Danuta Tarka, Anna Małgorzata Olszewska

Elementy statystyki
Opis statystyczny

Oficyna Wydawnicza Politechniki Białostockiej
Białystok 2018

Recenzenci:
dr hab. Józef Rogowski, prof. UwB
prof. dr hab. Wojciech Maciejewski

Redaktor wydawnictwa:
Elżbieta Dorota Alicka

Projekt okładki:
Agencja Wydawnicza EkoPress, wykorzystano grafikę SergeyNivens

Autor ilustracji:
Rafał Olszewski

ISBN 978-83-65596-66-6 ISBN 978-83-65596-67-3 (eBook)
DOI: 10.24427/978-83-65596-67-3

Redakcja techniczna, skład:
Oficyna Wydawnicza Politechniki Białostockiej

Druk:
EXDRUK s.c.

Oficyna Wydawnicza Politechniki Białostockiej
ul. Wiejska 45C, 15-351 Białystok
tel.: 85 746 91 37
e-mail: oficyna.wydawnicza@pb.edu.pl
www.pb.edu.pl

Spis treści

Wstęp

Niniejsza książka jest przede wszystkim podręcznikiem akademickim dla studentów tych kierunków studiów dziennych i zaocznych, na których obowiązuje minimalna liczba godzin obowiązkowych ze statystyki. Książka obejmuje materiał podstawowy ze statystyki opisowej, od którego zaczyna się wykładanie statystyki. Wyłożony tu materiał jest pomyślany jako wykład z podstaw statystyki realizowany przy minimum programowym 15 godzin samego wykładu lub 15 godzin wykładu i 15 godzin ćwiczeń.

Większość współczesnych podręczników statystyki pobieżnie traktuje tę część materiału, podając przede wszystkim wzory, bardzo skrótowo tłumacząc, co one opisują i jak interpretować wyniki uzyskane przy ich zastosowaniu. Być może autorzy tych podręczników uważają, że ta część wiedzy statystycznej jest bardzo prosta (to prawda) i student samodzielnie dojdzie do wniosków interpretacyjnych. Niestety, jak pokazuje wieloletnie doświadczenie autorek, kolejne „reformy" w systemie edukacyjnym (zwłaszcza coraz większe cięcia w podstawach programowych oraz liczbie godzin) skutkują coraz większą nieporadnością coraz większej liczby maturzystów w samodzielnym uczeniu się, czy bardziej ogólnie, w zdobywaniu wiedzy, oraz coraz mniejszą umiejętnością koncentracji. W efekcie na studia masowo przychodzą młodzi ludzie coraz słabiej przygotowani do studiowania. Stąd idea tej książki. Ma ona na celu wyłożenie jak najprostszym językiem podstaw statystyki przede wszystkim osobom, dla których będzie ona jednym z narzędzi do krytycznej analizy informacji, z jakimi będą się stykać w życiu zawodowym, czy też prywatnym.

W rozdziale pierwszym podajemy krótką historię statystyki, czyli jak wyewoluowała ona do rangi dyscypliny naukowej. W rozdziale drugim wprowadzamy podstawowe pojęcia i definicje używane w statystyce. Rozdział trzeci zapoznaje Czytelnika z problematyką zbierania materiału statystycznego oraz ze sposobami jego prezentacji. Jest z konieczności dosyć zwięzły i ma dać Czytelnikowi tylko pewne pojęcie o tej problematyce. Zbieranie i wstępne przetwarzanie danych statystycznych jest bardzo skomplikowanym zagadnieniem, będącym już *de facto* odrębnym i niełatwym działem statystyki. Waga tego problemu wynika z faktu, iż niepoprawnie zebrane i przetworzone dane statystyczne mogą prowadzić do fałszywych wniosków. Pozostałe rozdziały, od czwartego do szóstego, przedstawiają podstawowe miary używane w analizie materiału statystycznego. Są to kolejno: miary przeciętne, miary dyspersji oraz miary asymetrii i koncentracji.

Przedstawione tu podstawowe miary analizy statystycznej dotyczą, co podkreślamy, „dobrze zachowujących się" rozkładów jednej cechy. Przy pomocy zaprezentowanych w podręczniku miar Czytelnik będzie umiał scharakteryzować dany rozkład oraz określić, na ile jest on standardowy (normalny, czyli podobny do innych – co to dokładniej oznacza, będzie wyjaśniane w trakcie wykładu), jak też wyciągnąć wnioski w oparciu o analizowane dane dotyczące interesującego go zjawiska. By móc analizować mniej standardowe rozkłady (zbiory) danych, Czytelnik będzie musiał, jeśli zajdzie taka potrzeba, sięgnąć do bardziej zaawansowanego wykładu ze statystyki, przedstawionego w którymś z dostępnych na rynku podręczników.

W osobistej opinii autorek najlepszym polskim podręcznikiem dotyczącym statystyki opisowej jest podany w spisie literatury podręcznik B. Szulca. Ma on niestety podstawową wadę – jest trudno dostępny. Ostatnie jego wydanie pochodzi z lat siedemdziesiątych.

Autorki mają nadzieję, że prezentowany podręcznik okaże się przydatny studentom w opanowaniu podstaw statystyki. Na koniec pragną podziękować recenzentom za szczegółowe uwagi, które pomogły im poprawić klarowność prezentowanego tu wykładu.

1. Przedmiot i historia statystyki

Jak stwierdził jeden z najwybitniejszych statystyków XX wieku, C.R. Rao [1994, s. 51]: „Statystyka ma długą prehistorię, ale krótką historię. Jej pochodzenie można wywodzić od początków ludzkości, ale dopiero w ostatnich czasach okazała się ona przedmiotem o wielkim znaczeniu praktycznym".

Termin „statystyka" jest używany już od XVI wieku, jednak jego znaczenie ewoluowało, osiągając dzisiejsze rozumienie pojęcia dopiero w końcu XIX i w początkach XX wieku. Aby dojść do współczesnej definicji i przedmiotu statystyki, prześledźmy najpierw „prehistorię" i ewolucję tego pojęcia[1].

Starożytność:

2000 r. p.n.e., Chiny (Czasy dynastii Sia) – istnieją świadectwa o przeprowadzaniu spisów ludności.

1500 p.n.e. Stary Testament – Księga Liczb informuje o przeprowadzonych w tym okresie spisach ludności.

1122-256 r. p.n.e., Chiny (Dynastia Czou) – ustanowiono oficjalne stanowisko dla osoby odpowiedzialnej za prace statystyczne i nazwanej „Szih-su", czyli księgowy[2].

578-543 p.n.e., Rzym Serwiusz Tuliusz, król Rzymu, ustanawia obowiązek przeprowadzania spisu ludności. Urzędnicy rzymscy, zwani cenzorami[3], mieli sporządzać raz na pięć lat rejestr

[1] W większości polskich podręczników nie ma zbyt dużo, o ile w ogóle, informacji na ten temat. Polecamy tu prace Langego [1975, s. 58-69], Rao [1994, s. 51-56], Ostasiewicza [2011, s. 13-15], Puchalskiego [1978, s.15-17], Yule'a i Kendalla [1966], w oparciu o które prezentujemy tu skrócony rodowód statystyki. Podajemy tu Państwu tylko wybrane fakty i osoby z bogatej historii statystyki. Wkład w rozwój tej dyscypliny wniosło bardzo wielu uczonych i nie sposób podać ich wkładu do nauki w tak krótkim przeglądzie.

[2] Rao [1994, s. 52].

[3] Słowo to pochodzi od łacińskiego *censere* oznaczającego „szacowanie". Z czasem przekształciło się w słowo *census* i oznaczało „spis ludności", a współcześnie określa się nim np. każdy spis robiony w skali całego państwa. W Polsce np. dokonuje się tzw. spisów powszechnych, spisów rolnych, spisów mieszkań i innych. Samym słowem *censor* określano w starożytnym Rzymie jednego z dwóch urzędników wybieranych co pięć lat i nadzorujących przeprowadzanie spisów ludności i oszacowanie majątku.

	obywateli i stanu ich posiadania dla celów podatkowych i wojskowych[4].
300 p.n.e. Indie	Zarejestrowane na piśmie dowody na istnienie już wcześniej systemu rejestrów administracyjnych i statystyk urzędowych.
247-195 p.n.e. Chiny	Cesarz Liu Pang uznaje statystykę (spisy) za tak ważną, iż „(…) powierzył ją pieczy swojego pierwszego ministra – była to tradycja, która trwała w Chinach długi czas”[5].
74 p.n.e. Rzym	Ostatni regularny spis.
5 p.n.e. Rzym	Cezar August rozszerza spis ludności na całe Imperium Rzymskie. Od tego czasu: „Nie ma wzmianek o spisach ludności dokonywanych gdziekolwiek w świecie zachodnim przez wiele wieków po upadku Imperium Rzymskiego. Znane dzisiaj regularne okresowe spisy ludności zaczęły się dopiero w XVII w”[6].
Czasy nowożytne:	
1086 n.e. Anglia	Na rozkaz Wilhelma sporządzono księgę tzw. Doomsday Book, będącą spisem posiadłości wraz z informacją o ich obszarze, wartości itp.
IX-XII w. Rosja	Tworzone są spisy tzw. dymów, pługów itp. dla celów podatkowych, jako forma ewidencji posiadłości.

[4] Chodziło, jak we wszystkich tego typu spisach w starożytności i czasach późniejszych także, o określenie liczby mężczyzn zdolnych do służby wojskowej.
[5] Rao [1994, s. 52].
[6] Rao [1994, s. 53].

XIII Italia (Wenecja)	„(…) zaczęto prowadzić poprawną ewidencję, rachunkowość, które potrzebne były do pewnego przewidywania, oceny sytuacji i wyciągnięcia z tego wniosków do działania”[7].
1581 lub 1589[8] Włochy	G. Ghislini[9], włoski historyk i mąż stanu, używa po raz pierwszy słowa *statystyka*. Odniósł je do „(…) rachunku „civile, politics, statistica e militare scienza”[10]. Wiedzę o stanie państwa nazwał *statystyką*, tworząc nazwę od włoskiego słowa *stato* oznaczającego państwo. Generalnie uważa się, że właśnie stąd wywodzi się etymologia słowa statystyka[11], aczkolwiek jego znaczenie było inne niż współczesne.
1596-1597 Indie	Najlepiej znane zestawienie statystyki urzędowej o Indiach za rządów cesarza Akbara dokonane zostało przez jego ministra Abula Fazla. Zawiera szczegółowy opis tego, co posiadało Imperium oraz co się w nim działo od strony gospodarczej, jak np. ceny, wielkości plonów, zarobki w różnych zawodach, rodzaje zasiewów i hodowli.

Zarówno dotychczas sporządzane „spisy statystyczne”, jak i opisy stanu państwa na nich bazujące były, co najwyżej, „fotografiami” pewnego stanu państwa, nie wykorzystywano tych informacji do bardziej skomplikowanych analiz, na podstawie których można by było wyciągnąć ogólniejsze wnioski o społeczeństwie czy gospodarce. Były głównie informacjami dla rządzących wykorzystywanymi do celów podatkowych i militarnych. Rao określił funkcję statystyki z tamtych czasów jako „(…) oczy i uszy rządu”[12].

Jak widać z powyższego krótkiego przeglądu, statystyka wykształcała się jako nauka społeczna powiązana ze sprawami państwa, a zwłaszcza z analizami gospodarczymi.

---

[7] Puchalski [1978, s. 15].

[8] Pierwsza data wg Ostasiewicza [2011, s.14], wg Puchalskiego [1978, s. 15] druga.

[9] Pisownia nazwiska nie jest jasna, tu podajemy nazwisko za Ostasiewiczem [2011, s. 14], ale np. u Puchalskiego [1978, s. 15] podane jest nazwisko Gihlini G. Na włoskiej stronie Wikipedii nazwisko tego męża stanu jest podane w obu formach: http://it.wikipedia.org/wiki/Storia_della_statistica.

[10] Za Puchalskim [1978, s. 15].

[11] Słowo *stato* jest związane ze średniowiecznym łacińskim słowem *status* oznaczającym *państwo, stan*. Uważa się, że to łacińskie *status* zostało wykorzystane do ukucia nazwy *statystyka* jako dziedziny zajmującej się opisem stanu gospodarczego, społecznego i politycznego państwa. Świadczy o tym podobieństwo tej nazwy w większości języków europejskich: statistica (włoski), statistik (niemiecki), statystyka (polski) statistika (rosyjski), statistique (francuski), statistics (angielski).

[12] Rao [1994, s. 54].

XVII wiek, Europa – arytmetycy polityczni

1620-1674 Anglia **J. Graunt** jako pierwszy podjął próbę wykorzystania zbioru danych liczbowych do wyciągnięcia bardziej ogólnych wniosków niż tylko dotyczących obiektów opisywanych przez te liczby. Skonstruował on, w oparciu o rejestry kościelne dotyczące zgonów, pierwsze tablice umieralności, wykazując na ich podstawie pewne prawidłowości w zgonach ludności, czyli badanych obiektów.

1623-1687 Anglia **W. Petty** uważany jest za twórcę *arytmetyki politycznej*[13], czyli nauki, która „(…) zajmować się będzie liczeniem zjawisk społecznych oraz ustalaniem w tych zjawiskach liczbowo ujętych prawidłowości”[14].

Prace Graunta i Petty'ego[15] były podstawą powstania szkoły nazywanej szkołą arytmetyków politycznych. Jak stwierdza Lange[16]: „Arytmetycy polityczni byli właściwie pierwszymi statystykami”, dodajmy – w dzisiejszym tego słowa znaczeniu. Arytmetycy polityczni zastąpili bowiem opis werbalny opisem liczbowym oraz „zaproponowali”, by traktować analizowane zjawiska jako procesy masowe, a reprezentujące je liczby jako przejawy tych procesów i na tej podstawie szukać prawidłowości w procesach społecznych (w owym czasie głównie demograficznych). Wnioski przez nich wyciągane nie zawsze były poprawne, ale wynikało to najczęściej z braku wiarygodnych i kompletnych danych. Można powiedzieć, że wprowadzili oni wnioskowanie do analizy procesów społeczno-gospodarczych oparte na danych statystycznych. Wnioskowanie zaś jest uważane za jedną z podstawowych cech badania naukowego. Z czasem termin *arytmetyka polityczna* został zastąpiony pojęciem *statystyka*, używanym do czasów współczesnych.

1660 Niemcy **H. Conring** wygłasza po łacinie pierwszy udokumentowany wykład o nazwie *Statystyka*, jest to jednak wykład opisowy. Jak stwierdził K. Pearson[17], liczby w tych wykładach[18] pojawiły się tylko jako numery stron.

---

[13] „Polityczna” oznaczało wówczas ,,państwowa”, od greckiego słowa *polis* – państwo.

[14] Lange [1975, s. 60].

[15] W. Petty zaliczany jest do prekursorów klasycznej ekonomii politycznej, jako jeden z pierwszych zajął się bowiem badaniem (szukaniem) praw ekonomicznych rządzących gospodarką. Używał w tym celu prostej analizy danych statystycznych, stąd też uznaje się go za jednego z pierwszych statystyków w dzisiejszym tego słowa znaczeniu.

[16] Lange [1975, s. 60].

[17] Jeden z najznakomitszych statystyków przełomu XIX i XX wieku żył w latach 1857-1936, angielski uczony uważany za jednego z twórców statystyki matematycznej. Założył pierwszy w świecie wydział statystyki (*Department at University College London*).

[18] Opublikowano je kilkadziesiąt lat później.

1695 Anglia — **Ch. Davenant** (brytyjski polityk i urzędnik w randze generalnego inspektora handlu zagranicznego, autor dzieł z dziedziny ekonomiczno-militarnej) podaje definicję naukową arytmetyki politycznej. Jest „(…) to teoria lub zasady rozumowania o sprawach dotyczących rządzenia – nie jak u państwoznawców w oparciu o syntetyczne oceny o charakterze metafizyczno-moralnym i politycznym, lecz na podstawie danych liczbowych i w oparciu o specjalną metodę rozumowania”[19]. Statystyka jest tu więc nadal nauką o państwie, czy bardziej o społeczeństwie, ale wnioski są wyciągane w oparciu o dane liczbowe przy zastosowaniu specyficznej metody badawczej.

XVIII wiek

1707-1767 Niemcy — **J. P. Süssmilch**[20] (pastor niemiecki, uznawany za prekursora statystyki demograficznej w Niemczech) ogłasza pracę, w której porządkuje materiały liczbowe z rejestrów parafialnych, wykazując znane współcześnie prawidłowości demograficzne dotyczące umieralności i zawierania małżeństw. Przede wszystkim zaś wskazuje na „istnienie” prawa wielkich liczb, czyli na to, że są prawidłowości, które można wychwycić tylko poprzez analizę dużej ilości obserwacji tych samych zjawisk społecznych czy ekonomicznych (np. zgony, narodziny, konsumpcja itp.).

1719-1772 Niemcy — **G. Achenwall,** historyk i prawnik – jemu jako pierwszemu przypisuje się użycie (1749) słowa *statystyka* na określenie nauki o „osobliwościach państwowych”, do których zaliczał między innymi warunki naturalne, ustrój państwa, ludność, gospodarkę. Statystyka jest więc dla niego i jemu podobnych badaczy procesem gromadzenia danych liczbowych dotyczących państwa lub jego części i opisu werbalnego tych danych.

Tabelaryzm

Z czasem zauważono, że podawanie danych „surowych”, nieprzetworzonych nie sprzyja uzyskaniu ogólnego obrazu zjawiska czy wyłapaniu prawidłowości. Niektórzy badacze zaproponowali porządkowanie danych statystycznych dotyczących

[19] Za Luszniewicz [1973, s. 10].

[20] Można też spotkać zapis jego nazwiska jako Suessmilch w transkrypcji na niegermańskie języki.

państwa w postaci tablic podających najważniejsze informacje dotyczące ludności, obszaru, majątku, armii. Badaczy (statystyków) układających takie tabele nazwano tabelarystami.

1726-1727 Rosja **Kirgiłow** zapoczątkował *tabelaryzm*, opracowując jako pierwszy opis Rosji w postaci tabel.

1741 Dania **J. P. Anchersen** opracował tabele dla Danii.

Tabelaryści, w przeciwieństwie do arytmetyków politycznych, proponują opis liczbowy, ale nie proponują metody badawczej. Są więc raczej spadkobiercami nurtu opisu werbalnego państwa, choć ich propozycja używania danych w formie tabelarycznej jest aktualna do dziś. ***U tabelarystów następuje utożsamienie procesu gromadzenia i zestawiania danych w tabele z pojęciem statystyki.***

1770 Anglia Wychodzi angielskie wydanie książki barona **von Bielfelda**, w którym jeden z rozdziałów jest zatytułowany *Statystyka* i zawiera definicję tego pojęcia jako: „*Nauki, która uczy nas, jaki jest polityczny układ wszystkich współczesnych państw w znanym świecie*”[21]. W ten sposób termin *statystyka* trafia do Anglii.

1791-1795 Anglia **J. Sinclair,** angielski wydawca i organizator pierwszego rachunku statystycznego Szkocji wprowadza w swoim piśmiennictwie na grunt angielski przejęte z niemieckiego piśmiennictwa pojęcia *statystyka* i *statystyczny* jako *badania statystyczne, czyli badania dotyczące ludności, warunków politycznych, produkcji kraju i innych spraw państwa*[22]. Ponieważ Anglicy wyrazili zaskoczenie słowem „statystyka”, w jednej z prac tłumaczy się, dlaczego propaguje słowo niemieckie, a nie angielskie. Termin ten przyjął się, tak jak Sinclair miał nadzieję, na stałe w czasopiśmiennictwie angielskim, aczkolwiek jego znaczenie zmieniło się w ciągu pół wieku[23].

1796-1874 Belgia **A. Quetelet,** belgijski uczony uważany za pierwszego badacza, który zaproponował szerokie, systematyczne stosowanie statystyki jako metody badawczej do nauk społecznych, zbierał wszelakie dane o społeczeństwie i określał rozkład częstości występowania poszczególnych odmian wartości zjawiska, „tworząc” coś, co

[21] Cytat za Yule, Kendall [1966, s. 18].
[22] Cytat za Yule, Kendall [1966, s. 19].
[23] Bliżej patrz Yule, Kendall [1966, s. 19].

współcześnie nazywamy rozkładem normalnym cechy. Rao [1994, s. 55] przytacza przykład użycia przez niego takiego rozkładu do analizy zasięgu uchylania się od poboru do wojska we Francji. W tamtych czasach poborowi podlegali mężczyźni powyżej pewnego wzrostu. „Przez porównanie rozkładu wzrostu tych, którzy usłuchali wezwania do poboru, z aktualnym rozkładem wzrostu w populacji generalnej obliczył, że ponad 2000 mężczyzn uchyliło się od poboru, pozorując, że ich wzrost jest niższy od minimalnego". Był też propagatorem badań statystycznych na forum międzynarodowym, to pod jego wpływem Ch. Babbage założył towarzystwo statystyczne w Londynie (1834 r.). Jak stwierdza Rao [1994, s. 55]: „Współczesne pojęcia ekonomii i demografii, jak produkt narodowy brutto, tempo wzrostu i rozwoju oraz przyrost ludności, są spuścizną po Quetelecie i jego uczniach".

1797 Anglia — **Encyclopedia Britannica –** w wydaniu trzecim wymienia się pojęcie *statystyka* jako „słowo wprowadzone ostatnio, aby wyrazić obraz lub zwięzły opis jakiegoś królestwa, hrabstwa lub gminy"[24].

1800 Francja — Powstanie pierwszego „urzędu statystycznego" na świecie, czyli Centralnego Urzędu Statystycznego we Francji.

1821-1896 Niemcy — **E. Engel,** niemiecki statystyk i ekonomista, analizuje metodami statystycznymi dochody i wydatki rodzin robotniczych i określa występujące pomiędzy nimi prawidłowości, znane do dziś jako krzywe (prawa) Engla.

1830-1850 — Powstawanie urzędów statystycznych w wielu krajach europejskich.

1834 — Powstanie słynnego, istniejącego do dzisiaj, Royal Statistical Society[25] w Anglii.

Proces „ustalania się" definicji słowa *statystyka* nadal jednak nie jest zakończony. Artykuły pierwszego tomu czasopisma statystycznego[26] wydanego w latach 1838-1839 „(…) mają przeważnie charakter numeryczny, ale oficjalna definicja nie mówi nic o metodzie. „Statystyką", czytamy, „można nazwać, używając słów

[24] Za Rao [1994, s. 54].

[25] Królewskie Towarzystwo Statystyczne.

[26] „The Journal of the Royal Statistical Society"- najstarsze i jedno z najbardziej prestiżowych, wychodzące do dnia dzisiejszego, czasopismo statystyczne świata.

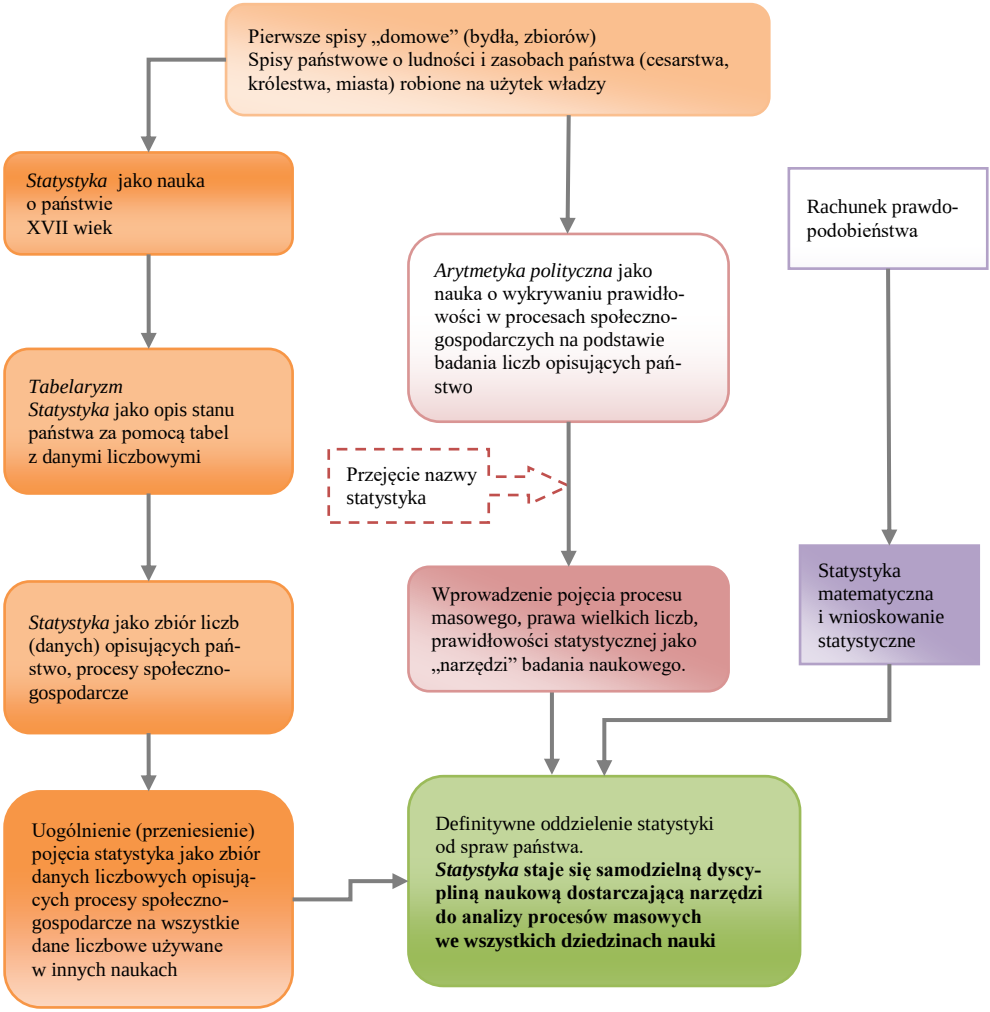
programu Royal Statistical Society, dziedzinę zajmującą się ustalaniem i zbieraniem tych faktów, które oblicza się dla przedstawienia warunków i perspektyw społeczeństwa", dodaje się jednak, że „statystyka najczęściej posługuje się liczbami i ujęciami tabelarycznymi"[27]. Uznaje się, że pod koniec XIX wieku i na początku XX słowo *statystyka* uzyskuje współczesne znaczenie (a właściwie znaczenia).

1853 **Odbywa się I Międzynarodowy Kongres Statystyczny w Brukseli**. Kongresy takie miały służyć (i do dzisiaj służą) przede wszystkim do osiągnięcia zgody co do pojęć i definicji zbieranych danych oraz ujednolicenia metod ich zbierania, tak by były one porównywalne między poszczególnymi państwami.

1885 Powstaje **Międzynarodowy Instytut Statystyczny** (*International Statistical Institute*), zadaniem którego jest realizacja uchwał kongresów statystycznych, czuwanie nad jednolitością w metodach opracowywania statystyk (zbiorów danych) i tworzenia wskaźników statystycznych.

W formie podsumowania, przedstawiono na rysunku 1.1 uproszczony schemat procesu tworzenia się pojęcia *statystyka* oraz dyscypliny naukowej zwanej statystyką we współczesnym tego słowa rozumieniu.

[27] Yule, Kendall [1966, s. 20].



Rysunek 1.1. Proces tworzenia się pojęcia *statystyka*

Źródło: opracowanie własne.

Jak wynika z powyższego schematu, statystyka zasadniczo składa się z dwóch podstawowych działów: statystyki opisowej (będącej przedmiotem wykładu w niniejszym podręczniku) oraz statystyki matematycznej.

Statystyka w Polsce

W Polsce, jak w wielu innych krajach, w 1470 roku sporządzono szczegółowy inwentarz dóbr kościelnych diecezji krakowskiej podany przez Jana Długosza. J. K. Haur[28] opracował i ogłosił w 1675 roku wzór zestawienia szacunku plonów, którą to pracę można uważać za pierwszą polską pracę ze statystyki w dzisiejszym tego słowa znaczeniu.

Uważa się jednak, że początki statystyki polskiej „(…) przypadają na okres polskiego Oświecenia (1780-1820)". Edward Rosset (1968) sądzi, że „(…) w tym właśnie czasie zakwitła w Polsce nowa gałąź wiedzy – statystyka w jej dzisiejszym rozumieniu, to znaczy nauka posługująca się językiem liczb". Za pionierów tak rozumianej statystyki uważa się Fryderyka Moszyńskiego (1755-1817), Tadeusza Czackiego (1765-1813), Stanisława Staszica (1755-1826) oraz Wawrzyńca Surowieckiego (1769-1827)"[29].

W okresie Sejmu Czteroletniego (1788-1792) przeprowadzono w Polsce, z inicjatywy Moszyńskiego, jeden z pierwszych na świecie spisów ludności, obejmujących ludność miejską i wiejską z pominięciem szlachty i duchowieństwa. Moszyński podjął też inicjatywę założenia i prowadzenia przez władze kościelne stałej ewidencji urodzeń i zgonów dla zapewnienia aktualizacji danych spisowych. W końcu XVIII wieku wiele miast polskich posiadało już spisy ludności (m.in. Kraków, Warszawa). „Swoimi inicjatywami doprowadził do powstania zorganizowanego systemu ewidencji ludności i niektórych rzeczowych elementów gospodarki narodowej"[30]. W 1807 roku Stanisław Staszic wydał (bezimiennie) pracę pt. *O statystyce Polski*, a w 1845 Tadeusz Czacki *Statystykę Polski*. Obaj autorzy posługiwali się, współcześnie uznanymi za poprawne, metodami zbierania danych liczbowych i grupowaniem statystycznym. W **1918** roku w Warszawie powstał **Główny Urząd Statystyczny**.

Niezależnie od rozwoju statystyki w końcu XVI i początkach XVII wieku rozwija się gałąź matematyki nazwana teorią prawdopodobieństwa[31]. Jej korzenie tkwią w ludzkich skłonnościach do hazardu. Gracze w kości lub karty, pragnąc wymyślić (ustalić) sposób na zdobycie wielkiej wygranej, czynili olbrzymie ilości obserwacji i próbowali dokonać uogólnień

[28] Za Puchalskim [1980, s. 13].

[29] Za Domańskim [2004 r, s. 67-75].

[30] Ibid. s. 66.

[31] W uproszczeniu jest to gałąź matematyki zajmująca się zdarzeniami losowymi, czyli takimi, wystąpienie których można określić tylko z jakimś prawdopodobieństwem.

pozwalających na zdobycie majątku. A że byli wśród nich wielcy matematycy, jak Galileusz (1564-1642), Blaise Pascal (1623-1662), Pierse de Fermat (1601-1665), Jan Bernoulli (1654-1705), Abraham de Moivre (1667-1754) rezultatem ich dociekań było powstanie nowej dyscypliny matematycznej – rachunku prawdopodobieństwa. W Polsce za pioniera rachunku prawdopodobieństwa uważa się Jana Śniadeckiego (1756-1830)[32].

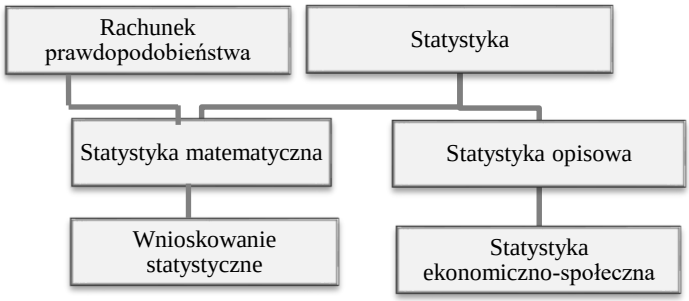
Istotny wzrost zainteresowania rachunkiem prawdopodobieństwa notuje się w XIX stuleciu, a wynikał on z potrzeby stosowania metod rachunku prawdopodobieństwa i wyodrębnionej z niego statystyki matematycznej do udzielania ubezpieczeń majątkowych i ubezpieczeń na życie.

Statystyka opisowa i statystyka matematyczna wykształcały się więc jako dwie odrębne dziedziny. Pierwsza na gruncie nauk społecznych, druga na gruncie matematyki. Z czasem zlały się tak silnie, że współcześnie uważa się je za jedną dyscyplinę – statystykę, a różnica w metodzie badania wynika z typu badania. W związku z powyższym wyodrębnić można dwa rodzaje badań statystycznych: **opis statystyczny** i **wnioskowanie statystyczne**.

Jeżeli celem badania jest opis całej zbiorowości, to obliczane są wówczas pewne parametry zbiorowości czy wskaźniki charakteryzujące tę zbiorowość (takie jak średnia lub średnie). Można wówczas określić stan interesującego zjawiska w badanej zbiorowości oraz wykryć prawidłowości w jej rozwoju. Użyta metoda nazywana jest **opisem statystycznym** lub można stwierdzić, że używa się narzędzi **statystyki opisowej** do analizy zjawiska w danej populacji. Jeżeli natomiast badana jest tylko część zbiorowości, a wyniki mają być uogólnione na całą zbiorowość, wówczas badanie bazuje na **wnioskowaniu statystycznym** i jest podstawowym narzędziem **statystyki matematycznej.**

Z czasem ze statystyki opisowej wyodrębniła się, przede wszystkim na użytek nauk społecznych, kolejna gałąź statystyki, nazwana **statystyką ekonomiczno-społeczną**, „zajmująca się" konstrukcją wskaźników ekonomicznych oraz dostarczająca metod analizy dynamiki zjawisk ekonomiczno-społecznych. Relacje pomiędzy poszczególnymi działami statystyki przedstawione zostały na schemacie zaprezentowanym na rysunku1.2.

[32] Zainteresowanych bliżej początkami rozwoju statystyki na ziemiach polskich odsyłamy do wspomnianego już artykułu prof. Domańskiego [2004 s. 67-75].



Rysunek 1.2. Relacje pomiędzy działami statystyki

Źródło: opracowanie własne.

Podsumowując, można stwierdzić, że statystyka rozwijała się trzema nurtami, które na początku XX wieku zlały się w jeden, niezależny od typu badań, czy to państwoznawczych, czy społecznych. Statystyka wyodrębniła się jako samodzielny dział nauki i zaczęła służyć wszystkim pozostałym dyscyplinom naukowym. *Mimo takiego zespolenia nadal występują trzy podstawowe sposoby rozumienia słowa „statystyka”.*

Pierwszy z nich traktuje tę dyscyplinę jako *zestaw danych liczbowych* odnoszących się do pewnej zbiorowości (np. statystyka przemysłu zawiera dane w postaci tablic opisujących rozwój i strukturę przemysłu w pewnym okresie i na określonym obszarze). Pierwotnie pojęcie to stosowano tylko do danych opisujących państwa, a więc do danych społeczno-gospodarczych. Dopiero z czasem tą nazwą zaczęto określać także dane liczbowe używane w innych dziedzinach. ***Współcześnie używa się go do zbioru uporządkowanych danych z dowolnej dziedziny.***

Drugi sposób rozumienia terminu statystyka **określa ją jako wartość charakteryzującą daną zbiorowość (parametr), która została otrzymana w wyniku operacji matematycznych na danych liczbowych.** Przykładem mogą być średnie wielkości dochodów ludności w poszczególnych państwach. Nazwy tej używa się głównie na gruncie statystyki matematycznej dla określenia parametrów opisu całej populacji.

Trzeci sposób rozumienia traktuje ją jako **„naukę o metodach badań poświęconych liczbowo wyrażalnym właściwościom zbiorowości”** [B. Szulc, s. 13].

W tym też ostatnim znaczeniu statystyka jest wykładana jako przedmiot akademicki, jak też jest przedmiotem niniejszego podręcznika.

Definicji statystyki jako nauki jest wiele i każdy autor definiuje ją nieco inaczej, aczkolwiek wszystkie te definicje mają wspólne jądro. Poniżej przytaczamy więc definicję encyklopedyczną, którą w następnym rozdziale rozwiniemy.

Statystyka[33] **jest to**: **„nauka zajmująca się metodami badania przedmiotów i zjawisk w ich masowych przejawach oraz ich ilościową lub jakościową analizą z punktu widzenia dyscypliny nauk, w której zakres wchodzą".**

Kluczowe pojęcia

Statystyka
Statystyka opisowa
Statystyka matematyczna
Opis statystyczny
Wnioskowanie statystyczne
Statystyka ekonomiczno-społeczna
Teoria (rachunek) prawdopodobieństwa

Pytania kontrolne

1. Skąd wywodzi się pojęcie „statystyka" i co ono pierwotnie oznaczało?
2. Jakie były etapy w rozwoju statystyki na świecie?
3. Na jakie podstawowe działy dzielimy statystykę?
4. Jak przebiegał rozwój statystyki w Polsce?
5. Z jakimi dziedzinami powiązana jest statystyka?
6. W czym tkwi podstawowa różnica pomiędzy statystyką opisową a matematyczną?
7. Co jest podstawowym celem statystyki jako nauki?

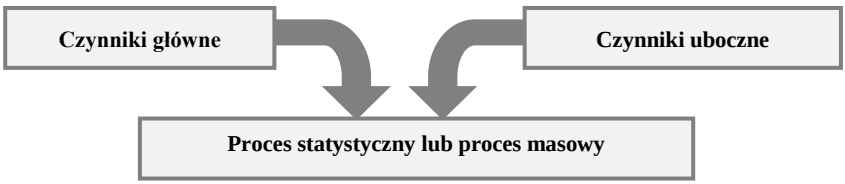
[33] Mała Encyklopedia Statystyki, s. 592.

2. Podstawowe pojęcia statystyki

Zanim zostaną przedstawione podstawowe narzędzia opisu statystycznego, konieczne jest wyjaśnienie kilku pojęć związanych z tą dziedziną nauki. Zacznijmy od pojęcia procesu masowego.

Procesy masowe są to zjawiska, które zachodzą często i/lub dotyczą wielu obiektów, jak np. urodzenia, zgony, akty kupna, sprzedaży, produkcja. Zjawiska te podlegają pewnym prawom (prawidłowościom), które można dopiero wychwycić, badając obiekty podlegające tym zjawiskom, tylko w znacznych ilościach, a nie pojedynczo. Przykładowo obserwacja jednego nabywcy (konsumenta) nic nie powie nam o prawidłowościach rządzących popytem na dobra konsumenckie. Badania przeprowadzone na tysiącach konsumentów pozwoliły uchwycić pewne prawidłowości[34] „rządzące" aktami kupna dóbr. Obserwacje tysięcy konsumentów pozwoliły na przykład stwierdzić, że generalnie popyt na dane dobro zależy ujemnie od ceny tego dobra, czyli podnosząc cenę dobra ograniczamy popyt (ilość nabywanego produktu) i odwrotnie – jeśli chcemy sprzedać więcej dobra, to powinniśmy obniżyć jego cenę[35].

Prawa, czy też prawidłowości, które można uchwycić tylko obserwując znaczną liczbę obiektów, nazywamy prawidłowościami statystycznymi (lub masowymi). Proces masowy można utożsamiać więc z prawidłowością masową, jeśli takowa istnieje. Na **proces masowy,** zwany też **procesem statystycznym lub stochastycznym,** oddziałują dwie grupy czynników kształtujących go: *czynniki główne* (zwane też *systematycznymi*) i *czynniki uboczne* (często nazywane *czynnikami losowymi lub przypadkowymi*), co ilustruje rysunek 2.1.



Rysunek 2.1. Proces statystyczny

Źródło: opracowanie własne.

[34] Dlatego będziemy nazywać je prawidłowościami statystycznymi.

[35] Przy założeniu tzw. warunku *ceteris paribus*, czyli że pozostałe czynniki wpływające na popyt pozostają niezmienione i możemy abstrahować od ich wpływu.

Czynniki główne są to czynniki działające jednakowo na wszystkie badane obiekty w analizowanym zjawisku. Przykładowo, badając popyt gospodarstw domowych na dany typ dobra, za czynnik główny określający, ile danego dobra konsument nabędzie w jednostce czasu, uznaje się cenę i/lub dochód konsumenta[36].

Druga grupa to **czynniki uboczne,** które nie działają jednakowo na wszystkie jednostki lub działają tylko na niektóre, co powoduje, że elementy zbiorowości mają zróżnicowane wartości cechy pomimo wpływu tych samych czynników głównych. Przykładem czynników ubocznych działających na popyt gospodarstw domowych, przy założeniu, że istnieje jeden czynnik główny (cena), są na przykład indywidualne upodobania konsumpcyjne członków gospodarstwa domowego czy ich struktura wiekowa.

Gdyby na zjawisko działały tylko przyczyny główne, wspólne wszystkim jednostkom, wówczas każdy pojedynczy proces przebiegałby identycznie i statystyka byłaby niepotrzebna. Wnioski o zjawisku można by było wyciągnąć w oparciu o obserwację jednego obiektu i przenieść na wszystkie obiekty zbiorowości. Byłaby to **zależność deterministyczna**. Z drugiej strony, gdyby dane zjawisko kształtowały tylko czynniki uboczne (losowe), wówczas nie bylibyśmy w stanie wykryć żadnych prawidłowości – **zależność** byłaby **chaotyczna** w potocznym tego słowa znaczeniu (czyli nie byłoby zależności).

Tam, gdzie obiekty badanej zbiorowości podlegają zarówno działaniu czynnika głównego (czynników głównych), jak i niejednakowym wpływom czynników ubocznych, wykrycie prawidłowości w badanym procesie wymaga zbadania jak największej liczby jednostek, by uchwycić prawidłowość przebijającą się przez działanie czynników ubocznych. Prawidłowość ta będzie reprezentowała składnik systematyczny zjawiska, zaś działanie czynników ubocznych będzie reprezentowane przez składnik losowy. O takiej zależności mówimy, że jest ***statystyczna*** lub ***stochastyczna***.

Prawidłowość statystyczna = składnik systematyczny + składnik losowy

Zarówno wówczas, gdy proces jest deterministyczny, jak i wtedy, gdy jest chaotyczny, statystyka nie znajduje zastosowania. W pierwszym przypadku jest niepotrzebna, ponieważ wystarczy obserwacja jednego obiektu, by wiedzieć, co będzie się działo z pozostałymi, zaś w drugim nie ma zastosowania, ponieważ w procesach chaotycznych przebadanie całej zbiorowości nie da nam informacji,

[36] Możemy prowadzić badanie, zakładając, że istnieje jeden czynnik główny, lub przyjąć, że jest ich więcej, gdyż zasadniczo sposób badania jest taki sam. O różnicach w sposobach badania wynikających z ilości czynników głównych będzie mowa w następnej części podręcznika, dotyczącej analizy więcej niż jednej cechy i związków między cechami.

jakim prawom ta zbiorowość podlega, gdyż takich praw tam nie ma. Oznacza to brak istnienia czynnika (czynników) głównego kształtującego zjawisko, a jedynie oddziaływanie realizowane przez znaczną ilość czynników ubocznych, z których żaden nie przeważa na tyle, by go nazwać głównym. Możemy teraz rozszerzyć pojęcie statystyki jako nauki w stosunku do definicji z poprzedniego rozdziału.

Statystyka *to nauka o metodach analizy zbiorowości podlegających działaniu prawidłowości masowych (statystycznych). Służy wykrywaniu prawidłowości, które istnieją, ale nie są bezpośrednio widoczne wskutek działania czynników losowych i ich uchwycenie wymaga zbadania dużej liczby obiektów. Im więcej obiektów zbadamy, z tym większą precyzją jesteśmy w stanie określić istniejącą prawidłowość w badanym procesie.*

W większości dyscyplin naukowych odkrywane tam prawa to prawa statystyczne określane poprzez analizę procesów masowych. Przykładem takich praw (zależności) są na przykład prawa Engla wykorzystywane w ekonomii, mówiące o zależnościach pomiędzy popytem na dobra a dochodami konsumentów, w fizyce prawa termodynamiki, w demografii zaś cała analiza opiera się na prawidłowościach statystycznych, podobnie jak w biologii.

Zdefiniujmy teraz kolejne podstawowe pojęcia, których będziemy używali w dalszych częściach podręcznika. Aby wykryć prawidłowości statystyczne w procesach masowych, badamy duże zbiory obiektów. Podstawowym wymogiem niezbędnym do poprawnego określenia istnienia prawidłowości statystycznej jest **jednorodność badanej populacji** ze względu na badaną cechę. Oznacza to, że wszystkie badane obiekty muszą podlegać tym samym czynnikom głównym określającym badaną cechę.

Zbiór obiektów podlegających badaniu nazywamy **zbiorowością całkowitą** lub **populacją generalną.** Jeżeli wszystkie obiekty podlegają badaniu, mówimy o **badaniu całościowym** (całkowitym), np. wszyscy studenci Politechniki Białostockiej zbadani pod kątem uzyskanych ocen w trakcie semestru czy wszyscy pracownicy pewnej korporacji analizowani ze względu na otrzymywane zarobki. Jeżeli badaniu, z różnych względów[37], podlega część populacji, mówimy o **badaniu częściowym**. Część zbiorowości podlegająca badaniu nazywana jest **zbiorowością częściową** (cząstkową, podpopulacją) lub **próbą**. Przykładem próby są studenci Wydziału Zarządzania, stanowiący podpopulację studentów Politechniki Białostockiej, czy pracownicy na przykład działu marketingu, stanowiący zbiorowość częściową wszystkich pracowników przedsiębiorstwa.

[37] Na ogół ze względu na koszty bada się tylko część populacji, gdy jest ona duża lub gdy badanie jest niszczące (np. jakość wyprodukowanych konserw, wytrzymałość liny wspinaczkowej).

Próby z kolei dzielimy na **losowe** i **nielosowe**. Ponieważ niniejszy podręcznik traktuje tylko o metodach badania całej populacji, nie będziemy omawiać bliżej rodzajów prób i problemów z tego wynikających[38].

Obiekty w populacji podlegające badaniu nazywamy jednostkami statystycznymi. Przykładem jednostki statystycznej jest każdy student Wydziału Zarządzania Politechniki Białostockiej czy każdy pracownik określonego przedsiębiorstwa, lub też każdy produkt podlegający kontroli jakości.

Każda populacja generalna badana jest pod kątem określonej właściwości, nazywanej cechą statystyczną. **Cecha statystyczna jest to interesująca badacza własność, która różnicuje jednostki statystyczne w populacji**. Przykładem może być ocena z egzaminu ze statystyki studentów danego roku zarządzania, zarobki każdego z pracujących w Polsce, liczba turystów przyjeżdżających w ciągu roku do poszczególnych krajów świata, oceny studentów uzyskane na zaliczeniach w konkretnym semestrze czy liczba wadliwych produktów w różnych ich partiach[39].

Kolejnym pojęciem powiązanym z badaniem statystycznym jest liczebność. Jako **liczebność** rozumiana jest **liczba jednostek statystycznych mających daną cechę w badanej populacji lub podpopulacji**. Przykładowo liczba studentów w PB, liczba studentów Wydziału Zarządzania PB, liczba studentów, którzy uzyskali konkretnego typu oceny na egzaminie z przedmiotu statystyka, liczba pracowników o danym typie wykształcenia.

Przyporządkowanie poszczególnym odmianom cechy ich liczebności lub częstości występowania w zbiorowości, czyli podanie (ustalenie), jak dana cecha rozkłada się w populacji, nazywane jest podaniem rozkładu statystycznego cechy.

Przykładem podania rozkładu cechy jest określenie, jakie uzyskali oceny (cecha) poszczególni studenci (jednostki statystyczne) na egzaminie ze statystyki w semestrze letnim roku 2017 spośród wszystkich 100[40] przebadanych studentów (populacja i jej liczebność) kierunku zarządzanie, co przedstawiono poniżej w tabeli 2.1.

---

[38] Próba losowa spełnia pewne wymogi, w związku z czym może służyć do wyciągania wniosków o prawidłowościach w całej populacji, a nie tylko w badanej części. Jak łatwo się domyślić, próba nielosowa nie może służyć do wyciągania wniosków o tym, co dzieje się poza nią. Bardziej dociekliwemu czytelnikowi rekomendujemy zajrzenie do dowolnego podręcznika o wnioskowaniu statystycznym.

[39] Zauważmy, że zawsze należy bardzo precyzyjnie określić, kto lub co jest obiektem badania, czyli jednostką statystyczną, i należy do badanej populacji oraz w jakim momencie lub przedziale czasu przeprowadza się badanie.

[40] Liczbę sto przyjęto dla wygody.

Tabela 2.1. Szereg statystyczny

Ocena x_i	Liczebność n_i	Szereg skumulowany n_{isk}
2	15	15
3	22	15+22=37
3,5	18	37+18=55
4	32	55+32=87
4,5	8	87+8=95
5	5	95+5=100

Źródło: dane umowne.

Powyższa tabela przedstawia głównego „bohatera" na pewnym etapie badania statystycznego. **Jest nim uporządkowany zbiór obserwacji danej cechy statystycznej tworzący szereg statystyczny**. **Rozkład statystyczny jest więc jednocześnie tożsamy z szeregiem statystycznym.**

Przedmiotem badania w powyższym przykładzie przedstawionym w tabeli 2.1 jest rozkład ocen studentów na egzaminie ze statystyki w semestrze letnim roku 2017. Badanie jest całkowite, bo dotyczy wszystkich studentów, którzy przystąpili do egzaminu w danym terminie, czyli jest to też badana populacja generalna. Jednostką statystyczną jest każdy student, a badaną cechą jest jego ocena z egzaminu.

Aby utworzyć szereg statystyczny (czyli określić rozkład cechy), należy dokonać pomiaru wartości cechy[41]) dla każdej jednostki statystycznej. Gdy mamy zrobiony pomiar cechy w zbiorowości, otrzymujemy nieuporządkowany, wstępny zbiór danych. Nieuporządkowane dane z pomiarów statystycznych nazywamy **danymi surowymi**. Trudno jest na ich podstawie znaleźć to, czego szukamy, czyli prawidłowość statystyczną. Surowe dane wyglądają zazwyczaj chaotycznie i należy je uporządkować. **W zależności od sposobu uporządkowania danych rozróżniamy dwa rodzaje szeregów statystycznych: *szeregi proste i szeregi rozdzielcze*.**

Szeregi proste – składają się z uporządkowanych monotonicznie[42] wartości cechy, na przykład ocen z kolokwium: 2, 2, 2, 2, 2, 2, 3, 3, 3, 3, 3, 3, 3, 3, 4, 4, 4, 4, 4, 4, 4, 4, 5, 5, 5, 5.

Szeregi rozdzielcze – są to szeregi uporządkowane, zawsze składające się dwóch kolumn[43]. W pierwszej kolumnie podaje się wszystkie odmiany cechy wy-

[41] Lub w przypadku cechy niemierzalnej określić jej odmianę.

[42] W przypadku cechy opisowej uporządkowanej według dowolnego klucza, np. alfabetycznie lub odnośnie poszczególnych odmian cechy, np. w przypadku koloru oczu: brązowy, brązowy, brązowy, …, brązowy, niebieski, niebieski, niebieski, …niebieski, zielony, zielony, zielony, zielony, …, zielony.

[43] Zakładamy tu badanie jednej cechy naraz.

stępujące w badanej populacji. W drugiej kolumnie podaje się, jak często poszczególne odmiany cechy występują. Częstość występowania określamy bądź liczbą jednostek statystycznych, bądź ich procentem (udziałem) w populacji. W powyższym przykładzie w tabeli 2.1 mamy jeszcze trzecią kolumnę. Zawiera ona skumulowane, czyli zsumowane kolejno, liczebności z kolumny drugiej. Szereg taki nazywamy **szeregiem skumulowanym.** Jest to monotonicznie uporządkowany zbiór odmian danej cechy statystycznej oraz dodawanych sukcesywnie częstości występowania kolejnych odmian cechy. Skumulowane częstości informują nas, ile jednostek statystycznych ma wartość cechy nie większą od danej wartości. Często nazywa się szereg skumulowany **dystrybuantą empiryczną.**

Analizując dane prezentowane w tabeli 2.1, można stwierdzić, że na przykład co najwyżej czwórkę otrzymało 87 studentów. Zauważmy też, że w ostatnim wierszu zapisana jest całkowita liczebność zbiorowości – w analizowanym przykładzie to 100 studentów. Jeżeli szereg jest zapisany w formie częstości względnych, a nie liczebności, wówczas częstość całkowita będzie równa jedynce lub stu procentom

Badając zbiorowości statystyczne, mamy do czynienia z różnymi rodzajami cech statystycznych. W zależności od przyjętego kryterium cechy statystyczne dzieli się na pewne typy i podział ten ma zasadniczy wpływ na sposób przetwarzania zebranego materiału statystycznego. Podstawowym, najbardziej istotnym podziałem jest podział na: **cechy mierzalne i niemierzalne inaczej nazywane odpowiednio ilościowymi i jakościowymi**.

Cechy niemierzalne to własności jednostek statystycznych zbiorowości, których nie da się zmierzyć, a jedynie można opisać. Poszczególne odmiany cechy niemierzalnej określamy słowami. Przykładami cech niemierzalnych są na przykład płeć, wykształcenie, kolor oczu, upodobania i odczucia konsumentów w badaniach rynkowych opisywane słowami typu: ładne, brzydkie, podoba się, nie podoba, nie mam zdania.

Cechy mierzalne to własności badanych obiektów, których wartość można bezpośrednio zmierzyć, czyli przypisać im liczbę. Mogą one występować w dwóch formach. Pierwszą z nich jest wartość wyrażona w liczbach bezwzględnych (często używa się pojęcia „w poziomach"), jak na przykład wzrost, waga, płaca, liczba przybyłych do danego regionu turystów w określonej jednostce czasu, wielkość zysku, wielkość produkcji, sprzedaży czy liczba dzieci w rodzinie. Druga forma to wskaźniki, czyli liczby względne, na przykład liczba studentów na tysiąc mieszkańców w poszczególnych krajach europejskich, wielkość dochodu narodowego na głowę mieszkańca, dochód na gospodarstwo domowe, liczba wadliwych produktów przypadających na milion wytworzonych produktów. W przypadku wskaźników należy dokonać podwójnego pomiaru w poziomach: odrębnie dla licznika i mianownika, a następnie podzielić je tak, aby uzyskać cechę w interesującej nas postaci. Cechy przedstawiane w postaci wskaźników (relacji) są bardzo użyteczne

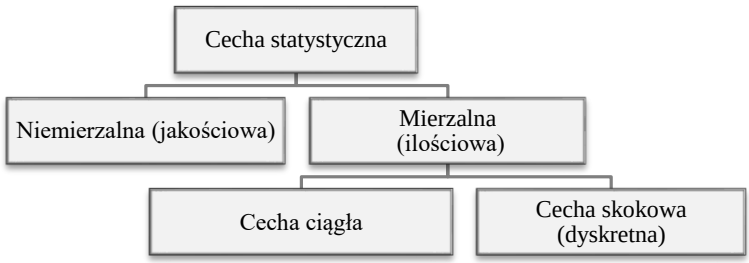
we wszelkich porównaniach, zasadniczo jednak ich analiza przebiega podobnie jak analiza cech wyrażonych w liczbach bezwzględnych, w związku z czym w dalszej części pracy będziemy wykorzystywać, dla uproszczenia, pojęcie cechy mierzalnej przedstawionej w postaci liczb bezwzględnych.

Wartości cech mierzalnych mogą pochodzić zarówno ze zbioru liczb rzeczywistych, jak i jego części, ograniczonej do liczb całkowitych lub naturalnych. Cechy mierzalne często nazywane są zmiennymi, choć pojęcie to wykorzystywane jest częściej na gruncie statystyki matematycznej. Jednak w sensie matematycznym nasze cechy także są zmiennymi, ponieważ przyjmują zmieniające się wartości[44].

Biorąc pod uwagę, jakie wartości mogą przyjmować, cechy ilościowe (mierzalne) dzielimy na **ciągłe i skokowe**.

Cechy ciągłe przyjmują wartości z przedziału liczb rzeczywistych, jak na przykład wzrost, waga, zysk, płace. Cechy te mogą więc mieć, przynajmniej teoretycznie, nieskończoną liczbę odmian.

Cechy skokowe, inaczej zwane dyskretnymi, przyjmują przeliczalną liczbę wartości, czyli z dziedziny liczb całkowitych, a w naukach społecznych najczęściej z dziedziny liczb naturalnych. Przykłady to chociażby liczba dzieci w rodzinie, liczba osób w gospodarstwie domowym, liczba sprzętu AGD w gospodarstwie domowym, liczba posłów w sejmach różnych krajów, liczba kwater turystycznych w gminach. Schematycznie klasyfikacja cech została przedstawiona na rysunku 2.2.



Rysunek 2.2. Klasyfikacja cech statystycznych

Źródło: opracowanie własne.

Podział na cechy ciągłe i skokowe w praktyce nie do końca jest jednoznaczny[45]. Wynika to z kilku faktów. Pierwszym z nich jest to, iż cechy, które z punktu widzenia matematycznego są ciągłe, w praktyce statystycznej stają się „nieciągłe",

[44] Oczywistym jest faktem, że gdyby wszystkie jednostki statystyczne miały tę samą wartość cechy, nie byłoby potrzeby użycia metod statystycznych.

[45] Szersze omówienie problemu można znaleźć np. w podręczniku Szulca [1967 i nast. wydania, rozdział 1].

ponieważ ***pomiar empiryczny dostarcza skończonej liczby obserwacji***. Biorąc przykładowo wagę człowieka, można uznać, że jest to cecha ciągła z przedziału od 0 do, powiedzmy, 300 kg, czyli może przyjmować nieskończenie wiele wartości w tym przedziale. W praktyce jednak, mierząc na przykład całą ludność Polski, otrzymamy skończoną liczbę pomiarów związaną z liczbą ludności, czyli zbiór „nieciągły".

Kolejnym czynnikiem wprowadzającym nieciągłość jest **dokładność pomiaru**. Pozostając przy przykładzie wagi człowieka, możemy określić ją tylko z pewną dokładnością związaną z dokładnością przyrządu pomiarowego. Jeżeli pomiaru dokonuje się z dokładnością do dekagramów, wówczas pomiędzy dwoma pomiarami wagi z dokładnością do jednego dekagrama nie będzie wyników pośrednich. Zatem obserwacja empiryczna materiału statystycznego wprowadza nieciągłość, z punktu widzenia matematyki.

Z odwrotną sytuacją możemy mieć do czynienia w przypadku cechy skokowej. Jeśli **cecha, teoretycznie skokowa, ma tak dużo odmian**, że jej analiza empiryczna jest, praktycznie rzecz biorąc, niemożliwa bądź kłopotliwa, wówczas w badaniu statystycznym przyjmuje się ją za ciągłą[46]. Przykładem tego są, praktycznie, wszystkie cechy prezentowane w ujęciu wartościowym, takie jak np. dochody, płace, wydatki. W tym przypadku pomiar dokonywany jest z dokładnością do jednego grosza[47] i w praktyce cechy te traktuje się jako ciągłe. Innym nieco przypadkiem jest traktowanie cechy dyskretnej jako ciągłej, jeśli jej liczba odmian jest bardzo duża, przy czym skrajne wartości występują bardzo rzadko. Wówczas, konstruując rozkład cechy (szereg), często tworzy się szereg jak dla cechy ciągłej, otrzymując „lepszy", z punktu widzenia badacza i celu badania, rozkład tejże cechy. Taką cechę określa się często jako quasi-ciągłą. Przykładem może być analiza liczby dzieci w rodzinie. Powiedzmy, że badanie wykazało, iż w badanej milionowej populacji najmniejsza liczba dzieci to zero, a największa to piętnaścioro, przy czym rodzin o liczbie dzieci ponad ośmioro jest bardzo mało, po kilka do kilkudziesięciu. Szereg dla piętnastu odmian cechy będzie mało czytelny i często konstruuje się taki rozkład jako rozkład cechy ciągłej[48]. Innym przykładem zmiennej quasi-ciągłej może być analiza liczby mieszkańców w miastach polskich wykonana w określonym momencie. W przykładzie tym występuje nieograniczona praktycznie liczba możliwych wariantów (od kilkudziesięciu tysięcy do kilku milionów mieszkańców), przy jednocześnie znacznie mniejszej liczbie miast. Ta zmienna

[46] W następnym rozdziale pokażemy, jaki to ma wpływ na tworzenie szeregu statystycznego.

[47] „Na przykład zarobki przybierają wartości równe pełnym groszom, ale w przedziale od 1000 do 2000 złotych liczba odmian, jakie mogą one przyjąć, wynosi aż 100 001 (…)" [B. Szulc 1967, s. 20].

[48] Jest to uproszczony, dla celów dydaktycznych, przykład uciąglenia zmiennej dyskretnej. Należy mieć świadomość, że przy tylko piętnastu odmianach cechy szereg powinno się przedstawić jako szereg cechy skokowej.

jest skokowa, bo zbiór wartości jest zbiorem przeliczalnym, ale powinna być analizowana jak ciągła. Bliżej o problemach konstrukcji rozkładu cechy będziemy mówić w następnym rozdziale.

Określenie więc, czy cecha mierzalna jest ciągła czy skokowa, nie zawsze jest zgodne z teorią matematyki i często jest to decyzja badacza, jak traktować daną zmienną. Zależy ona w znacznej mierze od zebranego materiału statystycznego oraz celu wykorzystania badania i związanej z tym dokładności, z jaką badacz chce opisać strukturę populacji pod względem badanej cechy.

Rozważmy następujący przykład, który ukaże pewne pułapki mogące mieć miejsce, jeżeli rozpoczyna się badanie statystyczne bez podstaw teoretycznych.

Przykład 2.1

Dziadek zabrał dwójkę swoich wnuków na spacer do zoo. Po drodze spotkali zapalonego statystyka-amatora. Ten z pasją postanowił pytać przechodniów o wzrost, by określić średni wzrost populacji bywalców ZOO. Ponieważ pierwszymi napotkanymi osobami byli właśnie nasi spacerowicze, więc podali następujące wartości: 176 cm, 110 cm, 98 cm. Statystyk-amator policzył momentalnie średnią[49], która wyniosła 128 cm. Czy ta średnia dała jakąś informację naszemu „statystykowi"? Te wyliczenia nie reprezentują przecież ani wzrostu dziadka, ani wzrostu wnuków.

Wykonana analiza jedynie potwierdziła amatorstwo naszego statystyka, gdyż nie wziął pod uwagę **warunków, w jakich stosuje się statystykę: koniecznej jednorodności populacji oraz dużej liczby obserwacji,** w związku z czym wynik jego badania nie ma sensu. Nie możemy na jego podstawie wyciągnąć sensownych wniosków. Analizowane dane nie reprezentują jednej populacji, lecz dwie: populację ludzi dorosłych i dzieci. Można w tym przypadku analizować je zamiast pytać o wiek, pytać o liczbę zjedzonych lodów czy wykonanych kroków od domu

[49] Pojęcie średniej szerzej zostanie omówione w kolejnym rozdziale, na tym etapie traktujemy ją jako sumę wartości podzieloną przez ich liczbę.

do zoo (pod warunkiem, że dzieci szły spokojnie, a nie skakały z radości wokół idącego wolnym krokiem dziadka). Jednak nadal populacja ta jest zbyt mała, by dało się uchwycić prawidłowość statystyczną.

Kluczowe pojęcia

Proces masowy
Proces statystyczny
Prawidłowość masowa lub statystyczna
Czynnik główny
Czynnik uboczny
Składnik systematyczny
Składnik losowy
Zbiorowość statystyczna
Populacja generalna
Zbiorowość częściowa
Podpopulacja
Jednostka statystyczna
Cecha statystyczna
Liczebność
Rozkład cechy
Szereg statystyczny
Szereg prosty
Szereg rozdzielczy
Szereg skumulowany
Cecha mierzalna
Cecha niemierzalna
Cecha jakościowa
Cecha ilościowa
Cecha ciągła i quasi-ciągła
Cecha skokowa

Pytania kontrolne

1. Jakie przyczyny składają się na powstanie zjawiska masowego?
2. Kiedy stosujemy statystykę opisową jako narzędzie badania?
3. Czy przyczyny uboczne, oddziałujące na poszczególne jednostki procesu masowego, mogą uniemożliwić sformułowanie prawidłowości statystycznej?
4. Kiedy czynnik uboczny, składający się na strukturę zjawiska masowego, może być uznany za nieistotny (nie mający wpływu) w procesie kształtowania prawidłowości statystycznej?
5. W czym tkwi podstawowa różnica pomiędzy statystyką opisową a matematyczną?
6. Na czym polega różnica pomiędzy ilościowymi cechami: ciągłą i skokową?
7. Jaką (kiedy, w jakich warunkach) cechę określamy jako quasi-ciągłą?
8. Od jakiego pojęcia łacińskiego pochodzi słowo *statystyka*?
9. Jakie znasz sposoby rozumienia słowa *statystyka* (podaj trzy)?
10. Na jakie trzy główne gałęzie dzieli się statystykę?
11. Co oznaczało pierwotnie słowo *statystyka*?
12. Podaj nazwisko jednego znanego statystyka polskiego.
13. Kiedy stosujemy statystykę matematyczną jako narzędzie badania?
14. Zdefiniuj czynnik główny procesu masowego i podaj przykład.
15. Zdefiniuj czynnik uboczny procesu masowego i podaj przykład.

16. Kiedy nie można zastosować statystyki jako metody badania zjawiska?
17. Co oznacza słowo *liczebność* w statystyce?
18. Co oznacza słowo *częstość* w statystyce (choć nie tylko w statystyce)?

Polecenia

Dokończ poniższe zdania:

1. Podstawowym celem statystyki jako nauki jest ……...
2. Podaj przykład zjawiska masowego ……..……...
3. Badanie pełne różni się od częściowego tym, że ……..
4. Badanie reprezentacyjne jest to badanie oparte na ……...
5. Badanie pełne przeprowadza się w oparciu o zbiorowość ……..
6. Badanie częściowe przeprowadza się w oparciu o zbiorowość ……...
7. Cecha skokowa jest to cecha, której wartości ……
8. Cecha ciągła jest to cecha, której wartości ……..……..……...
9. Cecha mierzalna jest to cecha ……...
10. Cecha niemierzalna jest to cecha ……..
11. Cecha statystyczna jest to …….……..……..
12. Jednostka statystyczna jest to ……...…….……...
13. Zbiorowość statystyczna jest to …….……...
14. Szereg statystyczny jest to ……..……..
15. Przykładem zbiorowości statystycznej jest ……..
16. Przykładem jednostki statystycznej jest ……...
17. Przykładem cechy statystycznej jest …….....
18. Przykładem procesu masowego jest ……
19. Przykładem cechy statystycznej jakościowej jest ……
20. Przykładem cechy statystycznej ilościowej ciągłej jest …….....
21. Przykładem cechy statystycznej ilościowej skokowej jest ……..
22. Przykładem cechy statystycznej niemierzalnej jest …….
19. Przykładem cechy statystycznej quasi-ciągłej jest ……..

3. Prezentacja danych statystycznych

Statystyka zajmuje się, jak to już zostało stwierdzone w poprzednich rozdziałach, analizą zjawisk zachodzących masowo. Zebranie i analiza dużej ilości danych wymaga zastosowania szeregu jasno sprecyzowanych metod i narzędzi. W niniejszym rozdziale zostały przedstawione podstawowe sposoby gromadzenia informacji statystycznej czyli danych, formy i metody prezentacji zgromadzonego materiału statystycznego, a przede wszystkim zasady konstrukcji poprawnego szeregu statystycznego, czyli rozkładu cechy.

3.1. Pomiar w statystyce, czyli skale pomiarowe

Pomiar w statystyce oznacza w pewnym uproszczeniu, że każdej jednostce zbiorowości statystycznej przyporządkowuje się „wartość" badanej cechy. Jednak nie zawsze jest to wartość liczbowa. W przypadku cech niemierzalnych słowo „wartość" oznacza werbalne określenie odmiany cechy (dlatego powyżej wzięto je w cudzysłów).

Biorąc pod uwagę, jakie są możliwe operacje na wynikach pomiaru, S. Stevens zaproponował[50] cztery podstawowe skale pomiarowe: nominalną, porządkową (inaczej rangową), przedziałową i ilorazową. Każda kolejna skala umożliwia dokonanie dodatkowych operacji w stosunku do skali poprzedniej (słabszej). Poniżej omówimy ich własności.

3.1.1. Skala nominalna

Najsłabsza ze skal – **skala nominalna** – **dotyczy cech jakościowych**, czyli niemierzalnych. Posługując się nią, można jedynie podzielić obiekty na podzbiory według badanej cechy, czyli określić, jaką odmianę cechy ma dany obiekt. Nie można w tym przypadku uporządkować obiektów wartościująco, np. badając koloru oczu grupy studentów, nie określimy, że ktoś ma więcej lub mniej „koloru" czy że ktoś jest lepszy ze względu na dany kolor oczu.

[50] Szerzej patrz np. H.M. Blalock [1975, rozdział 2].

Tabela 3.1. Rozkład płci pracowników zakładu A

Płeć	Liczba obiektów
kobieta	200
mężczyzna	300

Źródło: dane umowne.

Przykładami cech, dla których stosowana jest skala nominalna, są: płeć, wyznanie, kierunek studiów, stan cywilny, status zawodowy czy społeczny, grupa krwi, miejsce zamieszkania czy urodzenia. Tabele 3.1 i 3.2 prezentują przykłady cech, które zmierzono w skali nominalnej[51].

Tabela 3.2. Rozkład wyznania mieszkańców regionu XYZ

Wyznanie	Liczba osób
Katolickie	277
Prawosławne	322
Muzułmańskie	102
Inne	30
Bez wyznania	266

Źródło: dane umowne.

Skala nominalna jest wykorzystywana do opisania struktury zbiorowości, jak również do grupowania jednostek. Można jej użyć także w przypadku cech mierzalnych, gdy interesuje nas tylko struktura rozkładu cechy, a nie jej wartości.

3.1.2. Skala porządkowa

Skala porządkowa jest skalą mocniejszą od poprzedniej. Oznacza to, że można wykonywać w niej te same działania co w skali nominalnej, jak również dodatkowo umożliwia rangowanie, czyli porządkowanie obiektów przy pomocy relacji mniejsze lub większe według znaczenia, wielkości czy natężenia. Badane obiekty są wówczas uszeregowane w pewnym, ustalonym przez badacza, porządku. Pomiędzy cechami mierzonymi na skali porządkowej nie można określić odległości (różnicy) pomiędzy poszczególnymi wartościami cechy, a jedynie wskazać, który element jest mniejszy (niższy, gorszy) od innego. Przykładami cech mierzonych w skali porządkowej są: poziom wykształcenia, klasa społeczna, status zawodowy, siła reakcji na bodziec czy ocena produktu, opinie i poglądy na określony temat.

[51] Dodajmy też, że nie da się ich „zmierzyć" w żadnej innej skali.

Biorąc pod uwagę, przykładowo, rozkład poziomu wykształcenia przedstawiony powyżej w tabeli 3.3, można uporządkować go od najniższego do najwyższego, ale nie jest się w stanie stwierdzić, jaka jest różnica między kolejnymi poziomami w sensie następującym: czy wykształcenie wyższe jest „dalej" czy „bliżej" średniego niż wykształcenie gimnazjalne. Możemy tylko określić, że na przykład dwie jednostki mają takie samo wykształcenie lub jedna ma wykształcenie wyższe (niższe) od drugiej.

Tabela 3.3. Rozkład wykształcenia pracowników pewnego przedsiębiorstwa

Poziom wykształcenia pracowników	Liczba jednostek n_i
podstawowe	120
gimnazjalne	80
średnie	200
wyższe	50

Źródło: dane umowne.

Skala porządkowa wykorzystywana jest bardzo często w różnego rodzaju badaniach ankietowych związanych z preferencjami konsumentów. Przykład takiego zestawienia podano w tabeli 3.4. Tu również można określić, że dany produkt został oceniony bardzo wysoko lub bardzo nisko, ale nie ma możliwości wyznaczenia różnicy w ocenach.

Tabela 3.4. Rozkład ocen smaku jogurtu firmy X

Ocena smaku jogurtu	Liczba ocen (n_i)
Bardzo smaczny	110
Smaczny	200
Nie mam zdania	120
Niesmaczny	90
Bardzo niesmaczny	100

Źródło: dane umowne.

3.1.3. Skala przedziałowa

Skala przedziałowa, nazywana też interwałową, pozwala dodatkowo na wyliczanie odległości pomiędzy obiektami mierzonych różnicą wartości cechy. Oznacza to, że możliwe jest dodawanie i odejmowanie wartości cech, gdyż są one wyrażane

w zbiorze liczb rzeczywistych. Pozwala to również na porządkowanie obiektów z użyciem relacji: mniejsze, większe lub równe. Jednak należy pamiętać, że nie można przy użyciu pomiaru w tej skali liczyć ilorazów wartości cech, czyli określać, ile razy jedna wartość jest większa (mniejsza) od drugiej, ponieważ zero na tej skali jest określane umownie przez twórcę skali. Klasycznym przykładem takiej skali jest pomiar temperatury.

Tabela 3.5. Pomiar temperatury powietrza w określonym momencie

Miasto	Temperatura w ^{0}C	Temperatura w ^{0}F
Białystok	25	77
Gdynia	18	64,4
Katowice	32	89,6
Wenecja	38	100,4
Sewilla	40	104

Źródło: dane umowne.

W tabeli 3.5 mamy przykład pomiaru w skali przedziałowej. Zawarte są w niej pomiary temperatury w pięciu miastach. Pomiaru dokonano raz w skali Celsjusza (^{0}C), a drugi raz w skali Fahrenheita (^{0}F). Z prezentowanych danych dotyczących temperatury wyznaczonej w stopniach Celsjusza (^{0}C) wynika, że: pomiędzy Białymstokiem i Katowicami jest różnica temperatur taka sama jak pomiędzy Białymstokiem i Gdynią (7^{0}C), zaś w Sewilli temperatura jest wyższa niż w Gdyni i jest to największa różnica temperatur. Używając skali Fahrenheita[52], możemy nadal stwierdzić to samo co w przypadku skali Celsjusza, czyli: pomiędzy Białymstokiem i Katowicami jest różnica taka sama jak pomiędzy Białymstokiem i Gdynią (12,6^{0}F), zaś w Sewilli temperatura jest wyższa niż w Gdyni i nadal jest to największa różnica temperatur. *Nie można tu jednak stwierdzić, że* w Sewilli jest 1,6 (40/25 = 1,6) razy cieplej niż w Białymstoku. Skala Celsjusza daje inny wynik niż zastosowanie skali Fahrenheita. W ostatnim przypadku iloraz ten wynosi 1,35 (104/77 = 1,35).

Niemożność obliczania ilorazu w przypadku skali interwałowej wynika z braku zera absolutnego na skali przedziałowej. Jest ono na tej skali umowne. Brak zera absolutnego nie oznacza, że cecha o wartości zero nie występuje, na przykład temperatura 0^{0}C oznacza istnienie cechy temperatury o przypisanej wartości zero, przy której ma miejsce pewien proces fizyczny.

[52] Przekształcenie zamieniające temperaturę podaną w stopniach Celsjusza (T_c) na temperaturę w stopniach Fahrenheita (T_f) opisuje następujący wzór: $T_f = 32 + \frac{9}{5} \cdot T_c$. Analogiczne przekształcenie można zapisać, dokonując odwrotnej transformacji: $T_c = \frac{5}{9} \cdot (T_f - 32)$.

Innym przykładem skali przedziałowej jest rok urodzenia. Zero w kalendarzu chrześcijańskim oznacza istniejącą rzeczywiście datę, umownie przyjętą jako punkt odniesienia. Nie możemy powiedzieć, że zero oznacza niewystępowanie cechy, ale jest jednym z jej wariantów. Do innych przykładów pomiaru w skali interwałowej należą: data zajścia zjawiska czy indeksy społeczne (poziom ubóstwa, korupcji).

3.1.4. Skala ilorazowa

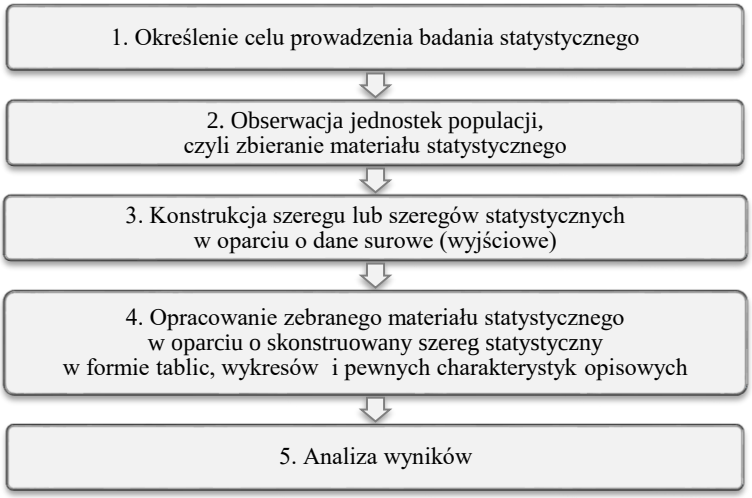
Najmocniejszą skalą jest **skala ilorazowa**, na której oprócz działań już wymienionych można obliczać ilorazy (stosunki) wartości cech. Zatem, oprócz dodawania i odejmowania, możliwe jest też mnożenie i dzielenie. Na tej skali istnieje zero absolutne niezależne od badacza. Zerowa wartość cechy oznacza w tym przypadku jej brak (niewystępowanie zjawiska). Przykładami zastosowania skali ilorazowej są pomiary: płac, dochodów, wydatków i tym podobnych danych ekonomicznych w ujęciu wartościowym, jak również takich cech, jak waga, wzrost, wiek, ceny towarów i usług, czas trwania zjawiska (produkcji, dostawy, wykonania usługi), liczba turystów, liczba hoteli itp. Szereg przedstawiony w tabeli 2.1 (rozkład ocen) jest przykładem cechy mierzonej w skali ilorazowej.

Ogólnie rzecz biorąc, skale pomiarowe, w zależności od ilości możliwych do wykonania działań, dzieli się na **słabe**, do których należą skale nominalna i porządkowa, oraz **skale mocne** (silne), tworzone przez skale przedziałową i porządkową. To, która skala pomiaru zostanie użyta, zależy zarówno od specyfiki cechy, jak też od celu badania. *Cechy, które można mierzyć na skali mocnej, można także mierzyć na skalach słabszych*, np. gdy nie interesują nas dokładne wartości cechy, a jedynie pewne przedziały (np. dochody: niskie, średnie, wysokie), wówczas pomiar jest dokonywany w skali porządkowej, a nie ilorazowej. ***Nigdy natomiast nie możemy dokonywać przejścia odwrotnego, czyli cechy mierzone tylko w skalach niższych traktować tak, jakby były mierzone w skalach wyższych.***

3.2. Materiał statystyczny – zbieranie i wstępne przetwarzanie

Zebranie materiału statystycznego i jego przetwarzanie powinno być realizowane w kilku następujących po sobie krokach. Kroki te ściśle związane są z zadaniami, jakie pełni statystyka. Każde badanie powinno być poprzedzone określeniem jego celu i zakresu. Dopiero wówczas można przeprowadzić obserwację jednostek populacji. Następnie na ich podstawie konstruowany jest szereg lub tworzone są szeregi statystyczne (gdy analizie podlegają przynajmniej dwie cechy), a uzyskany materiał jest opracowywany początkowo w formie tablic czy wykresów. W oparciu

o utworzony szereg obliczane są następnie pewne charakterystyki, których wyniki podlegają analizie. Schematycznie działania te zostały przedstawione na rysunku 3.1.



Rysunek 3.1. Etapy badania statystycznego

Źródło: opracowanie własne.

Czytając opisy poszczególnych etapów, można wysnuć wniosek, że ostatni krok jest najprostszym, bo najkrótszym w opisie elementem. Nic bardziej mylnego. Jest to etap, który wymaga od badacza posiadania ogromnej wiedzy, jak też dogłębnego przemyślenia, zwłaszcza w przypadku, gdy analizowane są nietypowe dane lub badaniu podlega szeroki ich zestaw. W dobie szybko postępującej komputeryzacji wiele działań może być zastąpionych przez odpowiednio dobrane programy, jednak mimo zastosowania nawet najlepszych i najnowocześniejszych z nich, nadal w gestii badacza pozostaje kilka kluczowych elementów, w tym przede wszystkim decyzja o sposobie zbierania danych, wybór obliczanych charakterystyk, a zwłaszcza analiza otrzymanych wyników. Żaden funkcjonujący obecnie program nie jest w stanie samodzielnie tego dokonać.

3.2.1. Rodzaje szeregów statystycznych

Pierwszym etapem badania statystycznego, jak wspomniano powyżej, jest zebranie materiału statystycznego. Po określeniu celu i zakresu badania mierzy się, w przypadku cechy ilościowej, wartość cechy u każdej jednostki statystycznej badanej populacji (np. wzrost każdego studenta PB) lub określa odmianę cechy w przypadku cechy jakościowej (np. kolor oczu każdego studenta PB). Otrzymujemy zbiór

danych określany mianem danych surowych. Surowy materiał statystyczny stanowi nieuporządkowany zbiór liczb i nie da się w oparciu o niego przeprowadzić analizy i wyciągać wnioski co do prawidłowości w rozkładzie cechy. Aby to było możliwe, należy dane surowe usystematyzować, czyli przedstawić w postaci szeregu statystycznego, a więc *de facto* skonstruować rozkład cechy. Sposób konstrukcji szeregu statystycznego zależy w głównej mierze od typu cechy statystycznej.

Podstawowym podziałem szeregów statystycznych jest kryterium czasu. Jeśli pomiar cechy dokonywany jest w kolejnych momentach na jednej i tej samej jednostce, wówczas mówimy o **szeregu czasowym lub inaczej dynamicznym (chronologicznym)**. Przykładem takiego szeregu może być pomiar dochodu budżetu państwa w Polsce w kolejnych latach, który został przedstawiony w tabeli 3.6.

Tabela 3.6. Dochód budżetu państwa w Polsce w latach 2005-2010

Rok	Dochód budżetu państwa [w mln zł]
2005	179 772
2006	197 640
2007	236 368
2008	253 547
2009	274 183
2010	250 303

Źródło: *Rocznik Statystyczny Rzeczypospolitej Polskiej 2011,* Główny Urząd Statystyczny, Warszawa 2011, s. 57.

Jeśli pomiar cechy dokonywany jest w jednym momencie lub okresie na wielu jednostkach statystycznych, wówczas mówimy o **szeregach (danych) przekrojowych**. Przykładem szeregu skonstruowanego dla tego typu danych może być pomiar przeciętnych miesięcznych wydatków przypadających na jedną osobę w Polsce w poszczególnych województwach w 2010 roku, przedstawiony w tabeli 3.7.

Tabela 3.7. Przeciętne miesięczne wydatki na jedną osobę w Polsce w 2010 roku

Województwo	Wydatki na 1 osobę [w zł]
Dolnośląskie	1018,93
Kujawsko-pomorskie	923,63
Lubelskie	825,53
Lubuskie	972,30
Łódzkie	1035,64

Województwo	Wydatki na 1 osobę [w zł]
Małopolskie	949,82
Mazowieckie	1299,85
Opolskie	1021,90
Podkarpackie	820,00
Podlaskie	850,56
Pomorskie	999,05
Śląskie	1005,93
Świętokrzyskie	831,70
Warmińsko-mazurskie	866,15
Wielkopolskie	901,27
Zachodniopomorskie	964,94

Źródło: *Rocznik Statystyczny Województw 2011,*
Główny Urząd Statystyczny, Warszawa 2012, s. 330.

Jeśli badanie łączy pomiar w przestrzeni (przekrojowy) z pomiarem w czasie, wówczas mówimy o **danych przekrojowo-czasowych** lub **panelowych**[53]. Przykładem będzie pomiar wytworzonego produktu narodowego brutto w poszczególnych województwach w latach 1990-2000, który przedstawiono w tabeli 3.8.

Tabela 3.8. Produkt narodowy brutto w poszczególnych województwach w latach 2005-2009

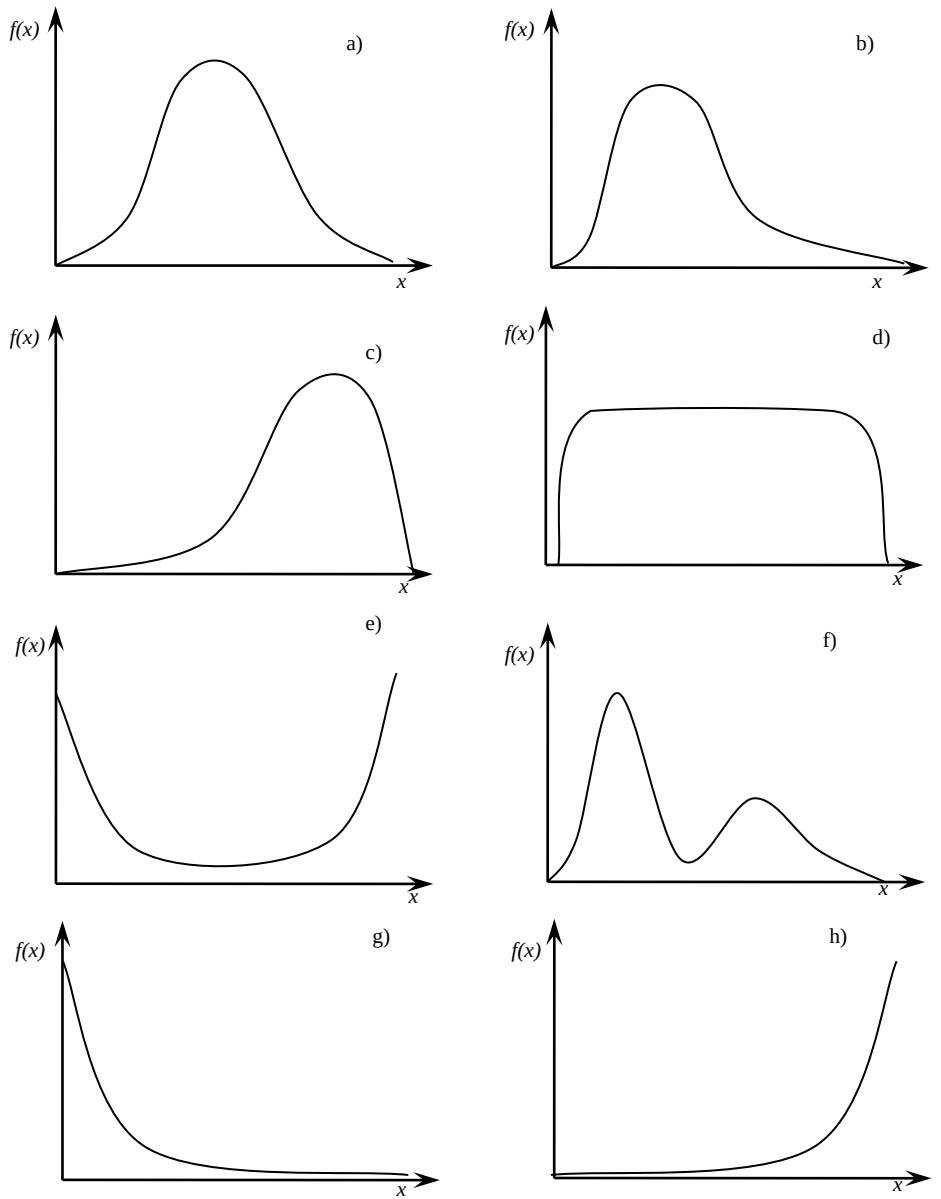
Województwo	Produkt narodowy brutto [w mln zł]				
	2005	2006	2007	2008	2009
Dolnośląskie	76943	85774	96 385	103 412	110 448
Kujawsko-pomorskie	46469	50217	55 414	59 624	61 721
Lubelskie	38388	40849	45 504	50 285	51 082
Lubuskie	23455	24942	27 473	28 893	30 358
Łódzkie	61110	65628	73 117	79 195	81 869
Małopolskie	71748	78789	86 974	95 016	99 509
Mazowieckie	210219	229212	255 560	275 375	293 974
Opolskie	22405	23338	26 478	29 240	29 680

[53] Wielu autorów używa pojęcia „dane" na określenie szeregu, czyli uporządkowanego zbioru danych. W bieżącej literaturze zamiast pojęcia „szereg" autorzy używają pojęcia dane. Autorki wolą używać słowa szereg jako kategorii o węższym znaczeniu, oznaczającej dane uporządkowane w tenże szereg, choć wielu autorów używa tych pojęć zamiennie.

Województwo	Produkt narodowy brutto [w mln zł]				
	2005	2006	2007	2008	2009
Podkarpackie	37319	39894	43 823	48 385	50 684
Podlaskie	22 909	24427	27 321	29 047	30 903
Pomorskie	55 602	60250	66 962	70 341	76 243
Śląskie	130 442	137959	153 066	167 823	175 324
Świętokrzyskie	24 794	27084	30 424	34 051	34 747
Warmińsko-mazurskie	28 153	29977	32 681	35 214	37 076
Wielkopolskie	92 813	98806	109 028	118 478	127 361
Zachodniopomorskie	40 533	42887	46 529	51 052	52 389

Źródło: *Rocznik Statystyczny Województw 2011,* Główny Urząd Statystyczny, Warszawa 2012, s. 90; *Rocznik Statystyczny Województw 2009,* Główny Urząd Statystyczny, Warszawa 2010, s. 88; *Rocznik Statystyczny Województw 2008,* Główny Urząd Statystyczny, Warszawa 2009, s. 84.

Uporządkowany materiał statystyczny podlega dalszej obróbce. Przy wielu jednostkach statystycznych konstruowany jest **szereg rozdzielczy,** inaczej **szereg strukturalny**. Przypominamy, że jest to szereg statystyczny przedstawiony w formie dwukolumnowej (dwuwierszowej) tablicy statystycznej. W pierwszej kolumnie podawane są poszczególne odmiany cechy lub czas pomiaru, w drugiej zaś kolumnie odnotowuje się, ile jednostek w badanej zbiorowości ma daną odmianę cechy. Tak uporządkowany szereg nazywany jest **empirycznym rozkładem cechy** w badanej populacji lub próbie. Przykłady różnych typów rozkładów w schematycznej postaci przedstawiono na rysunku 3.2.



Rysunek 3.2. Przykłady rozkładów: a) symetryczny, b) umiarkowanie asymetryczny (asymetria prawostronna umiarkowana), c) silnie asymetryczny (asymetria silna lewostronna), d) rozkład równomierny, e) rozkład typu U, f) rozkład bimodalny, g) lewostronny rozkład skrajnie asymetryczny (typu L), h) prawostronny rozkład skrajnie asymetryczny (typu J)

Źródło: opracowanie własne.

Poniżej przedstawiamy przykłady prezentujące różne typy rozkładów teoretycznych w oparciu o dane rzeczywiste. Prezentowane szeregi rozdzielcze zostały wzbogacone o wykresy, nazywane histogramami, których konstrukcja szczegółowo zostanie opisana w rozdziale następnym.

Przykład 3.1

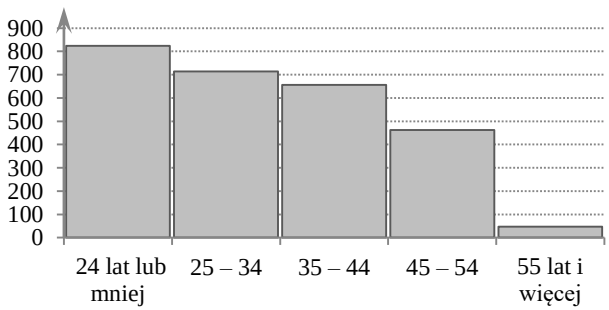
Poniższa tabela przedstawia wiek bezrobotnych zarejestrowanych w 2000 roku w Polsce.

Tabela 3.9. Bezrobotni zarejestrowani według wieku w 2000 roku w Polsce

Wiek (w latach)	Liczba bezrobotnych (w tys.)
24 lata lub mniej	823,5
25-34	713,2
35-44	656,5
45-54	461,9
55 lat i więcej	47,5

Źródło: *Rocznik Statystyczny Rzeczypospolitej Polskiej 2012,*
Główny Urząd Statystyczny, Warszawa 2012;
źródło dostępu:http://www.stat.gov.pl/cps/rde/xbcr/gus/RS_rocznik_statystyczny_rp_2012.pdf, s. 241.

W analizowanym przykładzie badaną populacją są osoby zarejestrowane jako bezrobotne w Polsce w 2000 roku. Jednostką statystyczną jest jeden człowiek, który w 2000 roku był zarejestrowanym bezrobotnym. Cechą, jaka jest brana pod uwagę w niniejszej analizie, jest wiek, zaś jednostką pomiaru cechy lata[54].



Wykres 3.1. Bezrobotni zarejestrowani według wieku w 2000 roku w Polsce

Źródło: opracowanie własne na podstawie; *Rocznik Statystyczny Rzeczypospolitej Polskiej 2012,*
Główny Urząd Statystyczny, Warszawa 2012;
źródło dostępu: http://www.stat.gov.pl/cps/rde/xbcr/gus/RS_rocznik_statystyczny_rp_2012.pdf, s. 241.

[54] Co oznacza, podkreślmy, że pomiar jest prowadzony z dokładnością do jednego roku.

Analizując liczebności badanego szeregu, można zauważyć, że maleją one systematycznie. Dodatkowo można to zaobserwować na wykresie 3.1. Taki kształt histogramu świadczy o skrajnej asymetrii[55].

■

Przykład 3.2

Poniższa tabela prezentuje dane dotyczące ludności w wieku 15 lat i więcej według wieku na podstawie spisu w 2011 roku.

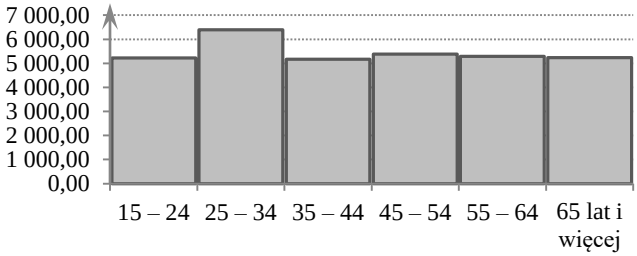
Tabela 3.10. Ludność w wieku 15 lat i więcej według wieku na podstawie spisu w 2011 roku

Wiek (w latach)	Liczba ludności (w tys.)
15-24	5 223,6
25-34	6 393,6
35-44	5 171,7
45-54	5 376,5
55-64	5 284,0
65 lat i więcej	5 230,2

Źródło: *Rocznik Statystyczny Rzeczypospolitej Polskiej 2012,*
Główny Urząd Statystyczny, Warszawa 2012;
źródło dostępu: http://www.stat.gov.pl/cps/rde/xbcr/gus/RS_rocznik_statystyczny_rp_2012.pdf, s. 225.

Analizowaną w tym przykładzie zbiorowością są osoby, które ukończyły co najmniej 15 lat i uczestniczyły w 2011 roku w spisie ludności. Jednostką statystyczną jest każda osoba uczestnicząca w spisie z 2011 roku, która ma 15 lub więcej lat. Analizowaną cechą jest wiek, zaś jednostką pomiaru wieku są lata.

[55] Komentarza wymaga domknięcie przedziałów – do tego wrócimy w kolejnym podrozdziale, gdzie zostaną omówione podstawy konstrukcji szeregu rozdzielczego.



Wykres 3.2. Ludność w wieku 15 lat i więcej według wieku na podstawie spisu w 2011 roku

Źródło: opracowanie własne na podstawie *Rocznik Statystyczny Rzeczypospolitej Polskiej 2012*, Główny Urząd Statystyczny, Warszawa 2012;
źródło dostępu: http://www.stat.gov.pl/cps/rde/xbcr/gus/RS_rocznik_statystyczny_rp_2012.pdf, s. 225.

Zarówno analiza liczebności zawartych w tabeli 3.10, jak też kształtu wykresu 3.2 wskazuje, że jest to rozkład prawie równomierny. Jedynie jeden słupek góruje (co prawda w niewielkim stopniu) nad pozostałymi słupkami. Przekracza on wartość 6 tys. osób, zaś pozostałe są prawie stałe, oscylując wokół liczebności 5 200-5 300 tys. osób.

■

Przykład 3.3

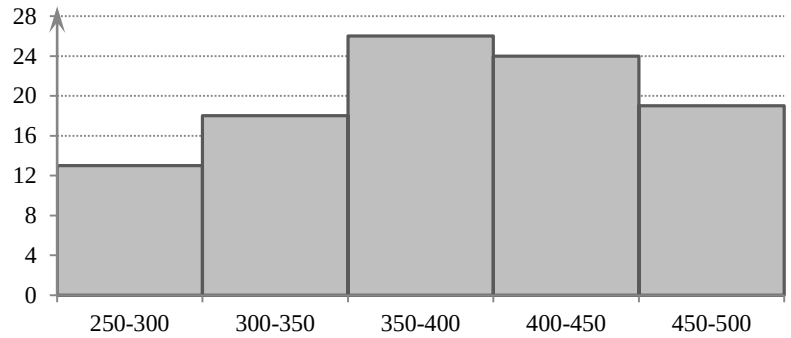
Poniższa tabela 3.11 prezentuje rozkład powierzchni [m^2] działek rekreacyjnych przeznaczonych na sprzedaż na Mazurach w 2000 roku.

Tabela 3.11. Powierzchnia [m^2] działek rekreacyjnych przeznaczonych na sprzedaż na Mazurach w 2000 roku

Powierzchnia	Liczba działek
250-300	13
300-350	18
350-400	26
400-450	24
450-500	19

Źródło: *Rocznik Statystyczny województwa podlaskiego 2002*, Warszawa 2002, s. 153.

W opisanym przykładzie analizowaną zbiorowością są działki przeznaczone na sprzedaż na Mazurach w 2000 roku. Jednostką statystyczną jest każda z działek oferowanych na sprzedaż w 2000 roku, zlokalizowana na Mazurach. Cechą statystyczną jest w tym przypadku powierzchnia działki, a jednostką pomiaru metry kwadratowe.



Wykres 3.3. Powierzchnia [m^2] działek rekreacyjnych przeznaczonych na sprzedaż na Mazurach w 2000 roku

Źródło: opracowanie własne na podstawie *Rocznik Statystyczny województwa podlaskiego 2002,* Warszawa 2002, s. 153.

Przedstawiony wykres 3.3 reprezentuje rozkład o umiarkowanej asymetrii lewostronnej (analizując kształt, można zauważyć, że obie strony wykresu nie są symetryczne względem jego środka – po lewej stronie wysokości słupków reprezentujących liczebności poszczególnych klas zaczynają się od niższej wartości i rosną szybciej niż po prawej stronie).

■

Przykład 3.4

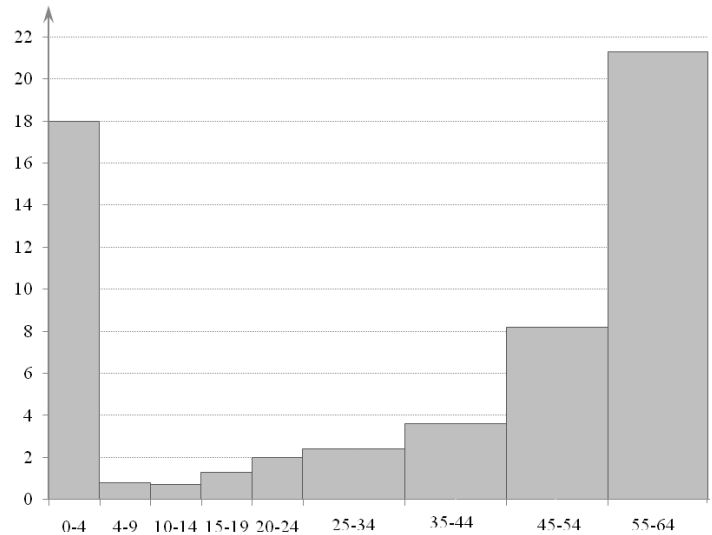
Tabela przedstawia rozkład zgonów mężczyzn według wieku w Polsce w 1959 roku.

Tabela 3.12. Rozkład zgonów mężczyzn według wieku w Polsce w 1959 roku

Wiek	Liczba mężczyzn (w tys.)
0-4	18,0
5-9	0,8
10-14	0,7
15-19	1,3
20-24	2,0
25-34	2,4
35-44	3,6
45-54	8,2
55-64	21,3

Źródło: *Rocznik statystyczny 1961*, GUS, Warszawa 1961, s. 36, tabl. 26 (46).

Analizowaną w podanym przykładzie zbiorowością są mężczyźni, którzy zmarli w 1959 roku w Polsce. Jednostką statystyczną jest każdy mężczyzna, który umarł w Polsce w 1959 roku. Cechą braną pod uwagę w podanym zestawieniu jest wiek tych osób, natomiast jednostką pomiaru wieku – lata.



Wykres 3.4. Rozkład zgonów mężczyzn według wieku w Polsce w 1959 roku

Źródło: opracowanie własne na podstawie: *Rocznik statystyczny 1961*, GUS, Warszawa 1961, s. 36, tabl. 26(46).

Podany powyżej wykres 3.1 prezentuje przykład rozkładu określanego jako rozkład U (rysunek 3.2e). Na wykresie można zaobserwować, że dwa skrajne przedziały mają największe liczebności, czyli najwięcej mężczyzn (czyli chłopców) umarło w 1959 roku w wieku od 0 do 4 lat i pomiędzy 55 a 64 rokiem życia. ■

Przykład 3.5

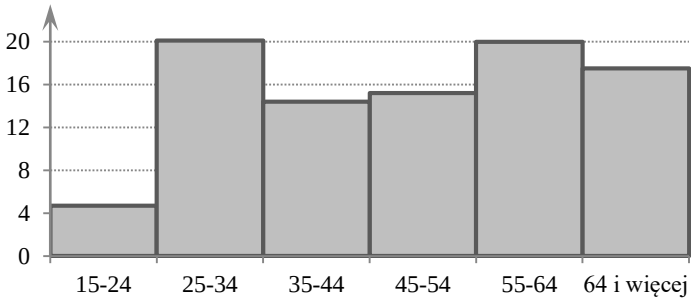
Struktura pracujących z wykształceniem wyższym według wieku w IV kwartale 2001 roku na 100 osób.

Tabela 3.13. Rozkład pracujących z wykształceniem wyższym według wieku w IV kwartale 2001 roku na 100 osób

Wiek pracujących z wykształceniem wyższym	Liczba osób pracujących z wykształceniem wyższym (na 100 osób)
15-24	4,7
25-34	20,1
35-44	14,4
45-54	15,2
55-64	20,0
64 i więcej	17,5

Źródło: *Mały Rocznik Statystyczny Polski 2002*, tabl. 5 (90), s. 142.

W przytoczonym przykładzie zbiorowością są osoby pracujące w IV kwartale 2001 roku podawane w przeliczeniu na 100 osób. Jednostką jest jeden zatrudniony w IV kwartale 2001 roku w przeliczeniu na 100 osób. Cechą statystyczną jest wiek pracujących, zaś jednostką pomiaru – lata.



Wykres 3.5. Pracujący z wykształceniem wyższym według wieku w IV kwartale 2001 roku na tys. osób

Źródło: opracowanie własne.

Wykres 3.5 przedstawia przykład rozkładu określanego jako bimodalny, czyli posiadający dwa szczyty (więcej informacji o tym znajdzie się w rozdziale o miarach przeciętnych).

■

Przykład 3.6

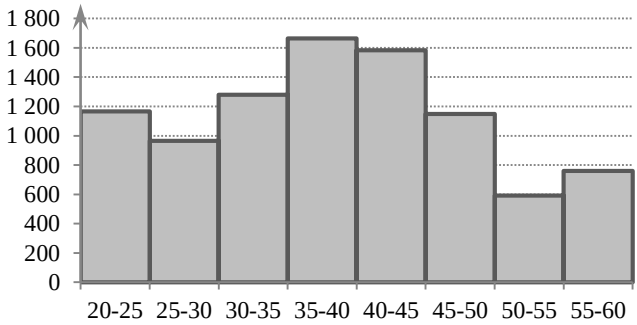
Tabela 3.14 prezentuje wiek kobiet skazanych prawomocnie przez sądy powszechne za przestępstwa ścigane z oskarżenia publicznego w 1992 roku.

Tabela 3.14. Rozkład wieku kobiet skazanych prawomocnie przez sądy powszechne za przestępstwa ścigane z oskarżenia publicznego w 1992 roku

Wiek skazanych kobiet	Liczba kobiet
20-25	1 166
25-30	965
30-35	1 280
35-40	1 664
40-45	1 584
45-50	1 149
50-55	591
55-60	760

Źródło: *Rocznik statystyczny GUS 1993*, GUS, tabl. 21 -(147), s. 94.

Przytoczone dane dotyczą zbiorowości kobiet skazanych w 1992 roku prawomocnie przez sądy powszechne za przestępstwa ścigane z oskarżenia publicznego. Jednostką statystyczną jest każda kobieta skazana prawomocnie w 1992 roku przez sąd powszechny za przestępstwo ścigane z oskarżenia publicznego. Badaną cechą jest wiek tych kobiet, zaś jednostką pomiaru są lata.



Wykres 3.6. Wiek kobiet skazanych prawomocnie przez sądy powszechne za przestępstwa ścigane z oskarżenia publicznego w 1992 roku

Źródło: opracowanie własne.

Powyższy przykład nietypowego rozkładu prezentuje rozkład, który uznaje się za złożenie dwóch rozkładów typu U.

3.2.2. Konstrukcja szeregu rozdzielczego

Celem otrzymania empirycznego rozkładu cechy należy dokonać klasyfikacji (grupowania) jednostek statystycznych w populacji według wartości cechy. W przypadku cech skokowych o niewielkiej liczbie możliwych odmian jednostki klasyfikuje się według wartości cechy, to znaczy, że każda odmiana cechy stanowi odrębną grupę jednostek statystycznych zwaną **klasą**. Przykładem takich cech mogą być: liczba dzieci w rodzinie, liczba izb w mieszkaniu, liczba osób w gospodarstwie domowym, ilość sprzętu AGD w użytkowaniu gospodarstwa domowego. Tworzenie szeregu statystycznego będzie w takim przypadku polegało na określeniu klas badanej cechy oraz podaniu, ile jednostek statystycznych należy do danej klasy. Jest to równorzędne podaniu empirycznego rozkładu cechy.

Przykład 3.7

Na pewnej niewielkiej uczelni postanowiono zbadać, z jak licznych rodzin pochodzą ich studenci, przeprowadzono zatem badanie polegające na zapytaniu każdego studenta, ile ma rodzeństwa. Po dodaniu do liczby podanej przez studenta jedności otrzymano, ile dzieci jest w danej rodzinie. Okazało się, że największa liczba rodzeństwa wynosiła cztery, co oznacza, że rozpiętość wartości badanej cechy (liczba dzieci w rodzinie) wynosi od jeden do pięciu. Po zliczeniu, ile osób znajduje się w poszczególnych klasach cechy, otrzymane dane przedstawiono w formie tabeli 3.15 prezentującej w pierwszych dwóch kolumnach szereg statystyczny rozkładu liczby dzieci w rodzinach studentów tej uczelni.

Tabela 3.15. Rozkład liczby dzieci w rodzinie studenta uczelni X

Wartość cechy x_i	Liczebność n_i	Częstość $w_i = \frac{n_i}{N}$	Częstość [w %]	Kumulacja liczebności $n_{k\,sk} = \sum_{i=1}^{k-1} n_i$	Kumulacja częstości $w_{k\,sk} = \sum_{i=1}^{k-1} w_i$
1	892	0,262	26,2	0	0,000
2	1 157	0,340	34,0	892	0,262
3	798	0,234	23,4	2 049	0,602
4	435	0,128	12,8	2 847	0,836
5	121	0,036	3,6	3 282	0,964
Razem	3 403	1,000	100,0	3 403	1,000

Źródło: dane umowne.

Kolumny pierwsza i druga powyższej tabeli prezentują rozkład cechy liczba dzieci w rodzinie studenta. Kolumny trzecia i czwarta to alternatywne możliwości przedstawienia liczebności poszczególnych odmian cechy w postaci częstości w zapisie dziesiętnym (kolumna trzecia) lub procentowym (kolumna czwarta). Przedostatnia i ostatnia kolumna (piąta i szósta) przedstawiają odpowiednio skumulowane liczebności i częstości, które interpretuje się jako liczbę jednostek statystycznych posiadających cechę do danej wartości x_i.

Liczba zero w pierwszym wierszu w piątej i szóstej kolumnie tabeli 3.15 oznacza, że nie ma studentów pochodzących z rodzin o liczbie dzieci mniejszej niż jeden (mniej niż jeden w podanym przypadku może oznaczać jedynie zerową liczbę dzieci – jest to sytuacja oznaczająca, że nie ma w danej rodzinie dzieci, co oznacza niemożliwe do zajścia w naszym przekładzie zdarzenie, bo badany student też jest dzieckiem swoich rodziców). Dla wartości $x_1 = 2$ liczebność skumulowana oznacza liczbę studentów o liczbie dzieci w rodzinie poniżej dwóch (a więc liczbę rodzin posiadających dokładnie jedno dziecko). W tym wierszu wielkość 892 oznacza liczbę studentów pochodzących z rodzin o liczbie dzieci do dwóch, ale bez dwóch. Następna wartość skumulowana 2049 oznacza liczbę studentów z rodzin o liczbie dzieci poniżej trzech itd.

W tabeli 3.15 podano szereg wyjściowy cechy oraz szereg skumulowany, gdyż oba najczęściej się podaje razem. Formalnie, zgodnie z jedną z dwóch możliwych definicji, **szereg skumulowany,** zwany też **dystrybuantą empiryczną,** można przedstawić tak jak w tabeli 3.16a lub częściej, gdy nie prowadzi to do nieporozumień związanych z interpretacją wyników, jak w tabeli 3.16b. Wyjaśniając różnice pomiędzy obiema tabelami, należy zaznaczyć, że nawias otwarty „)" kończący zapis w tabeli 3.16a oznacza, że kraniec przedziału nie jest włączony do przedziału, zaś nawias domknięty „]" oznacza, że obiekty o wartościach równych krańcowi przedziału są do niego zaliczane. Warto również podkreślić różnice w sposobie konstrukcji liczebności skumulowanej. I tak w przypadku postaci szeregu zamieszczonego w tabeli 3.16a sumowane są w kolumnie drugiej liczebności poprzedzające dany wiersz, zaś w wersji 3.16b brana jest pod uwagę, oprócz liczebności poprzedzających, również liczebność w bieżącym wierszu.

Tabela 3.16. Rozkład liczby dzieci w rodzinie studenta uczelni X – szeregi skumulowane

a.

Wartość cechy do bez x_i $[-\infty, x_i]$	Kumulacja liczebności $n_{k\,sk} = \sum_{i=1}^{k-1} n_i$
do 1	0
do 2	892
do 3	2 049
do 4	2 847
do 5	3 282
Razem lub do 6	3 403

b.

Wartość cechy do x_i łącznie z x_i $[-\infty, x_i]$	Kumulacja liczebności $n_{k\,sk} = \sum_{i=1}^{k} n_i$
do 1	892
do 2	2 049
do 3	2 847
do 4	3 282
do 5	3 403
Razem	

Źródło: dane umowne.

■

W przypadku cech skokowych o znacznej liczbie możliwych odmian oraz w przypadku cech ciągłych o teoretycznie nieskończonej liczbie odmian jednostki klasyfikujemy w pewne przedziały wartości cechy zwane **przedziałami klasowymi**. Wszystkie możliwe wartości cechy dzieli się wówczas na przedziały klasowe, przy czym **podział ten (1) musi być ciągły,** tzn. nie można w nim opuścić żadnej odmiany cechy. Następnie wszystkie jednostki populacji są zaliczane do któregoś z przedziałów klasowych. **Klasyfikacja jednostek musi być (2) rozłączna i (3) zupełna.** Rozłączność w tym przypadku oznacza, że jedna jednostka należy tylko i wyłącznie do jednego przedziału klasowego, zupełność zaś, że żadna jednostka nie zostanie pominięta w klasyfikacji (4). **Żaden też przedział klasowy nie może być pusty, czyli mieć zerową liczebność**.

Szeregi rozdzielcze z przedziałami klasowymi nazywa się **szeregami wielostopniowymi**. Konstrukcja takiego szeregu nie jest sprawą trywialną i o ile można znaleźć w literaturze wskazówki, jak konstruować szereg rozdzielczy, to nie są one jednoznacznie obowiązujące z powodu różnorodności obszarów badawczych oraz wiążących się z tym danych statystycznych. Nie ma „praw" czy „twierdzeń" określających sposób konstrukcji szeregu. Jest to zagadnienie, w którym decydujący w dużej mierze jest cel badania oraz doświadczenie i subiektywny osąd badacza konstruującego szereg.

Przykład 3.8

Zapytano uczniów jednej z białostockich szkół o ocenę, jaką otrzymali z ostatniego sprawdzianu z matematyki. Wyniki były następujące: 5; 4; 5; 3; 2; 3; 5; 5; 5; 6; 1; 2; 4; 4; 1; 1; 2; 5; 3; 3; 4; 3; 1; 2; 6; 4; 4; 3; 4; 3; 3; 3; 4; 1; 5; 3; 2; 2; 2; 3; 1; 4; 1; 6; 4; 5; 4; 3; 3; 2.

Chcąc dokonać analizy otrzymanych w ten sposób danych w pierwszym kroku należy określić zbiorowość, jednostkę statystyczną, jednostkę pomiaru i cechę oraz dokonać jej klasyfikacji.

Zatem badaną populację stanowią uczniowie jednej z białostockich szkół piszący określonego dnia sprawdzian z matematyki, zaś jednostką statystyczną jest każdy z piszących sprawdzian uczniów. Cechą, która zróżnicowała zbiorowość uczniów, jest ocena uzyskana ze sprawdzianu z matematyki wyrażona w postaci punktów skali ocen. Jest to cecha skokowa przyjmująca wartości ze skali ilorazowej.

Po dokonaniu powyższej klasyfikacji należy przystąpić do uporządkowania otrzymanych wartości w porządku rosnącym lub malejącym. W tym wypadku został zastosowany pierwszy z nich, co doprowadziło do otrzymania następującego zbioru danych: 1, 1, 1, 1, 1, 1, 1, 2, 2, 2, 2, 2, 2, 2, 2, 3, 3, 3, 3, 3, 3, 3, 3, 3, 3, 3, 3, 3, 4, 4, 4, 4, 4, 4, 4, 4, 4, 4, 4, 5, 5, 5, 5, 5, 5, 5, 5, 6, 6, 6.

Dzięki uporządkowaniu łatwiej jest skonstruować szereg rozdzielczy cechy. Widać, że w analizowanym przykładzie występuje sześć wariantów cechy: wartości oceny od 1 do 6 włącznie. Dlatego utworzona tabela będzie miała sześć wierszy numerowanych kolejnymi liczbami naturalnymi $i \in \{1, 2, 3, 4, 5, 6\}$. Wskazany numer wiersza z szeregu rozdzielczego cechy dyskretnej zawiera przynajmniej dwa podstawowe elementy: x_i (i-ta odmiana cechy i jej wartość) oraz n_i (liczebność i-tej odmiany cechy). Liczebność całej populacji oznaczana jest na ogół symbolem n lub $\boldsymbol{N}$. Często zamiast liczebności lub jako uzupełnienie informacji podaje się częstość występowania cechy (nazywaną także frakcją lub procentem), a oznaczana jako f_i lub w_i. częstość występowania cechy jest istotna, gdy porównujemy rozkłady o różnych ilościach jednostek statystycznych.

Tabela 3.17. Rozkład ocen ze sprawdzianu z matematyki

Ocena ze sprawdzianu z matematyki x_i	Liczba uczniów n_i	Częstość $w_i = \frac{n_i}{N}$
1	7	0,14
2	8	0,16
3	13	0,26
4	11	0,22
5	8	0,16
6	3	0,06
Razem	50	1,00

Źródło: dane umowne.

W omawianym przykładzie otrzymano szereg rozdzielczy cechy dyskretnej przedstawiony w tabeli 3.17. Widać, że najliczniejszą grupę stanowili uczniowie, którzy otrzymali ocenę 3, zaś najmniej było tych, którzy sprawdzian napisali na najwyższą notę.

■

Przykład 3.9

Dane dotyczą wagi 100 studentek kierunku zarządzanie (dane umowne):
52; 46; 50; 53; 59; 58; 54; 55; 53; 52; 52; 53; 49; 61; 59; 57; 49; 63; 57; 56; 54; 51; 61; 64; 56; 49; 59; 57; 63; 49; 48; 52; 67; 57; 56; 61; 51; 47; 50; 58; 50; 52; 52; 57; 54; 57; 55; 54; 64; 69; 53; 47; 61; 54; 61; 59; 58; 51; 48; 53; 55; 56; 58; 65; 55; 49; 54; 44; 53; 45; 57; 58; 53; 55; 54; 60; 49; 48; 57; 60; 52; 47; 57; 58; 60; 55; 46; 48; 54; 53; 57; 53; 55; 59; 48; 53; 53; 51; 52; 57.

Zbiorowość statystyczną stanowią tu wszystkie (w liczbie sto) studentki kierunku zarządzanie. Jednostkami statystycznymi są pojedyncze studentki, zaś mierzoną cechą statystyczną waga ich ciała. Jest to cecha mierzalna (ilościowa), ciągła. Jednostką pomiaru są kilogramy.

Biorąc pod uwagę fakt, że badaniu poddano 100 kobiet ze względu na cechę ciągłą, należy utworzyć szereg rozdzielczy przedziałowy.

Po ustawieniu danych w porządku rosnącym otrzymujemy:
44; 45; 46; 46; 47; 47; 47; 48; 48; 48; 48; 48; 49; 49; 49; 49; 49; 49; 50; 50; 50; 51; 51; 51; 51; 52; 52;52;52; 52; 52; 52; 52; 53; 53; 53; 53; 53; 53; 53; 53; 53; 53; 53; 54; 54; 54; 54; 54; 54; 54; 54; 55; 55; 55; 55; 55; 55; 56; 56; 56; 56; 57; 57; 57; 57; 57; 57; 57; 57; 57; 57; 57; 58; 58; 58; 58; 58; 58; 59; 59; 59; 59; 59; 60; 60; 60; 61; 61; 61; 61; 61; 63; 63; 64; 64; 65; 67; 69.

Warto tu zauważyć, o czym była mowa przy definiowaniu rodzajów cech, że waga jest cechą ciągłą, ale ze względu na ograniczenia wynikające z zastosowanego pomiaru z dokładnością do pełnych kilogramów[56] utworzono zestaw wartości dyskretnych. Ich różnorodność sprawia, że nie można ich traktować podczas grupowania w ten sam sposób jak w poprzednim przykładzie 3.2. W tym przypadku należy zbudować szereg rozdzielczy przedziałowy. Pierwszym krokiem przy bu-

[56] Intuicyjnie traktujemy wagę jako cechę ciągłą, albowiem teoretycznie nie ma żadnych ograniczeń na jej wartość (poza zerem). Waga przybiera wartości ze zbioru liczb rzeczywistych, jedynym ograniczeniem jest dokładność, z jaką jesteśmy w stanie ją zmierzyć.

dowie takiego szeregu jest określenie liczby klas[57], jakie wyodrębnimy w zbiorze. Można skorzystać z jednego z poniższych wzorów rekomendowanych w literaturze:

$$k = \sqrt{n}, \tag{3.1}$$

$$k = 1 + 3{,}322 \log n. \tag{3.2}$$

Przedstawione wzory (3.1) i (3.2) należy traktować jedynie jako podpowiedzi dotyczące liczby tworzonych klas. Do określenia liczby klas w analizowanym przykładzie wybrano wzór (3.2) i otrzymano za jego pomocą: $k = 1 + 3{,}322 \log 100 = 7{,}644 \approx 7$ klas.

Kolejnym krokiem jest ustalenie rozpiętości przedziałów, do czego wykorzystano wzór:

$$h = \frac{x_{\max} - x_{\min}}{k}, \tag{3.3}$$

gdzie:
x_{max} – wartość największa badanej cechy,
x_{min} – wartość najmniejsza badanej cechy.

Podstawiając dane, otrzymano:

$$h = \frac{69 - 44}{7} = 3{,}571 \approx 4.$$

Granice pierwszego przedziału utworzono w następujący sposób:

$$x_{1_0} = x_{\min} - \tfrac{1}{2} h,$$
$$x_{1_1} = x_{1_0} + h,$$

gdzie x_{1_0} oznacza początek pierwszego przedziału, zaś x_{1_1} jego koniec.

Kolejne przedziały (dla i = 2, 3, ..., k) wyznacza się następująco:

$$x_{i_0} = x_{(i-1)_1},$$
$$x_{i_1} = x_{i_0} + h,$$

przy czym x_{i_0} oznacza początek i-tego przedziału, zaś x_{i_1} jego koniec.

[57] Podajemy tu ogólny schemat konstrukcji, by dać czytelnikowi pewne pojęcie o procesie konstrukcji szeregu. Poprawna konstrukcja szeregu nie jest sprawą prostą i wymaga dużej wiedzy i doświadczenia. Informacje na ten temat można znaleźć w Bąk i in.[2001].

Zatem dla analizowanych danych otrzymano:

$$x_{1_0} = 44 - \tfrac{1}{2} \cdot 4 = 42,$$
$$x_{1_1} = 42 + 4 = 46,$$
$$x_{2_0} = x_{1_1} = 46,$$
$$x_{2_1} = 46 + 4 = 50,$$
$$x_{7_0} = x_{7_1} = 66,$$
$$x_{7_1} = 66 + 4 = 70.$$

Dla tak wyznaczonych granic przedziałów[58] pogrupowano dane, a wyniki przedstawiono w tabeli 3.18 (dwie pierwsze kolumny). Należy w tym miejscu przeprowadzić sprawdzenie poprawności przeprowadzonego grupowania. Konieczność przeprowadzenia takiego postępowania jest spowodowana tym, że wyboru zarówno liczby klas, jak i rozpiętości przedziałów dokonuje się, bazując w znacznej mierze na intuicji badacza i jego doświadczeniu.

Grupowanie jest przeprowadzone poprawnie, gdy różnica pomiędzy środkami poszczególnych przedziałów a średnimi arytmetycznymi wartości cech zakwalifikowanych do przedziałów nie przekracza połowy rozpiętości przedziału (dotyczy to wyłącznie szeregu o równej rozpiętości przedziałów)[59]. Wyliczenia z tym związane zostały przedstawione w tabeli 3.18.

Tabela 3.18. Wyliczenia niezbędne do utworzenia szeregu i sprawdzenia poprawności grupowania danych dotyczących wagi studentek

Przedziały $[x_{i_0} - x_{i_1}]$	Pomiary należące do poszczególnych przedziałów x_i	Liczebność n_i	Suma z przedziałów $\sum x_i$	Średnia przedziału $\bar{x}_i$	Środek przedziału $\dot{x}_i$	$\lvert\bar{x}_i - \dot{x}_i\rvert$
42-46	44; 45;	2	89	44,50	44	0,50
46-50	46; 46; 47; 47; 47; 48; 48; 48; 48; 48; 49; 49; 49; 49; 49;	16	767	47,94	48	0,06
50-54	50; 50; 50; 51; 51; 51; 51; 52; 52; 52; 52; 52; 52; 52; 52; 53; 53; 53; 53; 53; 53; 53; 53; 53; 53;	26	1 352	52,04	52	0,04
54-58	54; 54; 54; 54; 54; 54; 54; 54; 55; 55; 55; 55; 55; 55; 55; 56; 56; 56; 56; 57; 57; 57; 57; 57; 57; 57; 57; 57; 57.	30	1 668	55,60	56	0,40

[58] Opis sposobu domknięcia granic przedziału zostanie omówiony dalej.
[59] Bąk i in.[2001].

Przedziały $[x_{i_0} - x_{i_1}]$	Pomiary należące do poszczególnych przedziałów x_i	Liczebność n_i	Suma z przedziałów $\sum x_i$	Średnia przedziału $\bar{x}_i$	Środek przedziału $\dot{x}_i$	$\lvert\bar{x}_i - \dot{x}_i\rvert$
58-62	58; 58; 58; 58; 58; 58; 59; 59; 59; 59; 60; 60; 60; 61; 61; 61; 61; 61;	19	1 128	59,37	60	0,64
62-66	63; 63; 64; 64; 65;	5	319	63,80	64	0,20
66-70	67; 69;	2	136	68	68	0
Suma						1,84

Źródło: opracowanie własne na podstawie danych umownych.

Po zsumowaniu różnic występujących pomiędzy sumą poszczególnych wartości zakwalifikowanych do utworzonych poprzednio klas a ich środkiem otrzymano liczbę 1,84. Wynik ten jest niższy od połowy rozpiętości przedziału, czyli od wartości 2 ($0{,}5 \cdot h = 0{,}5 \cdot 4 = 2$). Zatem przeprowadzone grupowanie należy uznać za poprawne.

Tabela 3.19. Szereg rozdzielczy (rozkład empiryczny i dystrybuanta empiryczna) wagi studentek

$[x_{i_0} - x_{i_1}]$	n_i	f_i	n_{isk}	f_{isk}
42-46	2	0,02	2	0,02
46-50	16	0,16	18	0,18
50-54	26	0,26	44	0,44
54-58	30	0,30	74	0,74
58-62	19	0,19	93	0,93
62-66	5	0,05	98	0,98
66-70	2	0,02	100	1,00
Suma	100	1,00		

Źródło: opracowanie własne na podstawie danych umownych.

Pełną formę szeregu rozdzielczego przedziałowego utworzonego dla wagi analizowanej grupy kobiet przedstawiono w tabeli 3.19. Zawarta tam została również część dotycząca konstrukcji zarówno rozkładu empirycznego, jak też dystrybuanty. Zastosowano w tym przypadku jedną z kilku dostępnych dla badacza form grupowania danych związanych z domknięciem przedziałów. Pierwszą z opcji, zastosowaną powyżej, jest forma domknięcia przedziału od dołu, a pozostawienia otwartego od góry.

Tabela 3.20. Przykłady szeregu rozdzielczego wagi studentek przy różnym domknięciu przedziałów

a.

$[x_{i_0} - x_{i_1}]$	n_i
42-46	2
46-50	16
50-54	26
54-58	30
58-62	19
62-66	5
66-70	2
Suma	100

b.

$[x_{i_0} - x_{i_1}]$	n_i
42-45	2
46-49	16
50-53	26
54-57	30
58-61	19
62-65	5
66-69	2
Suma	100

Źródło: dane umowne.

Innym wyjściem jest pozostawienie otwartego dolnego krańca przedziału, a zamknięcie górnego. Wówczas szereg przyjmowałby postać przedstawioną w tabeli 3.20a. Można tak postąpić, gdyż dolny kraniec pierwszego przedziału był przesunięty o połowę rozpiętości.

Można również przedstawić szereg w postaci przedziałów domkniętych obustronnie – tabela 3.20b. Taką formę szeregu rozdzielczego stosuje się jedynie w odniesieniu do cech, które przyjmują całkowite wartości lub podawane są ze ściśle określoną dokładnością (np. z dokładnością do miejsc dziesiętnych). W tym przypadku należy pamiętać, że uprzednio wyznaczona rozpiętość h=4 oznacza liczbę jednostek pomiaru cechy mieszczących się w danym przedziale. Dlatego należy do określenia krańców (poza pierwszym dolnym, który wyznacza się tak, jak podano powyżej) zastosować następujący schemat:

$$x_{i_0} = x_{(i-1)_0} + h = x_{(i-1)_1} + 1,$$
$$x_{i_1} = x_{i_0} + h - 1.$$

Należy zaznaczyć, że w sytuacji gdy stykamy się ze zbudowanym szeregiem rozdzielczym, nie przeprowadzając samodzielnego grupowania, rozpiętość, którą wykorzystuje się, konstruując niektóre z podanych w dalszej części podręcznika miar statystycznych, powinna być wyznaczona następująco:

$$h = x_{i_0} - x_{(i-1)_0}.$$

Powyższy sposób wyznaczania rozpiętości można zastosować w przypadku każdego z powyżej zaprezentowanych sposobów domknięcia przedziałów w szere-

gu rozdzielczym. Jednak w odniesieniu do dwóch poprzednich (tabela 3.19 i 3.20a) można skorzystać również ze wzoru:

$$h = x_{i_1} - x_{i_0}.$$

Niezależnie od przyjętego sposobu grupowania każdorazowo dla danych szeregów rozpiętość w badanym przykładzie wynosi 4 kg.

■

3.3. Graficzna prezentacja danych statystycznych

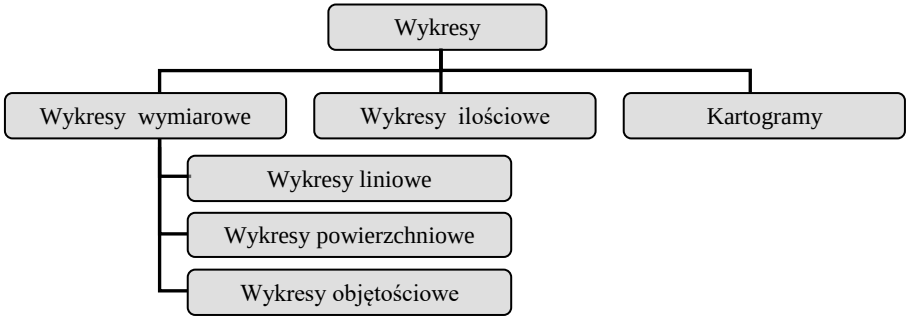
Niezmiernie ważną rolę w analizie statystycznej odgrywa graficzna prezentacja danych statystycznych. Dzięki wykresom można przeprowadzić nie tylko wstępną analizę zjawiska czy zbiorowości, ale także dokonać doboru odpowiednich parametrów opisu rozkładu cechy. Jest to przydatne zwłaszcza wtedy, gdy operujemy dużą liczbą danych.

Analizę wykresu można porównać do oglądania zdjęcia. Gdy zobaczymy zdjęcie, znacznie lepiej jesteśmy w stanie wyobrazić sobie daną osobę (jak ona wygląda), niż słuchając jedynie jej opisu. Poza tym wyobrażenie wyglądu powstałe w wyniku oglądania zdjęcia zazwyczaj znacznie bliższe jest prawdziwemu wizerunkowi. Podobnie jest z wykresami – pozwalają one „zobaczyć", jak wygląda rozkład danej cechy w zbiorowości (szeregi przekrojowe), czy też jak dane zjawisko zachowuje się w czasie (szeregi czasowe).

Wybór wykresu powinien być dostosowany do danych, jakie mają na nim zostać przedstawione. Graficzną prezentację danych można wykonywać zarówno dla cech ilościowych, jak i jakościowych. Jednak dla tak różnych typów danych dobiera się różne wykresy i inaczej się je interpretuje.

Istnieje wiele różnych typów wykresów. W literaturze przedmiotu można spotkać podział na wykresy wymiarowe, wykresy ilościowe i kartogramy[60]. Podział ten został przedstawiony schematycznie na rysunku 3.3.

[60] B. Szulc [1972, s. 47-57].



Rysunek 3.3. Rodzaje wykresów

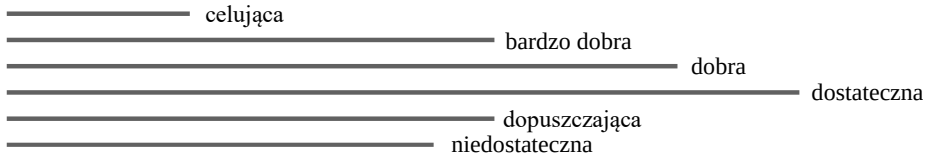
Źródło: opracowanie własne na podstawie B. Szulc [1972, s. 47].

Wykresy wymiarowe tworzone są przeważnie z różnego rodzaju figur geometrycznych proporcjonalnych do przedstawianych zjawisk. Do najczęściej wykorzystywanych należy zaliczyć odcinki (wykresy liniowe), kwadraty, prostokąty, koła i inne figury dwuwymiarowe lub obrazy symbolizujące badane zjawisko (wykresy powierzchniowe) oraz figury przestrzenne, takie jak prostopadłościany, walce, ostrosłupy (wykresy objętościowe). Tworzą tego typu wykresy, zwłaszcza w przypadku powierzchniowych i objętościowych, by określić, czy będzie zmianie ulegał jedynie jeden, dwa czy trzy wymiary prezentowanej figury.

W sytuacji gdy badacz decyduje się na wykres powierzchniowy, powinien pamiętać, że jeden z wymiarów zmieniany jest proporcjonalnie do wielkości zjawiska, zaś oba wymiary proporcjonalnie do pierwiastków kwadratowych tych wielkości. Pierwszy ze sposobów najczęściej prezentowany jest w postaci wykresu składającego się z prostokątów o jednakowych podstawach i zmieniających się wysokościach lub z prostokątów o jednakowych wysokościach i zmieniających się podstawach. W ten sposób ta forma wykresu przypomina wykres liniowy.

Przykład 3.4 (kontynuacja 3.2)

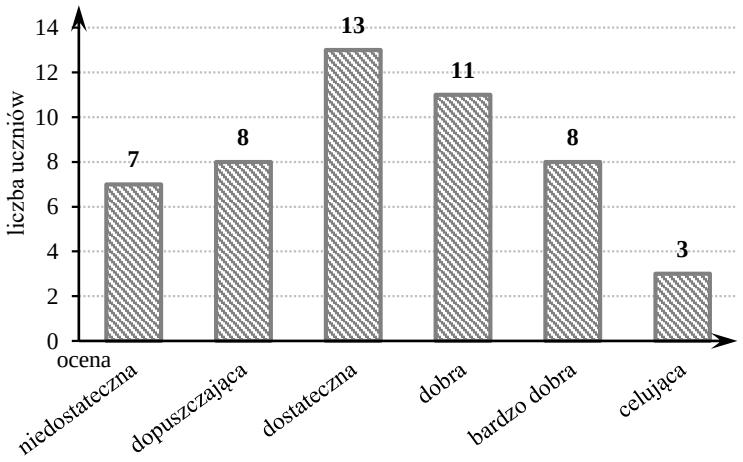
Biorąc pod uwagę dane dotyczące wyników sprawdzianu z matematyki (przykład 3.2), można wykonać różne wykresy wymiarowe (wykresy 3.7-3.9) obrazujące otrzymane wyniki. W celu przejrzystości zapis liczbowy wartości ocen zastąpiono ich opisem słownym.



Wykres 3.7. Oceny ze sprawdzianu z matematyki (wykres liniowy)

Źródło: opracowanie własne.

Wykres 3.7 jest wykresem liniowym, w którym liczba ocen uzyskanych przez uczniów jest obrazowana przez długość odcinka. Łatwo można z niego odczytać, iż najwięcej było ocen dostatecznych, zaś najmniej celujących.

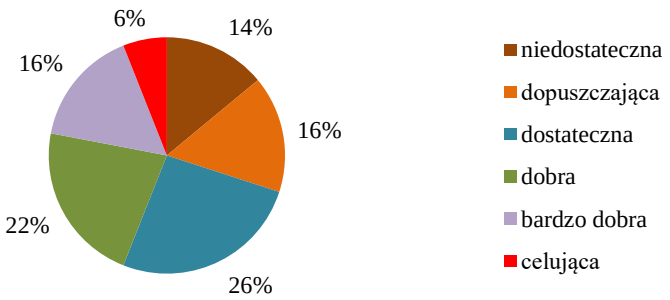


Wykres 3.8. Oceny ze sprawdzianu z matematyki
(wykres powierzchniowy – prostokąty o równych podstawach)

Źródło: opracowanie własne.

Kolejny wykres (3.8) prezentuje przykład wykresu powierzchniowego – zmianie ulega jedynie wysokość prostokątów, ich podstawy pozostają stałe. Tak przyjęta forma sprawia, że wykres ten jest zbliżony pod kątem jego interpretacji do wykresów liniowych. Wykres powierzchniowy, w którym zmianie ulega jedynie jeden z wymiarów, to jedna z najczęściej wykorzystywanych form graficznej prezentacji danych. Powodem jego popularności jest fakt, iż oko ludzkie najlepiej radzi sobie z porównywaniem figur pod kątem tylko jednego wymiaru. Znacznie trudniej przychodzi nam porównywanie powierzchni, nie wspominając już o porów-

nywaniu objętości. Dlatego te typy znacznie rzadziej są używane (wyjątkiem jest powierzchniowy wykres kołowy). Dlatego też, tworząc wykresy, warto pamiętać, żeby sztucznie nie dodawać wymiarów, tworząc wykresy objętościowe w sytuacji gdy wystarczy wykres powierzchniowy lub liniowy, czyli gdy zmianie ulega jedynie jeden wymiar.



Wykres 3.9. Oceny ze sprawdzianu z matematyki (wykres powierzchniowy)

Źródło: opracowanie własne.

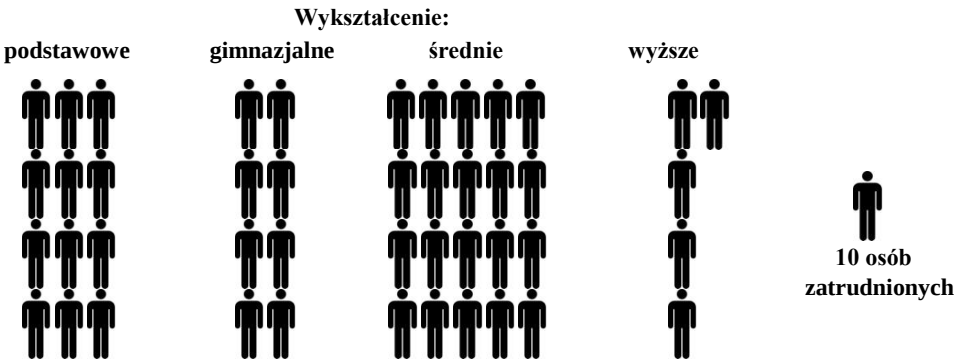
Ostatni z prezentowanych wykresów (3.9) jest szczególnym wykresem powierzchniowym, w którym porównuje się powierzchnie wycinków koła. Jego wyjątkowość polega na tym, że porównywanie powierzchni wycinka koła sprowadza się faktycznie jedynie do porównywania jednego wymiaru. Wynika to stąd, iż pola wycinków jednego koła (bo mają one ten sam promień) są zależne wprost proporcjonalnie jedynie od wielkości kąta (lub długości łuku), czyli ocena dokonywana jest jednowymiarowo.

Warto również zwrócić uwagę na dodatkowe zalety wykorzystania wykresów kołowych. Dają one możliwość prezentowania częstości występowania każdej z cech w odniesieniu do całej zbiorowości. Jeden rzut oka i widzimy strukturę rozkładu cechy w populacji.

■

Wykresy ilościowe powstają poprzez wielokrotne użycie odpowiedniego symbolu reprezentującego przyjętą jednostkę, zatem tworząc go, należy najpierw dobrać jej wielkość. Jako symbol najczęściej wybiera się figurę geometryczną lub obrazek przedstawiający poglądowo dane zjawisko (np. sylwetka samochodu jako jednostka produkcji samochodów w kolejnych latach w pewnej firmie). Bardzo jest to wygodne zwłaszcza przy prezentacji cech jakościowych.

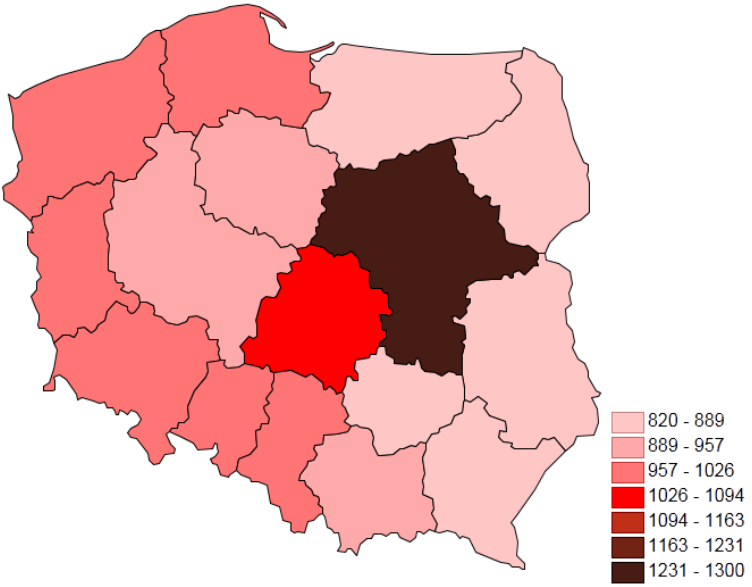
Na wykresie 3.10 przedstawiono przykład takiego wykresu opracowanego dla danych z tabeli 3.3, prezentującej poziom wykształcenia pracowników pewnego przedsiębiorstwa. Za jednostkę (jedna postać człowieka) przyjęto 10 zatrudnionych.



Wykres 3.10. Poziom wykształcenia pracowników pewnego przedsiębiorstwa (wykres ilościowy)

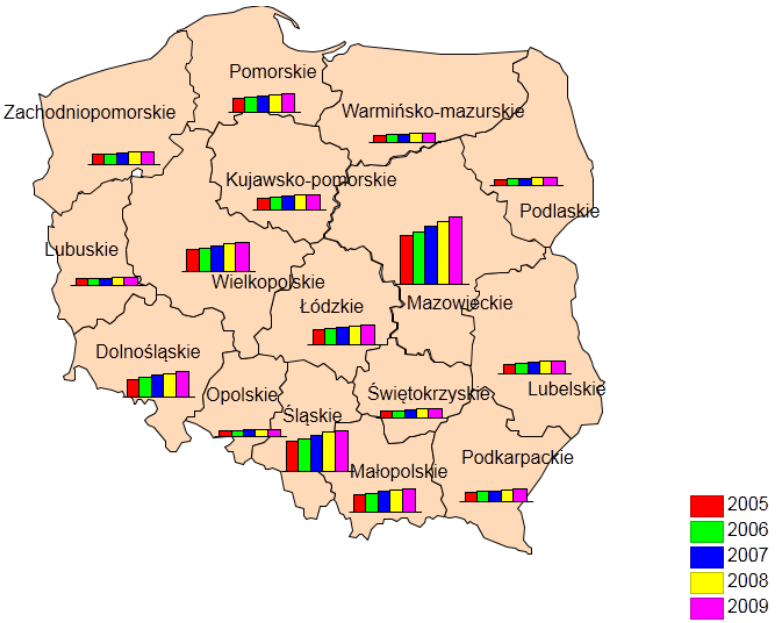
Źródło: opracowanie własne na podstawie danych umownych.

Kartogramy to rodzaj wykresu wykonywanego na mapach, obrazującego terytorialne występowanie danego zjawiska. Powstają one poprzez umieszczenie figur, symboli graficznych w odpowiedniej liczbie lub wielkości na poszczególnych jednostkach terytorialnych, lub też poprzez nadanie barw lub zakreskowanie obrazujące natężenie zjawiska. Przykłady kartogramów zostały przedstawione na wykresach 3.11 i 3.12.



Wykres 3.11. Przeciętne miesięczne wydatki na jedną osobę w Polsce w 2010 roku

Źródło: opracowanie własne na podstawie: *Rocznik Statystyczny Województw 2011*, Główny Urząd Statystyczny, Warszawa 2012, s. 330.



Wykres 3.12. Produkt narodowy brutto w poszczególnych województwach w latach 2005-2009

Źródło: opracowanie własne na podstawie: *Rocznik Statystyczny Województw 2011*, Główny Urząd Statystyczny, Warszawa 2012, s. 90; *Rocznik Statystyczny Województw 2009*, Główny Urząd Statystyczny, Warszawa 2010, s. 88; *Rocznik Statystyczny Województw 2008*, Główny Urząd Statystyczny, Warszawa 2009, s. 84.

Opisane powyżej typów wykresów mogą przybierać różne formy w zależności od specyfiki danych, które mają obrazować. I tak omówione poprzednio cechy, po pogrupowaniu ich wartości w szeregi rozdzielcze, mogą być przedstawione w szczególnej formie graficznej. Do tego celu najczęściej wykorzystywany jest wykres słupkowy (tzn. wymiarowy, powierzchniowy), nazywany histogramem. **Histogram jest „(…) uporządkowanym zbiorem prostokątów, których podstawy spoczywają na osi odciętych i są wyznaczone przez odcinki odpowiadające kolejnym przedziałom wartości zmiennej, a których wysokości odpowiadają natężeniom liczebności lub częstości w kolejnych klasach**”[61]. Należy podkreślić, że *histogram można konstruować tylko dla szeregów rozdzielczych przedziałowych.* Dla szeregów rozdzielczych punktowych stosuje się wykresy słupkowe (np. takie jak wykres 3.8), które nie mogą być nazwane histogramem, gdyż szerokość słupka nie ma w ich przypadku znaczenia. Przy pomocy wykresów słupkowych przedstawia się również strukturę populacji ze względu na cechę niemierzalną (wykres 3.13). Takiego wykresu także nie możemy nazwać histogramem.

[61] B. Szulc [1972, s. 95].

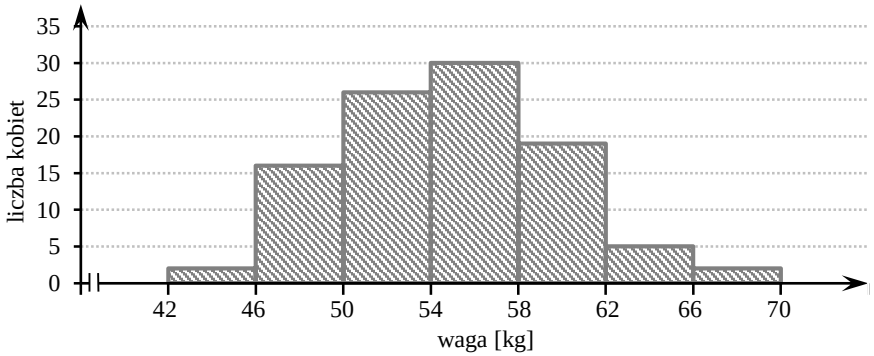
Wykres 3.13. Poziom wykształcenia pracowników pewnego przedsiębiorstwa (wykres słupkowy)

Źródło: opracowanie własne na podstawie danych umownych.

Poniżej przedstawmy przykład histogramu utworzonego dla danych wcześniej pogrupowanych w szereg rozdzielczy przedziałowy.

Przykład 3.5 (kontynuacja 3.3)

Wykorzystując dane dotyczące wagi studentek kierunku zarządzanie oraz pogrupowanie tych danych w szereg rozdzielczy (tabela 3.19), otrzymamy histogram przedstawiony na wykresie 3.14.

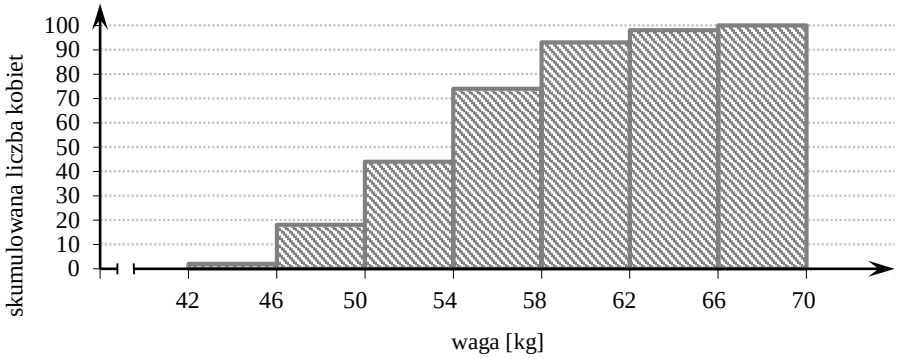


Wykres 3.14. Waga studentek kierunku zarządzanie (histogram zwykły)

Źródło: opracowanie własne na podstawie danych umownych.

Nie jest to jedyna forma wykresu, jaki może być utworzony dla tego typu szeregu. Do konstrukcji histogramu można wykorzystywać zarówno dane zgrupowane w szereg rozdzielczy przedziałowy zwykły, jak też w formie skumulowanej, tak jak to zostało przedstawione na wykresie 3.15. Ponadto w obrazowaniu danych przedstawianych w formie histogramu można stosować, jak wspomniano powyżej,

zarówno liczebności, jak i częstości. W obu przypadkach kształt wykresu pozostanie identyczny, zmianie ulega jedynie jednostka na osi liczbowej OY.



Wykres 3.15. Waga studentek kierunku zarządzanie (histogram skumulowany, reprezentujący także dystrybuantę empiryczną)

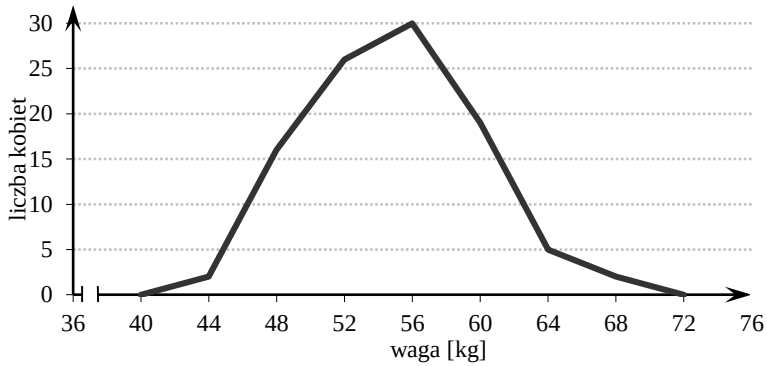
Źródło: opracowanie własne na podstawie danych umownych.

■

Innym typem wykresu tworzonego na podstawie danych z szeregu rozdzielczego przedziałowego jest **wielobok liczebności** (częstości). Tworzony jest on poprzez połączenie odcinkami punktów, których współrzędne (x, y) tworzone są przez środki przedziału wartości zmiennej w danej klasie (współrzędna x) oraz odpowiadającej przedziałowi liczebności lub częstości (współrzędna y). Dodatkowo dokładane są dwa punkty skrajne reprezentujące środek przedziału, który poprzedza pierwszy przedział lub następuje za ostatnim (współrzędna x), oraz zero (współrzędna $y = 0$).

Przykład 3.6 (kontynuacja 3.3)

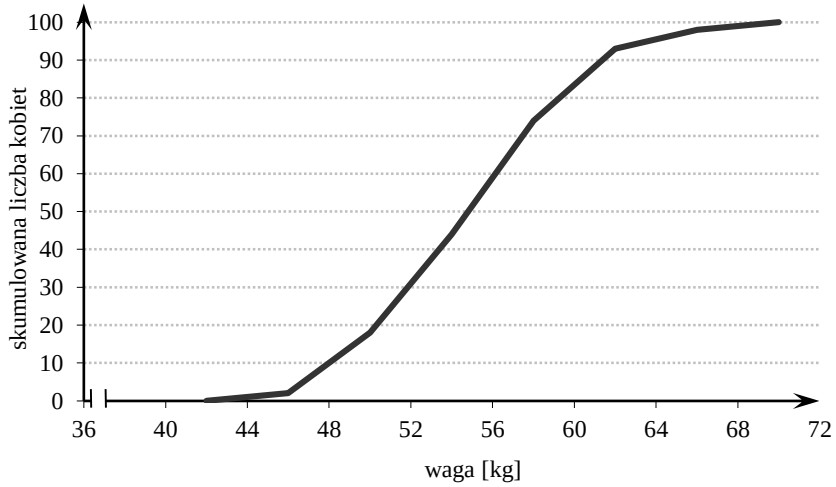
Korzystając z tabeli 3.14 i z wykresu 3.15, skonstruowano wielobok liczebności obrazujący rozkład wagi studentek kierunku zarządzanie. Został on przedstawiony na wykresie 3.16.



Wykres 3.16. Waga studentek kierunku zarządzanie (wielobok liczebności)

Źródło: opracowanie własne na podstawie danych umownych.

Analogiczny wielobok liczebności może zostać opracowany w postaci skumulowanej, z tą różnicą, że nie dodajemy punktu po ostatnim przedziale. Na wykresie 3.17 przedstawiono skumulowany wielobok liczebności dla analizowanego przykładu.



Wykres 3.17. Waga studentek kierunku zarządzanie (skumulowany wielobok liczebności, dystrybuanta empiryczna)

Źródło: opracowanie własne na podstawie danych umownych.

■

Wykonując wykresy, zwłaszcza na początku edukacji statystycznej, można popełnić szereg błędów zarówno koncepcyjnych, jak też związanych z interpretacją otrzymanych wykresów. Dotąd omawialiśmy konstrukcję szeregów o jednakowej rozpiętości przedziałów. Często jednak, ze względu na specyfikę cechy, konstruuje się szeregi o nierównych rozpiętościach przedziałów[62] Taka sytuacja sprzyja powstawaniu błędów interpretacyjnych przy konstrukcji histogramu oraz analizie szeregu opartej o liczebności poszczególnych przedziałów, bez uwzględnienia **różnic w rozpiętościach przedziałów**. Często prowadzi to do błędnych wniosków. Zilustrujmy to poniższym przykładem 3.7.

Przykład 3.7

Tabela 3.21 przedstawia powierzchnię uprawy ziemniaków w indywidualnych gospodarstwach rolnych (liczba gospodarstw uprawiających ziemniaki została podana w tys.) w Polsce w 2002 roku.

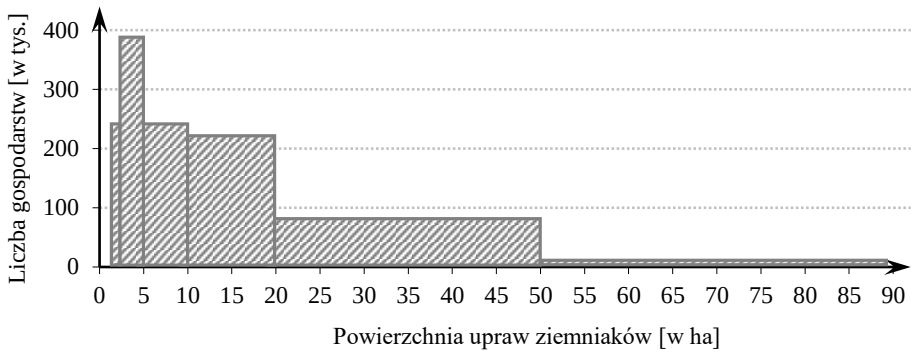
Tabela 3.21. Powierzchnia uprawy ziemniaków w indywidualnych gospodarstwach rolnych według grup obszarowych

Powierzchnia upraw ziemniaków [w ha]	Liczba gospodarstw uprawiających ziemniaki [w tys.]
1-2	227,2
2-5	385,3
5-10	325,0
10-20	217,1
20-50	74,0
>50	8,1

Źródło: *GUS. Zróżnicowanie regionalne rolnictwa, opracowanie pod kierunkiem prof. dr hab. Józefa Zegara,* Warszawa 2003, tabl. 7.5, s. 163.

W oparciu o dane zawarte w tabeli 3.21 wykonano wykres obrazujący rozkład powierzchni upraw ziemniaków w indywidualnych gospodarstwach rolnych (wykres 3.18).

[62] Pomijamy problem konstrukcji tego typu szeregów jako bardziej zaawansowany i skupimy się wyłącznie na interpretacji wyników.



Wykres 3.18. Powierzchnia uprawy ziemniaków w indywidualnych gospodarstwach rolnych według grup obszarowych

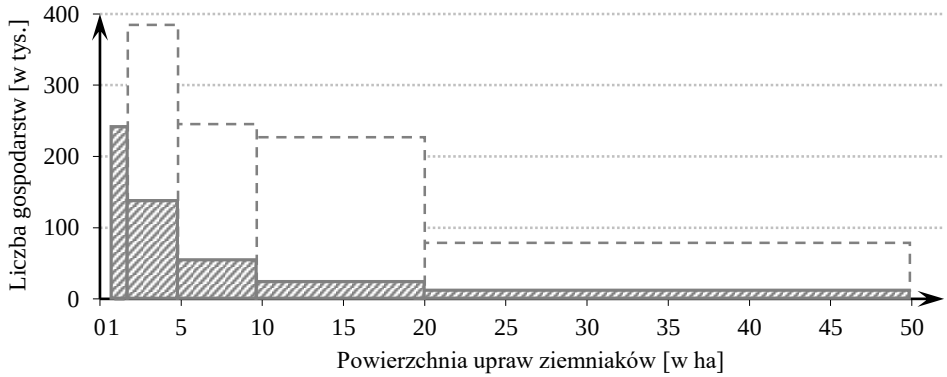
Źródło: opracowanie własne.

Analizując wykres 3.18, można błędnie wywnioskować, że najczęściej występują gospodarstwa mające 2-5 ha upraw ziemniaków. Jednak, aby móc to stwierdzić poprawnie, należy najpierw uwzględnić fakt, iż mamy tu do czynienia z nierówną rozpiętością przedziałów. Wobec tego być może przedział drugi jest liczniejszy niż pierwszy, ponieważ jest trzy razy większy. W takim przypadku należy przeliczyć liczebności z uwzględnieniem różnej szerokości przedziałów i dopiero wtedy można określić, które powierzchnie przeważają. Należy więc wówczas wyznaczyć histogram dla natężenia liczebności w przedziałach, czyli relacji liczby jednostek statystycznych do rozpiętości tych przedziałów. Zestawienie: histogramy liczebności z histogramem natężenia liczebności przedstawiono na wykresie 3.19. Przerywaną linią zaznaczono słupki, gdzie liczebność jest przedstawiona jako wysokość, zaś zakreskowane słupki przedstawiają natężenie liczebności w przeliczeniu na 1 ha powierzchni upraw ziemniaków w gospodarstwie.

Wykres zakreskowany powstał w następujący sposób: po podzieleniu przedziałów na jednakowe jednostki (w tym przypadku był to 1 ha) wielkości słupków zmniejszono tyle razy, ile razy ta jednostka mieściła się w szerokości przedziału cechy. Przykładowo drugi przedział miał rozpiętość od 2 do 5 ha, czyli odcinek odzwierciedlający 1 ha mieścił się w nim trzy razy. Zatem wysokość drugiego słupka została trzykrotnie zmniejszona.

Analizując nowo utworzony wykres, można stwierdzić, że w przedziałach jednohektarowych z zakresu od 2 do 5 ha powierzchni upraw ziemniaków liczba gospodarstw wynosiła średnio około 128,43 tys. gospodarstw. Należy też zauważyć, że największe natężenie liczebności było w przedziale gospodarstw o najmniejszych powierzchniach upraw od 1 do 2 ha. Widać stąd, że wysoka liczebność

w drugim przedziale związana była z jego szerokością, a nie z faktyczną największą występowalnością[63].



Wykres 3.19. Powierzchnia uprawy ziemniaków w indywidualnych gospodarstwach rolnych według grup obszarowych (zakres do 50 ha)

Źródło: opracowanie własne.

■

Oprócz wskazanego powyżej problemu konstrukcji wykresu i jego interpretacji mogą wystąpić również inne, dotyczące nie tylko histogramu, ale też innych typów diagramów. Przykładowo, chcąc porównać kilka wykresów, należy pamiętać o odpowiednim dobraniu skali (niezależnie od formy wykresu), gdyż różne skale mogą prowadzić do błędnych interpretacji. Poniższy przykład obrazuje, jakie błędy można wówczas popełnić.

Przykład 3.8

Tabela 3.22 przedstawia produkt krajowy brutto w Polsce w latach 1995-2003 w mln zł. Wykorzystując te dane, wykonano wykres, który następnie przedstawiono w dwóch różnych skalach.

[63] Bliżej o tej procedurze powiemy przy omawianiu dominanty w rozdziale 4.

Tabela 3.22. PKB w Polsce w latach 1995-2003

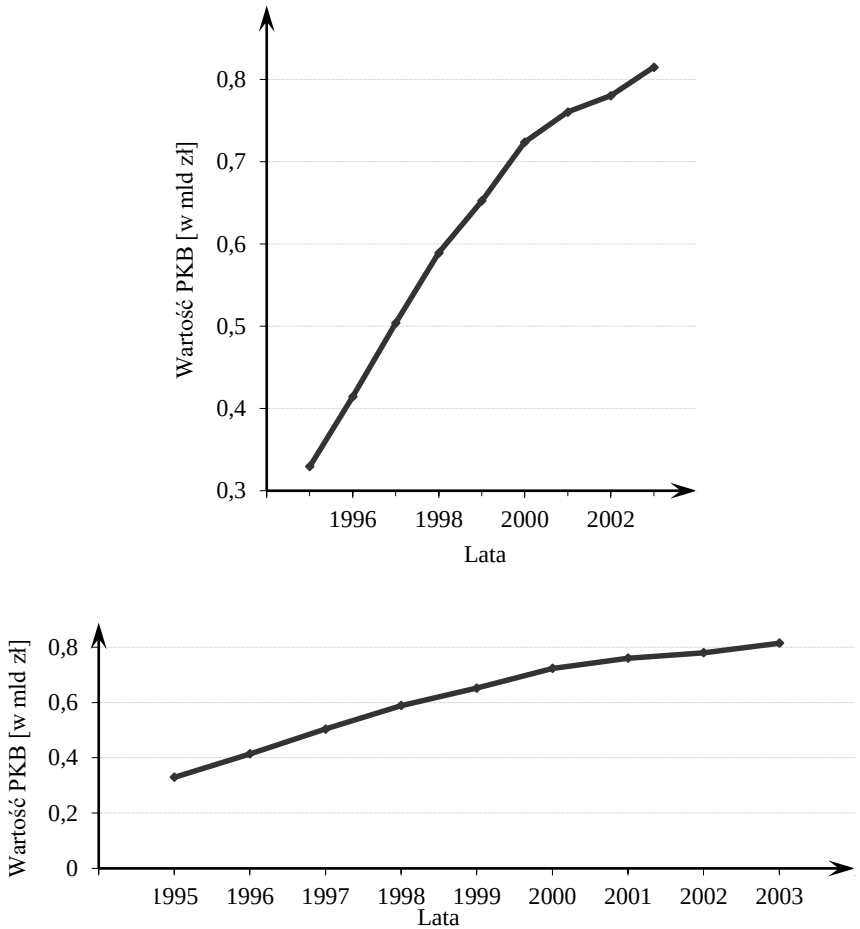
Lata	Produkt krajowy brutto [w mln zł]
1995	329 567,10
1996	414 424,70
1997	504 133,00
1998	589 361,30
1999	652 517,10
2000	723 886,30
2001	760 595,30
2002	780 449,60
2003	814 969,00

Źródło: opracowanie własne.

Analiza danych zawartych w tabeli 3.22 pozwala zauważyć, że wartość PKB w analizowanym przedziale czasowym systematycznie rośnie. Aby określić szybkość[64], z jaką zmieniało się PKB, przyjrzyjmy się dwóm wykresom przedstawionym na wykresie 3.20.

Pierwszy z dwu przedstawionych wykresów pozwala przypuszczać, że wzrost PKB w analizowanym przedziale czasowym był dość dynamiczny. Do odmiennego wniosku co do szybkości zmian można dojść, obserwując uważnie drugi z wykresów. Tym razem można byłoby sądzić, że wzrost PKB przebiegał systematycznie i powoli (przyrosty wydają się małe). Kąt nachylenia krzywej jest znacznie mniejszy niż na poprzednim wykresie, co może zostać zinterpretowane jako znacznie wolniejsze przyrosty. Jednak jest to błędne rozumowanie, gdyż w przypadku obu wykresów przebieg jest dokładnie taki sam. Dlatego też nie można zapominać o innych niż graficzne rodzajach analiz statystycznych. W tym przypadku, jedynie wyliczając odpowiednią miarę dynamiki, można stwierdzić, czy wzrost był szybki czy wolny.

[64] W następnym rozdziale podamy, jak mierzyć dynamikę określoną wstępnie jako szybkość zmiany.



Wykres 3.20. PKB w Polsce w latach 1995-2003 w dwóch różnych skalach

Źródło: opracowanie własne.

■

Należy w tym miejscu podkreślić, że wykonując wykresy, zwłaszcza w przypadku gdy będziemy je porównywać, należy pamiętać o jednakowych skalach. Zastosowanie różnorodnych formatów wykresów może prowadzić do mylnych wniosków. Podobnie rzecz ma się w sytuacji, gdy na jednym rysunku umieszczane są różne wykresy. Jeżeli są one niezwiązane ze sobą, wówczas może to powodować złudne skojarzenie o związku pomiędzy cechami.

Kluczowe pojęcia

Skale pomiarowe, skala nominalna, porządkowa, przedziałowa, ilorazowa
Badanie statystyczne, dane surowe, szereg uporządkowany
Szereg rozdzielczy: prosty, przedziałowy
Szereg czasowy, szereg przekrojowy
Dystrybuanta empiryczna
Histogram, wielobok, liczebność

Pytania kontrolne

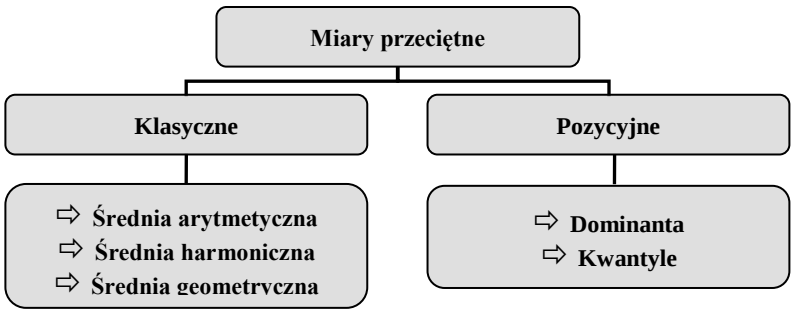
1. W jakich skalach dokonywane są pomiary?
2. Jakie są etapy badania statystycznego?
3. Co to jest szereg rozdzielczy?
4. Czym różni się szereg rozdzielczy cechy skokowej od szeregu rozdzielczego cechy ciągłej?
5. Jakie są typy wykresów i czym się one charakteryzują?
6. Co to jest histogram?
7. Dla jakich zmiennych możliwe jest wykonywanie histogramu?
8. W postaci jakich szeregów rozdzielczych przedstawiamy ilościowe cechy skokowe?
9. W postaci jakich szeregów rozdzielczych przedstawiamy ilościowe cechy ciągłe?

4. Miary przeciętne opisu zbiorowości

Pełny opis zbiorowości za pomocą wszystkich danych zawartych w tablicach oraz wykresów zawiera często zbyt wiele szczegółowych informacji. Nie nadaje się też do porównywania dwóch lub więcej zbiorowości czy cech. Aby uchwycić podstawowe własności populacji i zwięźle je przedstawić, używa się więc pewnych miar nazywanych **parametrami opisowymi** zbiorowości statystycznej. **Parametry w statystyce opisowej są to pewne stałe wartości charakteryzujące całą populację, obliczone na podstawie szeregu statystycznego[65].** Porównując populacje, porównujemy przede wszystkim te parametry.

Przykładowo, chcąc porównać poziom płac w dwóch województwach, nie możemy analizować po kolei poszczególnych płac czy przedziałów płacowych w szeregach statystycznych lub wybrać losowo po jednej osobie z województwa i porównać ich płace. Porównania takie nic nam nie powiedzą o różnicach w płacach w tych zbiorowościach jako całościach.

Do tego celu statystycy zaproponowali używanie parametrów, które zwięźle pokażą rozkład cechy w zbiorowości oraz umożliwią porównania. Podstawowe miary takiego opisu przedstawia schemat na rysunku 4.1.



Rysunek 4.1. Podział miar przeciętnych

Źródło: opracowanie własne.

[65] Zwracamy uwagę Czytelnika, że wartości parametrów obliczone na podstawie rozkładu empirycznego mogą nie reprezentować rzeczywistej cechy występującej w badanej populacji.

Podstawowymi miarami opisującymi całą zbiorowość oraz umożliwiającymi porównania zbiorowości w czasie i przestrzeni są miary przeciętne. Pomijając indywidualne różnice w płacach badanych jednostek obu województw, porównanie przeciętnych płac w badanych województwach jest znacznie efektywniejszym sposobem określenia, które województwo ma wyższy poziom płac[66]. Podział na miary średnie został przedstawiony na rysunku 4.1. W dalszej części zostanie przedstawiony szczegółowy opis każdej z nich.

4.1. Miary klasyczne

Miary przeciętne opisują zbiorowość statystyczną, abstrahując od różnic pomiędzy poszczególnymi jednostkami zbiorowości. Jeżeli używamy miar przeciętnych do porównań na przykład dwóch populacji, wówczas traktujemy te populacje tak, jakby wszystkie jednostki statystyczne miały tę samą wartość cechy – wartość przeciętną.

Średnie klasyczne reprezentują wypadkowe wszystkich występujących w zbiorowości wartości cechy, zarówno skrajnie wysokich czy niskich, jak i tych środkowych. Mają więc charakter abstrakcyjny. Mogą przyjmować wartości niespotykane w danej zbiorowości lub niemożliwe do wystąpienia w przypadku cechy skokowej, jak na przykład średnia liczba dzieci w rodzinie niebędąca liczbą całkowitą (np. wynosząca 2,7).

Jeżeli badana zbiorowość jest niejednorodna ze względu na badaną cechę, wówczas średnia nie będzie poprawnie odzwierciedlała tendencji centralnej, na przykład rozkładu wagi czy wzrostu w zbiorowości kobiet i mężczyzn, ponieważ zarówno waga ciała, jak i wzrost „zależą" od płci. Średnia w tym wypadku będzie sztucznym parametrem bez właściwej interpretacji. Zbiorowość niejednorodną ze względu na działanie przyczyn głównych, kształtujących badaną cechę, musimy podzielić tak, by do badania trafiły jednostki kształtowane przez ten sam czynniki główny. W przypadku wagi i wzrostu osób dorosłych należałoby przeprowadzić badanie odrębnie dla kobiet i mężczyzn.

4.1.1. Średnia arytmetyczna

Średnia arytmetyczna *jest to suma wartości cechy*[67] *wszystkich jednostek statystycznych podzielona przez liczbę jednostek w tej zbiorowości. Mówi więc nam, jaką wartość cechy miałaby każda jednostka statystyczna w populacji, gdyby*

[66] Aczkolwiek poprawne określenie, w którym województwie ludziom zarabia się lepiej, zależy jeszcze od paru innych miar, o których powiemy w dalszej części wykładu.

[67] Tę sumę będziemy nazywali funduszem cechy.

wszystkie jednostki musiały mieć tę samą wartość przy istniejącym funduszu cechy. Mówi nam, ile funduszu cechy przypadłoby na każdą jednostkę w zbiorowości, gdyby fundusz podzielono po równo.

Uczulamy jednak czytelnika w tym miejscu na to, że obliczane w badaniach statystycznych wartości parametrów (w tym przypadku średniej) w rzeczywistości mogą nie występować w badanym rozkładzie. Przykładem niech będzie średnia arytmetyczna ocen wynosząca 4,4. Wiadomo, że ocena 4,4 nie istnieje w przyjętej w Polsce skali ocen.

Średnią (nieważoną) z szeregu prostego liczy się jako sumę wartości cechy wszystkich jednostek podzieloną przez liczbę tych jednostek. Wyznaczona jest ona ze wzoru:

$$\bar{x} = \frac{x_1 + x_2 + ... + x_i + ... + x_N}{N} = \frac{1}{N}\sum_{i=1}^{N} x_i, \tag{4.1}$$

gdzie:

$\bar{x}$ – średnia arytmetyczna,

$x_1, x_2, ..., x_i, ..., x_N$ – obserwacje pewnej cechy X,

N – liczba obserwacji czyli, liczba jednostek statystycznych w badanej populacji,

$\sum_{i=1}^{N} x_i$ – suma po wszystkich wartościach cechy w zbiorowości lub jej części (zwana funduszem cechy).

Powyższy wzór powinien czytelnikowi być znajomy, albowiem właśnie w ten sposób liczy się średnią ocen na koniec semestru lub roku.

Przykład 4.1

Zapytano dziesięciu studentów, z ilu osób składa się ich rodzina (oni, rodzice i rodzeństwo). Otrzymano następujące odpowiedzi: 2, 3, 1, 4, 6, 4, 5, 4, 3, 4.

Średnia arytmetyczna (przeciętna) liczby osób w rodzinie w badanej zbiorowości wynosi więc:

$$\bar{x} = \frac{x_1 + x_2 + ... + x_i + ... + x_k}{N} = \frac{2+3+1+4+6+4+5+4+3+4}{10} = \frac{36}{10} = 3{,}6 \text{ osoby.}$$

Jak widać, w tym przypadku przeciętna w badanym zbiorze jest liczbą w rzeczywistości nie występującą.

■

Średnią arytmetyczną dla szeregu jednostopniowego (tzw. średnia ważona) wyznacza się ze wzoru:

$$\overline{x} = \frac{x_1 \cdot n_1 + x_2 \cdot n_2 + ... + x_i \cdot n_i + ... + x_k \cdot n_k}{N} = \frac{1}{N} \sum_{i=1}^{k} x_i n_i, \qquad (4.2)$$

gdzie:

k – ilość przedziałów cechy lub odmian wartości w przypadku cechy dyskretnej (i = 1, 2, …, k),

n_i – liczebności w poszczególnych przedziałach lub frakcja (procent) wystąpienia danej wartości cechy dyskretnej, przy czym $\sum_{i=1}^{k} n_i = N$.

Liczebności te nazywamy wagami cechy. Mówią one, jaka jest częstość występowania danej odmiany cechy w populacji, czyli jak ważna jest dana wartość cechy. Stąd też nazwa średnia ważona dla tak liczonej przeciętnej. Uważny czytelnik zauważy, że oba wzory różnią się tylko formą zapisu, a nie istotą liczenia średniej. W przypadku średniej ważonej używamy dodatkowo mnożenia zamiast tylko dodawania wartości cech[68].

Dla szeregu, w którym zamiast liczebności korzysta się z częstości, średnią arytmetyczną liczy się zgodnie ze wzorem:

$$\overline{x} = x_1 \cdot f_1 + x_2 \cdot f_2 + ... + x_i \cdot f_i + ... + x_k \cdot f_k = \sum_{i=1}^{k} x_i f_i, \qquad (4.3)$$

gdzie:

$f_i = \frac{n_i}{N}$ – częstość występowania danej wartości cechy dla i = 1, 2, …, k.

Wzór (4.3) powstał przez podzielenie każdego składnika sumy ze wzoru (4.2) przez liczebność ogólną N, co pokazano poniżej:

$$\overline{x} = x_1 \cdot \frac{n_1}{N} + x_2 \cdot \frac{n_2}{N} + ... + x_i \cdot \frac{n_i}{N} + ... + x_k \cdot \frac{n_k}{N} = x_1 \cdot f_1 + x_2 \cdot f_2 + ... + x_i \cdot f_i + ... + x_k \cdot f_k.$$

Przykład 4.2

Uporządkujmy dane surowe, na podstawie których liczyliśmy w przykładzie 4.1 średnią liczbę osób w rodzinie, tworząc następnie szereg jednostopniowy.

[68] W przypadku szeregu prostego, gdy mamy powtarzającą się wartość cechy, dodajemy je do siebie (np. 2 + 2 + 2 + 3 + 4 + 4 = 17), zaś w przypadku szeregu jednostopniowego najpierw mnożymy wartość cechy przez ilość wystąpień, a potem dodajemy do siebie wyniki mnożenia (2 x 3 + 3 x 1 + 4 x 2 = 17).

Tabela 4.1. Liczność rodzin grupy studentów

Wartość cechy x_i	Liczebności n_i	Łączna ilość osób $x_i n_i$	Częstość $f_i = n_i/N$	$x_i f_i$
1	1	1	0,1	0,1
2	1	2	0,1	0,2
3	2	6	0,2	0,6
4	4	16	0,4	1,6
5	1	5	0,1	0,5
6	1	6	0,1	0,6
Suma	10	36	1	3,6

Źródło: opracowanie własne na podstawie danych umownych.

Korzystając ze wzoru (4.2), można otrzymać:

$$\bar{x} = \frac{x_1 \cdot n_1 + x_2 \cdot n_2 + ... + x_i \cdot n_i + ... + x_6 \cdot n_6}{N} =$$

$$= \frac{1 \cdot 1 + 1 \cdot 2 + 3 \cdot 2 + 4 \cdot 4 + 5 \cdot 1 + 6 \cdot 1}{10} = \frac{36}{10} = 3{,}6 \text{ osoby.}$$

Podstawiając do wzoru (4.3), otrzymamy:

$$\bar{x} = x_1 f_1 + x_2 f_2 + ... + x_i f_i + ... + x_6 f_6 =$$

$$= 1 \cdot 0{,}1 + 2 \cdot 0{,}1 + 3 \cdot 0{,}2 + 4 \cdot 0{,}4 + 5 \cdot 0{,}1 + 6 \cdot 0{,}1 = 3{,}6 \text{ osoby.}$$

■

Podkreślmy to jeszcze raz, średniej ważonej używa się, gdy dane są uporządkowane w szereg jednostopniowy o powtarzających się wartościach cechy. Wagami w tym przypadku są liczebności lub częstości, z jakimi występują poszczególne odmiany cechy[69].

W przypadku cechy ciągłej informacje o rozkładzie podaje się w formie szeregu rozdzielczego przedziałowego. Jeżeli nie mamy dodatkowych informacji, wówczas nie wiemy, jakie konkretnie wartości cechy i z jaką częstością występują w danym przedziale. Bez tej dodatkowej informacji lub bez dodatkowego założenia nie jesteśmy w stanie określić funduszu wartości cechy w przedziale klasowym, a co za tym idzie, policzyć średniej. Standardowo zakłada się więc, że war-

[69] Nie jest to jedyny sposób określania wag przy konstrukcji średniej arytmetycznej. Przykładem niech będzie sposób liczenia średniej na dyplomie licencjackim, gdzie wagi są określone przez Radę Wydziału (średnia jest liczona z oceny z pracy dyplomowej, wyniku egzaminu i średniej ze studiów). Innym przykładem może być sposób liczenia średniej ze studiów na niektórych uczelniach. Liczy się ją jako średnią arytmetyczną ważoną, gdzie wagami są punkty ECTS.

tości cechy z danego przedziału są reprezentowane przez środek przedziału[70]. Mnoży się je przez liczebność przedziału, otrzymując przybliżoną sumę wartości cechy w danej klasie. Jest to przybliżony sposób obliczania średniej. Został on przedstawiony wzorem (4.4) przy wykorzystaniu liczebności jako wag oraz wzorem (4.5) z częstościami jako wagami:

$$\overline{x} = \frac{\overset{o}{x}_1 \cdot n_1 + \overset{o}{x}_2 \cdot n_2 + \ldots + \overset{o}{x}_i \cdot n_i + \ldots + \overset{o}{x}_k \cdot n_k}{N} = \frac{1}{N} \sum_{i-1}^{k} \overset{o}{x}_k n_i, \qquad (4.4)$$

$$\overline{x} = \overset{o}{x}_1 \cdot f_1 + \overset{o}{x}_2 \cdot f_2 + \ldots + \overset{o}{x}_i \cdot f_i + \ldots + \overset{o}{x}_k \cdot f_k = \sum_{i-1}^{k} \overset{o}{x}_i f_i, \qquad (4.5)$$

gdzie:

$\overset{o}{x}_i$ – środek *i*-tego przedziału (i = 1, 2, …, k).

(1) Przyjęcie środków przedziałów powoduje na ogół niedoszacowanie sum wartości cechy w klasach poprzedzających przedział z największą liczebnością, przeszacowanie zaś w klasach następujących po klasie o maksymalnej liczebności. W przypadku szeregu o symetrycznym rozkładzie cechy błędy te znoszą się. Im silniejsza więc jest asymetria[71] rozkładu cechy, tym większy popełniamy błąd, obliczając średnią w oparciu o środki przedziałów. Wzorów (4.4) i (4.5) nie należy zatem stosować do tzw. silnie asymetrycznych szeregów cechy ciągłej, czyli rozkładów typu L, J oraz U. Takie szeregi należy badać w oparciu o szersze informacje, najlepiej mając do dyspozycji dane wyjściowe[72], dla których szereg został skonstruowany. Środek przedziału jest dokładnym reprezentantem cechy jedynie wtedy, gdy poszczególne wartości cechy są rozmieszczone w przedziale równomiernie lub w środku tegoż przedziału. Mając natomiast dodatkowe informacje, możemy policzyć średnie wartości cechy w przedziałach, a następnie na ich podstawie obliczyć średnią dla całej zbiorowości.

---

[70] Przypomnijmy, że środek odcinka o krańcu dolnym x_a oraz górnym x_b wyznacza się jako ich średnia arytmetyczna: $\frac{1}{2}(x_a + x_b)$. W przypadku szeregów statystycznych dolnym krańcem przedziału jest jego początek, zaś jako koniec brany jest dolny kraniec przedziału kolejnego.

[71] Asymetrię, na razie, rozumiemy intuicyjnie jako brak symetrii. Pojęcie symetrii powinno być znane czytelnikowi z lekcji matematyki. Definicję i sposób pomiaru asymetrii w statystyce podamy później.

[72] Jeśli mamy dane surowe, to wszystkie parametry powinno się liczyć z tychże danych (poza dominantą). Zatem wzory 4.4 lub 4.5 i podane na następnych stronach stosujemy, gdy nie mamy dostępu do danych surowych.

(2) Kolejny problem przy analizie cechy ciągłej może wystąpić, gdy skrajne przedziały szeregu rozdzielczego nie są domknięte. Nie można w takim przypadku określić środka przedziału. Jeśli nie mamy informacji o rozkładzie cechy w otwartym przedziale, na przykład średniej wartości cechy w klasie otwartej, wówczas możemy domknąć sztucznie ten przedział *pod dwoma warunkami*:
   (2a) znamy górną (dolną) granicę rozkładu cechy, czyli wiemy, jak domknąć przedział;
   (2b) domykany przedział ma nieznaczną liczebność względną, więc nawet wyolbrzymiając rozpiętość przedziału, popełnimy mały błąd przy liczeniu średniej ze względu na bardzo małą wagę w obliczaniu całkowitej średniej, sztucznie ustalonego środka.

Powstaje pytanie, co można uznać za nieznaczną liczebność względną. Nie ma na to jednoznacznej odpowiedzi. W literaturze na ogół można spotkać pogląd, że nie powinna to być liczebność przekraczająca 5%-10% populacji. Wybór wartości granicznej jest subiektywną decyzją badacza, należy jednak pamiętać, że im większa jest liczebność względna domykanego przedziału, tym większym błędem będzie obarczona obliczona średnia.

Przykład 4.3 (kontynuacja 3.3)

Wróćmy do szeregu rozdzielczego wagi studentek, który skonstruowaliśmy w poprzednim rozdziale, i obliczmy ich średnią wagę. Jak można było zaobserwować, analizując wykres 3.4 populacja studentek jest zbiorem w miarę jednorodnym, o rozkładzie wagi posiadającym dosyć wyraźny punkt skupienia jednostek statystycznych oraz o domkniętych przedziałach skrajnych. Wobec tego możemy dla tego szeregu obliczyć średnią arytmetyczną. Najwygodniej jest dokonywać obliczeń w formie tablicy, w której kolejne kolumny reprezentują składowe użytego do obliczeń wzoru.

Tabela 4.2. Rozkład wagi studentek zarządzania – obliczanie średniej arytmetycznej

$\lfloor x_{i_0} - x_{i_1} \rfloor$	n_i	$\overset{o}{x}_i$	$\overset{o}{x}_i n_i$
42-46	2	44	88
46-50	16	48	768
50-54	26	52	1 352
54-58	30	56	1 680
58-62	19	60	1 140
62-66	5	64	320
66-70	2	68	136
Suma	100		5 484

Źródło: opracowanie własne.

Podstawiając do wzoru (4.4) i korzystając z wyliczeń zawartych w tabeli 4.2, otrzymamy następującą średnią:

$$\bar{x} = \frac{\sum_{i-1}^{k} x_i n_i}{N} = \frac{5484}{100} = 54{,}84 \approx 54{,}8 \text{ kg.}$$

Zatem z przedstawionych wyliczeń wynika, że średnia waga badanych studentek kierunku zarządzanie wynosi 54,8 kg.

■

Przykład 4.4

W tabeli 4.3 przedstawiono liczbę sklepów w Polsce według województw w 2002 roku.

Tabela 4.3. Liczba sklepów według województw w 2002 roku w Polsce

Liczba sklepów w województwach (w tys.)	Liczba województw
10-20	5
20-30	5
30-40	3
40-50	1
50-60	1
60-70	1

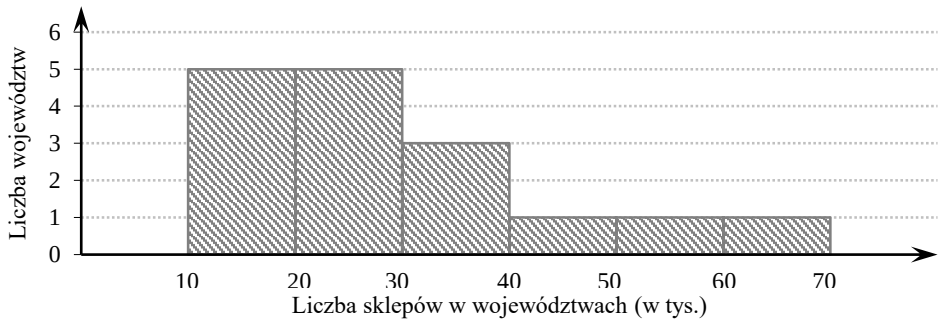
Źródło: opracowanie własne.

Wyznaczmy średnią liczbę sklepów w województwach w 2002 roku.

Mimo że wyznaczenie średniej jest formalnie możliwe, nie powinno się jej wyznaczać, gdyż w tym przypadku mamy do czynienia z szeregiem, w którym brak jest jednego wyraźnego punktu (przedziału) skupienia – ta zbiorowość nie jest jednolita. Praktycznie rzecz biorąc, można uznać, że szereg ten składa się z co najmniej dwóch populacji reprezentowanych przez jednostki w pierwszych trzech przedziałach jako jedna populacja i następne trzy jako druga, o znacznie różniących się liczebnościach i średnich poziomach wartości cechy. Powyższy szereg nazywamy szeregiem skrajnie asymetrycznym[73]. Jak widać na wykresie 4.1, liczebności kolejnych odmian cechy (przedziałów) są nierosnące w miarę wzrostu wartości cechy, jest to więc szereg monotoniczny[74].

[73] Dokładniej o tym zjawisku i jego pomiarze powiemy w rozdziale o asymetrii.

[74] W matematyce używa się pojęcia „funkcja monotoniczna" i stąd mamy pojęcie szeregu monotonicznego.



Wykres 4.1. Liczba sklepów według województw w 2002 roku w Polsce

Źródło: opracowanie własne.

Tworząc wielobok na podstawie powyższego histogramu, otrzymuje się krzywą nierosnącą i dla takich szeregów nie należy liczyć średniej arytmetycznej, jest ona bowiem przy takim rozkładzie miarą sztuczną, nie reprezentującą poprawnie całej zbiorowości. Nie jest ona w takim przypadku miarą skupienia zbiorowości, czyli punktem, wokół którego skupiają się jednostki (a zatem i wartości ich cechy), a tak interpretujemy średnią arytmetyczną. Takie rozkłady powinny być opisane (reprezentowane) innymi miarami tendencji centralnej niż średnia.

■

Przykład 4.5

Zmierzono czas biegu na 60 m dla 30 uczniów w pewnej szkole średniej. Otrzymane wyniki przedstawiono w tabeli 4.4.

Tabela 4.4. Czas biegu na 60 m uczniów szkoły średniej

Czas biegu na 60 m (w sekundach)	Liczba uczniów
6,0-7,5	2
7,5-8,0	3
8,0-8,5	6
8,5-9,0	8
9,0-9,5	5
9,5-10,0	4
10,0-10,5	2

Źródło: dane umowne.

Jaki był średni czas biegu na 60 metrów wśród ocenianych uczniów szkoły średniej?

Zanim można będzie podstawić do wzoru na średnią arytmetyczną ważoną szeregu rozdzielczego zmiennej ciągłej (czasu biegu na 60 m jest zmienną ciągłą), należy wyznaczyć środek każdego przedziału i przemnożyć je przez odpowiednią liczebność (zważyć). Ze względu na to, że nie było w treści przykładu zasugerowanego sposobu domknięcia przedziałów, przyjęto dolne domknięcie. Wyliczenia przedstawiono w tabeli 4.5.

Tabela 4.5. Czas biegu na 60 m uczniów szkoły średniej – wyznaczenie średniej

$\lfloor x_{i_0} - x_{i_1} \rfloor$	n_i	$\overset{o}{x_i}$	$\overset{o}{x_i}\, n_i$
6,0-7,5	2	6,75	13,50
7,5-8,0	3	7,75	23,25
8,0-8,5	6	8,25	49,50
8,5-9,0	8	8,75	70,00
9,0-9,5	5	9,25	46,25
9,5-10,0	4	9,75	39,00
10,0-10,5	2	10,25	20,50
Suma	30		262,00

Źródło: opracowanie własne na podstawie danych umownych.

Podstawiając otrzymane w tabeli 4.5 wyliczenia do wzoru (4.4), wyznaczamy średnią arytmetyczną:

$$\overline{x} = \frac{\sum_{i-1}^{k} \overset{o}{x_i}\, n_i}{N} = \frac{262}{30} = 8,733 \text{ s.}$$

Należy w tym miejscu zaznaczyć, że powyższa średnia jest zaniżona przez dwie obserwacje odbiegające od reszty (pierwszy, znacznie szerszy od pozostałych przedział). Dlatego też należy się zastanowić nad tym, czy w tym przypadku wszystkie obserwacje pochodzą z jednej populacji, uwzględniającej na przykład uczniów tylko jednej (rocznikowo) klasy. Jeżeli okaże się, że obserwacje odstające rzeczywiście należą do innej zbiorowości (np. są to uczniowie z niższej klasy niż większość badanych), wówczas analizie powinni zostać poddani jedynie ci, którzy stanowią większość (bez wartości odstających). Jednak w przypadku, gdy nie uda się takiego rozdzielenia dokonać, powinno się przeprowadzić badanie dla całej populacji, wykorzystując w tym celu miary pozycyjne.

Jeżeli w badanym przykładzie ustalono, dlaczego te dwie obserwacje odstają od reszty (np. pomiary dotyczą uczniów młodszych rocznikowo), należy wyeliminować pierwszy wiersz z tabeli (4.4)), a wówczas wyliczenia dotyczące średniej wyglądają następująco:

$$\bar{x} = \frac{\sum_{i-1}^{k} {}^{o} x_i n_i}{N} = \frac{248{,}5}{28} = 8{,}875\,\text{s}.$$

Zatem otrzymana wartość średniej, zgodnie z poprzednimi przypuszczeniami, jest wyższa (o 0,142 sekundy) od wyliczonej poprzednio. Podajemy dla ilustracji problemu bardzo uproszczony przykład, w rzeczywistości badacz musiałby dokonać bardziej dogłębnej analizy i być bardziej ostrożny we wnioskowaniu, że odrzucone obserwacje są nietypowe.

■

Pokazaliśmy czytelnikowi, jak obliczyć średnią arytmetyczną dla różnego rodzaju szeregów. Na zakończenie omawiania średniej arytmetycznej podamy kilka najważniejszych, i bardzo przydatnych w bardziej złożonych analizach własności tejże średniej.

Własności średniej arytmetycznej:

1. $\bar{c} = c.$

Średnia ze stałej jest równa tej stałej.

2. $\overline{(x \pm c)} = \bar{x} \pm c.$

Powyższa własność oznacza, że jeżeli do każdej wartości szeregu dodamy lub odejmiemy stałą, średnią nowego szeregu możemy policzyć jako średnią szeregu wyjściowego powiększoną lub pomniejszoną o tę stałą. Dodanie (odjęcie) stałej *c* powoduje przesunięcie szeregu o *c* jednostek na osi odciętych *X*, a tym samym przesunięcie średniej o tyle samo jednostek. Aby to udowodnić, wystarczy wykonać następujące przekształcenie:

$$\overline{(x \pm c)} = \frac{\sum_{i=1}^{k}(x_i \pm c)n_i}{\sum_{i=1}^{k} n_i} = \frac{\sum_{i=1}^{k}(x_i n_i \pm c n_i)}{\sum_{i=1}^{k} n_i} = \frac{\sum_{i=1}^{k} x_i n_i}{\sum_{i=1}^{k} n_i} \pm \frac{\sum_{i=1}^{k} c n_i}{\sum_{i=1}^{k} n_i} = \bar{x} \pm \frac{c\sum_{i=1}^{k} n_i}{\sum_{i=1}^{k} n_i} = \bar{x} \pm c.$$

3. $\sum_{i=1}^{k}(x_i - \bar{x})\frac{n_i}{N} = 0.$

Własność ta oznacza, iż suma ważonych odchyleń wartości cechy od średniej arytmetycznej tej cechy w zbiorowości jest zawsze równa zeru. Łatwo to wykazać, dokonując prostych przekształceń przedstawionych poniżej:

$$\sum_{i=1}^{k}(x_i-\bar{x})\frac{n_i}{N}=\sum_{i=1}^{k}(x_i\frac{n_i}{N}-\bar{x}\frac{n_i}{N})=\sum_{i=1}^{k}x_i\frac{n_i}{N}-\sum_{i=1}^{k}\bar{x}\frac{n_i}{N}=\bar{x}-\frac{\bar{x}}{N}\sum_{i=1}^{k}n_i=\bar{x}-\bar{x}=0.$$

4. $\bar{\bar{x}}=\overline{(\bar{x}_j)}=\dfrac{\sum_{j=1}^{m}\bar{x}_j n_{\bullet j}}{\sum_{j=1}^{m}n_{\bullet j}}=\dfrac{\sum_{j=1}^{m}\bar{x}_j n_{\bullet j}}{N}$ gdzie $n_{\bullet j}=\sum_{i=1}^{k}n_{ij}$ jest liczebnością *j*-tej grupy.

Własność ta oznacza, że jeżeli zbiorowość podzielono na *m* przedziałów wartości cechy i dla każdej z nich obliczono średnią arytmetyczną, wówczas średnią dla całej zbiorowości można obliczyć jako średnią arytmetyczną ze średnich grupowych[75], ważonych liczebnościami $n_{\bullet j}$ poszczególnych grup. Oznacza to, że średnia sumy jest równa ważonej sumie średnich. Jest to wynik następującego postępowania dowodowego. Załóżmy, że mamy *m* grup w populacji, w każdej grupie występuje *k* odmian cechy zmiennej x_i o liczebnościach (wagach) $n_{1j}, n_{2j}, ..., n_{kj}$ odpowiednio. Zatem każda średnia grupowa jest obliczana jako:

$$\bar{x}_j=\frac{\sum_{i=1}^{k}x_i n_{ij}}{\sum_{i=1}^{k}n_{ij}}=\frac{\sum_{i=1}^{k}x_i n_{ij}}{n_{\bullet j}}.$$

Średnią w całej populacji (dla wszystkich grup łącznie) obliczymy jako:

[75]Aby zaznaczyć, że średnia jest liczona jako średnia ze średnich grupowych, dodajemy jeszcze jedną kreskę nad symbolem średniej $\bar{\bar{x}}$ lub, jak niektórzy autorzy, można użyć symbolu $\bar{x}_{\bullet}$

$$\bar{x} = \frac{\sum_{i=1}^{k} x_i n_i}{N} = \frac{\sum_{i=1}^{k} x_i n_i}{\sum_{i=1}^{k} n_i} = \frac{\sum_{i=1}^{k} x_i \sum_{j=1}^{m} n_{ij}}{\sum_{j}^{m}\sum_{i=1}^{k} n_{ij}} = \frac{\sum_{i=1}^{k}\sum_{j=1}^{m} x_i n_{ij}}{\sum_{j}^{m}\sum_{i=1}^{k} n_{ij}}$$

$$= \frac{\sum_{j=1}^{m}\sum_{i=1}^{k} x_i n_{ij}}{\sum_{j=1}^{m}\sum_{i=1}^{k} n_{ij}} = \frac{\sum_{j=1}^{m} \frac{\sum_{i=1}^{k} x_i n_{ij}}{\sum_{i=1}^{k} n_{ij}} \sum_{i=1}^{k} n_{ij}}{\sum_{j=1}^{m}\sum_{i=1}^{k} n_{ij}} =$$

$$= \frac{\sum_{j=1}^{m}\left(\left(\frac{\sum_{i=1}^{k} x_i n_{ij}}{\sum_{i=1}^{k} n_{ij}}\right)\sum_{i=1}^{k} n_{ij}\right)}{\sum_{j=1}^{m}\sum_{i=1}^{k} n_{ij}} = \frac{\sum_{j=1}^{m}\left(\bar{x}_j \sum_{i=1}^{k} n_{ij}\right)}{\sum_{j=1}^{m}\sum_{i=1}^{k} n_{ij}} = \frac{\sum_{j=1}^{m}\left(\bar{x}_j n_{\bullet j}\right)}{\sum_{j=1}^{m} n_{\bullet j}} = \bar{\bar{x}}.$$

5. $\overline{cx} = c\bar{x}$.

Średnia dla szeregu wartości, który pomnożono przez stałą c, jest równa „starej" średniej szeregu wyjściowego pomnożonej tą stałą lub przez jej odwrotność (gdy $c<1$):

$$\overline{cx} = \frac{\sum_{i=1}^{k} cx_i n_i}{N} = \frac{\sum_{i=1}^{k} cx_i n_i}{\sum_{i=1}^{k} n_i} = \frac{c\sum_{i=1}^{k} x_i n_i}{\sum_{i=1}^{k} n_i} = c\bar{x}.$$

Podsumujmy teraz naszą wiedzę o średniej arytmetycznej:

1. W miarę dobrze charakteryzuje ona zbiorowość w przypadku, gdy populacja jest jednorodna względem badanej cechy, czyli poszczególne jednostki nie różnią się od siebie istotnie wartościami cechy. Średnia jest wówczas punktem, wokół którego skupiają się wartości cechy większości jednostek zbiorowości.
2. W przypadku gdy szereg jest wielostopniowy, przyjęcie środka przedziału zamiast średniej w przedziale powoduje, że popełniamy błąd *in minus* (niedo-

szacujemy sumy wartości cechy) w przedziałach przed średnią, zaś *in plus* (przeszacujemy) za średnią. W szeregach symetrycznych te błędy znoszą się i otrzymujemy poprawną wartość średniej dla szeregu wielostopniowego. W przypadku szeregów asymetrycznych popełnimy jednak poważny błąd szacunku, tym większy, im bardziej asymetryczny jest szereg. Nie można więc używać wzoru (4.4) i (4.5) do silnie asymetrycznych szeregów (należy sięgnąć do surowych danych lub, gdy ich brak, ograniczyć analizę do miar polecanych do takich szeregów).

3. Licząc średnią, wykorzystuje się wartości cechy każdej jednostki statystycznej. Brak danych o jakiejkolwiek wartości cechy wyklucza w zasadzie możliwość obliczenia średniej. Poważnym problemem zatem staje się szereg z otwartymi przedziałami skrajnymi, których nie możemy domknąć. Nie można domknąć przedziałów o nieznanych wartościach granicznych i/lub gdy przedziały te są zbyt liczne. Za określenie „liczne" przyjmuje się zwyczajowo liczebności wynoszące ponad 5% lub 10% całej populacji.
4. Nie ma większego sensu interpretacyjnego liczenie średniej dla skrajnie asymetrycznego szeregu jednostopniowego, choć jest to możliwe. Średnia nie reprezentuje w takim przypadku tendencji centralnej (punktu skupienia) zbiorowości. Po prostu takiej tendencji nie ma w tego typu szeregach.
5. Średnia arytmetyczna jest wrażliwa na skrajne wartości cechy w populacji (szeregu), które mogą być uznane za przypadkowe. W takich przypadkach należy ostrożnie interpretować średnią, mając świadomość, że jej wartość może być zaburzona przez te skrajne wartości. Wartość średniej przesuwa się bowiem w stronę tych skrajnych wartości. Jeżeli ich nietypowość wynika z przypadku, średnia kształtowana pod ich wpływem będzie przesunięta w stronę nietypowych (skrajnych) obserwacji i nie będzie dobrze reprezentowała tendencji centralnej w populacji.

4.1.2. Średnia harmoniczna

Kolejną przeciętną miarą klasyczną jest **średnia harmoniczna**. Jest ona miarą wykluczającą się ze średnią arytmetyczną, co znaczy, że jeśli zachodzą warunki do stosowania jednej z nich, to ***nie można*** zamiennie użyć drugiej.

Średnią harmoniczną stosujemy, gdy wartości cechy są podane w formie odwrotności, tj. gdy wartości jednej zmiennej są podane w przeliczeniu na stałą jednostkę innej zmiennej (czyli są wskaźnikami natężenia – np. osoba/km^2) lub wyrażone w postaci złożonej (np. 7 zł za sztukę, 80 km na godzinę), zaś wagi (czyli liczebności) są w jednostkach licznika tej cechy (w przypadku prędkości waga byłaby w km).

To, której średniej należy użyć, zależy od postaci danych, jakie mamy do dyspozycji. Bliżej pokażemy to na jednym z poniższych przykładów.

Tak jak w przypadku średniej arytmetycznej, średnią harmoniczną możemy przedstawić wzorem na średnią nieważoną, ważoną oraz, w przypadku cechy ciągłej, w oparciu o środki przedziałów.

Ogólny wzór na średnią harmoniczną ma postać[76]:

$$\bar{x}_H = H = \frac{N}{\sum_{i=1}^{k} \frac{n_i}{x_i}}, \tag{4.6}$$

gdzie $\frac{1}{x_i}$ to odwrotność badanej cechy, n_i – wagi (liczebności lub częstości, z jaką występują poszczególne odmiany cechy).

Po drobnym przekształceniu widzimy, że średnia harmoniczna jest odwrotnością średniej arytmetycznej liczonej dla odwrotności wartości cechy:

$$\frac{1}{H} = \frac{\sum_{i=1}^{k} \frac{n_i}{x_i}}{N}. \tag{4.7}$$

Przykład 4.6

Samochód jechał na kolejnych kilometrowych odcinkach drogi z następującymi prędkościami: 90, 100, 110, 120, 115, 80. Jaka była średnia prędkość samochodu?

Z definicji średniej jest to stała prędkość, z jaką samochód musiałby jechać, by przebyć tę samą drogę w tym samym czasie:
średnia prędkość ($\bar{v}$) = przebyta droga (*s*) / czas (*t*).

Przebytą drogę *s* znamy – samochód zmieniał prędkość co kilometr pięć razy, czyli *s* = 5 km. Nie wiemy, jak długo jechał, ale jest to możliwe do wyliczenia po przekształceniu wzoru na prędkość do postaci $t = \frac{s}{v}$. Z tego wzoru można wyliczyć czas, w jakim samochód przejeżdżał kolejne kilometrowe odcinki, które następnie należy zsumować:

[76] W literaturze używa się paru symboli na oznaczenie średniej harmonicznej. Podane tu symbole $\bar{x}_H$ oraz H są najczęściej spotykane.

$$t = t_1 + t_2 + t_3 + t_4 + t_5 + t_6 = \frac{1km}{90\frac{km}{godz}} + \frac{1km}{100\frac{km}{godz}} + \frac{1km}{110\frac{km}{godz}} + \frac{1km}{120\frac{km}{godz}} + \frac{1km}{115\frac{km}{godz}} + \frac{1km}{80\frac{km}{godz}}.$$

Zatem średnia prędkość wynosi:

$$\vec{v} = \bar{x}_H = \frac{1km + 1km + 1km + 1km + 1km + 1km}{\frac{1\cancel{km}}{90\frac{\cancel{km}}{godz}} + \frac{1\cancel{km}}{100\frac{\cancel{km}}{godz}} + \frac{1\cancel{km}}{110\frac{\cancel{km}}{godz}} + \frac{1\cancel{km}}{120\frac{\cancel{km}}{godz}} + \frac{1\cancel{km}}{115\frac{\cancel{km}}{godz}} + \frac{1\cancel{km}}{80\frac{\cancel{km}}{godz}}} =$$

$$= \frac{6\,km}{0{,}01111godz + 0{,}01godz + 0{,}0091godz + 0{,}0083godz + 0{,}0087 + 0{,}0125godz} =$$

$$= \frac{6\,km}{0{,}0597godz} = 100{,}5\,km/godz.$$

Z powyższych wyliczeń wynika, że średnia prędkość samochodu na tej pięciokilometrowej trasie wynosiła 100,5 km/godz. i została policzona jako średnia harmoniczna.

■

Aby zrozumieć konstrukcję średniej harmonicznej, należy przyjrzeć się dokonanemu podstawieniu do wzoru, z którego wyliczono średnią prędkość. W liczniku znajduje się N = 5 km i są to zsumowane jednostki statystyczne (kilometry), które obserwujemy pod względem posiadanej cechy, czyli prędkości. Jest to więc nasza zbiorowość statystyczna. W mianowniku mamy sumę odwrotności cechy (prędkość) mnożoną przez liczbę jednostek statystycznych, czyli liczbę kilometrów przebytych z tą prędkością. Ponieważ każda prędkość występowała tylko raz (zmiana co kilometr), możemy powiedzieć, że średnia prędkość samochodu została wyliczona jako nieważona średnia harmoniczna:

$$H = \frac{N}{\sum_{i=1}^{k} \frac{1}{x_i}}. \tag{4.8}$$

Przykład 4.7

Zmodyfikujmy przykład 4.6 w następujący sposób: samochód jechał 4 km z prędkością 90 km/godz., 6 km z prędkością 100 km/godz., 10 km z prędkością 110 km/godz., 20 km z prędkością 120 km/godz., 10 km z prędkością 115 km/godz. i 30 km z prędkością 80 km/godz. Jaka była średnia prędkość samochodu na badanym odcinku drogi?

Wykonując analogiczne do poprzedniego przykładu wyliczenia otrzymujemy:

$$v = \frac{4km + 6km + 10km + 20km + 10km + 30\cancel{km}}{\frac{4\cancel{km}}{90\frac{\cancel{km}}{godz}} + \frac{6\cancel{km}}{100\frac{\cancel{km}}{godz}} + \frac{10\cancel{km}}{110\frac{\cancel{km}}{godz}} + \frac{20\cancel{km}}{120\frac{\cancel{km}}{godz}} + \frac{10\cancel{km}}{115\frac{\cancel{km}}{godz}} + \frac{30\cancel{km}}{80\frac{\cancel{km}}{godz}}} =$$

$$= \frac{80km}{0{,}0444godz + 0{,}06godz + 0{,}0909godz + 0{,}1667godz + 0{,}087 + 0{,}375godz} =$$

$$= \frac{80km}{0{,}824godz} = 97{,}1\,km/godz.$$

W tym przypadku użyto średniej harmonicznej ważonej, wagami zaś były liczby kilometrów (n_i) przejechane z poszczególnymi prędkościami (x_i).

■

Jak już wspomniano, o tym, czy należy zastosować średnią arytmetyczną czy harmoniczną decyduje postać danych, na podstawie których jest ona obliczana. Określenie, którą średnią należy zastosować, bez pewnej wprawy w czytaniu danych może być kłopotliwe. Znacznie wygodniej jest podejść do problemu niejako od końca, czyli obliczyć średnią, biorąc pod uwagę definicję średniej dla badanej cechy, a następnie określić, którą średnią zastosowaliśmy, tak jak w powyższych przykładach.

Przykład 4.8

W pewnym domu okazało się o poranku, że nie ma cukru. Cała rodzina była sfrustrowana, ponieważ wszyscy słodzili i bardzo nie lubili gorzkich płynów. Drobna dyskusja, w nerwowej atmosferze pośpiechu, skończyła się ustaleniem, że cukier kupi wracający najwcześniej ze szkoły syn. Cała rodzina jednak uznała, wracając po kolei do domu, że nie należy do niego mieć zaufania, albowiem jest duże prawdopodobieństwo, że zapomni o kupnie cukru. Wobec tego na wszelki wypadek, wracając do domu, mama kupiła kilogram cukru za 4 zł, babcia kilogram za 3,6 zł, siostra delikwenta kupiła kilogram za 4,4 zł, dziadek kilogram za 4,8 zł i ojciec, który kupił kilogram najdrożej – za 6,4 zł. Syn, który miał wrócić najwcześniej,

wrócił jako ostatni, ponieważ wybrał się z kolegami na mecz i kompletnie zapomniał o cukrze i wcześniejszym powrocie. Pomińmy awanturę, jaka z tego tytułu rozpętała się w domu. Za karę pił gorzką herbatę przez następny tydzień. My zaś policzmy, ile średnio kosztował kilogram zakupionego przez całą rodzinę cukru.

Zacznijmy od definicji średniej ceny $\overline{p}$ [77]:

średnia cena ($\overline{p}$) = wartość towarów (w) / ilość towarów (q).

Policzmy najpierw, ile rodzina zapłaciła łącznie za cukier (czyli wartość towaru – w):

$$w = 1\text{kg} \cdot 4{,}0\ \text{zł/kg} + 1\text{kg} \cdot 3{,}6\ \text{zł/kg} + 1\text{kg} \cdot 4{,}4\ \text{zł/kg} + 1\text{kg} \cdot 4{,}8\ \text{zł/kg} + 1\text{kg} \cdot 6{,}4\ \text{zł/kg} = 23{,}2\ \text{zł}.$$

Ilość zakupionego cukru (q) wyniosła pięć kilogramów, bo pięć osób kupiło po kilogramie cukru, więc q = 5 kg. Zatem średnia cena wyniesie:

$$\overline{p} = \frac{1\cancel{\text{kg}} \cdot 4{,}0\ \text{zł/}\cancel{\text{kg}} + 1\cancel{\text{kg}} \cdot 3{,}6\ \text{zł/}\cancel{\text{kg}} + 1\cancel{\text{kg}} \cdot 4{,}4\ \text{zł/}\cancel{\text{kg}} + 1\cancel{\text{kg}} \cdot 4{,}8\ \text{zł/}\cancel{\text{kg}} + 1\cancel{\text{kg}} \cdot 6{,}4\ \text{zł/}\cancel{\text{kg}}}{5\text{kg}} =$$

$$= \frac{23{,}2}{5}\ \text{zł/kg} = 4{,}64\ \text{zł/kg}.$$

W tym przypadku, jak widać z podstawienia, należało użyć średniej arytmetycznej. Średnia cena cukru zakupionego przez poszczególnych członków rodziny wyniosła 4,64 zł. O tym, że użyjemy średniej arytmetycznej, wiemy już wcześniej, analizując dostępne dane. Nasza cecha „cena jednostkowa" jest ilorazem wartości do ilości, czyli zł/kg, zaś wagą jest ilość kupionych przez poszczególnych członków rodziny kilogramów cukru. Dla uproszczenia obliczeń przyjęto, że każdy kupował po kilogramie i jest to jednostka, która znajduje się w mianowniku jednostki pomiaru cechy (czyli ceny). Stąd wnioskujemy, że można i należy użyć średniej arytmetycznej. Gdyby problem był sformułowany następująco: *obliczyć średnią cenę kupionego cukru, wiedząc, ile każdy z rodziny zapłacił (waga), ale nie wiedząc, ile kilogramów cukru kupił,* wówczas musielibyśmy użyć średniej harmonicznej.

■

Przykład 4.9

Wszyscy w rodzinie (z poprzedniego przykładu), nie tylko lubią pić słodką popołudniową herbatę, ale uwielbiają też razem z tym napojem delektować się świeżo upieczoną szarlotką. Mama zaoferowała, że takową upiecze, pod warunkiem, że zostaną kupione jabłka. W ramach rekompensaty za poprzednie zapominalstwo i późne wracanie ze szkoły syn zdeklarował, że kupi jabłka. Po zajęciach wrócił

[77] Zamiast określenia cechy literą x użyjemy najbardziej popularnego i w ekonomii, i w statystyce symbolu ceny – p, od angielskiego *price*.

z siatką pełną owoców, na które wydał 4 zł. Pamiętał, że cena jabłek wynosiła 3,20 zł/kg. Pozostali członkowie rodziny, pamiętając poprzednie wydarzenia, również kupili jabłka w drodze powrotnej. I tak babcia wydała 5,10 zł (cena jabłek wynosiła 3,00 zł/kg), dziadek zapłacił 3,90 zł (cena 2,60 zł/kg), tata – 3,00zł (cena 3,00 zł/kg), siostra 3,50 zł (cena 2,80 zł/kg), zaś mama 8,50 zł (cena 3,40 zł/kg). Zakupiona ilość jabłek starczyła rodzinie nie na jedną, ale na kilka szarlotek upieczonych przez mamę w kolejnych dniach. Jednak sprawdźmy, jaka była średnia cena kupionych przez rodzinę jabłek.

Pamiętając, że cena jest to wartość towarów podzielona przez ich ilość, najpierw obliczmy, jaką kwotę łącznie wydali członkowie rodziny na jabłka do szarlotki:

$$w = 4{,}00\,\text{zł} + 5{,}10\,\text{zł} + 3{,}90\,\text{zł} + 3{,}00\,\text{zł} + 3{,}50\,\text{zł} + 8{,}50\,\text{zł} = 28{,}00\,\text{zł}\,.$$

Następnie wyznaczmy ilość zakupionych jabłek (ilość oblicza się jako sumę ilorazów wydanej kwoty i ceny jednostkowej):

$$q = \frac{4{,}00\ \text{zł}}{3{,}20\,\frac{\text{zł}}{\text{kg}}} + \frac{5{,}10\ \text{zł}}{3{,}00\,\frac{\text{zł}}{\text{kg}}} + \frac{3{,}90\ \text{zł}}{2{,}60\,\frac{\text{zł}}{\text{kg}}} + \frac{3{,}00\ \text{zł}}{3{,}00\,\frac{\text{zł}}{\text{kg}}} + \frac{3{,}50\ \text{zł}}{2{,}80\,\frac{\text{zł}}{\text{kg}}} + \frac{8{,}50\ \text{zł}}{3{,}40\,\frac{\text{zł}}{\text{kg}}}.$$

Te same wyliczenia otrzymujemy, podstawiając dane dotyczące kwoty wydanej na jabłka i ich ceny jednostkowej do wzoru na średnią harmoniczną:

$$\overline{p} = \frac{w}{q} = \frac{4{,}00\,\text{zł} + 5{,}10\,\text{zł} + 3{,}90\,\text{zł} + 3{,}00\,\text{zł} + 3{,}50\,\text{zł} + 8{,}50\,\text{zł}}{\frac{4{,}00\,\cancel{\text{zł}}}{3{,}20\,\frac{\cancel{\text{zł}}}{\text{kg}}} + \frac{5{,}10\,\cancel{\text{zł}}}{3{,}00\,\frac{\cancel{\text{zł}}}{\text{kg}}} + \frac{3{,}90\,\cancel{\text{zł}}}{2{,}60\,\frac{\cancel{\text{zł}}}{\text{kg}}} + \frac{3{,}00\,\cancel{\text{zł}}}{3{,}00\,\frac{\cancel{\text{zł}}}{\text{kg}}} + \frac{3{,}50\,\cancel{\text{zł}}}{2{,}80\,\frac{\cancel{\text{zł}}}{\text{kg}}} + \frac{8{,}50\,\cancel{\text{zł}}}{3{,}40\,\frac{\cancel{\text{zł}}}{\text{kg}}}} =$$

$$= \frac{28{,}00\ \text{zł}}{1{,}25\ \text{kg} + 1{,}70\ \text{kg} + 1{,}50\ \text{kg} + 1{,}00\ \text{kg} + 1{,}25\ \text{kg} + 2{,}50\ \text{kg}} = \frac{28{,}00\ \text{zł}}{9{,}20\ \text{kg}} = 3{,}04\,\frac{\text{zł}}{\text{kg}}.$$

Zatem średnia cena zakupionych jabłek wynosiła 3,40 zł/kg i jak łatwo zauważyć, użyto tu średniej harmonicznej. Wagami zaś są kwoty wydane na zakup poszczególnych ilości jabłek.

■

Przykład 4.10

Planowane jest połączenie miasta X z sąsiadującym miasteczkiem Y. W mieście X mieszka 100 000 osób, zaś w Y jedynie 5 000. Pierwsza z miejscowości ma gęstość zaludnienia na poziomie 2 500 osób/km^2, zaś druga 1 000 osób/km^2. Jaka będzie średnia gęstość zaludnienia połączonego miasta?

W tym przypadku wskazaniem do wyznaczenia przeciętnej gęstości zaludnienia jest zastosowanie średniej harmonicznej, gdyż nasza cecha „gęstość zaludnienia" jest wskaźnikiem natężenia (osoby/km^2), zaś wagi cechy (liczba ludności w poszczególnych miastach) są wyrażone w jednostkach z licznika cechy (osoby). Podobnie jak poprzednio, zacznijmy od zdefiniowania cechy „gęstość zaludnienia". Jest to, jak wiadomo, liczba ludzi przypadająca na jednostkę powierzchni danego terytorium.

$$\text{Gęstość} = \frac{\text{liczba osób}}{\text{powierzchnia}}.$$

Stosując wzór (4.6), otrzymujemy następującą gęstość zaludnienia:

$$H = \frac{N}{\sum_{i=1}^{k} \frac{n_i}{x_i}} = \frac{100\ 000 \text{ osób} + 5\ 000 \text{ osób}}{\frac{100\ 000 \text{ osób}}{2\ 500 \text{ osób/km}^2} + \frac{5\ 000 \text{ osób}}{1\ 000 \text{ osób/km}^2}} = \frac{105\ 000 \text{ osób}}{40 \text{ km}^2 + 5 \text{ km}^2} = 2333{,}33 \text{ osób/km}^2.$$

Zatem przedstawiony wynik oznacza, że w nowo utworzonej miejscowości średnia gęstość zaludnienia wynosić będzie 2 333 osoby/km^2.

■

Przykład 4.11

Rozważymy podobną sytuację jak w poprzednim przykładzie, ale dotyczącą połączenie miasta A z sąsiadującym miastem B. Miasto A zajmuje powierzchnię 30 km^2, zaś B 50 km^2. Gęstość zaludnienia w pierwszej z miejscowości wynosi 1 500 osób/km^2, a w drugiej 1 000 osób/km^2. Jaka będzie średnia gęstość zaludnienia miasta powstałego z połączenia obu miejscowości oraz włączenia do niej niezaludnionego obszaru 20 km^2 znajdującego się pomiędzy nimi?

W tym przykładzie do wyznaczenia przeciętnej gęstości zaludnienia powinniśmy zastosować średnią arytmetyczną, gdyż cecha gęstość zaludnienia (osoby/km^2) ma wagi wyrażone w jednostkach z mianownika cechy (czyli km^2). Stosujemy więc wzór:

$$\bar{x} = \frac{l.ludn.}{pow.} = \frac{30 \cancel{\text{km}^2} \cdot 1500 \text{ osób}/\cancel{\text{km}^2} + 50 \cancel{\text{km}^2} \cdot 1000 \text{ osób}/\cancel{\text{km}^2} + 20 \cancel{\text{km}^2} \cdot 0 \text{ osób}/\cancel{\text{km}^2}}{30 \text{ km}^2 + 50 \text{ km}^2 + 20 \text{ km}^2} =$$

$$= \frac{45\ 000 \text{ osób} + 50\ 000 \text{ osób} + 0 \text{ osób}}{100 \text{ km}^2} = \frac{95\ 000 \text{ osób}}{100 \text{ km}^2} = 950 \frac{\text{osób}}{\text{km}^2}.$$

W utworzonej miejscowości średnia gęstość zaludnienia będzie wynosić 950 osób/km^2.

■

4.1.3. Szereg czasowy – pojęcia wstępne

Ze względu na to, że średniej geometrycznej najczęściej używa się do analizy szeregów czasowych, zanim ją zdefiniujemy, wprowadzimy najpierw pewne niezbędne pojęcia i definicje związane z szeregami czasowymi. U podstaw analizy dynamiki procesów społeczno-gospodarczych leży konstrukcja szeregów czasowych, zwanych też chronologicznymi**. Szereg czasowy tworzy się w oparciu o momenty lub okresy.** *Pomiar zjawisk w momentach* dotyczy stanu zjawiska w określonym punkcie czasu, zmierzoną wielkość określamy jako *zasób*, np. spis ludności, spis rolny, spis mieszkań, liczba bezrobotnych w województwie zarejestrowanych w konkretnym dniu miesiąca jakiegoś roku. W każdym z takich badań ustala się liczbę jednostek (ludności, zwierząt, mieszkań, bezrobotnych) na konkretną datę. *Pomiar zjawisk w okresach* obejmuje sumę jednostek badanego zjawiska, które wystąpiły na danym obszarze w ciągu określonego czasu, np. roku, kwartału, miesiąca, dnia. Taki pomiar określa się jako pomiar *strumienia*, np. wielkość produkcji wytworzonej w ciągu kwartału, miesięczna sprzedaż w handlu detalicznym, ilość urodzeń w ciągu roku. Wybór jednostki czasu zależy od specyfiki zjawiska i od potrzeb, dla których prowadzi się badanie. Na ogół też pomiarów dokonuje się w tych samych momentach[78] i okresach jednakowej długości[79]. Ułatwia to porównania międzyokresowe.

Teoretyczny szereg czasowy przedstawia się jako ciąg, w którym każdej obserwacji cechy x u danej jednostki statystycznej przypisuje się numer (t) określający kolejność, w jakiej dokonano pomiaru, otrzymując zbiór postaci[80]:

$$x_1, x_2, x_3, \ldots\ldots x_{t-1}, x_t, x_{t+1}, \ldots, x_T,$$

gdzie x_t to poszczególne elementy szeregu czasowego, T – ostatni badany okres, czyli liczebność populacji okresów.

Jeżeli interesuje nas dynamika zjawiska, czyli jak zmieniały się wartości cechy z okresu na okres, wówczas możemy użyć na przykład *stopy wzrostu*[81]:

---

[78] Dwa najbardziej popularne momenty w pomiarach zjawisk społeczno-ekonomicznych to 30 VI i 31 II, podstawowe zaś okresy pomiaru to rok, kwartał, miesiąc na poziomie zarówno makro-, jak i mikroekonomicznym.

[79] Przy pomiarze prowadzonym w oparciu o okresy kalendarzowe należy pamiętać, że odstępy pomiędzy nimi nie są jednakowe. Luty na przykład różni się od stycznia o 3 dni, co stanowi prawie 10% długości stycznia. Tam, gdzie nie ma to większego znaczenia, możemy zaniedbać tę różnicę i stosować jednostkę miesiąc, ale w wielu zastosowaniach należy to korygować. Często, na przykład w bankowości, przyjmowano umownie, że każdy miesiąc ma 30 dni.

[80] Jeżeli interesuje nas pomiar dynamiki zmian stypendium studenta, musimy najpierw zaobserwować, ile wynosiło stypendium studenta w poszczególnych okresach jego studiów (załóżmy, że badamy 3 lata studiów licencjackich, co daje sześć obserwowanych semestrów). Otrzymamy wówczas np. szereg postaci: x_1=400 zł, x_2=400 zł, x_3=440 zł, x_4=462 zł, x_5=346,50 zł, x_6=415,80 zł, czyli T=6.

[81] Jest to podstawowa miara dynamiki procesów w czasie, choć nie jedyna.

$$I_{t/t-i} = \frac{x_t}{x_{t-i}},$$

gdzie:

x_t – poziom zjawiska w momencie *t*,

x_{t-i} – poziom zjawiska w momencie *t-i* opóźnionym o *i* okresów w stosunku do *t*.

Stopa wzrostu mówi nam, ile razy poziom cechy w momencie *t* wzrósł w stosunku do momentu (*t-i*) (punktu odniesienia) lub jaki procent wartości poprzedniej stanowi wartość w momencie *t*.

Ze stopą wzrostu jest ściśle związane tempo wzrostu.

Mówi ono nam, o ile coś przyrosło pomiędzy momentem *t* i (*t-i*).

$$r = \frac{x_t - x_{t-i}}{x_{t-i}} = \frac{x_t}{x_{t-i}} - \frac{x_{t-i}}{x_{t-i}} = \frac{x_t}{x_{t-i}} - 1 = I_{t/t-i} - 1$$

lub jeśli dane są w ujęciu procentowym, $r = I_{t/t-i} - 100$.

Przykład 4.12

Student Wydziału Zarządzania otrzymywał przez sześć semestrów stypendium naukowe wynoszące kolejno: $x_1 = 400$ zł, $x_2 = 400$ zł, $x_3 = 440$ zł, $x_4 = 460$ zł, $x_5 = 350$ zł, $x_6 = 420$ zł. Należy wyznaczyć, jak zmieniała się kwota stypendium.

Kolejne stopy wzrostu będą wynosiły odpowiednio:

$$I_{2/1} = \frac{x_2}{x_1} = \frac{400 \text{ zł}}{400 \text{ zł}} = 1;$$

$$I_{3/2} = \frac{x_3}{x_2} = \frac{440 \text{ zł}}{400 \text{ zł}} = 1,1;$$

$$I_{4/3} = \frac{x_4}{x_3} = \frac{460 \text{ zł}}{440 \text{ zł}} = 1,0455;$$

$$I_{5/4} = \frac{x_5}{x_4} = \frac{350 \text{ zł}}{460 \text{ zł}} = 0,7609;$$

$$I_{6/5} = \frac{x_6}{x_5} = \frac{420 \text{ zł}}{350 \text{ zł}} = 1,2.$$

Zastosowano tu jeden ze sposobów liczenia stopy wzrostu, czyli tzw. łańcuchową stopę wzrostu. Porównujemy tu okres bieżący do poprzedniego (czyli $i = 1$), co oznacza, że porównujemy zmiany stypendium każdorazowo w stosunku

do poprzedniego semestru. Zauważmy też, że na podstawie $T = 6$ obserwacji wartości stypendium skonstruowaliśmy $n = 5$ miar dynamiki, czyli szereg stóp wzrostu liczy teraz pięć obserwacji. Otrzymane miary interpretujemy następująco: stypendium w drugim semestrze w stosunku do pierwszego nie wzrosło ($I_{2/1} = 1$), w trzecim semestrze w stosunku do drugiego stypendium wzrosło $I_{3/2} = 1{,}1$ raza lub możemy powiedzieć, że stanowi 110% stypendium z poprzedniego okresu (lub 1,1 stypendium poprzedniego okresu). Analogicznie stypendium w czwartym semestrze wzrosło 1,0455 raza lub wynosiło 104,55% stypendium z okresu poprzedniego. W semestrze piątym stypendium spadło i wynosi 0,7609 lub 76,09% stypendium z semestru poprzedniego. W szóstym okresie stypendium znów relatywnie wzrosło 1,2 raza lub stanowiło 120% tego, co student otrzymywał w poprzednim semestrze (ale uwaga – wzrost tylko w stosunku do semestru piątego).

Tempa wzrostu najczęściej podaje się w ujęciu procentowym, co jest wygodniejsze w interpretacji, wobec tego tempa wzrostu stypendium w powyższym przykładzie będą wynosiły odpowiednio:

$r_1 = I_{2/1} - 1 = 1 - 1 = 0$ (stypendium nie zwiększyło się);

$r_2 = I_{3/2} - 1 = 1{,}1 - 1 = 0{,}1$ lub $r_2 = 110\% - 100 = 10\%$ (stypendium wzrosło o 10% w stosunku do poprzedniego okresu);

$r_3 = I_{4/3} - 1 = 1{,}0455 - 1 = 0{,}0455$ lub $r_3 = I_{4/3}\% - 100\% = 104{,}55\% - 100\% = 4{,}55\%$;

$r_4 = I_{5/4} - 1 = 0{,}7609 - 1 = 0{,}2391$ lub $r_4 = I_{5/4}\% - 100\% = 76{,}09\% - 100\% = -23{,}91\%$ (stypendium zmalało o 23,91% w stosunku do poprzedniego okresu);

$r_5 = I_{6/5} - 1 = 1 - 1{,}2 = 0{,}2$ lub $r_5 = I_{6/5}\% - 100\% = 120\% - 100\% = 20\%$ (stypendium wzrosło o 20% w stosunku do poprzedniego okresu).

4.1.4. Średnia geometryczna

Średnia geometryczna, tak jak poprzednio prezentowana średnia arytmetyczna, mówi nam, ile łącznego funduszu cechy przypada na jednostkę statystyczną zbiorowości, aczkolwiek w tym przypadku fundusz cechy ma inną definicję. Jeśli używamy średniej geometrycznej do obliczenia średniej stopy (lub tempa) wzrostu, wówczas mówi nam, jaką stopę wzrostu (lub tempo) musiałoby mieć badane zjawisko, by osiągnąć poziom okresu końcowego, rosnąc w kolejnych okresach tak samo. Funduszem cechy jest łączny iloczyn stóp (temp) wzrostu w badanym okresie.

Stosuje się ją w dwóch przypadkach:

1) gdy występują znaczne różnice między obserwowanymi wartościami cechy[82], gdyż jest mniej wrażliwa na wartości ekstremalne (nietypowe) niż średnia arytmetyczna, *można więc ją stosować zamiast średniej arytmetycznej*[83];
2) przede wszystkim jednak jest *używana do analizy zmiennych multiplikatywnych*[84], *zwłaszcza do danych w postaci liczb względnych*, np. stóp i temp wzrostu – jest to powszechnie używana miara dynamiki procesów.

Średnia geometryczna prosta (nieważona) wyznaczana jest z następującego wzoru:

$$\overline{x}_g = \sqrt[n]{x_1 \cdot x_2 \ldots \cdot x_i \ldots \cdot x_n}, \tag{4.9}$$

gdzie x_i – dane w postaci liczb bezwzględnych lub stóp wzrostu ($i = 1, 2, \ldots, n$).

Jeżeli poszczególne obserwacje występują więcej niż jeden raz, korzysta się, jak w przypadku poprzednio omawianych średnich, z wag. Wówczas wzór (4.9) ma następującą postać :

$$\overline{x}_g = \sqrt[n]{\prod_{i=1}^{k} x_i^{n_i}}, \tag{4.10}$$

gdzie x_i – jak wyżej, zaś n_i – częstość występowania odmiany x_i, czyli waga i $\sum_{i=1}^{k} n_i = n.$

Jeżeli dane są w postaci liczb bezwzględnych, a nas interesuje średnia stopa wzrostu, wówczas można zastosować następujący wzór łączący obliczanie stóp wzrostu i liczenie średniej:

---

[82] W szeregu są tak zwane wartości odstające, nietypowe, np. mamy sto osób o wzroście pomiędzy 1,6 m a 1,9 m oraz jedną o wzroście 2,1 m. Ta ostatnia osoba zostanie uznana za nietypową w badanej populacji. Uwzględnienie jej wzrostu wpłynie na podwyższenie średniej. Jeżeli jest to jednostka spoza badanej populacji (np. badamy wzrost polskich studentów, a zmierzono także wzrost studiującego w Polsce masajskiego studenta) lub błąd w pomiarze, jej uwzględnienie „zafałszuje" prawdziwą średnią i należałoby tę obserwację odrzucić. Z drugiej strony, jeśli odrzucimy tę jednostkę, a jest to poprawny pomiar i jednostka pochodzi z badanej populacji, wówczas popełnimy błąd, odrzucając taką obserwację.

[83] Interpretujemy ją analogicznie do średniej arytmetycznej: rozdziela ona równomiernie na wszystkie jednostki populacji fundusz cechy będący w tym przypadku iloczynem wszystkich wartości cechy.

[84] Zmienna multiplikatywna to zmienna, która dla danego zbioru elementów ma wartość funduszu cechy równy iloczynowi wartości przypisanych wszystkim elementom składowym zbioru. Średniej arytmetycznej używamy do zmiennych addytywnych, czyli takich, których wartość funduszu cechy jest sumą wartości cechy poszczególnych jednostek w zbiorze.

$$\overline{x}_g = \sqrt[n-1]{\frac{x_2}{x_1} \cdot \frac{x_3}{x_2} \cdot \ldots \cdot \frac{x_n}{x_{n-1}}}, \tag{4.11}$$

gdzie:

x_t – wartości analizowanej cechy w kolejnych jednostkach czasu t =1, 2, …, n.

Zauważmy, że powyższy wzór sprowadza się do postaci:

$$\overline{x}_g = \sqrt[n-1]{\frac{\cancel{x_2}}{x_1} \cdot \frac{\cancel{x_3}}{\cancel{x_2}} \cdot \ldots \cdot \frac{x_n}{\cancel{x_{n-1}}}} = \sqrt[n-1]{\frac{x_n}{x_1}}. \tag{4.11a}$$

Obliczanie stopy wzrostu za pomocą średniej geometrycznej na podstawie szeregu wartości bezwzględnych sprowadza się do obliczenia stopy wzrostu pomiędzy okresem pierwszym i ostatnim oraz wyciągnięcia pierwiastka stopnia równego ilości elementów mnożonych pod pierwiastkiem. Jeśli wyjściowo mamy n okresów, to na tej podstawie tworzymy (n-1) stóp wzrostu i z tego stopnia jest wyciągany pierwiastek.

Przykład 4.13 (kontynuacja przykładu 4.12)

W poprzednim przykładzie obliczyliśmy stopy wzrostu stypendium studenta w ciągu trzech lat. Ustalmy teraz, jaka była średnia stopa wzrostu tegoż stypendium, czyli jak szybko ono zmieniało się.

Mamy następujące stopy wzrostu: $I_{2/1} = 1$; $I_{3/2} = 1{,}1$; $I_{4/3} = 1{,}0455$; $I_{5/4} = 0{,}7609$; $I_{6/5} = 1{,}2$. Ponieważ wartości cechy są miarami dynamiki, musimy użyć średniej geometrycznej w postaci wzoru (4.11):

$$\overline{x}_g = \sqrt[n-1]{\frac{x_2}{x_1} \cdot \frac{x_3}{x_2} \cdot \ldots \cdot \frac{x_n}{x_{n-1}}} = \sqrt[n-1]{I_{2/1} \cdot I_{3/2} \cdot I_{4/3} \cdot I_{5/4} \cdot I_{6/5}} = \sqrt[6-1]{1 \cdot 1{,}1 \cdot 1{,}0455 \cdot 0{,}7609 \cdot 1{,}2} =$$

$$= \sqrt[5]{1{,}0501} = 1{,}0070.$$

Zauważmy, że to samo uzyskamy, gdy użyjemy danych wyjściowych w postaci poziomów stypendium:

$$\overline{x}_g = \sqrt[6-1]{\frac{x_2}{x_1} \cdot \frac{x_3}{x_2} \cdot \frac{x_4}{x_3} \cdot \frac{x_5}{x_4} \cdot \frac{x_6}{x_5}} = \sqrt[6-1]{\frac{400}{400} \cdot \frac{440}{400} \cdot \frac{460}{440} \cdot \frac{350}{460} \cdot \frac{420}{350}} = \sqrt[6-1]{\frac{420}{400}} =$$

$$\sqrt[6-1]{\frac{420}{400}} = \sqrt[5]{1{,}05} = 1{,}007.$$

Możemy więc powiedzieć, że stypendium studenta w tym przypadku rosło średnio co semestr przy stopie wzrostu 100,7% i aby nie łamać sobie języka, pro-

ściej jest powiedzieć, że średnio stypendium wzrastało o 0,7% (co wyliczyliśmy jako 100,7%-100%).

■

Podsumowując przedstawione w niniejszym podrozdziale podstawy wyznaczania średnich klasycznych, należy podkreślić, że są one miarami służącymi do sumarycznej charakterystyki rozkładu cechy w zbiorowości statystycznej. Swoje zadanie spełniają jednak tylko wtedy, gdy zbiorowość jest jednorodna pod względem badanej cechy, czyli jednostki zbiorowości skupiają się wokół przeciętnej. Jednorodność zbiorowości wynika z tego, że na wszystkie jednostki działają te same przyczyny główne. Wówczas średnie reprezentują w przybliżeniu działanie składnika systematycznego kształtującego rozkład wartości analizowanej cechy w badanej zbiorowości. Odchylenia cechy od poziomu średniego u poszczególnych jednostek zaś wynikają z działania przyczyn przypadkowych, ubocznych (losowych). Posługiwanie się średnią w zbiorowości niejednorodnej jest błędem, ponieważ średnia w tym przypadku da fałszywy obraz rzeczywistej struktury zbiorowości i może być podstawą błędnych wniosków co do przyczyn kształtujących rozkład danej cechy[85].

Przykład 4.14

Weźmy do naszych rozważań zbiorowość pracowników niewielkiego przedsiębiorstwa: dziesięciu szeregowych pracowników zarabia pomiędzy 600 zł a 1500 zł oraz dwoje (szef i księgowa) zarabiają po 5000 zł. Średnia płaca liczona dla wszystkich nie będzie dobrą miarą tego, jak wyglądają płace w tej firmie, ponieważ mamy do czynienia z małą, niejednorodną płacowo zbiorowością. Średnia dla całej zbiorowości wyniesie znacznie ponad 2000, gdy tymczasem szeregowy pracownik może liczyć na nie więcej niż 1500 zł (np. przy przyjmowaniu do pracy). W tym przykładzie mamy dwie odrębne grupy płacowe i dla każdej z nich należy policzyć średnią niezależnie. Jedna średnia będzie charakteryzowała poprawnie rozkład płac pracowników szeregowych, druga kierownictwa.

■

Przykład 4.15

Wróćmy do przykładu dziadka, który zabrał dwójkę swoich wnuków na spacer do zoo i po drodze spotkali zapalonego statystyka-amatora. Ten po nieudanej próbie analizy wzrostu z równie wielkim zapałem postanowił pytać przechodniów o wiek. Nasi spacerowicze podali następujące wartości: 62, 4, 6. Statystyk-amator ponownie policzył średnią, która wyniosła 24 lata. Czy możemy na podstawie tak

[85] Lange (1975) podaje interesujący przykład na ten temat (zacieranie różnic klasowych), s. 156 i nast.

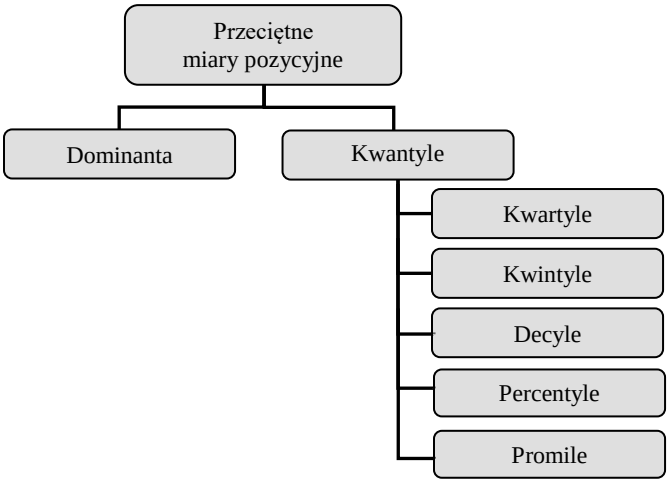
policzonej średniej wysnuć jakieś wnioski o populacji odwiedzających zoo? Nie. Wykonana analiza ponownie potwierdziła amatorstwo naszego statystyka, gdyż badana populacja jest zbyt mało liczna. Chcąc określić średnią wieku osób odwiedzających zoo, statystyk-amator powinien przebadać znacznie większą liczbę osób.

■

4.2. Miary pozycyjne

Przeciętne pozycyjne *są to parametry, których wartość jest związana z pozycją (miejscem) zajmowanym w zbiorowości.*

Rodzaje tych miar przedstawia schemat na rysunku 4.3. Wyodrębniamy dwa rodzaje miar pozycyjnych: dominantę i kwantyle.



Rysunek 4.2. Podział przeciętnych miar pozycyjnych

Źródło: opracowanie własne.

Sposób ich liczenia i interpretację zaczniemy od omówienia dominanty.

4.2.1. Dominanta

Dominanta jest to wartość cechy najczęściej występująca w badanej zbiorowości. Dominanta służy poprawnie jako miara przeciętna w rozkładach o wyraźnej tendencji do skupiania się jednostek zbiorowości wokół jakiejś jednej wartości cechy[86]. Nie może więc występować w skrajnych przedziałach. O takich szeregach (rozkładach), w których występuje jeden punkt skupienia, czyli dominanta, mówimy, że są jednomodalne[87].

W szeregach cechy skokowej wyznacza się ją bardzo prosto jako cechę, która jest reprezentowana najliczniej lub z największą częstotliwością. W przypadku cechy ciągłej problem jest bardziej skomplikowany. Przedstawmy więc najpierw przykłady określania dominanty w rozkładach cech skokowych.

Przykład 4.12

Przebadano łącznie sto gospodarstw domowych w Białymstoku, pytając o liczbę telefonów (odrębnych numerów) w rodzinie. Dane dotyczące badania zostały przedstawione w tabeli 4.6.

Tabela 4.6. Liczba aparatów telefonicznych w rodzinie

Liczba aparatów telefonicznych x_i	Liczba gospodarstw domowych n_i
0	5
1	10
2	20
3	45
4	20

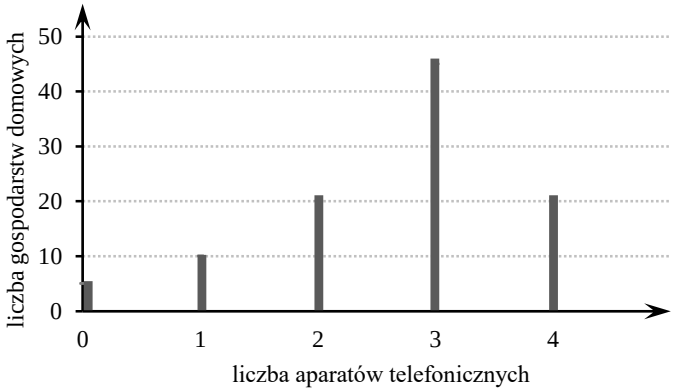
Źródło: dane umowne.

Wyznaczymy dominantę liczby aparatów telefonicznych w badanych gospodarstwach Białegostoku.

Analizując dane zawarte w tabeli 4.6, widać, że najczęściej występującą liczbą w badanej zbiorowości gospodarstw są trzy telefony. Łatwo to zauważyć, jeżeli spojrzy się na wykres 4.2 otrzymany z uzyskanych danych.

[86] Podobnie jak średnia arytmetyczna.

[87] Czyli jednodominantowe, jest to polski odpowiednik angielskiej nazwy dominanty *mode* wprowadzonej przez angielskiego statystyka K. Pearsona w 1895 r.



Wykres 4.2. Liczba aparatów telefonicznych w gospodarstwach domowych

Źródło: opracowanie własne.

Analizując wykres 4.2, widać, że badany szereg ma jeden, bardzo wyraźny, szczyt. Właśnie o tak wyglądającym szeregu mówi się, że jest jednomodalny.

■

Kolejny przykład zilustruje problem związany z wyznaczaniem dominanty w przypadku szeregu, który nie jest jednomodalny.

Przykład 4.13

W tablicy 4.7 przedstawiono rozkład liczby osób w gospodarstwach domowych w Polsce w 1998 roku.

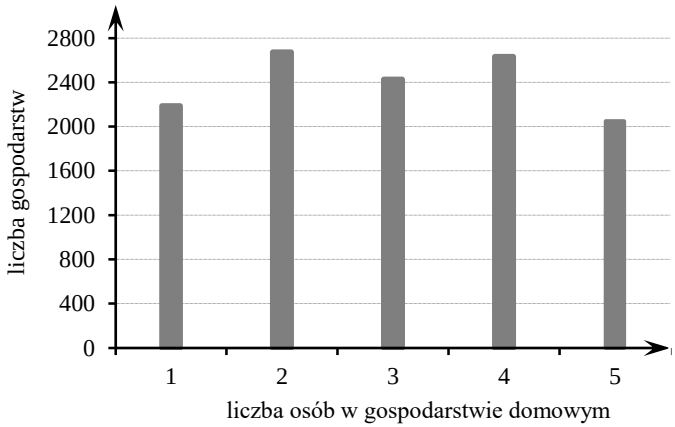
Tabela 4.7. Gospodarstwa domowe w Polsce w 1998 roku

Liczba osób w gospodarstwie x_i	Liczba gospodarstw domowych [w tys.] n_i
1	2188
2	2673
3	2427
4	2632
5 i więcej	2050

Źródło: *Mały rocznik statystyczny Polski 2002*, tabl. 9(71), s. 118.

Korzystając z przedstawionych danych, należy wyznaczyć dominantę liczebności gospodarstw domowych w Polsce w 1998 roku.

Na podstawie definicji dominanty można by stwierdzić, że w Polsce w 1998 roku dominowały gospodarstwa domowe dwuosobowe. Jednak zanim sformułuje się taki wniosek, należy przyjrzeć się wykresowi 4.3 przedstawiającemu ten szereg.



Wykres 4.3. Gospodarstwa domowe w Polsce w 1998 roku (w tys.)

Źródło: Obliczenia własne na podstawie danych z *Mały rocznik statystyczny Polski 2002*, tabl. 9(71), s. 118.

Poszczególne odmiany cechy występują z bardzo podobnymi liczebnościami (częstościami). Wyraźnie widać, że szereg ten nie jest jednoznacznie szeregiem jednomodalnym. Jest to przykład szeregu (rozkładu cechy), w którym nie ma zdecydowanie jednej wartości dominującej. Liczebności poszczególnych odmian cechy są, praktycznie rzecz biorąc, rozłożone równomiernie. Jednostki zbiorowości (tu: gospodarstwa domowe) nie wykazują tendencji do skupiania się wokół jednej wartości cechy. Do opisu takiego szeregu nie należy więc używać dominanty, albowiem tu jej tak naprawdę nie ma. Zauważmy także, iż liczenie średniej arytmetycznej jako punktu, wokół którego skupiają się jednostki statystyczne, także nie ma racji bytu, ponieważ nie ma tu takiego punktu, a więc choć średnia jest możliwa do policzenia, to nie będzie ona poprawnie reprezentowała zbiorowości.

■

W przypadku cechy ciągłej nie można podać, która odmiana cechy występuje najczęściej, a jedynie można określić dominujący przedział. Jest to przedział wartości cechy o największej liczebności lub częstości. Należy jednak pamiętać o bardzo ważnym założeniu, że **szereg jest podzielony na klasy o jednakowej rozpiętości**. Jeśli mamy przedziały o różnej rozpiętości, wówczas duża lub mała liczebność przedziału może wynikać z jego rozpiętości. Im szerszy jest przedział, tym więcej jednostek statystycznych może się w nim „zmieścić" i stąd może wynikać dominacja danego przedziału. W takim przypadku nie możemy wyznaczyć przedziału dominującego. Jak sobie poradzić z taką sytuacją, powiemy po omówieniu sposobu liczenia dominanty w szeregach o równej rozpiętości przedziałów.

Bardzo często interesuje nas nie dominanta w formie przedziału liczbowego, a jej punktowa, konkretna wartość. Możemy ją wówczas obliczyć, przy pewnych założeniach, ze wzoru interpolacyjnego przedstawionego poniżej:

$$D = x_{0D} + \frac{n_D - n_{D-1}}{(n_D - n_{D-1}) + (n_D - n_{D+1})} \cdot h_D, \tag{4.17}$$

gdzie:
x_{0D} – dolna granica przedziału dominanty,
n_D – liczebność przedziału dominanty,
n_{D-1} – liczebność przedziału poprzedzającego przedział dominanty,
n_{D+1} – liczebność przedziału następnego po przedziale dominanty,
h_D – rozpiętość przedziału dominanty.

Wzór interpolacyjny[88] jest skonstruowany przy założeniu, że przedział dominanty oraz dwa sąsiednie brane pod uwagę przedziały mają jednakową rozpiętość. Ponieważ wartość dominanty nie zależy od absolutnych liczebności, lecz od ich proporcji, z powyższego wzoru także możemy skorzystać, gdy zamiast liczebności mamy częstości w poszczególnych klasach:

$$D = x_{0D} + \frac{f_D - f_{D-1}}{(f_D - f_{D-1}) + (f_D - f_{D+1})} \cdot h_D, \tag{4.18}$$

gdzie:
f_D – częstość przedziału dominanty,
f_{D-1} – częstość przedziału poprzedzającego przedział dominanty,
f_{D+1} – częstość przedziału następnego po przedziale dominanty.

Przykład 4.14 (kontynuacja przykładu 3.3)

Wróćmy do przykładu 3.3 i wyznaczmy dominującą wagę studentek. Rozkład wagi przypominamy w tabeli 4.8.

[88] Konstrukcja wzoru interpolacyjnego oparta jest na założeniu, że dominanta jest liczona jako maksimum funkcji kwadratowej. Stąd wynikają pozostałe założenia i wnioski, będące pochodnymi sposobu konstrukcji wzoru: konieczność równej rozpiętości przedziałów czy niemożność liczenia dominanty, gdy najliczniejszy jest któryś z przedziałów skrajnych. W tym ostatnim przypadku nie jest spełnione matematyczne założenie o dominancie jako maksimum funkcji kwadratowej. Rozkład, w którym dominanta jest w skrajnym przedziale, reprezentuje matematycznie funkcję monotoniczną (czyli o wartościach stale niemalejących lub nierosnących).

Tabela 4.8. Rozkład wagi studentek kierunku zarządzanie

$\lfloor x_{i_0} - x_{i_1} \rfloor$	n_i	f_i
42-46	2	0,02
46-50	16	0,16
50-54	26	0,26
54-58	30	0,30
58-62	19	0,19
62-66	5	0,05
66-70	2	0,02
Suma	100	1,00

Źródło: opracowanie własne.

Biorąc pod uwagę dane przedstawione w tabeli 4.8, należy *w pierwszym kroku sprawdzić, czy spełnione są założenia wymagane przy wyznaczeniu dominanty, tj. równa rozpiętość przedziałów, jedno maksimum w rozkładzie (jednomodalność) i niewystępowanie dominanty w skrajnym przedziale*[89].

Wszystkie te warunki są spełnione (stała rozpiętość przedziałów wynosząca 4 kg, liczebności systematycznie rosną przy zwiększaniu wagi do pewnego przedziału w środku, a następnie sukcesywnie maleją, co sugeruje jednomodalność). Zatem można zastosować wzór interpolacyjny (4.17) do wyznaczenia dominanty:

$$D = x_{0D} + \frac{n_D - n_{D-1}}{(n_D - n_{D-1}) + (n_D - n_{D+1})} \cdot h_D = 54 + \frac{30 - 26}{(30 - 26) + (30 - 19)} \cdot 4 =$$

$$= 54 + \frac{4}{15} \cdot 4 = 54 + 1{,}067 = 55{,}067 = 55{,}1 \text{ kg.}$$

Wykorzystując częstość (wzór 4.18), otrzymuje się identyczną wartość:

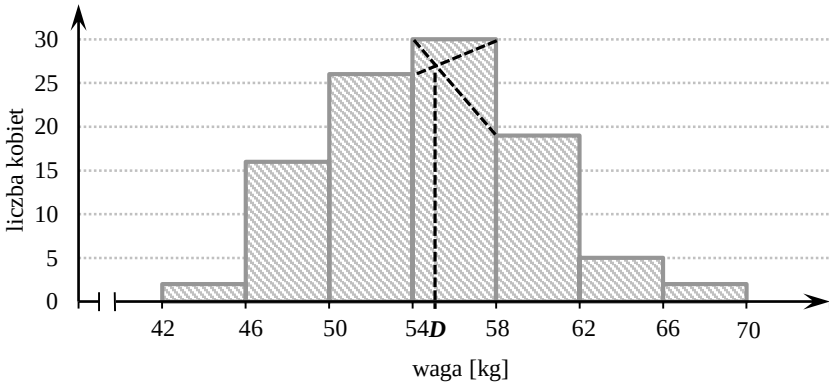
$$D = x_{0D} + \frac{f_D - f_{D-1}}{(f_D - f_{D-1}) + (f_D - f_{D+1})} \cdot h_D = 54 + \frac{0{,}30 - 0{,}26}{(0{,}30 - 0{,}26) + (0{,}30 - 0{,}19)} \cdot 4 =$$

$$= 54 + \frac{0{,}04}{0{,}15} \cdot 4 = 54 + 1{,}067 = 55{,}067 = 55{,}1 \text{ kg.}$$

[89] Jest to oznaką zjawiska zwanego skrajną asymetrią rozkładu, o czym będziemy mówić w jednym z dalszych rozdziałów.

Zatem dominantą, czyli wagą, wokół której skupiały się wagi badanych dziewcząt, było w zaokrągleniu[90] 55 kg. Jednak jest to pewne przybliżenie, bo gdyby wrócić do zebranych danych wyjściowych (przykład 3.3), to można by zauważyć, że waga 55 kg wystąpiła jedynie 7 razy i nie była to wartość najliczniejsza. Do najczęściej występujących należały dwie wartości: 53 i 57, które pojawiły się aż 11 razy. Zatem, jak już wspomniano powyżej, wartości wyznaczane ze wzoru 4.17 czy 4.18 są pewnym przybliżeniem (interpolacją) tego, co dzieje się w zbiorowości, są zwięzłą formą opisu zbiorowości, a nie wartością, która rzeczywiście wystąpiła najwięcej razy. Często może się wręcz zdarzyć, że otrzymana liczba w danym szeregu zupełnie nie wystąpiła, o czym już wspominaliśmy.

Oprócz wyznaczania dominanty z wzorów 4.17 czy 4.18 identyczny rezultat otrzymamy, stosując interpolację graficzną. Została ona przedstawiona na wykresie 4.4.



Wykres 4.4. Graficzne wyznaczenie dominanty

Źródło: opracowanie własne.

Graficzne wyznaczanie dominanty odbywa się etapami, z których pierwszym jest wskazanie najwyższego „słupka", czyli najliczniejszego przedziału cechy, przy jednakowych rozpiętościach przedziałów. Wewnątrz wybranego słupka łączy się wewnętrzne kąty w jego górnej części z miejscem styku przeciwległych boków z odpowiednio kolejnym i poprzednim słupkiem. Otrzymany w ten sposób punkt przecięcia rzutuje się na oś OX i otrzymaną na niej wartość odczytuje. Jest to właśnie poszukiwana wartość dominanty. Jak łatwo zauważyć, wartość zaznaczona na osi X wykresu 4.4 odpowiada wyliczonej na podstawie wzoru wartości dominanty $D = 55$ kg.

■

[90] Proszę zauważyć, iż pomiar był z dokładnością do kilograma stąd wynik zaokrąglamy do pełnych kilogramów.

Jak już wcześniej powiedziano, w przypadku nierównych przedziałów nie możemy bezpośrednio podać dominującego przedziału cechy, ponieważ występująca tu największa liczebność może wynikać z większej rozpiętości przedziału. Do wyznaczenia najliczniejszego przedziału użyjemy więc nie liczebności, a natężenia liczebności[91]. Najpierw należy obliczyć natężenia liczebności w poszczególnych klasach cechy, określić przedział dominanty, a następnie skorzystać z poniższego wzoru interpolacyjnego:

$$D = x_{0D} + \frac{k_D - k_{D-1}}{(k_D - k_{D-1}) + (k_D - k_{D+1})} \cdot h_D, \tag{4.19}$$

gdzie $k_i = n_i : \frac{h_i}{h_{\min}} = \frac{n_i h_{\min}}{h_i}$ oznacza gęstość *i*-tego przedziału, przy czym $h_{\min}$ – jest szerokością najmniejszego przedziału – traktujemy go jako szerokość jednostkową i przeliczamy wszystkie przedziały na jednostkę przedziału minimalnego. Czyli *de facto* stwierdzamy, ile razy dany przedział jest większy od minimalnego. Dzieląc przezeń liczebności, otrzymujemy zagęszczenie jednostek statystycznych przypadające na wzorcowy (jednostkowy) przedział. Jako punktu odniesienia możemy użyć także przedziału maksymalnego – wynik końcowy, czyli przedział dominujący, będzie ten sam.

Przykład 4.15

W tabeli 4.9 przedstawiono rozkład dochodów osób korzystających z pewnego biura turystycznego:

Tabela 4.9. Grupy dochodowe klientów biura turystycznego

Grupy dochodowe (w tys. zł)	Liczba klientów biura turystycznego
do 1 000	25
1 000-2 000	59
2 000-5 000	126
5 000-10 000	82
10 000-15 000	26
15 000-25 000	12
25 000 i więcej	1

Źródło: dane umowne.

[91] Można powiedzieć, że liczymy coś w rodzaju gęstości zaludnienia przedziału jednostkami statystycznymi. Najgęściej „zaludniony" przedział będzie przedziałem dominującym.

Jakie dochody dominowały wśród klientów tego biura turystycznego?

Aby wyznaczyć dominantę, należy w pierwszym kroku sprawdzić, czy podany szereg ma równe rozpiętości przedziałów. W tym przypadku warunek ten nie jest spełniony, podane liczebności należy więc przekształcić na natężenia liczebności (gęstość). W tym celu znajdujemy przedział o najmniejszej rozpiętości, w tym przypadku wynoszący 1 000 zł[92], a następnie wyznaczamy ilorazy bieżącej rozpiętości i minimalnej $\left(h_i / h_{\min}\right)$. Potem dzielimy liczebności w przedziałach przez iloraz rozpiętości, otrzymując natężenie liczebności będące podstawą wyboru przedziału dominanty. Obliczenia te przedstawiono w tabeli 4.10.

Tabela 4.10. Grupy dochodowe klientów biura turystycznego – wyznaczenie natężenia liczebności przedziałów

$x_{i_0} - x_{i_1}$	n_i	h_i	$\frac{h_i}{h_{\min}}$	$k_i = n_i : \frac{h_i}{h_{\min}}$
0-1 000	25	1 000	1	25,0
1 000-2 000	**59**	**1 000**	**1**	**59,0**
2 000-5 000	129	3 000	3	43,0
5 000-10 000	82	5 000	5	16,4
10 000-15 000	26	5 000	5	5,2
15 000-25 000	12	10 000	10	1,2
25 000 i więcej	1	?	?	?

Źródło: opracowanie własne.

Uwzględniając jedynie liczebności i nie biorąc pod uwagę różnych rozpiętości poszczególnych przedziałów, można dojść do mylnego wniosku, że dominanta w tym przykładzie znajduje się w trzecim przedziale. Jednak po wyznaczeniu natężenia liczebności (gęstości) widzimy, że najgęściej „zaludniony" jednostkami statystycznymi jest przedział drugi. Oznacza to, że to on dominuje i dominanta znajduje się w drugim przedziale o gęstości $k_i = 59$. Zatem, podstawiając do wzoru (4.19), otrzymujemy:

$$D = x_{0D} + \frac{k_D - k_{D-1}}{(k_D - k_{D-1}) + (k_D - k_{D+1})} \cdot h_D = 1000 + \frac{59 - 25}{(59 - 25) + (59 - 43)} \cdot 1000 =$$

$$= 1\,000 + \frac{34}{50} 1\,000 = 1\,680 \text{ zł.}$$

W badanej zbiorowości dominowały zatem dochody w wysokości 1680 zł.

■

[92] Zauważmy, że mamy otwarty przedział dolny, ale jak wiadomo, dochody nie mogą być ujemne, więc za dolną granicę przyjęto dochód zerowy. W przypadku górnego przedziału otwartego, ponieważ jest on najmniej liczny, a zdecydowanie szerszy od minimalnego, spokojnie możemy go wykluczyć jako przedział dominanty i zignorować fakt, że jest otwarty.

Przykład 4.16

Badano czas oczekiwania na wydanie posiłku w jednej z restauracji. Od 10 do 12 minut czekało 10 osób, natomiast od 14 do 16 minut 12 klientów. Wyliczono, że dominanta oczekiwania wynosiła 13,2 minut. Ile osób czekało na danie od 12 do 14 minut?

W zadaniu pojawia się informacja o najczęstszym czasie oczekiwania, zatem należy tu wykorzystać wzór 4.17, za pomocą którego wyznacza się dominantę. Podstawiając podane w treści zadania dane i dokonując na nich niewielkich przekształceń, można wyznaczyć liczbę osób czekających od 12 do 14 minut, czyli liczebność przedziału dominanty (n_D).

$$\left.\begin{array}{l} D = 13{,}2 \\ x_{0D} = 12 \\ n_{D-1} = 10 \\ n_{D+1} = 12 \\ h_D = 14 - 12 = 2 \end{array}\right\} \Rightarrow D = x_{0D} + \frac{n_D - n_{D-1}}{(n_D - n_{D-1}) + (n_D - n_{D+1})} \cdot h_D \Rightarrow$$

$$\Rightarrow 13{,}2 = 12 + \frac{n_D - 10}{(n_D - 10) + (n_D - 12)} \cdot 2 \Rightarrow n_D = 16.$$

Z przedstawionych powyżej obliczeń wynika, że od 12 do 14 minut czekało na danie 16 klientów restauracji.

■

Przykład 4.17

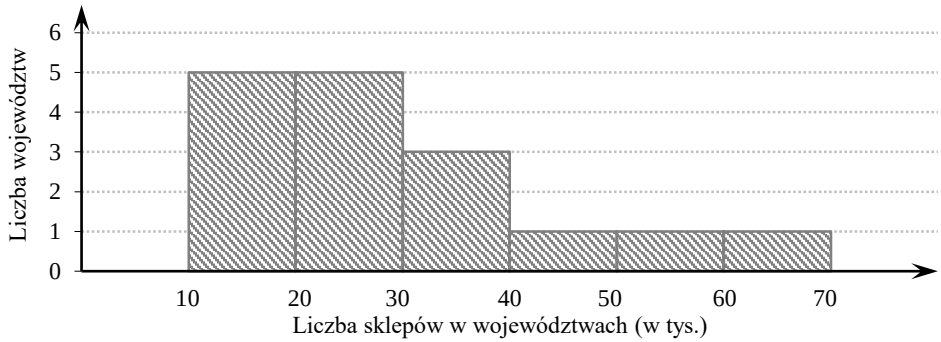
W tabeli 4.11 przedstawiono liczbę sklepów w Polsce według województw w 2002 roku. Należy podać dominującą liczbę sklepów.

Tabela 4.11. Liczba sklepów według województw w 2002 roku w Polsce

Liczba sklepów w województwach (w tys.)	Liczba województw
10-20	5
20-30	5
30-40	3
40-50	1
50-60	1
60-70	1

Źródło: opracowanie własne.

W powyższym szeregu nie można wyznaczyć dominanty, gdyż jest on skrajnie asymetryczny. W dwóch pierwszych przedziałach występują największe i identyczne liczebności, a w następnych liczebności monotonicznie maleją. Również fakt, iż jest to zbiorowość o małej liczebności (w Polsce jest jedynie 16 województw, a to one stanowią populację w tym przykładzie), stanowi dodatkowy problem podczas wyznaczania dominanty.



Wykres 4.5. Liczba sklepów według województw w 2002 roku w Polsce

Źródło: opracowanie własne.

Mała liczebność populacji sprawia, że gdyby pojawiła się różnica jednej jednostki[93] pomiędzy przedziałami, byłaby ona trudno interpretowalna w ujęciu bezwzględnym. Różnica taka mogłaby być znacząca, ale także mogłaby być wartością przypadkową, tzn. dana jednostka mogła trafić do innego przedziału cechy przy nieco innym sposobie konstrukcji szeregu.

■

Na zakończenie omawiania dominanty podsumujmy warunki jej wyznaczania:

- szereg wielostopniowy musi mieć równe przedziały klasowe, a jeśli nie ma, należy zastosować wzór z przekształconymi przedziałami (4.19);
- nie można dominanty wyznaczać, gdy cecha nie wykazuje tendencji do skupiania się jednostek wokół jakiejś wartości[94] lub wykazuje tendencję do skupiania się wokół kilku wartości (szeregi wielomodalne); nie można jej stosować w przypadku, gdy szereg jest skrajnie asymetryczny, co rozpoznajemy po największej częstości występowania cechy lub przedziału cechy skrajnych, gdyż jest to sprzeczne z definicją dominanty jako punktu skupienia; przy tym nie da się obliczyć wartości występującej najczęściej dla szeregu wielostopniowego skrajnie asymetrycznego o otwartych przedziałach klasowych;

[93]Sugerująca istnienie dominanty w danym przedziale.
[94]Gdy rozkład jest równomierny.

- w przypadku mało licznych przedziałów klasowych (i mało różniących się liczebności) szeregu rozdzielczego dominanta może być przypadkową wartością, należy więc być bardzo ostrożnym podczas jej obliczania i interpretacji, zwłaszcza w przypadku szeregów cechy ciągłej.

Zauważmy, że większość powyższych warunków (problemów) występuje także przy obliczaniu średniej, dlatego wiedząc, że występują problemy (zastrzeżenia) przy obliczaniu średniej możemy spodziewać się podobnych problemów przy konstrukcji dominanty.

4.2.2. Kwantyle

Kwantyle są to parametry, które dzielą zbiorowość na dwie części w określonych proporcjach.

Można je stosować zawsze[95], nawet gdy w szeregu wielostopniowym skrajne przedziały są otwarte lub nierówno rozpięte. Szereg należy jedynie uporządkować według rosnących (lub malejących) wartości cechy. Można powiedzieć, że kwantyle są ostatnią deską ratunku, gdy nie można policzyć pozostałych miar przeciętnych.

Ogólnie wzór[96] na dowolny kwantyl rzędu p ($0 < p < 1$) Q_p dla szeregu rozdzielczego przedziałowego można zapisać następująco:

$$Q_p = x_{0,p} + \frac{h_{Q_p}}{n_{Q_p}} \cdot \left(p \cdot N - \sum_{i=1}^{k_p - 1} n_i \right), \tag{4.20}$$

gdzie:

k_p – numer klasy, w której znajduje się kwantyl rzędu p,

$x_{0,p}$ – początek przedziału kwantyla,

h_{Q_p} – rozpiętość przedziału kwantyla,

n_{Q_p} – liczebność przedziału kwantyla,

$p \cdot N$ – pozycja kwantyla,

$\sum_{i=1}^{k_p - 1} n_i$ – suma liczebności poprzedzających przedział kwantyla.

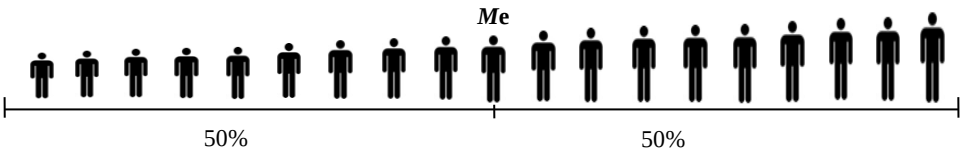
[95] Dokładniej mówiąc, prawie zawsze. Nie można ich policzyć tylko wtedy, gdy dany kwantyl znajdzie się w pierwszym, otwartym przedziale szeregu rozdzielczego i nie można go domknąć.

[96] Wzór jest wyprowadzony przy założeniu, że jednostki znajdujące się w klasie kwantyla są rozłożone w niej równomiernie, czyli że jest stała, taka sama różnica w wartościach cechy pomiędzy tymi jednostkami.

Do najbardziej popularnych kwantyli należą: kwartyle, kwintale, decyle i centyle, inaczej percentyle.

Kwartyle Q_i to wartości ćwiartkowe. W sumie wszystkie dzielą zbiorowość na cztery równe części. Pojedynczy kwartyl dzieli zbiorowość w proporcji 1:3 lub 2:2 lub 3:1. Są one najbardziej popularne w zbiorowości kwantyli.

Najpopularniejszym kwartylem jest kwartyl drugi, który jednocześnie jest piątym decylem i pięćdziesiątym centylem. **Kwartyl drugi $Me = Q_2$ zwany medianą, czyli wartością środkową,** dzieli szereg na dwie części pod względem liczebności (częstości) jednostek statystycznych. Połowa zbiorowości ma wartość cechy nie większą od wartości mediany, a druga połowa zbiorowości ma wartość cechy nie mniejszą od wartości mediany[97]. W przypadku cechy mierzalnej precyzyjne określenie mediany zależy od typu cechy (skokowa czy ciągła) oraz liczebności populacji (parzysta czy nieparzysta).



Wykres 4.6. Graficzne przedstawienie mediany (kwartyla drugiego)

Źródło: opracowanie własne.

W przypadku do cechy skokowej obliczenie mediany polega na wskazaniu jednostki środkowej w zbiorowości i określeniu, jaką wartość cechy ma badana jednostka. W przypadku zbiorowości o nieparzystej liczbie jednostek określamy środkową jednostkę bezpośrednio, zaś w przypadku parzystej liczebności zbiorowości jako medianę określamy, nieistniejącą w rzeczywistości, jednostkę o wartości cechy równej średniej z cech dwóch sąsiednich środkowych jednostek. Wzór, za pomocą którego określana jest mediana cechy skokowej, przedstawia się następująco:

$$Me = \begin{cases} x_{\frac{N+1}{2}}, \text{ dla } N \text{ nieparzystych} \\ \dfrac{x_{\frac{N}{2}} + x_{\frac{N+2}{2}}}{2}, \text{ dla } N \text{ parzystych} \end{cases}, \tag{4.21}$$

gdzie $x_{\frac{N+1}{2}}$ to wartość cechy jednostki statystycznej o numerze $(N+1)/2$.

[97] Powyższa definicja jest poprawna przy założeniu, że w zbiorowości wartości cechy nie powtarzają się.

Przy większych liczebnie zbiorowościach dane są zazwyczaj podane w formie szeregu rozdzielczego, czyli uporządkowane. Wówczas określenie mediany sprowadza się, w przypadku cechy skokowej, do podania klasy cechy, w której znajduje się jednostka środkowa lub, przy parzystej liczebności, jednostki środkowe. W tym ostatnim przypadku mogą się one znajdować zarówno w jednej klasie, jak i w dwóch sąsiednich.

Przykład 4.18

Nauczyciel uporządkował oceny z kolokwium dziewięciu studentów: 2; 2; 3; 3; 3,5; 4; 4; 4,5; 5. Dla $N = 9$ jednostką środkową jest student o numerze $(N+1)/2 = 5$. Piąta osoba w uporządkowanym powyżej szeregu otrzymała ocenę równą x_5 = 3,5. Medianą w tym szeregu jest więc ocena wynoszącą trzy z plusem. Oznacza to, że co najmniej połowa badanych studentów otrzymała ocenę trzy z plusem lub niższą i co najmniej połowa ocenę trzy z plusem lub więcej. Zauważmy, że piąty student jest częścią wspólną obu podzbiorów badanych studentów (zarówno tych, którzy otrzymali co najwyżej ocenę 3,5, i tych, co otrzymali co najmniej 3,5).

Przy interpretacji wyniku mediany należy pamiętać o tym, że poniżej i powyżej mediany jest tyle samo jednostek zbiorowości, bo to jest wartość cechy jednostki środkowej. Łącznie z jednostką medianową powinno to stanowić całość zbiorowości. Jest to istotne zwłaszcza w przypadku cechy skokowej. Gdyby tak w prezentowanym przykładzie powiedziano, błędnie, że połowa (50%) analizowanych studentów miała wartości ocen niższe od oceny dostatecznej plus, a połowa wyższe, wówczas słuchająca uważnie osoba mogłaby zapytać: A co z osobami mającymi trzy plus? Wynikałoby z tego bowiem, że badana zbiorowość ma więcej niż 100% jednostek, co jest oczywiście bzdurą[98]. Zatem wartość jednostki medianowej powinna być uwzględniona w obu częściach zbiorowości albo w żadnej[99].

■

Przykład 4.19

Kontynuując poprzedni przykład, załóżmy, że jeden student był nieobecny i pisał kolokwium w odrębnym terminie. Nauczyciel uwzględnił także jego ocenę w prezentowanym zestawieniu i otrzymał wówczas następujący szereg uporządkowa-

[98] 50% osób powyżej i 50% osób poniżej mediany plus jednostka medianowa dają razem 100% zbiorowości plus jedna osoba.

[99] Można wyliczoną wartość mediany wynosząca 3,5 zinterpretować jeszcze nieco inaczej. Ocenę poniżej 3,5 otrzymało tyle samo studentów co ocenę powyżej 3,5. Uważny czytelnik zauważy, że nie piszemy, ile jednostek statystycznych było powyżej i poniżej mediany. Było ich tyle samo, ale nie tworzą one całej populacji, gdyż brakuje w nich osoby mającej ocenę środkową.

nych ocen: 2; 2; 3; 3; 3,5; 4; 4; 4; 4,5; 5. Liczebność zbiorowości wzrosła o jedną jednostkę, czyli wynosi 10.

Teoretycznie połowa to pięć, ale jeśli przyjmie się jako medianę ocenę piątego studenta, wówczas należy uznać, że medianą jest ocena wynosząca 3,5. Z definicji mediany wynika, że z jej lewej strony musi być tyle samo jednostek co z prawej (aczkolwiek mogą mieć te same wartości cechy). W tym przypadku przed piątą jednostką w zbiorowości jest czterech studentów, zaś za nią pięciu. Tak określona mediana nie spełnia założenia o podziale zbiorowości na dwie równe części. Podobną sytuację otrzyma się, gdy za medianę przyjmie się ocenę szóstego studenta. Otrzymane zostaną wówczas proporcje o odwrotnych liczebnościach przed i za drugim kwartylem. Jeśli jednak jako wartość mediany przyjmiemy ocenę nieistniejącego studenta o numerze (5+6)/2 = 5,5 i przypiszemy mu wartość cechy wynoszącą $Me = X_{5,5} = (X_5+X_6)/2 = (3,5 + 4)/2 = 3,75$, otrzymamy parametr spełniający wymogi definicyjne. W sprawozdaniu nauczyciel napisze, że połowa studentów otrzymała ocenę co najwyżej 3,75, zaś druga połowa studentów w badanej grupie otrzymała ocenę wynoszącą co najmniej 3,75. Ponieważ taka ocena nie istnieje, bardziej precyzyjnie można zinterpretować, w tym przypadku, ten wynik następująco: połowa studentów otrzymała co najwyżej ocenę trzy z plusem, a druga połowa co najmniej cztery. Pierwsza interpretacja (liczba 3,75) jest bezpieczniejsza, bo zawsze prawdziwa, druga wymaga pewnej wprawy w analizie danych.

Uwaga:

W przypadku zmiennej skokowej o powtarzających się wartościach cechy niektórzy Czytelnicy mogą mieć problemy interpretacyjne z medianą, jak i ogólnie z kwantylami. Zmieńmy nieco liczby w powyższym przykładzie. Załóżmy, że studenci otrzymali następujące oceny: 2; 3; 3; 4; 4; 4; 4; 4; 5; 5. Łatwo zauważyć, że jednostka środkowa to, jak poprzednio, student o numerze piątym, ale mediana w tym przypadku wynosi cztery. Poniżej wartości mediany mamy dwie odmiany cechy (dwójka i trójka) i jedną większą od niej (piątka). Wobec tego zachodzi pytanie, czy ocena czwórkowa to wartość środkowa. Otóż tak. Proszę bowiem zauważyć, że to, co jest istotne w definicji mediany, to liczba jednostek statystycznych poniżej jednostki medianowej, która musi się równać liczbie jednostek powyżej jednostki medianowej. Wobec tego w tym przypadku ocenę co najwyżej cztery otrzymała przynajmniej połowa studentów, podobnie jak ocenę co najmniej cztery otrzymała również co najmniej połowa studentów. Zauważmy, że ocena cztery jest częścią wspólną obu tych części danej zbiorowości.

■

Przykład 4.20 (kontynuacja 4.13)

Wróćmy teraz do przykładu o liczebności gospodarstw domowych w Polsce w 1998 roku (przykład 4.13). Ustalono już, że w tym rozkładzie nie ma dominanty. Wyznaczmy teraz medianę liczby osób w gospodarstwach domowych. Najwygodniej jest to zrobić poprzez utworzenie skumulowanego szeregu liczebności. Klasa cechy, w której skumulowane liczebności osiągną numer środkowej jednostki, jest klasą mediany. Gdy liczba jednostek jest nieparzysta, medianą jest ta właśnie wartość cechy.

Tabela 4.12. Gospodarstwa domowe w Polsce w 1998 roku – wyliczenia pomocnicze

Liczba osób w gospodarstwie x_i	1	2	3	4	5 i więcej
Liczba gospodarstw domowych (w tys.) n_i	2 188	2 673	2 427	2 632	2 050
Liczebności skumulowane (w tys.) n_{sk}	2 188	4 861	7 288	9 920	11 970
Liczebności skumulowane (w %)	18,3	40,6	60,9	82,9	100

Źródło: obliczenia własne na podstawie danych z:
Mały rocznik statystyczny Polski2002, tabl. 9(71), s. 118.

W omawianym przykładzie analizowanych jest $N = 11\,970$ gospodarstw, więc brana pod uwagę liczebność jest liczbą parzystą. Zgodnie ze wzorem (4.21) środkowymi jednostkami będą więc gospodarstwa o numerach $n_1 = 11\,970/2 = 5\,985$ oraz $n_2 = (11\,970+2)/2 = 5\,986$. W klasie $x = 2$ jest skumulowanych 4 861 gospodarstw, zatem należy przejść do następnej klasy, w której zostaje przekroczona połowa liczebności populacji. Ponieważ jednak wartości cechy powtarzają się wśród jednostek statystycznych, medianą jest dowolna jednostka z tej klasy. Stąd w przypadku dużych zbiorów danych skokowych przedstawionych w postaci szeregu rozdzielczego numer jednostki środkowej możemy określić jako $n/2$, albowiem wszelkie obliczenia na podstawie danych pogrupowanych są przybliżone. Jeszcze lepiej widać to, obserwując ostatni wiersz tabeli 4.12, w którym znajdują się skumulowane częstości w ujęciu procentowym. Trzecia klasa zawiera skumulowane już ponad 60,9% populacji. Obie jednostki środkowe znajdują się więc w klasie o wartości cechy $x = 3$ osoby. Mają tę samą wartość cechy, zatem można powiedzieć, że co najmniej połowa gospodarstw domowych w Polsce w 1998 roku liczyła sobie co najwyżej trzy osoby, jak też co najwyżej połowa liczyła co najmniej trzy osoby[100].

■

[100] Co do interpretacji – patrz: uwaga do przykładu 4.18, gdyż mamy tu do czynienia z powtarzającymi się wartościami cechy.

Formalnie więc jednostkę środkową i medianę[101] znajduje się w klasie, dla której po raz pierwszy zachodzi poniższa nierówność: $n(x_i) \geq \frac{n}{2}$ lub alternatywnie w pierwszej klasie cechy, w której wartość skumulowanej częstości względnej (dystrybuanty empirycznej) jest równa co najmniej 0,5 (czyli 50%) $F_n(x_i) \geq 0{,}5$. W powyższym przykładzie te nierówności są spełnione (czyli przekraczają połowę liczebności w klasie trzeciej), stąd medianą jest troje dzieci.

W przypadku szeregu rozdzielczego przedziałowego bezpośrednio możemy określić tylko przedział mediany, jako ten, w którym liczebność jednostek statystycznych przekracza połowę populacji. Jeśli potrzebujemy mediany w postaci jednej liczby, musimy skorzystać ze wzoru interpolacyjnego 4.20 przystosowanego do podziału kwartylowego, czyli dla p=1/2, co przedstawia poniższy wzór:

$$Q_2 = Me = x_{0Q_2} + \frac{h_{Q_2}}{n_{Q_2}} \cdot \left(\frac{N}{2} - \sum_{i=1}^{k_2-1} n_i \right), \tag{4.22}$$

gdzie:

x_{0Q_2} – dolny kraniec przedziału, w którym znajduje się mediana;

h_{0Q_2} – rozpiętość przedziału mediany;

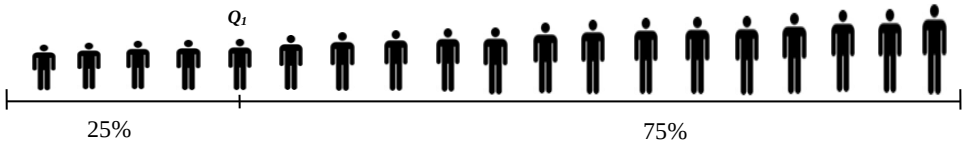
n_{0Q_2} – liczebność przedziału mediany;

$\sum_{i=1}^{k_2-1} n_i$ – suma liczebności przedziałów poprzedzających przedział mediany.

Przechodzimy teraz do wstępnego zdefiniowania pozostałych kwartyli (bez powtarzających się cech).

Kwartyl pierwszy Q_1 (zwany także dolnym) jest to ta wartość cechy, która dzieli zbiorowość w proporcji 1:3, czyli 25% zbiorowości ma wartość cechy do wartości kwartyla pierwszego włącznie, a 75% zbiorowości ma wartość cechy równą wartości kwartyla pierwszego lub wyższą.

[101] W większości znanych autorkom podręczników mediana jest definiowana standardowo, tak jak przedstawiono to we wstępnej, ogólnej definicji, bez uwzględnienia faktu, że wartości cechy skokowej mogą się powtarzać. Formułując powyższą definicję, uwzględniającą powtarzalność cechy wśród więcej niż jednej jednostki statystycznej, autorki wzorują się na sposobie definiowania mediany podanym w podręczniku J. Podgórskiego (s. 53).



Wykres 4.7. Graficzne przedstawienie kwartyla pierwszego

Źródło: opracowanie własne.

W przypadku cechy skokowej postępujemy podobnie jak w przypadku mediany. Wystarczy wskazać wartość cechy jednostki, która „stoi" w szeregu uporządkowanym od najmniejszej wartości cechy do największej, w miejscu dzielącym populację w proporcji1:3. Aczkolwiek ze względu na to, że dzielimy liczebność próby przez cztery, na ogół nie otrzymamy jako numer kwartyla liczby całkowitej. Kwartyl będzie więc wówczas reprezentowany przez nieistniejącą jednostkę o wartości cechy będącej średnią miejsc sąsiednich, tak jak w przypadku mediany przy parzystej liczebności próby[102].

W przypadku powtarzających się liczebności cechy, podobnie jak przy medianie, zdefiniujemy **kwartyl pierwszy formalnie** następująco: wartość kwartyla pierwszego znajduje się w klasie, dla której po raz pierwszy zachodzi poniższa nierówność: $n(x_i) \geq \frac{n}{4}$, lub alternatywnie w pierwszej klasie cechy, w której wartość skumulowanej częstości względnej (dystrybuanty empirycznej) jest równa co najmniej 0,25 (czyli 25%): $F_n(x_i) \geq 0{,}25$.

Natomiast w przypadku cechy ciągłej, tak jak przy medianie, bezpośrednio można podać jedynie przedział kwartyla, czyli granice w jakich zawiera się szukany parametr. Do określenia jego konkretnej wartości używamy wzoru interpolacyjnego 4.20 przekształconego do poniższego wzoru przy założeniu, że p=1/4:

$$Q_1 = x_{0Q_1} + \frac{h_{Q_1}}{n_{Q_1}} \cdot \left(\frac{N}{4} - \sum_{i=1}^{k_1-1} n_i \right), \tag{4.23}$$

gdzie:

x_{0Q_1} – dolny kraniec przedziału, w którym znajduje się kwartyl pierwszy;

h_{0Q_1} – rozpiętość przedziału kwartyla pierwszego;

n_{0Q_1} – liczebność przedziału kwartyla pierwszego;

[102] Można też wyznaczyć wartość kwartyla nie jako zwykłą średnią sąsiednich wartości cechy, pomiędzy którymi wypada miejsce kwartyla, a jako średnią ważoną, przy czym wagami są tu proporcje odległości numeru miejsca kwartyla od sąsiednich numerów jednostek.

$\sum_{i=1}^{k_1-1} n_i$ – suma wszystkich liczebności (kumulacja) przedziałów poprzedzających przedział kwartyla pierwszego.

Kwartyl trzeci Q_3 (zwany także górnym) jest to ta wartość cechy, która dzieli zbiorowość w proporcji 3:1, czyli 75% zbiorowości ma wartość cechy do wartości kwartyla trzeciego włącznie, a 25% zbiorowości ma wartość cechy równą lub powyżej wartości kwartyla trzeciego. Podobnie jak w przypadku poprzednich kwartyli, dla cechy skokowej wskazuje się wartość cechy jednostki dzielącej zbiorowość w podanych proporcjach.

W przypadku powtarzających się liczebności cechy, podobnie jak przy medianie i kwartylu pierwszym, zdefiniujemy **kwartyl trzeci formalnie** następująco: wartość kwartyla trzeciego znajduje się w klasie, dla której po raz pierwszy zachodzi poniższa nierówność: $n(x_i) \geq \frac{3n}{4}$, lub alternatywnie w pierwszej klasie cechy, w której wartość skumulowanej częstości względnej (dystrybuanty empirycznej) jest równa co najmniej 0,75 (czyli 75%): $F_n(x_i) \geq 0{,}75$.

W przypadku szeregu przedziałowego używamy przekształconego przy p=3/4 wzoru (4.20):

$$Q_3 = x_{0Q_3} + \frac{h_{Q_3}}{n_{Q_3}} \cdot \left(\frac{3N}{4} - \sum_{i=1}^{k_3-1} n_i \right), \tag{4.24}$$

gdzie:

x_{0Q_3} – dolny kraniec przedziału, w którym znajduje się kwartyl trzeci;

h_{0Q_3} – rozpiętość przedziału kwartyla trzeciego;

n_{0Q_3} – liczebność przedziału kwartyla trzeciego;

$\sum_{i=1}^{k_3-1} n_i$ – skumulowane liczebności przedziałów poprzedzających przedział kwartyla trzeciego.

Przykład 4.20 (kontynuacja)

Wyznaczmy pozostałe kwartyle rozkładu. Aby to zrobić, należy najpierw wyznaczyć ich pozycje. I tak kwartyl dolny znajduje się na miejscu $N/4$ = 11 970/4 = 2 992,5, zaś kwartyl górny na miejscu o numerze $3N/4$ = 3·11 970/4 = 8 977,5. Gospodarstwo domowe o numerze 2 992,5 znajduje się[103] w klasie cechy x = 2.W tym przypadku

[103] Ponieważ wynik dzielenia nie jest liczbą całkowitą, bierzemy pod uwagę gospodarstwa o numerze 2992 i 2993, które mają tę samą wartość cechy.

więc $Q_1 = 2$, co oznacza, że co najmniej 25% gospodarstw domowych składa się z co najwyżej dwóch osób, zaś co najwyżej 75% składa się z co najmniej dwóch osób.

Pozostaje jeszcze wyznaczyć wartość kwartyla trzeciego. Szukamy gospodarstwa o numerze 8 978 i znajdujemy w klasie cechy $x = 4$. Oznacza to, że co najmniej 75% wszystkich gospodarstw w Polsce w 1998 roku miało nie więcej niż czterech członków, a co najwyżej 25% co najmniej czterech.

■

Zauważmy, że w przypadku kwartyli najwygodniej jest korzystać z szeregu skumulowanego o liczebnościach w ujęciu procentowym (lub jako frakcje), albowiem te liczby pokazują nam na pierwszy rzut oka miejsca kwartyli.

Przykład 4.21 (kontynuacja przykładu 3.3)

Biorąc pod uwagę wagę studentek kierunku zarządzanie przedstawioną w tabeli 4.13, należy wyznaczyć wszystkie kwartyle.

Tabela 4.13. Rozkład wagi studentek – określenie pozycji kwartyli

$\left(x_{i_0} - x_{i_1}\right)$	n_i	n_{isk}
42-46	2	2
46-50	16	18
50-54	26	44
54-58	30	74
58-62	19	93
62-66	5	98
66-70	2	100
Suma	100	

Źródło: opracowanie własne.

Zanim przystąpimy do wyznaczenia kwartyli, należy ustalić ich pozycję. W analizowanym szeregu liczebność wynosiła 100, czyli jest parzysta, w tym przypadku mamy jednak do czynienia z cechą ciągłą o nieznanych dokładnie wartościach wewnątrz przedziałów. Branie więc dwóch sąsiednich jednostek statystycznych, jak w przypadku cechy skokowej przy parzystej zbiorowości, i uśrednianie wartości ich cech nie ma sensu[104], zatem miejsca poszczególnych kwartyli określamy po prostu jako:

[104] Chyba że jedna jednostka znajdzie się w tym przedziale, a druga w następnym, co teoretycznie może się zdarzyć.

$$N_{Q_1} = \frac{N}{4} = \frac{100}{4} = 25 \quad czyli \quad Q_1 = X_{25},$$

$$N_{Q_2} = \frac{N}{2} = \frac{100}{2} = 50 \quad czyli \quad Q_2 = X_{50},$$

$$N_{Q_3} = \frac{3N}{4} = \frac{3 \cdot 100}{4} = 75 \quad czyli \quad Q_3 = X_{75}.$$

Patrząc na tabelę 4.13, widzimy, że wartość kwartyla pierwszego znajduje się w trzecim przedziale, gdyż w pierwszym znajduje się łącznie osiem jednostek, element dziewiąty, dziesiąty itd. aż do osiemnastego włącznie tworzy przedział drugi, zaś dziewiętnasty, dwudziesty itd. do czterdziestego czwartego znajdują się w trzeciej klasie. Zatem jednostka dwudziesta piąta także się tam znajduje. Po określeniu pozycji kolejnym krokiem jest wyznaczenie wartości kwartyli i ich interpretacja. W odniesieniu do kwartyla pierwszego wykorzystujemy wzór (4.23):

$$Q_1 = x_{0Q_1} + \frac{h_{Q_1}}{n_{Q_1}} \cdot \left(\frac{N}{4} - \sum_{i=1}^{k_1 - 1} n_i \right) = 50 + \frac{4}{26} \cdot (25 - 18) = 50 + \frac{4}{26} \cdot 7 = 50 + 1{,}077 = 51{,}1 \text{kg}.$$

Wyznaczona powyżej wartość oznacza, że 25% przebadanych kobiet miało wagę 51,1 kg lub mniejszą, a 75% ważyło co najmniej 51,1 kg. Inaczej mówiąc, co czwarta przebadana studentka ważyła co najwyżej 51,1 kg.

Analogicznie wyliczamy medianę. Wiemy, że przy stu jednostkach kwartyl drugi jest na pozycji pięćdziesiątej. Zatem znajduje się on w czwartej klasie szeregu z tabeli 4.13 (znajdują się tam jednostki od czterdziestej czwartej do siedemdziesiątej trzeciej włącznie). Wartość mediany została wyliczona na podstawie wzoru (4.22) następująco:

$$Q_2 = Me = x_{0Q_2} + \frac{h_{Q_2}}{n_{Q_2}} \cdot \left(\frac{N}{2} - \sum_{i=1}^{k_2 - 1} n_i \right) = 54 + \frac{4}{30} \cdot (50 - 44) = 54 + \frac{4}{30} \cdot 6 = 54 + 0{,}800 = 54{,}800 \text{ kg}.$$

Otrzymany wynik oznacza, że 50% przebadanych studentek ważyło do 54,800 kg włącznie, a 50% ważyło 54,8 kg lub więcej. Zatem co druga studentka ważyła co najwyżej (co najmniej) 54,8 kg.

Pozycją trzeciego kwartyla jest siedemdziesiąty piąty element, więc jest to drugi element z piątej klasy szeregu z tabeli 4.16. W tym przypadku zastosowanie znajduje wzór (4.24):

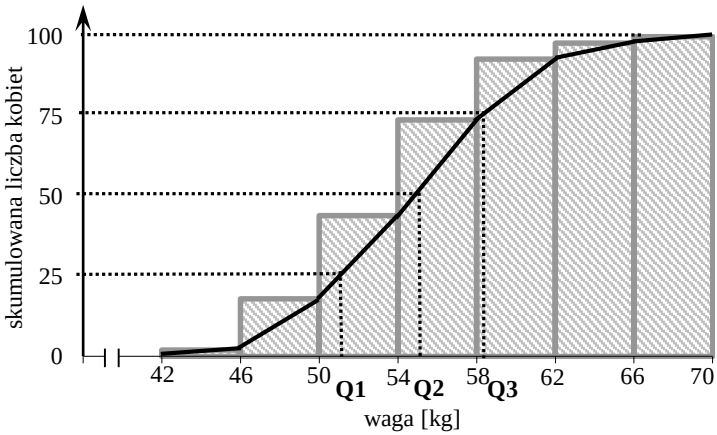
$$Q_3 = x_{0Q_3} + \frac{h_{Q_3}}{n_{Q_3}} \cdot \left(\frac{3N}{4} - \sum_{i=1}^{k_3 - 1} n_i \right) = 58 + \frac{4}{19} \cdot (75 - 74) = 58 + \frac{4}{19} \cdot 1 = 58 + 0{,}211 = 58{,}2 \text{ kg}.$$

Wartość wyznaczonego kwartyla trzeciego oznacza, że 75% wszystkich przebadanych kobiet ważyło 57,2 kg lub mniej, a 25% ważyło 57,2 kg lub więcej. Inaczej rzecz ujmując, co czwarta studentka ważyła co najmniej 57,2 kg.

Uwaga:

Zauważmy, że w przypadku zmiennej przedziałowej (ciągłej) interpretacja kwantyli jest prostsza. Stąd pewna różnica pomiędzy zapisem dla zmiennej skokowej w poprzednich przykładach i dla zmiennej ciągłej w niniejszym przykładzie. Wynika ona z faktu, że w przypadku zmiennej ciągłej prawdopodobieństwo zaobserwowania konkretnej wartości cechy jest praktycznie zerowe, czyli obserwacja równa dokładnie wartości kwartyla najprawdopodobniej nie została zarejestrowana.

Oprócz wartości wyznaczonych zgodnie z przedstawionymi wzorami estymacyjnymi można również wartości kwartyli obliczyć graficznie. W tym przypadku należy wykonać wykres liczebności lub częstości skumulowanej. W naszym przykładzie przedstawia go rysunek 4.8.



Wykres 4.8. Graficzne wyznaczanie kwartyli

Źródło: opracowanie własne.

Na wykresie przedstawiamy skumulowany histogram wagi. Następnie łączymy końce przedziałów, tak jak narysowano to na wykresie 4.8, tworząc krzywą skumulowaną. Następnie na osi rzędnych (OY) zaznaczamy odcinki o długości odpowiednio $N/4$ (25%), $2N/4$ (50%) i $3N/4$ (75%). Z każdego punktu podziału odcinka poprowadzimy prostą równoległą do osi odciętych (OX). Ostatnim etapem jest zrzutowanie punktów przecięcia prostych równoległych z krzywą skumulowaną na oś wartości cechy (OX). Odczytane stąd wartości dają kwartyle. Analizując

wyniki otrzymane na rysunku 4.8, widzimy, iż mają one wartości zgodne z otrzymanymi uprzednio wynikami.

■

Przykład 4.22 (kontynuacja 4.5)

Analizując dane dotyczące czasu biegu na 60 m dla 30 uczniów w pewnej szkole średniej, należy wyznaczyć miary pozycyjne. Aby te wyliczenia były możliwe do wykonania, w tabeli 4.14 przedstawiono liczebności skumulowane dla badanego szeregu. Przyjęto, podobnie jak przy wyliczeniach dotyczących średniego czasu biegu, że przedziały domknięte są od dołu.

Tabela 4.14. Czas biegu na 60 m uczniów szkoły średniej – wyliczenia dotyczące kwartyli

Czas biegu na 60 m (w sekundach)	Liczba uczniów	Skumulowana liczba uczniów
6,0-7,5	2	2
7,5-8,0	3	5
8,0-8,5	6	11
8,5-9,0	8	19
9,0-9,5	5	24
9,5-10,0	4	28
10,0-10,5	2	30

Źródło: dane umowne.

Wyliczenia należy rozpocząć od określenia pozycji danej miary, czyli pozycji uczniów, dla których czasy biegu są wartościami poszczególnych kwartyli. Wyznaczamy następująco:

$$N_{Q_1} = \frac{N}{4} = \frac{30}{4} = 7,5;$$

$$N_{Q_2} = \frac{N}{2} = \frac{30}{2} = 15;$$

$$N_{Q_3} = \frac{3N}{4} = \frac{3 \cdot 30}{4} = 22,5.$$

W pierwszym i drugim przedziale znajduje się łącznie pięciu uczniów, a więc uczeń o numerze n=7,5 jest dopiero w przedziale trzecim. Wartość kwartyla pierwszego jest więc wartością z trzeciego przedziału. Podobnie wyznaczamy pozycję mediany, którą znajdujemy w przedziale czwartym, zaś kwartyl trzeci

w piątym. Podstawiając odpowiednio do wzorów (4.23), (4.22) i (4.24), otrzymujemy następujące wartości kwartyli:

$$Q_1 = x_{0Q_1} + \frac{h_{Q_1}}{n_{Q_1}} \cdot \left(\frac{N}{4} - \sum_{i=1}^{k_1-1} n_i \right) = 8 + \frac{0{,}5}{6} \cdot (7{,}5 - 5) = 8 + 0{,}21 = 8{,}21\,\text{s};$$

$$Me = x_{0Q_2} + \frac{h_{Q_2}}{n_{Q_2}} \cdot \left(\frac{N}{2} - \sum_{i=1}^{k_2-1} n_i \right) = 8{,}5 + \frac{0{,}5}{8} \cdot (15 - 11) = 8{,}5 + 0{,}25 = 8{,}75\,\text{s};$$

$$Q_3 = x_{0Q_3} + \frac{h_{Q_3}}{n_{Q_3}} \cdot \left(\frac{3N}{4} - \sum_{i=1}^{k_3-1} n_i \right) = 9 + \frac{0{,}5}{5} \cdot (22{,}5 - 19) = 9{,}5 + 0{,}35 = 9{,}85\,\text{s}.$$

Otrzymana wartość kwartyla pierwszego oznacza, że 25% analizowanych uczniów dystans 60 m pokonuje w czasie 8,21 sekundy lub krótszym, a 75% w czasie co najmniej 8,21 sekund. Połowa badanych uczniów ukończyła bieg w czasie nie przekraczającym 8,75 sekundy, a pozostała połowa biegła co najmniej 8,75 sekundy. 75% uczniów potrzebowało na pokonanie dystansu 60 m co najwyżej 9,85 sekundy, a dla 25% ten czas wynosił 9,85 sekund lub więcej.

■

Przykład 4.23

Podczas corocznego badania szkolnego zebrano dane o wzroście grupy 8-letnich chłopców. Ilu ich było, jeżeli wiadomo, że połowa z nich miała co najmniej 137 cm wzrostu, a 55 miało co najwyżej 134 cm wzrostu? Chłopców o wzroście od 134 do 138 było 24.

W treści zadania pojawia się informacja o wzroście połowy 8-latków, czyli o medianie (*Me*). Wykorzystując wzór (4.22) i podstawiając do niego dane z zadania, można wyznaczyć liczbę przebadanych chłopców z badanej grupy wiekowej.

$$\left.\begin{array}{l} Me = 137 \\ x_{0Q_2} = 134 \\ \sum_{i=1}^{k_2-1} n_i = 55 \\ h_{Q_2} = 138 - 134 = 4 \\ n_{Q_2} = 24 \end{array}\right\} \Rightarrow \; Me = x_{0Q_2} + \frac{h_{Q_2}}{n_{Q_2}} \cdot \left(\frac{N}{2} - \sum_{i=1}^{k_2-1} n_i \right) \Rightarrow$$

$$\Rightarrow \quad 137 = 134 + \left(\frac{N}{2} - 55\right)\frac{4}{24} \quad \Rightarrow \quad N = 146.$$

Odpowiedź: Zbadano 146 chłopców mających osiem lat.

■

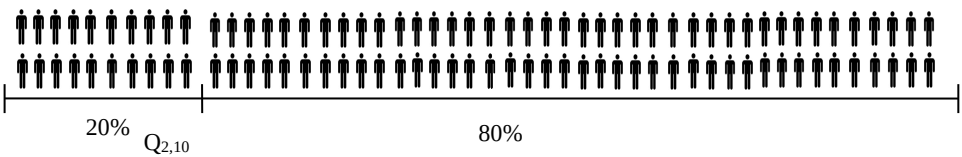
Podsumowanie omówionych miar średnich

- ⇨ Średnią arytmetyczną i dominantę należy liczyć tylko dla rozkładów symetrycznych lub o słabej asymetrii.
- ⇨ Kwartyle[105] można policzyć zawsze i nie ma żadnych ograniczeń na ich stosowanie (poza przypadkiem kwartyli w pierwszym przedziale, gdy jest on otwarty).

4.2.3. Przeciętne pozycyjne wyższych rzędów

Na rysunku 4.2 podano najbardziej popularne kwantyle. Jak już wspomniano, omówione powyżej kwartale dzielą zbiorowość na cztery części. Kwintyle z kolei analogicznie dzielą zbiorowość na pięć części, czyli skok przy podziale wynosi 20%. Decyle zaś dzielą populację na dziesięć części, percentyle (centyle) na sto, a promile na tysiąc.

Dla przykładu przedstawimy poniżej jeszcze jeden z popularniejszych sposobów podziału (i pomiaru) dużych zbiorowości, a mianowicie **podział decylowy**[106] czyli na dziesięć części. Poniższy rysunek przedstawia przykład definicji decyla drugiego.



Wykres 4.9. Graficzne przedstawienie decyla drugiego

Źródło: opracowanie własne.

Decyl drugi dzieli zbiorowość na dwie części w proporcji 2:8, czyli co najmniej 20% zbiorowości ma wartość cechy do wartości decyla drugiego włącznie, a co

[105] Dotyczy to także ogólnie wszystkich kwantyli.

[106] Znane czytelnikowi zastosowania decyli to określanie wyników egzaminu gimnazjalnego czy maturalnego (choć w tym ostatnim przypadku ma miejsce pewna korekta, co jest górną wartością wyniku – tworzy się tzw. podział staninowy, ale nie zmienia to idei podziału na decyle).

najmniej 80% zbiorowości ma wartość cechy równą lub powyżej wartości decyla drugiego.

Zatem wzór interpolacyjny na wartość decyla drugiego ma teraz następującą postać szczegółową:

$$Q_{2,10} = x_{0Q_{2,10}} + \frac{h_{Q_{2,10}}}{n_{Q_{2,10}}} \cdot \left(\frac{2N}{10} - \sum_{i=1}^{k_2-1} n_i \right), \tag{4.25}$$

gdzie k_2 – numer klasy decyla drugiego, zaś pozostałe oznaczenia są analogiczne do oznaczeń we wzorach na kwantyle.

Przykład 4.24

Obliczmy drugi decyl wagi studentek. Możemy skorzystać z szeregu skumulowanego użytego do obliczeń kwartyli, zamieszczonego w tabeli 4.14.

Decyl drugi to wartość cechy jednostki o numerze:

$$N_{Q_{2,10}} = \frac{2N}{10} = \frac{200}{10} = 20 \quad \text{czyli} \quad Q_{2,10} = X_{20},$$

która znajduje się w trzecim przedziale odmiany cechy, czyli 50-54 kg. Podstawiamy dane do wzoru 4.25 i otrzymujemy:

$$Q_{2,10} = 50 + \frac{4}{26} \cdot \left(\frac{2 \cdot 100}{10} - 18 \right) = 50,3 \text{kg}.$$

Wiemy teraz, że 20% studentek ważyło co najwyżej 50,3 kg, zaś 80% co najmniej tyle.

■

Kolejnym przykładem kwantyli są **centyle** (lub **percentyle**). Dzielą one zbiorowość na sto części, czyli są to znane czytelnikowi ze szkoły procenty.

Przykładowo centyl piąty dzieli zbiorowość na dwie części w proporcjach 5:95, czyli 5% zbiorowości ma wartości niższe lub równe wartości centyla piątego, zaś 95% populacji ma wartość równą lub większą od wartości centyla. Wzór interpolacyjny na wartość centyla piątego przyjmuje następującą postać szczegółową:

$$Q_{5,100} = x_{0Q_{5,100}} + \frac{h_{Q_{5,100}}}{n_{Q_{5,100}}} \cdot \left(\frac{5N}{100} - \sum_{i=1}^{k_5-1} n_i \right), \tag{4.26}$$

gdzie k_5 – numer klasy centyla piątego, zaś pozostałe oznaczenia są analogiczne do oznaczeń użytych we wzorach na pozostałe kwantyle.

Z centylami można się spotkać między innymi w gabinetach lekarzy pediatrów. Są one używane do opisu populacji dzieci na kolejnych etapach rozwoju i służą do porównań ich rozwoju. Lekarze wyznaczają takie parametry rozwoju dziecka, jak wzrost, waga, obwód głowy, zaznaczając je na siatkach centylowych dla dzieci określonego wieku. Siatki centylowe powstały na podstawie badań większej grupy dzieci. Przykładowo, jeżeli dziecko ze względu na wzrost osiąga centyl 75, oznacza to, że 75% dzieci z wcześniejszych badań miało wzrost wynoszący tyle co badane dziecko lub mniejszy, a jedynie 25% było takiego samego wzrostu lub wyższe. Siatki te służą także do sprawdzania, czy dane dziecko nie ma jakichś opóźnień w rozwoju.

Przykład 4.25

Sprawdźmy, jaką co najwyżej wagę z analizowanego przez nas powyżej przykładu miało 5% studentek. Obliczmy zatem wartość centyla piątego związanego z jednostką o numerze 5 (bo $5N/100=5$):

$$Q_{5,100} = 46 + \frac{4}{16} \cdot (5-2) = 46,75 \text{ kg.}$$

Zatem 5% studentek ważyło co najwyżej 46,75 kg lub mniej, zaś 95% co najmniej tyle lub więcej.

■

Ostatnim prezentowanym przykładem kwantyli są **promile** (łac. *pro mille* – na tysiąc). Dzielą one zbiorowość na tysiąc części, oznaczane symbolem ‰.

Przykładowo promil ósmy dzieli zbiorowość na dwie części w proporcji 8:992, co oznacza, że 8‰ *(lub* 0,8%) zbiorowości ma wartości niższe lub równe wartości tego promila, zaś 992‰ *(lub* 99,2%) populacji ma wartość równą lub większą od wartości promila ósmego. Wzór interpolacyjny na wartość promila ósmego przyjmuje następującą postać:

$$Q_{8,1000} = x_{0Q_{8,1000}} + \frac{h_{Q_{8,1000}}}{n_{Q_{8,1000}}} \cdot \left(\frac{8N}{1000} - \sum_{i=1}^{k_8-1} n_i \right), \tag{4.27}$$

gdzie k_8 – numer klasy promila ósmego, zaś pozostałe oznaczenia są analogiczne do oznaczeń we wzorach na kwantyle.

Promile używane są do analizy dużych zbiorowości oraz opisu rzadkich procesów. Są bardzo popularne w analizach demograficznych.

Przykład 4.26

Wyznacz na podstawie poniższej tabeli, dotyczącej wieku kobiet skazanych prawomocnie przez sądy powszechne za przestępstwa ścigane z oskarżenia publicznego w 1992 roku, promil dwudziesty pierwszy i trzysta piętnasty.

Tabela 4.15. Rozkład wieku kobiet skazanych prawomocnie przez sądy powszechne za przestępstwa ścigane z oskarżenia publicznego w 1992 roku

Wiek skazanych kobiet	Liczba kobiet	Skumulowana liczba kobiet
20-25	1 166	1 166
25-30	965	2 131
30-35	1 280	3 411
35-40	1 664	5 075
40-45	1 584	6 659
45-50	1 149	7 808
50-55	591	8 399
55-60	760	9 159

Źródło: opracowanie własne na podstawie: *Rocznik statystyczny GUS 1993*, GUS, tabl. 21 (147), s. 94.

Poniżej zostały zaprezentowane wyliczenia dotyczące wartości dwudziestego pierwszego promila związanego z jednostką o numerze 192 (bo $21 \cdot N/1000 = 192{,}34$), która znajduje się w pierwszym przedziale:

$$Q_{21,1000} = 20 + \frac{5}{1166} \cdot (192{,}34 - 0) = 20{,}82.$$

Zatem 2,1% skazanych prawomocnie przez sądy powszechne za przestępstwa ścigane z oskarżenia publicznego w 1992 roku kobiet miało co najwyżej 20,8 roku.

Podobne wyliczenia można wykonać w odniesieniu do trzechsetnego piętnastego promila, który dotyczy jednostki o numerze 2885 (bo $315 \cdot N/1000 = 2885{,}085$), znajdującej się w trzecim przedziale:

$$Q_{315,1000} = 30 + \frac{5}{1280} \cdot (2885{,}085 - 2131) = \ 32{,}95.$$

Oznacza to, że 31,5% analizowanych kobiet miało mniej niż 33 lata.

■

Kluczowe pojęcia

Parametry opisowe
Średnie klasyczne: średnia arytmetyczna, średnia harmoniczna, średnia geometryczna
Średnie pozycyjne: dominanta, moda, kwartyle, mediana, decyle, centyle, promile
Stopa wzrostu, tempo wzrostu

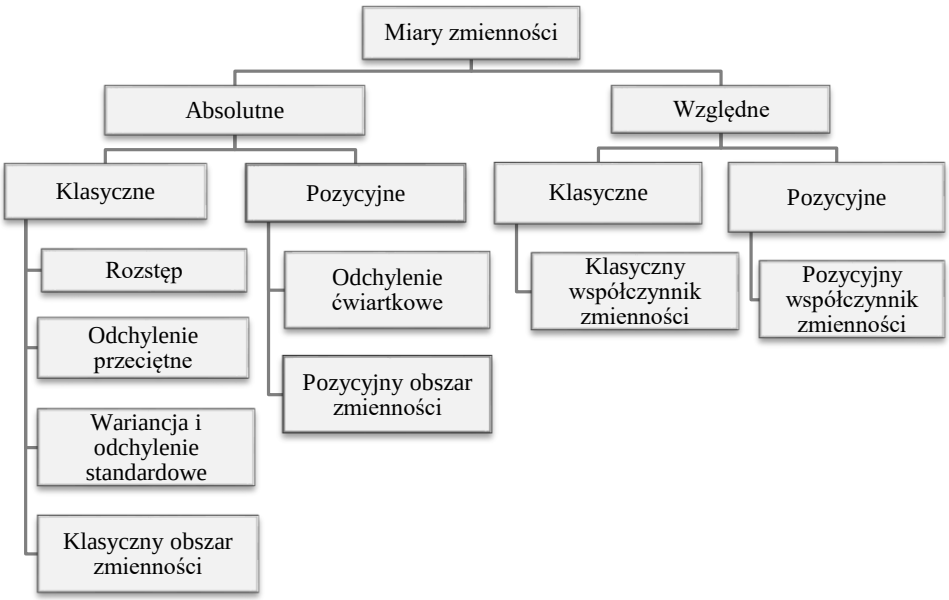
Pytania kontrolne

1. Czy do opisu szeregu czasowego można użyć średniej harmonicznej?
2. Czy średniej arytmetycznej i geometrycznej można używać zamiennie w szeregach przekrojowych?
3. Czy średniej arytmetycznej i harmonicznej można używać zamiennie w szeregach przekrojowych?
4. Jakiej średniej należy użyć, gdy szereg czasowy jest w postaci stóp lub temp wzrostu?
5. Jak nazwiemy rozkład, w którym nie ma dominanty? Podaj co najmniej jeden typ.
6. Jak nazwiemy rozkład, w którym jest więcej niż jedna dominanta?
7. Jaki procent populacji zawiera się między kwartylem pierwszym i trzecim?
8. O czym mówi nam średnia stopa wzrostu?
9. Jaka jest różnica pomiędzy tempem a stopą wzrostu?
10. Kiedy nie można do opisu populacji użyć średniej arytmetycznej?
11. Kiedy nie można do opisu populacji użyć dominanty?
12. Kiedy nie można do opisu populacji użyć mediany?
13. Kiedy nie można do opisu populacji użyć kwartyla pierwszego?
14. Kiedy nie można policzyć średniej arytmetycznej?
15. Kiedy nie można policzyć dominanty?
16. Kiedy nie można policzyć mediany?
17. Kiedy nie można policzyć kwartyla pierwszego?
18. Kiedy nie można policzyć kwartyla trzeciego?

5. Miary dyspersji

Miary dyspersji, czyli rozproszenia, *są to takie charakterystyki opisowe rozkładu cechy mierzalnej, które mierzą stopień zróżnicowania wartości badanej cechy w tym rozkładzie.*

Mówiąc o zróżnicowaniu, mamy na myśli odpowiedź na pytanie, jak bardzo różnią się wartości zmiennej w badanej populacji lub jak te różnice są duże. Różnice absolutne mogą być bowiem identyczne, ale mogą mieć zupełnie inne znaczenie względne, np. stuzłotowa różnica w zarobkach w populacji o średnim zarobku 500 zł i taka sama różnica w populacji o średniej płacy 1000 zł to z punktu widzenia zainteresowanych zarobkami różne wartości.



Rysunek 5.1. Podział miar zmienności

Źródło: opracowanie własne.

Stosuje się więc dwa rodzaje miar rozproszenia: *absolutne* (wyrażone w jednostkach naturalnych badanej cechy) oraz *względne, czyli stosunkowe* (wyrażone w relacji do jakiegoś poziomu wartości cechy w badanej populacji). Zarówno miary względne, jak i absolutne dzielimy na miary klasyczne i pozycyjne w zależności od punktu odniesienia. W miarach klasycznych punktem odniesienia jest jakaś **średnia klasyczna**, najczęściej średnia arytmetyczna**,** zaś w drugim przypadku punktem odniesienia są **średnie pozycyjne**, najczęściej kwartyle, a przede wszystkim **mediana**. Podstawowy podział najbardziej popularnych miar dyspersji przedstawia schemat umieszczony na rysunku 5.1.

Przedstawione na rysunku 5.1 miary nie są jedynymi, dzięki którym możemy określić zmienność badanej cechy. W literaturze można spotkać wiele innych, ale te opisywane w niniejszym podręczniku są uznawane za podstawowe i najczęściej wyznaczane.

5.1. Klasyczne miary dyspersji

Najprostszą z miar dyspersji, a więc i dającą najmniej informacji, jest **rozstęp**. Jest to różnica pomiędzy największą i najmniejszą wartością cechy w zbiorowości:

$$R = x_{\max} - x_{\min}. \tag{5.1}$$

Przykład 5.1

Stopa rejestrowanego bezrobocia na koniec roku w Polsce w latach 2003-2012 wynosiła odpowiednio (w %)[107]: 20,0; 19,0; 17,6; 14,8; 11,2; 9,5; 12,1; 12,4; 12,5; 13,4. Wyznacz rozstęp między rejestrowanymi na koniec każdego roku stopami bezrobocia w badanym okresie.

Największą stopą bezrobocia rejestrowaną na koniec roku w analizowanym okresie czasu była wartość 20%, zaś najniższą 11,2%. Zatem rozstęp wynosił:

$$R = x_{\max} - x_{\min} = 20\% - 11,2\% = 8,8\%.$$

Różnica między najniższą a najwyższą stopą rejestrowanego na koniec każdego roku bezrobocia w latach 2003-2012 wynosiła 8,8 punktu procentowego[108].

■

[107] Stopa bezrobocia jest to, w uproszczeniu, liczba osób bezrobotnych podzielona przez liczbę osób w wieku produkcyjnym aktywnych zawodowo. Źródło danych: http://www.stat.gov.pl/gus/wskazniki_makroekon_PLK_HTML.htm.

[108] Zwracamy uwagę czytelnika, że jeżeli dane są w postaci jednostek %, wówczas, jak w podanym przykładzie, różnicę między wartościami podajemy jako punkty procentowe. Punkty procentowe oznaczają *różnicę* w ujęciu procentowym, podczas gdy słowo „procent" oznacza iloraz dwóch liczb.

Klasyczne miary zmienności stanowią wypadkową odchyleń wszystkich wartości cechy od jakiegoś ustalonego poziomu. Można więc je przedstawić w postaci średniej – są to więc średnie odchylenia od pewnej stałej, na ogół średniej arytmetycznej. Najwygodniejszą i najprostszą miarą rozproszenia cechy wokół poziomu przeciętnego byłoby więc średnie odchylenie wszystkich wartości cechy od średniej w populacji. Jednak, jak już wykazaliśmy w poprzednim rozdziale, własnością średniej arytmetycznej jest, iż suma odchyleń cechy od tejże średniej jest zawsze równa zeru, gdyż odchylenia dodatnie i ujemne zawsze się znoszą. Mamy więc pewien problem z konstrukcją takiej miary, co ilustruje poniższy przykład.

Przykład 5.2

Zebrano dane dotyczące wydatków ponoszonych przez grupę pięciu przyjaciół na śniadanie. Otrzymano następujący zestaw danych (w zł): 5, 3, 2, 4, 4, dla którego średnia wyniosła:

$$\bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i = \frac{5+3+2+4+4}{5} = \frac{18}{5} = 3{,}6 \text{ zł.}$$

Popatrzmy, co będzie się działo, gdy zsumujemy odchylenia od tej średniej:

$$\sum_{i=1}^{n}\left(x_i - \bar{x}\right) = \left[\left(5-3{,}6\right)+\left(3-3{,}6\right)+\left(2-3{,}6\right)+\left(4-3{,}6\right)+\left(4-3{,}6\right)\right] =$$

$$= \left(1{,}4+(-0{,}6)+(-1{,}6)+0{,}4+0{,}4\right) = 0 \text{ zł.}$$

■

Jak widać z powyższych obliczeń, dodane odchylenia poszczególnych wydatków wzajemnie się znoszą i w sumie otrzymujemy, zgodnie z teorią, wartość zerową odchyleń. Liczenie średniego odchylenia wydatków od wydatku średniego nie ma sensu, albowiem po podzieleniu przez liczbę przyjaciół nadal otrzymujemy zero. Sytuacja taka zawsze ma miejsce, dlatego aby określić stopień zróżnicowania, należałoby zastosować „sztuczkę" eliminującą problem znoszenia się znaków. Możemy to uczynić, gdyż nie interesuje nas w tej chwili, czy poszczególne wartości cechy są powyżej czy poniżej średniej (a więc znak odchylenia). Nas interesuje, jak bardzo one odchylają się od ustalonego poziomu, czyli na przykład średniej arytmetycznej. W matematyce znak liczby usuwa się na dwa sposoby. Pierwszy z nich *to wzięcie modułu danej liczby*, drugi zaś *to podniesienie tej liczby do kwadratu*.

5.1.1. Odchylenie przeciętne

Jak wspomniano wyżej, aby „obejść" problem znaków odchyleń od średniej, powodujących zerowanie się sumy odchyleń od średniej arytmetycznej, można użyć dwóch sposobów, co w efekcie daje nam dwie miary dyspersji. Pierwszy z nich to pominięcie znaków, czyli sumowanie wartości bezwzględnych odchyleń. Badacza nie interesuje bowiem, czy odchylenie jest *in plus* (wartość cechy jest powyżej średniej arytmetycznej) czy *in minus* (wartość cechy poniżej średniej arytmetycznej). Interesująca jest jedynie wielkość odchylenia. Otrzymuje się w ten sposób miarę zwaną ***odchyleniem przeciętnym***. Miara ta jest intuicyjna i wygodna w interpretacji.

Odchylenie przeciętne mierzy średnią bezwzględnej wartości odchyleń wartości cechy od jej poziomu średniego.

W zależności od postaci szeregu, z jakiego jest wyznaczane, korzysta się z różnych postaci wzoru (podobnie jak przy średniej arytmetycznej). I tak dla prostego szeregu punktowego stosowany jest następujący wzór:

$$d = \frac{1}{n}\sum_{i=1}^{n}\left|x_i - \overline{x}\right|. \tag{5.2}$$

W przypadku szeregu rozdzielczego punktowego korzystamy z poniższej postaci (ważonej):

$$d = \frac{1}{n}\sum_{i=1}^{k}\left|x_i - \overline{x}\right| \cdot n_i, \tag{5.3}$$

zaś dla przedziałowego z następującej:

$$d = \frac{1}{n}\sum_{i=1}^{k}\left|{}^{o}x_i - \overline{x}\right| \cdot n_i. \tag{5.4}$$

Zauważmy, iż przedstawione wzory na odchylenie przeciętne są prawie identyczne z wzorami przedstawionymi w poprzednim rozdziale, omawiającym pojęcie średniej arytmetycznej. Widzimy więc, że odchylenie przeciętne jest też średnią arytmetyczną. Jedyna różnica polega na tym, że cecha, dla której liczymy teraz średnią, jest w postaci bezwzględnych odchyleń od średniej, czyli jest szeregiem przekształconym.

Przykład 5.3 (kontynuacja 5.2)

Biorąc pod uwagę dane dotyczące wydatków grupy przyjaciół na śniadanie, uśrednione różnice modułów poszczególnych wartości wynoszą:

$$d = \frac{1}{n}\sum_{i=1}^{n}|x_i - \bar{x}| = \frac{1}{5}\left[|5-3,6|+|3-3,6|+|2-3,6|+|4-3,6|+|4-3,6|\right] =$$

$$= \frac{1}{5}(1,4+0,6+1,6+0,4+0,4) = \frac{1}{5}\cdot 4,4 = 0,88 \text{ zł.}$$

Oznacza to, że wydatki co do wartości bezwzględnej każdego z przyjaciół odchylały się przeciętnie od średniej arytmetycznej dla całej grupy o 88 groszy.

■

Przykład 5.4 (kontynuacja 4.18)

Wróćmy do przykładu ocen otrzymanych przez studentów na kolokwium i określmy stopień ich zróżnicowania. Przypomnijmy, że otrzymane oceny były następujące: 2; 2; 3; 3; 3,5; 4; 4; 4; 4,5; 5.

Zanim wyznaczona zostanie miara dyspersji, musimy wyznaczyć średnią z ocen:

$$\bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i = \frac{1}{10}(2+2+3+3+3,5+4+4+4+4,5+5) = \frac{1}{10} \cdot 35 = 3,5.$$

Zatem przeciętnie studenci otrzymali ocenę dostateczną z plusem. Aby można było wyznaczyć odchylenie przeciętne, należy policzyć różnice pomiędzy poszczególnymi wartościami a otrzymaną średnią ocen. Wyliczenia te, dla wygody, przedstawiliśmy w tabeli 5.1.

Tabela 5.1. Oceny z kolokwium – wyliczenia związane z odchyleniem przeciętnym

x_i	$x_i - \bar{x}$	$\lvert x_i - \bar{x}\rvert$
2	–1,5	1,5
2	–1,5	1,5
3	–0,5	0,5
3	–0,5	0,5
3,5	0,0	0,0
4	0,5	0,5
4	0,5	0,5
4	0,5	0,5
4,5	1,0	1,0
5	1,5	1,5
Suma	0,0	8,0

Źródło: opracowanie własne.

Uwzględniając otrzymany w tabeli 5.1 wynik końcowy, odchylenie przeciętne wyznaczane ze wzoru (5.2) jest następujące:

$$d = \frac{1}{n}\sum_{i=1}^{n}\left|x_i - \bar{x}\right| = \frac{1}{10}\cdot 8{,}0 = 0{,}8.$$

Wyliczona wartość oznacza, że przeciętnie oceny badanych studentów co do wartości bezwzględnej różniły się od średniej o 0,8 punktu (stopnia).

■

Przykład 5.5 (kontynuacja przykładu 3.3)

Biorąc pod uwagę wagę studentek kierunku zarządzanie przedstawioną w tabeli 4.8, należy wyznaczyć odchylenie przeciętne. W przykładzie 4.3 wyznaczono średnią wagi, która wyniosła 54,84 kg. Wyliczenia dotyczące odchylenia przeciętnego przedstawiono w tabeli 5.2.

Tabela 5.2. Rozkład wagi studentek – obliczenia związane z odchyleniem przeciętnym

$\left(x_{i_0} - x_{i_1}\right)$	n_i	x_i^o	$x_i^o - \bar{x}$	$\left\|x_i^o - \bar{x}\right\| n_i$
42-46	2	44	–10,84	21,68
46-50	16	48	– 6,84	109,44
50-54	26	52	– 2,84	73,84
54-58	30	56	1,16	34,80
58-62	19	60	5,16	98,04
62-66	5	64	9,16	45,80
66-70	2	68	13,16	26,32
Suma	100			409,92

Źródło: opracowanie własne.

Aby wyznaczyć odchylenie przeciętne, należy otrzymaną sumę 409,92 podstawić do wzoru (5.4), dzieląc ją przez liczbę obserwacji:

$$d = \frac{1}{n}\sum_{i=1}^{k}\left|x_i^o - \bar{x}\right| \cdot n_i = \frac{1}{100}\cdot 409{,}92 = 4{,}0992 \approx 4{,}10\ \text{kg}.$$

Otrzymany wynik oznacza, że przeciętna waga studentek kierunku zarządzanie co do wartości bezwzględnej odchyla się od średniej wagi o 4,10 kg.

■

Odchylenie przeciętne jest jednak rzadko używane, zwłaszcza gdy chce się użyć miary dyspersji do konstrukcji innych miar. Ze względu na moduł nie jest ono wygodne w dalszych operacjach matematycznych, choć jest to, podkreślmy, miara bardzo intuicyjna w interpretacji i w sumie prosta obliczeniowo. Można jej śmiało używać, jeśli nie zamierzamy jej wykorzystać do budowy innych miar.

5.1.2. Wariancja i odchylenie standardowe

Drugim sposobem ominięcia problemu znaków odchyleń od średniej jest, jak już powyżej zaznaczono, ich eliminacja poprzez podniesienie poszczególnych różnic do kwadratu. Otrzymuje się w ten sposób kolejną miarę dyspersji cechy wokół poziomu średniego. Sumując podniesione do kwadratu odchylenia od średniej i dzieląc przez liczbę jednostek statystycznych (czyli uśredniając), otrzymujemy najbardziej popularną miarę rozproszenia – ***wariancję***. Nie nadaje się jednak ona do interpretacji oraz do porównań ze średnią, ponieważ jej miana są zawsze w kwadratach jednostki cechy. Bezpośrednio interpretowalną miarą rozproszenia wartości cechy jest pierwiastek z wariancji zwany ***odchyleniem standardowym***. Ma ono wartość tego samego rzędu co badana cecha oraz jest w takich samych jednostkach miary.

Jeżeli badacz chce porównać dyspersję cechy różnych zbiorowości, wówczas wystarczy użyć do tego celu wariancji. Pierwiastkowanie sprowadzające wariancję do odchylenia standardowego nie jest niezbędne, ponieważ własności odchylenia standardowego wynikają z własności wariancji[109]. Zatem wariancja jest jednym z najważniejszych parametrów używanych w statystyce i dyscyplinach ze statystyką związanych. Śmiało można powiedzieć, że jest to jeden z najważniejszych parametrów w analizie statystycznej, zarówno w teorii, jak i w praktyce.

Omówienie wariancji należy zacząć od podania jej definicji. **Wariancja jest to średnia arytmetyczna z kwadratów odchyleń wartości cechy od średniej arytmetycznej tej cechy w całej zbiorowości**. Wyznaczana jest ona, podobnie jak miary przeciętne, różnie, w zależności od postaci szeregu statystycznego. I tak dla szeregu prostego cechy skokowej stosowany jest wzór:

$$\sigma_x^2 = D^2(X) = s_x^2 = \frac{1}{n}\sum_{i=1}^{n}\left(x_i - \overline{x}\right)^2, \tag{5.5}$$

zaś dla szeregu rozdzielczego jeden z dwóch poniższych:

[109] Zbiorowość o większej wariancji ma także większe odchylenie standardowe niż druga porównywana populacja.

$$\sigma_x^2 = D^2(X) = s_x^2 = \frac{1}{n}\sum_{i=1}^{k}\left(x_i - \overline{x}\right)^2 n_i, \tag{5.6}$$

$$\sigma_x^2 = D^2(X) = \hat{s}_x^2 = \frac{1}{n-1}\sum_{i=1}^{k}\left(x_i - \overline{x}\right)^2 n_i, \tag{5.7}$$

gdzie $\sigma_x^2, D^2(X), s_x^2, \hat{s}_x^2$ – alternatywne oznaczenia wariancji w literaturze; używamy tylko jednego z tych symboli, by „poinformować" , że liczymy wariancję.

Pierwszego z wymienionych (wzór (5.6)) należy używać w przypadku, gdy szereg ma dużą liczebność ($n > 30$), zaś wzoru (5.7) do szeregów o niewielkiej liczebności ($n \leq 30$). W tym drugim przypadku uwzględniamy fakt, iż wykorzystaliśmy już raz dany szereg do policzenia średniej arytmetycznej. Mówimy, że straciliśmy jeden stopień swobody. Nieuwzględnienie tej poprawki powoduje tzw. obciążenie wariancji, czyli obarczenie jej błędem. Będzie ona zaniżona. W przypadku dużych zbiorowości nie gra to roli, albowiem błąd jest nieznaczny. W przypadku małej liczby badanych jednostek błąd może być na tyle duży, że może prowadzić do niepoprawnych wniosków bazujących na tej wariancji. Generalnie problem ten występuje w przypadku badań prowadzonych w oparciu o próby statystyczne, wówczas należy brać pod uwagę wzór (5.7). W niniejszym podręczniku mamy do czynienia wyłącznie z opisem statystycznym, czyli całymi populacjami, wobec tego nie będziemy korzystać ze wzoru (5.7). Sygnalizujemy tu jedynie istnienie tego typu problemu.

W przypadku szeregów rozdzielczych przedziałowych stosujemy wzór:

$$D^2(X) = s_x^2 = \frac{1}{n}\sum_{i=1}^{k}\left(\overset{o}{x}_i - \overline{x}\right)^2 n_i. \tag{5.8}$$

Ponieważ użycie środków przedziałów, podobnie jak w przypadku średniej, wprowadza pewien błąd[110], rekomenduje się użycie (zwłaszcza przy małej liczebności szeregu) tzw. poprawki Shepparda:

$$s_{skor}^2 = s_x^2 - \frac{h^2}{12}, \tag{5.9}$$

gdzie h – rozpiętość przedziału. Poprawki tej używa się do dokładniejszego policzenia wariancji w przypadku szeregów przedziałowych.

[110] Który jednak nie znosi się, jak w przypadku średniej, ze względu na podnoszenie różnic do kwadratów i eliminację ujemnych odchyleń.

Podstawowe własności wariancji:

1. $D^2(X) \geq 0$ – nieujemność.

   Własność ta wynika z faktu, że sumowane są kwadraty odchyleń od średniej. Wariacja równa się zeru jedynie w przypadku, gdy szereg jest stały, czyli wszystkie jednostki statystyczne mają tę samą wartość, co powoduje, że odchylenia od średniej są zerowe.
2. Wariancję można rozłożyć na różnicę średnich arytmetycznych:

$$D^2(X) = \overline{x^2} - \left(\overline{x}\right)^2$$

   Jest to różnica pomiędzy średnią policzoną dla kwadratów cechy a kwadratem średniej szeregu wyjściowego. Jest to jedna z najczęściej wykorzystywanych własności wariancji. Poniżej zostanie przeprowadzony dowód tej własności:

$$D^2\left(X\right) = \frac{1}{n}\sum_{i=1}^{k}(x_i - \overline{x})^2 n_i = \frac{1}{n}\sum_{i=1}^{k}(n_i x_i^2 - 2n_i x_i \overline{x} + n_i\left(\overline{x}\right)^2) =$$

$$= \frac{1}{n}\left[\sum_{i=1}^{k} n_i x_i^2 - 2\sum_{i=1}^{k} n_i x_i \overline{x} + \sum_{i=1}^{k} n_i\left(\overline{x}\right)^2\right] = \frac{\sum_{i=1}^{k} n_i x_i^2}{n} - 2\frac{\sum_{i=1}^{k} n_i x_i \overline{x}}{n} + \frac{\sum_{i=1}^{k} n_i\left(\overline{x}\right)^2}{n} =$$

$$= \overline{x^2} - 2\frac{\overline{x}\sum_{i=1}^{k} n_i x_i}{n} + \left(\overline{x}\right)^2 \cdot \frac{\sum_{i=1}^{k} n_i}{n} = \overline{x^2} - 2\overline{x}\cdot\overline{x} + \left(\overline{x}\right)^2 \cdot 1 = \overline{x^2} - \left(\overline{x}\right)^2.$$

3. $D^2\left(\frac{X}{c}\right) = \frac{D^2(X)}{c^2} \quad \left(\text{lub} \quad D^2\left(cX\right) = c^2 \cdot D^2(X)\right).$

   Własność ta oznacza, że podzielenie (pomnożenie) szeregu przez stałą c (c>1) powoduje zmniejszenie (zwiększenie) wariancji o tę samą wartość c, ale w drugiej potędze[111]. Dowód tej własności przedstawiono poniżej (dla uproszczenia przeprowadzono go w odniesieniu do szeregu nieważonego):

$$D^2\left(\frac{X}{c}\right) = \frac{1}{n}\sum_{i=1}^{n}\left[\frac{x_i}{c} - \overline{\left(\frac{x}{c}\right)}\right]^2 = \frac{1}{n}\sum_{i=1}^{n}\left[\frac{x_i}{c} - \frac{1}{c}\overline{(x)}\right]^2 = \frac{1}{n}\sum_{i=1}^{n}\left[\frac{1}{c}\left(x_i - \overline{(x)}\right)\right]^2 =$$

$$= \frac{1}{c^2}\frac{1}{n}\sum_{i=1}^{n}\left[\left(x_i - \overline{(x)}\right)\right]^2 = \frac{D^2(X)}{c^2}.$$

[111] W przypadku c<1 będzie to odpowiednio przy dzieleniu zwiększenie wariancji, a przy mnożeniu zmniejszenie wariancji.

4. $D^2(X \pm c) = D^2(X)$ – wariancja szeregu, do którego dodano (odjęto) stałą, nie zmienia się, dodanie stałej do każdej wartości cechy przesuwa wartości cechy w układzie współrzędnych wzdłuż osi X, nie zmieniając zakresu jej zmienności[112]:

$$D^2(X \pm c) = \frac{1}{n}\sum_{i=1}^{n}\left((x_i \pm c) - \overline{(x \pm c)}\right)^2 = \frac{1}{n}\sum_{i=1}^{n}\left((x_i \pm c - \overline{(x)} \pm c\right)^2 = \frac{1}{n}\sum_{i=1}^{n}\left(x_i - \bar{x}\right)^2 = D^2(X).$$

5. $D^2(X) = \overline{D_j^2(X)} + D^2(\overline{X_j})$ [113] – jeżeli badaną zbiorowość podzielono na *m* grup (podpopulacji), to ogólna wariancja cechy X może być rozłożona na dwie składowe: średnią arytmetyczną z wariancji wewnątrzgrupowych oraz wariancję grupowych średnich arytmetycznych (wariancję międzygrupową). Wagami są liczebności poszczególnych grup:

$$D^2(X) = \overline{D_j^2(X)} + D^2(\overline{X_j}) = \frac{\sum_{j=1}^{m} D_j^2(X) n_{\bullet j}}{\sum_{j=1}^{m} n_{\bullet j}} + \frac{\sum_{j=1}^{m}\left(\overline{X_j} - \overline{\overline{X}}\right)^2 n_{\bullet j}}{\sum_{j=1}^{m} n_{\bullet j}}. \qquad (5.10)$$

Inaczej można też ten wzór zapisać jako:

$$s_x^2 = \frac{1}{n}\sum_{j=1}^{m}\left(\bar{x}_j - \overline{\bar{x}}\right)^2 n_j \quad + \frac{1}{n}\sum_{j=1}^{m} s_j^2 n_j. \qquad (5.10\text{ a})$$

lub gdy grupy są numerowane (indeksowane) wierszami, używamy w zapisie numeru wiersza *i*:

$$s_x^2 = \frac{1}{n}\sum_{i=1}^{m}\left(\bar{x}_i - \overline{\bar{x}}\right)^2 n_i \quad + \frac{1}{n}\sum_{i=1}^{m} s_i^2 n_i. \qquad (5.10\text{ b})$$

Oba zapisy są równoprawne, albowiem to, jak układamy dane w tabeli, jest wyborem badacza i nie wpływa na wyniki obliczeń, co oznacza, że możemy sumować po wierszach (*i*) lub kolumnach (*j*) w zależności od sposobu konstrukcji tabeli z danymi. Przypomnijmy też, że $\sum_{j=1}^{m} n_{\bullet j} = \sum_{i=1}^{m} n_{i\bullet} = n$, zaś $\overline{X}_j$ to średnie

[112] Przypomnijmy, że zmieni średnią arytmetyczną o tę stałą *c* (zwiększy lub zmniejszy), ale nie zmieni rozrzutu wartości cechy wokół średniej.

[113] Jest to bardzo istotna własność, mająca zastosowanie w analizie statystycznej dwóch (lub więcej) cech.

z poszczególnych grup, $\overline{\overline{\overline{X}}}$ – średnia dla całej populacji, a $D_j^2(X)$ to wariancje w poszczególnych grupach.

Jak już wspomniano, wariancja nie jest miarą, której opis można interpretować. Właściwą interpretowalną miarą dyspersji jest odchylenie standardowe wyznaczone zgodnie poniższym wzorem:

$$D(X) = \sqrt{D^2(X)} \text{ lub } s = \sqrt{s^2}. \tag{5.11}$$

Formalnie interpretacja odchylenia standardowego nie jest zbyt intuicyjna, albowiem jest to pierwiastek ze średniego kwadratu odchyleń od średniej arytmetycznej. Jeżeli rozkład jest w miarę standardowy (zbliżony do normalnego), wówczas można stosować, z dużą dozą ostrożności, powszechnie występującą w podręcznikach poniższą interpretację odchylenia standardowego.

Odchylenie standardowe mówi nam, *o ile średnio* wartości cechy poszczególnych jednostek zbiorowości różnią się od średniej w zbiorowości.

W podanych poniżej przykładach będziemy wobec tego stosować uproszczoną interpretację.

Przykład 5.6 (kontynuacja 5.2)

Analizując dane związane z wydatkami grupy przyjaciół na śniadanie, wariancja wyznaczana jest następująco (zgodnie ze wzorem (5.5)):

$$s^2 = \frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})^2 = \frac{1}{5}\left[(5-3,6)^2 + (3-3,6)^2 + (2-3,6)^2 + (4-3,6)^2 + (4-3,6)^2\right] =$$

$$= \frac{1}{5}\left(1,4^2 + 0,6^2 + 1,6^2 + 0,4^2 + 0,4^2\right) = \frac{1}{5}\cdot(1,96 + 0,36 + 2,56 + 0,16 + 0,16) = \frac{1}{5}\cdot 5,2 = 1,04 \text{ zł}^2.$$

Otrzymany wynik wyrażony jest w złotówkach do kwadratu, dlatego trudno go zinterpretować merytorycznie. W związku z tym należy otrzymaną wartość spierwiastkować, obliczając w ten sposób odchylenie standardowe:

$$s_x = \sqrt{s_x^2} = \sqrt{1,04} = 1,0198 \approx 1,02 \text{ zł}.$$

Obliczyliśmy, że średnio wydatki przyjaciół różniły się o 1,02 zł od średniego wydatku wynoszącego (przypomnijmy) 3,6 zł. Informacja o średniej i odchyleniu standardowym w zbiorowości daje nam dobre pojęcie o tym, co się dzieje z cechą u większości jednostek populacji. Bliżej o tym za chwilę, gdy zdefiniujemy tzw. typowy obszar zmienności cechy.

■

Przykład 5.7 (kontynuacja 4.18)

Wróćmy do przykładu ocen otrzymanych przez studentów na kolokwium i określmy stopień ich zmienności. Przypomnijmy, że oceny przedstawiały się następująco: 2; 2; 3; 3; 3,5; 4; 4; 4; 4,5; 5. Ich średnia wyniosła 3,5.

Wyliczenia związane z wariancją w tym przykładzie, podobnie jak w poprzednim, bazują na wzorze 5.5. Analogicznie do odchylenia przeciętnego należy zacząć od wyznaczenia różnic pomiędzy poszczególnymi ocenami a średnią ocen. Wyliczenia zostały przedstawione w tabeli 5.3.

Tabela 5.3. Oceny z kolokwium – wyliczenia związane z wariancją

x_i	$x_i - \bar{x}$	$(x_i - \bar{x})^2$
2	–1,5	2,25
2	–1,5	2,25
3	–0,5	0,25
3	–0,5	0,25
3,5	0,0	0,00
4	0,5	0,25
4	0,5	0,25
4	0,5	0,25
4,5	1,0	1,00
5	1,5	2,25
Suma	0,0	**9,00**

Źródło: opracowanie własne.

Otrzymaną w tabeli 5.3 sumę kwadratów odchyleń od średniej wstawiamy do wzoru na wariancję (5.5) i otrzymujemy:

$$s_x^2 = \frac{1}{n}\sum_{i=1}^{k}(x_i - \bar{x})^2 = \frac{1}{10} \cdot 9 = 0{,}9 \text{ (stopnia)}^2,$$

a następnie wyznaczamy odchylenie standardowe:

$$s_x = \sqrt{s_x^2} = \sqrt{0{,}9} = 0{,}94868 \approx 0{,}95 \text{ stopnia}.$$

Wyliczona wartość oznacza, że przeciętnie oceny dziesięciu studentów różniły się od średniej o 0,95 stopnia.

■

Przykład 5.8 (kontynuacja przykładu 3.3)

Wróćmy do przykładu (3.3), gdzie badano rozkład wagi studentek. Wyznaczmy teraz, o ile średnio ta waga różniła się od wagi średniej wynoszącej 54,84 kg. Zaczniemy, jak zawsze, od obliczenia wariancji. Pomocnicze obliczenia zawarte zostały w tabeli 5.4.

Tabela 5.4. Rozkład wagi studentek – obliczenia związane z wariancją

$(x_{i_0} - x_{i_1})$	n_i	$\overset{o}{x}_i$	$\overset{o}{x}_i - \bar{x}$	$\left(\overset{o}{x}_i - \bar{x}\right)^2 n_i$
42-46	2	44	– 10,84	235,01
46-50	16	48	– 6,84	748,57
50-54	26	52	– 2,84	209,71
54-58	30	56	1,16	40,37
58-62	19	60	5,16	505,89
62-66	5	64	9,16	419,53
66-70	2	68	13,16	346,37
Suma	100			**2 505,45**

Źródło: opracowanie własne.

Podstawiając sumę z tabeli 5.4 do wzoru (5.8), otrzymujemy:

$$s_x^2 = \frac{1}{n}\sum_{i=1}^{k}\left(\overset{o}{x}_i - \bar{x}\right)^2 n_i = \frac{1}{100}\cdot 2505{,}45 = 25{,}0545\ (\text{kg})^2.$$

Wynik ten należy jeszcze przekształcić na odchylenie standardowe:

$$s_x = \sqrt{s_x^2} = \sqrt{25{,}0545} = 5{,}00545 \approx 5{,}01\,\text{kg}.$$

Obliczyliśmy, że średnio studentki mają wagę różniącą się od średniej populacyjnej wynoszącej 54,84 kg o 5,01 kg.

■

Przykład 5.9 (kontynuacja 4.4)

Biorąc pod uwagę dane dotyczące czasu biegu na 60 m dla 28 uczniów w pewnej szkole średniej, należy wyznaczyć odchylenie standardowe otrzymanych wyników. Wiadomo, że średnia czasu biegu wynosi 8,875 sekundy (przykład 4.5).

Aby wskazać, jaka była dyspersja czasu biegu uczniów, należy rozpocząć od wyliczenia wariancji. Wyniki z tym związane przedstawiono w tabeli 5.5.

Tabela 5.5. Czas biegu na 60 m uczniów szkoły średniej – wyznaczenie wariancji

$(x_{i_0} - x_{i_1})$	n_i	$\overset{o}{x}_i$	$\overset{o}{x}_i - \bar{x}$	$\left(\overset{o}{x}_i - \bar{x}\right)^2 n_i$
7,5-8,0	3	7,75	– 1,125	3,797
8,0-8,5	6	8,25	– 0,625	2,344
8,5-9,0	8	8,75	– 0,125	0,125
9,0-9,5	5	9,25	0,375	0,703
9,5-10,0	4	9,75	0,875	3,063
10,0-10,5	2	10,25	1,375	3,781
Suma	28			13,813

Źródło: opracowanie własne na podstawie danych umownych.

Podstawiając otrzymane w tabeli 5.5 wyliczenia do wzoru (5.8), wyznaczymy wariancję:

$$s_x^2 = \frac{1}{n}\sum_{i=1}^{k}\left(\overset{o}{x}_i - \bar{x}\right)^2 n_i = \frac{1}{28}\cdot 13{,}813 = 0{,}493\,\text{s}^2,$$

a następnie odchylenie standardowe:

$$s_x = \sqrt{s_x^2} = \sqrt{0{,}493} = 0{,}702\,\text{s}.$$

Otrzymana wartość odchylenia standardowego oznacza, że przeciętnie czasy biegu analizowanych uczniów różniły się od średniej biegu o 0,702 sekundy.

■

Przykład 5.10

Przedsiębiorstwo X posiada trzy linie produkcyjne. Na każdej z nich wytwarzany jest ten sam wyrób. Poddano analizie czas produkcji wybranego elementu na poszczególnych li

Tabela 5.6. Czas produkcji wybranego elementu

Linia	Średni czas wykonania [w min.] $\bar{x}_i$	Odchylenie standardowe czasu wykonania [w min.] s_i	Liczba pomiarów [w szt.] n
I	25	2,1	240
II	20	2,2	160
III	22	1,9	100

Źródło: dane umowne.

Korzystając ze zgromadzonych w ten sposób danych, należy wyznaczyć średnią i odchylenie standardowe czasu wytwarzania analizowanego elementu dla całego przedsiębiorstwa.

Pierwszą miarą, jaką należy wyznaczyć, jest średnia arytmetyczna czasu produkcji wybranego elementu, jednak ze względu na fakt, że liczba pomiarów dla każdej linii była inna, użyjemy średniej ważonej (wzór (4.2)), gdzie wagami są ilości wytworzonych produktów na każdej linii produkcyjnej. Wyliczenia pomocnicze przedstawiono w tabeli 5.7.

Tabela 5.7. Czas produkcji wybranego elementu – wyliczenia pomocnicze

Linia	$\bar{x}_i$	s_i	n_i	$\bar{x}_i n_i$	$\bar{x}_i - \bar{\bar{x}}$	$\left(\bar{x}_i - \bar{\bar{x}}\right)^2 n_i$	$s_i^2 n_i$
I	25	2,1	240	6 000	2,2	1 161,6	1 058,4
II	20	2,2	160	3 200	-2,8	1 254,4	774,4
III	22	1,9	100	2 200	-0,8	64,0	361,0
Suma			**500**	**11 400**		**2 480,0**	**2 193,8**

Źródło: opracowanie własne na podstawie danych umownych.

Podstawiając do wzoru (4.2), otrzymujemy następującą średnią czasu wytwarzania produktu dla całego przedsiębiorstwa:

$$\bar{\bar{x}} = \frac{1}{n}\sum_{i=1}^{k} \bar{x}_i n_i = \frac{1}{500} \cdot 11\,400 = 22{,}8\,\text{min}.$$

Wykorzystaliśmy tu trzecią własność średniej arytmetycznej, czyli możliwość jej policzenia jako średniej ze średnich grupowych. W kolejnym kroku policzono wariancje średnich grupowych (dla poszczególnych linii) i średnią z wariancji czasu realizacji badanego wyrobu na kolejnych liniach.

Wariancja średnich, czyli dyspersja średnich czasów produkcji dla poszczególnych linii wokół średniego czasu produkcji dla całego zakładu, wyniosła:

$$D^2(\bar{X}_i) = s^2(\bar{X}_i) = \frac{1}{n}\sum_{i=1}^{k}\left(\bar{x}_i - \bar{\bar{x}}\right)^2 n_i = \frac{1}{500} \cdot 2\,480{,}0 = 4{,}96\ \text{(min)}^2.$$

Średnia z wariancji grupowych, czyli uśredniona wariancja dla wszystkich linii łącznie, wynosi:

$$\overline{D_i^2(X)} = \frac{1}{n}\sum_{i=1}^{k} s_i^2 n_i = \frac{1}{500} \cdot 2193{,}8 = 4{,}39\ \text{(min)}^2.$$

Suma tych dwóch składowych tworzy wariancję czasu realizacji wyrobu dla całego przedsiębiorstwa (wzór (5.10)) i wynosi:

$$s_x^2 = \frac{1}{n}\sum_{i=1}^{m}\left(\bar{x}_i - \bar{\bar{x}}\right)^2 n_i \; + \frac{1}{n}\sum_{i=1}^{m} s_i^2 n_i = 4{,}96 + 4{,}39 = 9{,}35\,(\text{min})^2,$$

zaś odchylenie standardowe:

$$s_x = \sqrt{s_x^2} = \sqrt{9{,}35} = 3{,}06\,\text{min}.$$

Możemy więc stwierdzić, że średni czas produkcji badanego wyrobu w przedsiębiorstwie wynosi 22,8 minut, zaś przeciętnie poszczególne czasy produkcji różniły się od czasu średniego o 3,06 min.

■

Przejdźmy teraz do następnego pojęcia, w którym połączymy informacje o średniej i odchyleniu standardowym w nową miarę, dającą jeszcze więcej informacji o rozkładzie cechy w zbiorowości.

5.1.3. Typowy obszar zmienności i współczynnik zmienności

W oparciu o średnią i odchylenie standardowe lub przeciętne wyznaczany jest **typowy obszar zmienności** badanej cechy. Jest on konstruowany jako przedział, krańce którego wyznaczamy jako: $\bar{x} \pm s_x$ (lub $\bar{x} \pm d_x$), czyli średnia plus i minus odchylenie standardowe (średnia plus i minus odchylenie przeciętne), co pełniej zapisuje się jako:

$$\bar{x} - s_x < x_{typ} < \bar{x} + s_x, \qquad 5.12a)$$

$$\bar{x} - d_x < x_{typ} < \bar{x} + d_x. \qquad (5.12b)$$

Łącząc informację o średniej z jej odchyleniem standardowym (lub przeciętnym), otrzymujemy znacznie pełniejszy obraz rozkładu.

Obszar typowy mówi nam, z jakiego przedziału wartości przyjmuje ponad połowa jednostek statystycznych badanej zbiorowości, czyli co najmniej pięćdziesiąt procent plus jedna jednostka[114].

Jeżeli rozkład badanej cechy jest rozkładem normalnym[115], wówczas klasyczny typowy obszar zmienności zawiera 2/3 jednostek statystycznych z populacji.

[114] Jest to powszechnie przyjmowana interpretacja obszaru typowego, o ile w ogóle autorzy podręczników go interpretują. Nie do końca jednak jest jasne, skąd się wzięło podane 50%. Autorki nie znalazły źródła tego stwierdzenia.

[115] Dokładniej o rozkładzie normalnym Czytelnik może dowiedzieć się z podręczników statystyki matematycznej. Podanie postaci funkcyjnej rozkładu wymagałoby wprowadzenia pojęć z zakresu rachunku prawdopodobieństwa i statystyki matematycznej, więc to pominiemy. Kształt krzywej

Uczulamy Czytelnika, że jest to prawdą tylko w przypadku rozkładu normalnego. W wielu książkach niestety nie podaje się tego założenia, co może wprowadzać Czytelnika w błąd i wyrobić przekonanie, że jest to zawsze prawdą.

Absolutne miary dyspersji niezbyt nadają się do przeprowadzania porównań dyspersji tej samej cechy w różnych zbiorowościach i kompletnie nie nadają się do porównań dyspersji różnych cech w tej samej zbiorowości ze względu na różniące się poziomy średnie zjawiska i różne miana cech. Sama wariancja w miarę poprawnie pokaże, w którym szeregu ta sama cecha jest bardziej lub mniej zróżnicowana, pod warunkiem że średnie poziomy w obu szeregach znacznie się nie różnią

W przypadku takich porównań *używa się względnych miar zmienności*, które są niemianowane i odnoszą się do przeciętnego poziomu zjawiska (zjawisk) w populacji (populacjach). Najczęściej poziom przeciętny jest reprezentowany przez średnią arytmetyczną, ale w wielu wypadkach rekomenduje się użycie mediany, zwłaszcza tam, gdzie średnia niezbyt dobrze reprezentuje badane zjawisko[116]. W klasycznej postaci współczynnik zmienności występuje w dwóch wersjach, w zależności od tego, czy użyto odchylenia przeciętnego, czy standardowego jako miary dyspersji. I tak, gdy obliczono odchylenie przeciętne stosowany jest następujący wzór:

$$V_d = \frac{d}{|\bar{x}|}100\%, \tag{5.13}$$

zaś w przypadku, gdy użyto odchylenia standardowego, mamy[117]:

$$V_s = \frac{s_x}{|\bar{x}|}100\%. \tag{5.14}$$

Im wyższa jest, wartość współczynnika zmienności, tym cecha jest silniej zróżnicowana w danej zbiorowości. Podkreślmy, że tych współczynników używa się zamiennie, albowiem mają podobną interpretację. Mówią nam, jak cecha rozkłada się wokół średniej arytmetycznej. W badaniach empirycznych często chcemy określić siłę zróżnicowania zbiorowości w kategoriach słabe, umiarkowane, silne. Nie ma w literaturze jednoznacznego podziału wartości współczynnika zmienności

---

rozkładu normalnego przypomina dzwon, często więc nazywa się go krzywą dzwonową lub krzywą Gaussa. Okazało się, że wiele zjawisk zachodzących w przyrodzie, a następnie w naukach społecznych ma rozkład zgodny z krzywą tego rozkładu, dlatego określono go jako „normalny" (rozumiany jako norma). Bez dalszych szczegółowych wyjaśnień przyjmujemy więc, że jest to pewien wzorcowy, często spotykany typ rozkładu, do którego wiele zjawisk jest odnoszonych.

[116] Przypomnijmy, że ten problem był omawiany przy opisie średniej arytmetycznej i dominanty.

[117] Warto w tym miejscu podkreślić, że współczynnik zmienności nie jest liczony, gdy średnia wynosi zero, co wynika z podstawowych zasad matematyki.

na przedziały, które będziemy interpretować w kategoriach siły. Na ogół przyjmuje się, że współczynnik zmienności do 10% oznacza słabe zróżnicowanie cechy w zbiorowości, współczynniki powyżej 100% uznaje się za informujące o bardzo silnym zróżnicowaniu cechy. Wartości pomiędzy tymi liczbami interpretuje się na ogół jako zmienność umiarkowaną do silnej. Interpretacja ta jest oparta na badaniach empirycznych, a więc siłą rzeczy rozmyta i dyskusyjna[118]. Na tym jednak etapie opisu pozostawimy ją w takiej postaci.

Przykład 5.11 (kontynuacja 5.2)

Kontynuując analizę wydatków na śniadanie grupy przyjaciół, wyznaczymy obszar typowy i współczynnik zmienności tych wydatków. Przypomnijmy, że średnia i odchylenie standardowe wynosiły odpowiednio 3,6 zł oraz 1,02 zł. Korzystając ze wzoru (5.12), otrzymujemy:

$$\bar{x} - s_x < x_{typ} < \bar{x} + s_x,$$

$$3,60 - 1,02 < x_{typ} < 3,60 + 1,02,$$

$$2,58 < x_{typ} < 4,62 \text{ zł}.$$

Zatem większość przyjaciół (czyli ponad połowa badanych) wydało na śniadanie od 2,59 do 4,62 zł. Biorąc pod uwagę obserwacje uwzględniane w obliczeniach: 5, 3, 2, 4, 4, łatwo można zauważyć, że trzy z pięciu obserwacji znajduje się w obliczonym obszarze. Stanowi to 60% obserwacji.

Obliczmy teraz za pomocą wzoru (5.14) współczynnik zmienności:

$$V_s = \frac{s_x}{\bar{x}} 100\% = \frac{1,02}{3,60} 100\% = 28,3\%.$$

Otrzymana wartość oznacza, że zróżnicowanie przyjaciół ze względu na wydatki na śniadanie jest umiarkowane. Otrzymany wynik daje możliwość porównania dyspersji wydatków na śniadanie na przykład z kosztem obiadu czy kolacji. Zakładając, że wydatki na obiad po analogicznych obliczeniach dały średni koszt obiadu 12 zł z odchyleniem standardowym 1,8 zł, daje to współczynnik zmienności V_s = 15%. Współczynnik zmienności wydatków na śniadanie jest zdecydowanie

[118] Choć część autorów wyraża pogląd, że ze słabą dyspersją mamy do czynienia, gdy $V_s \leq 20\%$. Nie jest to kwestia jednoznaczna i zależy od typu badania, rodzaju badanych cech i subiektywnej oceny badacza. Z problemem tym badacz spotyka się zwłaszcza w badaniach porównawczych typu taksonomicznego i więcej na ten temat czytelnik znajdzie w podręcznikach z analizy wielowymiarowej czy w monografiach przedstawiających konkretne badania porównawcze wielu cech i zbiorowości, np. Panek [2009], Młodak [2006].

wyższy niż współczynnik wydatków na obiad, co oznacza, że koszt śniadania silniej różnicuje badanych przyjaciół, przy czym zróżnicowanie to jest umiarkowane. W przypadku obiadów przyjaciele wydają znacznie bardziej zbliżone do siebie sumy i można uznać, że koszty te są słabo zróżnicowane.

Gdybyśmy z kolei chcieli porównać otrzymany wynik z inną grupą przyjaciół i gdyby dla tej drugiej grupy współczynnik zmienności wydatków na śniadanie wyniósł na przykład 46%, to moglibyśmy stwierdzić, że grupa druga jest bardziej zróżnicowana pod względem wydatków niż nasza wyjściowa grupa przyjaciół. Wydatki w grupie drugiej bowiem różnią się znacznie bardziej pomiędzy przyjaciółmi niż w grupie pierwszej.

■

Przykład 5.12 (kontynuacja przykładu 3.3)

Biorąc pod uwagę wagę studentek kierunku zarządzanie przedstawioną w tabeli 4.11, należy wyznaczyć obszar typowy oraz współczynnik zmienności. W przykładzie 4.3 wyznaczono średnią wagi, która wyniosła 54,84 kg. Wyliczenia dotyczące odchylenia standardowego, które wyniosło 5,01 kg, przedstawiono w przykładzie 5.7. Stosując wzór (5.12), otrzymujemy:

$$\bar{x} - s_x < x_{typ} < \bar{x} + s_x,$$

$$54,84 - 5,01 < x_{typ} < 54,84 + 5,01,$$

$$49,83 < x_{typ} < 59,85 \text{ kg}.$$

Otrzymany zakres oznacza, że ponad połowa badanych studentek kierunku zarządzania ma w zaokrągleniu wagę w przedziale od 49,8 kg do 59,9 kg. Chcąc sprawdzić, czy rzeczywiście ponad połowa badanych miała wagę należącą do otrzymanego przedziału, należy wrócić do danych szczegółowych (przykład 3.3). Do wskazanego obszaru wagowego należało 67 dziewcząt ze 100 branych pod uwagę, co daje 67% wszystkich obserwacji. Zatem stwierdzamy, że znacznie ponad połowa badanych miała wagę z obliczonego typowego obszaru zmienności.

Wyznaczmy także współczynnik zmienności:

$$V_s = \frac{s_x}{\bar{x}} 100\% = \frac{5,01}{54,84} 100\% = 9,1\%.$$

Otrzymana wartość oznacza, że zróżnicowanie analizowanych studentek ze względu na ich wagę jest bardzo słabe czyli studentki mają wagi dosyć zbliżone do średniej populacyjnej, a co za tym idzie, i do siebie.

■

Przykład 5.13 (kontynuacja 4.4)

Uwzględniając dane dotyczące czasu biegu na 60 m dla 28 uczniów w pewnej szkole średniej, należy wyznaczyć obszar typowy i współczynnik zmienności. Wiadomo, że średnia czasu biegu wynosi 8,875 sekundy (przykład 4.5), zaś odchylenie standardowe 0,702 sekundy (przykład 5.8). Obszar typowy wyznaczamy następująco:

$$\overline{x} - s_x < x_{typ} < \overline{x} + s_x,$$

$$8,875 - 0,702 < x_{typ} < 8,875 + 0,702,$$

$$8,173 < x_{typ} < 9,577 \text{ sek.}$$

Otrzymany przedział wskazuje, że większość (ponad połowa) branych pod uwagę uczniów pokonało dystans 60 metrów w czasie od 8,173 sekundy do 9,577 sekund.

Wskaźnik zmienności wynosi:

$$V_s = \frac{s_x}{\overline{x}} 100\% = \frac{0,702}{8,875} 100\% = 7,9\%.$$

Otrzymany wynik oznacza bardzo słabe zróżnicowanie analizowanej grupy uczniów ze względu na czas biegu na 60 metrów. Otrzymany wynik można porównać z innymi osiągnięciami tych młodych ludzi w innych dziedzinach sportu (np. długość skoku w dal, czas biegu na 1000 metrów, długość rzutu piłką lekarską), czy też odnieść je do wyników uczniów innych klas.

■

Przykład 5.14

Dane przedstawione w tabeli 5.8 przedstawiają powierzchnię (cecha X) i liczbę mieszkańców poszczególnych województw (cecha Y).

Tabela 5.8. Podstawowe dane dotyczące województw w 2010 roku

Województwo	Powierzchnia [w tys. km^2]	Liczba ludności [w tys. osób]
Dolnośląskie	19,947	2 878
Kujawsko-pomorskie	17,972	2 070
Lubelskie	25,122	2 152
Lubuskie	13,988	1 011
Łódzkie	18,219	2 534

Województwo	Powierzchnia [w tys. km^2]	Liczba ludności [w tys. osób]
Małopolskie	15,183	3 310
Mazowieckie	35,558	5 243
Opolskie	9,412	1 029
Podkarpackie	17,846	2 104
Podlaskie	20,187	1 188
Pomorskie	18,310	2 240
Śląskie	12,333	4 636
Świętokrzyskie	11,711	1 266
Warmińsko-mazurskie	24,173	1 427
Wielkopolskie	29,826	3 419
Zachodniopomorskie	22,892	1 693
OGÓŁEM	**312,679**	**38 200**

Źródło: Rocznik Statystyczny Województw, 2011.

Na podstawie podanych informacji należy określić, która z badanych cech silniej różnicuje województwa: liczba ludności czy powierzchnia.

Wyznaczenie poszczególnych miar należy rozpocząć od średniej. Wiedząc, że łączna powierzchnia Polski to 312 679 km^2, którą w 2010 roku zamieszkiwało 38 200 tys. ludności, średnia powierzchnia wyznaczona ze wzoru 4.1 wynosi:

$$\bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i = \frac{1}{16} \cdot 312{,}679 = 19{,}542 \text{ tys. km}^2,$$

zaś średnia liczba ludności:

$$\bar{y} = \frac{1}{n}\sum_{i=1}^{n} y_i = \frac{1}{16} \cdot 38\,200 = 2\,387{,}5 \text{ tys.}$$

Przypomnijmy interpretację średniej arytmetycznej. Otrzymane wyniki oznaczają, że gdyby każde województwo miało taką samą powierzchnię i liczbę ludności, to wartości te byłyby równe wyliczonym średnim arytmetycznym, czyli średnia powierzchnia jednego województwa w Polsce to 19 542 km^2. W 2010 roku średnio województwo zamieszkiwało 2 387,5 tys. osób.

W kolejnym kroku policzmy wariancję i odchylenie standardowe. Niezbędne wyliczenia pomocnicze zostały ujęte w tabeli 5.9.

Tabela 5.9. Dane dotyczące województw – wyliczenia pomocnicze

Województwo	x_i	$x_i-\bar{x}$	$(x_i-\bar{x})^2$	y_i	$y_i-\bar{y}$	$(y_i-\bar{y})^2$
Dolnośląskie	19,947	0,4	0,164	2 878	490,5	240 590,25
Kujawsko-pomorskie	17,972	-1,6	2,465	2 070	-317,5	100 806,25
Lubelskie	25,122	5,6	31,136	2 152	-235,5	55 460,25
Lubuskie	13,988	-5,6	30,847	1 011	-1 376,5	1 894 752,25
Łódzkie	18,219	-1,3	1,75	2 534	146,5	21 462,25
Małopolskie	15,183	-4,4	19,001	3 310	922,5	851 006,25
Mazowieckie	35,558	16,0	256,512	5 243	2 855,5	8 153 880,25
Opolskie	9,412	-10,1	102,617	1 029	-1 358,5	1 845 522,25
Podkarpackie	17,846	-1,7	2,876	2 104	-283,5	80 372,25
Podlaskie	20,187	0,6	0,416	1 188	-1 199,5	1 438 800,25
Pomorskie	18,310	-1,2	1,518	2 240	-147,5	21 756,25
Śląskie	12,333	-7,2	51,97	4 636	2 248,5	5 055 752,25
Świętokrzyskie	11,711	-7,8	61,325	1 266	-1 121,5	1 257 762,25
Warmińsko-mazurskie	24,173	4,6	21,446	1 427	-960,5	922 560,25
Wielkopolskie	29,826	10,3	105,761	3 419	1 031,5	1 063 992,25
Zachodniopomorskie	22,892	3,4	11,223	1 693	-694,5	482 330,25
OGÓŁEM	**312,679**		**701,027**	**38 200**		**23 486 806,00**

Źródło: opracowanie własne na podstawie: Rocznik Statystyczny Województw, 2011.

Korzystając z przedstawionych w tabeli 5.9 wyliczeń, obliczamy wariancję i w konsekwencji odchylenie standardowe powierzchni województw:

$$s_x^2 = \frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})^2 = \frac{1}{16} \cdot 701{,}027 = 43{,}814 \text{ (tys. km}^2)^2,$$

$$s_x = \sqrt{s_x^2} = \sqrt{43{,}814} = 6{,}619 \text{ tys. km}^2$$

oraz liczby ludności:

$$s_y^2 = \frac{1}{n}\sum_{i=1}^{n}(y_i - \bar{y})^2 = \frac{1}{16} \cdot 23\,486\,806 = 1\,467\,925{,}38 \text{ (tys. osób)}^2,$$

$$s_y = \sqrt{s_y^2} = \sqrt{1\,467\,925{,}38} = 1\,211{,}58 \text{ tys. osób.}$$

Otrzymana wartość odchylenia standardowego oznacza, że przeciętnie województwa różnią się od średniej powierzchni (średniej liczby ludności) o plus/minus 6,619 tys. km^2 (o plus/minus 1 211,58 tys. osób). Zatem współczynniki zmienności dla powierzchni województw i liczby ludności wynoszą odpowiednio:

$$V_s^{pow} = \frac{s_x}{\bar{x}} 100\% = \frac{6,619 \text{ tys. km}^2}{19,542 \text{ tys. km}^2} 100\% = 33,9\%,$$

$$V_s^{lud} = \frac{s_y}{\bar{y}} 100\% = \frac{1\,211,58 \text{ tys. os.}}{2\,387,50 \text{ tys. os.}} 100\% = 50,7\%.$$

Z przedstawionych wyliczeń wynika, że województwa są silniej zróżnicowane ze względu na liczbę zamieszkującej je ludności niż powierzchnię. Zaś w przypadku obu cech dyspersje te są umiarkowane.

■

5.2. Pozycyjne miary dyspersji

Zmienność, czy też dyspersję można także badać, dysponując jedynie pozycyjnymi miarami opisu zbiorowości. W tym celu wykorzystuje się przede wszystkim kwartyle. **Podstawową miarą określającą zróżnicowanie w oparciu o kwartyle jest odchylenie ćwiartkowe.** Wyznaczane jest ono ze wzoru:

$$Q = \frac{Q_3 - Q_1}{2} = \frac{(Me - Q_1) + (Q_3 - Me)}{2}, \tag{5.15}$$

gdzie Q_1, Me, Q_3 to odpowiednio kwartyl pierwszy, mediana i kwartyl trzeci.

Otrzymana wartość odchylenia ćwiartkowego wyznaczana jest jako połowa różnicy pomiędzy skrajnymi wartościami cechy w środkowej połowie uporządkowanej zbiorowości. *Inaczej jest to średnia rozpiętość cechy w drugiej i trzeciej ćwiartce. Bada ona jedynie dyspersję cechy u 50% środkowych jednostek zbiorowości, co sprawia, że jest rzadko stosowana*.

Znajduje ona wykorzystanie głównie wówczas, gdy analizowany szereg posiada otwarte przedziały lub jest skrajnie asymetryczny (czyli gdy nie możemy lub nie powinniśmy liczyć średniej arytmetycznej).

Korzystając z wyliczeń związanych z odchyleniem ćwiartkowym, można wyznaczyć pozostałe miary zmienności, takie jak pozycyjny obszar typowy czy pozycyjny współczynnik zmienności. W obu tych miarach środkiem ciężkości (punktem odniesienia) jest mediana. Zastępując nią średnią ze wzoru (5.12), zaś odchy-

leniem ćwiartkowym odchylenie przeciętne lub standardowe, otrzymujemy **typowy pozycyjny obszar zmienności:**

$$Me - Q < x_{typ} < Me + Q. \tag{5.16}$$

Otrzymany przedział wskazuje, w jakim zakresie znajduje się większość obserwacji z dwóch środkowych ćwiartek badanej zbiorowości[119].

Zamieniając miary klasyczne na pozycyjne we wzorze (5.14), wyznaczamy **pozycyjny współczynnik zmienności:**

$$V_Q = \frac{Q}{Me} 100\%. \tag{5.17}$$

Współczynnik ten informuje, jakie jest zróżnicowanie, ze względu na analizowaną cechę środkowej części badanej zbiorowości, czyli jednostek pomiędzy kwartylem pierwszym a trzecim.

Przykład 5.15 (kontynuacja 4.13)

Wracając do przykładu o liczebności gospodarstw domowych w Polsce w 1998 roku (przykład 4.13), wyznaczymy odchylenie ćwiartkowe, pozycyjny obszar zmienności i pozycyjny współczynnik zmienności. Na podstawie wcześniejszych wyliczeń wiadomo, że mediana *(Me)* wyniosła 3 osoby, kwartyl pierwszy (Q_1) to 2 osoby, zaś trzeci (Q_3) to 4 osoby w gospodarstwie domowym. Podstawiając te dane do wzoru (5.15), otrzymujemy:

$$Q = \frac{Q_3 - Q_1}{2} = \frac{4-2}{2} = \frac{2}{2} = 1 \text{ osoba}.$$

Otrzymany wynik oznacza, że przeciętnie liczba osób w gospodarstwie domowym w Polsce w 1998 roku w obszarze zawężonym do dwóch środkowych ćwiartek różniła się od mediany o jedną osobę. Podstawiając otrzymany wynik do wzoru (5.16), otrzymujemy obszar typowy wynoszący:

$$Me - Q < x_{typ} < Me + Q,$$

$$3 - 1 < x_{typ} < 3 + 1,$$

$$2 < x_{typ} < 4 \text{ osób}.$$

[119] Przy czym jeżeli rozkład jest asymetryczny, o czym za chwilę, to *Me* nie leży w środku odcinka (Q1;Q3), a to oznacza, że ten przedział zawiera również obserwacje spoza centralnej połowy.

Oznacza on, że większość gospodarstw domowych z dwóch środkowych ćwiartek, czyli ponad 25% populacji, w 1998 roku składało się z 3 osób. Współczynnik zmienności wyznaczymy zaś następująco:

$$V_Q = \frac{Q}{Me}100\% = \frac{1}{3}100\% = 33,3\%.$$

Otrzymany wynik świadczy, że zróżnicowanie liczebności gospodarstw domowych w obszarze pomiędzy pierwszym a trzecim kwartylem jest umiarkowane.

■

Przykład 5.16 (kontynuacja przykładu 3.3)

Biorąc pod uwagę wagę studentek kierunku zarządzanie oraz korzystając z wyliczeń dotyczących kwartyli ($Q_1 = 51,08$ kg, $Me = 54,80$ kg, $Q_3 = 58,21$ kg), należy wyznaczyć pozycyjne miary zmienności.

Odchylenie ćwiartkowe wynosi:

$$Q = \frac{Q_3 - Q_1}{2} = \frac{58,21 - 51,08}{2} = \frac{7,13}{2} \approx 3,57 \text{ kg}.$$

Otrzymany wynik oznacza, że przeciętnie waga studentek z zakresu pomiędzy pierwszym a trzecim kwartylem różniła się od mediany wagi o 3,57 kg. Wobec tego pozycyjny obszar typowy wynosi:

$$Me - Q < x_{typ} < Me + Q,$$

$$54,80 - 3,57 < x_{typ} < 54,80 + 3,57,$$

$$51,23 < x_{typ} < 58,37 \text{ kg}.$$

Większość studentek, których waga zawierała się w obszarze pomiędzy pierwszym a trzecim kwartylem, ważyła od 51,23 kg do 58,37 kg[120]. Uwzględniając dane wyjściowe służące do konstrukcji szeregu, stwierdzamy, że 55 studentek ma wagę z wyznaczonego pozycyjnego obszaru typowego (tabela 3.18). Jest też on

[120] Zauważmy, że obszar typowy w tym przypadku *de facto* to odcinek pomiędzy Q1 a Q3. Górny kraniec obszaru typowego nieznacznie wychodzi poza Q_3 (czyli poza 50% centralnych obserwacji), co się często zdarza w przypadku szeregów asymetrycznych. W omawianym przypadku przekroczenie jest nieznaczne (niewiele ponad Q_3=58,21) i może wynikać z zaokrągleń. Istotne przekroczenie jednego z krańców środkowego odcinka (środkowych 50% obserwacji) świadczy o asymetrii środkowej rozkładu.

znacznie węższy od obszaru typowego wyznaczonego na podstawie miar klasycznych ($49{,}83 < x_{typ} < 59{,}85$ kg).

Pozycyjny współczynnik zmienności wyznaczany jest następująco:

$$V_Q = \frac{Q}{Me}100\% = \frac{3{,}57kg}{54{,}80kg}100\% = 6{,}5\%.$$

Zróżnicowanie studentek ze względu na wagę w zakresie powstałym po odrzuceniu 25% najniższych i 25% najwyższych wartości jest bardzo słabe. Przy czym zmienność w środkowej części zbiorowości jest niższa niż przy uwzględnianiu wszystkich pomiarów $(V_s = 9{,}1\%)$.

■

Przykład 5.17 (kontynuacja 4.5)

Analizując dane dotyczące czasu biegu na 60 m dla 30 uczniów w pewnej szkole średniej, należy wyznaczyć pozycyjne miary zmienności. Dla podanego szeregu wyznaczono wartości kwartyli, które wyniosły: $Q_1 = 8{,}21$ s, $Me = 8{,}75$ s, $Q_3 = 9{,}85$ s.

Odchylenie ćwiartkowe czasu biegu na 60 metrów wyznaczane jest następująco:

$$Q = \frac{Q_3 - Q_1}{2} = \frac{9{,}85 - 8{,}21}{2} = \frac{1{,}64}{2} = 0{,}82 \text{ s}.$$

Otrzymana wartość odchylenia ćwiartkowego oznacza, że przeciętnie uczniowie, którzy pokonali dystans 60 metrów, mieli wyniki różniące się od mediany o 0,82 sekundy. Wartość ta zostanie wykorzystana do wyznaczenia pozycyjnego obszaru typowego:

$$Me - Q < x_{typ} < Me + Q,$$

$$8{,}75 - 0{,}82 < x_{typ} < 8{,}75 + 0{,}82,$$

$$7{,}93 < x_{typ} < 9{,}57 \text{ s}.$$

Większość uczniów, których czas biegu na 60 metrów zawierał się w obszarze pomiędzy pierwszym a trzecim kwartylem, ukończyła ten bieg w czasie pomiędzy 7,93 sekundy a 9,75 sekundy.

Pozycyjny współczynnik zmienności wynosi:

$$V_Q = \frac{Q}{Me}100\% = \frac{0{,}82}{8{,}75}100\% = 9{,}4\%.$$

Zatem zróżnicowanie uczniów w środkowej części zbiorowości ze względu na czas biegu na 60 metrów jest bardzo słabe.

■

Kluczowe pojęcia

Zmienność, dyspersja, zróżnicowanie
Rozstęp
Odchylenie przeciętne
Wariancja
Odchylenie standardowe
Odchylenie ćwiartkowe
Klasyczny typowy obszar zmienności
Pozycyjny typowy obszar zmienności
Klasyczny współczynnik zmienności
Pozycyjny współczynnik zmienności

Pytania kontrolne

1. Co badają miary zmienności?
2. Jak klasyfikuje się miary dyspersji?
3. Jakie założenia towarzyszą klasycznym miarom zmienności?
4. Jaka jest treść poznawcza rozstępu?
5. Jaka część populacji jest analizowana za pomocą odchylenia ćwiartkowego?
6. Ile jednostek statystycznych mieści się w klasycznym, typowym obszarze zmienności?
7. Ile jednostek statystycznych mieści się w typowym medianowym obszarze zmienności?

6. Miary asymetrii i koncentracji

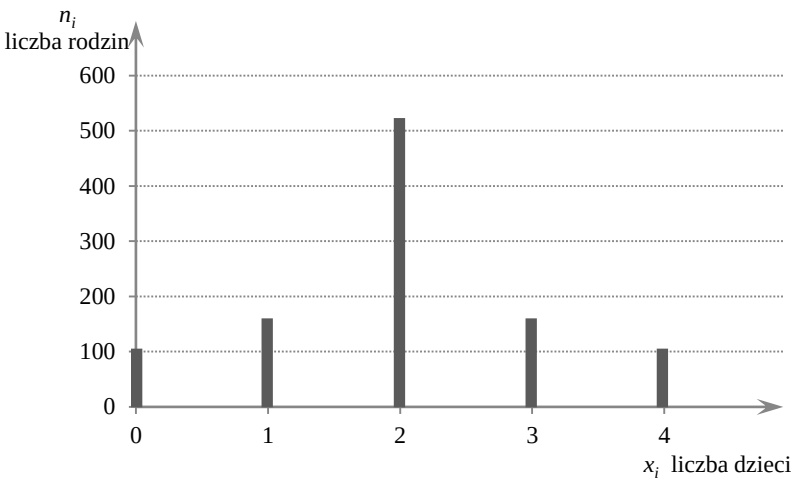
Przedstawione w poprzednich rozdziałach miary statystyczne miały na celu określenie tendencji centralnej rozkładu cechy bez wchodzenia w szczegóły danego rozkładu[121] oraz sposobu rozkładu poszczególnych odmian cechy wokół miary centralnej, czyli siły rozproszenia wartości cechy. Kolejnym etapem analizy jest określenie sposobu rozproszenia cechy wokół miary centralnej, czyli zbadanie, jaka część jednostek statystycznych ma wartości cechy przekraczające (powyżej lub poniżej) poziom przeciętny oraz w jakim stopniu są one skoncentrowane względem poziomu przeciętnego lub czy są rozłożone równomiernie w całej badanej zbiorowości. O sposobie rozłożenia poszczególnych wartości cechy wokół miary centralnej mówią nam tzw. miary asymetrii, zaś o stopniu skupienia wokół wartości przeciętnej lub równomierności rozłożenia wartości cechy w zbiorowości miary koncentracji.

6.1. Miary asymetrii

Pojęcie symetrii czy jej braku w rozkładzie (co w statystyce nazywamy asymetrią) jest terminem wziętym z geometrii. Używaliśmy już tego pojęcia intuicyjnie w rozdziale 4. Podstawową informację o symetryczności lub asymetryczności rozkładu można uzyskać, porównując miary tendencji centralnej. ***Szereg symetryczny charakteryzuje się tym, że występuje w nim tyle samo jednostek poniżej co powyżej wartości średniej arytmetycznej*. Sprowadzając problem do analizy parametrów, możemy powiedzieć, że dominują wartości cechy równe średniej arytmetycznej[122].**

Popatrzmy na wykres 6.1 przedstawiający rozkład liczby dzieci w rodzinie w przebadanej populacji $n = 1000$ rodzin. Zauważmy, że częstość występowania poszczególnych odmian cechy przed średnią rozkładu jest zwierciadlanym odbiciem częstości powyżej średniej.

[121] Chcąc wiedzieć, w którym mieście ludzie mają wyższe zarobki, porównamy średnie zarobki, a nie zarobek każdej jednostki statystycznej jednego miasta z zarobkami jednostek drugiego miasta.
[122] Przypominamy, że w niniejszym podręczniku omawiamy rozkłady proste, czyli jednomodalne.



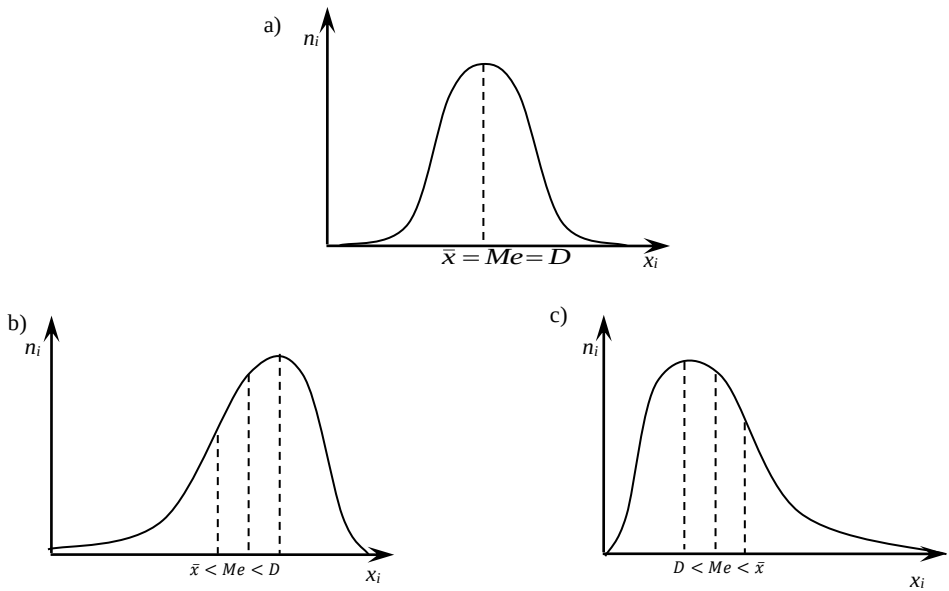
Wykres 6.1. Rozkład liczby dzieci w rodzinie

Źródło: opracowanie własne na podstawie danych umownych.

Jak widać z powyższego wykresu (6.1), średnia (oraz równa jej dominanta i mediana) wynosi dwoje dzieci i takich rodzin jest pięćset. Po jednym dziecku ma tyle samo rodzin co trójkę dzieci, czyli po sto pięćdziesiąt. Tyle samo też rodzin (sto) nie ma dzieci co ma ich czworo. *Taki rozkład nazywamy symetrycznym*, można powiedzieć, że występuje tu idealna symetria (co wynika z faktu, że jest to sztuczny przykład). W rzeczywistości społeczno-gospodarczej nie występują rozkłady idealnie symetryczne. Te ostatnie występują w teorii i stąd mówimy o rozkładach teoretycznych i empirycznych. Te ostatnie opisują realne procesy, obserwacja których jest często wycinkowa, obarczona błędem pomiaru i agregacji danych związanej między innymi z konstrukcją szeregu statystycznego. Stąd rozkłady empiryczne nie są idealne, takie jak zakłada teoria. Problem ten dotyczy zarówno rozkładów cech skokowych, jak i cechy ciągłej.

Graficzna prezentacja problemu symetrii i jej braku dla cechy ciągłej została przedstawiona na rysunku 6.2 a. Widać na nim, że wartości cechy rozkładają się z jednakową częstotliwością po obu stronach wartości średniej, która jest jednocześnie wartością dominującą oraz medianą $(\bar{x} = Me = D)$. Przykładem takiego rozkładu symetrycznego ciągłego jest **rozkład normalny**, nazywany również rozkładem Gaussa. Jego wykres nosi często nazwę, ze względu na swój kształt, krzywej dzwonowej. Jest to jeden z najważniejszych rozkładów w badaniach statystycznych, jako że odzwierciedla on rozkład wielu naturalnych cech, takich jak wzrost czy waga ludzi określonego wieku i płci czy określonego gatunku zwierząt. Wiele zjawisk ekonomicznych również ma kształt krzywej dzwonowej.

Jest to rozkład, własności którego zostały bardzo szczegółowo opracowane od strony teoretycznej. Duża część narzędzi analizy statystycznej bazuje na nim lub na rozkładach pochodnych. Posiada on wygodne własności, do których zalicza się fakt symetryczności, mówiący, iż jego średnia, mediana i dominanta są równe, tworzące maksimum rozkładu. W oddaleniu od tej wartości znajdują się punkty przegięcia, które obejmują obszar zawierający 2/3 całej zbiorowości (ok. 68%). Natomiast w zakresie pomiędzy wartościami $\bar{x}-3s_x$ i $\bar{x}+3s_x$ (średnia plus/minus trzy odchylenia standardowe) mieści się prawie cała zbiorowość (99,7%)[123]. Zatem, jeżeli nie zna się obszaru zmienności danego rozkładu, a posiada jedynie informacje o jego podstawowych parametrach, to zgodnie z zasadą trzech sigm można przyjąć, że jest to wskazany zakres zmienności powstały przez przesunięcie średniej o trzy odchylenia standardowe w obie strony[124].



Rysunek 6.1. Przedstawienie różnych typów rozkładów: a) rozkładu symetrycznego, b) rozkładu o asymetrii lewostronnej, c) rozkładu o asymetrii prawostronnej

Źródło: opracowanie własne.

[123] Tzw. twierdzenie o trzech sigmach (sigma – litera oznaczająca odchylenie standardowe).
[124] Szczegółowo rozkład ten jest omawiany w podręcznikach wnioskowania statystycznego.

Szereg, który nie jest symetryczny, nazwiemy szeregiem asymetrycznym. Będzie więc to każdy szereg, w którym odmiany cechy powyżej (lub poniżej) średniej będą występowały częściej niż poniżej (lub odpowiednio powyżej).

Cechą charakterystyczną takich szeregów jest fakt, iż dominanta i mediana nie są równe średniej arytmetycznej. Istnienie tej różnicy będzie więc dobrym wskaźnikiem istnienia bądź nie symetrii w szeregu statystycznym (rozkładzie).

Jeżeli dominanta i średnia arytmetyczna (oraz automatycznie mediana) pokrywają się, wówczas mówimy, że szereg jest symetryczny. Jeżeli dominanta i średnia arytmetyczna różnią się od siebie, oznacza to, że szereg jest asymetryczny.

Rozróżniamy w związku z tym dwa **rodzaje asymetrii**:

- asymetrię dodatnią, często zwaną prawostronną,
- asymetrię ujemną często zwaną lewostronną.

Asymetria rozkładu jest dodatnia, gdy w rozkładzie przeważają (czyli dominują) jednostki o wartości cechy poniżej średniej, co można zapisać jako: $\bar{x} > D,$ czyli $\bar{x} - D > 0,$ co oznacza, że bardzo wysokie wartości cechy występują rzadziej niż niskie. Graficznym wyrazem tego jest długi „ogon" na wykresie rozkładu, taki jak na rysunku 6.1c. Na tym wykresie „ogon" usytuowany jest z prawej strony patrzącego, stąd częste określenie rozkład prawostronny.

Analogicznie **asymetria rozkładu jest ujemna, gdy w rozkładzie przeważają (czyli dominują) jednostki o wartości cechy powyżej średniej,** co można zapisać jako: $\bar{x} < D,$ czyli $\bar{x} - D < 0,$ co oznacza, że bardzo niskie wartości cechy występują rzadziej niż wysokie, częściej występują wartości powyżej średniej. Graficznym wyrazem tego jest długi „ogon" na wykresie rozkładu, taki jak na rysunku 6.1b. Ogon ten na wykresie jest z lewej strony patrzącego, stąd częste określenie rozkład lewostronny.

Podsumowując, możemy mieć trzy rodzaje ustawień średniej i dominanty (a co za tym idzie – i mediany) względem siebie, co wykorzystuje się do konstrukcji podstawowego wskaźnika asymetrii:

$$W = \bar{x} - D. \tag{6.1}$$

1. Jeśli $\bar{x} = D, \quad to \quad W = 0,$ co oznacza, że rozkład jest **symetryczny**[125], przy czym $(\bar{x} = Me = D)$.
2. Gdy $W = \bar{x} - D < 0,$ oznacza to, że rozkład jest **asymetryczny ujemnie** (średnia z lewej strony dominanty), czyli występuje asymetria lewostronna, przy czym zachodzi relacja $\bar{x} < Me < D,$ co przedstawia rysunek 6.1 b.

[125] Jest to pewne uproszczenie, albowiem jeżeli rozkład jest symetryczny, to $W = 0$, ale $W = 0$ nie jest równoważne symetrii.

3. Gdy $W = \bar{x} - D > 0$, oznacza to, że rozkład jest **asymetryczny dodatnio** (średnia z prawej strony dominanty), czyli występuje asymetria prawostronna, przy czym zachodzi relacja[126] $\bar{x} > Me > D$, co przedstawia rysunek 6.1 c.

Za pomocą tego wskaźnika można określić istnienie symetrii rozkładu lub asymetrii i jej rodzaj (można też spotkać pojęcie „kierunek asymetrii"). Jednak ze względu na to, iż wartość wskaźnika zależy od jednostek pomiaru cechy (wyrażony jest w jednostkach bezwzględnych), nie nadaje się do porównań oraz do określania siły tejże asymetrii.

W tym celu przekształcamy go do postaci względnej poprzez podzielenie przez jedną z dwóch klasycznych miar dyspersji, eliminując w ten sposób jednostki miary oraz uzyskując pewien punkt odniesienia umożliwiający, do pewnego stopnia, ocenę siły asymetrii. W ten sposób tworzony jest współczynnik asymetrii. Miara ta informuje nie tylko o kierunku, ale również o sile asymetrii. Ze względu na fakt ujęcia w niej zarówno miar klasycznych (średniej i odchylenia standardowego), jak i pozycyjnych (dominanty) nazywany jest **mieszanym współczynnikiem asymetrii** (skośności) i wyznaczany następująco:

$$A_s = \frac{\bar{x} - D}{s_x} \tag{6.2}$$

lub

$$A_d = \frac{\bar{x} - D}{d}. \tag{6.3}$$

Jak widać, punktem odniesienia może być odchylenie standardowe (najczęściej) lub odchylenie przeciętne. Pamiętając, że odchylenie przeciętne jest zawsze mniejsze od standardowego, wnioskujemy, że współczynnik asymetrii oparty na odchyleniu przeciętnym będzie zawsze wykazywał większą asymetrię niż oparty na odchyleniu standardowym. Współczynnik ten dla szeregu symetrycznego ma wartość równą zero. Nie ma on górnej granicy, w związku z czym interpretacja siły asymetrii jest umowna. Im dalej od zera, tym asymetria rozkładu jest silniejsza. Przypominamy, że znak mówi nam o kierunku asymetrii, a nie o sile. Jak twierdzi Szulc (1967, s. 246), miara ta jedynie przy bardzo silnej asymetrii przekracza wartość ±1. Jednak należy pamiętać, że nie daje on poprawnych wyników, gdy w analizowanym szeregu występują problemy z dominantą i/lub średnią omówione w rozdziale czwartym.

[126] Eleganckie i proste wyjaśnienie tych relacji czytelnik znajdzie w podręczniku Szulca (1967, s. 196-200).

Omawiany miernik jest bezużyteczny[127], gdy rozkład jest skrajnie asymetryczny, czyli gdy dominanta jest wartością skrajną rozkładu, co oznacza de facto jej brak, biorąc pod uwagę definicję dominanty.

Podczas wyznaczania stopnia asymetrii można również skorzystać z innej własności rozkładu, a mianowicie tej, że *w rozkładach symetrycznych wszystkie nieparzyste momenty centralne są równe zeru,* stąd też wielkość każdego momentu nieparzystego może służyć jako miara braku symetrii, czyli asymetrii.

Wyjaśnijmy sobie wobec tego pojęcie momentu. **Moment jest to parametr opisujący rozkład. W statystyce istnieją dwa podstawowe typy momentów: *zwykłe i centralne*.**

Momentem zwykłym *r*-tego rzędu nazywamy parametr opisu rozkładu zmiennej będący średnią arytmetyczną wartości cechy podniesionych do *m*-tej potęgi. Dla prostego szeregu punktowego zapisujemy moment zwykły jako:

$$m_r = \frac{1}{n}\sum_{i=1}^{n} x_i^r. \tag{6.4}$$

Dla punktowego szeregu rozdzielczego przybiera postać wzoru na średnią ważoną:

$$m_r = \frac{1}{n}\sum_{i=1}^{k} x_i^r n_i \quad \text{lub} \quad m_r = \sum_{i=1}^{k} x_i^r f_i. \tag{6.5}$$

W przypadku szeregów rozdzielczych przedziałowych stosujemy jeden ze znanych już czytelnikowi zapisów:

$$m_r = \frac{1}{n}\sum_{i=1}^{k}\left(\overset{o}{x}_i\right)^r n_i \quad \text{lub} \quad m_r = \sum_{i=1}^{k}\left(\overset{o}{x}_i\right)^r f_i. \tag{6.6}$$

Warto zaznaczyć, że dla $r = 1$ moment zwykły pierwszego rzędu ma postać $m_1 = \frac{1}{n}\sum_{i=1}^{k} x_i n_i$, czyli jest to znana czytelnikowi średnia arytmetyczna. W odniesieniu do kolejnej wartości $r = 2$ moment zwykły drugiego rzędu ma postać $m_2 = \frac{1}{n}\sum_{i=1}^{k} x_i^2 n_i$ (ten moment, jak i następne nie mają swych nazw). Dla $r = 3$ moment zwykły trzeciego rzędu ma postać $m_3 = \frac{1}{n}\sum_{i=1}^{k} x_i^3 n_i$ i tak dalej.

[127] Daje bowiem niepoprawny wynik lub jest nie do policzenia (w przypadku braku dominanty).

Momenty zwykłe rzędów wyższych niż jeden są wykorzystywane do obliczania tzw. momentów centralnych (o czym za chwilę) w prostszy sposób niż z definicji.

Jeżeli znane są momenty zwykłe rzędu drugiego i pierwszego, to mogą one służyć do wyliczenia wariancji, gdyż wariancja jest różnicą momentu zwykłego rzędu drugiego i kwadratu momentu zwykłego rzędu pierwszego (średniej):

$$s_x^2 = m_2 - m_1^2. \tag{6.7}$$

Momentem centralnym rzędu *r*-tego nazywamy średnią arytmetyczną z odchyleń wartości cechy od jakiejś stałej *c* podniesionych do potęgi *r*:

$$\mu_r = \frac{1}{n}\sum_{i=1}^{n}\left(x_i - c\right)^r. \tag{6.8}$$

Stała *c* jest wybranym punktem odniesienia (punktem centralnym) analizy rozkładu. Najczęściej jest to zwykła średnia arytmetyczna, ale może to być dowolne wybrana stała[128]. We wzorach podanych poniżej za tę stałą przyjmiemy więc najpopularniejszy punkt centralny, czyli średnią arytmetyczną.

Podobnie jak w przypadku momentów zwykłych, zapis momentu centralnego zależy od typu cechy i szeregu. Dla prostego szeregu cechy skokowej wzór wygląda następująco:

$$\mu_r = \frac{1}{n}\sum_{i=1}^{n}\left(x_i - \bar{x}\right)^r.$$

Dla punktowego szeregu rozdzielczego

$$\mu_r = \frac{1}{n}\sum_{i=1}^{k}\left(x_i - \bar{x}\right)^r n_i \quad \text{lub} \quad \mu_r = \sum_{i=1}^{k}\left(x_i - \bar{x}\right)^r f_i. \tag{6.9}$$

W przypadku szeregów przedziałowych stosowany jest jeden z poniższych wzorów:

$$\mu_r = \frac{1}{n}\sum_{i=1}^{k}\left(\overset{o}{x}_i - \bar{x}\right)^r n_i \quad \text{lub} \quad \mu_r = \sum_{i=1}^{k}\left(\overset{o}{x}_i - \bar{x}\right)^r f_i. \tag{6.10}$$

Wobec tego pierwszy moment centralny dla $r = 1$ zawsze wynosi $\mu_1 = \frac{1}{n}\sum_{i=1}^{n}\left(x_i - \bar{x}\right) = 0,$ co wynika z omawianych w rozdziale 4 własności średniej

[128] Np. dominanta; często w przypadku silnej asymetrii rozkładu rekomendowana jest mediana.

arytmetycznej. Drugi moment centralny ($r = 2$) jest, jak łatwo zauważyć, znaną już czytelnikowi wariancją: $\mu_2 = D^2(x) = \frac{1}{n}\sum_{i=1}^{k}\left(x_i - \bar{x}\right)^2 n_i$.

W rozdziale piątym, przy omawianiu własności wariancji, było powiedziane, że miarę tę można rozłożyć na różnicę dwóch średnich. Teraz możemy wyrazić się precyzyjniej, mówiąc, iż można ją rozłożyć na różnicę dwóch momentów zwykłych[129]: $\mu_2 = s^2(x) = D^2(X) = \overline{x^2} - \left(\bar{x}\right)^2$. Jest to, jak widać, różnica pomiędzy drugim momentem zwykłym a pierwszym podniesionym do kwadratu, aczkolwiek do zapamiętania proponujemy formułę: *wariancję liczymy jako różnicę pomiędzy średnią z kwadratów cechy a kwadratem średniej arytmetycznej rozkładu.*

Jak wspomniano powyżej, pierwszy moment centralny jest zawsze równy zero, drugi jest zawsze nieujemny (gdyż jest sumą kwadratów), więc do badania asymetrii rozkładu możemy użyć dopiero momentu trzeciego. Przyjmowanie przez dany parametr wartości zarówno dodatnich, jak i ujemnych jest nam niezbędne, jeśli dany moment ma określać rodzaj asymetrii. Przypomnijmy bowiem, że z asymetrią mamy do czynienia wówczas, gdy jednostki w rozkładzie skupiają się bliżej dolnej lub górnej granicy tegoż rozkładu. Stąd odchylenia od średniej w rozkładzie asymetrycznym nie będą takie same powyżej i poniżej tej średniej. Pierwszy moment centralny, jak wiemy, zeruje się zawsze, stąd aby wychwycić, po której stronie średniej odchylenia jednego znaku przeważają nad odchyleniami drugiego znaku, musimy do analizy wziąć trzeci moment centralny. Zainteresowany czytelnik znajdzie w podręczniku Szulca (1967, s. 234-236) analizę pokazującą, że znak trzeciego momentu centralnego zależy od rodzaju asymetrii rozkładu. Szulc pokazuje też, że wartość bezwzględna trzeciego momentu centralnego rośnie wraz ze wzrostem asymetrii rozkładu. *Reasumując, wartość trzeciego momentu centralnego mówi nam o sile asymetrii, zaś jego znak o kierunku.*

Moment centralny rzędu trzeciego jako miara bezwzględna (w jednostkach cechy podniesionych do trzeciej potęgi) nie jest dobry do porównań rozkładów, zwłaszcza że jego wartość zależy od stopnia rozproszenia rozkładu. Wobec tego do analizy używamy miary względnej, czyli *współczynnika asymetrii* postaci:

$$A = \frac{\mu_3}{s_x^3}. \tag{6.11}$$

[129] Nie będziemy bliżej tego omawiać, ale wszystkie momenty centralne można przedstawić jako sumę momentów zwykłych, co jest wykorzystywane w obliczeniach, jako że momenty zwykłe są prostsze obliczeniowo. Własność ta była istotna zwłaszcza wówczas, gdy obliczeń dokonywano ręcznie, a nie za pomocą programów obliczeniowych.

Dzieląc trzeci moment centralny przez odchylenie standardowe w trzeciej potędze likwidujemy miana cechy oraz, co ważniejsze, eliminujemy wpływ stopnia rozproszenia cechy na wielkość[130] momentu centralnego μ_3. Otrzymaliśmy w ten sposób **klasyczny współczynnik asymetrii**.

W zależności od postaci szeregu moment centralny rzędu trzeciego wyznaczany jest według jednego z poniższych wzorów:

$$\mu_3 = \frac{1}{n}\sum_{i=1}^{n}\left(x_i - \bar{x}\right)^3, \tag{6.12}$$

$$\mu_3 = \frac{1}{n}\sum_{i=1}^{k}\left(x_i - \bar{x}\right)^3 n_i \quad \text{lub} \quad \mu_3 = \sum_{i=1}^{k}\left(x_i - \bar{x}\right)^3 f_i, \tag{6.13}$$

$$\mu_3 = \frac{1}{n}\sum_{i=1}^{k}\left(\overset{o}{x}_i - \bar{x}\right)^3 n_i \quad \text{lub} \quad \mu_3 = \sum_{i=1}^{k}\left(\overset{o}{x}_i - \bar{x}\right)^3 f_i. \tag{6.14}$$

Jak już powiedziano, momenty centralne można wyznaczyć za pomocą momentów zwykłych. Moment centralny rzędu trzeciego w zależności od momentów zwykłych wyznaczamy według poniższej formuły[131]:

$$\mu_3 = m_3 - 3m_1m_2 + 2m_1^3. \tag{6.15}$$

Niestety, klasyczny współczynnik asymetrii ma dolną, lecz podobnie jak mieszany nie ma górnej granicy wartości, czyli zawiera się pomiędzy $0 \le |\mu_3| < \infty$. W praktyce[132]najczęściej przyjmuje on wartości z przedziału [-2; 2]. Podobnie jak w przypadku mieszanego współczynnika, zero oznacza symetrię rozkładu, zaś wartości różne od zera jej brak. Przy czym wartość np. ±1 oznacza tę samą siłę asymetrii, ale różniące się kierunki. Wartości przekraczające 2 lub mniejsze od –2 informują o bardzo silnej asymetrii badanego rozkładu.

Ostatnią z miar asymetrii, jaką tu omówimy, będzie **pozycyjny miernik asymetrii bazujący na kwartylach**. Idea miary jest następująca: pomiędzy każdym kwartylem znajduje się 25% jednostek populacji, w przypadku symetrii rozkładu te jednostki rozłożone są na równych odcinkach obszaru zmienności cechy. Jeżeli

---

[130] Zauważmy, że im cecha jest bardziej rozproszona (ma większy zakres zmienności), tym wartość parametru μ_3 jest większa. Porównując więc dwa rozkłady o tej samej asymetrii, ale różnej dyspersji, możemy wyciągnąć mylny wniosek, że różnią się one siłą asymetrii.

[131] Czytelnik może łatwo przekonać się o tym, wychodząc od definicji momentu i podnosząc do sześcianu zawartość nawiasu. Po drobnych przekształceniach otrzymamy sumę momentów zwykłych. Przekształcenia te można znaleźć w podręczniku Szulca [1967, s. 237-238].

[132] Patrz szerzej: Szulc [1967, s. 242].

odcinki nie są równe i na przykład kwartyl pierwszy jest bardziej oddalony od mediany niż kwartyl trzeci, oznacza to, że jednostki te, zajmując dłuższy odcinek obszaru zmienności, występują rzadziej pomiędzy kwartylem pierwszym i drugim. Występuje więc tu po lewej stronie mediany „ogon", jednostki statystyczne są tu rzadziej zagęszczone niż w ćwiartce trzeciej (czyli pomiędzy medianą a kwartylem trzecim). Stąd pozycyjny wskaźnik asymetrii przyjmuje postać:

$$W_Q = (Q_3 - Me) - (Me - Q_1). \tag{6.16}$$

Oznacza to istnienie asymetrii w rozkładzie, ale uwaga: jest to asymetria w środkowej części zbiorowości. Miara ta „nie widzi", co się dzieje poniżej kwartyla pierwszego i powyżej trzeciego. Ponieważ, podobnie jak przy poprzednich miarach, wskaźnik pozycyjny zależy od dyspersji cechy oraz jako miara bezwzględna jest wyznaczany w jednostkach cechy, nie nadaje się on do porównań i interpretacji w kategoriach siły. Dlatego też przekształca się go do postaci względnej dzieląc przez podwojone odchylenie ćwiartkowe. Otrzymujemy w ten sposób **pozycyjny współczynnik asymetrii**:

$$A_Q = \frac{W_Q}{2Q} = \frac{(Q_3 - Me) - (Me - Q_1)}{2\frac{(Q_3 - Me) + (Me - Q_1)}{2}} = \frac{Q_3 - 2Me + Q_1}{(Q_3 - Me) + (Me - Q_1)}. \tag{6.17}$$

Należy podkreślić, że pozycyjny współczynnik asymetrii można stosować, gdy szereg jest skrajnie asymetryczny. Można powiedzieć, że jest to miara będąca ostatnią deską ratunku, gdy nie jesteśmy w stanie policzyć średniej i/lub dominanty. *Miara ta ma ograniczony zakres zmienności* i przyjmuje wartości z zakresu od zera do ± 1 $\left(0 \le \left|A_Q\right| \le 1\right)$. Oznacza to, że można precyzyjnie określić nie tylko kierunek, ale i siłę asymetrii. Analogicznie do poprzednich miar otrzymana **wartość zero oznacza brak asymetrii w środkowej części rozkładu**. Skrajną asymetrię w środkowej części rozkładu wyznacza jej wartość równa ± 1, ujemna wartość miary oznacza asymetrię ujemną, czyli lewostronną, dodatnia zaś asymetrię prawostronną. *Należy jeszcze raz podkreślić, że pozycyjny miernik asymetrii wskazuje na istnienie bądź brak asymetrii jedynie w środkowej części zbiorowości* (pomiędzy kwartylem trzecim i pierwszym).

Przykład 6.1 (kontynuacja 3.3)

Odwołując się do przedstawionego poprzednio przykładu dotyczącego wagi studentek kierunku zarządzanie, określmy stopień jego asymetrii. Wiemy już, że średnia waga studentek wyniosła 54,84 kg (przykład 4.3), zaś odchylenie stan-

dardowe 5,01 kg (przykład 5.8). Wyliczenia związane z momentem centralnym rzędu trzeciego przedstawiono w tabeli 6.1.

Tabela 6.1. Rozkład wagi studentek – obliczenia związane z momentem centralnym rzędu trzeciego

$\left(x_{i_0} - x_{i_1}\right)$	n_i	$\overset{o}{x}_i$	$\overset{o}{x}_i - \bar{x}$	$\left(\overset{o}{x}_i - \bar{x}\right)^3 n_i$
42-46	2	44	–10,84	–2 547,52
46-50	16	48	–6,84	–5 120,22
50-54	26	52	–2,84	–595,56
54-58	30	56	1,16	46,83
58-62	19	60	5,16	2 610,37
62-66	5	64	9,16	4 558,24
66-70	2	68	13,16	4 558,24
Suma	100			2 795,02

Źródło: opracowanie własne.

Ponieważ otrzymana w ostatniej kolumnie suma jest dodatnia, oznacza to, że asymetria rozkładu jest dodatnia, czyli prawostronna. Świadczy o tym dodatni znak momentu centralnego rzędu trzeciego:

$$\mu_3 = \frac{1}{n}\sum_{i=1}^{k}\left(\overset{o}{x}_i - \bar{x}\right)^3 n_i = \frac{1}{100}\cdot 2\,795{,}02 = 27{,}95\,(\text{kg})^3.$$

Aby określić, jaka jest siła asymetrii, należy otrzymany wynik podzielić przez odchylenie standardowe w trzeciej potędze:

$$A = \frac{\mu_3}{s_x^3} = \frac{27{,}95}{(5{,}01)^3} = 0{,}222$$

lub dla wygody interpretacyjnej $A = \frac{\mu_3}{s_x^3}100\% = 22{,}2\%.$

Wartość 0,222 świadczy o słabej asymetrii rozkładu wagi studentek kierunku zarządzanie. Oznacza to, że większość studentek ma wagę niższą, niż wskazuje średnia, jednak różnica ta nie jest zbyt duża, gdyż siła asymetrii jest niewielka.

Korzystając z wyznaczonej poprzednio dominanty wagi studentek wynoszącej 55,07 kg (przykład 4.14), wyznaczamy mieszany współczynnik asymetrii:

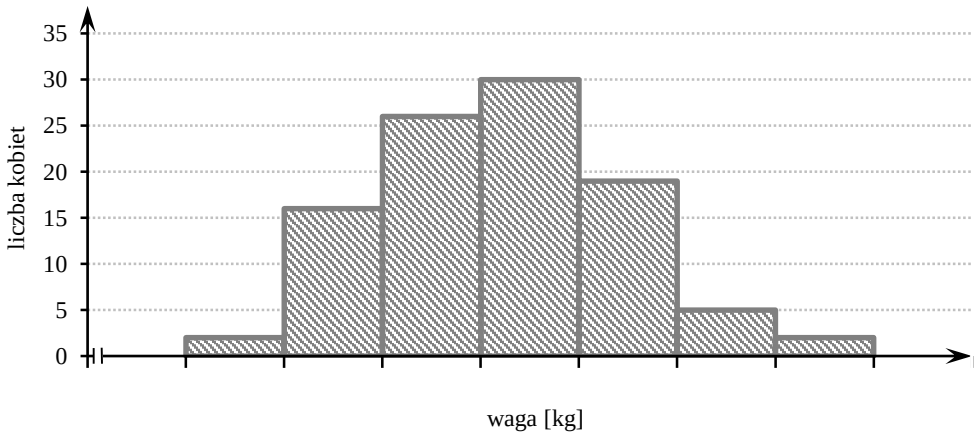
$$A_s = \frac{\bar{x} - D}{s_x} = \frac{54{,}84 - 55{,}067}{5{,}01} = \frac{-0{,}227}{5{,}01} = -0{,}045.$$

Otrzymany w tym wypadku wynik świadczy o nieznacznej asymetrii lewostronnej rozkładu wagi analizowanych studentek. W sumie można uznać rozkład za symetryczny, o wyniku różnym od zera mogły bowiem zadecydować błędy zaokrągleń.

Pozostaje jeszcze sprawdzić, jaka jest asymetria w środkowej części rozkładu. W tym celu, korzystając z wyliczonych uprzednio kwartyli $Q_1 = 51{,}08$ kg, $Me = 54{,}80$ kg, $Q_3 = 58{,}21$ kg (przykład 4.21) i odchylenia ćwiartkowego $Q = 3{,}57$ kg (przykład 5.16), wyznaczamy pozycyjny współczynnik asymetrii:

$$A_Q = \frac{Q_3 - 2Me + Q_1}{2Q} = \frac{58{,}21 - 2 \cdot 54{,}80 + 51{,}08}{2 \cdot 3{,}57} = -0{,}043.$$

Otrzymana wartość współczynnika wskazuje, że w środkowej części rozkładu, pomiędzy pierwszym a trzecim kwartylem, wagę studentek cechuje znikoma asymetria lewostronna.



Wykres 6.2. Histogram zwykły wagi studentek

Źródło: opracowanie własne na podstawie danych umownych.

Podsumowując otrzymane wyniki, warto zwrócić uwagę, że klasyczny i mieszany współczynnik asymetrii różnią się od siebie. Jednak obie miary różnią się na tyle mało od zera, że można je zinterpretować w kategoriach braku asymetrii, przy czym zauważmy także, że różnice między nimi mogą częściowo wynikać z błędów zaokrągleń oraz faktu, że szereg jest przedziałowy. Potwierdza to również ponownie zamieszczony powyżej wykres rozkładu wagi studentek (rysunek 6.2).

■

Przykład 6.2

Tabela 6.2 przedstawia liczbę pracujących wyłącznie lub głównie w gospodarstwach rolnych w czerwcu 2010 roku.

Tabela 6.2. Pracujący wyłącznie lub głównie w gospodarstwach rolnych w czerwcu 2010 roku

Gospodarstwa o powierzchni użytków rolnych (w ha)	Liczba pracujących (w tys. osób)
do 1,00	312,1
1,01-1,99	268,7
2,00-4,99	551,7
5,00-9,99	498,6
10,00-14,99	255,4
15,00-19,99	132,5
20,00-49,99	194,2
50,00 i więcej	92,3

Źródło: [Dokument elektroniczny] *Rocznik Statystyczny Rolnictwa 2011*, [źródło dostępu] http://www.stat.gov.pl/cps/rde/xbcr/gus/rs_rocznik_rolnictwa_2011.pdf, s. 114.

Na podstawie otrzymanych danych należy wyznaczyć asymetrię tego rozkładu.

Analizując tabelę 6.2, można zauważyć, że szereg ma oba krańce wartości cechy (zarówno dolny, jak i górny) otwarte. Ponieważ pierwszy zakres obejmuje gospodarstwa rolne o powierzchni do 1 ha, więc można go domknąć od dołu wartością 0 ha (gospodarstwo musi posiadać powierzchnię, aby nazywać je gospodarstwem). Dla celów dydaktycznych[133] domkniemy górny przedział wielkością 100 ha (możemy tak postąpić, gdyż liczebność tej klasy nie przekracza 10% liczebności pracujących).

W tabeli 6.3 przedstawiono wyliczenia pomocnicze do konstrukcji klasycznego współczynnika asymetrii.

[133] W realnym badaniu domknięcie to nie byłoby takie oczywiste (i raczej niepoprawne) bez dodatkowej informacji o wielkości największego gospodarstwa.

Tabela 6.3. Pracujący wyłącznie lub głównie w gospodarstwach rolnych w czerwcu 2010 roku – wyliczenia pomocnicze

$(x_{i_0} - x_{i_1})$	f_i	$\overset{o}{x}_i$	$\overset{o}{x}_i f_i$	$\overset{o}{x}_i - \bar{x}$	$\left(\overset{o}{x}_i - \bar{x}\right)^2 f_i$	$\left(\overset{o}{x}_i - \bar{x}\right)^3 f_i$
0,00-1,00	0,135	0,5	0,068	-10,544	15,009	-158,252
1,01-1,99	0,117	1,5	0,176	-9,544	10,657	-101,713
2,00-4,99	0,239	3,5	0,837	-7,544	13,602	-102,613
5,00-9,99	0,216	7,5	1,620	-3,544	2,713	-9,615
10,00-14,99	0,111	12,5	1,388	1,456	0,235	0,343
15,00-19,99	0,058	17,5	1,015	6,456	2,417	15,607
20,00-49,99	0,084	35,0	2,940	23,956	48,207	1 154,841
50,00-100,00	0,040	75,0	3,000	63,956	163,615	10 464,148
Suma	1,000		11,044		256,455	11 262,746

Źródło: opracowanie własne na podstawie [Dokument elektroniczny] *Rocznik Statystyczny Rolnictwa 2011*, [źródło dostępu] http://www.stat.gov.pl/cps/rde/xbcr/gus/rs_rocznik_rolnictwa_2011.pdf, s. 114.

Wyliczenia związane z wyznaczeniem asymetrii rozkładu rozpoczęto od obliczenia niezbędnych miar klasycznych. W drugiej kolumnie wyznaczono częstości, w trzeciej środki przedziałów, następnie zaś iloczyny częstości i środków przedziałów. Otrzymana w ten sposób czwarta kolumna posłużyła do obliczenia średniej arytmetycznej rozkładu, która wyniosła $\bar{x} = 1{,}044$ ha. Wykorzystując tę informację w kolejnych kolumnach, wyznaczono różnicę pomiędzy wartością cechy reprezentowaną przez środek przedziału a średnią arytmetyczną. Różnica ta podniesiona odpowiednio do potęgi drugiej i trzeciej oraz po pomnożeniu przez odpowiadające częstości dała momenty centralne rzędu drugiego i trzeciego. Zatem

$$\mu_2 = s_x^2 = \sum_{i=1}^{n} (x_i - \bar{x})^2 f_i = 256{,}455 \text{ ha}^2, \quad \text{zaś} \quad \mu_3 = \sum_{i=1}^{n} (x_i - \bar{x})^2 f_i = 11262{,}746 \text{ ha}^3.$$

Moment centralny rzędu drugiego jest wariancją, więc odchylenie standardowe wyniosło $s_x = \sqrt{\mu_2} = \sqrt{256{,}455} = 16{,}014$ ha. Oznacza to, że przeciętnie wielkość gospodarstw, w których zatrudniani są pracujący tam wyłącznie lub głównie, różni się od średniej powierzchni o ±16,014 ha. Odchylenie standardowe jest większe od średniej, zatem zmienność wielkości gospodarstw, w których zatrudniani są pracujący, jest bardzo wysoka, co pokazuje także współczynnik zmienności:

$$V = \frac{s_x}{\bar{x}} 100\% = \frac{16{,}014}{11{,}044} 100\% = 145\%.$$

Korzystając z otrzymanych w tabeli 6.3 wyliczeń, obliczamy klasyczny współczynnik:

$$A = \frac{\mu_3}{s_x^3} = \frac{11262,746}{(16,014)^3} = 2,742.$$

Asymetria rozkładu powierzchni gospodarstw, w których pracują zatrudnieni tam głównie lub wyłącznie, jest bardzo silna i prawostronna (dodatnia). Oznacza to, że dominuje zatrudnienie w gospodarstwach o wielkości poniżej średniej arytmetycznej.

Warto w tym miejscu rozważyć sytuację, gdy posiadamy informacje dotyczące największego gospodarstwa w Polsce w 2010 roku. Wiedząc, że wartość maksymalna wyniosła 6950 ha, czyli przyjmując jako górny kraniec badanego szeregu 7 000 ha, otrzymuje się wartość współczynnika asymetrii równą 4,693. Wynik ten oznacza jeszcze silniejszą (skrajną) asymetrię prawostronną.

Kontynuując analizę przykładu w tabeli 6.4, przedstawiamy wyliczenia pomocnicze niezbędne do wyznaczenia współczynnika asymetrii mieszanego i pozycyjnego.

Tabela 6.4. Pracujący wyłącznie lub głównie w gospodarstwach rolnych w czerwcu 2010 roku – wyliczenia pomocnicze

$(x_{i_0} - x_{i_1})$	f_i	h_i	k_i	f_{isk}
0,00-1,00	0,135	1	0,135	0,135
1,01-1,99	0,117	1	0,117	0,252
2,00-4,99	0,239	3	0,080	0,491
5,00-9,99	0,216	5	0,043	0,707
10,00-14,99	0,111	5	0,022	0,818
15,00-19,99	0,058	5	0,012	0,876
20,00-49,99	0,084	30	0,003	0,960
50,00-100,00	0,040	50	0,001	1,000
Suma	1,000			

Źródło: opracowanie własne na podstawie
[Dokument elektroniczny] *Rocznik Statystyczny Rolnictwa 2011*,
[źródło dostępu] http://www.stat.gov.pl/cps/rde/xbcr/gus/rs_rocznik_rolnictwa_2011.pdf, s. 114.

Podczas obliczenia mieszanego współczynnika asymetrii musimy wyznaczyć dominantę. Jednak, jak łatwo zauważyć, nie możemy użyć standardowego wzoru interpolacyjnego, ponieważ szereg ma przedziały o nierównej rozpiętości. Musimy

więc przekształcić liczebności rozkładu do natężenia liczebności, by móc skorzystać ze wzoru 4.19. W tabeli 6.4 wyznaczono w czwartej kolumnie gęstości k_i. W trakcie obliczeń rozpiętości przedziałów zostały one zaokrąglone do liczb całkowitych (różnice sięgały jednej setnej). Analizując wyliczone gęstości w poszczególnych przedziałach, stwierdzamy, że rozkład jest monotoniczny, czyli skrajnie asymetryczny. Najwyższa gęstość zatrudnienia znajduje się w pierwszym przedziale, co oznacza brak dominanty. W efekcie stwierdzamy, że wyznaczenie mieszanego współczynnika asymetrii jest niemożliwe.

Ostatnia kolumna w tabeli 6.4 prezentuje częstości skumulowane. Wynika z niej, że kwartyl dolny znajduje się w drugim przedziale, mediana w czwartym, zaś kwartyl górny w piątym. Ich wartości wynoszą odpowiednio:

$$Q_1 = x_{0Q_1} + \frac{h_{Q_1}}{f_{Q_1}} \cdot \left(0{,}25 - \sum_{i=1}^{k_1-1} f_i \right) = 1 + \frac{1}{0{,}117} \cdot (0{,}25 - 0{,}135) = 1{,}983 \text{ ha};$$

$$Me = x_{0Me} + \frac{h_{Me}}{f_{Me}} \cdot \left(0{,}5 - \sum_{i=1}^{k_2-1} f_i \right) = 5 + \frac{5}{0{,}216} \cdot (0{,}5 - 0{,}491) = 5{,}208 \text{ ha};$$

$$Q_3 = x_{0Q_3} + \frac{h_{Q_3}}{f_{Q_3}} \cdot \left(0{,}75 - \sum_{i=1}^{k_3-1} f_i \right) = 10 + \frac{5}{0{,}111} \cdot (0{,}75 - 0{,}707) = 11{,}937 \text{ ha}.$$

Pomocniczo wyznaczono również odchylenie ćwiartkowe:

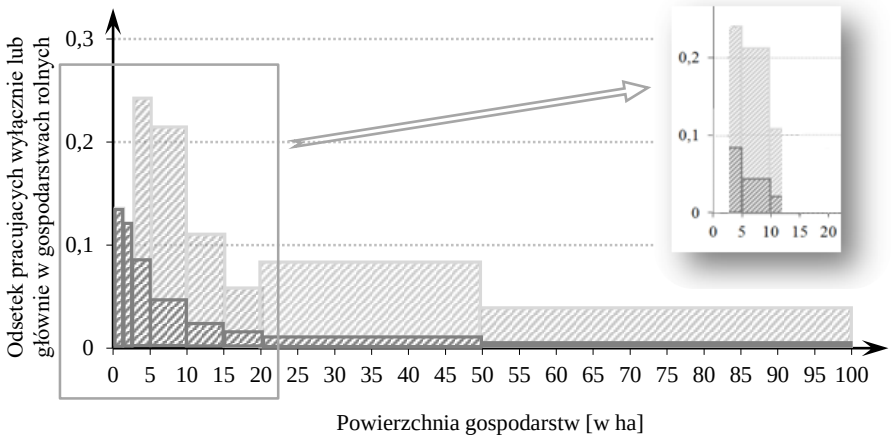
$$Q = \frac{Q_3 - Q_1}{2} = \frac{11{,}937 - 1{,}983}{2} = 4{,}977 \text{ ha}.$$

Ostatecznie współczynnik pozycyjny asymetrii wynosi:

$$A_Q = \frac{(Q_3 - Me) - (Me - Q_1)}{2Q} = \frac{Q_3 - 2Me + Q_1}{Q_3 - Q_1} = \frac{11{,}937\text{-}2 \cdot 5{,}208 + 1{,}983}{2 \cdot 4{,}977} = 0{,}352.$$

Wynik dodatni oznacza, że w zawężonym do dwóch środkowych ćwiartek obszarze powierzchni gospodarstw rozkład liczby tam zatrudnionych posiada umiarkowaną asymetrię prawostronną.

W przedstawionym przykładzie wartości obu współczynników wyraźnie różnią się od siebie. Wynika to z faktu, iż jak wspomniano powyżej, opisują one asymetrię w różnych zakresach. Pierwszy obejmuje całą populację, drugi tylko jej środkową część. Dobrą ilustracją będzie w tym momencie przyjrzenie się histogramowi zwykłemu (jasnoszary) i histogramowi gęstości (ciemnoszary) wykonanymi dla badanego rozkładu, które został przedstawione na wykresie 6.3.



Wykres 6.3. Histogram pracujących wyłącznie lub głównie w gospodarstwach rolnych w czerwcu 2010 roku według powierzchni gospodarstwa

Źródło: opracowanie własne na podstawie: [Dokument elektroniczny] *Rocznik Statystyczny Rolnictwa 2011*, [źródło dostępu] http://www.stat.gov.pl/cps/rde/xbcr/gus/rs_rocznik_rolnictwa_2011.pdf, s. 114.

Analizując ten wykres, dostrzegamy, że w obu przypadkach[134] wykres rozkładu gęstości liczby pracujących wyłącznie lub głównie w gospodarstwach rolnych ze względu na ich powierzchnię jest skrajnie asymetryczny. Ograniczając ogląd do obszaru pomiędzy pierwszym a trzecim kwartylem, widzimy, że rzeczywiście asymetria tego obszaru jest niższa.

■

Przykład 6.3

Tabela 6.5 przedstawia rozkład wieku posłów VI kadencji do Sejmu Rzeczypospolitej Polskiej. Korzystając z tych danych, wyznaczmy asymetrię rozkładu tego wieku.

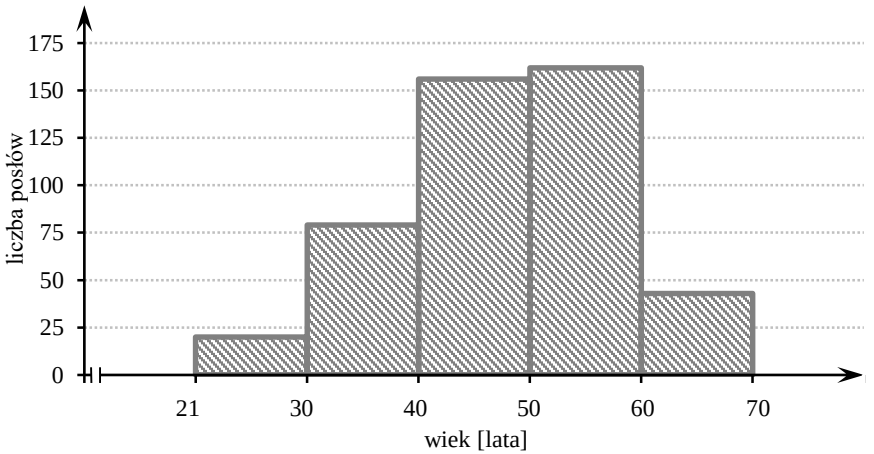
Tabela 6.5. Posłowie VI kadencji Sejmu Rzeczypospolitej Polskiej według wieku (stan na początek kadencji)

Wiek (w latach)	Liczba posłów
21-29	20
30-39	79
40-49	156
50-59	162
60 lat i więcej	43

Źródło: [Dokument elektroniczny] *Rocznik Statystyczny Rzeczypospolitej Polskiej 2012*, [źródło dostępu] http://www.stat.gov.pl/cps/rde/xbcr/gus/RS_rocznik_statystyczny_rp_2012.pdf, s. 120.

[134] Zwykły histogram i oparty na gęstości.

Zanim zostaną wyznaczone poszczególne współczynniki asymetrii, warto przyjrzeć się histogramowi wieku posłów, który przedstawiono na wykresie 6.4. Ponieważ górny przedział był niedomknięty, ograniczono go dla celów dydaktycznych wartością 69 lat (w przedziale znajduje się mniej niż 10% liczebności posłów).



Wykres 6.4. Histogram wieku posłów VI kadencji Sejmu Rzeczypospolitej Polskiej (stan na początek kadencji)

Źródło: opracowanie własne na podstawie:
[Dokument elektroniczny] *Rocznik Statystyczny Rzeczypospolitej Polskiej 2012*,
[źródło dostępu] http://www.stat.gov.pl/cps/rde/xbcr/gus/RS_rocznik_statystyczny_rp_2012.pdf, s. 120.

Analizując wykres 6.4, łatwo można zauważyć, że rozkład wieku posłów VI kadencji cechuje niezbyt silna asymetria lewostronna. Aby dokładnie określić jej wartość, należy wyznaczyć klasyczny współczynnik asymetrii. Wyliczenia pomocnicze z tym związane przedstawiono w tabeli 6.6.

Tabela 6.6. Posłowie VI kadencji Sejmu Rzeczypospolitej Polskiej według wieku (stan na początek kadencji) – wyliczenia pomocnicze

$\lfloor x_{i_0} - x_{i_1} \rfloor$	f_i	$\overset{o}{x}_i$	$\overset{o}{x}_i f_i$	$\overset{o}{x}_i - \bar{x}$	$\left(\overset{o}{x}_i - \bar{x}\right)^2 f_i$	$\left(\overset{o}{x}_i - \bar{x}\right)^3 f_i$	f_{isk}
21-29	0,044	25,5	1,122	-22,302	21,885	-488,072	0,044
30-39	0,172	35,0	6,020	-12,802	28,189	-360,879	0,216
40-49	0,339	45,0	15,255	-2,802	2,662	-7,458	0,555
50-59	0,352	55,0	19,360	7,198	18,238	131,274	0,907
60-69	0,093	65,0	6,045	17,198	27,507	473,061	1,000
Suma	1,000		47,802		98,481	-252,074	

Źródło: opracowanie własne na podstawie:

[Dokument elektroniczny] *Rocznik Statystyczny Rzeczypospolitej Polskiej 2012*,
[źródło dostępu] http://www.stat.gov.pl/cps/rde/xbcr/gus/RS_rocznik_statystyczny_rp_2012.pdf, s. 120.

Wyliczenia pomocnicze rozpoczęto od obliczenia, podobnie jak we wcześniejszym przykładzie, częstości (dla wygody obliczeń), a następnie środków przedziałów i średniej arytmetycznej. Wyniosła ona $\bar{x} = 47{,}8$ lat, co oznacza, że średnio posłowie na IV kadencję Sejmu mieli prawie 48 lat. W kolejnych kolumnach policzone zostały momenty centralne rzędu drugiego i trzeciego. Wyniosły one odpowiednio: $\mu_2 = 98{,}48$ lat 2 i $\mu_3 = -252{,}07$ lat^3. Pierwiastek z wariancji (momentu centralnego rzędu drugiego) stanowi odchylenie standardowe, które wyniosło: $s_x = \sqrt{\mu_2} = \sqrt{98{,}48} = 9{,}92 \approx 10$ lat. Otrzymana wartość oznacza, że przeciętnie wiek posłów różnił się od średniej ich wieku o 10 lat, co oznacza, że zróżnicowanie posłów ze względu na wiek jest dosyć słabe $\left(V = \frac{s_x}{\bar{x}} 100\% = \frac{9{,}92}{47{,}80} 100\% = 21\% \right)$.

Powracając do asymetrii, klasyczny współczynnik asymetrii wyznaczono następująco:

$$A = \frac{\mu_3}{s_x^3} = \frac{-252{,}07}{(9{,}92)^3} = -0{,}258.$$

Powyższy wynik oznacza, że rozkład wieku posłów VI kadencji jest lewostronnie asymetryczny, zaś siła asymetrii jest niewielka. Dominują posłowie powyżej średniego wieku, aczkolwiek różnica nie jest znaczna. Przedstawione wyliczenia potwierdziły wnioski z analizy wykresu 6.4.

Następnie wyliczamy dominantę, która znajduje się w czwartym przedziale (możemy przeprowadzić te wyliczenia, nie przekształcając liczebności w częstość, gdyż jedynie przedział pierwszy ma inną, chociaż niewiele odbiegającą od pozostałych rozpiętość, i nie sąsiaduje on z przedziałem branym pod uwagę podczas wyznaczania dominanty):

$$D = x_{0D} + \frac{f_D - f_{D-1}}{(f_D - f_{D-1}) + (f_D - f_{D+1})} \cdot h_D = 50 + \frac{0{,}352 - 0{,}339}{(0{,}352 - 0{,}339) + (0{,}352 - 0{,}093)} \cdot 10 = 50{,}48$$

lat.

Zatem dominujący wiek analizowanych posłów skupiał się wokół wartości 50,5 lat. Łącząc tę informację z klasyczną średnią i odchyleniem standardowym, otrzymujemy mieszany współczynnik asymetrii:

$$A_s = \frac{\bar{x} - D}{s_x} = \frac{47.80 - 50{,}48}{9{,}92} = -0{,}270.$$

Otrzymana wartość jest zbliżona do wyznaczonej poprzednio miary klasycznej, oznaczając słabą lewostronną asymetrię wieku posłów. Zanim wyznaczymy wartość pozycyjnego współczynnika asymetrii, policzymy poszczególne kwartyle

(dolny znajduje się w trzecim przedziale, podobnie jak mediana, a kwartyl górny w czwartym):

$$Q_1 = x_{0Q_1} + \frac{h_{Q_1}}{f_{Q_1}} \cdot \left(0{,}25 - \sum_{i=1}^{k_1-1} f_i \right) = 40 + \frac{10}{0{,}339} \cdot (0{,}250 - 0{,}216) = 41{,}00 \text{ lat};$$

$$Me = x_{0Me} + \frac{h_{Me}}{f_{Me}} \cdot \left(0{,}5 - \sum_{i=1}^{k_2-1} f_i \right) = 40 + \frac{10}{0{,}339} \cdot (0{,}500 - 0{,}216) = 48{,}38 \text{ lat};$$

$$Q_3 = x_{0Q_3} + \frac{h_{Q_3}}{f_{Q_3}} \cdot \left(0{,}75 - \sum_{i=1}^{k_3-1} f_i \right) = 50 + \frac{10}{0{,}352} \cdot (0{,}750 - 0{,}555) = 55{,}54 \text{ lat}.$$

Pomocniczo wyznaczone zostało również odchylenie ćwiartkowe:

$$Q = \frac{Q_3 - Q_1}{2} = \frac{55{,}54 - 41{,}00}{2} = 7{,}27 \text{ lat}.$$

Ostatecznie współczynnik pozycyjny asymetrii wyniósł:

$$A_Q = \frac{Q_3 - 2Me + Q_1}{2Q} = \frac{55{,}54 - 2 \cdot 48{,}38 + 41{,}00}{2 \cdot 7{,}27} = -0{,}015.$$

Asymetria rozkładu wieku posłów VI kadencji w środkowej części zbiorowości jest w znikomym stopniu lewostronna.

■

Przykład 6.3[135]

Zmierzmy różnymi sposobami asymetrię rozkładu wysokości połowów w pewnej spółdzielni rybackiej.

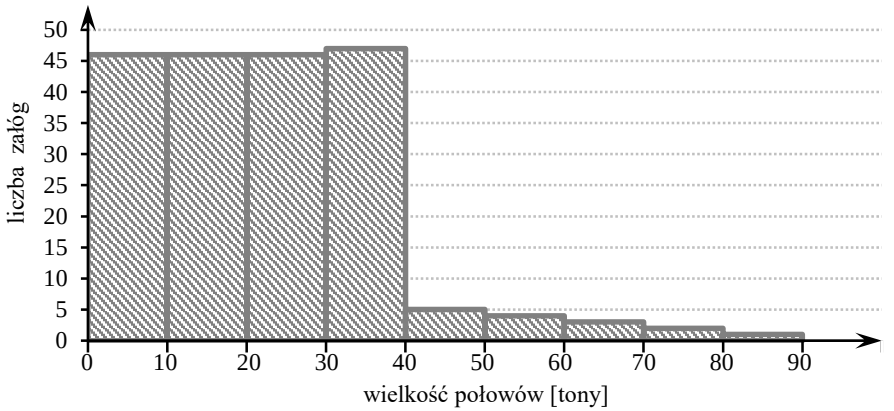
Tabela 6.7. Załogi spółdzielni rybackiej nr 3 według wysokości połowów

$(x_{0i} - x_{1i}]$	0-10	10-20	20-30	30-40	40-50	50-60	60-70	70-80	80-90	Razem
n_i	46	46	46	47	5	4	3	2	1	200

Źródło: B. Szulc, *Statystyka dla ekonomistów. Opis statystyczny*, Państwowe Wydawnictwo Ekonomiczne, Warszawa 1967, s. 247.

Badanie asymetrii należy rozpocząć od utworzenia wykresu. Został on przedstawiony na wykresie 6.5.

[135]Zadanie pochodzi z podręcznika Szulca (1967, s. 247).



Wykres 6.5. Histogram wielkości połowów spółdzielni rybackiej

Źródło: opracowanie własne na podstawie: B. Szulc, *Statystyka dla ekonomistów. Opis statystyczny*, Państwowe Wydawnictwo Ekonomiczne, Warszawa 1967, s. 247.

Zacznijmy od wyznaczenia poszczególnych miar. W tabeli 6.8 przedstawiono niezbędne wyliczenia.

Tabela 6.8. Wyliczenia dotyczące danych o załogach spółdzielni rybackiej

$(x_{i_0} - x_{i_1})$	n_i	$\overset{o}{x}_i$	$\overset{o}{x}_i n_i$	$\overset{o}{x}_i - \bar{x}$	$\left(\overset{o}{x}_i - \bar{x}\right)^2 n_i$	$\left(\overset{o}{x}_i - \bar{x}\right)^3 n_i$	n_{isk}
0 – 10	46	5	230	–17,95	14 821,32	–266 042,69	46
10 – 20	46	15	690	–7,95	2 907,32	–23 113,19	92
20 – 30	46	25	1 150	2,05	193,32	396,31	138
30 – 40	47	35	1 645	12,05	6 824,52	82 235,47	185
40 – 50	5	45	225	22,05	2 431,01	53 603,77	190
50 – 60	4	55	220	32,05	4 108,81	131 687,36	194
60 – 70	3	65	195	42,05	5 304,61	223 058,85	197
70 – 80	2	75	150	52,05	5 418,41	282 028,24	199
80 – 90	1	85	85	62,05	3 850,20	238 904,91	200
Suma	200		4 590		45 859,52	722 759,03	

Źródło: opracowanie własne na podstawie: B. Szulc, *Statystyka dla ekonomistów. Opis statystyczny*, Państwowe Wydawnictwo Ekonomiczne, Warszawa 1967, s. 247.

Średnia wyniosła: $\bar{x} = \frac{4\,590}{200} = 22{,}95$ tony, co oznacza, że średnio załogi łowiły 22,95 tony ryb. Momenty centralne rzędu drugiego i trzeciego wyniosły odpowiednio: $\mu_2 = \frac{45\,859{,}52}{200} = 229{,}30$ tony^2 i $\mu_3 = \frac{722\,759{,}03}{200} = 3\,613{,}80$ tony^3. Zatem odchylenie standardowe wyniosło: $s_x = \sqrt{3\,613{,}80} = 15{,}14$ tony. Oznacza to, że przeciętnie wielkość połowów różniła się od średniej połowu dla badanych załóg o 15,14 tony, co oznacza, że zróżnicowanie załóg ze względu na wielkość połowów jest umiarkowane $\left(V = \frac{15{,}14}{22{,}95} 100\% = 66\% \right)$. Klasyczny współczynnik asymetrii wynosi:

$$A = \frac{3\,613{,}80}{(15{,}14)^3} = 1{,}04.$$

Wynik ten oznacza silną asymetrię prawostronną, co można zaobserwować na wykresie 6.5.

Aby można było wyznaczyć mieszany współczynnik asymetrii, należy wyliczyć dominantę. Jednak patrząc na wykres 6.5, musimy dojść do wniosku, że wartość dominanty nie powinna być wyznaczana, gdyż nie ma wyraźnie dominującego przedziału. Różnica pomiędzy czwartym przedziałem a trzema poprzednimi jest znikoma i może być przypadkowa. Jednak na potrzeby niniejszego zadania wyznaczmy ją:

$$D = 30 + \frac{47 - 46}{(47 - 46) + (47 - 5)} \cdot 10 = 30{,}23 \text{ tony}.$$

Zatem mieszany współczynnik asymetrii wynosi:

$$A = \frac{22{,}95 - 30{,}23}{15{,}14} = -0{,}481.$$

Wyliczenia wskazują na umiarkowaną asymetrię lewostronną, co jest nieprawdziwe, gdy spojrzy się na wykres 6.5. Wniosek ten jest błędny, gdyż bazuje na dominancie, która w tym rozkładzie *de facto* nie występuje (błędnie przyjęto przedział, w którym się ona znajduje).

Wyliczmy teraz kwartyle, na podstawie których wyznaczymy pozycyjny współczynnik skośności. Kwartyl pierwszy znajduje się w drugim przedziale, mediana w trzecim, a kwartyl trzeci w czwartym przedziale. Poniżej przedstawiliśmy wyliczenia dotyczące wartości kwartyli:

$$Q_1 = 10 + \frac{10}{46} \cdot (50 - 46) = 10{,}87 \text{ tony};$$

$$Me = 20 + \frac{10}{46} \cdot (100 - 92) = 21{,}74 \text{ tony};$$

$$Q_3 = 30 + \frac{10}{47} \cdot (150 - 138) = 32{,}55 \text{ lat.}$$

Stąd odchylenie ćwiartkowe wynosi:

$$Q = \frac{32{,}55 - 10{,}87}{2} = 10{,}84 \text{ lat.}$$

Zatem pozycyjny współczynnik asymetrii wyniósł:

$$A_Q = \frac{32{,}55 - 2 \cdot 21{,}74 + 10{,}87}{2 \cdot 10{,}84} = -0{,}003 \approx 0.$$

Otrzymana wartość wskazuje na rozkład symetryczny, co jest widoczne na wykresie 6.5. Jeżeli odrzucimy 25% najniższych wielkości połowów i 25% najwyższych, to w tym obszarze widoczny jest szereg symetryczny.

■

Podany powyżej przykład wskazuje, że należy ostrożnie dobierać i interpretować różne miary opisujące zbiorowość. Zawsze przed ich wyznaczeniem należy się zastanowić, czy powinniśmy je liczyć. Należy również zwracać uwagę na to, jak wygląda wykres, gdyż wnioski z wykresu i obliczeń nie powinny być sprzeczne.

6.2. Miary koncentracji

Koncentracja w potocznym języku oznacza skupienie czegoś. W podobnym znaczeniu będziemy używali tego pojęcia w statystyce. Słowa tego używa się jednak dwojako na określenie dwóch różnych problemów.

Pierwsze znaczenie obejmuje stopień skupienia wartości cechy wokół średniej arytmetycznej danego rozkładu – miarą tego stopnia jest parametr zwany **kurtozą** i związany z nim parametr zwany **ekscesem**. Służą one do porównania stopnia skupienia badanego rozkładu cechy z pewnym rozkładem uznanym za wzorcowy – czyli rozkładem normalnym.

W drugim, bardziej powszechnym zastosowaniu termin „koncentracja" jest używany w statystyce na określenie nierównomierności rozłożenia ogólnej sumy wartości cechy (funduszu cechy) pomiędzy poszczególne jednostki zbiorowości –

podstawową miarą tak rozumianej koncentracji są **krzywa Lorenza** i współczynnik Lorenza.

Poniżej omówimy oba rodzaje „koncentracji", aczkolwiek od razu zaznaczmy, że kurtoza i eksces mają mniejsze znaczenie w analizach statystycznych. Używane są na ogół jako miary dodatkowe w analizie rozkładu cechy.

6.2.1. Miara skupienia względem wartości przeciętnej – kurtoza i eksces

Do określania stopnia skupienia bądź rozproszenia cechy wokół średniej służy czwarty moment centralny. Bezpośrednio, podobnie jak moment drugi i trzeci, nie nadaje się on do interpretacji i porównań. Przekształcamy go wobec tego do postaci względnej, dzieląc przez odchylenie standardowe w czwartej potędze, co umożliwia uproszczenie jednostek pomiaru oraz normalizuje ten moment ze względu na dyspersję cechy.[136] Miernik ten nosi nazwę kurtozy i zapisujemy go w postaci:

$$k = \frac{\mu_4}{s_x^4}, \tag{6.17}$$

gdzie czwarty moment centralny, w zależności od typu cechy, wyznaczany jest jako czwarta potęga różnic pomiędzy wartościami cechy lub środkami przedziałów i średnią arytmetyczną.

Przypomnijmy więc, że dla szeregu prostego czwarty moment centralny obliczamy według wzoru:

$$\mu_4 = \frac{1}{n}\sum_{i=1}^{n}\left(x_i - \bar{x}\right)^4. \tag{6.18}$$

W przypadku szeregu rozdzielczego cechy skokowej stosujemy jeden z poniższych wzorów:

$$\mu_4 = \frac{1}{n}\sum_{i=1}^{k}\left(x_i - \bar{x}\right)^4 n_i \quad \text{lub} \quad \mu_4 = \sum_{i=1}^{k}\left(x_i - \bar{x}\right)^4 f_i, \tag{6.19}$$

zaś do cechy ciągłej przedstawionej w postaci szeregu rozdzielczego przedziałowego stosujemy:

[136] Podobnie jak przy poprzednich miarach względnych, odchylenie standardowe jest tu jednostką odległości od średniej.

$$\mu_4 = \frac{1}{n}\sum_{i=1}^{k}\left(x_i^o - \bar{x}\right)^4 n_i \quad \text{lub} \quad \mu_4 = \sum_{i=1}^{k}\left(x_i^o - \bar{x}\right)^4 f_i. \tag{6.20}$$

Moment centralny rzędu czwartego może być również wyznaczony jako kombinacja momentów zwykłych rzędu od pierwszego do czwartego zgodnie z poniższym wzorem:

$$\mu_4 = m_4 - 4m_3m_1 + 6m_2m_1^2 - 3m_1^4. \tag{6.21}$$

Ponieważ kurtoza nie jest miarą unormowaną, czyli nie jest znany zakres możliwych jej wartości (ze względu na parzystą potęgę wiadomo jedynie, że jej wartości są większe od zera[137]), więc nie można określić, czy stopień koncentracji jednostek populacji wokół średniej jest duży czy mały. W tym celu porównuje się wyliczoną kurtozę z rozkładem uznanym za wzorcowy, czyli z rozkładem normalnym, dla którego ma ona wartość równą liczbie trzy. Różnica pomiędzy wartością kurtozy rozkładu badanego a rozkładu normalnego nazywa się ekscesem:

$$K = k - 3. \tag{6.22}$$

Wobec tego jeżeli kurtoza rozkładu badanego $k = 3$, to eksces jest równy zeru, co informuje nas, że badany rozkład ma taką samą koncentrację jednostek wokół średniej jak wzorcowy rozkład normalny.

Jeżeli kurtoza rozkładu badanego $k > 3$, to eksces jest dodatni: $K = k - 3 > 0$, co oznacza, że analizowany szereg jest **leptokurtyczny,** czyli smuklejszy niż rozkład normalny. Oznacza to, że wartości cechy są bardziej skupione wokół średniej niż w rozkładzie normalnym[138].

Jeżeli kurtoza rozkładu badanego $k < 3$, to eksces jest ujemny: $K = k - 3 < 0$, co oznacza, że rozkład jest **platokurtyczny,** czyli bardziej spłaszczony niż rozkład normalny. Oznacza to, że wartości cechy są mniej skupione wokół średniej niż w rozkładzie normalnym[139] .

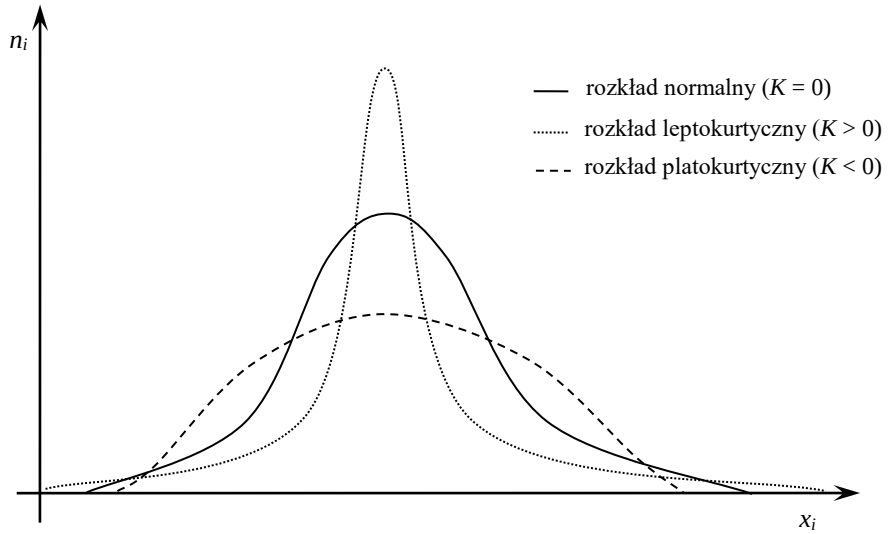
---

[137] Kurtoza równa zero oznacza trywialny przypadek rozkładu, w którym wszystkie jednostki mają tę samą wartość cechy, co oznacza zerowe odchylenia od średniej, co jest mało interesujące i nie ma tu co badać. Jako że w liczniku współczynnika koncentracji mamy odchylenia od średniej, w parzystej potędze suma odchyleń jest zawsze dodatnia.

[138] Bystry czytelnik jednak zauważy, patrząc na rysunek 6.2, że zakres zmienności cechy w rozkładzie leptokurtycznym jest większy niż w rozkładzie normalnym. Jak widać na rysunku, „ogony" rozkładu leptokurtycznego sięgają dalej niż rozkładu normalnego, ale mają jednocześnie niższą wysokość, co oznacza, że jest w nich mniej jednostek statystycznych niż w rozkładzie normalnym.

[139] Odwrotnie więc niż w rozkładzie leptokurtycznym mamy w rozkładzie platokurtycznym do czynienia z mniejszym zakresem zmienności („ogony" są krótsze, ale „grubsze" niż w rozkładzie normalnym), ale większym rozproszeniem wokół średniej.

Na rysunku 6.2 przedstawiono porównanie rozkładów: normalnego, platokurtycznego oraz leptokurtycznego.



Rysunek 6.2. Przedstawienie różnych typów rozkładów: normalnego, platokurtycznego i leptokurtycznego

Źródło: opracowanie własne.

O rozkładzie platokurtycznym mówi się, że jest on spłaszczony w porównaniu do rozkładu normalnego w środkowej części (niższa i szersza środkowa część), ale zawężony (krótszy) w „ogonach" wykresu. Rozkład leptokurtyczny jest węższy i wyższy w porównaniu do rozkładu normalnego w środkowej części i jednocześnie wydłużony na brzegach wykresu.

Niektórzy autorzy[140] zamiast pojęcia miara koncentracji sugerują użycie określenia **miary spłaszczenia** rozkładu jako lepiej oddających to, o czym mówi kurtoza i eksces.

Podsumowując, należy zaznaczyć, że kurtoza jest miarą należącą do przedziału jednostronnie domkniętego $(0 \leq k < \infty)$[141], czyli nie ma górnej granicy i dlatego bezpośrednio nie jesteśmy w stanie zinterpretować wyniku w kategoriach stopnia (siły) koncentracji wokół średniej. Do interpretacji używamy więc ekscesu, stosując jako punkt odniesienia *kurtozę rozkładu normalnego* wynoszącą trzy $(k = 3)$. W ten sposób jesteśmy tylko w stanie określić, czy badany rozkład jest

[140] Patrz: Zeliaś i in., Steczkowski, Malina.

[141] Istnieje ściślejsze ograniczenie kurtozy, a mianowicie jest ona niewiększa od liczby obserwacji.

mniej lub bardziej skupiony (bardziej lub mniej spłaszczony) wokół swojej średniej niż *rozkład normalny*.

Rozkład leptokurtyczny jest wysmuklony w stosunku do normalnego, co oznacza, że wartości cechy jednostek takiej populacji są skupione silniej wokół średniej w środkowej części rozkładu niż w rozkładzie normalnym, ale większa jest dyspersja cechy na krańcach. Zatem średnia w rozkładzie leptokurtycznym dobrze reprezentuje taką zbiorowość.

Rozkład platokurtyczny jest spłaszczony w stosunku do normalnego, czyli wartości cechy jednostek takiej populacji są mniej skupione wokół średniej rozkładu środkowej części niż w rozkładzie normalnym, ale dyspersja cechy jest mniejsza niż w rozkładzie normalnym. Zatem średnia w rozkładzie platokurtycznym słabiej reprezentuje taką zbiorowość.

Stopień koncentracji rozkładu wokół średniej bada się dla rozkładów symetrycznych i umiarkowanie asymetrycznych. W przypadku silnej asymetrii liczenie koncentracji nie ma sensu, ponieważ średnia arytmetyczna jest, w przypadku cechy ciągłej, przekłamana z powodu sztucznie przyjętych środków przedziałów jako reprezentantów cechy. W przypadku obu rodzajów cech średnia nie odzwierciedla także poprawnie rozkładu cechy w populacji. Gdy rozkład jest skrajnie asymetryczny lub równomierny, średnia arytmetyczna jest sztucznym tworem niczego nie reprezentującym.

Przykład 6.4 (kontynuacja 3.3)

Odwołując się ponownie do przykładu opisującego wagę studentek kierunku zarządzanie określmy stopień koncentracji tej wagi wokół wartości średniej. Wiadomo, że średnia waga studentek wyniosła 54,84 kg (przykład 4.3), zaś odchylenie standardowe 5,01 kg (przykład 5.8). Wyliczenia niezbędne do wyznaczenia momentu centralnego rzędu czwartego przedstawiono w tabeli 6.9.

Tabela 6.9. Rozkład wagi studentek – obliczenia związane z momentem centralnym rzędu czwartego

$(x_{i_0} - x_{i_1})$	n_i	$\overset{o}{x}_i$	$\overset{o}{x}_i - \bar{x}$	$\left(\overset{o}{x}_i - \bar{x}\right)^4 n_i$
42-46	2	44	– 10,84	27 615,13
46-50	16	48	– 6,84	35 022,28
50-54	26	52	– 2,84	1 691,40
54-58	30	56	1,16	54,32
58-62	19	60	5,16	13 469,53
62-66	5	64	9,16	35 200,75
66-70	2	68	13,16	59 986,50
Suma	100			173 039,91

Źródło: opracowanie własne.

Otrzymana w ostatniej kolumnie suma po podzieleniu przez liczebność określa wartość momentu centralnego rzędu czwartego:

$$\mu_3 = \frac{1}{n}\sum_{i=1}^{k}\left(\overset{o}{x}_i - \bar{x}\right)^4 n_i = \frac{1}{100}\cdot 173\,039{,}91 = 1\,730{,}40 \text{ kg}^4.$$

Wyznaczając kurtozę, należy otrzymany wynik zestandaryzować, czyli podzielić przez odchylenie standardowe w czwartej potędze:

$$k = \frac{\mu_4}{s_x^4} = \frac{1\,730{,}40}{(5{,}01)^4} = 2{,}747.$$

Otrzymany wynik może być jeszcze przekształcony na eksces:

$$K = k - 3 = 2{,}747 - 3 = -0{,}253.$$

Ujemna wartość ekscesu informuje nas, że badany rozkład jest platokurtyczny, czyli spłaszczony wokół wartości średniej w porównaniu do rozkładu normalnego. Oznacza to, że koncentracja wartości wagi studentek wokół średniej wagi w badanej grupie jest w niewielkim stopniu mniejsza, niż miałoby to miejsce, gdyby rozkład był zgodny z rozkładem normalnym.

■

Przykład 6.5 (kontynuacja 6.2)

Przypomnijmy przykład 6.2 dotyczący rozkładu wielkości gospodarstw, w jakich pracują zatrudnieni wyłącznie lub głównie w gospodarstwach rolnych w 2010 roku. Pamiętając, że średnia rozkładu wyniosła $\bar{x} = 1{,}044$ ha, zaś odchylenie standardowe $s_x = 16{,}014$ ha, określmy stopień koncentracji wokół średniej wielkości gospodarstwa.

Analizując wykres 6.1 tego rozkładu, widzimy, że jest to rozkład jednomodalny, chociaż nie można było dla niego wyznaczyć dominanty. Jednomodalność oznacza, że możliwe jest określenie stopnia koncentracji wokół wartości średniej. Wyliczenia należy rozpocząć od określenia momentu centralnego rzędu czwartego. Pomocnicze obliczenia przedstawiono w tabeli 6.10.

Tabela 6.10. Pracujący wyłącznie lub głównie w gospodarstwach rolnych w czerwcu 2010 roku – wyliczenia pomocnicze

$(x_{i_0} - x_{i_1})$	f_i	$\overset{o}{x}_i$	$\overset{o}{x}_i - \bar{x}$	$\left(\overset{o}{x}_i - \bar{x}\right)^4 f_i$
0,00-1,00	0,135	0,5	-10,544	1 668,612
1,01-1,99	0,117	1,5	-9,544	970,750
2,00-4,99	0,239	3,5	-7,544	774,113
5,00-9,99	0,216	7,5	-3,544	34,074
10,00-14,99	0,111	12,5	1,456	0,499
15,00-19,99	0,058	17,5	6,456	100,759
20,00-49,99	0,084	35,0	23,956	27 665,371
50,00-100,00	0,040	75,0	63,956	669 245,049
Suma	1,000			700 459,227

Źródło: opracowanie własne na podstawie: [Dokument elektroniczny] *Rocznik Statystyczny Rolnictwa 2011*, [źródło dostępu] http://www.stat.gov.pl/cps/rde/xbcr/gus/rs_rocznik_rolnictwa_2011.pdf, s. 114.

Suma z ostatniej kolumny, ze względu na wykorzystanie częstości, tworzy moment centralny rzędu czwartego: $\mu_4 = 700\,459{,}227$ ha^4. Obliczając kurtozę, należy wartość momentu podzielić przez czwartą potęgę odchylenia standardowego:

$$k = \frac{\mu_4}{s_x^4} = \frac{700\ 459{,}227}{(16{,}014)^4} = 10{,}651,$$

a następnie przekształcić kurtozę w eksces:

$$K = k - 3 = 10{,}651 - 3 = 7{,}951.$$

Otrzymaliśmy dość wysoką dodatnią wartość ekscesu, co świadczy o tym, że badany rozkład jest wysmuklony w środkowej części i rozciągnięty w ogonach, czyli jest to rozkład leptokurtyczny. Pracujący wyłącznie lub głównie w gospodarstwach byli zatrudnieni w gospodarstwach o powierzchni znacznie bardziej zbliżonej do średniej powierzchni, niż miałoby to miejsce w przypadku rozkładu normalnego.

■

Przykład 6.6 (kontynuacja 6.3)

Wracając do przykładu 6.3 o wieku posłów VI kadencji do Sejmu, należy zbadać, czy średnia dobrze odzwierciedla badany rozkład, czyli policzyć stopień koncentracji wieku. Obliczyliśmy już, że średnia wynosi $\bar{x} = 47{,}80$ lat, a odchylenie stan-

dardowe $s_x = 10$ lat. Wyliczenia pomocnicze związane z wyznaczeniem momentu centralnego rzędu czwartego przedstawiono w tabeli 6.11.

Tabela 6.11. Posłowie VI kadencji Sejmu Rzeczypospolitej Polskiej według wieku (stan na początek kadencji) – wyliczenia pomocnicze

$(x_{i_0} - x_{i_1})$	f_i	$\overset{o}{x}_i$	$\overset{o}{x}_i - \bar{x}$	$\left(\overset{o}{x}_i - \bar{x}\right)^4 f_i$
21-29	0,044	25,5	-22,30	10 881,08
30-39	0,172	35,0	-12,80	4 617,09
40-49	0,339	45,0	-2,80	20,84
50-59	0,352	55,0	7,20	945,96
60-70	0,093	65,0	17,20	8 139,48
Suma	1,000			24 604,45

Źródło: opracowanie własne na podstawie: [Dokument elektroniczny] *Rocznik Statystyczny Rzeczypospolitej Polskiej 2012*, [źródło dostępu] http://www.stat.gov.pl/cps/rde/xbcr/gus/RS_rocznik_statystyczny_rp_2012.pdf, s. 120.

Ponieważ w tym przypadku również podczas wyliczeń skorzystaliśmy z częstości zamiast liczebności, suma z ostatniej kolumny jest już momentem centralnym rzędu czwartego $\mu_4 = 24\,604{,}45$ (lat)^4. Iloraz obliczonej wartości i czwartej potęgi odchylenia standardowego tworzy wartość kurtozy:

$$k = \frac{\mu_4}{s_x^4} = \frac{24\,604{,}45}{(9{,}92)^4} = 2{,}541.$$

Zatem eksces wynosi:

$$K = k - 3 = 2{,}541 - 3 = -0{,}459.$$

Ujemna wartość ekscesu świadczy o tym, że badany rozkład jest spłaszczony w środkowej części i pogrubiony w ogonach. Rozkład wieku posłów na sejm VI kadencji jest więc platokurtyczny. Ma on nieco niższą koncentrację wokół wartości średniej niż rozkład normalny, czyli wiek poszczególnych posłów jest mniej skupiony wokół średniej, niż byłby w przypadku normalności rozkładu.

6.2.2. Miara Lorenza[142]

Omówmy teraz sposób mierzenia koncentracji w drugim tego słowa znaczeniu, bardziej zbliżonym do intuicyjnego rozumienia tego pojęcia, czyli jako stopnia nierównomierności[143] rozkładu wartości funduszu cechy pomiędzy poszczególne jednostki badanej populacji. Jest to miara o znacznie większym znaczeniu niż kurtoza i może być stosowana do każdego rodzaju rozkładu, bez względu na stopień jego asymetrii.

Wprowadźmy najpierw pewne pojęcie, które ułatwi dalsze objaśnienia.

Fundusz cechy jest to wartość zsumowanych poszczególnych wartości cechy posiadanej przez każdą jednostkę statystyczną w zbiorowości. W przypadku omawianego tu stale przykładowego rozkładu wagi studentek funduszem cechy „waga" jest suma wag poszczególnych kobiet, czyli ile one ważą łącznie. Właśnie funduszu cechy użyliśmy do policzenia średniej arytmetycznej tejże wagi.

Badanie koncentracji jest to więc badanie, jak fundusz cechy w populacji jest podzielony pomiędzy poszczególne jednostki statystyczne populacji.

Jeżeli całym funduszem cechy dysponuje jedna jednostka, mówi się wówczas o *koncentracji absolutnej*, natomiast jeżeli wszystkie jednostki dysponują taką samą sumą wartości cechy, wówczas występuje *brak koncentracji*, czyli fundusz cechy rozkłada się równomiernie pomiędzy poszczególne jednostki.

Aby lepiej zrozumieć pojęcie koncentracji absolutnej i jej braku, rozważmy poniższy przykład. Powiedzmy, że wydział dostaje od sponsora dwa tysiące złotych miesięcznie jako fundusz stypendialny z przeznaczeniem na stypendia naukowe dla dziesięciu najlepszych studentów. Jak zostanie podzielona ta suma, zależy tylko od wydziału. Jeśli owe dwa tysiące złotych (czyli cały fundusz) będzie dostawała jedna, najlepsza osoba, wówczas wszystkie pieniądze przeznaczone na stypendia będą skoncentrowane absolutnie na jednym studencie. Jeśli wydział podzieli tę sumę równo pomiędzy dziesięciu najlepszych studentów, wówczas każdy student dostaje tyle samo, czyli dwieście złotych. Oznacza to, że fundusz stypendialny jest rozłożony równomiernie. Jeżeli wysokość stypendium zostanie zróżnicowana w zależności od jakiegoś kryterium[144], otrzyma się wówczas pośredni stopień koncentracji funduszu stypendialnego.

Jest to najprostszy przykład wyjaśniający pojęcie koncentracji. Oczywiście w praktycznych zastosowaniach sprawa jest nieco bardziej skomplikowana, po-

[142] W literaturze można spotkać dwa sposoby pisania tego nazwiska. Np. Lange [1975, s. 188 i nast.] pisze o krzywej Lorentza, w Małej Encyklopedii Statystyki [1976, s. 278 i nast.] oraz w Słowniku terminów statystycznych Kendalla i Bucklanda [1986, s.77] jest mowa o krzywej Lorenza, podobnie w późniejszych polskich podręcznikach autorzy używają zapisu krzywa (lub miara) Lorenza.

[143] Lub równomierności jako zwierciadlanego odbicia problemu.

[144] Na przykład wysokości średniej oceny z egzaminów w poprzednich dwóch semestrach.

nieważ analizowane dane są szeregami statystycznymi znacznie bardziej rozbudowanymi, często w postaci przedziałowej, co komplikuje nieco tego typu badanie, niezbędne są bowiem pewne dodatkowe informacje, nie zawsze podawane przez konstruktorów szeregów cechy ciągłej. W przypadku jednak szeregów, co do których dysponujemy pełnymi informacjami, zbadanie koncentracji funduszu cechy miarą Lorenza jest sprawa dosyć prostą.

Najpopularniejszymi problemami makroekonomicznymi, przy analizie których bada się stopień koncentracji zjawiska (cechy), są: stopień koncentracji dochodów ludności, bogactw ludności, wielkości gospodarstw rolnych w danym kraju, liczby ludności w miastach czy wielkość produkcji w zakładach danej branży. Na szczeblu mikroekonomicznym bardzo popularnymi zagadnieniami, zwłaszcza w dydaktyce, są: badanie stopnia koncentracji mocy elektrowni, tonażu floty morskiej, łóżek szpitalnych w szpitalach określonej wielkości, badanie wielkości sprzedaży w określonej wielkości sklepach i sieciach handlowych. W badaniach rynku tego typu miary koncentracji można używać do analizy stopnia opanowania rynku przez określone marki czy firmy.

Idea pomiaru stopnia koncentracji cechy według Lorenza jest bardzo prosta. Porównuje się częstość występowania jednostek statystycznych w różnych przedziałach[145] wartości cechy z udziałami sumy wartości cechy w poszczególnych przedziałach[146] odniesionymi do funduszu cechy w całej zbiorowości. Czyli porównuje się, jakim procentem funduszu cechy całej zbiorowości dysponuje jaki procent jednostek statystycznych uporządkowanych według wartości. W tym celu tworzymy dwa szeregi skumulowane: odsetek funduszu cechy w poszczególnych przedziałach cechy oraz odsetek jednostek statystycznych z tych przedziałów. Następnie na osi odciętych (oś X) odkładamy skumulowane odsetki jednostek statystycznych a na osi rzędnych (oś Y) odkładamy skumulowany procent funduszu cechy w dyspozycji tych jednostek. Jeżeli *rozkład jest równomierny*, wówczas danemu odsetkowi jednostek statystycznych będzie towarzyszył taki sam odsetek funduszu cechy. Wykresem takiej sytuacji będzie przekątna kwadratu[147] 100% na 100%.

[145] Lub o różnych odmianach cechy, gdy jest skokowa.

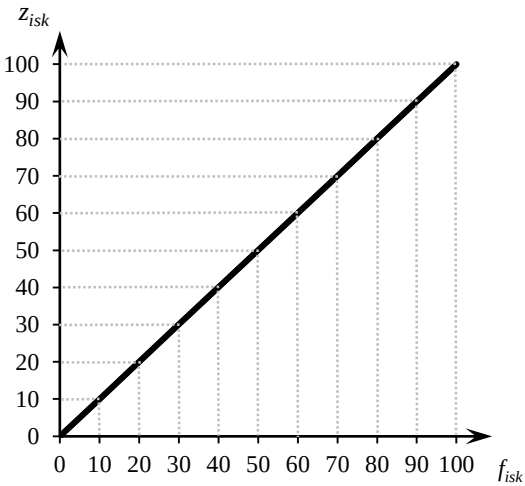
[146]Opiszemy badanie koncentracji na podstawie szeregu przedziałowego, albowiem jest to najczęściej występujący typ badania. Dla cech skokowych obliczenia wyglądają analogicznie.

[147] Najczęściej dla wygody odsetki podaje się w procentach; jeżeli pozostajemy przy zapisie dziesiętnym, wówczas należy o tym pamiętać i skorygować obliczenia, przesuwając odpowiednio przecinki, powstanie bowiem kwadrat 1x1, a nie 100%x100%.

Rysunek 6.3. Przykład koncentracji zerowej

Źródło: opracowanie własne.

Rysunek 6.3 ilustruje sytuację, gdy wszyscy studenci z naszego przykładu dostają tęą samą sumę jako stypendium, f_{isk} – skumulowany odsetek studentów otrzymujących stypendium danej wysokości, z_{isk} – skumulowany odsetek funduszu stypendialnego otrzymywanego przez dany procent funduszu stypendialnego. Pierwszych 10% studentów dostaje 10% funduszu stypendialnego, następne 10% studentów dostaje również 10% funduszu, co po kumulacji daje odpowiednio 20% populacji otrzymującej 20% funduszu i tak dalej. W efekcie, jak widzimy na wykresie 6.6, *krzywą ilustrującą brak koncentracji jest właśnie przekątna kwadratu koncentracji, zwana krzywą równomiernego rozkładu*.



f_{isk}	z_{sk}
10%	10%
20%	20%
30%	30%
40%	40%
50%	50%
60%	60%
70%	70%
80%	80%
90%	90%
100%	100%

Wykres 6.6. Krzywa Lorenza dla koncentracji zerowej

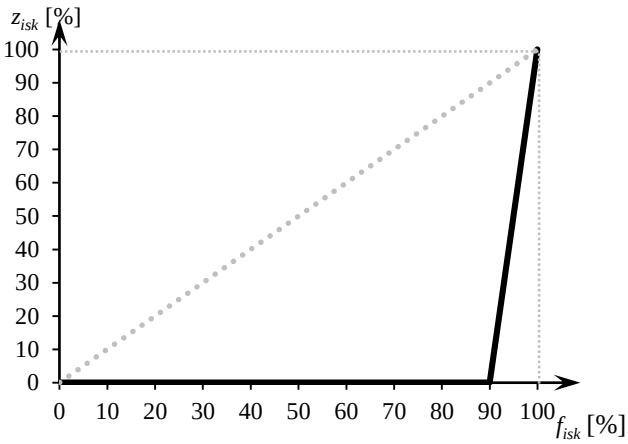
Źródło: opracowanie własne.

Na rysunku 6.4 przedstawiono z kolei drugą, krańcową sytuację *koncentracji absolutnej*.

Rysunek 6.4. Przykład koncentracji absolutnej

Źródło: opracowanie własne.

Na rysunku 6.4 widzimy, że 10% studentów otrzymuje cały fundusz stypendialny (100% pieniędzy), a pozostali nic. Krzywą ilustrująca tę sytuację jest łamana oparta o bok kwadratu na osi X oraz o niemal prostopadły do niego prawy bok kwadratu (wykres 6.7).



f_{isk}	z_{sk}
10%	0%
20%	0%
30%	0%
40%	0%
50%	0%
60%	0%
70%	0%
80%	0%
90%	0%
100%	100%

Wykres 6.7. Krzywa Lorenza dla koncentracji absolutnej

Źródło: opracowanie własne.

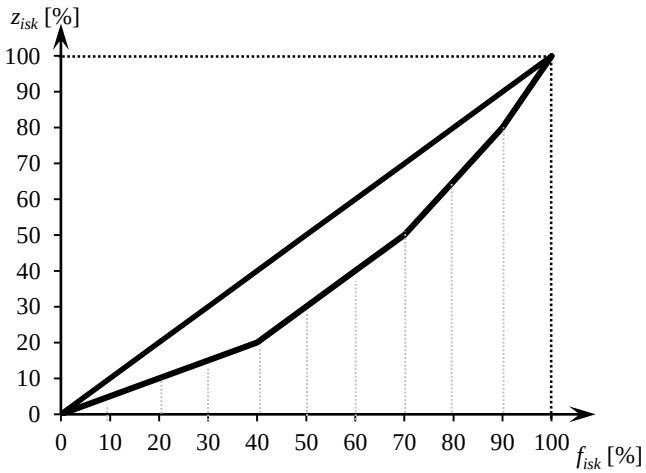
Na ogół mamy do czynienia z sytuacją pośrednią, kiedy fundusz stypendialny nie jest ani równomiernie rozłożony, ani absolutnie skoncentrowany. Przykładową sytuację przedstawiono na rysunku 6.5.

Rysunek 6.5. Przykład koncentracji pośredniej

Źródło: opracowanie własne.

Utworzona krzywa znajdzie się pomiędzy linią równomiernego podziału a krzywą absolutnej koncentracji. Im bardziej krzywa koncentracji odbiegnie od linii równomiernego podziału (wybrzuszy się w stronę osi X i prawego boku kwadratu), tym silniejszą koncentrację cechy w zbiorowości będzie reprezentowała. Wielkości tego odchylenia od linii braku koncentracji możemy więc użyć jako miary siły koncentracji cechy. *Dobrą miarą wielkości koncentracji jest więc pole pomiędzy przekątną kwadratu koncentracji (brak koncentracji) a krzywą koncentracji pośredniej, zwaną krzywą Lorenza.*

Pole utworzone dla omawianego przykładu przedstawiono na wykresie 6.8.



f_{isk}	z_{sk}
10%	5%
20%	10%
30%	15%
40%	20%
50%	30%
60%	40%
70%	50%
80%	65%
90%	80%
100%	100%

Wykres 6.8. Krzywa Lorenza dla koncentracji pośredniej

Źródło: opracowanie własne.

Ideę tej miary koncentracji zwięźle objaśnia Lange[148] [1975, s. 191]: „Krzywa koncentracji (albo wielobok koncentracji) oddala się od linii równomiernego rozkładu tym więcej, im bardziej nierównomierny jest rozkład wartości zmiennej

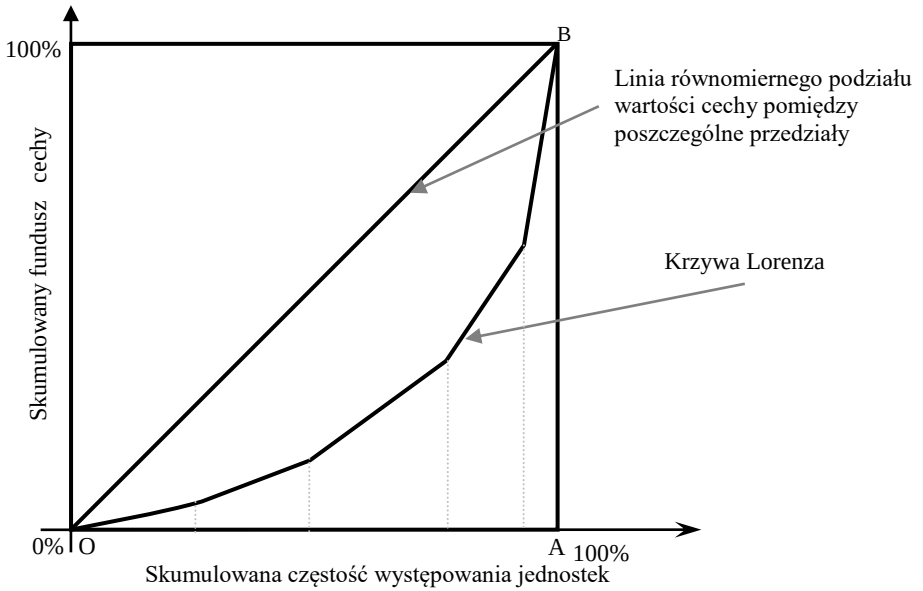
[148] Jeden z najbardziej znanych polskich ekonomistów.

między jednostki zbiorowości statystycznej. Jest tak dlatego, że im większy jest stopień koncentracji tych wartości u niektórych jednostek, tym powolniejszy jest wzrost krzywej koncentracji z początku, gdy dodajemy jednostki o małych wartościach zmiennej, a tym silniejszy później, gdy zaczynamy dodawać jednostki, u których są skoncentrowane wysokie wartości zmiennej. Dlatego przy większej koncentracji krzywa koncentracji jest „bardziej wklęsła" w stosunku do linii równomiernego rozkładu".

W skrajnym przypadku krzywa „idzie" po osi X i dopiero w punkcie $f_{isk} = 100\%$ skacze do punktu o współrzędnej $z_{isk} = 100\%$, pokazując, że cały fundusz cechy jest skoncentrowany w „rękach" jednej jednostki lub zbioru jednostek. Wiemy więc, że gdy brak jest koncentracji, wartość miary powinna wynosić zero, a w przypadku absolutnej koncentracji wynosi ona niemal 100%. Aby odkryć to, wystarczy popatrzeć na szeregi skumulowane liczebności i funduszu cechy oraz na wykres krzywej. W przypadku koncentracji pośredniej jesteśmy w stanie, patrząc na wykres, stwierdzić, że takowa istnieje, ale jeśli chcemy określić stopień koncentracji dokładniej, musimy wyznaczyć miarę liczbową[149].

Jak już wspomniano, im większa jest koncentracja, tym bardziej krzywa Lorenza odchyla się od linii równomiernego podziału. Stąd dosyć oczywisty wniosek, że wielkość pola pomiędzy nimi będzie dobrze reprezentować siłę koncentracji, że im większe jest pole, tym silniejsza koncentracja. Znacznie jednak wygodniejsze jest przekształcenie bezwzględnej miary, jaką jest pole figury, do postaci względnej. Pole koncentracji badanej cechy odnosimy do pola koncentracji absolutnej, jako maksymalnie możliwej do uzyskania, i w ten sposób otrzymujemy miarę względną mówiącą nam, jak „daleko" otrzymany wynik jest „odległy" od wartości maksymalnej wynoszącej sto procent (lub jeden w zapisie dziesiętnym).

[149] A tak jest najczęściej. Zwłaszcza w przypadku porównań i ocen wygodniej jest posłużyć się liczbami, by określić, czy coś jest duże czy małe albo która zbiorowość lub która cecha ma silniejszą (słabsza) koncentrację.



Rysunek 6.6. Krzywa Lorenza

Źródło: opracowanie własne.

Zatem aby określić, jaka jest siła koncentracji, należy obliczyć pole figury znajdującej się pomiędzy krzywą Lorenza a linią równomiernego podziału OB (rysunek 6.6). Jeżeli wielkość tego pola określimy jako P_k, fundusz cechy i częstość skumulowaną przedstawimy w postaci procentowej, wówczas miara Lorenza przyjmie postać:

$$K_L = \frac{P_k}{5\,000}, \tag{6.23}$$

gdzie wartość z mianownika oznacza wielkość pola trójkąta poniżej linii równomiernego podziału, to jest połowę pola kwadratu koncentracji $\left(P_\Delta = \frac{1}{2} \cdot a^2 = \frac{1}{2} \cdot 100\% \cdot 100\% = 5\,000\%\% \right)$.

Niestety, na ogół nie znamy postaci analitycznej krzywej Lorenza, a dysponujemy jedynie wielobokiem (krzywą łamaną) koncentracji i nie możemy bezpośrednio obliczyć interesującego nas pola koncentracji (P_k). Obchodzimy ten problem, obliczając najpierw pole pod krzywą koncentracji w sposób przybliżony, a następnie, pamiętając, że maksymalne pole koncentracji to $P = 5000$, interesujące nas pole P_k obliczamy jako różnicę pomiędzy polem maksymalnym a polem pod krzywą koncentracji, co zapiszemy jako:

$$K_L = \frac{P_{\max} - P}{P_{\max}} = \frac{P_k}{P_{\max}} = \frac{5000 - P}{5000} = 1 - \frac{P}{5000}$$

lub w zapisie dziesiętnym

$$K_L = \frac{P_{\max} - P}{P_{\max}} = \frac{P_k}{P_{\max}} = \frac{0{,}5 - P}{0{,}5} = 1 - 2P.$$

Na podstawie powyższego zapisu łatwo też zauważyć i zapamiętać, dlaczego miara koncentracji Lorenza jest ograniczona i w jakich znajduje się granicach. Minimalne pole figury w geometrii euklidesowej wynosi zero (pole prostej), a maksymalne w tym przypadku to pole połowy kwadratu o boku o wymiarze 100%. Stąd, przypomnijmy, miara Lorenza zawiera się w przedziale: $0 \le K_L \le 100\%$ lub w alternatywnym zapisie w przedziale: $0 \le K_L \le 1$.

Pozostaje pytanie, jak obliczyć pole *P* pod wielobokiem koncentracji. Problem jest dosyć prosty do rozwiązania. Jak łatwo zauważyć na rysunku, pole pod wielobokiem (krzywą łamaną) możemy rozłożyć na sumę pól trapezów. Podstawami tych trapezów są odcięte przeprowadzone w poszczególnych punktach kumulacji funduszu cechy, a wysokościami są liczebności względne (odsetki) poszczególnych przedziałów cechy (lub odmian cechy skokowej). Łatwiej będzie to zrozumieć na poniższym przykładzie 6.8.

Przykład 6.8

Tabela 6.11 przedstawia dane dotyczące polskich miast według ich wielkości i liczby ludności je zamieszkującej (stan na dzień 31.12.2011 r.). Korzystając z zawartych tam informacji, należy zbadać siłę koncentracji ludności w polskich miastach, stosując z miarę Lorenza.

Tabela 6.12. Miasta według liczby mieszkańców (stan na dzień 31.12.2011 r.)

Miasta o liczbie ludności (w tys.)	Liczba miast	Ludność w danym typie miasta (w tys.)
do 5	317	970,8
5-10	183	1 304,0
10-20	185	2 700,9
20-50	136	4 264,0
50-100	48	3 250,4
100-200	22	3 005,6
200 i więcej	17	7 890,1
Razem	908	23 385,8

Źródło: [Dokument elektroniczny] *Rocznik Demograficzny 2012*,
[źródło dostępu] http://www.stat.gov.pl/cps/rde/xbcr/gus/rs_rocznik_demograficzny_2012.pdf, s. 82.

Wyliczenia związane z określeniem poziomu koncentracji należy rozpocząć od wyznaczenia częstości (f_i) i odsetka całkowitego funduszu cechy w danym typie miasta (z_i) przedstawionych w formie procentowej. Częstość wyznaczana jest jako iloraz liczebności z poszczególnych klas i ogólnej liczebności zbiorowości. W analizowanym przykładzie w pierwszym wierszu częstość ta wyznaczana jest jako iloraz liczby miast mających poniżej 5 tys. mieszkańców przez liczbę wszystkich miast $\left(f_1 = \frac{317}{908} \cdot 100\% = 34{,}9\% \right)$, w drugim – liczby miast, w których jest od 5 do 10 tys. mieszkańców, podzielona przez liczbę wszystkich miast $\left(f_2 = \frac{183}{908} \cdot 100\% = 20{,}1\% \right)$ itd.

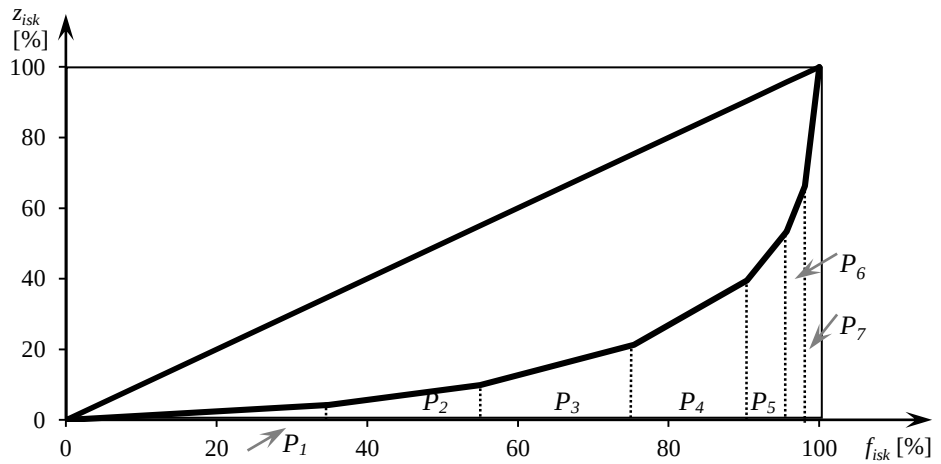
Analogicznie należy postąpić z łącznym funduszem cechy. W analizowanym przykładzie funduszem cechy jest łączna liczba ludności zamieszkująca poszczególne rodzaje miast. I tak w pierwszym rodzaju miast, czyli mających mniej niż 5 tys. mieszkańców, mieszka łącznie 970,8 tys. osób, czyli $z_1 = \frac{970{,}8}{23\,385{,}8} \cdot 100\% = 4{,}2\%$, co oznacza, że w 2011 roku w miastach o liczbie ludności poniżej pięciu tysięcy mieszkało 4,2% ogółu ludności mieszkającej w polskich miastach. Łączna liczba mieszkańców miast mających od 5 do 10 tys. mieszkańców wynosi 1 304,0 tys. osób, więc $z_2 = \frac{1304{,}0}{23\,385{,}8} \cdot 100\% = 5{,}6\%$, czyli w tej grupie znajduje się 5,6% ogółu ludności miejskiej. Analogiczne obliczenia wykonuje się dla pozostałych klas cechy. Wszystkie niezbędne obliczenia, łącznie z opisanymi powyżej, przedstawiono w tabeli 6.13 w czwartej i piątej kolumnie.

Tabela 6.13. Miasta według liczby mieszkańców (stan na dzień 31.12.2011 r.) – wyliczenia pomocnicze

Miasta o liczbie ludności (w tys.)	Liczba miast	Ludność w miastach (w tys.)	Liczba miast w %	Ludność w danym typie miast w % ludności miejskiej ogółem	Skumulowana liczba miast w % ogółem miast	Skumulowana ludność w miastach w % ludności miejskiej
	n_i	Σx_i	f_i	z_i	f_i^{sk}	z_i^{sk}
do 5	317	970,8	34,9	4,2	*34,9*	*4,2*
5-10	183	1 304,0	20,1	5,6	*55,0*	*9,8*
10-20	185	2 700,9	20,4	11,5	*75,4*	*21,3*
20-50	136	4 264,0	15,0	18,2	*90,4*	*39,5*
50-100	48	3 250,4	5,3	13,9	*95,7*	*53,4*
100-200	22	3 005,6	2,4	12,9	*98,1*	*66,3*
200 i więcej	17	7 890,1	1,9	33,7	*100,0*	*100,0*
Razem	**908**	**23 385,8**	100,0	100,0	**Oś X**	**Oś Y**

Źródło: opracowanie własne na podstawie: [Dokument elektroniczny] *Rocznik Demograficzny 2012*, [źródło dostępu] http://www.stat.gov.pl/cps/rde/xbcr/gus/rs_rocznik_demograficzny_2012.pdf, s. 82.

Otrzymaną częstość i łączny fundusz cechy w ujęciu procentowym dla poszczególnych odmian cechy kumulujemy. Kumulacje te umieszczono w dwóch ostatnich kolumnach tabeli 6.11. W oparciu o te dwie ostatnie kolumny wykreślamy krzywą[150] Lorenza, co przedstawia wykres 6.9, przy czym przypomnijmy, że teraz, nietypowo w stosunku do poprzednich wykresów[151], skumulowaną częstość f_i^{sk} odkładamy na osi X , zaś skumulowany fundusz cechy z_i^{sk} na osi Y.



Wykres 6.9. Krzywa Lorenza koncentracji liczby mieszkańców w polskich miastach

Źródło: opracowanie własne na podstawie: [Dokument elektroniczny] *Rocznik Demograficzny 2012*, [źródło dostępu] http://www.stat.gov.pl/cps/rde/xbcr/gus/rs_rocznik_demograficzny_2012.pdf, s. 82.

Analizując wykres 6.9, można zauważyć, że pole figury utworzonej przez krzywą Lorenza i linię równomiernego podziału zajmuje więcej niż połowę powierzchni trójkąta znajdującego się poniżej podziału przekątnej kwadratu, co oznacza, że obliczana miara Lorenza powinna wynosić więcej niż 50% (lub 0,5 w zapisie dziesiętnym) i że jakaś koncentracja ludności w miastach występuje. Aby ją dokładnie określić, należy policzyć pole figury koncentracji P_k. Jednak, jak już powiedziano, zdecydowanie łatwiejszym zadaniem jest wyznaczenie pól trapezów poniżej krzywej Lorenza, oznaczonych na rysunku 6.6 jako symbole od P_1do P_7 (przy czym pierwszy z nich jest trapezem zdegenerowanym, czyli trójkątem).

[150] Właściwie bardziej poprawnie należałoby powiedzieć: wielobok Lorenza.

[151] Nietypowość polega na tym, że na poprzednich wykresach częstości, liczebności, kumulacje liczebności czy kumulacje częstości były odkładane na osi OY.

Pierwszą figurą znajdującą się poniżej krzywej Lorenza jest trójkąt. Zatem jego pole wyznaczane jest następująco[152]:

$$P_1 = \frac{1}{2}ah = \frac{1}{2}z_1^{sk} f_1 = \frac{1}{2}\cdot 4,2\cdot 34,9 = 73,29.$$

Pozostałe pola obliczamy, stosując wzór na pole trapezu[153]:

$$P_2 = \frac{1}{2}(a+b)h = \frac{1}{2}\cdot(z_1^{sk} + z_2^{sk})\, f_2 = \frac{1}{2}\cdot(4,2+9,8)\cdot 20,1 = 140,70;$$

$$P_3 = \frac{1}{2}\cdot(z_2^{sk} + z_3^{sk})\, f_3 = \frac{1}{2}\cdot(9,8+21,3)\cdot 20,4 = 317,22;$$

$$P_4 = \frac{1}{2}\cdot(z_3^{sk} + z_4^{sk})\, f_4 = \frac{1}{2}\cdot(21,3+39,5)\cdot 15,0 = 456,00;$$

$$P_5 = \frac{1}{2}\cdot(z_4^{sk} + z_5^{sk})\, f_5 = \frac{1}{2}\cdot(39,5+53,4)\cdot 5,3 = 246,19;$$

$$P_6 = \frac{1}{2}\cdot(z_5^{sk} + z_6^{sk})\, f_6 = \frac{1}{2}\cdot(53,4+66,3)\cdot 2,4 = 143,64;$$

$$P_7 = \frac{1}{2}\cdot(z_6^{sk} + z_7^{sk})\, f_7 = \frac{1}{2}\cdot(66,3+100,0)\cdot 1,9 = 157,99.$$

Całkowita suma wyznaczonych pól to:

$$P = \sum_{i=1}^{7} P_i = 1\,535,03.$$

Zatem pole koncentracji pomiędzy linią równomiernego podziału a krzywą Lorenza wynosi:

$$P_k = 5\,000 - P = 500 - \sum_{i=1}^{7} P_i = 5\,000 - 1\,535,03 = 3\,464,97.$$

Stąd miara Lorenza obliczona ze wzoru (6.23) wynosi:

[152] Należy skorzystać ze wzoru na pole trójkąta, obliczane jako połowa iloczynu długości podstawy i wysokości.

[153] Przypomnijmy, pole trapezu to połowa iloczynu sumy długości podstaw i wysokości.

$$K_L = \frac{P_k}{5\,000} \cdot 100\% = \frac{3\,464{,}97}{5\,000{,}00} \cdot 100\% = 0{,}693 \cdot 100\% = 69{,}3\%$$

Koncentracja na poziomie 69,3% oznacza, że ludność w Polsce w 2011 roku wykazywała sporą koncentrację w dużych miastach.

■

Przykład 6.9 (kontynuacja 3.3)

Wracając do prezentowanego uprzednio przykładu dotyczącego wagi studentek, obliczmy poziom koncentracji wagi w rozumieniu miary Lorenza. Cofnijmy się do przykładu 3.3, w którym konstruowaliśmy szereg rozdzielczy przedziałowy. Aby można było wyznaczyć miarę Lorenza, należy wyznaczyć fundusz cechy dla każdego przedziału[154]. W analizowanym przykładzie był on obliczany jako element badania poprawności grupowania. W tabeli 6.14 przedstawiono szereg rozdzielczy z uwzględnieniem funduszu cechy.

Tabela 6.14. Waga studentek kierunku zarządzanie – przypomnienie

Przedziały	Pomiary należące do poszczególnych przedziałów	Liczebność	Suma łącznej wagi w poszczególnych przedziałach
$(x_{i_0} - x_{i_1})$	x_i	n_i	$\sum x_i$
42-46	44; 45	2	89
46-50	46; 46; 47; 47; 47; 48; 48; 48; 48; 48; 49; 49; 49; 49; 49; 49	16	767
50-54	50; 50; 50; 51; 51; 51; 51; 52; 52; 52; 52; 52; 52; 52; 52; 53; 53; 53; 53; 53; 53; 53; 53; 53; 53; 53; 53; 53	26	1 352
54-58	54; 54; 54; 54; 54; 54; 54; 54; 55; 55; 55; 55; 55; 55; 55; 56; 56; 56; 56; 57; 57; 57; 57; 57; 57; 57; 57; 57; 57; 57	30	1 668

[154] Zauważmy, że w wyjściowym szeregu informacji o łącznej wadze w przedziałach nie posiadamy. Musieliśmy ją utworzyć na podstawie danych surowych. To pokazuje problem z zastosowaniem miary Lorenza. Jeśli mamy szereg bez tej informacji i nie mamy dostępu do danych surowych, nie jesteśmy w stanie użyć miary Lorenza. „Sztuczka" ze środkami przedziałów da przekłamane wyniki i nie rekomendujemy jej, wbrew niektórym podręcznikom.

Przedziały	Pomiary należące do poszczególnych przedziałów	Liczebność	Suma łącznej wagi w poszczególnych przedziałach
$(x_{i_0} - x_{i_1})$	x_i	n_i	$\sum x_i$
58-62	58; 58; 58; 58; 58; 58; 59; 59; 59; 59; 59;60; 60; 60; 61; 61; 61; 61; 61	19	1 128
62-66	63; 63; 64; 64; 65	5	319
66-70	67; 69	2	136

Źródło: opracowanie własne na podstawie danych umownych.

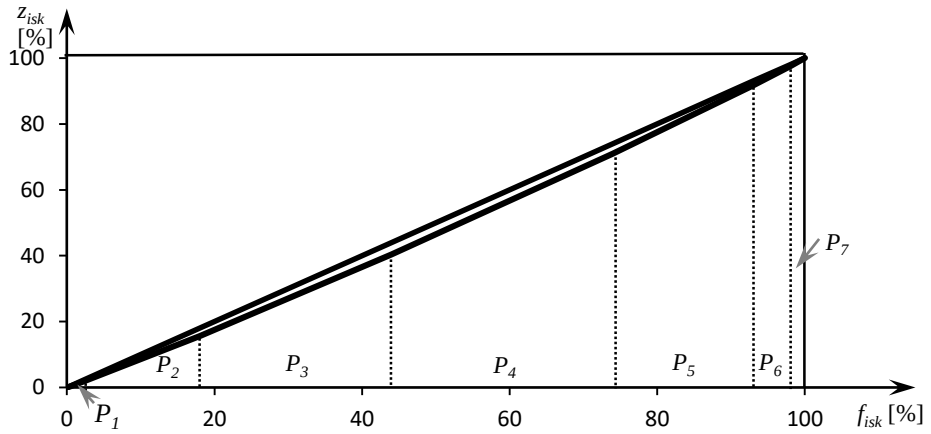
Przedstawione w tabeli 6.14 wyniki mogą teraz zostać wykorzystane do określenia poziomu koncentracji wagi studentek w badanej zbiorowości. Częstość i łączny fundusz cechy w procentowej postaci zostały przedstawione w tabeli 6.15.

Tabela 6.15. Waga studentek kierunku zarządzanie – wyliczenia pomocnicze miary Lorenza

$(x_{i_0} - x_{i_1})$	n_i	$\sum x_i$	f_i[%]	z_i[%]	f_{isk} [%]	z_{isk} [%]
42-46	2	89	2	1,6	2	1,6
46-50	16	767	16	14,0	18	15,6
50-54	26	1 352	26	24,8	44	40,4
54-58	30	1 668	30	30,6	74	71,0
58-62	19	1 128	19	20,7	93	91,7
62-66	5	319	5	5,8	98	97,5
66-70	2	136	2	2,5	100	100,0
Suma	100	5 459	100	100,0		

Źródło: opracowanie własne na podstawie danych umownych.

Dwie ostatnie kolumny z tabeli 6.14 są współrzędnymi punktów umieszczonych na wykresie 6.4. Punkty te po połączeniu linią utworzyły krzywą Lorenza.



Wykres 6.10. Krzywa Lorenza koncentracji wagi studentek kierunku zarządzanie

Źródło: opracowanie własne na podstawie danych umownych.

Analizując wykres 6.10, można zauważyć, że pole utworzone pomiędzy linią równomiernego podziału a krzywą Lorenza jest niewielkie, zatem waga dość równomiernie rozkłada się w badanej grupie kobiet. Aby precyzyjnie określić stopień koncentracji, należy wyznaczyć wielkość pola utworzonej figury. Zastosowano w tym przypadku, podobnie jak w poprzednim przykładzie, schemat postępowania polegający na wyznaczeniu pól figur powstałych poniżej krzywej Lorenza i odjęciu ich sumy od pola trójkąta. Pierwszą figurą znajdującą się poniżej krzywej Lorenza zawsze jest trójkąt, zaś kolejne to trapezy. Ich pola zostały obliczone następująco:

$$P_1 = \frac{1}{2} \cdot 2 \cdot 1,6 = 1,60;$$

$$P_2 = \frac{1}{2} \cdot (1,6 + 15,6) \cdot 16 = 137,60;$$

$$P_3 = \frac{1}{2} \cdot (15,6 + 40,4) \cdot 26 = 728,00;$$

$$P_4 = \frac{1}{2} \cdot (40,4 + 71,0) \cdot 30 = 1\,671,00;$$

$$P_5 = \frac{1}{2} \cdot (71,0 + 91,7) \cdot 19 = 1\,545,65;$$

$$P_7 = \frac{1}{2} \cdot (97,5 + 100,0) \cdot 2 = 197,5.$$

Zatem pole utworzonej figury pomiędzy linią równomiernego podziału a krzywą Lorenza wynosi:

$$P_k = 5\ 000 - \sum_{i=1}^{7} P_i = 5\ 000 - 4\ 754{,}35 = 245{,}65,$$

zaś miara Lorenza wyznaczona ze wzoru (6.23) to:

$$K_L = \frac{P_k}{5\ 000} = \frac{245{,}65}{5\ 000{,}00} = 0{,}05.$$

Otrzymany wynik na poziomie 0,05 (lub 5%) informuje nas, że waga studentek w badanej zbiorowości, praktycznie rzecz biorąc, nie jest skoncentrowana w jakimś przedziale cechy i jest rozłożona dość równomiernie w badanej populacji. Można to zauważyć, śledząc, jakim częstościom występowania poszczególnych odmian cechy odpowiadają jakie udziały funduszu cechy, czyli patrząc na kolumny czwartą i piątą tabeli 6.13. Łatwo zauważyć, że procenty udziałów są w obu kolumnach zbliżone, co odpowiada brakowi koncentracji.

■

Kluczowe pojęcia

Rozkład (szereg) symetryczny
Asymetria dodatnia, prawostronna
Asymetria ujemna, lewostronna
Rozkład normalny
Wskaźnik skośności
Współczynnik asymetrii
Rozkład leptokurtyczny
Rozkład platokurtyczny
Kurtoza
Eksces
Linia równomiernego podziału
Krzywa Lorenza, wielobok liczebności Lorenza

Pytania kontrolne

1. Czym charakteryzuje się szereg symetryczny?
2. Jakie znasz wskaźniki asymetrii rozkładu?
3. Czym się różnią wskaźniki i współczynniki asymetrii rozkładu?

4. Jaka relacja zachodzi między dominantą, średnią arytmetyczną i medianą w przypadku: asymetrii lewostronnej rozkładu, asymetrii prawostronnej rozkładu, rozkładu symetrycznego? Podaj interpretację, ilustrując ją przykładem.
5. Zbadano ze względu na wzrost dwie drużyny koszykówki. W pierwszej wskaźnik asymetrii wyniósł 0,52, zaś w drugiej 0,45. Którą drużynę koszykówki (biorąc pod uwagę jedynie wzrost) powinien wybrać trener? Odpowiedź uzasadnij.
6. Zbadano zawartość substancji smolistych (szkodliwe dla zdrowia) w wypalanych papierosach tej samej marki wytwarzanych przez dwie różne fabryki. Otrzymany rozkład zawartości dla pierwszego wytwórcy cechuje umiarkowana asymetria ujemna (lewostronna), zaś dla drugiego – dodatnia (prawostronna). Mając wybór wytwórcy, którego należy wybrać i dlaczego?
7. Co to jest moment zwykły i centralny w statystyce? Czym się one różnią?
8. Co nazywamy kurtozą?
9. Do czego służy kurtoza (co mierzy)?
10. Co jest punktem odniesienia w interpretacji kurtozy badanego rozkładu?
11. Co nazywamy krzywą Lorenza?
12. Co nazywamy miarą (wskaźnikiem) Lorenza?

Spis tabel

Spis rysunków

Spis wykresów

Literatura cytowana

[1] Aczel A. D., *Statystyka w zarządzaniu*, PWN, Warszawa 2000.

[2] Bąk I., Markowicz I., Mojsiewicz M., Wawrzyniak K., *Statystyka w zadaniach.* Część I. *Statystyka opisowa*, Wydawnictwo Naukowo-Techniczne, Warszawa 2002.

[3] Blalock H. M., *Statystyka dla socjologów,* PWN, Warszawa 1975.

[4] Domański Cz., *Początki statystyki na ziemiach polskich*, „Wiadomości Statystyczne" nr 9, 2004.

[5] Kendall M.G., Buckland W.R., *Słownik terminów statystycznych*, PWE, Warszawa 1986.

[6] Lange O., *Dzieła. Statystyka,* t. 4, PWE, Warszawa 1975.

[7] Luszniewicz A., *Statystyka ogólna*, PWE, Warszawa 1973.

[8] *Mała Encyklopedia Statystyki*, PWE, Warszawa 1976.

[9] Młodak A., *Analiza taksonomiczna w statystyce regionalnej*, „Difin", Warszawa 2006.

[10] Ostasiewicz W., *Badania statystyczne*, Oficyna Wolters Kluwer Business, Warszawa 2011.

[11] Panek T., *Statystyczne metody wielowymiarowej analizy porównawczej*, Szkoła Główna Handlowa - Oficyna Wydawnicza, Warszawa 2009.

[12] Podgórski J., *Statystyka dla studiów licencjackich*, Polskie Wydawnictwo Ekonomiczne, wyd. 2 zmienione, Warszawa 2005.

[13] Puchalski T., *Statystyka opisowa*, wyd. II zmienione, PWN, Warszawa 1978.

[14] Puchalski T., *Statystyka. Wykład podstawowych zagadnień*, PWN, Warszawa 1980.

[15] Rao C.R., *Statystyka i prawda*, Wydawnictwo Naukowe PWN, Warszawa 1994.

[16] Szulc B., *Statystyka dla ekonomistów. Opis statystyczny*, PWE, Warszawa 1972.

[17] Yule G.U., Kendall M.G., *Wstęp do teorii statystyki*, PWN, Warszawa 1966.

Literatura dostępna

[1] Abt S., *Metody analizy statystycznej*, Akademia Ekonomiczna w Poznaniu, „Materiały Dydaktyczne" nr 41, Poznań 1999.

[2] Adamkiewicz H.G., *Statystyka. Zastosowania w ekonomii: analiza struktury zbiorowości statystycznej, szeregów czasowych oraz współzależności zjawisk*, Gdańsk: Ośrodek Doradztwa i Doskonalenia Kadr, 1996.

[3] Augustyniak H., *Statystyka opisowa z elementami demografii*, Przedsiębiorstwo Wydawnicze Ars Boni et Aequi, Poznań 2007.

[4] Bąk I., Markowicz I., Mojsiewicz M., Wawrzyniak K., *Statystyka opisowa: przykłady i zadania*, CeDeWu, Warszawa 2015.

[5] Bielecki J., Jurkiewicz B., Szymanowska Z., *Zbiór zadań ze statystyki ogólnej i matematycznej*, PWN, Warszawa 1978.

[6] Bielecki J., Makać W., Szymanowska Z., *Statystyka społeczno-ekonomiczna w zadaniach*, PWN, Warszawa 1975.

[7] Buga J., Kassyk-Rokicka H., *Podstawy statystyki opisowej*, Vizja PRESS & IT, Warszawa 2008.

[8] Clegg F., *Po prostu statystyka*, Wydawnictwa Szkolne i Pedagogiczne, Warszawa 1994.

[9] Hozer J., (red.) *Statystyka. Opis statystyczny*, Katedra Ekonometrii i Statystyki, Uniwersytet Szczeciński, Szczecin 1998.

[10] Jóźwiak J., Podgórski J., *Statystyka od podstaw*, PWE, Warszawa 2006 i wcześniejsze wydania.

[11] Kassyk-Rokicka H. (red.), *Statystyka. Zbiór zadań*, PWE, Warszawa 1994 i nast. wyd.

[12] Kassyk-Rokicka H., *Statystyka nie jest trudna. Mierniki statystyczne*, PWE, Warszawa 2011.

[13] Kendall M.G., Buckland W.R., *Słownik terminów statystycznych*, PWE, Warszawa 1986.

[14] Konopacki, G., & Worwa, K., *Wybrane zagadnienia ze statystyki opisowej*, Wyższa Szkoła Ekonomiczna w Warszawie, Warszawa 2002.

[15] Kukuła K., *Elementy statystyki w zadaniach*, PWN, Warszawa 1998.

[16] Kulczycki R., *Rachunek indeksowy*, PWE, Warszawa 1983.

[17] Luszniewicz A., Słaby T., *Statystyczne analizy z użyciem pakietu Statistica PL*, Sylabusy 1, 2, 3, 4, 5, 6, Wyższa Szkoła Handlu i Prawa, Warszawa 2000.

[18] Luszniewicz A., Słaby T., *Statystyka stosowana*, PWE, Warszawa 1996.
[19] Luszniewicz A., Słaby T., *Statystyka. Zadania testowe oraz sylabusy komputerowe*, SGH, Warszawa 1995
[20] Luszniewicz A., Słaby T., *Statystyka z użyciem pakietu Statistica PL*, Wydawnictwo C.H. Beck, Warszawa 2003.
[21] Łapkowska-Baster B., *Miary struktury zbiorowości w statystyce opisowej: przykłady i zadania*, Wydawnictwo Uniwersytetu Jagiellońskiego, Kraków 2007.
[22] Makać W., Urbanek-Krzysztofiak D., *Metody opisu statystycznego*, Wydawnictwo Uniwersytetu Gdańskiego, Gdańsk 1996.
[23] Maksimowicz-Ajchel A., *Wstęp do statystyki: metody opisu statystycznego*, Wydawnictwo Uniwersytetu Warszawskiego, Warszawa 2007.
[24] Ostasiewicz S. i in. *Statystyka, Materiały do ćwiczeń*. Książka Ekonomiczna, Wrocław 1994.
[25] Ostasiewicz W., *Myślenie statystyczne*, Oficyna Wolters Kluwer Business, Warszawa 2012.
[26] Paradysz J. (red.), *Statystyka w przykładach i zadaniach*, AE Poznań, Poznań 1996.
[27] Pusz P., Zaręba L., *Elementy statystyki*, Wydawnictwo Oświatowe FOSZE, Rzeszów 2006.
[28] Rogoziński Z., *Wstęp do statystyki społecznej*, PWN, Warszawa 1969.
[29] Sobczyk M., *Statystyka opisowa*, Wydawnictwo C.H. Beck, Warszawa 2010.
[30] Sobczyk M., *Statystyka*, Wydawnictwo PWN, Warszawa 1996 .
[31] Starzyńska W., *Statystyka praktyczna*, Wydawnictwo Naukowe PWN, Warszawa 2005.
[32] Steczkowski J., *Opis statystyczny: pozyskiwanie, przetwarzanie i analizowanie informacji*, Wydawnictwo Wyższej Szkoły Informatyki i Zarządzania z siedzibą w Rzeszowie, Rzeszów 2005.
[33] Stokowski F., *Statystyka. Teoria – przykłady – zadania*, Wydawnictwa SGPiS, Warszawa 1991.
[34] Walker J. A., McLean M.M., *Statystyka dla każdego*, WSiP, Warszawa 1994.
[35] Wasilewska E., *Statystyka opisowa nie tylko dla socjologów*, Wydawnictwo SGGW, Warszawa 2008.
[36] Wasilewska E., *Statystyka opisowa od podstaw: podręcznik z zadaniami,* wyd. 2 popr. i rozsz., Wydawnictwo SGGW, Warszawa 2011.
[37] Wieczorkowska G., Wierzbiński J., *Statystyka. Analiza badań społecznych*, Wydawnictwo Naukowe SCHOLAR, Warszawa 2017.

[38] Woźniak M. (red.), *Statystyka ogólna*, Akademia Ekonomiczna w Krakowie, Kraków 1997.
[39] Zając K., *Zarys metod statystycznych*, PWE, Warszawa 1994.
[40] Zeliaś A., Pawełek B., Wanat S., *Metody statystyczne. Zadania i sprawdziany*, PWE, Warszawa 2002.
[41] Zeliaś A., *Metody statystyczne*, Polskie Wydawnictwo Ekonomiczne, Warszawa 2000.

Indeks pojęć