ELEMENTY STATYSTYKI

1. DANE

W badaniach statystycznych populacją nazywamy grupę osób, zwierząt, roślin lub przedmiotów badanych. Interesują nas przy tym pewne wybrane cechy tych populacji. Takie cechy nazywamy zmiennymi i oznaczamy dużymi literami X, Y, Z itd. Nas będą interesować głównie cechy wyrażające się za pomocą liczb (cechy ilościowe). Wartościami zmiennych nazywamy danymi. Dane zmiennej X oznaczamy przez x_1 , x_2 ,..., x_n , dane zmiennej Y oznaczamy przez y_1 , y_2 , ..., y_n itd.

PRZYKŁADY

1. Wzrost uczniów waszej klasy jest zmienną. Dane stanowią liczby wyrażające ten wzrost np. w milimetrach.

2. Ocena jaką otrzymał z matematyki każdy uczeń waszej klasy w poprzednim roku nauki, jest zmienną. Danymi w tym przypadku mogły być oceny: 1, 2, 3, 4, 5, 6. Oceny najniższej nie życzymy nikomu (hehe).

W badaniach statystycznych bardzo ważną czynnością jest policzenie częstości występowania poszczególnych danych.

PRZYKŁAD

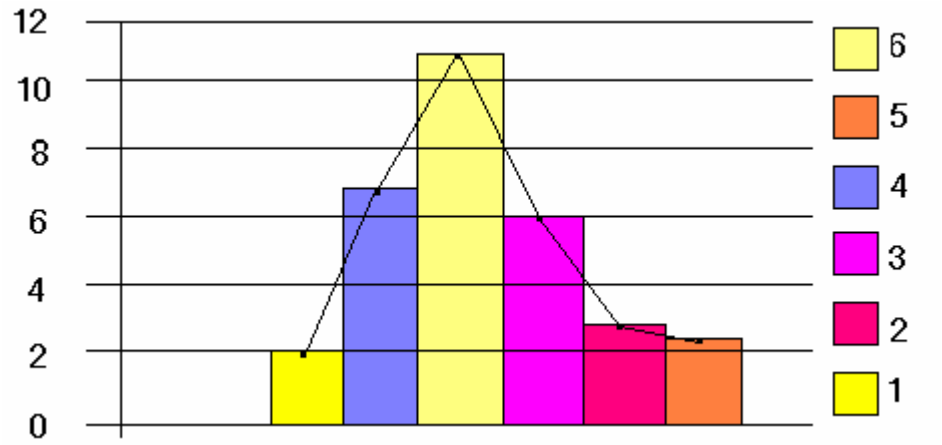
1. W pewnej klasie uczniowie otrzymali następujące oceny ze sprawdzianu z matematyki

3	4	2	3	3	1	5	5	4	3
1	2	2	3	3	3	4	6	2	2
3	4	5	4	4	3	3	3	2	2

Powyższe zestawienie jest mało czytelne. Policzymy, ile ocen każdego rodzaju w nim występuje. Otrzymamy zestawienie nowego typu, które w przyszłości będziemy nazywać tabelą rozkładu częstości (lub krócej tabelą częstości). Stosunek częstości występowania danej do liczby wszystkich danych nazywamy częstością względną. Daną o największej częstości będziemy nazywać daną modalną albo modą (w naszym przypadku będzie to ocena 3)

Ocena	1	2	3	4	5	6
Częstość	2	7	11	6	3	1
Częstość względna	$\frac{1}{15}$	$\frac{7}{30}$	$\frac{11}{30}$	$\frac{1}{5}$	$\frac{1}{10}$	$\frac{1}{30}$

Tego typu dane możemy przedstawiać w formie graficznej. Najpopularniejszą formą przedstawienia graficznego danych jest histogram. Rysujemy go w układzie współrzędnych. Na osi poziomej odkładamy kolejne dane. Na osi pionowej odkładamy częstości występowania poszczególnych danych. Każdej danej przyporządkowany jest słupek o stałej szerokości oraz o wysokości równej częstości tej danej. Łącząc środki górnych krawędzi słupków histogramu, otrzymujemy wielokąt częstości.



Często dane występują w postaci zbyt szczegółowej. Aby je lepiej analizować, łączymy je w klasy. Na przykład, wprowadzając w powyższym przykładzie dwie klasy danych: 1-3, 4-6, otrzymamy tabelę:

Klasa	1-3	4-6
Częstość	20	10

PRZYKŁAD

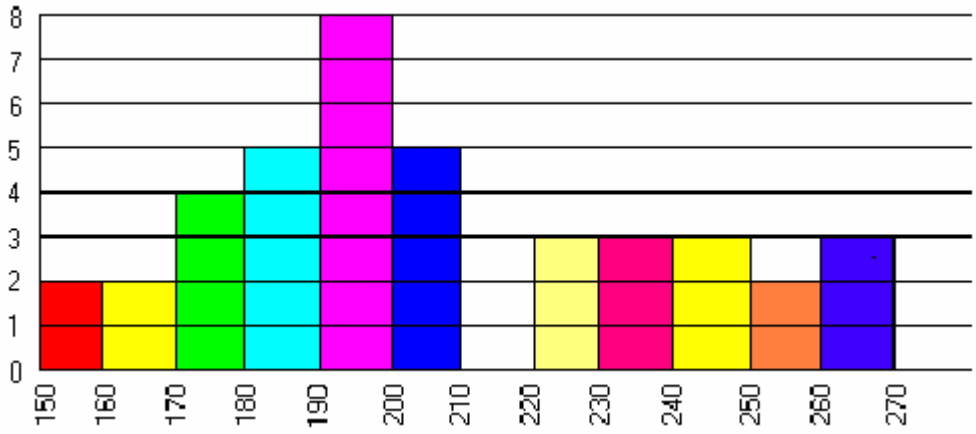
1. Oto zestawienie liczby dni deszczowych w wybranych stacjach meteorologicznych Wielkiej Brytanii w ciągu jednego roku:

260	234	173	197	182	209	241	187	191	171
265	188	206	164	194	207	199	226	135	156
194	205	243	268	185	241	253	228	199	158
236	209	199	226	164	250	198	177	175	180

Utworzymy tablicę częstości, stosując przedziały klasowe o rozpiętości 10, począwszy od klasy 150 – 159

Klasa	150-159	160-169	170-179	180-189	190-198	200-209	210-219	220-229	230-239	240-249	250-259	260-269
Częstość	2	2	4	5	8	5	0	3	3	3	2	3

Klasa 190 – 199 ma tutaj częstość największą. Taką klasę nazywamy klasą modalną.



ZADANIA

1. Poniżej podaliśmy przyrost naturalny w Polsce w promilach w wybranych latach okresu lat 1950 – 1990:
   19 20 15 10 9 10 10 8 4
   Narysuj histogram częstości występowania tych danych oraz wielokąt częstości.

2. Policzono, że w pewnej powieści, na pewniej stronicy znajduje się w kolejnych wierszach

   11 10 6 8 12 1 9 5 7 8
   8 8 8 9 10 8 8 8 3 7
   13 5 10 9 7 9 6 8 6 8
   11 9 9 8 9 11 3 6

   wyrazów. Przedstaw dane w postaci tablicy częstości. Ile wynosi moda wyrazów w wierszu? Narysuj histogram częstości danych. Oblicz częstości względne poszczególnych danych.

ŚREDNIA ARYTMETYCZNA

Przy opracowywaniu danych bardzo często stajemy przed koniecznością podania liczby charakteryzującej w jakiś sposób cały zbiór danych. Takie liczby nazywamy średnimi. Najpopularniejszą średnią jest średnia arytmetyczna. Obliczamy ją, dodając wszystkie dane, a następnie dzieląc otrzymaną sumę przez liczbę danych. Jeśli zmienną będziemy oznaczać przez X, liczbę tych danych przez N i średnią arytmetyczną zmiennej przez to powyższe słowa możemy zapisać za pomocą następującego wzoru symbolicznego:

$$\overline{X} = \frac{\Sigma X}{N}$$

We wzorze tym ΣX oznacza sumę wszystkich danych zmiennej

$$X = x_1 + x_2 + \ldots + x_n$$

PRZYKŁADY

1. Średnią liczb:
   1 3 5 7 9 11 13 15 17 19 jest liczba

$$\frac{1+3+5+7+9+11+13+15+17+19}{10} = \frac{100}{10} = 10$$

W naszym przypadku średnia nie jest żadną z liczb danych i leży dokładnie w środku między wartościami krańcowymi danych. Zauważmy, że każda dana występuje tutaj tylko jeden raz.

2. Jeżeli obliczyliśmy częstości danych, to średnią można obliczyć nieco łatwiej. Na przykład średnią arytmetyczną danych:
   1 5 7 2 5 5 7 2 2 1 1 7
   obliczamy, sumując iloczyny danych przez ich częstości, a następnie dzieląc otrzymaną sumę przez liczbę danych:

$$\frac{1\cdot 3+2\cdot 3+5\cdot 3+7\cdot 3}{12} = 3\frac{3}{4}$$

3. Weźmy pod uwagę dane:
   8 7 10 8 8 10 8 10 10 2
   8 8 10 10 10 9 9 9 13 11
   9 8 11 10 10 4
   określające, ile wyrazów mają kolejne wiersze wstępu do pewnego podręcznika. Średnią arytmetyczną tych danych będziemy obliczać inaczej niż poprzednio. Zauważmy, że liczba 8 znajduje się stosunkowo blisko większości danych. Każdą daną przedstawmy jako sumę liczby 8 oraz pewnej liczby (zwanej odchyleniem):
   0 –1 2 0 0 2 0 2 2 –6
   0 0 2 2 2 1 1 1 5 3
   1 0 3 2 2 -4

Średnia arytmetyczna danych jest wtedy równa sumie liczby 8 oraz średniej arytmetycznej odchyleń:

$$8 + \frac{0\cdot 7+1\cdot 4+2\cdot 9+3\cdot 2+5+(1)+(4)+(6)}{26} = 8\ \frac{11}{13}$$

Posłużyłyśmy się tutaj rozkładem częstości odchyleń.

4. Średnią arytmetyczną można obliczać również dla danych zgrupowanych. Otrzymana średnia nie jest wtedy na ogół równa średniej arytmetycznej wyjściowych danych. W szczególności grupując dane z poprzedniego przykładu w klasy 1 - 5, 6 - 10, 11 – 15, otrzymujemy częstości:

Klasa	1 - 5	6 - 10	11 - 15
Częstości	2	21	3

Wartościami środkowymi kolejnych klas są:

$$\frac{1+5}{2}=3 \qquad \frac{6+10}{2}=8 \qquad \frac{11+15}{2}=13$$

Średnia arytmetyczna danych zgrupowanych jest równa ilorazowi sumy iloczynów wartości środkowych klas i ich częstości przez liczbę danych:

$$\frac{3\cdot 2+8\cdot 21+13\cdot 3}{26} = 8\ \frac{5}{26}$$

Sposób obliczania średniej arytmetycznej z przedostatniego przykładu jest wart uwagi. Można go wypowiedzieć następująco: na podstawie wyglądu danych przyjmujemy pewną wartość jako średnią „przewidywaną”. Następnie wypisujemy odchylenia poszczególnych danych od średniej przewidywanej i obliczamy średnią arytmetyczną odchyleń.

Tw.

Średnia arytmetyczna danych jest sumą średniej przewidywanej oraz średniej arytmetycznej odchyleń.

DOWÓD.

Jeśli dane są postaci:

$a + r_1$, $a + r_2$, ... , $a + r_n$, to zachodzą następujące równości

$$\frac{(a+r_1)+(a+r_2)+\ldots+(a+r_n)}{n} = \frac{na+r_1+r_2+\ldots+r_n}{n} = a + \frac{r_1+r_2+\ldots+r_n}{n}$$

ZADANIA

1. Poniżej podano średnią temperaturę powietrza w wybranych stacjach meteorologicznych Polski w latach 1951 – 1980 w stopniach Celsjusza:
   7,4 7,5 6,0 6,8 7,2 8,3 6,8 7,6 7,2 8,1
   8,0 7,7 7,2 8,2 7,8 7,6 7,4 8,2 6,9 7,2
   7,7 0,4 7,2 8,3 7,7 7,6 8,0 7,7 7,9 5,0
   Oblicz średnią temperaturę powietrza w Polsce w tych latach. Następnie pogrupuj je, stosując przedziały klasowe:5,0 – 5,9 ; 6,0 – 6,9 itd. Utwórz tabelę częstości klas i korzystając z niej, oblicz jeszcze raz, średnią temperaturę powietrza.

2. Sporządź tablicą częstości i oblicz, ile średnio dzieci przypada w rodzinach kolegów i koleżanek Twojej klasy.

3. Niżej podana tabela częstości przedstawia liczbę bramek strzelonych podczas jesiennej Rundy rozgrywek piłki nożnej w Anglii:

Bramki	0	1	2	3	4	5
Częstości	16	15	6	4	2	1

   Oblicz średnią liczbę uzyskanych bramek.

RÓŻNE RODZAJE ŚREDNICH I ICH INTERPRETACJE

Poznaliśmy dotychczas dwa rodzaje średnich: modę oraz średnią arytmetyczną. Średnia arytmetyczna nie zawsze oddaje charakter danych. Świadczy o tym następujący przykład.

PRZYKŁAD

Przypuśćmy, że w pewnym zakładzie pracy jest zatrudnionych 100 pracowników, z których:
90 zarabia po 500 zł miesięcznie
5 zarabia po 1000 zł miesięcznie
4 zarabia po 4000 zł miesięcznie
1 zarabia po 10000zł miesięcznie.

Średnia arytmetyczna płac wyniesie wtedy

$$\frac{90\cdot 500 + 5\cdot 1000 + 4\cdot 4000 + 10000}{100} = \frac{76000}{100} = 760 \text{ zł}$$

Jest to kwota przewyższająca najniższą płacę o 52%. Ponieważ najniżej zarabiający jest aż90%, więc średnia ta nie oddaje charakteru danych. Znacznie lepszą średnią jest moda, która w tym wypadku jest równa 500 zł. Stosując ją, nie obrazimy uczuć tych najsłabiej uposażonych.

Inną średnią jest mediana, czyli wartość środkowa. Znajdujemy ją w następujący sposób. Dane porządkujemy według ich wielkości liczbowych. Jeśli liczba danych jest nieparzysta to bierzemy tę, która leży w środku. Jeżeli liczba jest parzysta, to bierzemy średnią arytmetyczną dwóch środkowych danych.

PRZYKŁADY

1. Medianą wzrostu siedmiu uczniów pewnej klasy (liczonego w centymetrach):
152 147 155 149 151 148 157
jest liczba 151, bo ustawiając je zgodnie z rosnącą wielkością mamy:
147 148 149 151 152 155 157

2. Medianę możemy obliczać również, mając dany rozkład:

Wynik	151	152	153	154	155	156
Częstość	3	4	7	6	3	1

mamy 24 dane. Wyniki 12 i 13 odpowiadają danej 153. Zatem mediana jest równa 153.

3. Dość skomplikowane jest obliczanie mediany dla danych zgrupowanych. Rozkład częstości wagi uczniów pewnej klasy wynosił (z dokładnością do 1 kg):

Waga [kg]	35 - 39	40 - 44	45 - 49	50 - 54	55 - 59	60 - 64
Częstości	1	5	9	9	4	2

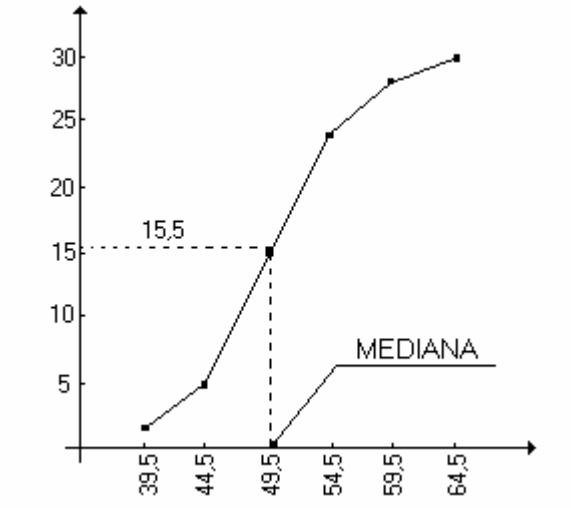
Tworzymy najpierw tabelę tak zwanych częstości skumulowanych:

Waga [kg]	35 - 39	40 - 44	45 - 49	50 - 54	55 - 59	60 – 64
Częstość	1	6	15	24	28	30

Następnie rysujemy krzywą częstości skumulowanych. W tym celu znajdujemy najpierw tak zwane górne granice klas, czyli dla klasy 35 – 39 środek 39,5 odcinka 39 – 40, dla klasy 40 – 44 środek 44,5 odcinka 44 – 45 itd.
Następnie łączymy punkty (39,5 ; 1)(44,5 ;6)(49,5 ; 15)(54,5 ; 24)(59,5 ; 28)(64,5 ; 30)
Otrzymujemy krzywą o charakterystycznym wyglądzie. Danych jest 30. Mediana jest to wartość liczbowa, odpowiadającą punktowi na osi częstości leżącemu w środku między 15 i 16, czyli 15,5. Jest to

$$49{,}5 + 0{,}5 \cdot \frac{5}{9} = 49\frac{7}{9}$$

(bo granicy 49,5 odpowiadającej częstości 15 dodajemy wartość odpowiadającą przyrostowi częstości 0.5, a wiemy, że przyrostowi częstości 9 odpowiada przyrost danej 5).



Mediana nie ulega zmianie z powodu jednej lub kilku nienormalnie dużych lub małych danych. Warto ją stosować, jeśli odległości między większością danych są niewielkie w porównaniu z wielkością samych danych. W niektórych sytuacjach mediana nie jest jednak dobrą średnią.

PRZYKŁAD

1. Weźmy pod uwagę dane: 1 1 1 1 1 3 4 6 7
   Medianą jest tu 1 i nie oddaje ona charakteru danych. Średnia arytmetyczna wynosi 3 i lepiej je charakteryzuje.
   Czasem się zdarza, że niektóre dane mają większe znaczenie od innych. W takich przypadkach możemy stosować tzw. **średnią ważoną**. Mając dane $d_1, d_2, \ldots, d_n$ wiążemy z każdą z nich wagi $\omega_1, \omega_2, \ldots, \omega_n$ i obliczamy stosunek

$$\frac{d_1\omega_1 + d_2\omega_2 + \ldots + d_n\omega_n}{\omega_1, \omega_2, \ldots, \omega_n}$$

A oto kilka przykładów ilustrujących zastosowanie tej średniej.

PRZYKŁADY

1. Przypuśćmy, że pewien nauczyciel matematyki przeprowadził w ciągu semestru trzy sprawdziany, z których drugi jest dwa razy trudniejszy od pierwszego, a trzeci jest dwa razy trudniejszy od drugiego. Postanowił przy tym na podstawie tych sprawdzianów wystawić uczniom oceny na semestr. Pewien uczeń otrzymał z pierwszego sprawdzianu 4, a z drugiego 3 i z trzeciego 2. Jaką ocenę semestralną otrzyma?
Zauważmy, że trzeci sprawdzian jest cztery razy trudniejszy od pierwszego. Zastosujemy więc średnią ważoną o wagach 1, 2, 4:

$$\frac{1\cdot 4 + 2\cdot 3 + 4\cdot 2}{1 + 2 + 4} = \frac{18}{6} = 3$$

Uczeń powinien otrzymać ocenę dostateczną.

1. W sklepie sprzedawano żarówki 60 W w cenie 1 zł, 75 W w cenie 1,5 zł oraz 100 W w cenie 2 zł. Kierownik sklepu ustalił, że za każde cztery sprzedane żarówki 60 W przypadają trzy sprzedane żarówki 75 W oraz jedna 100 W. Jaki jest średni przychód ze sprzedaży tych żarówek, liczony na jednostkę towaru?
Obliczamy średnią ważoną o wagach 4, 3, 1:

$$\frac{4\cdot 1 + 3\cdot 1{,}5 + 1\cdot 2}{4 + 3 + 1} = \frac{10{,}5}{8} = 1{,}31$$

Średni przychód wyniósł 1,31 zł.

ZADANIA

1. Oblicz średnią arytmetyczną, medianę i modę następujących zbiorów danych:
   a) 6 6 10 10 10 14 14
   b) 122 134 152 173 180 180 180
   c) 24 24 25 31 40 50 63 79
2. Uczeń otrzymał następujące oceny z kolejnych sprawdzianów:
   3 2 2 1 3 5 4 2 3 2
   Oblicz medianą, średnią arytmetyczną i modę tych danych.
   Która ze średnich preferować będzie uczeń, mówiąc o przeciętej ocen?
   Która ze średnich najbardziej zainteresuje rodziców?

3. Poniżej dane stanowią liczby liter w trzydziestu kolejnych wyrazach pewnej książki:
   35 61 27 55 47 73 36 61 42 61
   46 52 49 54 60 32 42 36 36 54
   26 32 29 51 61 49 57 28 32 28
   a) Jaka jest modalna liczba liter w wyrazie?
   b) Jaka jest średnia liczba liter w wyrazie?
   c) Znajdź medianę tego zbioru danych

4. Znajdź średnią ważoną wartości:
   a) 25 i 50 o wagach 2 i 3
   b) 20, 28, 36 i 40 o wagach odpowiednio 1, 2, 3, 4 i 5

MIARY ROZPROSZENIA

Średnie opisują zbiory danych, ale nie jest to opis pełny. Bardzo istotne jest, w jaki sposób dane są ułożone dokoła średnich. Ilustruje to następujący przykład.

PRZYKŁAD

1. Weźmy pod uwagę cztery zakłady pracy A, B, C i D, z których każda zatrudnia po 10 ludzi. W poniższej tabeli podaliśmy liczbę pracowników, którzy zarabiają pewne sumy pieniędzy:

Zarobek	A	B	C	D
200	2	5	1	0
400	2	0	2	5
600	2	0	4	0
800	2	0	2	5
100	2	5	1	0

Średnia arytmetyczna jest we wszystkich zakładach równa 600 zł. Jednak sposób, w których poszczególne dane są rozłożone wokół tej średniej, jest zupełnie inny. Jednak ze sposobów ich porównania jest tak zwane odchylenie średnie, obliczane jako średnia arytmetyczna wartości bezwzględnej odchyleń poszczególnych danych od średniej. Odchylenia te są równe:
dla 200 - 400
dla 400 - 200
dla 600 0
dla 800 200
dla 1000 400
Z tego zestawienia wynika, że odchylenie średnie jest równe:

$$\text{dla zakładu A } \frac{2\cdot 400 + 2\cdot 200 + 2\cdot 0 + 2\cdot 200 + 2\cdot 400}{10} = 240$$

$$\text{dla zakładu B } \frac{5\cdot 400 + 0\cdot 200 + 0\cdot 0 + 0\cdot 200 + 5\cdot 400}{10} = 400$$

$$\text{dla zakładu C } \frac{1\cdot 400 + 2\cdot 200 + 4\cdot 0 + 2\cdot 200 + 1\cdot 400}{10} = 160$$

$$\text{dla zakładu D } \frac{0\cdot 400 + 5\cdot 200 + 0\cdot 0 + 5\cdot 200 + 0\cdot 400}{10} = 200$$

W każdym przypadku mamy inne odchylenie średnie. Im większe jest odchylenie średnie, tym większe jest rozproszenie danych. Zatem w zakładzie C rozproszenie jest najmniejsze. Nie jest ono jednak najmniejsze z możliwych. Gdyby wszyscy pracownicy zarabiali po 600 zł, to odchylenie średnie byłoby równe zero

Wartość odchylenia średniego możemy zapisać symbolicznie

$$\frac{\Sigma |X - \overline{X}|}{N}$$

Zauważmy, że jeśli odchylenie średnie jest równe zero, to wszystkie dane równe są średniej arytmetycznej. Mamy więc wtedy dane skupione w jednym punkcie. Im większe jest odchylenie średnie tym bardziej dane mogą być odchylone od średniej arytmetycznej. Znacznie lepszymi, pod względem rachunkowym , narzędziami do określania miary rozproszenia danych są wariancja i odchylenie standardowe. W żadnej z nich nie stosuje się uciążliwej pod względem rachunkowym wartości bezwzględnej.

Def.

Wariancją σ^2 zmiennej X nazywamy średnią arytmetyczną kwadratów odchyleń danych

$$\sigma^2 = \frac{\Sigma |X - \overline{X}|}{N}$$

od średniej arytmetycznej. Zapisujemy to symbolicznie

Def.

Odchyleniem standardowym σ zmiennej X nazywamy pierwiastek kwadratowy ze średniej arytmetycznej kwadratów odchyleń danych od średniej arytmetycznej. Zapisujemy:

$$\sigma = \sqrt{\frac{\Sigma |X - \overline{X}|}{N}}$$

PRZYKŁADY

1. Obliczymy odchylenie standardowe dla danych:
   3 8 1 3 6 4 2 2 7
   Zaczniemy od obliczenia średniej arytmetycznej. Jest ona równa 4. Następnie budujemy tabelę odchyleń oraz kwadratów odchyleń danych od średniej:

Dana	3	8	1	3	6	4	2	2	7
Odchylenie	-1	4	-3	-1	2	0	- 2	-2	3
Kwadrat odchylenia	1	16	9	1	4	0	4	4	9

Odchylenie standardowe jest pierwiastkiem kwadratowym ze średniej arytmetycznej kwadratów odchyleń:

$$\sqrt{\frac{1 + 16 + 9 + 1 + 4 + 0 + 4 + 4 + 9}{9}} = \frac{4\sqrt{3}}{3} = 2{,}3$$

ZADANIA

1. Oblicz średnią oraz odchylenie standardowe każdego z poniższych zbiorów danych:
   a) 3 4 5 6 7 9 10 12
   b) 21 20 23 20 29 27 24
   c) 60 60 64 66 63 67 69
   d) 35 37 40 37 30 33 34 38

2. Czy odchylenie standardowe może być równe zero, jeśli nie wszystkie dane są równe zero, a średnia arytmetyczna jest równa zero?

3. Wykaż, że odchylenie standartowe jest nie mniejsze od odchylenia średniego.