255 argumentów liczbowych,

Dominanta

d = wariant cechy występujący najczęściej (o ile taki istnieje).

Np. dla danych

2, 3, 4, 3, 2, 5, 3, 2, 3

dominantą jest 3.

Natomiast dla danych 2, 3, 4, 3, 2, 5, 3, 2, 3, 2 dominanta nie jest określona (mówimy, że jest to rozkład dwumodalny).

EXCEL:

WYST.NAJCZĘŚCIEJ

Zwraca wartość najczęściej występującą lub powtarzającą się w tablicy albo w zakresie danych.

Składnia

WYST.NAJCZĘŚCIEJ(**liczba1**;liczba2;...)

Liczba1; liczba2;... to 1 do 255 argumentów,

Zamiast listy argumentów rozdzielonych średnikami można zastosować także pojedynczą tablicę lub odwołanie do tablicy.

Jeśli zbiór danych nie zawiera zduplikowanych punktów danych, funkcja WYST.NAJCZĘŚCIEJ zwraca wartość błędu #N/D!.

Mediana (wartość środkowa)

Jeśli $x_1 \leq x_2 \leq \ldots\ldots \leq x_n$ dane uporządkowane to

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

Przykład
Dla danych (po uporządkowaniu)
2, 2, 3, 3, 4, 5, 5, 5, 5 medianą jest 4.

Dla danych (po uporządkowaniu)
2, 2, 2, 3, 3, 4, 5, 5, 5, 5
medianą jest 3,5.

EXCEL:

MEDIANA

Zwraca wartość mediany dla podanych liczb. Mediana jest liczbą w środku zbioru liczb.

Składnia

MEDIANA(**liczba1**;liczba2;...)

Liczba1; liczba2;... to 1 do 255 liczb, dla których należy wyznaczyć medianę.

Podział co 50% - mediana,
Podział co 25% - kwartyle, q_1, $q_2 = m_e$, q_3,
Podział co 10% - decyle,
Podział co 1% - percentyle (centyle),

Obliczanie **kwartyli**.

Jeśli $x_1 \le x_2 \le \ldots\ldots \le x_n$ dane uporządkowane to

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

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

EXCEL:

KWARTYL

Zwraca kwartyl zbioru danych.

Składnia

KWARTYL(**tablica**;**kwartyl**)

Tablica to tablica lub zakres komórek wartości liczbowych, dla których chcemy obliczyć wartość kwartylu.

Kwartyl wskazuje wartość, która ma być zwrócona.

Jeżeli kwartyl równa się	funkcja KWARTYL zwraca
0	Wartość minimalna
1	Pierwszy kwartyl (25. percentyl)
2	Wartość mediany (50. percentyl)
3	Trzeci kwartyl (75. percentyl)
4	Wartość maksymalna

EXCEL:

PERCENTYL

Zwraca k-ty percentyl wartości w zakresie.

Składnia

PERCENTYL(**tablica**;**k**)

Tablica to tablica lub zakres danych, który określa względną pozycję.

k to wartość percentylu w zakresie od 0 do 1 włącznie.

Wariancja

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

Np. dla danych 2, 5, 3, 4, 6, średnia wynosi 4. Aby wyznaczyć wariancję liczymy sumę kwadratów odchyleń poszczególnych danych od średniej:
$(2 – 4)^2 + (5 – 4)^2 +(3 – 4)^2 +(4 – 4)^2 +(6 – 4)^2 = 4 + 1 + 1 + 0 + 4 = 10$
otrzymana sumę dzielimy przez 5 (liczba danych). Zatem wariancja dla powyższych danych wynosi 2.

EXCEL:

WARIANCJA.POPUL

Oblicza wariancję na podstawie całej populacji.

Składnia

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

Liczba1; liczba2;... to od 1 to 255 argumentów liczbowych,

Uwaga

$$s^2 = \frac{1}{n}\sum_{i=1}^{n} x_i^{\,2} - \left(\overline{x}\right)^2$$

Wariancja mierzy rozrzut (zróżnicowanie) danych statystycznych (punktem odniesienia jest średnia) lecz miara ta wyrażona jest w kwadratach jednostek rozpatrywanych danych statystycznych co utrudnia interpretację, dlatego w praktyce częściej stosujemy pierwiastek z wariancji nazywany **odchyleniem standardowym**.

Odchylenie standardowe

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

EXCEL:

ODCH.STANDARD.POPUL

Oblicza odchylenie standardowe dla całej populacji podanej w postaci argumentów. Odchylenie standardowe jest miarą tego, jak szeroko wartości są rozproszone od wartości średniej.

Składnia

ODCH.STANDARD.POPUL(**liczba1**;liczba2;...)

Liczba1; liczba2;... to od 1 do 255 argumentów odpowiadających populacji.

Zamiast argumentów rozdzielonych średnikami można użyć pojedynczej tablicy lub odwołania do tablicy.

Odchylenie przeciętne

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

EXCEL:

ODCH.ŚREDNIE

Zwraca wartość średnią odchyleń **bezwzględnych** punktów danych od ich wartości średniej.

Składnia

ODCH.ŚREDNIE(**liczba1**;liczba2;...)

Liczba1; liczba2;... to od 1 do 255 argumentów, dla których należy wyznaczyć średnią odchyleń bezwzględnych.

Współczynnik zmienności

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

(niekiedy wynik jest podawany w procentach)

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

Przedział typowych wartości

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

Jest to przedział do którego należy większość danych statystycznych, interpretacja ta jest uzasadniona wtedy gdy cecha ma rozkład zbliżony do rozkładu normalnego.

Rozstęp

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

Współczynnik asymetrii

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

lub $a_1 = \frac{\bar{x} - d}{s}$ (wskaźnik asymetrii)

Wskaźnik asymetrii można wyznaczać tylko gdy dominanta jest określona.

EXCEL:

SKOŚNOŚĆ

Zwraca skośność rozkładu.

Skośność charakteryzuje stopień asymetrii rozkładu wokół jego średniej. Skośność dodatnia określa rozkład z asymetrią rozciągającą się w kierunku wartości dodatnich. Skośność ujemna określa rozkład z asymetrią rozciągającą się w kierunku wartości ujemnych.

Składnia

SKOŚNOŚĆ(**liczba1**;liczba2;...)

Liczba; liczba2;... to od 1 do 255 argumentów, dla których należy obliczyć skośność. Zamiast argumentów rozdzielonych średnikami można użyć pojedynczej tablicy lub odwołania do tablicy.

Jeśli liczba punktów danych jest mniejsza niż trzy lub jeśli odchylenie standardowe równe jest zero, funkcja SKOŚNOŚĆ zwraca wartość błędu #DZIEL/0!.

- Wzór obliczający skośność:

$$\hat{a} = \frac{n^2}{(n-1)(n-2)} \frac{\frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})^3}{\hat{s}^3}$$

gdzie

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

Jest **wariancją nieobciążoną (z próby) (funkcja WARIANCJA w EXCELU)**

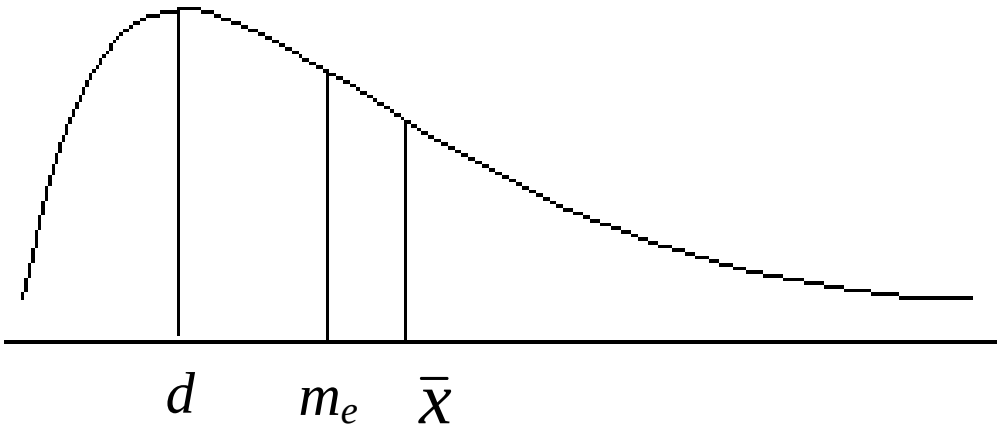
Uwaga

Znak współczynnika asymetrii wskazuje na kierunek asymetrii natomiast jego wartość bezwzględna określa siłę asymetrii.

Ponieważ współczynnik asymetrii jest wielkością niemianowaną to dobrze nadaje się do porównywania dwóch cech lub tej samej cechy w różnych populacjach

Miary asymetrii mają poniższą interpretację tylko w przypadku rozkładów z jedną dominującą wartością (rozkład jednomodalny) wtedy mediana plasuje się miedzy dominantą a średnią tzn. $d \leq m_e \leq \bar{x}$ lub $\bar{x} \leq m_e \leq d$.

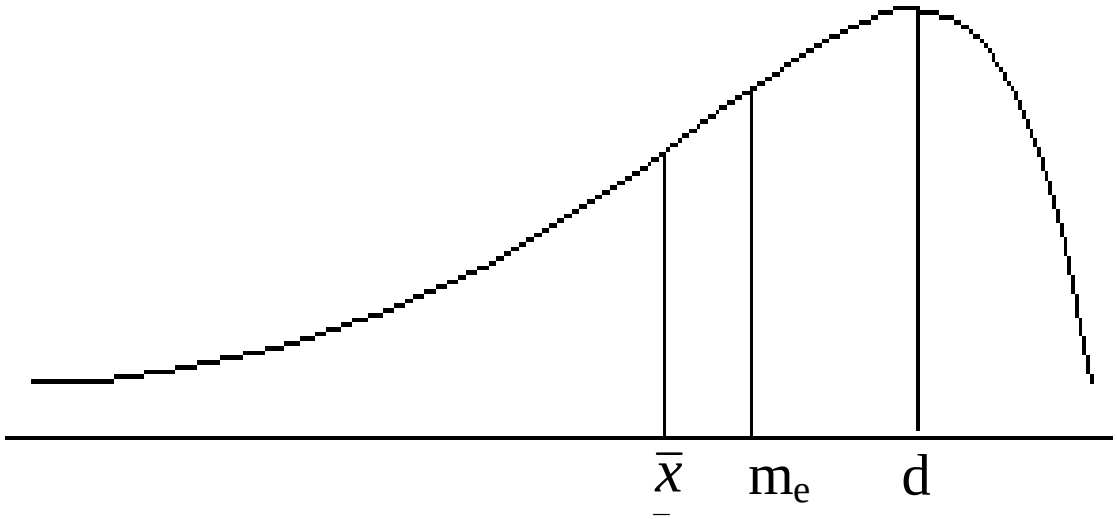
a) Jeśli $a = 0$ to mówimy, że cecha ma rozkład symetryczny
b) Jeśli $a > 0$ to mówimy, że cecha ma rozkład asymetryczny (asymetria dodatnia lub prawostronna)



Asymetria dodatnia (prawostronna)

Ponieważ mediana dzieli badaną strukturę na dwie równe części a średnia jest większa od mediany to **mniej niż połowa danych ma wartości większe od średniej.**

c) Jeśli $a < 0$ to mówimy, że cecha ma rozkład asymetryczny (asymetria ujemna lub lewostronna).



Asymetria ujemna (lewostronna)

Ponieważ mediana dzieli badaną strukturę na dwie równe części a średnia jest mniejsza od mediany to ponad **połowa danych ma wartości większe od średniej.**

Współczynnik skupienia (kurtoza)

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

Uwaga

Kurtoza mierzy skupienie (koncentrację) wartości cechy wokół średniej arytmetycznej.

W praktyce silne skupienie oznacza, że średnia arytmetyczna dobrze reprezentuje badaną zbiorowość bowiem wiele jej elementów ma wartości zbliżone do średniej.

- Jeśli $k = 3$ to skupienie jest normalne (takie skupienie ma rozkład normalny – będzie omawiany i stosowany później). W tym przypadku można przyjmować, że w typowym obszarze zmienności mieści się około 68% obserwacji.
- Jeśli $k < 3$ to rozkład jest spłaszczony (platokurtyczny). W tym przypadku można przyjmować, że w typowym obszarze zmienności mieści się mniej niż 68% obserwacji.
- Jeśli $k > 3$ to rozkład jest wysmukły (leptokurtyczny). W tym przypadku można przyjmować, że w typowym obszarze zmienności mieści się ponad 68% obserwacji.

Wskaźnik kurt ozy

$$k' = k - 3$$

EXCEL:

KURTOZA

Zwraca kurtozę zbioru danych.

Kurtoza charakteryzuje względne spłaszczenie rozkładu w porównaniu z rozkładem normalnym. Dodatnia kurtoza oznacza rozkład o stosunkowo małym spłaszczeniu. Ujemna kurtoza oznacza rozkład stosunkowo płaski.

Składnia

KURTOZA(**liczba1**;liczba2;...)

Liczba1; liczba2;... to od 1 do 255 argumentów, dla których jest obliczana kurtoza. Zamiast argumentów rozdzielonych średnikami można zastosować pojedynczą tablicę lub odwołanie do tablicy.

Jeżeli jest mniej niż cztery punkty danych lub jeśli standardowe odchylenie próbki jest równe zero, funkcja KURTOZA zwraca wartość błędu #DZIEL/0!.

- Wzór obliczający Kurtozę:

$$\hat{k} = \frac{n^2(n+1)}{(n-1)(n-2)(n-3)} \frac{\frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})^4}{\hat{s}^4} - 3\frac{(n-1)^2}{(n-2)(n-3)}$$

gdzie

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

Jest **wariancją nieobciążoną (z próby) (funkcja WARIANCJA w EXCELU)**

Uwaga.
W EXCELU można globalnie obliczyć charakterystyki korzystając z opcji

STATYSTYKA OPISOWA

w module

ANALIZA DANYCH

(zakładka DANE).

PRZYKŁAD

dane

18
26
35
12
38
45
25
54
32
15
28
22
15
18
48
42
55
14
36
16

Kolumna1		
Średnia	29,7	
Błąd standardowy	3,096772513	(błąd średniej)
Mediana	27	
Tryb	18	(dominanta)!
Odchylenie standardowe	13,8491877	(z próby)!
Wariancja próbki	191,8	
Kurtoza	-1,016690605	!
Skośność	0,460970165	!
Zakres	43	(rozstęp)
Minimum	12	
Maksimum	55	
Suma	594	
Licznik	20	

Szereg rozdzielczy punktowy (stosujemy gdy dane się powtarzają),

w_i	n_i	s_i
w_1	n_1	n_1
w_2	n_2	$n_1 + n_2$
...	...	...
w_r	n_r	$n_1 + n_2 + ... + n_r = n$
razem	n	---

(ostatnia kolumna umieszczona dodatkowo)

Oznaczenia:

X – badana cecha,

n–liczba danych statystycznych,

x_i – dane statystyczne ($i = 1, 2, ..., n$),

r –liczba wariantów,

w_i – warianty cechy ($i = 1, 2, ..., r$),

n_i – liczebność wariantu w_i ($i = 1, 2, ..., r$),

$$\boldsymbol{(n = n_1 + n_2 + ...+ n_r)}$$

s_i – liczebności skumulowane ($s_i = n_1 + n_2 + ...+ n_i$).

Niekiedy liczebności poszczególnych wariantów nazywa się **częstościami**.

Przykład

W 25 osobowej grupie studentów na egzaminie ze statystyki zarejestrowano następujące wyniki:

3, 2, 4, 3, 2, 5, 3, 3, 3, 2, 3, 4, 5, 3, 5, 3, 3, 2, 4, 3, 3, 4, 3, 2, 3.

Szereg rozdzielczy punktowy

w_i	n_i	s_i
2	5	5
3	13	18
4	4	22
5	3	25
—	25	—

EXCEL:

CZĘSTOŚĆ

Oblicza, jak często wartości występują w określonym zakresie wartości, a następnie zwraca tablicę liczb w układzie pionowym. Ponieważ funkcja CZĘSTOŚĆ zwraca tablicę, musi być wprowadzona jako formuła tablicowa.

Składnia

CZĘSTOŚĆ(**tablica_dane**;**tablica_przedziały**)

Tablica_dane to tablica lub odwołanie do zbioru wartości, dla których mają być liczone częstości.
Tablica_przedziały to tablica lub odwołanie do interwałów, w których mają być grupowane wartości argumentu tablica_dane.

Liczba elementów w zwróconej tablicy jest o jeden większa niż liczba elementów w argumencie tablica_przedziały. Ten dodatkowy element zwraca liczbę wszystkich wartości istniejących poza najwyższym interwałem..

UWAGA Formuła musi być wprowadzona jako formuła tablicowa. Należy zaznaczyć zakres komórek wyniku, nacisnąć klawisz F2, a następnie nacisnąć klawisze CTRL+SHIFT+ENTER. Jeżeli formuła nie jest wprowadzana jako formuła tablicowa, to będzie tylko jeden wynik w komórce.

Średnia (arytmetyczna)

$$\bar{x} = \frac{1}{n}\sum_{i=1}^{r} n_i w_i$$ gdy dane się powtarzają.

Przykład

Dla danych 2, 3, 4, 3, 2, 5, 3, 2, 4, 5, 3, 4, 2, 2, 3 możemy wykorzystać ich krotność (unikamy wielokrotnego dodawania tych samych składników) i zanim policzymy średnią sporządzamy zestawienie danych w szeregu rozdzielczym punktowym.

Ostatnia kolumna zawiera pomocnicze obliczenia sumy $\sum_{i=1}^{r} n_i w_i$.

w_i	n_i	$w_i\, n_i$
2	5	10
3	5	15
4	3	12
5	2	10
razem	15	47

Dzieląc sumę ostatniej kolumny przez liczbę danych otrzymujemy wartość średniej

47/15 = 3,13.

Wariancja

$$s^2 = \frac{1}{n}\sum_{i=1}^{r} n_i \left(w_i - \overline{x}\right)^2$$

Uwaga

$$s^2 = \frac{1}{n}\sum_{i=1}^{r} n_i w_i^{\,2} - \left(\bar{x}\right)^2$$

Odchylenie standardowe

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

Odchylenie przeciętne

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

Współczynnik asymetrii

$$a = \frac{\frac{1}{n}\sum_{i=1}^{r} n_i \left(w_i - \overline{x}\right)^3}{s^3}$$

Współczynnik skupienia (kurtoza)

$$k = \frac{\frac{1}{n}\sum_{i=1}^{r} n_i \left(w_i - \bar{x}\right)^4}{s^4}$$

Przykład

W 25 osobowej grupie studentów na egzaminie ze statystyki zarejestrowano następujące wyniki:

3, 2, 4, 3, 2, 5, 3, 3, 3, 2, 3, 4, 5, 3, 5, 3, 3, 2, 4, 3, 3, 4, 3, 2, 3.

w_i	n_i	s_i	n_i / n	$w_i * n_i$	$(w_i - \bar{x}) * n_i$	$(w_i - \bar{x})^2 * n_i$	$(w_i - \bar{x})^3 * n_i$	$(w_i - \bar{x})^4 * n_i$
2	5	5	0,2	10	– 6	7,2	– 8,64	10,368
3	13	18	0,52	39	– 2,6	0,52	– 0,104	0,021
4	4	22	0,16	16	3,2	2,56	2,048	1,638
5	3	25	0,12	15	5,4	9,72	17,496	31,493
—	25	—	1	80	0,00	20,00	10,80	43,52