Statystyka w szkole

Piotr Gumienny

statystyka (łac. *status* ‘stan rzeczy, państwo’)
1. nauka zajmująca się ilościowymi technikami badania prawidłowości procesów masowych.
2. wykazy informacji liczbowych.

Statystyka, nauka zajmująca się ilościowymi metodami badania prawidłowości zjawisk (procesów) masowych. Jej celem jest poznanie występujących prawidłowości, ich ilościowe wyrażenie oraz wyodrębnienie w nich składnika systematycznego i przypadkowego. Wyróżnia się statystykę opisową i statystykę matematyczną.
Pierwsza zajmuje się metodami zbierania i prezentacji informacji statystycznych i ich sumarycznego opisu przy wykorzystaniu takich parametrów, jak miary średnie oraz miary dyspersji. Druga, oparta na rachunku prawdopodobieństwa, umożliwia uogólnienie wyników badań, ocenę stopnia dokładności i wiarygodności wyników. Statystyka znajduje obecnie zastosowanie w wielu naukach, zarówno technicznych, jak i społecznych.

Dyspersja, rozproszenie, zmienność, zróżnicowanie jednostek zbiorowości statystycznej z uwagi na pewną cechę mierzalną. Im bardziej wartości cechy jednostek są skupione dookoła swej średniej, tym mniejsza jest dyspersja i odwrotnie – im bardziej są rozproszone, tym większa jest dyspersja. Liczbową ocenę dyspersji przeprowadzamy za pomocą miar dyspersji, z których najpowszechniej stosowane są: wariancja, odchylenie standardowe lub przeciętne, rozstęp (obszar zmienności), odchylenie ćwiartkowe (kwartylowe), współczynnik zmienności.

Statystyka matematyczna, nauka o budowie reguł wnioskowania o właściwościach badanej zbiorowości statystycznej na podstawie danych dotyczących części tej zbiorowości, wybranej w sposób losowy.

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 centymetrach.

2. Ocena jaką otrzymał ze sprawdzianu każdy uczeń, jest zmienną. Danymi w tym przypadku mogły być oceny: 1, 2, 3, 4, 5, 6.
   Oceny najniższej nie życzymy nikomu (hehe ☺).

Częstość

Częstość - charakteryzuje, ile razy określona dana wystąpiła w badaniu statystycznym.

Częstość względna – jest to stosunek częstości występowania danej do liczby wszystkich danych.

PRZYKŁAD

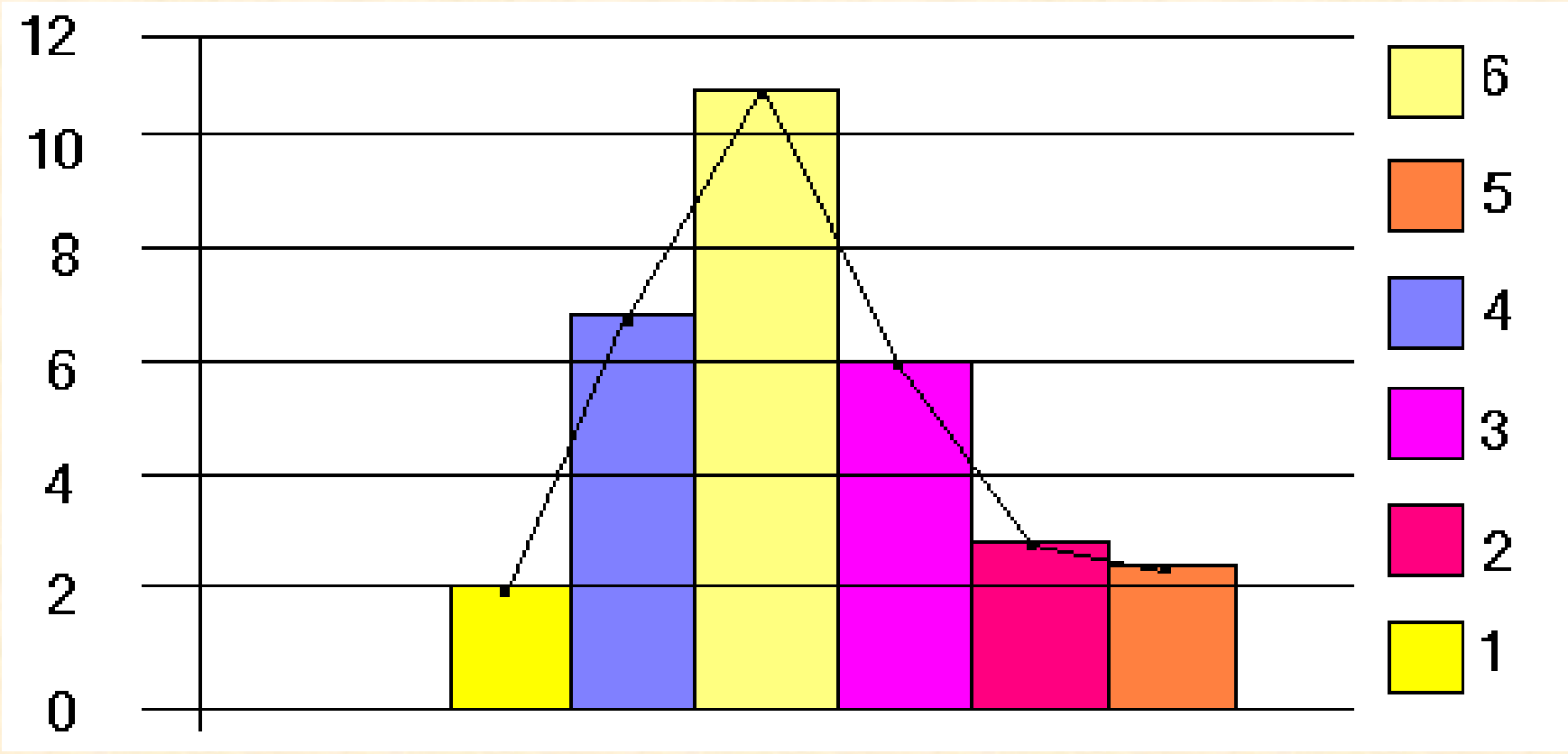
W pewnej klasie uczniowie otrzymali następujące oceny ze sprawdzianu:

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

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

Graficzna prezentacja danych

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**.


12
10
8
6
4
2
0
6
5
4
3
2
1

Często dane występują w postaci zbyt szczegółowej. Aby je lepiej analizować, łączymy je w tzw. **klasy**.

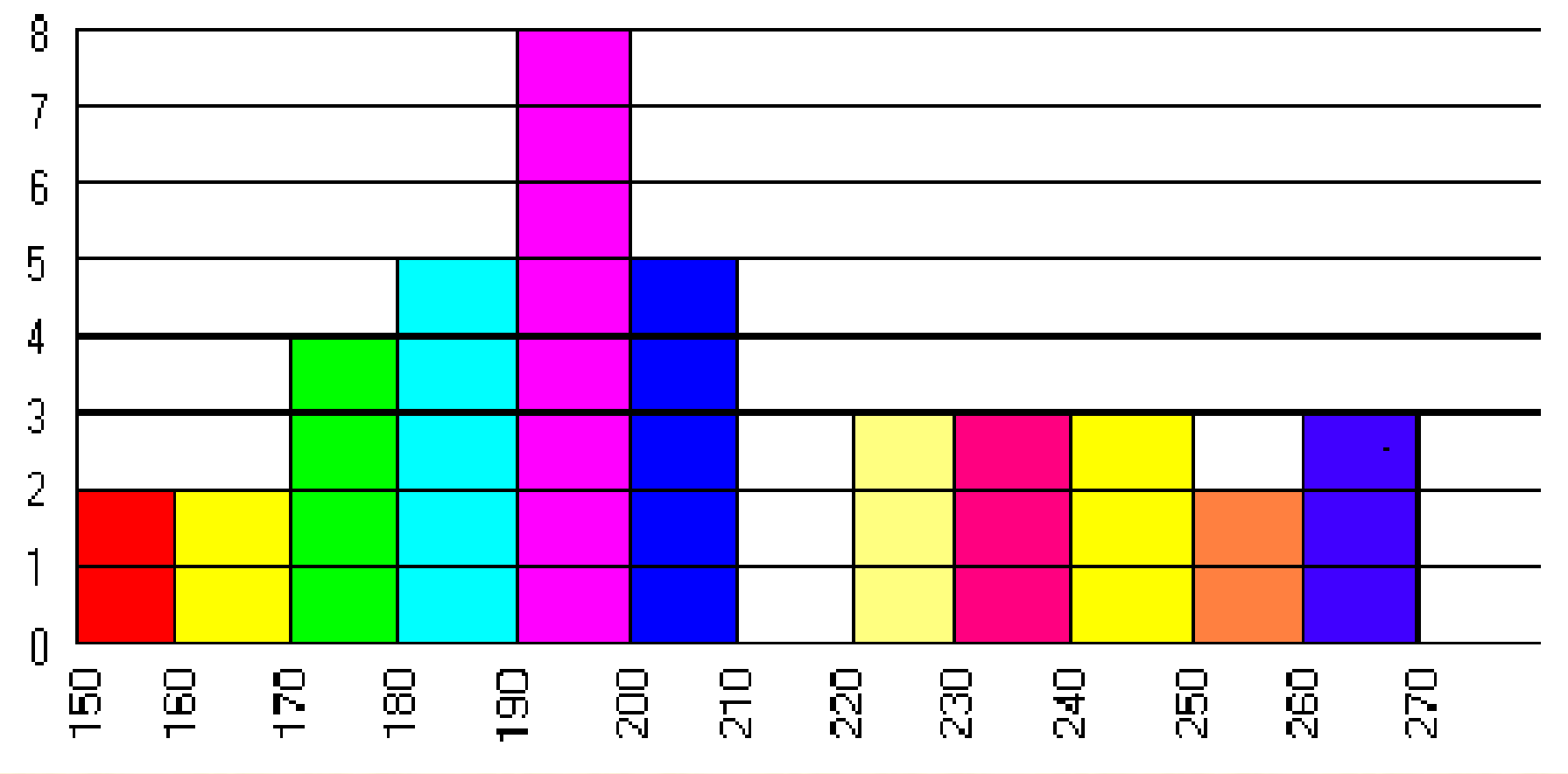
PRZYKŁAD

Oto zestawienie liczby dni deszczowych w wybranych stacjach meteorologicznych 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**


8
7
6
5
4
3
2
1
0
150
160
170
180
190
200
210
220
230
240
250
260
270

Miary tendencji centralnej

Przy opracowywaniu danych bardzo często stajemy przed koniecznością podania liczby charakteryzującej w jakiś sposób cały zbiór danych. Takie liczby nazywamy **średnimi**.

ŚREDNIA ARYTMETYCZNA

Średnią arytmetyczną obliczamy, dodając wszystkie dane, a następnie dzieląc otrzymaną sumę przez liczbę danych.

Jeśli zmienną będziemy oznaczać przez X, liczbę tych danych przez N to powyższe słowa możemy zapisać za pomocą następującego wzoru symbolicznego:

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

Mediana

Medianę, **czyli wartość środkową** znajdujemy 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.

Modalna

Modalna (moda, dominanta) - to wartość (dana) występująca najliczniej spośród wszystkich danych.

PRZYKŁAD

Uczeń otrzymał następujące oceny sprawdzianów:

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

1. Oblicz medianę, średnią arytmetyczną i modę tych danych.
2. Sporządź histogram.
3. Wyznacz częstość i częstość względną.
4. Która ze średnich preferować będzie uczeń, mówiąc o przeciętej/średniej ocen?

Średnie nie oddają charakteru danych i często prowadzą do „zafałszowania” rzeczywistości.

Przykład

Z punktu widzenia fizyka osoba trzymająca jedną nogę w wiadrze ze wrzątkiem, a drugą w wiadrze z zimna wodą „średnio statystycznie” powinna być zadowolona i powinno jej być „średnio ciepło”.

Przykład

W pewnym zakładzie pracuje 9 pracowników i dyrektor. Pracownicy zarabiają po 1000 zł, a dyrektor 15.000 zł.

„Średnia” płaca w tym zakładzie to 2.400 zł. !?!?

Miary dyspersji (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

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

Zarobek (w jednostkach monetarnych)	A	B	C	D
200	2	5	1	0
400	2	0	2	5
600	2	0	4	0
800	2	0	2	5
1000	2	5	1	0

Średnia arytmetyczna jest we wszystkich zakładach równa 600. Jednak sposób, w których poszczególne dane są rozłożone wokół tej średniej, jest zupełnie inny. Jednym 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

Wartość odchylenia średniego możemy zapisać symbolicznie

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

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

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

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

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, to odchylenie średnie byłoby równe zero.

Odchylenie standardowe

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

$$\sigma = \sqrt{\frac{\sum \left(X - \overline{X} \right)^2}{N}}$$

PRZYKŁAD

Obliczyć 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:

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

Odchylenie standardowe można też liczyć ze wzoru (równoważnego poprzedniemu)

$$\sigma = \sqrt{\frac{\sum X^2}{N} - \overline{X}^2}$$

PRZYKŁAD

Dla poprzedniego przykładu:

Dana	**3**	**8**	**1**	**3**	**6**	**4**	**2**	**2**	**7**
Kwadrat danej	**9**	**64**	**1**	**1**	**36**	**16**	**4**	**4**	**49**

Średnia arytmetyczna = 4

Kwadrat średniej arytmetycznej =16

$$\sigma = \sqrt{\frac{\sum X^2}{N} - \overline{X}^2} = \sqrt{\frac{9+64+1+9+36+16+4+4+49}{9} - 16}$$

$$= \sqrt{\frac{64}{3} - 16} = \frac{4\sqrt{3}}{3}$$

Interpretacja odchylenia standardowego

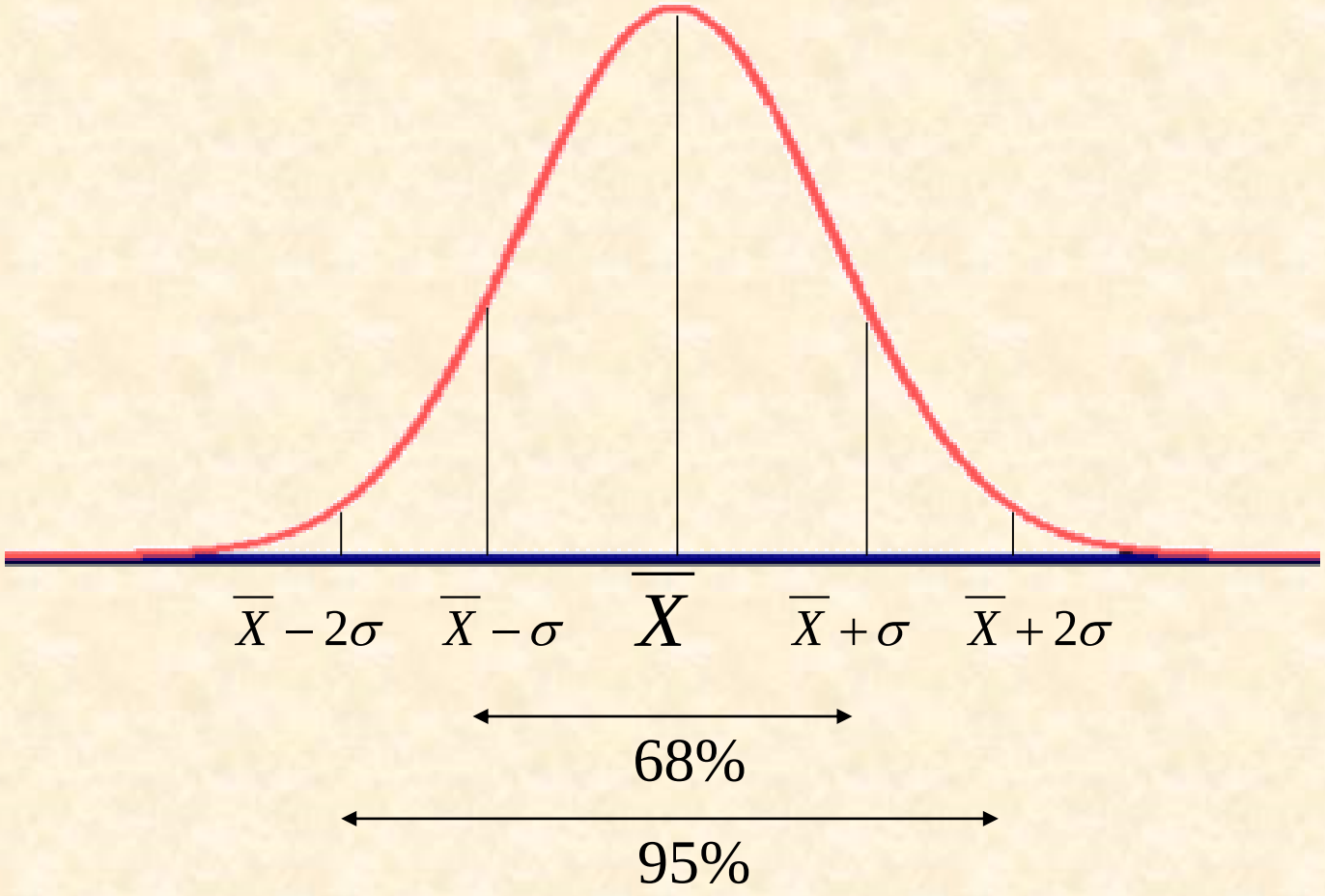
Zgodnie z klasyczną interpretacją dla tzw. rozkładu normalnego w przedziale

$\left(\overline{X}-\sigma, \overline{X}+\sigma\right)$ „mieści się” 68% wyników

A w przedziale

$\left(\overline{X}-2\sigma, \overline{X}+2\sigma\right)$ „mieści się” 95% wyników

Krzywa dzwonnowa Gaussa



PRZYKŁAD

CKE opublikowała wyniki egzaminu maturalnego z pewnego przedmiotu podając wynik średni 62% z odchyleniem standardowym 10%. Oceń wyniki ucznia, który uzyskał :

a) 65 %

b) 85 %

c) 45 %

CKE opublikowała wyniki egzaminu maturalnego z pewnego przedmiotu podając wynik średni 45% z odchyleniem standardowym 10%. Oceń wyniki ucznia, który uzyskał :

a) 65 %

b) 85 %

c) 45 %

Miary położenia w ujęciu pozycyjnym

Średnie i miary dyspersji opisują zbiory danych, ale nadal nie jest to opis pełny. Na przykład dla ucznia ważny jest nie tylko wynik, ale „pozycja” jaką z tym wynikiem zajmuje w grupie (populacji).

PRZYKŁAD

Rozpatrzmy wyniki ucznia (w pkt.) na tle pewnej populacji z dwóch przedmiotów A i B. Wyniki badanego ucznia zaznaczono kolorem

PRZEDMIOT A

9,11,15,20,24,26,30,35,35,36,37,37,38,39,**40**,44,48,50

PRZEDMIOT B

26,**40**,41,41,42,43,44,44,45,46,46,46,47,48,49,49,50,50

Wynik punktowy ucznia w obu sprawdzianach jest taki sam, jednak jego POZYCJA w grupie (na tle populacji) jest zupełnie inna !!!

Bazując na pokazanym przykładzie, łatwo sobie wyobrazić sytuację, że uczeń z jednego przedmiotu (sprawdzianu) uzyskał więcej punktów niż z innego, a mimo to OSIĄGNĄŁ ZNACZNIE GORSZY WYNIK na tle populacji !!!

Dlatego stosujemy miary położenia …

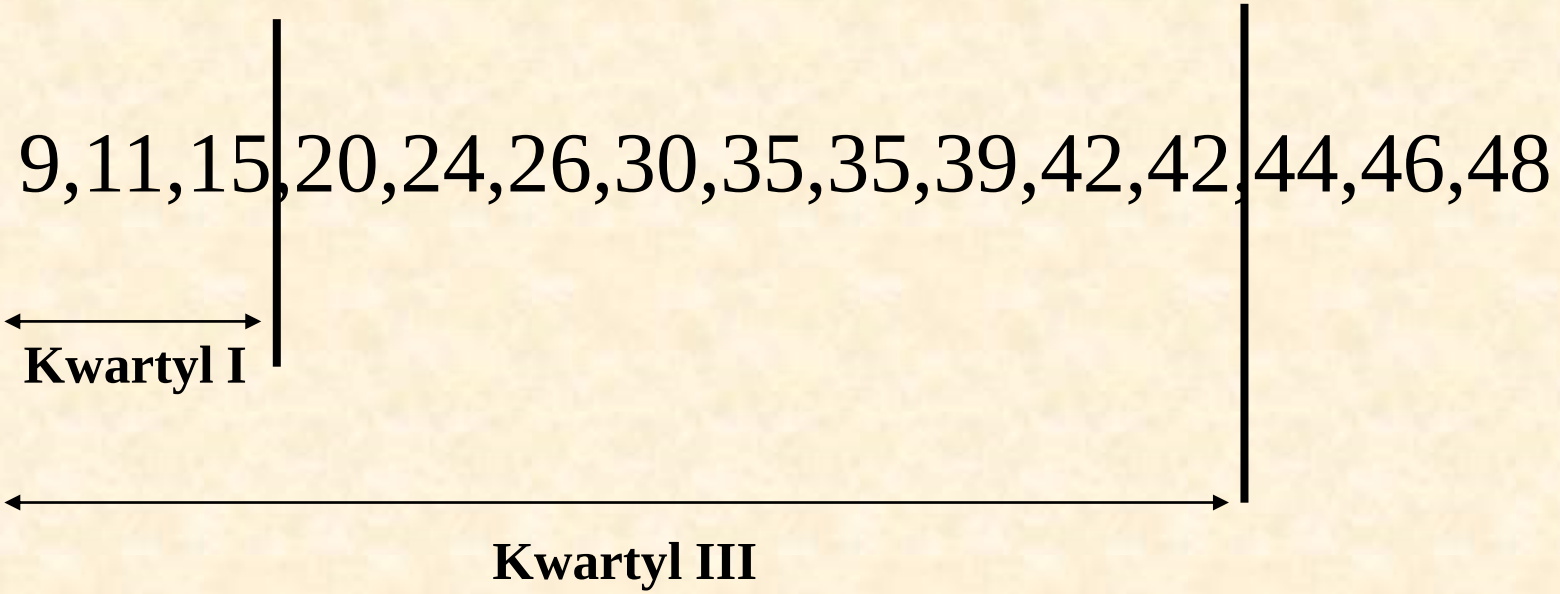
Kwartyle

Kwartyle dzielą zbiorowość pod względem liczebności na ćwiartki.

Kwartyl I identyfikuje „dolne” 25% jednostek (wyników)

Kwartyl III identyfikuje „górne” 25% jednostek (wyników)

Przykład



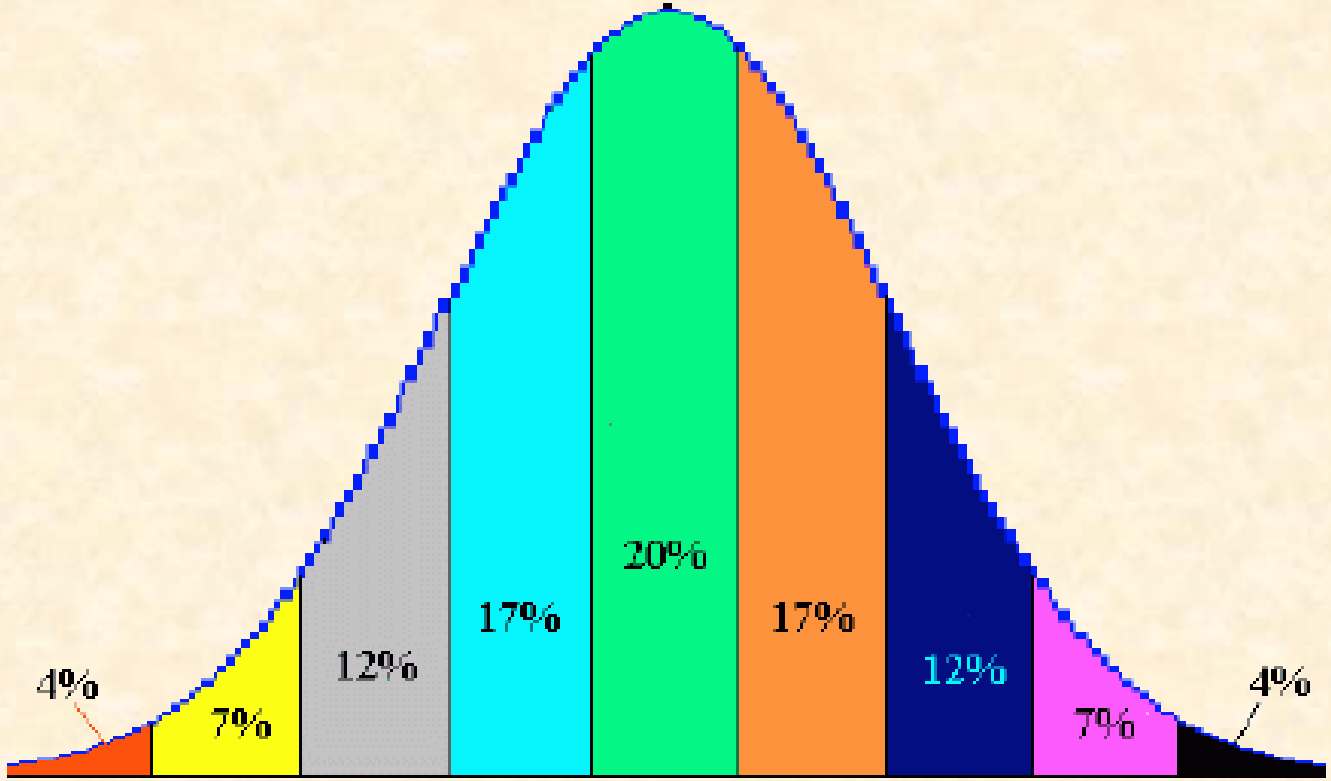
Na analogicznej zasadzie (tj. „pozycji w grupie” a nie „wartości wyniku”) opierają się inne skale (podziały) – np. centyle (podział na 100)

Z modyfikacji tej skali dla potrzeb m.in. pedagogiki skonstruowano tzw. **skalę staninową**

Skala staninowa wprowadza 9 przedziałów wyników. Poniższa tabela określa jaka **wyrażona procentowo liczba** wyników uszeregowanych rosnąco, zawiera się w określonym przedziale.

nazwa stanina	rangi w % liczby wyników	procent wyników zawarty w przedziale
1- najniższy	poniżej 4	4
2- bardzo niski	4-10	7
3- niski	11-22	12
4- niżej średni	23-39	17
5- średni	40-59	20
6- wyżej średni	60-76	17
7- wysoki	77-88	12
8- bardzo wysoki	89-95	7
9- najwyższy	powyżej 95	4

Graficzna interpretacja skali staninowej w kontekście krzywej dzwonnowej Gaussa (rozkład normalny)



Stanin 1 2 3 4 5 6 7 8 9

Zalety skali staninowej (Przy właściwym zrozumieniu i stosowaniu ☺)

- Łatwo czytelny poziom osiągnięć ucznia na tle grupy bez względu wynik punktowy/procentowy.
- Możliwość „porównania wyników” bez względu na przyjęte skale punktowe.
- Możliwość porównania osiągnięć uczniów z różnych przedmiotów i różnych lat.

Wady skali staninowej

- Test/sprawdzian musi być wystandaryzowany dla danej grupy (odpowiedni poziom trudności zadań, by wyniki „mogły się właściwie rozłożyć”)
- Nie określają poziomu umiejętności/wiadomości/wiedzy ucznia.

Dziękuję za uwagę ...