INTERPRETACJA DANYCH STATYSTYCZNYCH

LITERATURA

1. Statystyka. Elementy teorii i zadania.
2. S. Ostasiewicz, Z. Rusnak, U. Siedlecka, Wydawnictwo UE we Wrocławiu, Wrocław 2006.
3. Statystyka w zarządzaniu
4. A. Aczel, PWN, Warszawa 2011.
5. Statystyka
6. M. Sobczyk,, PWN, Warszawa 2014
7. Statystyka opisowa dla studentów administracji i prawa
8. A. Malinowski,, LIBER Warszawa 2006.

POJĘCIA WSTĘPNE

Statystyka jako dziedzina wiedzy to nauka o metodach zbierania, opracowywania i analizy danych dotyczących zjawisk i procesów masowych. Przedmiotem statystyki są ilościowe metody badania prawidłowości w zjawiskach masowych, jest to nauka podejmowania decyzji w warunkach niepewności, stąd podział na:

- statystykę opisową,
- statystykę matematyczną.

Statystyka opisowa to dział zajmujący się metodami opisu danych statystycznych uzyskanych podczas badania statystycznego. Celem stosowania metod statystyki opisowej jest podsumowanie zbioru danych i wyciągnięcie pewnych podstawowych wniosków i uogólnień na temat zbioru. Stanowi wstępne opracowanie wyników pomiarów z próby bez posługiwania się rachunkiem prawdopodobieństwa mającym na celu wykrywanie zachodzących prawidłowości.

Natomiast **statystyka matematyczna** zajmuje się podejmowanie poprawnych decyzji dotyczących populacji generalnej z wykorzystaniem rachunku prawdopodobieństwa.

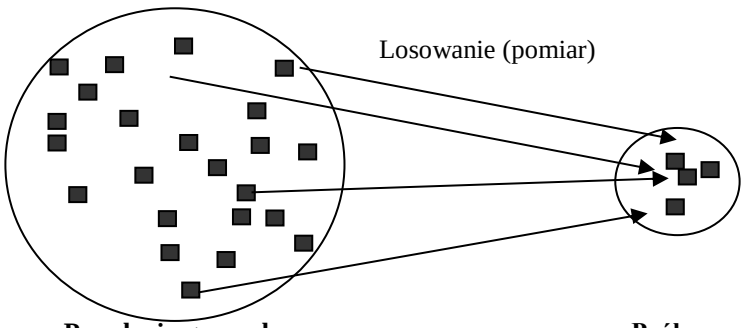
Statystyczne metody wykrywania prawidłowości i podejmowania decyzji w warunkach niepewności zakładają, że podjęcie decyzji dotyczącej badanego zjawiska jest poprzedzone **badaniem statystycznym** czyli wykonaniem dostatecznej liczby obserwacji w podobnych warunkach i określonym czasie.

Rozróżnia się:

- Badanie pełne (całkowite), które obejmuje wszystkie jednostki zbiorowości statystycznej;
- Badanie niepełne (częściowe), które obejmuje niektóre jednostki zbiorowości statystycznej, które są elementami próby wyodrębnionej w określony sposób.

Jednak nie zawsze badanie pełne oparte na całej zbiorowości jest przeprowadzane. Decyzja o przeprowadzeniu badania częściowego może być podjęta ponieważ:

- Zbiorowość statystyczna jest zbyt liczna (koszty i czas);
- Badanie ma charakter niszczący;
- Chodzi jedynie o wyniki orientacyjne.



Zbiorowością statystyczną lub populacją generalną nazywa się zbiór wszystkich jednostek statystycznych mających przynajmniej jedną cechę stałą oraz pewną liczbę cech zmiennych. Cechy stałe decydują o zaliczeniu jednostek do określonej zbiorowości statystycznej, natomiast cechy zmienne powodują zróżnicowanie poszczególnych jednostek zbiorowości. Warto podkreślić, że cel badania statystycznego determinuje zbiorowość statystyczną np.: (tzn. jeżeli chcemy zbadać wartość produkcji przedsiębiorstw przemysłu chemicznego w kraju to zbiorowością są wszystkie zakłady chemiczne w Polsce natomiast jeśli badanie dotyczy wysokości zarobków pracowników przemysłu chemicznego, to zbiorowość stanowią wszystkie osoby zatrudniane w tej branży). Zbiorowości statystyczne można scharakteryzować bardziej szczegółowo wyróżniając populacje:

- skończone i nieskończone,
- jednorodne i niejednorodne,
- symetryczne i niesymetryczne.

Próbą lub populacją próbną nazywamy wyodrębnioną przy pomocy odpowiedniej metody statystycznej część zbiorowości statystycznej. Podzbiór ten podlega badaniu, a uzyskane wyniki są uogólniane na populację generalną. Liczbę elementów próby nazywaną liczebnością próby oznacza się n, przy czym $n < N$.

Ogół metod doboru próby do badania dzielimy na:

- metody statystyczne (metody doboru losowego),
- metody nie statystyczne (metody doboru nielosowego)

Metody statystyczne związane są z losowym doborem próby charakteryzującym się tym, że dokonując losowania poszczególnych jednostek, każda jednostka zbiorowości musi mieć takie same szanse wejścia do próby. Próba spełniająca postulat losowego wyboru nazywana jest próbą prostą i jej struktura jest podobna do struktury całej zbiorowości. **Metody nie statystyczne** to metody w których wybór opiera się na osądzie statystyka, wynikającym z jego doświadczenia oraz wiedzy. Polegają na wyborze konkretnych jednostek o ustalonych z góry charakterystykach – stąd często wymagana jest znajomość parametrów charakteryzujących populację. W metodach tych występuje duża doza subiektywizmu związanego z samym dobieraniem jednostek do próby, przy jednoczesnym zachowaniu obiektywizmu wyboru kryteriów doboru. Nie rządzi tu zatem przypadek, ponieważ dobór nielosowy jest związany z takim sposobem postępowania, w którym ostateczny wybór jednostek powierza się subiektywnym decyzjom osób przeprowadzających badania. Zastosowanie tego typu metod nie pozwala na uogólnienie na całą zbiorowość.

Podstawowym i oczywistym postulatem współczesnej metodologii badania opinii publicznej jest gwarancja reprezentatywności próby (wyboru grupy osób dla przeprowadzenia sondażu). Próbą reprezentacyjną jest próbą losową dobrze odzwierciedlającą strukturę i relacje zachodzące w danej zbiorowości, wówczas wyniki badania dostarczą jak najwięcej informacji o prawidłowościach w badanej populacji.
Aby mieć próbę reprezentacyjną muszą być spełnione dwa warunki:

- próba powinna być dostatecznie liczna,
- każda jednostka danej zbiorowości statystycznej powinna mieć jednakową szansę trafienia do próby.

W celu przeprowadzenia badań statystycznych wyodrębniane są obiekty (np.: osoby, rzeczy, zjawiska) nazywane **jednostkami statystycznymi**. Warto jednocześnie podkreślić, że cel badania determinuje sposób wyodrębniania jednostek (tzn. jeżeli chcemy zbadać wartość produkcji przedsiębiorstw przemysłu chemicznego w kraju to jednostkami statystycznymi są poszczególne zakłady chemiczne w Polsce natomiast jeśli badanie dotyczy wysokości zarobków pracowników przemysłu chemicznego, to jednostkami statystycznymi będą osoby zatrudniane w tej branży). Niedokładne określenie jednostek statystycznych może spowodować nieporównywalność otrzymanych wyników. Właściwość jednostek statystycznych, która podlega badaniu statystycznemu nazywana jest **cechą statystyczną**. W zależności od celu badania w analizach uwzględnia się tylko te cechy, które są istotne dla zjawisk będących przedmiotem analiz.
Cechy statystyczne dzielą się na:

- ilościowe (mierzalne),
- jakościowe (niemierzalne).

Cechą ilościową nazywa się taką cechę, która może być wyrażona za pomocą liczby pochodzącej z pomiaru lub obliczenia. Wśród cech ilościowych wyróżnia się cechy skokowe (dyskretne) oraz ciągłe. Cecha skokowa

może przyjmować wartości z pewnego skończonego lub przeliczalnego zbioru liczb. Cecha ciągła jest cechą przyjmującą dowolne wartości liczbowe z pewnego nieprzeliczalnego zbioru (waga ciała, temperatura).
Cechą jakościową nazywa się taką cechę, która może być wyrażona jedynie za pomocą wyrażenia słownego. Wśród cech niemierzalnych (jakościowych) wyróżnia się tzw. cechy quasi-mierzalne tzn. są to takie cechy, które w sposób prosty można przekodować na cechy mierzalne jak np. płeć jest cechą 0-1.

Dana cecha ilościowa przyjmuje na ogół różne wartości dla różnych jednostek statystycznych, podobnie cecha jakościowa u różnych jednostek występuje w różnych kategoriach. Jeśli w określonym zbiorze jednostek statystycznych cecha przyjmuje różne wartości liczbowe lub różne kategorie jakościowe to jest to **cecha zmienna**. Jeśli cecha przyjmuje tę samą wartość liczbową lub występuje w tej samej kategorii jakościowej to jest to **cecha stała**. Wyróżnia się również cechy tzw. **quasi-stałe** charakteryzujące się zbyt małym zróżnicowaniem.

Zbiorowość statystyczna	**Jednostka statystyczna**	**Cecha statystyczna**	**Charakter cechy statystycznej**
Pracownicy firmy X w Polsce w 2013 roku	pracownik	• staż pracy, • Wiek (w latach), • wydajność pracy, • płaca, • Wykształcenie	• mierzalna - ciągła • mierzalna -skokowa • mierzalna - ciągła • mierzalna - ciągła • niemierzalna
Studenci studiów inżynierskich	Student studiów inżynierskich	• Wzrost (cm), • Waga (kg), • płeć, • średnia ocen	• mierzalna -skokowa • mierzalna -skokowa • quasi-mierzalna • mierzalna - ciągła
Gospodarstwa domowe w Polsce w 2013 roku	gospodarstwo domowe	• dochód, • wydatki, • liczba dzieci	• mierzalna - ciągła • mierzalna - ciągła • mierzalna -skokowa

ORGANIZACJA BADAŃ STATYSTYCZNYCH.

Badania statystyczne (zarówno całkowite jak i częściowe) są niejednokrotnie złożonym przedsięwzięciem organizacyjnym, składającym się z odrębnych etapów:

- Przygotowanie badania (określenie celu, zbiorowości/próby, jednostki statystycznej objętej badaniem oraz źródła danych i metody badania)
- Pomiar i zebranie materiału statystycznego.
- Opis statystyczny.
- Analiza materiału statystycznego i wnioskowanie statystyczne.

Materiał statystyczny to zbiór obserwacji, który ze względu na źródła informacji dzieli się na materiał pierwotny i wtórny. **Pierwotne** źródła gromadzenia informacji obejmują te wszystkie źródła, które zostały przygotowane specjalnie dla badania wybranego problemu. Podstawowymi pierwotnymi źródłami informacji są przede wszystkim studia empiryczne, takie jak obserwacja i badania wykorzystujące kwestionariusze (ankiety). W przypadku źródeł pierwotnych należy dokonać pomiaru wartości cechy z wykorzystaniem różnych skal pomiarowych: nominalna; porządkowa (rangowa); przedziałowa; ilorazowa. **Skala nominalna** daje najmniej precyzyjny sposób pomiaru, ponieważ liczby (symbole) w tej skali pełnią rolę umowną i służą do identyfikacji i klasyfikacji jednostek statystycznych. Podstawą zaliczenia jednostki do danej kategorii jest fakt posiadania określonego wariantu cech (można tu tylko powiedzieć, że warianty te są lub nie są jednakowe):

- płeć,
- trasy autobusów,
- numer sali wykładowej.

Skala porządkowa posiada wszystkie cechy skali nominalnej, a dodatkowo pozwala na porządkowanie jednostek w ramach poszczególnych kategorii pod względem natężenia cechy np.:

- 10-stopniowa skala Mohsa mierząca twardość minerałów,
- 9-stopniowa skala Richtera mierząca siłę trzęsienia ziemi,
- Stopnie wojskowe.

Skala przedziałowa – ma wszystkie cechy skali porządkowej, ale pozwala na mierzenie dystansu między jednostkami. Jednakowym różnicom między stopniami własności badanych jednostek odpowiadają jednakowe różnice w liczbach opisujących to natężenie np.:

- lata kalendarzowe;
- Skale temperatur (Celsjusz, Fahrenheit);
- Indeks cen.

Skala ilorazowa (stosunkowa) ma własności poznanych trzech skal. Dodatkowo charakteryzuje się posiadaniem naturalnego punktu zero, który oznacza brak danej cechy np.:

- wiek;
- ceny towarów;
- długość;
- ciężar

Wtórne źródła gromadzenia informacji obejmują te wszystkie źródła, które nie zostały opracowane z myślą o badanym problemie. Głównymi wtórnymi źródłami informacji są przede wszystkim:

- publikacje organów państwowych,
- publikacje placówek naukowo-badawczych,
- materiały wewnętrzne przedsiębiorstw,
- biuletyny agencji badań opinii publicznej lub badań rynkowych.

Korzystając ze sprawozdawczości statystycznej (źródła wtórne) nie dokonujemy pomiarów tylko gromadzimy dane.

AGREGACJA DANYCH

W wyniku obserwacji statystycznej otrzymujemy zbiór danych, zwanych danymi statystycznymi. Należy przy tym podkreślić że w przypadku cech ilościowych wartość cechy oznacza wartość liczbową (liczbę), natomiast w przypadku cech jakościowych termin ten oznacza dowolną z możliwych kategorii opisu słownego. Wartości cechy statystycznej X oznacza się małymi literami: x_i, $i = 1, \dots, n$. Natomiast jeśli jednostki statystyczne zbadano pod względem kilku cech wówczas obserwacją statystyczną nazywa się odpowiedni wektor wartości $[x_i, y_i, \dots, z_i]$ gdzie $i = 1, \dots, n$.

Analizę materiału statystycznego musi poprzedzić jego **opis statystyczny**, którego elementami są:

- systematyzacja - porządkowanie i grupowanie, polega na (mniej lub bardziej zróżnicowanym) podziale niejednorodnej zbiorowości na możliwie jednorodne grupy według obranych kryteriów, charakteryzujących poszczególne grupy, i odpowiednim zestawieniu danych statystycznych.
- klasyfikacja - przeprowadza się zazwyczaj według wybranych cech, których prawidłowa analiza jest możliwa dopiero w ramach otrzymanych jednorodnych grup.
- streszczenie materiału statystycznego za pomocą kilku miar (wskaźników),
- prezentacja graficzna.

Zbiór wyników obserwacji uporządkowanych według określonych cech (kryteriów) nazywamy szeregiem statystycznym. Najczęściej wyróżnia się dwa kryteria podziału szeregów:

- **kryterium formalne**, związane z budową szeregu, na podstawie którego możemy wyodrębnić: szeregi szczegółowe, szeregi rozdzielcze i szeregi kumulacyjne,
- **kryterium merytoryczne**, wynikające z typu badanej cechy zbiorowości, wg którego wyróżnia się szeregi czasowe i szeregi przestrzenne.

Podziały te jednak nie wykluczają się wzajemnie, gdyż np.: szereg rozdzielczy może być jednocześnie szeregiem czasowym lub przestrzennym.

Szeregiem szczegółowym nazywamy uporządkowany, wyłącznie według wartości badanej cechy, zbiór danych. Porządkowanie polega na ustawieniu wartości określonej cechy danej zbiorowości lub próby według kolejności rosnącej lub malejącej. Szereg szczegółowy obejmuje wartości zmiennych występujących u wszystkich jednostek badanej zbiorowości

Np.

1800,1900, 1900,1900, 1000, 2100, 2100, 2100, 2500, 2500, 2500, 2500, 2500, 2500, 3000, 3500.

Szereg rozdzielczy to zbiór wartości liczbowych uporządkowanych wg wariantów pewnej badanej cechy, przy czym poszczególnym wariantom zmiennej przyporządkowane są odpowiadające im liczebności. Szeregi rozdzielcze mogą dotyczyć zarówno cechy jakościowej, jak i ilościowej. Charakteryzują one strukturę danej zbiorowości stąd nazywane są czasem szeregami strukturalnymi.

Otrzymuje się go dzieląc zbiorowość statystyczną na klasy zbiorcze według pewnej cechy i podając liczebności każdej z tych klas, zwane liczebnościami klasowymi i oznaczonymi $n_i, i = 1, \dots, k$ gdzie k oznacza liczbę klas.

Szeregi rozdzielcze punktowe budowane są dla cechy skokowej:

x_i	n_i
19	1
20	4
21	2
22	1
23	1
19	1

Ocena lokalizacji	Liczba punktów sprzedaży
bardzo dobra	42
dobra	178
przeciętna	213
zła	41
bardzo zła	20

natomiast **szeregi rozdzielcze przedziałowe** przede wszystkim dla cechy ciągłej:

Liczba pracowników	Liczba punktów sprzedaży
do 3	290
3 - 4	85
5 -8	42
powyżej 8	77

Liczba osób w rodzinie	Liczba gospodarstw domowych	Gospodarstwa domowe (%)
1	3254736	23,99
2	3483408	25,67
3	2741982	20,21
4	2203028	16,24
5 i więcej	1884844	13,89
Razem	13567999	100,00

Budując szeregi rozdzielcze należy zdecydować o:

- liczbie klas,
- ich rozpiętości i sposobie określania granic przedziałów.

Należy pamiętać, że dobra klasyfikacja powinna spełniać dwa podstawowe warunki:

- musi być przeprowadzona w sposób rozłączny, co oznacza, że poszczególne jednostki o określonych cechach powinny być w sposób jednoznaczny przydzielone do poszczególnych klas (grup),
- musi być przeprowadzona w sposób zupełny, co oznacza, że klasy powinny objąć wszystkie cechy występujące w danej zbiorowości.

Do ustalenia orientacyjnej liczby klas (k) w zależności od liczebności próby (n) wykorzystuje się następujące reguły:

$$\boldsymbol{k \leq 5 \log n}$$

$$\boldsymbol{k \approx \sqrt{n}}$$

$$\boldsymbol{k = 1 + 3,332 \log n}$$

Rozpiętość (h) przedziału oblicza się wówczas według:

$$\boldsymbol{h = \frac{x_{max} - x_{min}}{k}}$$

Przykład

PRZYKŁAD 1:
Dany jest szereg szczegółowy o miesięcznych zyskach 60 zakładów pracy (w tys.zł):

2	27	18	32	36
6	20	17	39	34
5	19	16	29	23
12	18	15	15	40
7	15	17	29	39
10	10	19	30	50
5	12	23	33	45
15	11	25	12	43
14	16	44	34	8
13	9	28	42	37
7	19	33	34	39
11	27	31	46	42

W celu dokonania agregacji danych należy wyznaczyć:
Liczbę klas (w trzech wariantach w zależności od sposobu) otrzymano równą odpowiednio:

$$\boldsymbol{k \leq 5 \log n = 5 \log 60 = 8,89 \approx 9}$$
$$\boldsymbol{k = \sqrt{n} = \sqrt{60} = 7,74 \approx 8}$$
$$\boldsymbol{k = 1 + 3,332 \log n = 1 + 3,332 \log 60 = 6,92 \approx 7}$$

Rozpiętość:

$$\boldsymbol{h = \frac{x_{max} - x_{min}}{k}}$$

$$\boldsymbol{h = \frac{50-2}{9} = 5,33 \approx 5} \qquad \mathbf{h = \frac{50-2}{8} = 6} \qquad \boldsymbol{h = \frac{50-2}{7} = 6,85 \approx 7}$$

Otrzymane szeregi rozdzielcze:

k=9, h=5	
x_i	n_i
2-7	4
7-12	8
12-17	11
17-22	8
22-27	3
27-32	7
32-37	7
37-42	5
42-50	7

k=8, h=6	
x_i	n_i
2-8	6
8-14	10
14-20	14
20-26	4
26-32	7
32-38	8
38-44	7
44-50	4

k=7, h=7	
x_i	n_i
2-9	7
9-16	14
16-23	10
23-30	8
30-37	9
37-44	8
44-51	4

Dodatkowo do prezentacji danych mogą służyć wykresy statystyczne. Do najbardziej popularnych należą:

- histogram, czyli zbiór prostokątów, których podstawy wyznaczone są na osi OX przez rozpiętość poszczególnych przedziałów, a wysokości określone są na osi OY przez liczebności odpowiadające poszczególnym przedziałom;
- diagram, który otrzymuje się w wyniku połączenia punktów będących środkami przedziałów i odpowiadających im liczebności;

Histogram to sposób przedstawiania rozkładu empirycznego cechy statystycznej. Składa się z szeregu prostokątów umieszczonych na osi współrzędnych. Prostokąty te są z jednej strony wyznaczone przez przedziały klasowe wartości cechy, natomiast ich wysokość jest określona przez:

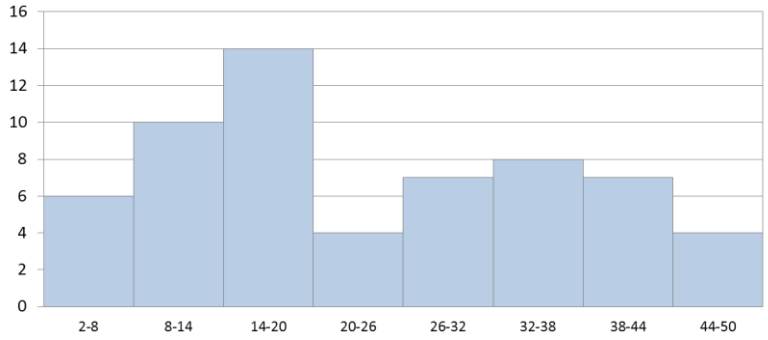
- liczebności
- częstości,
- gęstość prawdopodobieństwa

elementów należących do określonego przedziału klasowego.

CD. PRZYKŁAD 1:
Dla otrzymanych szeregów rozdzielczych otrzymano następujące histogramy:

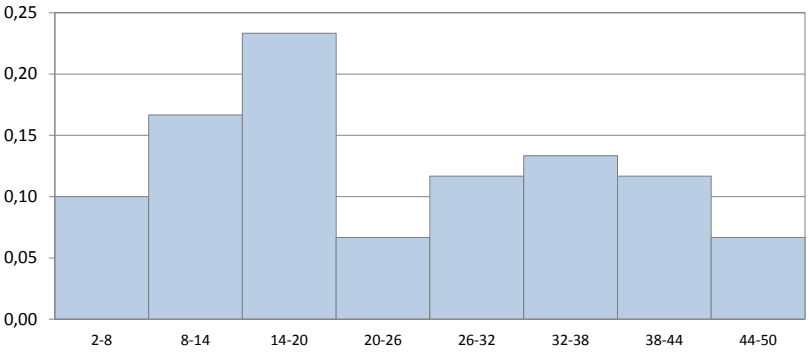
Liczebności:

k=8, h=6	
x_i	n_i
2-8	6
8-14	10
14-20	14
20-26	4
26-32	7
32-38	8
38-44	7
44-50	4



Częstości

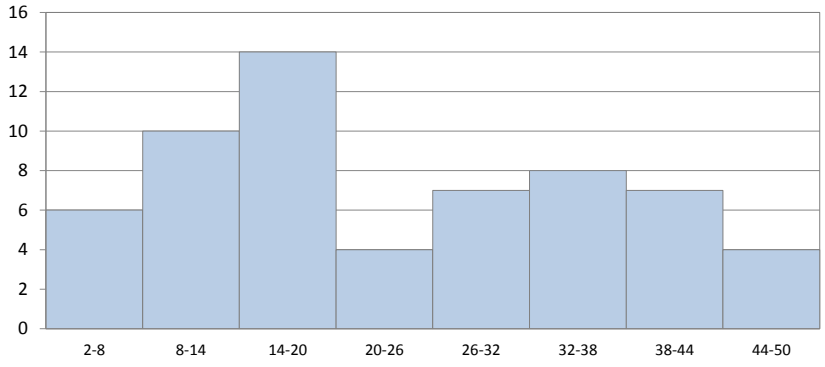
k=8, h=6	
x_i	ω_i
2-8	0,10
8-14	0,17
14-20	0,23
20-26	0,07
26-32	0,12
32-38	0,13
38-44	0,12
44-50	0,07



Wysokość „słupków" to częstość, którą wyznaczamy wg wzoru: $\omega_i = \frac{n_i}{n}$ - ich suma =1.

Prawdopodobieństwa

k=8, h=6	
x_i	p_i
2-8	0,02
8-14	0,03
14-20	0,04
20-26	0,01
26-32	0,02
32-38	0,02
38-44	0,02
44-50	0,01



W histogramie tym pole powierzchni poszczególnych „słupków" =1.
A prawdopodobieństwo wyznacza się ze wzoru: $p_i = \frac{\omega_i}{h_i}$

ANALIZA STRUKTURY
SYNTETYCZNY OPIS ZBIOROWOŚCI STATYSTYCZNEJ

Celem analizy rozkładu jest syntetyczna charakteryzacja zbioru danych niezależnie od tego czy zbiór danych dotyczy wszystkich jednostek badanej populacji, czy też jej części. Obecnie chcemy jedynie ukazać, jak syntetycznie i czytelnie zaprezentować to co najistotniejsze w analizowanym zbiorze danych. Najczęściej wykorzystywane charakterystyki opisowe to:

1. Miary położenia – służące do opisywania takiej wartości zmiennej wokół której skupiają się pozostałe wartości;
2. Miary rozproszenia - służące do badania stopnia zróżnicowania wartości zmiennej;
3. Miary asymetrii – służące do badania kierunku zróżnicowania;
4. Miary koncentracji - służące do badania stopnia skupienia poszczególnych jednostek wokół średniej

MIARY POŁOŻENNIA

Podstawowymi parametrami opisującymi w sposób syntetyczny zasadnicze właściwości pewnych zbiorowości są miary położenia, zwane też miarami tendencji centralnej. Przedstawiają one centrum, środek lub najbardziej typową wartość w zbiorze danych liczbowych. Najprostszym, choć niedoskonałym, opisem badanej zbiorowości może być określenie średniego stanu wartości badanych cech za pomocą jednej syntetycznej miary. Miary położenia można obliczać dla wszystkich elementów badanej zbiorowości (populacji) lub dla jej części. Od celu i założeń określonego badania zależy czy dany zbiór jest traktowany jako populacja, czy też jako próba.

MIARY POŁOŻENIA	
KLASYCZNE	**POZYCYJNE**
• średnia arytmetyczna • średnia harmoniczna • średnia geometryczna	• dominanta • kwantyle (mediana, kwartyle, decyle, percentyle)

ŚREDNIA ARYTMETYCZNA

Najprostszą miarą położenia jest średnia arytmetyczna, którą otrzymuje się przez podzielenie sumy wartości odpowiadających wszystkim elementom zbioru (wszystkich obserwacji) przez liczbę elementów, które występują w tym zbiorze. Jeżeli średnią z wartości $x_1, x_2, x_3, ..., x_n$ oznaczymy symbolem $\overline{X}$ to obliczamy ją według jednego ze wzorów w zależności od sposobu agregacji danych:

Szereg szczegółowy:

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

Szereg rozdzielczy punktowy (k – klas):

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

Szereg rozdzielczy przedziałowy:

$$\overline{x} = \frac{\sum_{i=1}^{k} n_i \dot{x}_i}{n}$$

gdzie: x_i – kolejne wartości badanego zbioru,

n – liczebność zbioru danych.

Średnią wykorzystujemy, gdy chcemy scharakteryzować zbiór wyników obserwacji (zbiorowość) jako całość za pomocą jednej wartości. Jej obliczanie jest przydatne tylko wtedy, gdy zbiór wyników obserwacji jest jednorodny tzn. gdy wchodzące do niego obserwacje (jednostki) są tego samego rodzaju. Jednorodnym zbiorem wyników są np. dane dotyczące kontroli jakości jednego typu wędlin we wszystkich

sklepach mięsnych we Wrocławiu. Natomiast zbiór niejednorodny tworzą wyniki kontroli jakości wszystkich rodzajów wędlin.

PRZYKŁAD 1a:
Mając informacje o wieku osób pracujących w sekcji płac pewnej firmy w postaci zbioru liczb: 27, 32, 38, 38, 38, 43, 47 i 49 lat, średnią wieku obliczamy według podanego wzoru sumując:

$$27+32+38+38+38+43+47+49 = 312,$$

a następnie dzieląc sumę przez 8 gdyż tyle osób jest zatrudnionych w tej sekcji, co daje

$$312: 8= 39.$$

Zatem średni wiek osoby pracującej w sekcji płac w badanej firmie wynosi **39** lat.

PRZYKŁAD 1b:
Odnotowano następujące kursy sprzedaży USD w dziesięciu kantorach w Legnicy (powiedzmy 20 sierpnia 2004 r.): 3,61; 3,62; 3,63; 3,63; 3;66; 3,65; 3,65; 3,65; 3;66; 3,67.
Średni kurs USD w tym dniu w 10 kantorach Legnicy kształtował się na poziomie **3,643** zł.

Zauważmy, że otrzymany wynik 3,643 jest większy od zanotowanego najniższego kursu USD oraz mniejszy od zanotowanego najwyższego kursu USD dla wybranych kantorów Legnicy. Możemy to zapisać:

$$3,61=X_{min} < \overline{X} = 3,643 < X_{max} = 3,67$$

Taka nierówność jest zawsze spełniona, gdyż analizowany zbiór danych zawiera przynajmniej dwie różne liczby.

Średnie są wartościami umownymi, które faktycznie mogą wcale nie występować w badanym zbiorze. W żadnym z dziesięciu kantorów kurs USD nie pokrywał się ze średnim kursem. Analogiczna sytuacja wystąpiła w przykładzie dotyczącym przeciętnego wieku pracowników sekcji płac. Wyznaczony przeciętny wiek 39 lat nie odpowiada żadnej konkretnej osobie w zbiorze pracowników.

PRZYKŁAD 1c:
Po przeprowadzonym Narodowym Spisie Powszechnym Ludności i Mieszkań w 2002 roku postanowiono obliczyć przeciętną liczbę osób mieszkających w gospodarstwach domowych w Polsce. Dane zaczerpnięte z ostatniego Spisu Powszechnego przedstawiono w tabeli 2.

Gospodarstwa domowe w Polsce w 2002 r. według liczby osób

Liczba osób w gospodarstwie domowym x_i	Liczba gospodarstw domowych (w tys.) n_i	Obliczenia pomocnicze $x_i * n_i$
1	3307	3307
2	3097	6194
3	2654	7962
4	2405	9620
5	1086	5430
6	462	2772
7*	326	2282
sumy	**13337**	**37567**

Źródło: Rocznik Statystyczny 2003, GUS Warszawa, s…..
* ostatni wariant został domknięty, gdyż jego liczebność stanowiła niewiele ponad 2 procent ogólnej liczebności

Kolumna pierwsza podaje liczbę osób w gospodarstwie domowym (od jednej do siedmiu). Natomiast w kolumnie drugiej zamieszczono liczbę gospodarstw domowych, które zawierały kolejno jedną, dwie, trzy, cztery, pięć, sześć i siedem osób.
Przeciętną liczbę osób w gospodarstwach domowych obliczamy dzieląc ogólną sumę liczby osób we wszystkich gospodarstwach domowych przez sumę gospodarstw domowych:

$$\overline{X} = \frac{37567}{13337} = 2,82.$$

A więc przeciętna wielkość gospodarstwa domowego mierzona liczbą osób wynosiła w Polsce w 2002 roku 2,82 osoby.

Otrzymany wynik może budzić zastrzeżenia, gdyż faktyczna liczba mieszkańców gospodarstwa domowego wyrażona jest liczbami naturalnymi 1, 2, i.t.d. Wielkość średnia, którą otrzymaliśmy umożliwia nam analizę porównawczą wielkości gospodarstw domowych w krajach Unii Europejskiej, lub w jednym kraju w różnych latach np. 1970, 1980, 1990 i 2000. Wartość ta sama w sobie bez zamysłu porównawczego nie przedstawia wartości informacyjnej.

PRZYKŁAD 2:
Postanowiono określić przeciętny wiek kobiety rodzącej dziecko w 2002 roku w Polsce. W tym celu skorzystano z danych statystycznych zawartych w *Roczniku Demograficznym GUS*, które zamieszczono w tabeli 3. Informacje o liczbie urodzeń żywych według wieku matki są publikowane przez GUS w tabelach, gdzie wiek jest pogrupowany w przedziały klasowe (kolumna 1). W drugiej kolumnie tabeli 3 podano liczbę urodzeń w Polsce odpowiadającą danej grupie wiekowej.

Urodzenia żywe w Polsce w 2002 r. według wieku matki

Wiek matki (w latach)	Liczba urodzeń (w tys.)
19 lat i mniej	24,3
20-25	106,4
25-30	127,1
30-35	63,0
35-40	25,8
40-45	6,7
45 i więcej lat	0,3

Źródło: Rocznik demograficzny 2003, GUS Warszawa,s.250

Mając w ten sposób pogrupowane dane statystyczne średniej arytmetycznej nie możemy obliczyć. Możemy natomiast wyznaczyć jej przybliżoną wartość przyjmując środki przedziałów jako możliwe wartości wieku i następnie dla nich obliczyć średnią ważoną. Sposób ten prowadziłby do dokładnego wyznaczenia średniej gdyby istotnie wszystkie dane zaliczone do podanych grup znajdowały się w środku odpowiedniego przedziału.

Jeżeli środek i-tego przedziału oznaczymy jako $\dot{x}_i$, to teraz nasz wzór na średnią arytmetyczną ważoną przybierze postać:

$$\overline{X} = \frac{\sum_{i=1}^{k} \dot{x}_i \cdot n_i}{\sum_{i=1}^{k} n_i}$$

Obliczmy teraz przybliżoną wartość średniego wieku kobiet rodzących w Polsce w 2002 roku na podstawie pogrupowanych danych przedstawionych w kolumnie 1 i 2 tabeli 4 (do przykładu 4). W tabeli tej w pozostałych kolumnach ujęto obliczenia pomocnicze.

Tabela 4

Urodzenia żywe w Polsce w 2002 r. według wieku matki- obliczanie średniego wieku

Nr przedziału klasowego (i)	Wiek matki (w latach) $<x_i ; xj)$	Liczba urodzeń (w tys.) n_i	środki przedziałów $\dot{x}_i$	$\dot{x}_i \cdot n_i$
1	15-20	24,3	17,5	425,25
2	20-25	106,4	22,5	2394,00
3	25-30	127,1	27,5	3495,25
4	30-35	63,0	32,5	2047,50
5	35-40	25,8	37,5	967,50
6	40-45	6,7	42,5	284,75
7	45-50	0,3	47,5	14,25
x	suma	353,6	x	9628,5

Źródło: obliczenia własne na podstawie Rocznika demograficznego 2003

Przybliżoną wartość średniej arytmetycznej wieku rodzących kobiet wyznaczamy w następujący sposób:

$$\overline{X} = \frac{\sum_{i=1}^{7} \dot{x}_i \cdot n_i}{\sum_{i=1}^{7} n_i} = \frac{9628,5}{353,6} = 27,23$$

Przeciętny wiek kobiety rodzącej dziecko w 2002 roku w Polsce wynosił 27,23 roku.

Stosując metodę bazującą na średniej arytmetycznej ważonej, gdy wartości cechy zostały pogrupowane w przedziały klasowe, przybliżenie jest tym dokładniejsze, im mniejsza jest rozpiętość przedziałów klasowych. Interesującym może być to, że średni wiek kobiet rodzących wzrósł znacznie od roku 1950 co jest skutkiem daleko idących zmian społecznych i kulturowych.

Własność 1: Na wartość średnią mogą mieć duży wpływ wartości skrajne.

Własność 2: Jeżeli każdą wartość w zbiorze danych zwiększymy (lub zmniejszymy) o pewną stałą k , to wartość średniej arytmetycznej zwiększy się (lub zmniejszy) o tę stałą.

Własność 3: Jeżeli każdą wartość zbioru danych pomnożymy (podzielimy) przez stałą liczbę k, to wartość średniej arytmetycznej zwiększy się (zmniejszy) k razy.

Własność 4: Suma odchyleń poszczególnych wartości w zbiorze od ich średniej arytmetycznej jest równa zeru tzn. $\sum_{i=1}^{N}\left(X_i - \overline{X}\right) = 0$.

UWAGA: Im mniejsze występują różnice między wartościami w badanym zbiorze, tym średnia arytmetyczna lepiej charakteryzuje średni poziom interesującego nas zjawiska. Wartość średnia może ulec istotnej zmianie przy zmianach wartości ekstremalnych. Uważa się to za negatywną cechę średniej arytmetycznej. Miarami położenia, które nie posiadają tej wady są tzw. statystyki pozycyjne.

ŚREDNIA HARMONICZNA I GEOMETRYCZNA

Średnia harmoniczna jest miarą rzadko wykorzystywaną. Stosujemy ją wówczas, gdy wyniki obserwacji są podane w jednostkach względnych (w przeliczeniu na stałą jednostkę), np. w kg/osobę, km/godz., osoby/km2, lub gdy staramy się uchwycić przeciętną intensywność zmian. Jest ona używana do obliczania średniej szybkości pojazdów, średniego czasu potrzebnego do wykonania jednostki wyrobu, średniej gęstości zaludnienia, średniej szybkości obrotów pieniężnych. W treści poznawczej jest identyczna ze średnią arytmetyczną, różni się natomiast sposobem obliczenia z uwagi na odmienność danych dotyczących badanego zjawiska. Średnią harmoniczną obliczamy według wzoru:

$$H = \frac{n}{\frac{1}{x_1} + \frac{1}{x_2} + ... + \frac{1}{x_N}} = \frac{n}{\sum_{i=1}^{n}\frac{1}{x_i}},$$

gdzie: $x_{1,}\, x_{2,}\ldots\, x_{N,}$ - wyniki obserwacji w postaci liczb stosunkowych natężenia (w jednostkach względnych)

Średnia harmoniczna jest równa odwrotności średniej arytmetycznej obliczonej dla odwrotności poszczególnych wyników obserwacji badanego zjawiska.

Średnia geometryczna podobnie jak średnia harmoniczna jest znacznie rzadziej stosowana niż średnia arytmetyczna do obliczania przeciętnego poziomu wartości badanego zjawiska. Stosujemy ją szczególnie w zbiorach danych, w których wartości obserwacji są przedstawione w liczbach względnych oraz gdy występują znaczne różnice między wartościami (jest ona mniej wrażliwa na wartości ekstremalne niż średnia arytmetyczna). Znajduje zastosowanie przy obliczaniu przeciętnego tempa badanego zjawiska (przy badaniu kształtowania się zjawiska w czasie –t.j. przy szeregach czasowych).
Średnią geometryczną obliczmy według następującego wzoru:

$$G = \sqrt[n]{x_1 x_2 ... x_n} = \sqrt[n]{\prod_{i=1}^{n} x_i}$$

gdzie:
x_i – kolejne wartości badanego zbioru,
n – liczebność zbioru danych.

PRZYKŁAD 7A:
Stwierdzono, że aby podłączyć do sieci telefonicznej każdy nowo wybudowany dom jednorodzinny przy pewnej ulicy we Wrocławiu do sieci telefonicznej 3 pracowników telekomunikacji wykonuje tę samą czynność w różnym czasie: Kowalski – w ciągu 2 godzin, Walewski – w ciągu 4 godz., Kotyrba – w ciągu 8 godzin. Wszystkie domy są usytuowane w jednakowej odległości od jezdni. Zastanówmy się, jak obliczyć, ile czasu zużywają średnio pracownicy telekomunikacji na podłączenie domu do sieci telefonicznej.

Pracując równocześnie w jednym dniu przez 8 godzin podłączą: Kotyrba jeden dom, Wawelski dwa domy (4+4), a Kowalski podłączy najwięcej bo cztery domy (2+2+2=2). A więc razem potrzebowali 3*8 godz. = 24 godziny na podłączenie siedmiu domów. Czyli średnio pracownicy na podłączenie jednego domu do sieci zużywają: $\frac{24}{7} = 3\frac{3}{7}$ godziny.

Korzystając ze wzoru na średnią harmoniczną otrzymujemy analogiczny wynik :

$$H = \frac{3}{\frac{1}{2}+\frac{1}{4}+\frac{1}{8}} = \frac{3}{\frac{4}{8}+\frac{2}{8}+\frac{1}{8}} = \frac{24}{7} = 3\frac{3}{7} \text{ godziny.}$$

Natomiast stosując zwykłą średnią arytmetyczną otrzymalibyśmy : (2+4+8)/ 3 = $\frac{14}{3}$ =$4\frac{2}{3}$ godziny. Podając w ten sposób obliczony średni czas można niechcący (lub celowo) zniekształcić informację, która może stanowić podstawę w naliczaniu płac z wykonania zleceń.

PRZYKŁAD 7B:
Znając gęstość zaludnienia w trzech miastach 100 tysięcznych : 300 osób/km^2, 400 osób/km^2 i 600 osób/km^2 obliczmy jaka byłaby średnia gęstość zaludnienia dla trzech aglomeracji miejskich połączonych razem. Podstawiając do wzoru na średnią harmoniczną otrzymujemy:

$$H = \frac{3}{\frac{1}{300}+\frac{1}{400}+\frac{1}{600}} = 400 \text{ osób/km}^2.$$

Błędem byłoby dodanie tych trzech wartości i podzielenie przez 3: [(300+400+600)/3]= **433,3** osoby/km^2.
Znając liczbę mieszkańców tych trzech miast (100 tys. *3 = 300 tys.) i dzieląc ją przez powierzchnię jaką zajmują wszystkie miasta razem, możemy sprawdzić, która z otrzymanych średnich gęstości zaludnienia jest prawidłowa.
Wyliczmy więc powierzchnię dla każdego z tych miast:

I miasto: 100 000 osób : 300 osób/km^2 = 333,33 km^2
II miasto: 100 000 osób : 400 osób/km^2 = 250 km^2
II miasto: 100 000 osób : 600 osób/km^2 = 166,67 km^2

Razem miasta zajmują powierzchnię 750 km^2, a więc prawidłowa średnia gęstość zaludnienia w tych trzech miastach wynosi: 300 000 osób : 750 km^2 = **400** osób/km^2.

Jak wcześniej wspomniano średnia harmoniczna jest parametrem, którym dość rzadko się posługujemy. Wystarczy bowiem posiadać więcej informacji na temat badanego zjawiska, aby odtworzyć wielkości bezwzględne opisujące jego rozmiary, a zatem wykorzystać dla oceny sytuacji przeciętnej średnią arytmetyczną.

KWANTYLE

Kwantyle - definiuje się jako wartości cechy badanej zbiorowości, które dzielą zbiorowość na określone części pod względem liczby jednostek, części te pozostają do siebie w określonych proporcjach

Kwartyl pierwszy Q_1	dzieli zbiorowość na dwie części w ten sposób, że 25% jednostek zbiorowości ma wartości cechy niższe bądź równe kwartylowi pierwszemu Q_1, a 75% równe bądź wyższe od tego kwartyla
Kwartyl drugi (mediana Me)	dzieli zbiorowość na dwie równe części; połowa jednostek ma wartości cechy mniejsze lub równe medianie, a połowa wartości cechy równe lub większe od Me; stąd nazwa wartość środkowa

Kwartyl trzeci Q_3	dzieli zbiorowość na dwie części w ten sposób, że 75% jednostek zbiorowości ma wartości cechy niższe bądź równe kwartylowi pierwszemu Q_3, a 25% równe bądź wyższe od tego kwartyla

MEDIANA

Drugim, po średniej arytmetycznej, najczęściej używanym parametrem jest mediana (Me), w literaturze nazywana także wartością środkową.

Jeżeli $X_1, X_2, X_3, ..., X_n$ oznaczymy wyniki obserwacji, to mediana oznacza, mówiąc nieformalnie, wartość najbardziej centralną w uporządkowanym zbiorze tych samych obserwacji: $X_{(1)}, X_{(2)}, ..., X_{(n)}$. Symbol $X_{(1)}$ oznacza najmniejszą co do wielkości wartość w zbiorze, z kolei $X_{(2)}$ drugą co do wielkości itd. Sposób wyznaczania lub obliczania mediany zależy od tego, w jaki sposób ujęty został materiał statystyczny, którym dysponujemy.

Dla szeregu szczegółowego pozycją mediany to $N_{Me} = \frac{n+1}{2}$ i jej podstawie wyznacza się wartość środkową według wzoru:

$$Me = \begin{cases} x_{\frac{n+1}{2}} & \text{gdy } n \text{ jest nieparzyste} \\ \frac{1}{2}(x_{\frac{n}{2}} + x_{\frac{n}{2}+1}) & \text{gdy } n \text{ jest parzyste} \end{cases}$$

Szereg powinien być najpierw uporządkowany (w kolejności rosnącej) – następnie odczytujemy wartość wyrazu środkowego – dla parzystej liczby obserwacji, albo liczymy średnią arytmetyczną wyrazów sąsiadujących.

PRZYKŁAD 8A:
W piętnastoosobowym zespole pracowników firmy DINFO zajmującej się doradztwem finansowym dla sektora małych i średnich przedsiębiorstw wpłaty do urzędów skarbowych z tytułu podatku dochodowego od osób fizycznych za rok 2003 były następujące (w zł): 2802, 698, 1505, 2584, 938, 4580, 1030, 1278, 5322, 1350, 1860, 1044, 2056, 3668, 856.
Chcąc znaleźć medianę, należy w pierwszej kolejności uporządkować posiadany zbiór wyników, a następnie odszukać wartość obserwacji środkowej:
698, 856, 938, 1030, 1044, 1278, 1350, 1505, 1860, 2056, 2584, 2802, 3668, 4580, 5322

W naszym przykładzie $M_e = X_{\frac{15+1}{2}} = X_8 = 1505$ zł.

Oznacza to, że połowa pracowników tej firmy dokonała za 2003 rok wpłaty do urzędów skarbowych z tytułu podatku dochodowego od osób fizycznych w wysokości co najmniej 1505 zł lub inaczej, połowa spośród analizowanych osób zapłaciła podatek w wysokości nie większej niż 1505 zł.

PRZYKŁAD 8B:
Wróćmy do uporządkowanego zbioru wyników miesięcznych wynagrodzeń brutto ośmiu pracowników z przykładu 7:

750, 850, 880, 900, 930, 990, 1000, 2500.

Mediana miesięcznego wynagrodzenia w tej grupie osób zgodnie z wzorem wynosi:

$$M_e = \frac{1}{2}(X_{\frac{12}{2}} + X_{\frac{12}{2}+1}) = \frac{1}{2}(X_6 + X_7) = \frac{1}{2}(900+930) = 915 \text{ zł.}$$

Oznacza to, że połowa pracowników działu Inwentaryzacji zarabia miesięcznie nie więcej niż 915 zł., a druga połowa nie mniej niż 915 zł. Mediana okazała się tutaj lepszą miarą położenia charakteryzującą średnią płacę pracowników niż średnia arytmetyczna (1100 zł) jaką policzono w przykładzie 7.
Na podstawie przedstawionych tu przykładów wyraźnie widać, że w przeciwieństwie do średniej arytmetycznej, na wartość mediany nie mają wpływu wyniki obserwacji skrajnych (krańcowych).

Jeżeli wyniki obserwacji zostały pogrupowane w klasy bez przedziałów – to wyznaczenie mediany sprowadza się do wskazania jednostki środkowej i odczytania wartości jaka jej odpowiadaW tym celu wyznacza się pozycję mediany:

$$N_{Me} = \frac{n}{2}$$

. Przy danych pogrupowanych odnalezienie środkowej jednostki ułatwia skumulowanie liczebności.

PRZYKŁAD 9:
Pewien wykładowca matematyki skrzętnie notował nieobecności swoich studentów na jego wykładzie w I semestrze. Otrzymane wyniki przedstawił (ujął) w formie tabelarycznej.

Nieobecności studentów na wykładzie z matematyki

nr klasy	liczba nieobecności x_i	liczba studentów n_i	skumulowana liczba studentów (n_{sk})	
1	0	35	35	
2	1	39	74	przedział M_e
3	2	22	96	
4	3	12	107	
5	4	3	110	
6	5	1	111	
7	6	1	112	
x	x	113	x	

Aby znaleźć medianę liczby nieobecności studentów na wykładzie z matematyki należy wskazać pozycję mediany w uporządkowanym zbiorze, czyli jednostkę środkową. Spośród 113 studentów (n=113) uczęszczających na wykład, jednostkę środkową znajdujemy :

$$\frac{n+1}{2} = \frac{113+1}{2} = 57 \text{ (pozycja mediany).}$$

A więc 57 student w naszym uporządkowanym zbiorze znajduje się (patrząc na liczebności skumulowane – kolumna 3) w drugiej klasie i jego liczba opuszczonych wykładów wynosi 1. Czyli mediana = 1 nieobecność. Oznacza to, że połowa studentów w czasie I semestru co najwyżej opuściła wykład 1 raz, lub połowa studentów nie uczestniczyła w wykładzie co najmniej 1 raz.

Natomiast, gdy wyniki obserwacji są pogrupowane w klasy z przedziałami, wówczas medianę wyznaczamy metodą analityczną, opartą na interpolacji, wykorzystując wzór:

$$Me = x_{Me} + \frac{N_{Me} - S_{i-1}}{n_{Me}} \cdot h$$

gdzie: x_{Me} - dolna granica przedziału, w którym jest mediana,

h - rozpiętość przedziału, w którym jest mediana,

N_{Me} - pozycja mediany liczona jako połowa zbioru,

S_{i-1} - skumulowane (zsumowane) liczebności przedziałów poprzedzających przedział mediany, (suma liczebności przedziałów poprzedzających przedział mediany)

n_{Me} - liczebność rzeczywista przedziału, w którym jest mediana.

PRZYKŁAD 10:
W tabeli przedstawiono liczbę kobiet zawierających związek małżeński w Polsce w 2002 roku według wieku kobiety. Z tak przedstawionych danych statystycznych nie możemy wyznaczyć średniego wieku kobiety zawierającej związek małżeński używając średniej arytmetycznej. Możemy natomiast posłużyć się medianą.

Kobiety zawierające związek małżeński w Polsce w 2002 r.

wiek kobiety X_i	liczba kobiet n_i	skumulowana liczba kobiet (n_{sk})
19 i mniej	19499	19499
→ **20-29**	**144425**	163924
30-39	14628	178552
40-49	5754	184306
50-59	3803	188109
ponad 60	2393	190502
suma	**190502**	**x**

Obliczając (szacując) medianę w pierwszej kolejności musimy ustalić jej pozycję (numer) tzn. połowa liczebności wszystkich kobiet: $= 95251$. Podstawiając dane możemy ustalić medianę wieku kobiety zgodnie z podanym wcześniej wzorem:

$$Me = 20 + \frac{95251 - 19499}{144425} \cdot 10 = 20 + 5{,}24 = 25{,}24$$

Otrzymany wynik oznacza, że przeciętny wiek kobiety zawierającej związek małżeński w Polsce w 2002 roku to wiek 25,24 lat, a dokładniej, połowa kobiet zawierających związek małżeński w Polsce w 2002 roku nie przekroczyła wieku 25,24 lat.

Mediana jest mniej wrażliwa na zmiany wartości skrajnych obserwacji, niż średnia arytmetyczna. Jest obliczana wszędzie tam, gdzie nie ma możliwości obliczenia średniej arytmetycznej, np. gdy dane są przedstawione w postaci szeregu rozdzielczego z niedomkniętym pierwszym i ostatnim przedziałem klasowym. Lepiej charakteryzuje badaną populację, gdy średnia arytmetyczna nie plasuje się na pozycji centralnej w posiadanych wynikach (tzn. kiedy nasze wyniki wykazują wyraźną asymetrię). Trzeba pamiętać, że im bardziej średnia arytmetyczna różni się od mediany, tym lepiej mediana wyraża tendencję centralną.

KWARTYL PIERWSZY Q1 i KWARTYL TRZECI Q3

Oprócz mediany w statystyce opisowej stosowane są też inne miary pozycyjne o podobnym znaczeniu. Dzielą one uporządkowany zbiór wyników obserwacji na więcej niż dwie jednakowo liczne części. Kiedy dzielimy taki uporządkowany zbiór wyników na 10 części – mamy do czynienia z decylami, a kiedy dzielimy na 100 części – z percentylami. Ogólnie miary pozycyjne tego typu nazywamy kwantylami. Jednak najczęściej w statystyce opisowej stosowane są kwartyle (zwane wartościami ćwiartkowymi).

Dla szeregu szczegółowego kwartyl pierwszy i trzeci wyznacza się w ten sposób, że w dwóch częściach zbiorowości, które powstały po wyznaczeniu mediany, ponownie wyznacza się medianę; mediana w pierwszej części odpowiada kwartylowi pierwszemu, a w drugiej kwartylowi trzeciemu.
Pozycja kwartyla pierwszego i trzeciego

$$N_{Q1} = \frac{n+1}{4}, \quad N_{Q3} = \frac{3(n+1)}{4}$$

PRZYKŁAD 10:
W zbiorze uporządkowanych danych dotyczących wpłat do urzędów skarbowych z tytułu podatku dochodowego pracowników firmy DINFO już wcześniej znaleźliśmy 50-ty percentyl (czyli medianę). Znajdźmy teraz w naszym szeregu kwartyl pierwszy Q_1, zwany też kwartylem dolnym:

698, 856, 938, 1030, 1044, 1278, 1350, 1505, 1860, 2056, 2584, 2802, 3668, 4580, 5322

Me= Q_2

Kwartyl pierwszy Q_1 jest wartością tego elementu w zbiorze uporządkowanym, który jest medianą z połowy obserwacji położonych poniżej kwartyla drugiego Q_2. A więc szukamy mediany z wartości w zbiorze poniżej 1505 zł:

698, 856, 938, 1030, 1044, 1278, 1350

Q_1

Ponieważ Q_1 to też 25-ty percentyl, więc jego pozycję możemy też znaleźć wykorzystując wzór:

$$\frac{(n+1)\cdot 25}{100} = \frac{(15+1)\cdot 25}{100} = 4,$$

czyli na czwartej pozycji (X_4) w naszym zbiorze znajduje się wartość kwartyla pierwszego Q_1= 1030 zł.
Oznacza to, że 25% (1/4) osób w firmie DINFO zapłaciła za 2003 rok podatek dochodowy w wysokości nie większej niż 1030 złotych, a 75% (3/4) osób w wysokości nie mniejszej niż 1030 złotych.

Pozostał nam jeszcze do odszukania kwartyl trzeci Q_3, zwany kwartylem górnym. Jest on wartością tego elementu w uporządkowanym zbiorze, który jest medianą z połowy obserwacji położonych powyżej kwartyla drugiego Q_2.
W naszym przykładzie należy więc szukać kwartyla Q_3 z wartości elementów:

1860, 2056, 2584, 2802, 3668, 4580, 5322

Q_3

Ponieważ kwartyl trzeci jest równocześnie 75-tym percentylem więc jego pozycję znajdujemy zgodnie ze wzorem:

$$\frac{(n+1)\cdot 75}{100} = \frac{(15+1)\cdot 75}{100} = 12,$$

czyli na dwunastej pozycji (X_{12}) w naszym całym uporządkowanym zbiorze wpłat z tytułu podatku dochodowego znajduje się wartość Q_3 = 2802.
Oznacza to, że 75% (3/4) osób w firmie DINFO dokonało wpłat z tytułu podatku dochodowego w wysokości nie większej od 2802 zł., a 25% (1/4) osób w wysokości nie mniejszej niż 2802 zł.

Jeśli materiał statystyczny został pogrupowany w klasy bez przedziałów (przykład 3, 10) wówczas pozycję kwartyla pierwszego wyznaczamy :

$$N_{Q1} = \frac{n}{4},\ N_{Q3} = \frac{3n}{4}$$

Dla szeregu rozdzielczego wyznaczenie kwartyli poprzedza się ustaleniem ich pozycji:

Kwartyl pierwszy Q1

$$Q_1 = x_{0m} + \frac{N_{Q1} - \sum_{i=1}^{m-1} n_i}{n_m} \cdot h_m$$

$$Q_1 = x_{Q_1} + \frac{N_{Q_1} - S_{i-1}}{n_{Q_1}} \cdot h$$

Kwartyl trzeci Q3

$$Q_3 = x_{0m} + \frac{N_{Q_3} - \sum_{i=1}^{m-1} n_i}{n_m} \cdot h_m$$

$$Q_3 = x_{Q_3} + \frac{N_{Q_3} - S_{i-1}}{n_{Q_3}} \cdot h$$

gdzie: x_Q - dolna granica przedziału, w którym jest kwartyl,

h - rozpiętość przedziału, w którym jest mediana,

S_{i-1} - skumulowane (zsumowane) liczebności przedziałów poprzedzających przedział kwartyla,

n_Q - liczebność rzeczywista przedziału, w którym jest kwartyl.

PRZYKŁAD 11:
W tabeli 10 podano odsetek kobiet pełnozatrudnionych z wyższym wykształceniem według przedziałów wynagrodzeń w pewnym województwie w październiku 2002 roku, natomiast w tabeli 11 odsetek mężczyzn. Ponieważ przedstawione dane statystyczne są pogrupowane w przedziały, ponadto dwa z nich (pierwszy i ostatni) są otwarte, dlatego przy opisie tej zbiorowości należałoby wykorzystać parametry pozycyjne.

Kobiety z wyższym wykształceniem, pełnozatrudnione według wysokości wynagrodzenia w październiku 2002 roku w jednym z województw

wynagrodzenie (w zł) X_i	odsetek kobiet w_i'	skumulowany odsetek kobiet w_{sk}'
do 800 zł	2,1	2,1
800 – 1000	4,3	6,4
1000 – 1500	6,6	13,0
1500 – 2000 przedział Q_1	15,5	28,5
2000 – 2500 przedział Q_2	28,1	56,6
2500 – 3000	17,4	74,0
3000 – 3500 przedział Q_3	17,0	91,0
3500 – 4000	3,8	94,8
4000 – 4500	2,5	97,3
4500 – 5000	1,2	98,5
5000 – 5500	0,9	99,4
5500 – 6000	0,4	99,8
powyżej 6000	0,2	100

Źródło: szacunek własny na podstawie danych US we Wrocławiu

Mężczyźni z wyższym wykształceniem, pełnozatrudnieni według wysokości wynagrodzenia w październiku 2002 roku w jednym z województw

wynagrodzenie (w zł) X_i	odsetek mężczyzn w_i'	skumulowany odsetek mężczyzn w_{sk}'
do 800 zł	0,2	0,2
800 – 1000	1,3	1,5
1000 – 1500	4,3	5,8
1500 – 2000 przedział Q_1	20,1	25,9
2000 – 2500	15,7	41,6
2500 – 3000 przedział Q_2	11,0	52,6
3000 – 3500	8,7	61,3
3500 – 4000	8,3	69,6
4000 – 4500 przedział Q_3	8,5	78,1
4500 – 5000	7,7	85,8
5000 – 5500	5,5	91,3
5500 – 6000	5,1	96,4
powyżej 6000	3,6	100

PRZYKŁAD cd 11:
Przeciętne wynagrodzenie kobiet jak i mężczyzn z wyższym wykształceniem w rozpatrywanym województwie w październiku 2002 r. określi mediana, czyli kwartyl drugi:

- dla kobiet: $Me = 2000 + \frac{50-28,5}{28,1} \cdot 500 = 2000 + 382,6 = 2382,6$ zł.
- dla mężczyzn: $Me = 2500 + \frac{50-41,6}{11,0} \cdot 500 = 2500 + 381,8 = 2881,8$ zł.

Jak widzimy przeciętny poziom wynagrodzeń mężczyzn z wyższym wykształceniem w październiku 2002 roku w rozpatrywanym województwie był wyższy od przeciętnego poziomu dla kobiet o prawie 500 złotych. Połowa pełnozatrudnionych kobiet z wyższym wykształceniem nie przekroczyła wynagrodzenia 2382,6 zł, natomiast połowa mężczyzn nie przekroczyła wynagrodzenia 2881,8 zł., pozostała połowa kobiet i mężczyzn otrzymywała wynagrodzenie powyżej mediany.
Z kolei na podstawie obliczonego kwartyla pierwszego:

- dla kobiet: $Q_1 = 1500 + \frac{25-13}{15,5} \cdot 500 = 1500 + 387,1 = 1887,1$ złotych oraz
- dla mężczyzn: $Q_1 = 1500 + \frac{25-5,8}{20,1} \cdot 500 = 1500 + 477,6 = 1977,6$ złotych,

stwierdzamy, że 25% kobiet badanej populacji otrzymywała wynagrodzenie poniżej 1887 złotych a 75% tych kobiet przekroczyła ten poziom płac. Wyższy poziom kwartyla pierwszego wynagrodzeń dla mężczyzn świadczy o tym, że większość mężczyzn (75%) zarabia więcej niż kobiety t.j. powyżej 1977,6 zł.
Jeżeli policzymy jeszcze kwartyl trzeci:

- dla kobiet: $Q_3 = 3000 + \frac{75-74}{17} \cdot 500 = 3000 + 29,4 = 3029,4$ zł oraz
- dla mężczyzn: $Q_3 = 4000 + \frac{75-69,6}{8,5} \cdot 500 = 4000 + 317,6 = 4317,6$ zł

to widzimy, że wartość trzeciego kwartyla wynagrodzeń dla mężczyzn przewyższa o 1882 złotych poziom odpowiedniego kwartyla wynagrodzeń dla kobiet. 75% pełnozatrudnionych mężczyzn z wyższym wykształceniem w październiku 2002 roku otrzymywało wynagrodzenie poniżej 4317,6 zł., natomiast ten sam odsetek kobiet nie pobierał większego wynagrodzenia niż 3029,4 złotych.

DOMINANTA /MODALNA

Często wykorzystywanym przeciętnym parametrem pozycyjnym jest **dominanta**, czyli wartość najczęstsza (zwana też w literaturze **modalną**). Jest to ta wartość, która w zbiorze danych statystycznych ($X_1, X_2, X_3, ..., X_N$) występuje najczęściej (najliczniej) i możemy wówczas powiedzieć, że ona dominuje nad innymi wartościami.

PRZYKŁAD 12:
Akcje Spółki Akcyjnej „FEMAG" w czasie kolejnych notowań giełdowych w ostatnim kwartale 2003 roku osiągnęły następujące wartości (w zł): 36, 41, 29, 39, 40, 36, 36, 31, 30, 30, 32, 36, 33, 38, 31, 31, 33, 28, 35, 35, 38, 41, 36, 36, 31, 30, 30, 31, 32, 35, 38, 29, 36, 39, 41, 40, 29, 28, 29, 31, 31, 33, 36, 34, 38, 31, 37, 35, 30, 37, 28, 29, 34, 39, 36, 41, 38, 40, 31, 30, 30, 28, 32, 35, 36, 32,31, 33.

Jeżeli zbiór danych statystycznych jest liczny, wówczas aby znaleźć wartość dominującą, w naszym przykładzie cenę akcji o największej liczbie powtórzeń) można wartości zbioru uporządkować: 28,28,28,28, 29,29,29,29,29, 30,30,30,30,30,30,30, 31,31,31,31,31,31,31,31,31,31, 32,32,32,32, 33,33,33,33, 34,34, 35,35,35,35,35, 36,36,36,36,36,36,36,36,36,36, 37,37, 38,38,38,38,38, 39,39,39, 40,40,40, 41,41,41,41. Jak wynika z wyliczeń w ostatnim kwartale roku 2003 notowania Spółki „FEMAG" najczęściej osiągały (bo aż 10 razy) cenę 31 zł i cenę 36 zł.
Obie te wartości uznajemy jako dominanty: D_1=31 zł i D_2=36 zł.

W przypadku danych pogrupowanych w przedziały klasowe wyznaczenie dominanty sprowadza się do wskazania przedziału, w którym znajduje się dominanta, a następnie obliczenia jej wartości w oparciu o interpolację. Chcąc ją oszacować rozpiętości przedziałów klasowych muszą być równe. Przedziałem dominanty jest przedział o największej liczebności. Z kolei wartość dominanty wyznaczamy w sposób przybliżony, wewnątrz przedziału dominanty wykorzystując wzór:

$$D = x_D + \frac{n_D - n_{D-1}}{(n_D - n_{D-1}) + (n_D - n_{D+1})} \cdot h ,$$

gdzie: D - wartość dominanty,

x_D - dolna granica przedziału dominanty,

n_D - liczebność przedziału dominanty,

n_{D-1} - liczebność przedziału poprzedzającego przedział dominanty,

n_{D+1} - liczebność przedziału następującego po przedziale dominanty,

PRZYKŁAD 14:
Jeżeli przypatrzymy się zbiorowi danych (z przykładu 3), to stwierdzamy, że przedział 25-30 lat odznacza się największą liczebnością (w przykładzie liczbą urodzeń żywych), a wiec w tym przedziale mieści się wartość dominanty (co zaznaczono poniżej).

Wiek matki (w latach) $<X_i ; Xj)$	Liczba urodzeń (w tys.) n_i	odsetek urodzeń w_i
15-20	24,3	6,9
20-25	106,4	30,1
25-30 przedział dominanty	127,1 $n_{max} = n_D$	35,9
30-35	63,0	17,8
35-40	25,8	7,3
40-45	6,7	1,9
45-50	0,3	0,1

I właściwie możemy poprzestać na wskazaniu tego przedziału. Jednak, jeżeli chcemy dokładniej oszacować wartość dominanty, posługujemy się podanym wcześniej wzorem (…). Wartość dominanty będzie większa od granicy dolnej a mniejsza od granicy górnej przedziału, w którym się znajduje. Podstawiając do wzoru otrzymujemy:

$$D = 25 + \frac{127{,}1 - 106{,}4}{(127{,}1 - 106{,}4) + (127{,}1 - 63)} \cdot 5 = 25 + 1{,}2 = 26{,}2$$

Otrzymany wynik oznacza, że największa liczba kobiet rodzących w Polsce w 2002 roku charakteryzowała się wiekiem 26,2 roku. A zatem wyznaczona wartość dominanty mieści się w wyznaczonym przedziale.

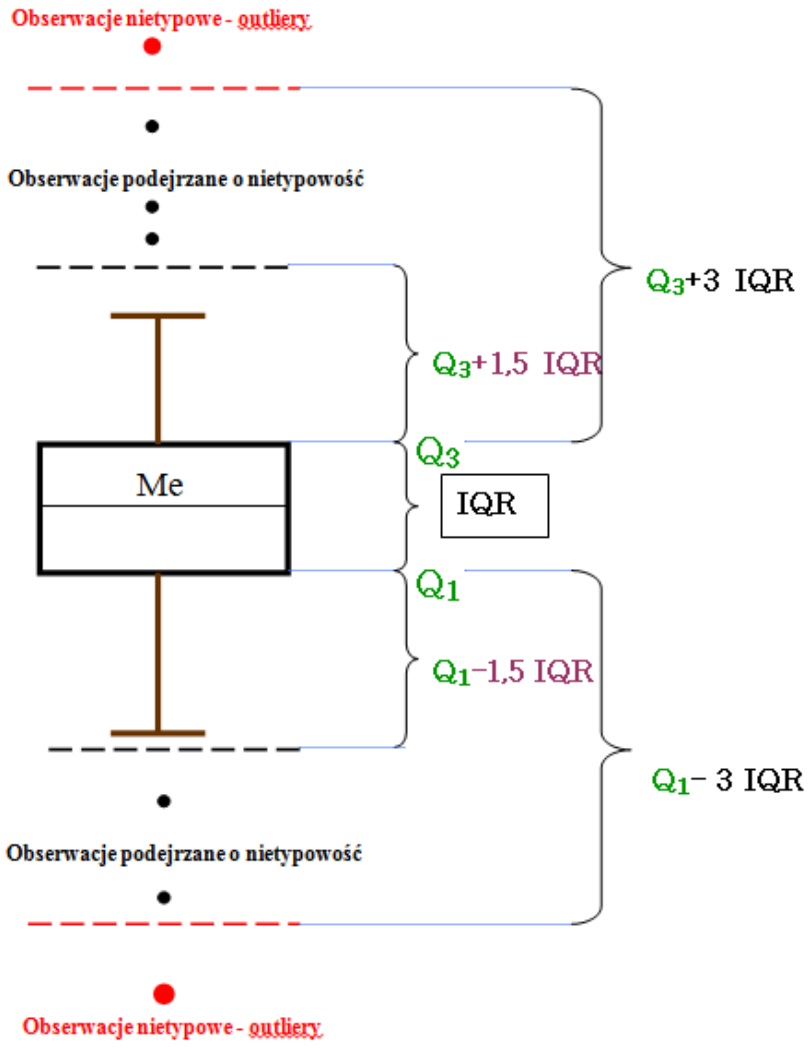
Na uwagę zasługuje też fakt, że wartość dominanty nie zmieni się, jeżeli w pogrupowanym zbiorze danych wystąpią odsetki zamiast liczebności absolutnych. Podstawiając w miejsce liczebności odsetki zamieszczone w 3 kolumnie tabeli …. otrzymujemy:

$$D = 25 + \frac{35{,}9 - 30{,}1}{(35{,}9 - 30{,}1) + (35{,}9 - 17{,}8)} \cdot 5 = 25 + 1{,}22 = 26{,}2$$

Ponieważ wzór, którym posłużyliśmy się do oszacowania wartości dominanty, zbudowany jest na podstawie założenia, iż w obrębie klasy najliczniejszej i dwóch sąsiednich rozkład cechy jest opisany funkcją kwadratową, więc liczebność przedziałów znajdujących się bezpośrednio przed przedziałem dominanty i po nim nie może wynosić zero. A więc nie jest wskazane obliczanie dominanty z danych pogrupowanych w przedziały klasowe, jeżeli największa liczebność znajduje się w przedziale skrajnym.

WYKRES PUDEŁKOWY

Najwięcej informacji przydatnych do analitycznego badania zbiorów danych dostarcza nam wykres pudełkowy (box plot), zwany też „pudełkiem z wąsami" (box-and-whisker plot). Wprowadził go stosunkowo niedawno (w 1977 r.) J.Tukey. Wykres pudełkowy, bardzo przydatny w opisie statystycznym, upowszechnił się wraz z pakietami statystycznymi. Są na nim przedstawione podstawowe statystyki opisowe, takie jak: mediana, kwartyl pierwszy i trzeci, wartości podejrzane o nietypowe oraz tzw. wartości odstające (ekstremalne) czyli takie, które wyraźnie odbiegają od pozostałych wartości zbioru danych. Ponadto na podstawie wykresu pudełkowego możemy określić rozproszenie wartości i asymetrię rozkładu.



MIARY ROZPROSZENIA

Przy opisie statystycznym różnych zbiorów danych (zbiorowości) nie wystarczy poprzestać na zastosowaniu miar położenia, lecz należy też określić stopień zróżnicowania tych danych (jednostek). Parametry opisowe, które tutaj wykorzystamy określamy mianem miar rozproszenia lub zmienności. Miary zmienności określają liczbowo stopień zróżnicowania obserwacji /elementów/ w badanym zbiorze danych (t.j. stopień w jakim poszczególne wartości zbioru odbiegają od wartości średniej).

Bywa tak, że średnie wyników obserwacji dwóch zbiorów są jednakowe, a mimo to zbiory te różnią się między sobą stopniem zmienności i skupieniem poszczególnych wartości wokół średniej. W takim przypadku wnioskowanie na podstawie tylko średnich arytmetycznych tych zbiorów jest niewystarczające.

PRZYKŁAD 15:
W związku z napływem do Wrocławia kapitału japońskiego, jedna z wrocławskich szkół języków obcych, uruchomiła kurs nauki języka japońskiego. Utworzono dwie grupy 7 osobowe. Pierwszą stanowiły osoby w wieku 20-35 lat, drugą osoby powyżej 35 roku życia. Na zakończenia I semestru nauki kursanci otrzymali następujące wyniki z testu :
I grupa: 100, 115, 130, 140, 150, 165, 180;
II grupa: 125, 130, 135, 140, 145, 150, 155.
Jak się okazało średnia oraz mediana liczby uzyskanych punktów z testu w obu grupach była jednakowa, równa 140 punktów, a jednak wyniki poszczególnych osób w tych grupach znacznie się różnią (rys.1).

Porównanie wartości testu w I i II grupie

100	105	110	**115**	120	125	**130**	135	**140**	145	**150**	155	160	**165**	170	175	**180**
					125	**130**	**135**	**140**	**145**	**150**	**155**					

Zauważmy, że wyniki testu osiągnięte przez osoby młodsze są bardziej rozproszone niż wyniki osób starszych. Wyniki tej grupy leżą dalej od średniej liczby punktów (równej 140) niż wyniki grupy drugiej. Wyniki testu grupy drugiej są skupione blisko siebie, a więc mało zróżnicowane.

Z przedstawionego przykładu wynika, że aby gruntowniej opisać zbiory analizowanych danych, należy oprócz obliczenia średniej arytmetycznej czy mediany, również ustalić w jakim stopniu wyniki poszczególnych obserwacji różnią się od siebie, a więc należy dokonać pomiaru ich rozproszenia. Przy opisie statystycznym średnia jest parametrem prawidłowym w odniesieniu do zbioru o niewielkim zróżnicowaniu wyników obserwacji. Gdy występuje wzrost zróżnicowania badanej cechy, to średnia traci swoją wartość poznawczą i wówczas opis powinien być uzupełniony o parametry rozproszenia.

Do pomiaru rozproszenia /zmienności/ wartości w badanym zbiorze danych służą parametry rozproszenia, zwane też parametrami zmienności. Zaliczamy do nich m.in.:

- rozstęp,
- odstęp międzykwartylowy,
- wariancję,
- odchylenie standardowe i ćwiartkowe,
- współczynnik zmienności.

ROZSTĘP I ODSTĘP MIĘDZYKWARTYLOWY

Rozstęp jest różnicą między wartością największą i najmniejszą w analizowanym zbiorze danych:

$$R = X_{max} - X_{min}$$

Podobnym parametrem rozproszenia, aczkolwiek częściej używanym jest odstęp międzykwartylowy – różnica między kwartylem trzecim Q_3 i pierwszym Q_1 :

$$IQR = Q_3 - Q_1.$$

PRZYKŁAD 15B:
W przykładzie w zbiorach wyników testu językowego rozstępy wynoszą odpowiednio:

100	105	110	**115**	120	125	**130**	135	**140**	145	**150**	155	160	**165**	170	175	**180**
					125	**130**	**135**	**140**	**145**	**150**	**155**					

dla grupy I: $R_I = 180-100=80$,
dla grupy II: $R_{II} = 155-125 = 30$. Widać więc, że rozstęp wartości punktów w grupie osób młodszych był większy niż wśród osób starszych wiekiem (grupa II). Grupa I wykazuje większą zmienność wartości wyników testu – większe rozproszenie.
Rozstęp jest parametrem rozproszenia zależnym tylko od dwóch skrajnych wartości zbioru danych, które często różnią się istotnie od pozostałych wartości. Dlatego jest to miara o małej wartości poznawczej.
Odstep międzykwartylowy w przytoczonym przykładzie 16 dla pierwszej grupy wynosi IQR=165-115 = 50, natomiast dla drugiej grupy: IQR=150-130 =20.

Jak widać, z podanych dwóch parametrów rozproszenia badanego zbioru danych odstęp międzykwartylowy jest mniej wrażliwy na wyniki skrajne.

WARIANCJA I ODCHYLENIE STANDARDOWE

Wariancja jest parametrem rozproszenia, który mierzy przeciętny kwadrat odchylenia poszczególnych obserwacji badanej cechy od średniej arytmetycznej w zbiorze danych statystycznych.
Jeżeli wyniki obserwacji z populacji skończonej oznaczymy jako $X_1, X_2, X_3,..., X_n$, to wariancję obliczymy ze wzoru:

$$S^2 = \frac{1}{n} \cdot \sum_{i=1}^{N} (X_i - \overline{X})^2,$$

Natomiast, gdy wyniki obserwacji dotyczące populacji zostały uporządkowane i pogrupowane w k- klas wówczas wariancję obliczamy według wzoru:

$$S^2 = \frac{1}{n} \cdot \sum_{i=1}^{k} (X_i - \overline{X})^2 \cdot n_i$$

Jeżeli wyniki obserwacji zostały pogrupowane w przedziały klasowe, wówczas analogicznie jak przy obliczaniu średniej arytmetycznej wykorzystujemy środki przedziałów klasowych ($\dot{x}_i$).

Kiedy nasze wyniki obserwacji pochodzą z małej próby, to obliczając wariancję w mianowniku we wzorze zamiast „n” umieszcza się „$n-1$”- jako liczebność tej próby. Jak wynika ze wzoru, wariancja jest zawsze wielkością nieujemną i mianowaną. Jej mianem jest kwadrat jednostki w jakiej dokonano pomiaru wyników obserwacji. Im większa jest jej wartość, tym silniejsze jest zróżnicowanie wartości badanej cechy.

Przy ocenie rozproszenia wykorzystujemy w istocie pierwiastek kwadratowy z wariancji – zwany odchyleniem standardowym. Jest to najczęściej używany i najbardziej precyzyjny parametr rozproszenia, oznaczany jako:

$$S = \sqrt{S^2}\ .$$

W typowym zbiorze danych 95% obserwacji leży w przedziale dwóch odchyleń standardowych wokół średniej:

$$(\overline{X} - 2S,\ \overline{X} + 2S).$$

PRZYKŁAD 15B:
Zwróćmy uwagę na wyniki testów średnia oraz mediana liczby uzyskanych punktów z testu w obu grupach była jednakowa, równa 140 punktów. Wyznaczmy rozproszenie otrzymywanych wyników.

100	105	110	**115**	120	125	**130**	135	**140**	145	**150**	155	160	**165**	170	175	**180**
					125	**130**	**135**	**140**	**145**	**150**	**155**					

Podstawiając do wzoru na wariancję otrzymujemy:

$S^2 = \frac{1}{17}[(-40)^2 + (-35)^2 + (-30)^2 + (-25)^2 + (-20)^2 + \ldots + 30^2 + 35^2 + 40^2] = 600\text{pkt}.$

Natomiast odchylenie standardowe płac pracowników tego działu wynosi:

$S = \sqrt{S^2} = \sqrt{600} = 24{,}5\text{pkt}.$

Oznacza ono, że wyniki z testu różnią średnio od przeciętnego wyniku o 24,5pkt.

W grupie drugiej:

$S^2 = \frac{1}{17}[(-15)^2 + (-10)^2 + (-5)^2 + (0)^2 + (5)^2 + 10^2 + 35^2 + 15^2] = 100\text{pkt}.$

Natomiast odchylenie standardowe płac pracowników tego działu wynosi:

$S = \sqrt{S^2} = \sqrt{100} = 10\text{pkt}.$

Oznacza ono, że wyniki z testu w II grupie różnią średnio od przeciętnego wyniku o 10pkt.

WSPÓŁCZYNNIK ZMIENNOŚCI

Ponieważ średnie płace w tych grupach nie różnią się więc porównując wielkość odchylenia standardowego można ocenić wielkość zróżnicowania. Jednak gdy średnie znacznie się różnią, więc dla porównania stopnia zróżnicowania należy skorzystać ze stosunkowego /względnego/ parametru rozproszenia jakim jest współczynnik zmienności, liczony według wzoru:

$$V = \frac{S}{\overline{X}} \cdot 100.$$

Współczynnik zmienności mierzy wielkość zróżnicowania obserwacji w zbiorze danych w postaci liczby niemianowanej, która jest odsetkiem wartości odchylenia standardowego w stosunku do średniej arytmetycznej.

Jego wartość bliska zeru świadczy, że obserwacje w badanym materiale statystycznym są jednorodne. Natomiast im bardziej wartości są zróżnicowane, tym większy jest współczynnik zmienności. Ponadto współczynnik zmienności może być przydatny przy porównywaniu zróżnicowania takich wielkości jak wydajność pracy, czas pracy, absencja w pracy, w różnych przedsiębiorstwach i instytucjach czy działach jednego przedsiębiorstwa. Wynika to z faktu, że odchylenie standardowe jest miarą bezwzględną, więc nie pozwala na porównanie zmienności cech o różnych mianach.

Podstawowe własności odchylenia standardowego:

Własność 1: Jego wartość nie ulega zmianie gdy:

a) liczebności w zbiorze zawierającym dane pogrupowane zostaną wyrażone w liczbach względnych (częstości, procenty),
b) dodamy lub odejmiemy od wszystkich wartości w zbiorze jakakolwiek (tę samą) liczbę.

Własność 2: Jeżeli wszystkie wartości w materiale statystycznym pomnożymy lub podzielimy przez jakąkolwiek (tę samą) liczbę różną od zera, to odchylenie standardowe będzie tylokrotnie mniejsze lub większe.

Własność 3: Jest parametrem najbardziej precyzyjnym spośród parametrów rozproszenia.

ODCHYLENIE ĆWIARTKOWE I POZYCYJNY WSPÓŁCZYNNIK ZMIENNOŚCI

Jeżeli nie możemy lub nie chcemy posłużyć się średnią arytmetyczną, bo n.p. w badanym zbiorze danych występują duże odchylenia wartości ekstremalnych czy też przedziały w danych pogrupowanych są otwarte, wówczas możemy posłużyć się odpowiednio pozycyjnymi miarami zróżnicowania: odchyleniem ćwiartkowym i kwartylowym współczynnikiem zmienności opartym na parametrach pozycyjnych:

$$Q = \frac{\left[\left(Q_3 - M_e\right) + \left(M_e - Q_1\right)\right]}{2} = \frac{IQR}{2}$$

$$V_Q = \frac{Q}{M_e} \cdot 100$$

PRZYKŁAD 16:
Wróćmy do przykładu o miesięcznych wynagrodzeniach pracowników działu Inwentaryzacji. Przeciętna płaca wynosiła 1100 zł. Wyznaczmy rozproszenie otrzymywanych wynagrodzeń. Podstawiając do wzoru na wariancję otrzymujemy:

$$S^2 = \frac{1}{8}[(-350)^2 + (-220)^2 + (-250)^2 + (-200)^2 + (-170)^2 + (-110)^2 + (-100)^2 + 1400^2] = 285550 zł^2.$$

Natomiast odchylenie standardowe płac pracowników tego działu wynosi:

$$S = \sqrt{S^2} = \sqrt{285550} = 534{,}37 \text{ zł.}$$

Oznacza ono, że płace tychże pracowników różnią średnio od przeciętnej płacy o 534,37 zł.

Nasuwa się pytanie czy płace pracowników innych działów w tym urzędzie są tak samo zróżnicowane? Na podstawie listy płac obliczono średnią płacę i odchylenie standardowe pracowników w dwóch kolejnych działach, tj. w dziale Księgowości i dziale Promocji Miasta. Otrzymano następujące wyniki:

dla działu Księgowości - $\overline{X}_K = 1800$ zł , $S_K = 535$ zł ,

dla działu Promocji Miasta - $\overline{X}_P = 1650$ zł , $S_P = 125$ zł .

PRZYKŁAD c.d 16:
Więc w celu porównania rozproszenia płac w trzech działach obliczono współczynniki zmienności:

Dla działu Inwentaryzacji: : $V_I = \frac{534{,}37}{1100} \cdot 100 = 48{,}6\%,$

Dla działu Księgowości: $V_K = \frac{535}{1800} \cdot 100 = 29{,}7\%,$

Dla działu Promocji Miasta: $V_P = \frac{125}{1650} \cdot 100 = 7{,}6\%,$

Najmniej zróżnicowana wysokość wynagrodzeń wystąpiła wśród pracowników działu Promocji Miasta. Zatem zróżnicowanie płac pracowników działu Inwentaryzacji jest ponad sześć razy większe od zmienności płac pracowników działu Promocji Miasta.

MIARY ASYMETRII

Miary rozproszenia poszerzają naszą wiedzę o strukturze zbiorowości poprzez wskazanie, w jakim stopniu wartości poszczególnych elementów w zbiorze danych koncentrują się wokół wielkości centralnej tego zbioru. Natomiast nie opisują nierównomierności rozłożenia obserwacji badanego zbioru wokół wartości średniej arytmetycznej. Problem ten rozwiązują miary asymetrii.
Miara asymetrii obliczana z wartości wszystkich obserwacji zbioru, to tzw. klasyczny współczynnik asymetrii obliczany według wzoru:

$$K_s = \frac{\frac{1}{n}\sum_{i=1}^{n}(x_i - \overline{X})^3}{S^3},$$

Jeżeli natomiast wartości obserwacji zostały pogrupowane w szereg rozdzielczy, wówczas wzór na współczynnik asymetrii przedstawia się następująco:

$$K_s = \frac{\frac{1}{n}\sum_{i=1}^{n}(x_i - \overline{X})^3 \cdot n_i}{S^3},$$

W przypadku grupowania wartości zbioru danych w przedziały klasowe, wówczas analogicznie jak przy obliczaniu średniej arytmetycznej wykorzystujemy środki przedziałów klasowych ($\dot{x}_i$).

Współczynnik ten określa kierunek i natężenie (siłę) asymetrii. Może być wykorzystany do porównywania asymetrii zbiorów danych wyrażonych w różnych jednostkach miary. Jednak wadą tej miary jest brak określonych granic jej zmienności.

Jeżeli $K_s>0$ wówczas występuje asymetria prawostronna, w badanym zbiorze przeważają obserwacje, których wartości są mniejsze od średniej. Z kolei, gdy $K_s<0$ to mamy asymetrię lewostronną, w badanym zbiorze przeważają obserwacje, których wartości są większe od średniej arytmetycznej.

PRZYKŁAD 17:
Wykorzystując wcześniejsze obliczenia dotyczące przeciętnej liczby osób w gospodarstwach domowych w Polsce w 2002 roku oraz odchylenia standardowego sprawdźmy asymetrię rozkładu tych gospodarstw.

Obliczenia pomocnicze przy wyznaczaniu asymetrii

Liczba osób w gospodarstwie domowym (X_i)	Liczba gospodarstw domowych (n_i)	$(X_i - \overline{X})^3$	$(X_i - \overline{X})^3 \cdot n_i$	$(X_i - \overline{X})^4$	$(X_i - \overline{X})^4 \cdot n_i$
1	3307	-6,03	-19941,21	10,97	36277,79
2	3097	-0,55	-1703,35	0,45	1393,65
3	2654	0,01	26,54	0,001	2,65
4	2405	1,64	3944,20	1,94	4665,70
5	1086	10,36	11250,96	22,59	24532,74
6	462	32,16	14857,92	102,26	47244,12
7	326	73,03	23807,78	305,28	99521,28
sumy	**13337**	**x**	**32242,84**	**X**	**213637,93**

Podstawiając do wzoru otrzymujemy:

$$K_s = \frac{\frac{1}{n}\sum_{i=1}^{n}(X_i - \overline{X})^3 \cdot n_i}{S^3} = \frac{\frac{1}{13337} \cdot 3224{,}284}{1{,}55^3} = \frac{2{,}418}{3{,}724} = 0{,}649$$

Wynik ten świadczy o tym, iż rozkład gospodarstw domowych według liczby osób charakteryzuje się słabą asymetrią prawostronną, tzn. przeważają gospodarstwa o liczbie osób większej niż 2,88.

W sytuacji, gdy zbiór wyników jest pogrupowany i występują w nim przedziały otwarte, utrudniające obliczenie średniej arytmetycznej i odchylenia standardowego, wówczas dla określenia asymetrii możemy posłużyć się tzw. pozycyjnym współczynnikiem skośności, opartym na kwartylach:

$$K_Q = \frac{Q_3 + Q_1 - 2M_e}{IQR}.$$

Jest to miara unormowana, która przyjmuje wartości z przedziału [-1; 1]. A więc może być też wykorzystana do mierzenia natężenia asymetrii. W zbiorze danych o asymetrii prawostronnej (dodatniej - $K_Q>0$) kwartyl pierwszy znajduje się bliżej mediany niż kwartyl trzeci, a przy asymetrii lewostronnej (ujemnej - $K_Q<0$) zachodzi sytuacja odwrotna , czyli kwartyl pierwszy jest bardziej oddalony od mediany niż kwartyl trzeci.
Należy jednak pamiętać, że współczynnik skośności oparty na kwartylach bada asymetrię tylko na połowie wyników obserwacji, tylko tych które mają wartości między Q_3 i Q_1.

PRZYKŁAD 18:
Wykorzystując obliczone parametry pozycyjne położenia w przykładzie 15 dotyczącym wynagrodzenia kobiet i mężczyzn z wykształceniem wyższym w pewnym województwie, w październiku 2002 r., zbadano asymetrię wynagrodzeń dla obu rozpatrywanych grup:
- dla kobiet

$$K_Q = \frac{Q_3 + Q_1 - 2M_e}{IQR} = \frac{3029{,}4 + 1887{,}1 - 2 \cdot 2382{,}6}{1142{,}3} = 0{,}13,$$

- dla mężczyzn

$$K_Q = = \frac{4317{,}6 + 1977{,}6 - 2 \cdot 2881{,}8}{2340} = 0{,}23.$$

Obliczone współczynniki skośności są dodatnie, zatem obie badane zbiorowości pod względem wynagrodzeń charakteryzuje asymetria prawostronna. Aczkolwiek jej natężenie dla zbiorowości kobiet jest słabsze.