Barbara Olbrych
Sylwia Rudecka-Gutkowska

STATYSTYKA OPISOWA W PRAKTYCE

Radom 2024

Recenzent
prof. dr hab. inż. Zbigniew Kosma

Procedura recenzowania jest zgodna z wytycznymi Ministerstwa Nauki i Szkolnictwa Wyższego, które zawarto w broszurze pt. ”Dobre praktyki w procedurach recenzyjnych w nauce”, Warszawa 2011

Projekt okładki
M&Z Frankiewicz s.c.

Redakcja językowa
Barbara Jaworska

Skład i łamanie
Cezary Majewski

e - ISBN 978-83-62491-82-7

Wydawnictwo: Akademia Handlowa Nauk Stosowanych w Radomiu
26-600 Radom, ul. Mazowieckiego 7a, tel: (48) 363 22 90
e-mail: wydawnictwo@ahns.pl,www.ahns.pl

Celem obliczeń nie są same liczby,
lecz ich zrozumienie.

Richard Hamming

SPIS TREŚCI

WSTĘP

Złożoność współczesnej rzeczywistości znajduje potwierdzenie w różnorodności procesów i zjawisk, często nienazwanych, które wywierają istotny wpływ na życie jednostek ludzkich, funkcjonowanie grup społecznych oraz całych państw. Procesy te i zjawiska nie tylko zmieniają nasze otoczenie, ale wymuszają konieczność adaptacji do nowo tworzących się warunków. Pojawienie się koncepcji społeczeństwa ponadnarodowego, wyłaniającego się pod wpływem handlu międzynarodowego, migracji ludności oraz oddziaływania organizacji spowodowało uniwersalizację życia społeczeństw. Niewątpliwie wielce pomocne w rozumieniu współczesnego życia jest gromadzenie i przetwarzanie informacji dotyczących określonej dziedziny społeczno-gospodarczej. A to wiąże się z pojęciem „statystyka", co w najprostszym rozumieniu tego słowa oznacza zbiór danych liczbowych opisujących wybrane zjawiska, przede wszystkim o charakterze masowym.

Statystyka jest nauką o metodach badania prawidłowości, występujących w zbiorowościach, charakteryzując te prawidłowości za pomocą liczb. Prawidłowości statystyczne są wypadkową oddziaływania na zjawiska masowe, czyli wtedy, gdy badaniu podlega duża liczba jednostek zbiorowości. Początkowo statystyka kojarzona była ze zbieraniem i porządkowaniem danych w postaci zestawień tabelarycznych. Z czasem nasunęła się konieczność wyrażenia właściwości całego materiału liczbowego za pomocą jednej liczby lub kilku liczb. Wtedy statystyka przekształciła się w metodę naukową, opartą na teorii rachunku prawdopodobieństwa, stała się częścią matematyki stosowanej. Obecnie statystyka zajmuje się nie tylko gromadzeniem i przedstawianiem danych w postaci tablic i wykresów, lecz służy także do podejmowania decyzji w warunkach niepewności. Dlatego w statystyce można wyróżnić dwa pozornie odrębne obszary: statystykę opisową (opis statystyczny) oraz wnioskowanie statystyczne (statystykę matematyczną). Statystyka opisowa jest dyscypliną zajmującą się metodami gromadzenia, opracowywania, prezentacji i analizy danych liczbowych, dotyczących badanych zbiorowości, osób, rzeczy lub zdarzeń. Wnioskowanie statystyczne polega na uogólnianiu wyników badania części zbiorowości zwanej próbą losową na całą zbiorowość (populację), z której ta część pochodzi z równoczesnym szacowaniem wielkości popełnianego błędu tego uogólniania.

W efekcie pandemii wiele obszarów działalności człowieka zostało przeniesionych do świata wirtualnego. Coraz większą rolę odgrywa cyberprzestrzeń, w której ludzkość stała się społeczeństwem uzależnionym od informatycznych środków przetwarzania danych. Zmiany spodziewane i te losowe, przypadkowe, można opisać za pomocą narzędzi statystycznych. W sposób niekwestionowany wzrasta zatem znaczenie metod statystycznych w analizach, a rozwój środków elektronicznego przetwarzania danych określa nowe możliwości. Opis niepewności wyprowadza się z zachowań przeszłych (danych dotyczących tych zachowań), co w konsekwencji pozwala na rozumienie zjawisk i budowanie prognoz ich przebiegu. Zasoby danych, informacje wyjściowe do analiz są ograniczone, należy więc liczyć się z ich złożonością, niedoskonałością, subiektywnymi kryteriami ocen.

Podstawę uzyskania informacji o przebiegu procesów masowych stanowią badania statystyczne. Instytucją powołaną w Polsce tylko do tego celu jest Główny Urząd Statystyczny, który publikuje wyniki swoich badań w rocznikach statystycznych i innych opracowaniach. W Unii Europejskiej jest to Eurostat. Badania statystyczne – ilościowe i jakościowe – nie są korelatywne, dostarczają bowiem informacji o odmiennym charakterze, na ogół nieporównywalnych. Są one jednak względem siebie komplementarne, uzupełniając wzajemnie wiedzę o badanej rzeczywistości. Zdarza się, że badania jakościowe stanowią wstępny etap badań ilościowych, identyfikując szczegółowe obszary do analiz. Analizy prowadzone metodami badań jakościowych przechodzą często w analizy wspomagane przez metody ilościowe bądź też następuje łączenie obu rodzajów analiz. W obu przypadkach badania statystyczne mają inny zakres oraz wykorzystywane są inne metody.

Celem niniejszej książki jest zachęcenie Czytelników do zdobywania umiejętności praktycznego stosowania metod statystycznych. Zaprezentowane metody mogą służyć nie tylko studentom, ale także badaczom do syntetycznego ujęcia analizowanych zjawisk i procesów zarówno ekonomicznych, jak i społecznych. Skomplikowane wzory i wywody teoretyczne zostały pominięte na korzyść przykładów liczbowych dotyczących różnych problemów. Liczne wskazówki ułatwią zdobycie wiedzy, a zestaw przykładów pomoże opanować sposoby interpretacji uzyskanych wyników. Z uwagi na pracochłonność obliczeń w przykładach i zadaniach występuje niewielka liczba obserwacji, co sprzyja osiągnięciu celów dydaktycznych. Zrezygnowano z pokazywania czasochłonnych technik obliczeniowych ze względu na dostępność komputerowych pakietów statystycznych i arkuszy kalkulacyjnych, zawierających programy obliczeniowe.

Rozdział pierwszy przedstawia podstawowe pojęcia i metody gromadzenia, opracowywania danych statystycznych, a także opis rodzajów badań. W rozdziale drugim zawarto metody analizy struktury zbiorowości z licznymi przykładami zastosowań. Natomiast rozdział trzeci zawiera wywody teoretyczne i przykłady dotyczące szeregów czasowych, rachunku indeksowego oraz analizy szeregów czasowych.. W kolejnym, czwartym rozdziale zaprezentowano metody i przykłady stosowania analizy korelacji i regresji zjawisk.

Autorki kierują podziękowania do Kierownictwa Akademii Handlowej Nauk Stosowanych w Radomiu za wsparcie w realizacji tego zamierzenia. Recenzentowi dziękujemy za istotne komentarze i sugestie. Podziękowania należą się również kolegom, przyjaciołom i studentom, którzy wykazali praktyczne podejście do statystyki, aktywnie współpracowali w zakresie badań i analiz statystycznych, zachęcając do napisania książki. Oddana do rąk Czytelników książka zawiera niezbędne elementy teorii z zakresu statystyki opisowej i dużą liczbę praktycznych przykładów wraz z opisaniem sposobów ich rozwiązywania oraz podaniem wyniku końcowego. Książka jest głęboko osadzona w rzeczywistym świecie. Starannie wybrane przykłady i problemy ilustrują zagadnienia z takich dziedzin, jak: transport, turystyka, finanse, zarządzanie, ekonomia, medycyna, produkcja, administracja publiczna, moda, reklama i inne. Wiele przykładów i zadań pochodzi z profesjonalnych publikacji, jak np.: Rocznik Statystyczny Rzeczypospolitej Polskiej, inne są efektem poszukiwań autorek. Celem było zaprezentowanie przykładów, które nie tylko zilustrują zagadnienia, ale również zainteresują Czytelnika. Autorki, na podstawie wieloletnich doświadczeń dydaktycznych, proponują do samodzielnego rozwiązania przykłady i zadania demonstrujące różne wątki zastosowań analiz ilościowych i jakościowych. Praktyczne przykłady są tak zróżnicowane, jak różnorodna i bogata jest otaczająca nas rzeczywistość.

Barbara Olbrych, Sylwia Rudecka-Gutkowska

ROZDZIAŁ 1
WPROWADZENIE DO BADAŃ STATYSTYCZNYCH

Badanie statystyczne jest to ogół prac mających na celu poznanie struktury określonej zbiorowości. Termin „statystyka" pochodzi od łacińskiego słowa „status", które oznacza państwo lub stan rzeczy. W piśmiennictwie określenie statystyka zostało użyte po raz pierwszy przez Gottfrieda Achenwala, profesora uniwersytetu w Marburgu i Getyndze w połowie XVIII wieku i oznaczało zbiór szeroko ujmowanych wiadomości o państwie. Współcześnie terminu statystyka używa się w kilku różnych znaczeniach:

- jako nazwy zbioru danych liczbowych (np. w tym znaczeniu mówi się o statystyce ludności, statystyce przemysłu itp.),
- jako nazwy czynności zbierania i opracowywania danych liczbowych (przykładowo: statystyka zatrudnienia polega na opracowaniu zestawień liczbowych, dotyczących poziomu zatrudnienia w określonym miejscu i czasie),
- jako nazwy pewnych charakterystyk liczbowych obliczanych ze zbiorowości próbnych, np. średnia arytmetyczna z próby czy odchylenie standardowe z próby,
- jako nazwy zbioru metod, które służą do badania prawidłowości w zbiorowościach (terminem statystyka w takim znaczeniu określa się dyscyplinę naukową, wykładaną na wyższych uczelniach).

W dalszej części rozważań przyjmuje się, że statystyka to nauka, która traktuje o ilościowych metodach badania procesów (zjawisk) masowych. Przedmiotem statystyki jako dyscypliny naukowej są zjawiska (procesy) masowe, występujące w otaczającej nas rzeczywistości. Tylko zdarzenia, powtarzające się wielokrotnie, tworzą tzw. procesy masowe, np. proces obsługi klientów w supermarketach, czy proces uprawy zbóż. Obserwacja pojedynczych zdarzeń, mających na ogół charakter losowy (przypadkowy), nie pozwala na formułowanie wniosków co do ich istoty. Dopiero analiza dużej liczby przypadków losowych ujawnia prawidłowości zachodzące pomiędzy zdarzeniami. Wiadomo z doświadczeń badawczych, że masowość występowania jednostek nie zawsze przesądza o konieczności stosowania metod statystycznych. Metody te można stosować tylko w przypadku zjawisk

masowych, w których poszczególne jednostki wykazują różnice indywidualne, a ujęte w masie –wykazują pewne prawidłowości. Przykładem mogą być plony pszenicy w gospodarstwach rolnych, które mogą zmieniać się zarówno w czasie, jak i w przestrzeni. Powodem tych zmian są czynniki, takie jak rodzaj gleby, odmiana ziarna użytego do zasiewu czy poziom nawożenia. Plony pszenicy w poszczególnych gospodarstwa rolnych różnią się, ale badane w określonym przedziale czasowym cechują się pewną prawidłowością, jak wzrost czy spadek, bądź mają niezmieniony poziom. A zatem, metody statystyczne nie mogą być stosowane, gdy zjawiska masowe składają się z identycznych jednostek. Pod pojęciem prawidłowości należy rozumieć zjawiska masowe, które wywołują dwojakiego rodzaju przyczyny:

- przyczyny główne (systematyczne) – oddziałują jednakowo lub podobnie na wszystkie jednostki zbiorowości, są właśnie powodem występowania określonych prawidłowości,
- przyczyny uboczne (przypadkowe) – działają na jednostki zbiorowości różnokierunkowo, powodując odchylenia od ogólnej tendencji rozwojowej badanego zjawiska.

Rezultaty oddziaływania przyczyn ubocznych mają tendencję do znoszenia się, a znoszą się tym dokładnie, im większa jest zbiorowość podlegająca badaniu. Wraz ze wzrostem liczebności zbiorowości dokładniej ujawnia się składnik systematyczny jako rezultat działania przyczyn głównych. Oznacza to wówczas, że działa prawo wielkich liczb. Przykładem może być liczba dzieci w rodzinach miejskich i wiejskich i stwierdzenie, że rodziny wiejskie są liczniejsze. Taką prawidłowość można zaobserwować jedynie wtedy, gdy badaniu podlega duża liczba rodzin. W pojedynczych przypadkach zbyt silnie działają przyczyny uboczne, które powodują odchylenia od składnika systematycznego.

1.1. Podstawowe pojęcia i definicje statystyki

Aby zrozumieć istotę i sposób realizacji badań statystycznych należy zapoznać się z podstawowymi pojęciami statystyki oraz rodzajem i organizacją badań. Przedmiotem badań statystycznych są określone zbiorowości osób, rzeczy lub zdarzeń.

Zbiorowość statystyczna (populacja generalna) jest to zbiór dowolnych elementów, nieidentycznych, stanowiących jedną, logiczną całość. Zbiorowość statystyczna może być skończona (1, 2, ..., n) lub nieskończona, ale przeliczalna (1, 2, 3,.....). Aby zbiór jednostek był zbiorowością

statystyczną powinien być jednocześnie identycznym ze względu na jedną cechę stałą oraz różnić się przynajmniej ze względu na jedną cechę zmienną. Elementy składowe badanej zbiorowości noszą nazwę jednostek statystycznych. Wybór jednostki statystycznej do badania zależy od celu badania. Przykładem może być badanie demograficzne, w którym jednostką statystyczną mogą być osoby, rodziny lub gospodarstwa domowe.

Typy zbiorowości statystycznych:

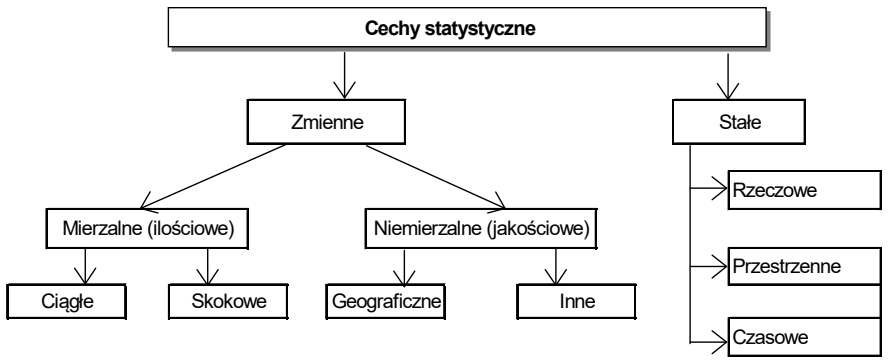
- skończenie liczna (składa się z ograniczonej, możliwej do przeliczenia liczby elemen*tów(np. liczba absolwentów szkół wyższych w danym roku akademickim),*
- *nieskończenie liczna,*
- statyczna (składa się z jednostek istniejących w ściśle określonym momencie (np. liczba pracowników marketu 20 grudnia 2021 r.),
- dynamiczna (obejmuje jednostki, które wystąpiły w pewnym przedziale czasowym, np. liczba mieszkań oddanych do użytku w I kwartale 2021 r.) zbiorowością mogą być złowione w danym dniu karasie),
- jednorodna (obejmuje jednostki o jednakowym typie, gatunku, rodzaju, a różnice między jednostkami mają charakter wyłącznie ilościowy, nie jakościowy, np. zbiorowością mogą być złowione w danym dniu karasie),
- niejednorodna (składa się z jakościowo odmiennych jednostek lub grup jednostek, a różnice między jednostkami mają charakter jakościowy).

Przestawiony podział jest umowny i względny. Ta sama zbiorowość w jednym badaniu może być jednorodna, w innym niejednorodna, np. robotnicy w firmie Y mają w 2020 roku tę samą płeć, ale różnią się stażem pracy. Jednostki statystyczne różnią się pewnymi właściwościami, które nazywają się cechami statystycznymi.

Cecha statystyczna jest to właściwość jednostek zbiorowości, która może służyć jako kryterium pozwalające odróżnić od siebie poszczególne jednostki zbiorowości. Wyróżnia się cechy statystyczne ***zmienne*** oraz ***stałe*****.**

Właściwości, które pozwalają na rozróżnianie poszczególnych jednostek zbiorowości są cechami zmiennymi. Cechy te dzielą się na ***ilościowe*****,** których wartości zmienne są mierzalne tzn. można im przypisać wartość liczbową oraz ***jakościowe*****,** których wartości zmienne (warianty) można opisać słownie lub za pomocą skal, np. numerycznych. Cechy ilościowe mogą być ***ciągłe****,* gdy przyjmują dowolne wartości z określonego przedziału liczbowego [a, b], lub ***skokowe*** (dyskretne), gdy przyjmują skończony lub przeliczalny zbiór wartości. Cechy jakościowe przedstawiane są w sposób opisowy (wielowartościowy) lub w układzie dwuwartościowym (zero-jedynkowym).

Właściwości wspólne dla wszystkich elementów zbiorowości statystycznej stanowią cechy stałe. Wśród stałych cech statystycznych wyróżnia się: cechy rzeczowe (przedmiotowe) określające przedmiot badania, cechy przestrzenne opisujące rozmieszczenie elementów zbiorowości statystycznej na terenie obszaru administracyjnego oraz czasowe określające, z jakiego okresu lub momentu pochodzą badane elementy zbiorowości statystycznej. Rysunek 1.1.1 ukazuje klasyfikację cech statystycznych.



Rys. 1.1. Klasyfikacja cech statystycznych
Źródło: Opracowanie własne

Cechy mierzalne są to właściwości jednostek statystycznych, które można wyrazić za pomocą liczb mianowanych (można je zmierzyć). Wyrażane są w różnych jednostkach, np. tony, zł, sztuki itp.; występują z różnym natężeniem w poszczególnych jednostkach zbiorowości, opisywane są w skali interwałowej, albo ilorazowej.

Cechy niemierzalne przedstawiają właściwości jednostek zbiorowości, których nie można zmierzyć, a jedynie stwierdzić, który wariant cechy występuje u badanej jednostki (np. płeć, zawód). Cechy te opisuje się w skali nominalnej lub porządkowej. Liczba wariantów cechy niemierzalnej jest zawsze skończona.

W tabeli 1.1. pokazano przykłady zbiorowości i zmiennych cech statystycznych.

Tab. 1.1. Przykłady zbiorowości i zmiennych cech statystycznych

Zbiorowość	Jednostka obserwacji	Mierzalna cecha skokowa	Mierzalna cecha ciągła	Niemierzalna cecha opisowa
Przedsiębiorstwa	Przedsiębiorstwo	Kooperanci	Zyski	Branża
Województwa	Województwo	Miasta	Powierzchnia	Region
Bezrobotni	Bezrobotny	Wiek	Zasiłek	Płeć
Urodzenia	Urodzenie	Kolejność	Waga	Szpital
Kraje	Kraj	Ludność	PKB	Rządy

Źródło: Opracowanie własne

Dla potrzeb pomiaru cech statystycznych stosuje się następujące rodzaje skal: nominalną, porządkową, interwałową oraz ilorazową.

Skala nominalna to skala wykorzystująca wyłącznie opis słowny w celu identyfikacji jednostki (np. kobieta, mężczyzna); wyklucza się prowadzenie działań arytmetycznych na danych opisanych na skali nominalnej.

Skala porządkowa służy do klasyfikowania danych (np. ranking z punktu widzenia wysokości wydatków gospodarstw domowych na żywność); podobnie jak w skali nominalnej nie można prowadzić działań arytmetycznych na danych w tej skali.

Skala interwałowa ma własności skali porządkowej, bo możliwe jest systematyzowanie jednostek statystycznych opisanych w tej skali, a jednocześnie możliwe jest określenie interwału (przedziału) liczbowego, który zawiera te obserwacje. Dane, które są opisane w tej skali są wyrażone w postaci wartości liczbowych (np. temperatura). Działania arytmetyczne, takie jak dodawanie i odejmowanie są możliwe, zaś zero przyjmuje się arbitralnie.

Skala ilorazowa ma cechy skali interwałowej, a ponadto iloraz ma tu określoną interpretację. W skali ilorazowej zwraca uwagę zero, które jest absolutne. Dane opisane w tej skali zawsze przyjmują wartości liczbowe (np. wzrost, waga);na danych można wykonywać działania arytmetyczne: dodawanie, odejmowanie, mnożenie, dzielenie.

Obserwacja statystyczna – proces zbierania danych statystycznych.

Próba – skończony lub nieskończony podzbiór elementów populacji, które zamierzamy poddać obserwacji empirycznej ze względu na badaną cechę.

Liczebność próby – liczba elementów populacji generalnej wybranych do próby. Z uwagi na liczebność próby dzieli się je na małe i duże. Próby małe mają nie więcej niż 30 elementów.

Próba losowa – próba wybrana z populacji w sposób losowy tzn. tak, że tylko przypadek decyduje o zaliczeniu elementu populacji generalnej do próby.

Próba reprezentatywna – próba losowa, której struktura (budowa) pod względem badanej cechy nie różni się istotnie od struktury zbiorowości generalnej ze względu na tę cechę. Próby reprezentatywne uzyskuje się za pomocą tzw. schematu losowania, czyli algorytmu losowania elementów do próby z populacji generalnej.

Wyniki z próby – zaobserwowane wartości badanej cechy statystycznej elementów próby losowej. Wyniki uzyskane z obserwacji dużych prób (zawierających powyżej 30 elementów) grupuje się najczęściej w klasy,

tworząc tym samym tzw. rozdzielcze szeregi statystyczne, które są głównym narzędziem opisu struktury zbiorowości statystycznej.

Materiał liczbowy, zebrany w wyniku przeprowadzonego badania statystycznego, należy odpowiednio zaprezentować. Temu celowi służą szeregi statystyczne.

Szereg statystyczny – zestawienie wartości zmiennych badanej cechy (wyników obserwacji), uporządkowanych według logicznego kryterium, z przyporządkowanymi im częstościami ich występowania.

Szereg statystyczny szczegółowy (prosty) – ciąg wartości badanej cechy statystycznej uporządkowany rosnąco lub malejąco. Taki sposób porządkowania informacji ma miejsce wtedy, gdy przedmiotem badania jest niewielka liczba jednostek.

Szereg rozdzielczy – zbiorowość statystyczna podzielona na części (klasy), według określonych wariantów cechy z podaniem liczebności wyodrębnionych klas.

Szereg strukturalny – zestawienie wyników w postaci szeregu rozdzielczego z cechą jakościową.

Rozkład empiryczny – zestawienie wyników w postaci szeregu rozdzielczego z cechą mierzalną. Wyróżnia się rozkład empiryczny ***punktowy (skokowy)***, gdy $n_i = f(x_i)$ oraz **przedziałowy**, gdy $n_i = f(x_{d_i}, x_{g_i})$, gdzie i oznacza liczbę grupowań wartości zmiennych badanej cechy statystycznej X.

Szereg statystyczny czasowy – ciąg wartości cechy statystycznej (wyników obserwacji) uporządkowanych w czasie. Wyróżnia się szeregi czasowe momentów, gdy cecha statystyczna przyjmuje wartości dla danego momentu, na przykład na koniec roku, oraz szeregi czasowe okresów, gdy cecha statystyczna przyjmuje wartości za pewien okres, na przykład miesięcznej produkcji.

Parametry – to charakterystyki liczbowe rozkładu zbiorowości według badanej cechy. Przy badaniu pełnym mówimy o parametrach populacji, przy badaniu częściowym – o parametrach z próby, zwanych też miernikami statystycznymi.

1.2. Rodzaje badań statystycznych

Badaniem statystycznym jest każdy zespół czynności związany ze zbieraniem, przetwarzaniem, analizowaniem i prezentacją danych liczbowych. Aby określone badanie można było uznać za statystyczne, musi

dotyczyć kształtowania się zmiennej (zmiennych) w tej zbiorowości. Ponadto badanie powinno spełniać warunki:

- dotyczyć zbiorowości (masy) statystycznej,
- określać prawidłowości charakteryzujące całą zbiorowość,
- prawidłowości muszą dotyczyć zmiennych (cech) występujących w tej zbiorowości.

Każde badanie statystyczne wymaga odpowiedniej metody. Wybór metody zależy od celu badania, rodzaju zbiorowości statystycznej, tematu badania, jego szczegółowości oraz środków jakimi dysponujemy (to są ludzie, środki finansowe, możliwości techniczne i materiałowe, inne).

Badania statystyczne mogą dotyczyć całej zbiorowości lub jej części (próby). Mogą być prowadzone w czasie rzeczywistym, ciągle lub okresowo, bądź też przebieg badanego procesu może być odtwarzany *post factum* (po pewnym czasie) na podstawie dokumentacji warsztatowej, księgowej itp. Informacje mogą być rejestrowane automatycznie przez odpowiednio skonstruowane czujniki i rejestratory albo zbierane przez osoby, tzw. obserwatorów przy wykorzystaniu ankiet lub innych dokumentów badawczych. Wynika stąd, że istnieje wiele metod prowadzenia badań statystycznych.

W zależności od przyjętego kryterium (płaszczyzny) podziału można wyróżnić następujące rodzaje badań:

- ***pełne*** (wyczerpujące) – obserwacji podlegają wszystkie jednostki (obiekty) populacji generalnej. Są to badania najbardziej wiarygodne, zarazem najdroższe i często czasochłonne, ale wyniki obserwacji stanowią charakterystykę badanej zbiorowości, przykłady: spisy, rejestracja bieżąca, sprawozdawczość,
- ***częściowe*** – gdy obserwacji podlega część jednostek zbiorowości, czyli próba losowa. Za pomocą tej metody można ograniczyć koszty badań, ale wyniki obserwacji charakteryzują tylko część elementów populacji i charakterystyki zbiorowości statystycznej można jedynie oszacować na poziomie określonego prawdopodobieństwa. Przykładowo metodę tę wykorzystuje się, gdy badany obiekt musi ulec zniszczeniu.

Badania częściowe dzielą się na:

- ✓ ***ankietowe*** – informacje o zbiorowości statystycznej zbierane są przy użyciu ankiet (dokumentów badawczych) wypełnianych przez informatorów: do ściśle określonych osób, instytucji lub przedsiębiorstw, na przykład ankiety rozsyłane do firm, lub osób,
- ✓ ***monograficzne***, polegające na szczegółowym, wieloaspektowym opisie i analizie wybranego elementu zbiorowości statystycznej; wybrana jed-

nostka powinna być typowa, powszechnie występująca lub wskazująca kierunek rozwoju, a więc przodująca,

- ✓ ***reprezentacyjne*** **–** badania prowadzone są na próbie reprezentacyjnej, która jest częściowym badaniem statystycznym, opartym na próbie pobranej ze zbiorowości generalnej w sposób losowy (z teoretycznego i praktycznego punktu widzenia metoda ta jest najbardziej prawidłową formą badania częściowego, bowiem zastosowanie rachunku prawdopodobieństw umożliwia przenoszenie wyników z próby losowej na całą zbiorowość statystyczną oraz określenie wielkości popełnionego błędu).

Należy pamiętać, że możliwości oszacowania błędu, popełnionego w trakcie badania, nie daje ani metoda ankietowa, ani monograficzna, a tylko metoda reprezentacyjna. W praktyce stosujemy głównie badania częściowe (reprezentacyjne, monograficzne, ankietowe).

Badania są prowadzone według pewnego schematu, tzw. planu badań wskazującego, które z wymienionych metod mogą występować łącznie, np.:

[badania pełne $\Rightarrow$ doraźne $\Rightarrow$ ankietowe].

Zarówno badania pełne, jak i częściowe mogą być:

- ✓ ***ciągłe*** (monitorowanie)– obserwacja obiektu prowadzona jest nieprzerwanie, np. badanie czasu pracy kierowców samochodów ciężarowych przy użyciu kart tachografu, ochrona banków za pomocą kamer telewizyjnych itp.,
- ✓ ***okresowe*** – obserwacje prowadzone są co pewien czas przez ściśle określony okres, np. wszelkiego rodzaju spisy,
- ✓ ***doraźne*** **–** badania podyktowane koniecznością poznania jakiegoś zjawiska, takiego jak wzrost umieralności, spadek popytu, opinia społeczna i inne.

W sytuacji, gdy nie chcemy albo nie możemy w sposób bezpośredni (tzn. na podstawie badania pełnego albo częściowego)uzyskać informacji o interesującej nas zbiorowości, wtedy możemy przeprowadzić postępowanie zwane szacunkiem statystycznym. W celu lepszego zrozumienia tego zagadnienia przytoczone zostaną określenia:

- ✓ ***szacunki statystyczne*** **–** gdy informacje o badanej zbiorowości ustala się na podstawie znanej zbiorowości, która ma z nią logiczny związek, np. zachowanie higieny można oceniać na podstawie popytu na mydła i środki piorące,
- ✓ ***szacunki interpolacyjne*** **–** nieznane wartości cechy statystycznej oceniane są na podstawie znanych wartości sąsiednich (wcześniejszych i późniejszych),

- ✓ ***szacunki ekstrapolacyjne*** – nieznane wartości cechy statystycznej oceniane są poza przedziałem wartości znanych.

Z punktu widzenia pozyskiwania źródeł informacji, w badaniach empirycznych wyróżniamy ***badania wtórne i pierwotne.*** W badaniach wtórnych wykorzystywane są dane zastane, czyli takie, które znajdują się w już istniejących zasobach. Badania pierwotne, zwane również terenowymi, mają na celu zebranie danych indywidualnych, które podlegają następnie dalszemu przetworzeniu w informację wynikową, pozwalającą zweryfikować hipotezy badawcze.

Dane zastane są to dane dostępne w różnych zasobach, tj. archiwach, bazach danych, opublikowane w opracowaniach naukowych raportujących wyniki zrealizowanych badań itp. Ich wspólną cechą jest to, że pochodzą z wcześniej zrealizowanych procesów badawczych. Dane wtórne natomiast to dane zebrane w innym celu niż aktualnie rozwiązywany problem badawczy. Dane ze źródeł wtórnych mają zawsze charakter historyczny, a ich wykorzystanie nie wymaga kontaktu z obiektem badania. Dane zastane od danych pierwotnych odróżnia wiele właściwości. Dotyczą one celu badania, sposobu realizacji, kosztów pozyskania informacji, czasu trwania badania oraz wiedzy o procesie badawczym i braku możliwości identyfikowania badanych obiektów (anonimowość).

Dane pierwotne to dane zebrane przez badacza specjalnie po to, by odpowiedzieć na określony problem badawczy. Utworzenie danych pierwotnych jest zawsze związane z przeprowadzeniem pomiarów w terenie, co wiąże się z dużymi nakładami pracy i czasu, związanymi z przygotowaniem narzędzi badawczych oraz stworzeniem zaplecza organizacyjnego, technicznego, a także finansowego prowadzonych działań.

Istotną odmiennością danych pierwotnych względem danych wtórnych jest ich różne ulokowanie na osi czasu, czyli moment powstania obu zasobów. Dane wtórne zawsze dotyczą przeszłości, dane pierwotne odnoszą się do obecnego okresu lub wręcz dotyczą czasu rzeczywistego.

1.3. Organizacja badań statystycznych

Jedną z pierwszych czynności, jaką należy wykonać przystępując do realizacji badań, od której zależy zarówno wielkość nakładów ponoszonych na badania, jak i wiarygodność uzyskiwanych wyników, jest optymalne zorganizowanie badań. Całość prac związanych z organizacją badań statystycznych składa się z kilku etapów, które obejmują:

- określenie celu badań,

- zdefiniowanie zbiorowości statystycznej i jednostki badawczej (przedmiotu i podmiotu badań),
- wybór metody badania,
- określenie źródła informacji (pierwotne lub wtórne),
- opracowanie formularzy statystycznych (dokumentów badawczych) oraz tablic wynikowych,
- opracowanie metody kontroli wiarygodności zbieranego materiału statystycznego.

Ogólnie celem badań statystycznych jest uzyskanie (zebranie) wiarygodnych danych przydatnych w podejmowaniu konkretnych decyzji. Precyzyjne określenie zbioru decyzji (obszaru decyzyjnego), któremu mają służyć zebrane informacje ma istotne znaczenie dla organizacji badań, ponieważ to ukierunkowuje (ogranicza lub rozszerza) zakres kolejnych prac. Cel badań przesądza o przedmiocie i podmiocie badań oraz wyznacza zbiór badanych cech statystycznych.

Przedmiot badań stanowi zbiorowość statystyczna, która może składać się z rzeczy, osób lub zdarzeń. Ze zdefiniowanej na podstawie celu badań zbiorowości statystycznej należy wybrać elementy, które będą podlegały badaniu, czyli będą podmiotem badań. Wyboru elementów do próby dokonuje się na drodze jednego z wymienionych poniżej schematów losowania:

- indywidualnego, gdy losujemy pojedyncze elementy z całej populacji generalnej,
- zespołowego, gdy wydzielamy grupy (zespoły) elementów, np. elementy grupy zawodowej do próby losujemy spośród tych zespołów,
- wielostopniowego, gdy populację dzielimy na grupy, a następnie podgrupy, z których losujemy elementy do próby, np. zakłady $\Rightarrow$ wydziały $\Rightarrow$ grupy zawodowe,
- ograniczonego, gdy losujemy z poszczególnych części populacji podzielonej na tzw. warstwy, będące podzbiorami populacji generalnej utworzonymi według przyjętego kryterium, którym np. w przypadku zbiorowości samochodów może być zużycie resursu eksploatacyjnego, mierzonego kilometrami przejechanymi przez samochody.

Przedstawione schematy losowania próby dotyczą skończonych, rzeczywistych zbiorowości statystycznych. Ponumerowane elementy populacji tworzą tzw. operat losowania, z którego elementy do próby wybiera się przy użyciu jednej z trzech technik losowania:

- wybór na chybił trafił,
- z wykorzystaniem tablic liczb losowych,
- z zastosowaniem generatorów liczb losowych.

Metoda badań wynika bezpośrednio z założonego celu badań, do którego należy dobrać odpowiedni plan badań, biorąc pod uwagę czas i koszty ich przeprowadzenia. Jeżeli celem badań jest szybkie uzyskanie danych o popularności podjętej przez rząd decyzji, należy wybrać metodę badań doraźnych, reprezentacyjnych, ankietowych, tzn. przeprowadzanych doraźnie na reprezentacyjnej próbie osób za pomocą ankiet.

Organizując badania należy również określić z jakiego źródła będą pochodzić zbierane dane. Czy będą to informacje zbierane podczas realizacji badanych procesów, czy też będą odtwarzane na podstawie istniejącej dokumentacji. W pierwszym przypadku mamy do czynienia ze źródłami pierwotnymi, do których zalicza się dane uzyskiwane na podstawie obserwacji bezpośredniej, wywiadu lub ankiety. W drugim przypadku dane pochodzą z wtórnych źródeł, takich jak sprawozdawczość firm, publikacje statystyczne, prace naukowo-badawcze itp.

Materiał statystyczny, uzyskany w wyniku obserwacji statystycznej, jest to zbiór szczegółowych informacji o właściwościach jednostek statystycznych. W celu sensownego wykorzystania zebranych danych niezbędne jest opracowanie, które obejmuje grupowanie i zliczanie. Grupowanie polega na wyodrębnieniu jednorodnych lub względnie jednorodnych części w ramach większej i zróżnicowanej zbiorowości statystycznej. Zadaniem grupowania jest przejście od informacji o właściwościach poszczególnych jednostek zbiorowości do informacji o całej zbiorowości lub wyodrębnionej jej części (próby). Przykładem częstego grupowania może być grupowanie według płci, wieku czy miejsca zamieszkania. Choć statystyka nie podaje sztywnych reguł, to jednak zaleca się korzystanie z dwóch kryteriów:

- jednostki zaliczane do tej samej grupy nie powinny być zbyt zróżnicowane z punktu widzenia badanej cechy,
- liczba grup nie powinna być zbyt duża.

W badaniach jednorazowych grupowanie materiału statystycznego przeprowadza się odmiennie, w zależności od wyznaczonego celu. W badaniach ciągłych i periodycznych, w których ważną rolę odgrywa problem porównywalności danych, należy dążyć do zastosowania jednolitych zasad grupowania. Jednolity system grupowania nazywa się klasyfikacją. Przykładem klasyfikacji może być podział gospodarki narodowej według działów, gałęzi itd.

Możemy podzielić materiał statystyczny na grupy według jednej cechy (grupowanie proste) albo według więcej niż jednej cechy (grupowanie złożone). Z punktu widzenia celu badania, wyróżniamy grupowanie typologiczne i wariancyjne. Przykładem grupowania typologicznego jest podział ludności czynnej zawodowo według grup społeczno-ekonomicznych (cecha

jakościowa). Podstawą grupowania wariancyjnego jest cecha ilościowa, przy czym warianty nie różnią się w sposób istotny. Jako przykład można tu podać podział zmarłych niemowląt według wieku (dni) na grupy: 0 dni, 1-6 dni, 7-13 dni, 14-20 dni, 21-27 dni.

Zbierane dane muszą być przez informatorów w wybranej formie zapamiętywane (zapisywane). Celowi temu służą dokumenty badawcze (formularze statystyczne). W dokumencie badawczym zapisywane są wartości badanych stałych i zmiennych cech statystycznych. Dokument badawczy powinien zawierać tylko te cechy statystyczne, które są niezbędne dla zidentyfikowania badanego obiektu oraz zapewniające zrealizowanie celu i zakresu badań. Nie należy zapisywać informacji, które mogą być wygenerowane (obliczone) w oparciu o inne zebrane informacje, np. gromadząc dane dotyczące ilości i ceny nie ma potrzeby zbierania informacji o wartości.

Zbierane informacje (materiał statystyczny) należy kontrolować zarówno w trakcie jego gromadzenia, jak i przed przystąpieniem do jego przetwarzania, ponieważ błędy popełnione na etapie zbierania danych rzutują w sposób nieodwracalny na wyniki badań. Dlatego organizacja badań obejmuje również opracowanie metody oceny wiarygodności zebranego materiału statystycznego. Metoda ta musi uwzględniać kontrolę formalną i merytoryczną. Kontrola formalna dotyczy przeważnie kompletności zebranego materiału statystycznego. Do oceny merytorycznej należy opracować komputerowe algorytmy wykorzystujące pewne specyficzne zależności pomiędzy wartościami poszczególnych cech statystycznych, pozwalające między innymi na wyeliminowanie błędów pomiaru oraz błędów systematycznych. Logika takich algorytmów wynika z istoty badanych zjawisk.

Materiał statystyczny może być zbierany przez obserwatorów lub rejestrowany automatycznie przy użyciu odpowiednich czujników lub kamer. Jest to obserwacja statystyczna. Zbieranie informacji przez obserwatorów polega na ich wpisywaniu do przygotowanych na etapie organizacji badań dokumentów badawczych.

Obserwatorem może być odpowiednio przeszkolona osoba, np. rachmistrz spisowy lub pracownik firmy wypełniający dokumenty badawcze w ramach dodatkowego zajęcia. Automatycznie informacje statystyczne zapisywane są na papierowych nośnikach informacji, takich jak tarcze tachografów samochodowych, na taśmach magnetycznych lub komputerowych nośnikach informacji. Informacje mogą pochodzić ze źródeł wtórnych lub pierwotnych.

Najważniejszym problemem występującym na etapie zbierania informacji jest uniknięcie błędnego zidentyfikowania i zapisu wartości cech

statystycznych. Błędy te są trudne, a czasem wręcz niemożliwe do zauważenia podczas kontroli wiarygodności wyników badań, co w istotny sposób rzutuje na wyniki analizy statystycznej.

Uzyskany z badań materiał statystyczny zawiera szczegółowe dane liczbowe o wartościach cech badanych obiektów (jednostek obserwacji). Jeżeli badamy jedną zbiorowość N o przeliczalnej liczbie n obiektów z uwagi na jedną cechę statystyczną, czyli prowadzimy badania w przestrzeni jednowymiarowej, wyniki obserwacji (warianty zmiennej cechy statystycznej) x_i tworzą tzw. wektor obserwacji o liczbie obserwacji równej n.

$$[N] = [x_1, x_2, ..., x_i, ..., x_n]$$

W przypadku, gdy dla jednej zbiorowości *N* o przeliczalnej liczbie jednostek prowadzimy badania kilku cech statystycznych (badania w przestrzeni wielowymiarowej), warianty cechy zmiennej x_{ij} mają postać tzw. macierzy obserwacji o rozmiarach $w \times n$:

$$(\mathrm{N}) \Rightarrow \begin{bmatrix} x_{11}, x_{12}, ..., x_{1n} \\ x_{21}, x_{22}, ..., x_{2n} \\ .... \\ x_{w1}, x_{w2}, ..., x_{wn} \end{bmatrix}$$

gdzie:
n – liczba elementów zbiorowości ($i = 1,2,...,$n),
w – liczba badanych cech statystycznych ($j = 1,2,...,w$).

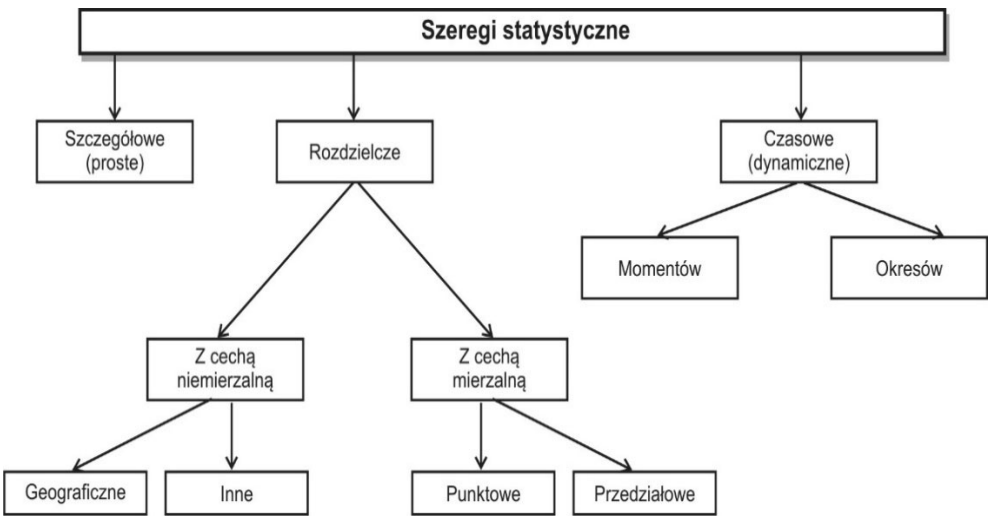
Jeżeli badamy kilka zbiorowości statystycznych N_i; $i = 1, 2, ...d$ o przeliczalnej liczbie jednostek ze względu na jedną cechę statystyczną, warianty cechy zmiennej można przedstawić w następującej postaci:

$$\begin{bmatrix} N_1 \\ N_2 \\ \vdots \\ N_d \end{bmatrix} \Rightarrow \begin{bmatrix} x_{11}, x_{12}, ..., x_{1n_1} \\ x_{21}, x_{22}, ..., x_{2n_2} \\ ..... \\ x_{d1}, x_{d2}, ..., x_{dn_d} \end{bmatrix}$$

Uporządkowane (zidentyfikowane) w powyższy sposób wyniki obserwacji, głównie obrazujące liczebność materiału statystycznego, nie mają większego praktycznego znaczenia. W celu wykorzystania wyników badań do analizy struktury zbiorowości zebrany materiał statystyczny należy usystematyzować, pogrupować i przedstawić w postaci szeregów statystycznych.

Szeregiem statystycznym nazywamy zbiór wyników obserwacji jednostek według pewnej cechy. Jednostkowe wartości cechy, zapisane według kolejności badania jednostek, tworzą nieuporządkowany zbiór. Te same wartości, ale uporządkowane w określony sposób (np. rosnąco lub malejąco), stanowią uporządkowany szereg statystyczny. Klasyfikację szeregów statystycznych przedstawia rysunek 1.2.

Szereg szczegółowy jest to materiał statystyczny uporządkowany wyłącznie według wartości badanej cechy. Można porządkować rosnąco lub malejąco. W pierwszej kolejności należy, w oparciu o stałe cechy statystyczne (przestrzenne, rzeczowe i czasowe), wyróżnić grupy wyników obserwacji różniących się jakościowo. Następnie do wyróżnionych grup wyników należy przyporządkować liczebności wartości zarejestrowanych zmiennych cech statystycznych.



Rys. 1.2. Klasyfikacja szeregów statystycznych
Źródło: Opracowanie własne

Szereg rozdzielczy powstaje na bazie wszystkich informacji o badanej zbiorowości, które to są przydzielane do określonych przedziałów wartości wyróżnionej cechy. W ten sposób tworzy się kolejną kategorię zwaną liczebnością n_i elementów danego przedziału klasowego. Przystępując do budowy szeregu rozdzielczego, należy zadeklarować tzw. liczbę przedziałów klasowych k, ich rozpiętość oraz określić sposób ograniczenia przedziałów.

Szeregiem czasowym nazywamy uporządkowany względem czasu ciąg wartości cechy (np. wartość kapitału, cena na określone dobro, stopa procentowa itd.). Wartości cechy są zarejestrowane w pewnych momentach lub okresach (miesięcznych, kwartalnych, rocznych). Stąd dalszy podział szeregów czasowych na szeregi czasowe momentów i szeregi czasowe okresów.

Przykład 1.3.1

Zmierzono wagę dziesięciu sportowców i uzyskano następujący zbiór informacji w kg: 66, 90, 84, 100, 57, 70, 85, 86, 85, 77. Z uzyskanych danych zbudować szereg szczegółowy.

Rozwiązanie

Uporządkowany rosnąco (lub malejąco), ale nie pogrupowany materiał statystyczny tworzy szereg szczegółowy. W tym przypadku dane liczbowe uporządkowano rosnąco.

Cecha statystyczna	Wartości zmienne cechy statystycznej (kg)
Waga sportowców	57,66,70,77,84,85,85,86,90,100

Źródło: Opracowanie własne

Przykład 1.3.2

W pewnym kurorcie górskim odnotowano roczne wielkości liczby przebywających tam kuracjuszy w okresie 2013-2021. Jak nazywa się przedstawiony poniżej szereg?

Lata	2013	2014	2015	2016	2017	2018	2019	2020	2021
Liczba kuracjuszy w tys.	7	8	9	10	13	12	8	8	9

Źródło: Opracowanie własne

Rozwiązanie

W tabeli przedstawione są informacje dotyczące skali pewnego zjawiska(liczba kuracjuszy), które badano w pewnym okresie(dziewięć lat). Jest to przykład szeregu czasowego. Ponieważ liczba kuracjuszy w poszczególnych latach jest sumą roczną, mamy więc do czynienia z szeregiem czasowym okresów.

Przykład 1.3.3

W skład Sejmu Rzeczypospolitej Polskiej VII kadencji wchodziło 460 posłów, wśród których 207 pochodziło z Platformy Obywatelskiej, 157 z Prawa i Sprawiedliwości, 40 z Ruchu Palikota, 28 z Polskiego Stronnictwa Ludowego, 27 z Sojuszu Lewicy Demokratycznej oraz 1 z Mniejszości Niemieckiej. Zbudować szereg punktowy.

Rozwiązanie

Rozwiązaniem jest szereg strukturalny, który powstał w wyniku grupowania według cechy jakościowej.

Cecha statystyczna	Warianty cechy statystycznej	Liczba posłów
Rodzaj komitetu wyborczego	PO	207
	PiS	157
	Ruch Palikota	40
	PSL	28
	SLD	27
	Mniejszość Niemiecka	1

Źródło: Mały Rocznik Statystyczny 2014, ss. 63-64

Dla ilościowych cech statystycznych wyniki obserwacji można przedstawiać w postaci szeregów rozdzielczych przedziałowych i punktowych. Szeregi rozdzielcze przedziałowe buduje się dla cech mierzalnych oraz skokowych z dużą liczbą wariantów cechy. W tym celu należy utworzyć klasy (przedziały) wartości wariantów cechy i przypisać należące do tych klas liczebności wyników obserwacji. Cechy statystyczne ustala się (wyróżnia) na etapie organizacji badań. Szereg rozdzielczy jest podstawowym narzędziem analizy rozkładu cechy.

Aby zbudować szereg rozdzielczy przedziałowy należy określić liczbę klas oraz rozpiętość przedziałów klasowych. Istnieją różne sposoby ustalania klas oraz ich rozpiętości, przy czym wybór powinien być uzależniony od liczby obserwacji n i od zmienności cechy. Liczbę klas k można ustalić z zależności:

$$k = \sqrt{n}$$

Rozpiętość h przedziału klasowego wynosi:

$$h = \frac{R}{k-1} = \frac{x_{max} - x_{min}}{k-1}$$

gdzie:

R – przedział zmienności cechy.

Klasy ustalamy tak, aby objąć wszystkie informacje oraz zagwarantować ich rozłączność. Lewostronną wartość przedziału klasowego x_1 wyznacza się z zależności:

$$x_1 = x_{min} - \frac{1}{2}h$$

Przyjmując lewostronnie domknięte przedziały klasowe i przyporządkowując obserwacje określonym przedziałom oraz zliczając je w każdej klasie otrzymujemy szereg rozdzielczy. Przy budowie szeregów rozdzielczych przedziałowych nie ma jednoznacznie określonych zasad postępowania w ustaleniu liczby klas i rozpiętości przedziałów klasowych. Nie ma też reguł rozstrzygających, czy przedziały mają być otwarte, czy zamknięte. Należy jednak pamiętać, aby liczebności w poszczególnych klasach nie odbiegały zbytnio od siebie, a także przestrzegać, by rozkład empiryczny charakteryzował się jednym maksimum.

Przykład 1.3.4

Miesięczne wydatki na jedzenie w grupie 35 studentów (w zł) są następujące:

405	420	411	427	480	440	378
468	437	452	421	414	402	422
462	428	431	414	437	405	390
425	425	400	432	447	385	419
400	425	458	439	360	405	369

Na podstawie powyższych danych zbudować odpowiedni szereg.

Rozwiązanie

Należy zbudować szereg rozdzielczy przedziałowy. W tym celu musimy określić liczbę klas oraz rozpiętość przedziałów klasowych.

Liczba klas:

$$k = \sqrt{n} = \sqrt{35} \cong 6$$

Rozpiętość przedziału (klasy):

$$h = \frac{x_{max} - x_{min}}{k - 1} = \frac{480 - 360}{6 - 1} = \frac{120}{5} = 24$$

Dolna granica przedziału:

$$x_1 = x_{min} - \frac{1}{2}h = 360 - \frac{1}{2} \cdot 24 = 360 - 12 = 348$$

Przyjmując lewostronnie domknięte przedziały klasowe oraz przyporządkowując obserwacje określonym przedziałom, a następnie zliczając je w każdej klasie otrzymamy szereg rozdzielczy.

Cecha statystyczna	Numer klasy i	Warianty cechy statystycznej (w zł) $< x_{d_i}, x_{g_i})$	Liczba osób n_i
Miesięczne wydatki na jedzenie	1	<348, 372)	2
	2	<372, 396)	3
	3	<396,420)	10
	4	<420,444)	14
	5	<444,468)	4
	6	<468,492)	2
	Razem		35

Źródło: Opracowanie własne

W przypadku zmiennych cech statystycznych ilościowych, przyjmujących wartości skokowe, można również utworzyć punktowy szereg rozdzielczy pod warunkiem, że liczba wariantów cechy nie jest duża. Szereg ten buduje się w ten sposób, że poszczególnym wartościom cechy przyporządkowuje się liczebności ich występowania. Liczba klas równa jest w tym przypadku liczbie wariantów przyjmowanych przez cechę statystyczną.

Przykład 1.3.5

Badano liczbę błędów w kodzie źródłowym 25 programistów. Otrzymano wyniki: 3, 2, 1, 3, 4, 5, 3, 1, 0, 2, 6, 3, 4, 5, 3, 1, 5, 3, 0, 1, 2, 2, 4, 3, 4. Zbudować szereg.

Rozwiązanie

Mamy do czynienia z cechą ilościową skokową, możemy skonstruować szereg rozdzielczy punktowy.

Cecha statystyczna	Numer klasy i	Warianty cechy statystycznej x_i	Liczba osób n_i
Liczba błędów	1	0	2
	2	1	4
	3	2	4
	4	3	7
	5	4	4
	6	5	3
	7	6	1
	Razem		25

Źródło: Opracowanie własne

Na uwagę zasługują szeregi rozdzielcze zbudowane na podstawie cechy jakościowej, jaką jest rozmieszczenie przestrzenne. Są to szeregi geograficzne.

Przykład 1.3.6
Rozmieszczenie Domów Pomocy Społecznej (DPS) w miastach na prawach powiatu w województwie mazowieckim.

Cecha statystyczna	Warianty cechy statystycznej	Liczba DPS-ów
Miasta	Ostrołęka	2
	Płock	3
	Radom	4
	Siedlce	1

Źródło: Opracowanie własne na podstawie rejestru DPS województwa mazowieckiego – stan na dzień 22.02.2021

1.4. Wskaźniki struktury, natężenia i podobieństwa struktur

Najprostszą formą opisu sposobu lokalizacji jednostek zbiorowości statystycznej ze względu na wartości zmienne badanej cechy są wskaźniki struktury. Są one szczególnie przydatne przy porównywaniu dwóch lub więcej struktur o różnej liczebności badanej zbiorowości.

Wskaźnikiem struktury (frakcją, odsetkiem) występowania danego wariantu cechy nazywa się stosunek liczby jednostek o danej wartości cechy do liczebności próby. Jest to zatem stosunek części zbiorowości do całej zbiorowości. Wskaźnik struktury określa zatem częstość (liczebność względną), z jaką dany wariant cechy zmiennej występuje w ogólnej liczbie wszystkich wartości cechy.

Wartość wskaźnika struktury oblicza się z zależności:

$$w_i = \frac{n_i}{n} \cdot 100\%, \qquad i = 1, 2, \ldots, k$$

gdzie:

n_i – liczebność wariantu cechy,

n – ogólna liczba obserwacji,

k – liczba wariantów cechy,

przy czym:

$$\sum_{i=1}^{k} w_i = 1.$$

$$0 \le w_i \le 1$$

Sumując (kumulując) częstości względne uzyskuje się skumulowane wartości wskaźnika struktury dla wartości cechy x mniejszej od danej wartości cechy x_i:

$$w_{sk} = \sum_{x<x_i} w_i$$

Skumulowana wartość wskaźnika struktury wskazuje, jak często wartość obserwowanej cechy kształtuje się poniżej pewnego poziomu.

W celu otrzymania skumulowanej liczebności przekształca się szereg rozdzielczy w kumulacyjny rozkład liczebności przez sumowanie liczebności kolejnych wariantów cechy. Skumulowane wartości liczebności lub częstości wyznaczają wartości charakterystyki zbiorowości statystycznej nazywanej dystrybuantą.

Dystrybuantą empiryczną nazywamy przyporządkowanie kolejnym wartościom cechy statystycznej (zmiennej) odpowiadających im częstości skumulowanych (względnie liczebności skumulowanych).

$$F(x) = \sum_{x<x_i} w_i \qquad \text{lub} \qquad F(x) = \sum_{x<x_i} n_i$$

Inaczej można powiedzieć, że dystrybuanta empiryczna określa częstości względne (liczebności) dla wszystkich wariantów cechy mniejszych od przyjętej wartości x_i.

Przykład 1.4.1

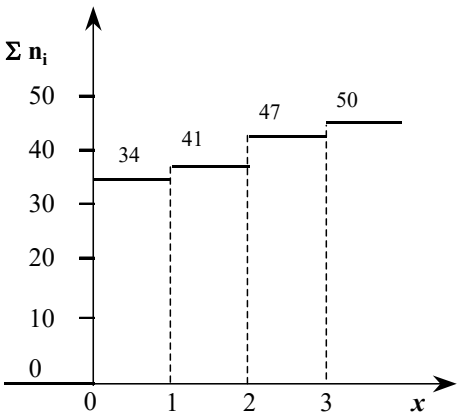
W magazynie pewnego przedsiębiorstwa wylosowano 50 sztuk wyrobów i oceniono liczbę usterek. W 34 wyrobach nie stwierdzono usterek, w 7 stwierdzono 1 usterkę, w 6 wyrobach 2 usterki, a 3 w pozostałych. Podać dystrybuantę empiryczną rozkładu.

Rozwiązanie

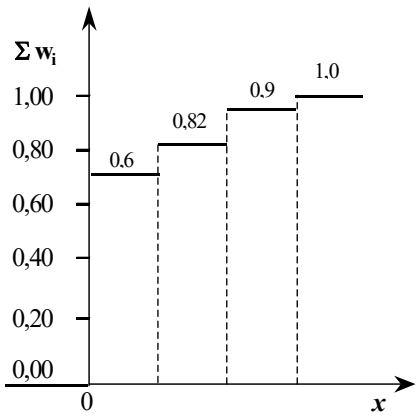
Badaną cechą jest liczba usterek. Jest to cecha mierzalna, skokowa. Należy zbudować szereg rozdzielczy punktowy liczby usterek, obliczyć wskaźniki struktury, a następnie dokonać kumulacji.

Warianty cechy statystycznej (liczba usterek) x_i	Numer klasy (przedziału) i	Liczebność wariantu cechy (liczba wyrobów) n_i	Wskaźnik struktury w_i	Dystrybuanta empiryczna liczebności $\sum_{x<x_i} n_i$	Dystrybuanta empiryczna częstości $\sum_{x<x_i} w_i$
0	1	34	0,68	34	0,68
1	2	7	0,14	41	0,82
2	3	6	0,12	47	0,94
3	4	3	0,06	50	1,00

Źródło: Opracowanie własne



Rys. 1.3. Wykres dystrybuanty liczebności



Rys. 1.4. Wykres dystrybuanty częstości

W przypadku, gdy uzyskane z badań wartości cechy statystycznej wykazują dużą koncentrację wartości w jednej grupie, wtedy należy stosować różne rozpiętości przedziałów klasowych. Aby ocenić strukturę zbiorowości przy różnych rozpiętościach przedziałów klasowych, konieczne jest stosowanie zamiast liczebności (częstości) wskaźnika zwanego gęstością liczebności (gęstością częstości).

Gęstość liczebności f_{n_i} (częstości f_{w_i}) jest to stosunek liczebności (częstości) danej klasy do rozpiętości przedziału klasowego. Wartości tych wskaźników oblicza się z zależności:

– gęstość liczebności:

$$f_{n_i} = \frac{n_i}{h_i}, \qquad i = 1, 2, \ldots, k$$

– gęstość częstości:

$$f_{w_i} = \frac{w_i}{h_i}, \qquad i = 1, 2, \ldots, k$$

gdzie:

k – liczba przedziałów klasowych,

h_i – rozpiętość i-tego przedziału klasowego.

Jeżeli przedmiotem badań są zmiany zjawiska w czasie, to wyniki obserwacji przedstawia się w postaci szeregów czasowych.

Przykład 1.4.2

Badając liczbę osób przebywających na zwolnieniu lekarskim w pewnej firmie otrzymano w kolejnych miesiącach 2020 roku wyniki: 7, 5, 8, 5, 7, 9, 12, 6, 7, 3, 4, 9. Zbudować szereg czasowy.

Rozwiązanie

Miesiąc t	I	II	III	IV	V	VI	VII	VIII	IX	X	XI	XII
Liczba osób na zwolnieniu	7	5	8	5	7	9	12	6	7	3	4	9

Źródło: Opracowanie własne

Wyniki obserwacji mogą być również prezentowane w postaci wskaźników natężenia, które znajdują duże zastosowanie w analizach społeczno-gospodarczych.

Wskaźniki natężenia są to wielkości stosunkowe, wyrażające kształtowanie się jednego zjawiska na tle innego, logicznie z nim powiązanego, np.:

- gęstość zaludnienia (liczba ludności przypadająca na 1 km^2 powierzchni),
- stopa bezrobocia (stosunek liczby bezrobotnych do liczby ludności czynnej zawodowo),
- wskaźnik rozwoju gospodarczego (produkt krajowy brutto/netto do liczby ludności),
- wskaźnik rentowności (zysk do wielkości sprzedaży),
- wskaźnik efektywności (zysk do zaangażowanego kapitału).

Często zdarza się, że ocena rozmiarów badanego zjawiska uwarunkowana jest wcześniejszym obliczeniem odpowiedniego wskaźnika natężenia.

Pomiar podobieństwa struktur może być określony w sposób syntetyczny za pomocą tzw. wskaźnika podobieństwa struktur[1], który uogólniono dla potrzeb niniejszej pracy, przyjmując, że $2 \leq n < N$:

$$\omega_p = \sum_{i=1}^{k} min(\omega_{1i}, \omega_{2i}) \text{ przy czym } 0 < \omega_p \leq 1.$$

Im ω_i jest bliższe jedności, tym struktury badanych zbiorowości są bardziej podobne.

Jak już wiadomo, materiał statystyczny grupuje się w postaci szeregów strukturalnych także dla cech niemierzalnych.

Przykład 1.4.3

Przeprowadzono badanie dotyczące gospodarowania budżetem domowym w wybranych losowo 2600 gospodarstwach domowych. Otrzymano wyniki (p. tab. 1.2).

Tabela 1.2. Sposób gospodarowania budżetem domowym

Wyszczególnienie	Częstość odpowiedzi (%)			
	ogółem	miasto	wieś	minimum $(\omega_{1i}, \omega_{2i})$
1. Pieniędzy wystarczy bez oszczędzania	4,1	3,9	4,3	3,9
2. Żyją oszczędnie, wystarcza na wszystko	24,6	24,5	25,1	24,5
3. Żyją bardzo oszczędnie, by odłożyć na ważniejsze zakupy	44,2	44,9	42,2	42,2
4. Pieniędzy wystarcza tylko na tanie jedzenie i ubranie	18,4	18,3	18,9	18,3
5. Pieniędzy wystarcza tylko na najtańsze jedzenie	7,2	7,2	7,1	7,1
6. Pieniędzy nie wystarcza nawet na najtańsze jedzenie i ubranie	1,5	1,2	2,4	1,2

Źródło: Dane umowne

Wariantami cechy niemierzalnej są sposoby gospodarowania budżetem gospodarstw domowym. Badanie nie wykazało dużych różnic między odczuciami gospodarstw miejskich i wiejskich. Najwyższą wartość wskaźnika struktury uzyskała odpowiedź trzecia, natomiast odsetek osób żyjących w biedzie w tej zbiorowości był większy na wsi, niż w mieście. Wskaźnik

[1] T. Słaby (red.), *Konsumpcja elit ekonomicznych w Polsce – ujęcie empiryczne*, SGH, Warszawa 2006, s. 21.

podobieństwa struktur ω_p był bardzo wysoki i wynosił w tym przypadku $\omega_p = 0{,}972$.

Przykład 1.4.4

Strukturę gospodarstw domowych według liczby osób w tym gospodarstwie i miejsca zamieszkania przedstawia tabela 1.3. Z przedstawionych danych[2] wynika, że cechą statystyczną jest tu liczba osób w gospodarstwie domowym. Cechę tę analizujemy w dwóch zbiorowościach, z których pierwsza dotyczy gospodarstw domowych w mieście, a druga na wsi. Można także rozpatrywać strukturę gospodarstw domowych w Polsce, wtedy zbiorowością będzie liczba gospodarstw ogółem. Wyniki obliczeń zostały zamieszczone w tabeli 1.4.

Tabela 1.3. Gospodarstwa domowe według liczby osób w gospodarstwie domowym

Liczba osób w gospodarstwie	Liczba osób w tysiącach		
	miasto	wieś	ogółem
x_i	n_{1i}	n_{2i}	n_{oi}
1	1594	594	2181
2	1852	821	2673
3	1724	703	2427
4	1821	811	2632
5	611	560	1171
6	180	334	514
7 i więcej	82	283	365
Razem	7864	4106	11970

Źródło: S. Ostasiewicz, Z. Rusnak, U. Siedlecka, *Statystyka*, Wyd. AEwe Wrocławiu, Wrocław 1995, s. 28

Z uwagi na dużą liczbę gospodarstw domowych w mieście i na wsi, dla porównania struktury według miejsca zamieszkania należy obliczyć wskaźniki struktury według wzoru:

$$w_i = \frac{n_i}{n}$$

[2] S. Ostasiewicz, Z. Rusnak, U. Siedlecka, *Statystyka. Elementy teorii i zadania,* Wyd. AE Wrocław, Wrocław 1995, s. 28.

Tabela 1.4. Wskaźniki struktury gospodarstw domowych ze względu na liczbę osób w gospodarstwie

Liczba osób w gospodarstwie	Wskaźniki struktury		Minimum(ω_{1i}, ω_{2i})
x_i	n_{1i}	n_{2i}	n_{oi}
1	0,202	0,145	0,145
2	**0,236**	**0,200**	0,200
3	0,219	0,171	0,171
4	0,232	0,198	0,198
5	0,078	0,136	0,078
6	0,023	0,081	0,023
7 i więcej	0,010	0,069	0,010
Razem	1	1	0,825

Źródło: S. Ostasiewicz, ibidem

Miarą podobieństwa struktur jest wskaźnik podobieństwa struktur $\omega_{p,}$ który po obliczeniach wynosi 0,825. Wysoka wartość bliska jedności świadczy o dużym podobieństwie struktur. Zarówno w mieście, jak też na wsi największy był odsetek gospodarstw dwuosobowych (odpowiednio 23,6%, 20%). Aż 89% gospodarstw miejskich i 71,4% gospodarstw wiejskich to gospodarstwa nie większe niż czteroosobowe.

1.5. Formy prezentacji wyników obserwacji statystycznej

Podstawową formą prezentacji rezultatów obserwacji statystycznej są tablice statystyczne. Przystępując do projektowania wzorów tablic wynikowych należy zapoznać się z wymaganiami odbiorców wyników oraz ich przeznaczeniem. Należy brać również pod uwagę sposób opracowywania materiału statystycznego. Jeżeli materiał statystyczny jest obszerny i będzie opracowywany w ośrodkach informatycznych, wówczas wzory tablic wynikowych mają istotny wpływ na założenia systemu przetwarzania tego materiału. Ponadto tablice wynikowe powinny być tak zaprojektowane, aby zawierały maksymalną ilość informacji niezbędnych dla zrealizowania celu badań, przedstawionych w zawartej i syntetycznej formie. Przy projektowaniu wzorów tablic wynikowych należy przestrzegać następujących elementarnych zasad :

- z tablicy musi wynikać dokładnie, jaki jest przedmiot badania, tzn. jakie jednostki badania zostały zaliczone do danej populacji,
- tytuł tablicy powinien określać cechy, według których badano poszczególne jednostki populacji,

- tablica musi zawierać informacje jednego momentu lub okresu, którego badanie dotyczy,
- z tytułu lub z treści boczku lub główki tablicy musi wynikać, jaki zakres terytorialny populacji prezentowany jest w tablicy,
- w jakich jednostkach miary podawane są poszczególne wielkości występujące w tablicy,
- jakie jest źródło danych, na podstawie których została opracowana tablica.

Tablice zawierają opis liczbowy badanych zbiorowości statystycznych według jednej lub kilku cech. Oprócz części liczbowej tablice zawierają część opisową, na którą składają się: tytuł tablicy, nazwy wierszy (tzw. boczek), nazwy kolumn (tzw. główka), źródło danych oraz ewentualne uwagi odnoszące się do przedstawionych liczb.

W tablicach używa się pewnych znaków umownych:

- kreska (–) oznacza, że dane zjawisko nie występuje,
- zero (0) oznacza, że dane zjawisko występuje, ale w ilościach mniejszych niż pół jednostki miary przyjętej w tablicy,
- kropka (.) oznacza brak informacji,
- znak (x) oznacza, że danej rubryki nie można wypełnić,
- „w tym" oznacza, że nie podaje się wszystkich składników sumy ogólnej.

Materiał statystyczny można zaprezentować graficznie za pomocą wykresów. Wykresy, chociaż najczęściej nie podają wszystkich wartości liczbowych badanej cechy statystycznej, pozwalają szybko wizualnie ocenić zmiany zachodzące w badanym procesie. Są bardziej sugestywne i lepiej przemawiają do odbiorcy od zestawień tabelarycznych. Wykresy stanowią często uzupełnienie tabel wynikowych.

Do najczęściej stosowanych w statystyce graficznych form prezentacji wyników obserwacji należą wykresy: punktowe, słupkowe (histogramy), liniowe (diagramy), obrazkowe, powierzchniowe, mapowe. Wybór jest uzależniony od rodzaju prezentowanego materiału statystycznego, a także od charakteru prawidłowości, jakie wykres ma pokazywać.

Aby wykres był komunikatywny dla czytelnika musi zawierać opis czego dotyczy, objaśnienia użytych symboli, kolorów i skali oraz precyzować źródło danych.

Przykład 1.5.1

W poniższej tablicy statystycznej zastosowano znak umowny kropka (.) i znak (×).

Lata / Absolwenci szkół w tys.	1946	1950	1960	1970	1980	1990	2000
podstawowych	79,7	270	392	660	508	592	1214
zasadniczych zawodowych	.	64,3	35,6	164	252	231	192
liceów ogólnokształcących	9,7	23,7	30,2	6,4	92,4	87,0	178
techników	6,0	24,2	19,5	94,9	121	106	189
Artystycznych ogólnokształcących	.	.	.	.	.	.	2,0
policealnych	×	×	9,5	39,5	59,8	39,5	86,4

Źródło: Rocznik Statystyczny Rzeczypospolitej Polskiej 2019, s. 44

Na wykresach sporządzonych w układzie współrzędnych kartezjańskich oś X(odciętych) prezentuje najczęściej wartości cechy, a oś Y(rzędnych) liczebności lub częstości występowania wariantów cechy. W przypadku prezentacji szeregów czasowych na osi odciętych zaznacza się jednostki czasu t.

Poniżej zaprezentowane zostaną przykłady wybranych, często stosowanych w praktyce wykresów, sporządzonych w oparciu o arkusz kalkulacyjny Excel.

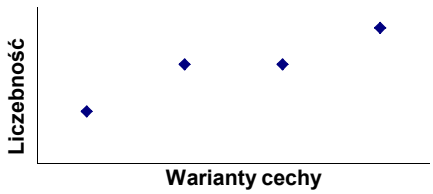
Wykres punktowy tworzy zbiór punktów, które dla szeregów szczegółowych prezentują jedną obserwację, natomiast dla szeregów przedziałowych odpowiadają liczbie obserwacji mających ten sam wariant cechy lub należących do tego samego przedziału wariantu cechy. Wykres w postaci histogramu (*wykres słupkowy*) składa się z prostokątów, których podstawy znajdują się na osi odciętych i odzwierciedlają rozpiętości przedziałów klasowych, natomiast wysokości są odkładane na osi rzędnych i przedstawiają liczebności lub częstości należące do tych przedziałów.

Wykresy liniowe (diagramy, wieloboki liczebności) otrzymuje się przez połączenie punktów, których współrzędnymi są środki przedziałów klasowych z odpowiadającymi tym przedziałom: liczebnością, częstością lub gęstością.

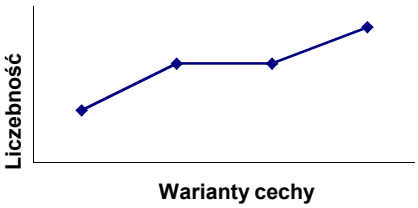
Na wykresach obrazkowych zbiorowość statystyczna przedstawiana jest w postaci symboli (obrazków) różniących się wielkością, kolorem lub liczbą. Wykresy te mają praktyczne zastosowanie przy prezentacji szeregów strukturalnych.

Wykresy powierzchniowe prezentują wyniki obserwacji statystycznej w postaci płaskich figur geometrycznych (kół, prostokątów, trójkątów) i mogą być stosowane do szeregów rozdzielczych, czasowych strukturalnych i przestrzennych.

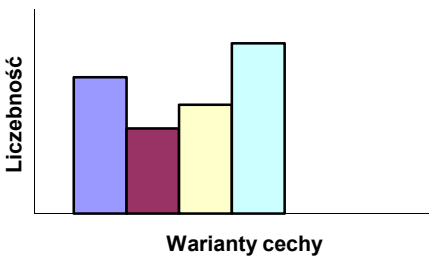
Wykresy mapowe są graficzną formą przestrzennego przedstawiania zróżnicowania wariantów cech. Na wykresach tych przedstawia się za pomocą barw (kartogram), lub połączenia mapy z wykresami (kartodiagram) różnice w natężeniu (na jednostkę powierzchni lub 1000 mieszkańców) badanego zjawiska.



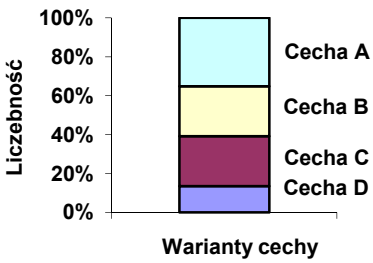
Wykres punktowy



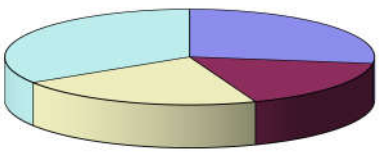
Wykres liniowy



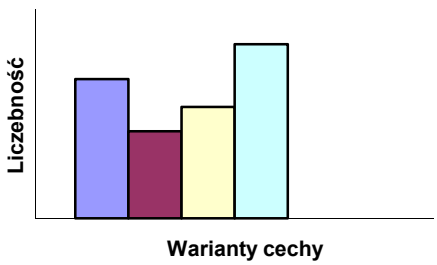
Wykres kolumnowy



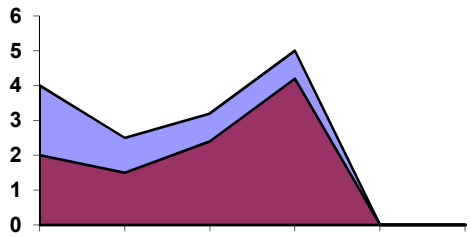
Wykres kolumnowy skumulowany



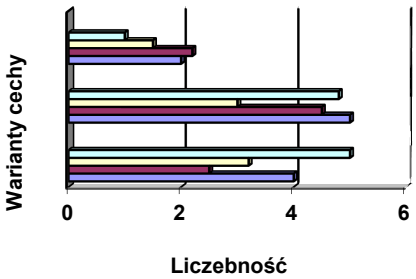
Wykres kołowy



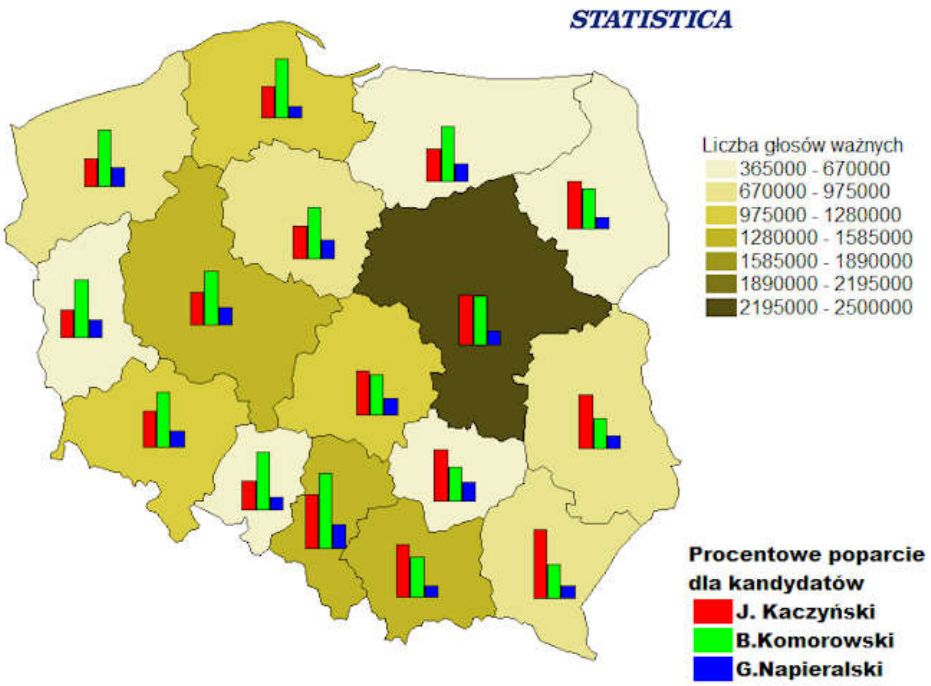
Wykres słupkowy



Wykres warstwowy



Wykres słupkowy grupowany



Wykres mapowy wyników I tury wyborów prezydenckich z 2010 roku
Źródło: https://maps-for-excel.com/blog/presentation-of-geographical-data/(dostęp: 6. 10. 2022)

1.6. Przykłady praktyczne

Przykład 1.6.1

Na poczcie przeprowadzono badanie wagi paczek (w kg) i otrzymano informacje:

2	5	2	5	4	10
3	4	3	6	4	2
4	10	4	2	3	4
6	8	6	5	4	2

Zbudować szereg rozdzielczy punktowy.

Rozwiązanie

Mamy do czynienia z cechą ilościową, skokową, możemy zatem zbudować szereg rozdzielczy punktowy.

Waga paczek	Liczba paczek
2	5
3	3
4	7
5	3
6	3
8	1
10	2
Razem	24

Przykład 1.6.2

Czas oczekiwania (w miesiącach) na oddanie do użytku domu jednorodzinnego przez pewnego dewelopera wśród 35 wybranych losowo osób, oczekujących na domek, przedstawia się następująco:

3,0	2,5	2,6	3,1	4,8	4,0	5,0
2,6	3,1	4,8	2,5	5,5	5,0	4,8
2,6	3,0	4,8	4,8	5,0	3,1	2,6
3,0	5,5	5,0	4,8	4,0	4,0	3,0
3,0	3,0	3,5	5,5	3,0	4,0	3,1

Na podstawie powyższych danych zbudować odpowiedni szereg.

Rozwiązanie

Należy zbudować szereg rozdzielczy przedziałowy. W tym celu określimy liczbę klas oraz rozpiętość przedziałów klasowych.

Liczba klas:

$$k = \sqrt{n} = \sqrt{35} \cong 6$$

Rozpiętość przedziału (klasy):

$$h = \frac{x_{max} - x_{min}}{k - 1} = \frac{5{,}5 - 2{,}5}{5} = 0{,}6$$

Dolna granica przedziału:

$$x_1 = x_{min} - \frac{1}{2}h = 2{,}5 - \frac{1}{2} \cdot 0{,}6 = 2{,}2$$

Przyjmując lewostronnie domknięte przedziały klasowe, przyporządkowując obserwacje określonym przedziałom i zliczając je w każdej klasie otrzymamy szereg rozdzielczy.

Cecha statystyczna	Numer klasy i	Warianty cechy statystycznej [miesiące] $< x_{d_i}, x_{g_i} >$	Liczba osób n_i
Czas oczekiwania na mieszkanie	1	<2,2 – 2,8)	6
	2	<2,8 – 3,4)	11
	3	<3,4 – 4,0)	1
	4	<4,0 – 4,6)	4
	5	<4,6 – 5,2)	10
	6	<5,2 – 5,8)	3
	×	Razem	35

Źródło: Opracowanie własne

Przykład 1.6.3

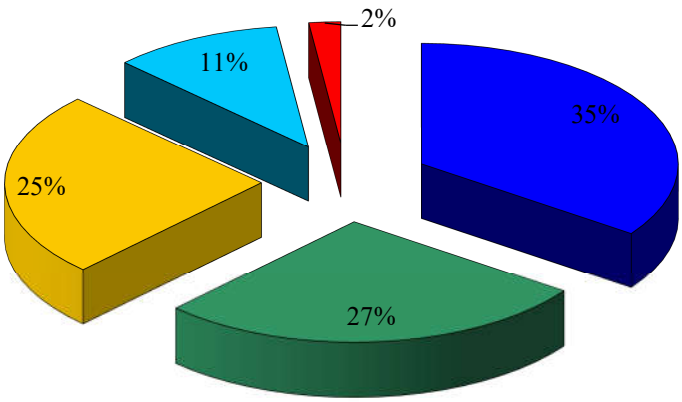
Obliczyć wskaźniki struktury i wykonać wykres kołowy, przedstawiający udziały procentowe dla podanych w tabeli danych.

Bezrobocie według wieku (dane umowne)

Wiek w latach x_i	Liczba bezrobotnych (w tys.) n_i	Obliczenia w_i (%)
do 24	24,0	34,8
25 – 34	18,9	27,4
35 – 44	17,3	25,1
45 – 54	7,4	10,7
55 i więcej	1,4	2,0
Ogółem	69,0	100,0

Źródło: Opracowanie własne

Uwaga: $w_i = \frac{n_i}{n} \cdot 100\%$



Wykres kołowy udziałów procentowych bezrobocia według wieku

Przykład 1.6.4

Dla danych umownych przedstawionych w tabeli dokonać kumulacji wskaźników struktury.

Liczba usterek	Wskaźnik struktury	Wskaźniki skumulowane
1	0,3	0,3
2	0,4	0,7
3	0,1	0,8
4	0,2	1,0
Razem	1,0	×

Przykład 1.6.5

Poniżej zapisane dane pogrupować w sześciu klasach.

14,0	14,8	15,0	15,5	16,1	16,5	16,6	17,0	17,0	17,3	18,1	18,4
18,7	19,1	19,1	19,5	19,6	19,9	20,0	20,1	20,7	20,8	20,8	21,0
21,0	21,1	21,4	21,6	21,8	21,8	22,0	22,1	22,4	22,4	22,5	23,3
23,6	23,6	23,9	24,1	24,8	25,0	25,3	25,7	26,1	26,9	27,8	28,7

Wyznaczyć liczebności skumulowane, wskaźniki struktury i skumulowane wskaźniki struktury.

Numer klasy	Przedział $<x_i;x_{i+1})$	Liczebność n_i	Liczebność skumulowana n_{isk}	Wskaźnik struktury w_i	Wskaźniki skumulowane
1	<14; 16,5)	5	5	0,1041	0,1041
2	<16,5; 19)	8	13	0,1667	0,2708
3	<19; 21,5)	14	27	0,2917	0,5625
4	<21.5; 24)	12	39	0,25	0,8125
5	<24; 26,5)	6	45	0,125	0,9375
6	<26,5; 29)	3	48	0,0625	1
Razem	×	48	×	1	×

Przykład 1.6.6

Dla danych umownych w tabeli wyznaczyć częstość występowania ocen ze statystyki wśród badanych studentów w_i w procentach.

Ocena ze statystyki	Liczba studentów	$w_i = \frac{n_i}{n} \cdot 100\%$
2	10	11,11
3	40	44,44
4	25	27,78
4,5	8	8,89
5	7	7,78
Razem	90	100,00

Ocena 3,0 występowała najczęściej, uzyskało ją aż 40 studentów (44,44%). Ocenę 5 uzyskało jedynie 7 studentów, co stanowiło 7,78% ogółu badanych.

ROZDZIAŁ 2
ANALIZA STRUKTURY ZBIOROWOŚCI

Analizę struktury zjawisk masowych można przeprowadzić za pomocą charakterystyk opisowych. Strukturę zbiorowości z punktu widzenia określonej cechy odzwierciedla jej rozkład, ustalany na podstawie obserwacji.

Rozkładem empirycznym danej zmiennej nazywamy przyporządkowanie kolejnym wartościom zmiennej x_i odpowiadających im liczebności n_i.

Analizując rozkład cechy mierzalnej należy brać pod uwagę następujące jego własności: tendencję centralną (przeciętny poziom), dyspersję (zróżnicowanie), asymetrię (skośność) i koncentrację (skupienie). Do oceny tych rozkładów służą charakterystyki zwane parametrami rozkładu. Rozróżnia się:

- parametry klasyczne, obliczane na podstawie wszystkich obserwacji,
- parametry pozycyjne, wyznaczane na podstawie ich miejsca w szeregu lub częstości występowania.

Parametry klasyczne stosuje się głównie do analizy rozkładów, które charakteryzują się tendencją centralną, tzn. takich, w których punkt skupienia znajduje się w środku rozkładu (symetryczne) lub w pobliżu środka rozkładu (umiarkowanie asymetryczne). Parametry pozycyjne można natomiast stosować do badania każdego typu rozkładu. Są one najbardziej przydatne w przypadku szeregów silnie asymetrycznych, a także takich, w których są otwarte przedziały klasowe (nie ma pełnych informacji o rozkładzie).

Parametry rozkładu mogą wystąpić jako liczby absolutne, wyrażone w tych samych jednostkach miary, co cecha zmienna, np. kg, m^2, ha, ale mogą też przyjąć postać liczb względnych, wyrażonych w ułamku lub w procentach. Parametry względne są przydatne przy porównywaniu rozkładów różnych cech w tej samej zbiorowości (np. porównanie rozkładu wysokości składek ubezpieczeniowych i rozkładu wysokości odszkodowań w PZU SA w Radomiu), bądź rozkładów tej samej cechy w różnych zbiorowościach (np. rozkład liczby dni absencji chorobowej w zbiorowości pracowników i zbiorowości studentów AHNS w Radomiu).

Do charakterystyk najczęściej stosowanych przy opisie struktury zbiorowości należą:

- ***momenty rozkładu***, miary klasyczne, obliczane są na podstawie wszystkich obserwacji,
- ***miary położenia*** (zwane też miarami średnimi, miarami przeciętnymi lub miarami poziomu wartości zmiennej); służą do określenia tej wartości zmiennej opisanej przez rozkład, wokół której skupiają się wszystkie pozostałe wartości,
- ***miary dyspersji*** (zróżnicowania, rozproszenia, zmienności); służą do badania stopnia zróżnicowania wartości zmiennej,
- ***miary asymetrii*** (skośności); wykorzystywane są do badania kierunku zróżnicowania wartości zmiennej,
- ***miary koncentracji*** stosowane są do badania stopnia nierównomierności rozkładu ogólnej sumy wartości zmiennej pomiędzy poszczególne jednostki zbiorowości; mogą też służyć do analizy stopnia skupienia poszczególnych jednostek wokół średniej.

Charakterystyki opisowe pozwalają w sposób syntetyczny określić właściwości badanych rozkładów i dokonać porównania różnych zbiorowości. Wyróżniamy na ogół dwa typy porównań:

- porównanie dwóch różnych zbiorowości, ale pod względem tej samej cechy, np. struktura bezrobotnych według wieku mężczyzn i kobiet,
- porównanie w ramach jednej zbiorowości, ale w stosunku do dwóch różnych cech, np. struktura urodzeń żywych według kolejności urodzenia dziecka i wieku matki.

2.1. Momenty rozkładu

W analizie rozkładu cechy mierzalnej ważną rolę odgrywają charakterystyki liczbowe, zwane momentami, zaliczane do miar klasycznych i obliczane na podstawie wszystkich wartości badanej cechy.

Momentem rzędu r nazywamy średnią arytmetyczną z podniesionych do potęgi r wartości cechy od pewnej stałej.

Moment zwykły otrzymamy, jeżeli przyjmiemy 0 jako stałą:

$$m_r = \frac{\sum_{i=1}^{k}(x_i - 0)^r \cdot n_i}{n} = \frac{\sum_{i=1}^{k} x_i^r \cdot n_i}{n}$$

Moment centralny uzyskamy, gdy przyjmiemy średnią arytmetyczną jako stałą:

$$\mu_r = \frac{\sum_{i=1}^{k}(x_i - \bar{x})^r \cdot n_i}{n} = \frac{\sum_{i=1}^{k} x_i^r \cdot n_i}{n}$$

Chcąc opisać własności rozkładu, należy skorzystać z czterech kolejnych momentów, które przedstawimy poniżej:

– moment pierwszy zwykły jest średnią arytmetyczną:

$$m_1 = \frac{\sum_{i=1}^{k} x_i \cdot n_i}{n}$$

– moment drugi zwykły:

$$m_2 = \frac{\sum_{i=1}^{k} x_1^2 \cdot n_i}{n}$$

– moment trzeci zwykły:

$$m_3 = \frac{\sum_{i=1}^{k} x_1^3 \cdot n_i}{n}$$

– moment czwarty zwykły:

$$m_4 = \frac{\sum_{i=1}^{k} x_1^4 \cdot n_i}{n}$$

– moment pierwszy centralny:

$$\mu_1 = \frac{\sum_{i=1}^{k}(x_i - \bar{x}) \cdot n_i}{n}$$

– moment drugi centralny nosi nazwę wariancji i jest miarą dyspersji:

$$\mu_2 = \frac{\sum_{i=1}^{k}(x_i - \bar{x})^2 \cdot n_i}{n}$$

– moment trzeci centralny jest miarą asymetrii:

$$\mu_3 = \frac{\sum_{i=1}^{k}(x_i - \bar{x})^3 \cdot n_i}{n}$$

– moment czwarty centralny jest miarą koncentracji:

$$\mu_4 = \frac{\sum_{i=1}^{k}(x_i - \bar{x})^4 \cdot n_i}{n}$$

Do obliczania momentów centralnych wykorzystuje się momenty zwykłe. Wzory powstają w oparciu o rozwinięcie wielomianu $(a+b)^r$, tak więc:

$$\mu_2 = m_2 - m_1^2$$

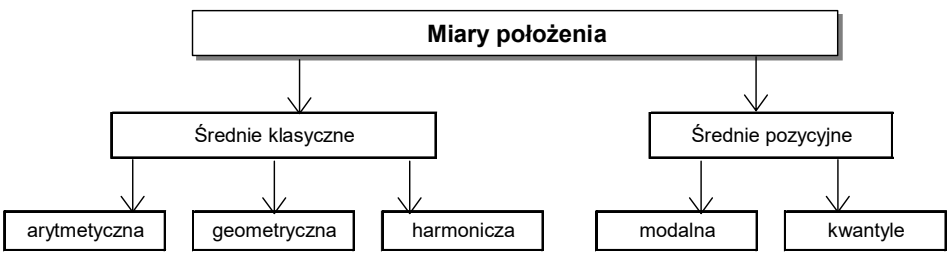
$$\mu_3 = m_3 - 3m_2m_1 + 2m_1^3$$

$$\mu_4 = m_4 - 4m_3m_1 + 6m_2m_1^2 - 3m_1^4$$

Zastosowanie przedstawionych wzorów zostanie pokazane praktycznie w przypadku obliczania wariancji, współczynnika asymetrii i współczynnika koncentracji.

2.2. Miary położenia

Zbiorowości rozpatrywane według cech ilościowych charakteryzują się na ogół pewną koncentracją wokół wartości centralnej badanej cechy. Miary położenia podają za pomocą jednej liczby charakterystykę poziomu wartości zmiennej badanej cechy. Miary te charakteryzują zbiorowość statystyczną jako całość, informują o przeciętnym poziomie cechy, abstrahując od różnic pomiędzy poszczególnymi jednostkami. Są to miary mianowane, które pozwalają ocenić średni lub typowy poziom wartości cechy. Charakterystyki liczbowe obliczane w oparciu o wszystkie wartości zmiennej nazywamy miarami klasycznymi, zaś te, które nie obejmują wszystkich realizacji zmiennej nazywamy miarami pozycyjnymi. Obie grupy miar wzajemnie się uzupełniają, każda z nich opisuje poziom wartości cechy zmiennej z innego punktu widzenia. Są jednak przypadki, gdy układ informacji liczbowych nie pozwala na obliczanie miary średniej. Ten problem zostanie wyjaśniony w dalszej kolejności.



Rys. 2.1. Klasyfikacja miar położenia

Załóżmy, że warianty cechy mierzalnej (zmiennej) występują w badanej zbiorowości n razy i przyjmują wartości $x_1, x_2, \ldots, x_n$. Jest to szereg statystyczny prosty.

Średnia arytmetyczna jest ilorazem sumy poszczególnych wartości badanej cechy i liczby obserwacji.

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

Ta formuła przedstawia średnią ***arytmetyczną prostą***. Obliczamy ją, gdy obserwacje przedstawione są w postaci szeregu szczegółowego.

Przykład 2.2.1

Informacje o liczbie chorych dzieci przyjętych przez lekarzy pediatrów w jednej z radomskich przychodni w kolejnych dniach tygodnia zawarto w następującym zestawieniu:

Dzień tygodnia	Liczba przyjęć
Poniedziałek	100
Wtorek	85
Środa	90
Czwartek	95
Piątek	80
Razem	450

Wyznaczyć średnią liczbę przyjęć w tej przychodni.

Rozwiązanie

Korzystamy ze wzoru:

$$\bar{x} = \frac{1}{n} \sum_{i=1}^{n} x_i = \frac{1}{5} \cdot 450 = 90$$

W badanej przychodni średnia liczba przyjętych dzieci wynosiła 90.

Jeżeli zbiorowość jest podzielona na klasy, a poszczególnym wartościom cechy x_i odpowiadają liczebności n_i, wówczas mamy do czynienia z szeregiem rozdzielczym. W takim przypadku stosujemy formułę ***średniej arytmetycznej ważonej***, gdzie częstości występowania poszczególnych wartości cechy nadają znaczenie (wagę) tym wartościom.

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

Średnią ważoną możemy obliczyć, przyjmując jako wagi liczebności bezwzględne częstości lub wskaźniki struktury.

Przykład 2.2.2

W pewnej firmie przeprowadzono badanie jakościowe wybranej partii śrub. W tym celu wylosowano 50 śrub i sprawdzono liczbę usterek. Okazało się, że 2 śruby spełniają wszystkie wymogi jakościowe, więc nie mają ich wcale,11 śrub ma jedną usterkę, 25 posiada ich dwie, zaś w 8 zauważono 3 usterki. W 4 kolejno sprawdzanych śrubach wykryto aż cztery usterki. Jaka była średnia liczba usterek w badanej partii?

Rozwiązanie

Wyniki obserwacji zestawimy w postaci szeregu rozdzielczego punktowego, a następnie obliczymy średnią arytmetyczną liczby usterek.

Liczba usterek x_i	Liczba śrub n_i	$x_i\, n_i$
0	2	0
1	11	11
2	25	50
3	8	24
4	4	16
Razem	50	101

Źródło: Dane umowne

$$\bar{x} = \frac{1}{50} \cdot 101 = 2{,}02$$

Na wybraną do badania jakościowego partię śrub przypadają przeciętnie biorąc, 2 usterki.

Przykład 2.2.3

Obliczyć wskaźniki struktury kapitału zagranicznego zakładów ubezpieczeń według krajów pochodzenia (stan w dniu 31 grudnia 2018 r.) oraz średni kapitał.

Wyszczególnienie	Kapitał (mln zł)	w_i	$x_i \cdot w_i$
Austria	1356,0	0,339	459,684
Belgia	19,0	0,005	0,095
Francja	703,0	0,176	123,728
Holandia	742,8	0,186	138,161
Japonia	58,2	0,015	0,873
Kanada	169,2	0,042	7,106
Luksemburg	30,1	0,008	0,241
Niemcy	677,9	0,170	115,243
Stany Zjednoczone	60,0	0,015	0,900
Wielka Brytania	153,8	0,038	5,844
Razem	3996,0	1,000	851,875

Źródło: Opracowanie na podstawie: Rocznik Statystyczny Rzeczypospolitej Polskiej 2019

Rozwiązanie

Jeżeli w miejscu liczebności n_i występują wskaźniki struktury w_i, to średnia arytmetyczna wyraża się wzorem:

$$\bar{x} = \sum_{i=1}^{k} x_i \cdot w_i$$

Średni kapitał zagraniczny zakładów ubezpieczeń według krajów pochodzenia w badanym okresie wyniósł 851,88 mln zł.

Jeżeli obserwacje (dane) są w postaci szeregu rozdzielczego, wówczas obliczamy średnią arytmetyczną ważoną. W szeregu rozdzielczym przedziałowym wartość cechy badanej nie jest podana w postaci jednej liczby, należy więc dla każdego przedziału wybrać jedną wielkość reprezentującą wszystkie wartości tego przedziału. Tą wielkością będzie środek przedziału, oznaczony przez $\dot{x}_i$. Poszczególne liczebności pojawiają się z różną częstotliwością, wagami są liczebności. Wówczas wzór średniej arytmetycznej przyjmuje postać:

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

Ponieważ $\sum_{i=1}^{k} n_i = n$.

Przykład 2.2.4

Zbadać, jak kształtuje się przeciętne miesięczne zużycie wody w rodzinach mieszkających w 100 wylosowanych domkach jednorodzinnych w miejscowości M.

Miesięczne zużycie wody (w m^3) x_i	Liczba rodzin n_i	Obliczenia pomocnicze	
		$\dot{x}_i$	$\dot{x}_i \cdot n_i$
10 – 12	5	11	55
12 – 14	18	13	234
14 – 16	29	15	435
16 – 18	35	17	595
18 – 20	13	19	247
Razem	100	x	1566

Źródło: Opracowanie własne

Rozwiązanie

$$\bar{x} = \frac{1}{n}\sum_{i=1}^{k} \dot{x}_i n_i = \frac{1566}{100} = 15{,}66\, m^3$$

Przeciętne miesięczne zużycie wody wynosi 15,66 m^3.

Wybrane własności średniej arytmetycznej:

1. Suma wartości cechy *X* jest równa średniej arytmetycznej pomnożonej przez liczebność:

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

2. Suma odchyleń poszczególnych wartości cechy X od średniej arytmetycznej jest równa zeru:

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

3. Suma kwadratów odchyleń poszczególnych wartości cechy *X* od średniej arytmetycznej jest mniejsza niż suma kwadratów odchyleń od jakiejkolwiek innej liczby, np. „*z*”:

$$\sum_{i=1}^{n} (x_i - \bar{x})^2 < \sum_{i=1}^{n} (x_i - z)^2 \qquad z \neq \bar{x}$$

4. Średnia arytmetyczna jest większa od najmniejszej wartości cechy i mniejsza od jej największej wartości:

$$\min_i \{x_i\} < \bar{x} < \max_i \{x_i\}$$

Uwagi:

1. Średniej arytmetycznej nie można obliczać dla szeregu o otwartych przedziałach klasowych, wtedy należy umownie przyjąć granice tych przedziałów bądź stosować inną miarę, np. medianę.
2. Średniej arytmetycznej nie należy obliczać, gdy w zbiorowości występują wartości skrajne (duże lub małe). Możemy posłużyć się wtedy średnią geometryczną, która jest mniej czuła na wartości ekstremalne.
3. Średniej arytmetycznej nie obliczamy na podstawie szeregu rozdzielczego, gdy jest on skrajnie asymetryczny (tj. gdy największe liczebności skupiają się wokół najwyższych wartości lub najniższych wartości cechy).
4. Średnią arytmetyczną możemy obliczyć, jeżeli liczebność w otwartym przedziale klasowym stanowi niewielki odsetek badanej zbiorowości (do 5%), możliwe jest wówczas zamknięcie takiego przedziału.

5. Średnie klasyczne obliczane są na podstawie wszystkich wartości szeregu.

Średnią harmoniczną stosujemy, gdy wartości jednostek zbiorowości statystycznej są podane w formie odwrotności, tj. gdy wartości jednej zmiennej są podane w przeliczeniu na stałą jednostkę innej zmiennej (np. 80 km/godz.) lub wyrażone w postaci złożonej (np. obrót = cena × ilość). Miary tej używamy w przypadkach obliczania:

- przeciętnej szybkości pojazdów mechanicznych (km/godz.),
- przeciętnego czasu potrzebnego do wykonania pewnych czynności (ton/godz.),
- przeciętnej ceny towarów, których cena jest wyrażona w liczbie jednostek towaru za jednostkę pieniężną,
- przeciętnej szybkości obrotów pieniężnych (obrotu funduszu).

$$\bar{x}_H = \frac{n}{\sum_{i=1}^{n} \frac{1}{x_i}}$$ średnia harmoniczna prosta

$$\bar{x}_H = \frac{\sum_{i=1}^{k} n_i}{\sum_{i=1}^{k} \frac{n_i}{x_i}}$$ średnia harmoniczna ważona

Przykład 2.2.5

W ciągu 8 godzin pracy obserwowano pracę trzech robotników. Robotnik A zużywał na wykonanie jednego elementu 4 minuty, robotnik B – 6 minut, robotnik C – 12 minut. Ile czasu zużywają średnio ci robotnicy na wykonanie jednego elementu.

Rozwiązanie

$$n = 3$$

$$x_i : 4, 6, 12$$

$$\bar{x}_H = \frac{3}{\frac{1}{4} + \frac{1}{6} + \frac{1}{12}} = 6 \text{ minut}$$

Robotnicy zużywają średnio 6 minut na wykonanie jednego elementu.

Średnią geometryczną obliczamy, gdy w szeregu występują znaczne różnice między obserwacjami:

$$\bar{x}_G = \sqrt[n]{x_1 \cdot x_2 \cdot \ldots \cdot x_n} = \sqrt[n]{\bigcap_{i=1}^{n} x_i}$$ średnia geometryczna prosta

$$\bar{x}_G = \sqrt[\sum_{i=1}^{k} n_i]{x_1^{n_1} \cdot x_2^{n_2} \cdot \ldots \cdot x_k^{n_k}}$$ średnia geometryczna ważona

n_i – liczebność poszczególnych klas.

Średnia geometryczna jest obliczana, gdy w zbiorze informacji występują wartości skrajne, znacząco różniące się od większości zaobserwowanych wartości cechy. W tych przypadkach średnia geometryczna lepiej niweluje wpływ tych wartości skrajnych na ocenę średniego poziomu. Jest stosowana również w analizie szeregów czasowych jako miara średniego przyrostu.

Przykład 2.2.6

Właściciel działki dokonał pięciu pomiarów wysokości drzew otrzymując wyniki [w metrach]: 2; 2,5; 2; 1,5; 14. Ocenić średnią wysokość drzewa. Którą miarę zastosować?

Rozwiązanie

$$\bar{x} = \frac{1}{5}(2 + 2{,}5 + 2 + 1{,}5 + 14) = 4{,}4$$ średnia arytmetyczna

$$\bar{x}_G = \sqrt[5]{2 \cdot 2{,}5 \cdot 2 \cdot 1{,}5 \cdot 14}) = 2{,}9$$ średnia geometryczna

Zauważmy, że zastosowanie średniej geometrycznej dało wynik bardziej zbliżony do tych wartości cechy, przy których występowało skupienie obserwacji.

Przykład 2.2.7

Zbadać, jak kształtuje się dzienne zużycie energii elektrycznej w rodzinach mieszkających w pewnej niewielkiej miejscowości.

Dzienne zużycie energii elektrycznej w kWh x_i	Liczba rodzin n_i	Obliczenia pomocnicze	
		$\dot{x}_i$	$\dot{x}_i \cdot n_i$
2 – 4	6	3	18
4 – 6	10	5	50
6 – 8	30	7	210
8 – 10	40	9	360
10 – 12	10	11	110
12 – 14	4	13	52
Razem	100	x	800

Źródło: Opracowanie własne

Rozwiązanie

$$\bar{x} = \frac{1}{n}\sum_{i=1}^{k} \dot{x}_i n_i = \frac{1}{100} \cdot 800 = 8\ kWh$$

Średnie dzienne zużycie energii elektrycznej wynosi 8 kWh.

Przykład 2.2.8

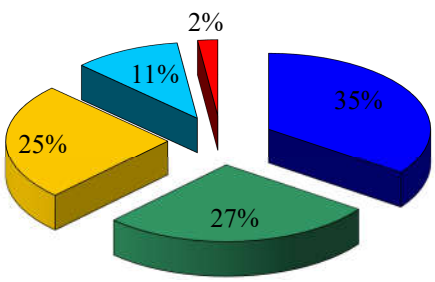
Obliczyć wskaźniki struktury i wykonać wykres kołowy przedstawiający udziały procentowe.

Bezrobocie wg wieku w wybranym województwie (dane umowne)

Wiek w latach x_i	Liczba bezrobotnych (w tys.) n_i	Obliczenia w_i (%)
do 24	24,0	34,8
25 – 34	18,9	27,4
35 – 44	17,3	25,1
45 – 54	7,4	10,7
55 i więcej	1,4	2,0
Ogółem	69,0	100,0

Źródło: Opracowanie własne

Udziały procentowe bezrobocia według wieku pokazuje rysunek 2.2.



Rys. 2.2. Wykres kołowy udziałów procentowych bezrobocia według wieku

Przykład 2.2.9

Jaka jest liczba nieobecności na zajęciach ze statystyki w semestrze zimowym, zaobserwowanej w grupie studentów liczącej 35 osób?

Liczba nieobecności w semestrze zimowym x_i	Liczba studentów n_i	Obliczenia pomocnicze $x_i \cdot n_i$
0	10	0
1	12	12
2	8	16
3	3	9
4	1	4
5	1	5
Razem	35	46

Źródło: Opracowanie własne

Rozwiązanie

$$\bar{x} = \frac{1}{n}\sum_{i=1}^{k} x_i \cdot n_i = \frac{1}{35} \cdot 46 = 1{,}3$$

Na każdego studenta przypada średnio 1,3 nieobecności w semestrze zimowym.

Modalna (dominanta, moda) jest to wartość cechy statystycznej, która w danym rozkładzie empirycznym występuje najczęściej.

W szeregach szczegółowych i rozdzielczych punktowych jest to ta wartość cechy, której odpowiada największa liczebność (częstość). W szeregach rozdzielczych przedziałowych modalną wyznacza się ze wzoru interpolacyjnego:

$$M_0 = x_{0m} + \frac{n_m - n_{m-1}}{(n_m - n_{m-1}) + (n_m - n_{m+1})} \cdot h_m$$

gdzie:

m – numer przedziału (klasy) modalnej,

x_{0m} – dolna granica przedziału modalnej,

n_m – liczebność przedziału modalnej,

n_{m-1}, n_{m+1} – liczebności klas: poprzedzającej przedział modalnej i następujący po przedziale modalnej,

h_m – rozpiętość przedziału modalnej.

Uwagi:

1. Wyznaczanie modalnej w szeregach rozdzielczych ma sens wtedy, gdy rozkład empiryczny jest jednomodalny (występuje jedno wyrażenie zaznaczone maksimum).
2. Przedział, w którym występuje modalna oraz dwa sąsiadujące z nim przedziały muszą mieć takie same rozpiętości.
3. Jeżeli rozkład cechy jest skrajnie asymetryczny, wówczas modalnej nie można wyznaczyć analitycznie.
4. Przy interpretacji modalnej należy pamiętać, że charakteryzuje ona jednostki o typowym poziomie cechy, nie zaś wszystkie badane jednostki.

Przykład 2.2.10

Na poczcie przeprowadzono badanie wagi paczek (w kg) i otrzymano informacje:

2	5	2	5	4	10
3	4	3	6	4	2
4	10	4	2	3	4
6	8	6	5	4	2

Ocenić dominującą wagę paczek.

Rozwiązanie

Zbudujemy szereg rozdzielczy punktowy.

Waga paczek	Liczba paczek	
2	5	
3	3	
→ 4	7	M_0
5	3	
6	3	
8	3	
10	2	
Razem	24	

Patrząc na liczebności zauważamy, że wartość najwyższa jest 7, a zatem dominująca waga paczek wynosi 4 kg.

Przykład 2.2.11

W pewnej spółce handlującej mieszkaniami sporządzono następujący rejestr:

Powierzchnia mieszkań (w m^2)	Liczba mieszkań	
30 – 50	8	
50 – 70	19	
→ 70 – 90	23	M_0
90 – 110	14	
110 i więcej	16	

Jaka była najczęściej spotykana powierzchnia mieszkań?

Rozwiązanie

Stwierdzamy, że najliczniejszy przedział to: 70 – 90 m^2, a rozpiętości przedziałów są jednakowe i wynoszą 20. Zastosujemy wzór interpolacyjny:

$$M_0 = 70 + \frac{23-19}{(23-19)+(23-14)} \cdot 20 = 76{,}2\ m^2$$

Najczęściej spotykana, czyli dominująca powierzchnia mieszkań wynosi 76,2 m^2, a obliczona wartość dominanty mieści się w przedziale najliczniejszym.

Przykład 2.2.12

Wyznaczyć dominantę dla danych wartości liczbowych w trzech przypadkach:

a) 5, 0, 5, 1, 5, 7, 0, 2,
b) 1, 2, 4, 5, 1, 4, 3, 2, 7, 5, 4, 2, 8, 3,
c) 1, 2, 3, 1, 2, 3, 1, 2, 3,

Rozwiązanie

a) Najczęściej występującą jest liczba 5 (występuje 3 razy), a zatem $M_o = 5$.
b) W tym zestawie danych liczba 2 występuje 3 razy i liczba 4 występuje 3 razy. Zestaw jest dwumodalny, czyli $M_0 \in \{2{,}4\}$
c) Nie ma wartości liczbowej występującej najczęściej. Dominanta tych danych nie istnieje.

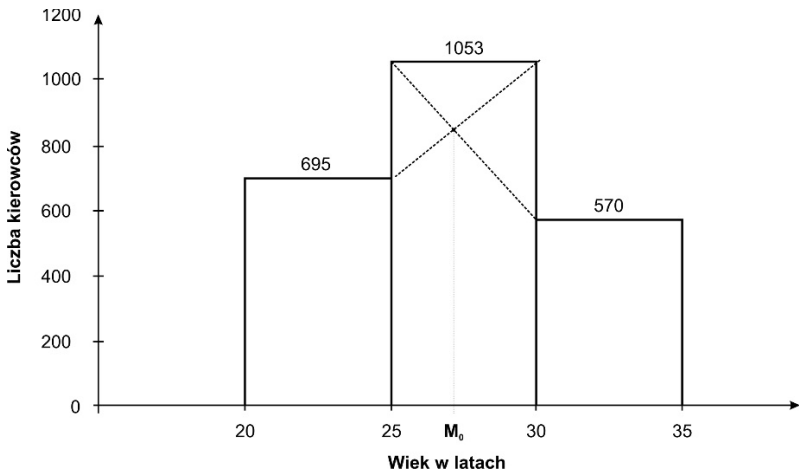
Modalną można wyznaczyć graficznie. Metoda sprowadza się do wykreślenia histogramu liczebności z trzech przedziałów klasowych: przedziału, w którym znajduje się wartość dominująca oraz dwóch sąsiednich. Ilustruje to przykład 2.2.13.

Przykład 2.2.13

Wiek kierowców w pewnym mieście, ukaranych mandatem a wykroczenia drogowe, którzy nie dokonali wpłaty i są dłużnikami (dane umowne). Wyznaczyć graficznie modalną wieku tych kierowców.

Lp	Wiek w latach	Liczba kierowców
1	do 20	31
2	20-25	695
3	25-30	1053
4	30-35	571
5	35-40	421
6	40-50	730
7	50 i więcej	399

Graficzną metodę wyznaczania modalnej wieku tych kierowców pokazuje rysunek 2.3.



Rys. 2.3. Graficzna metoda wyznaczania modalnej

Kwantyle definiuje się jako wartości cechy badanej zbiorowości statystycznej, przedstawionej w postaci szeregu statystycznego, które dzielą zbiorowość na określone części pod względem liczby jednostek. Części te mogą być równe, lub pozostawać w stosunku do siebie w określonych proporcjach. Szeregi, z których wyznacza się kwantyle muszą być uporządkowane według rosnących lub malejących wartości cechy statystycznej. Do najczęściej stosowanych kwantyli należą ***kwartyle*** *i* ***decyle.***

***Kwartyl pierwszy*Q_1** dzieli zbiorowość na dwie części w ten sposób, że 25% jednostek zbiorowości ma wartości cechy niższe bądź równe kwartylowi pierwszemu Q_1,a 75% równe bądź wyższe od tego kwartyla.

***Kwartyl drugi*Q_2 (*mediana* M_e)** dzieli zbiorowość na dwie równe części; połowa jednostek ma wartości cechy mniejsze lub równe medianie, a połowa wartości cechy równe lub większe od M_e. Medianę nazywa się wartością środkową.

***Kwartyl trzeci* Q_3** dzieli zbiorowość na dwie części w ten sposób, że 75% jednostek ma wartości cechy niższe bądź równeQ_3, a 25% równe bądź wyższe od kwartyla trzeciego.

W szeregach szczegółowych medianę wyznaczamy według formuły:

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

Kwartyle pierwszy i trzeci wyznacza się analogicznie, jak medianę.

W szeregach rozdzielczych wyznaczamy kwartyle według wzoru interpolacyjnego:

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

gdzie:

x_{0_m} –wartość dolnej granicy przedziału kwartyla,

L – numer kwartyla,

m – numer przedziału (klasy) kwartyla,

n_m –liczebność przedziału odpowiedniego kwartyla,

$\sum_{i=1}^{m-1} n_i$ –suma liczebności poprzedzających przedział odpowiedniego kwartyla (liczebność skumulowana),

h_m – rozpiętość przedziału kwartyla,

N_{QL} – pozycja kwartyla.

$$N_{Q1}=\frac{n}{4};\qquad N_{Q2}=\frac{n}{2};\qquad N_{Q3}=\frac{3n}{4}\qquad n-parzyste$$

$$N_{Q1}=\frac{n+1}{4};\quad N_{Q2}=\frac{n+1}{2};\quad N_{Q3}=\frac{3(n+1)}{4}\quad n-nieparzyste$$

Uwagi:

1. Mediana jest obok średniej arytmetycznej najczęściej stosowanym parametrem statystycznym. Może być obliczana w przypadkach, gdy szereg ma otwarte przedziały klasowe, a rozpiętości przedziałów klasowych są różne.
2. Mediana nie reaguje na zmiany wartości cech skrajnych jednostek, na tzw. obserwacje nietypowe.
3. Gdy badaną zbiorowość traktujemy jako próbę pobraną z populacji generalnej, wówczas przy zmianie próby mediana ulega większym zmianom niż średnia arytmetyczna.
4. W szeregu rozdzielczym punktowym medianą jest wartość cechy we wskazanym przedziale, natomiast w szeregu przedziałowym stosujemy wzór interpolacyjny.

Przykład 2.2.14

Dostawy czereśni do siedmiu punktów skupu (w tonach) kształtowały się następująco: 19,9, 15,2, 25,0, 16,3 ; 20,4, 23,0, 21,9 . Wyznaczyć medianę.

Rozwiązanie

Porządkujemy dane rosnąco: 15,3; 16,3; 19,9; 20,4; 21,0; 23,0; 25,0 i wskazujemy wartość środkową, a zatem mediana wynosi 20,4 tony.

Przykład 2.2.15

Wyznaczyć medianę liczby nieobecności studentów na zajęciach ze statystyki, mając dane:

Liczba nieobecności w semestrze (x_i)	0	1	2	3	4	5
Liczba studentów (n_i)	10	12	8	3	1	1

Rozwiązanie

W celu wyznaczenia mediany należy skumulować liczebności, aby określić pozycję mediany.

Liczba nieobecności w semestrze (x_i)	Liczba studentów (n_i)	Liczebność skumulowana	
0	10	10	
1	12	22	
→ 2	8	30	M_e
3	3	33	
4	1	34	
5	1	35	
Razem	35	×	

Wynik oznacza, że połowa badanych studentów była nieobecna nie więcej niż jeden raz w semestrze, a druga połowa nie mniej.

Przykład 2.2.16

Obliczyć medianę wieku kobiet zawierających związek małżeński (dane umowne):

Wiek kobiet w latach	Odsetek kobiet	Skumulowane częstości względne	
do 19	0,21	0,21	
→ 20 – 24	0,56	0,77	M_e
25 – 29	0,13	0,90	
30 – 34	0,04	0,94	
35 – 39	0,02	0,96	
40 – 49	0,02	0,98	
50 – 59	0,01	0,99	
60 i więcej	0,01	1,00	
Razem	1,00	×	

Rozwiązanie

Określamy pozycję mediany i wskazujemy przedział, a następnie wykorzystujemy wzór interpolacyjny:

$$M_e = 20 + \frac{0,5 - 0,21}{0,56} \cdot 5 \approx 22,59\ lat$$

Wynik oznacza, że połowa kobiet zawierających związek małżeński miała mniej niż 22,59 lat, zaś druga połowa więcej.

Przykład 2.2.17

Przeprowadzono anonimowy sondaż, dotyczący liczby wpłat na konto wspomagające chorego tytułem darowizny i przedstawiono informacje w tabeli (dane umowne). Obliczyć kwartyle.

Lp.	Wysokość wpłat (w zł)	Liczba darczyńców	Liczebność skumulowana	
1	Poniżej 100	10	10	
2	100 – 200	20	30	Q_1
3	200 – 300	27	57	Q_2
4	300 – 400	21	75	Q_3
5	400 – 500	13	81	
6	Powyżej 500	9	100	
	Razem	100	×	

Rozwiązanie

Po skumulowaniu liczebności wyznaczamy pozycje poszczególnych kwartyli:

$$N_{Q1} = \frac{100}{4} = 25; \quad N_{Q2} = \frac{100}{2} = 50; \quad N_{Q3} = \frac{3 \cdot 100}{4} = 75$$

Odszukujemy w liczebności skumulowanej przedziały, w których wyznaczamy kwartyle:

$$Q_1 = x_{02} + \frac{N_{Q1} - \sum_{i=1}^{2-1} n_i}{n_2} \cdot h_2 = 100 + \frac{25 - 10}{20} \cdot 100 = 175\,zł$$

$$Q_2 = x_{03} + \frac{N_{Q2} - \sum_{i=1}^{3-1} n_i}{n_3} \cdot h_3 = 200 + \frac{50 - 30}{27} \cdot 100 = 274\,zł$$

$$Q_3 = x_{04} + \frac{N_{Q3} - \sum_{i=1}^{4-1} n_i}{n_4} \cdot h_4 = 300 + \frac{75 - 57}{21} \cdot 100 = 386\,zł$$

Otrzymane wyniki oznaczają, że 25% darczyńców wpłaciło nie więcej niż 175 zł, a pozostałe 75% nie mniej. Połowa wpłaciła nie więcej niż 274 zł, zaś druga połowa nie mniej. 75% darczyńców wpłaciło nie więcej niż 386 zł, a pozostałe 25% nie mniej.

Przykład 2.2.18

Rozkład urodzeń żywych według wieku matki w latach (dane umowne).

Wiek matki	Odsetek urodzeń
do 24	0,43
24 – 29	0,28
29 – 34	0,16
34 – 39	0,08
39 – 44	0,03
44 i więcej	0,02
	1,00

Ocenić za pomocą miar przeciętnych urodzenia według wieku matki.

Rozwiązanie

Zastosujemy miary pozycyjne. Średniej arytmetycznej nie możemy obliczyć ze względu na otwarte przedziały i dużą asymetrię rozkładu. Nie możemy też wyznaczyć kwartyla pierwszego, ponieważ otwarty pierwszy przedział klasowy grupuje ponad 25% zbiorowości. Dominującą jest grupa matek, które urodziły dziecko w wieku do 24 lat. W tej grupie wiekowej matek mamy 43% ogólnej liczby narodzin. Możemy obliczyć medianę i kwartyl trzeci.

$$M_e = x_{02} + \frac{N_{Q2} - \sum_{i=1}^{2-1} n_i}{n_2} \cdot h_2 = 24 + \frac{0,5 - 0,43}{0,28} \cdot 5 = 25,25 \ \ lat$$

Połowa badanych dzieci została urodzona przez matki w wieku nie przekraczającym 25 lat.

$$Q_3 = x_{03} + \frac{N_{Q3} - \sum_{i=1}^{3-1} n_i}{n_3} \cdot h_3 = 29 + \frac{0,75 - 0,71}{0,16} \cdot 5 = 30,25 \ lat$$

Około 75% badanych dzieci było urodzonych przez matki w wieku nie przekraczającym 30 lat.

Przykład 2.2.19

Mediana stażu pracy (w latach) 90 pracowników znajdowała się w przedziale od 12 do 16 lat. W tym przedziale znajdowało się 30 pracowników. Wartość mediany wynosiła około 12,3 roku. Ilu pracowników miał staż poniżej 12 lat.

Rozwiązanie

Korzystając z informacji zawartych w zadaniu tworzymy wzór, wyznaczający wartość mediany:

$$12{,}3 = 12 + \frac{45 - \sum_{i=1}^{k-1} n_i}{30} \cdot 4$$

Nieznaną we wzorze wielkością jest liczebność zbiorowości zsumowana do przedziału mediany, a więc jest to liczba pracowników, którzy mieli staż pracy mniejszy niż 12 lat. Po przekształceniu wzoru i dokonaniu odpowiednich obliczeń otrzymujemy:

$$\sum_{i=1}^{k-1} n_i \approx 43 \text{ pracowników}$$

Około 43 pracowników miało staż pracy mniejszy od 12 lat.

Przykład 2.2.20

Przedsiębiorstwo utworzyło nowe stanowiska pracy w pierwszym półroczu (dane umowne).

Miesiąc	Liczba nowych stanowisk
Styczeń	3
Luty	2
Marzec	4
Kwiecień	2
Maj	6
Czerwiec	0

Wyznaczyć poznane średnie.

Rozwiązanie

Możemy obliczyć średnią arytmetyczną prostą:

$$\bar{x} = \frac{3+2+4+2+6+0}{6} = 2{,}83$$

Przedsiębiorstwo utworzyło około 3 stanowisk na miesiąc w pierwszym półroczu, średnio biorąc.

Patrząc na liczbę nowych stanowisk zauważamy, że $M_0 = 2$.

Ponieważ dysponujemy parzystą liczbą obserwacji, medianę obliczamy jako średnią z dwóch środkowych. W tym celu porządkujemy obserwacje rosnąco, a następnie znajdujemy środek: 0, 2, 2, 3, 4, 6:

$$M_e = \frac{1}{2}(x_{\frac{n}{2}} + x_{\frac{n}{2}+1}) = \frac{1}{2}(x_{\frac{6}{2}} + x_{\frac{6}{2}+1}) = \frac{1}{2}(x_3 + x_4) = \frac{1}{2}(2+3) = 2{,}5$$

W połowie badanego okresu utworzono nie więcej niż 2,5 stanowiska.

2.3. Miary dyspersji

Obliczenie wartości średniej badanej cechy jest pewnym kryterium poznania zbiorowości, ale nie informuje, jaka jest zmienność cechy. Wartości średnie nie dają wyczerpującej charakterystyki struktury zbiorowości. Na przykład, jeżeli średnia płaca dwóch brygad jest na tym samym poziomie, to nie znaczy to, że zarobki w obu przypadkach są jednakowe. Zróżnicowanie płac w każdej brygadzie może być inne.

Dyspersją (rozproszeniem) nazywamy zróżnicowanie jednostek zbiorowości statystycznej ze względu na wartość badanej cechy. Siłę dyspersji oceniamy za pomocą klasycznych i pozycyjnych miar zmienności.

Grupę miar dyspersji można także podzielić na bezwzględne i względne. Do miar bezwzględnych (mianowanych) zaliczamy obszar zmienności, wariancję, odchylenie standardowe, przeciętne i ćwiartkowe. Względną miarą dyspersji jest współczynnik zmienności, wyrażany w procentach.

2.3.1. Klasyczne miary dyspersji

***Wariancja*(S^2)** jest średnią arytmetyczną z kwadratów odchyleń wartości cechy od średniej arytmetycznej:

$$S^2(x) = \frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})^2 \qquad \text{dla szeregu szczegółowego}$$

$$S^2(x) = \frac{1}{n}\sum_{i=1}^{k}(x_i - \bar{x})^2 \cdot n_i \qquad \text{dla szeregu rozdzielczego}$$

Im większa jest wariancja, tym silniejsze jest zróżnicowanie badanej cechy. Ponieważ wariancja nie ma logicznej interpretacji, przy ocenie dyspersji posługujemy się odchyleniem standardowym, będącym pierwiastkiem kwadratowym z wariancji.

Odchylenie standardowe (S) jest średnią z odchyleń wartości cechy od jej średniej arytmetycznej:

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

Odchylenie standardowe określa, o ile wszystkie jednostki danej zbiorowości różnią się średnio od średniej arytmetycznej badanej zmiennej. Jest to liczba mianowana (zł, t, m), uniemożliwia to bezpośrednie porównywanie kilku zbiorowości.

Uwagi:

1. Odchylenie standardowe jest wielkością obliczoną na podstawie wszystkich obserwacji w danym szeregu.
2. Jego wartość nie zmieni się, jeśli liczebność szeregu wyrazimy w liczbach względnych (procentach) dokładnie ustalonych.
3. Jego wartość nie zmieni się, jeśli do wszystkich wartości zmiennej w szeregu dodamy pewną stałą liczbę.
4. Jeśli wszystkie wartości szeregu pomnożymy przez pewną stałą liczbę większą od zera, to odchylenie standardowe będzie również tylokrotnie większe.
5. Odchylenie standardowe możemy wykorzystać do konstrukcji typowego obszaru zmienności. W obszarze tym mieści się około $\frac{2}{3}$ wszystkich jednostek badanej zbiorowości statystycznej, bo jest on zawarty
   w granicach dwóch odchyleń standardowych.

$$\bar{x} - S < x_{typ} < \bar{x} + S$$

Przykład 2.3.1

Liczbę abonentów telefonii komórkowej na 1000 ludności w krajach wstępujących do Unii Europejskiej w 2004 roku przedstawia tabela.

Kraj	Cypr	Estonia	Litwa	Łotwa	Malta	Polska	Republika Czeska	Słowacja	Słowenia	Węgry
Liczba abonentów	448	543	293	267	292[a)]	249	677	399	758	488

a) 2000 r.

Źródło: Polska – Unia Europejska, GUS, Warszawa 2003

Wyznaczyć odchylenie standardowe.

Rozwiązanie

Tablica obliczeniowa

Lp.	Kraj	Liczba abonentów	$x_i - \bar{x}$	$(x_i - \bar{x})^2$
1.	Cypr	448	6,6	43,56
2.	Estonia	543	101,6	10 322,56
3.	Litwa	293	-148,4	22 022,56
4.	Łotwa	267	-174,4	30 415,36
5.	Malta	292	-149,4	22 320,36
6.	Polska	249	-192,4	37 017,76
7.	Republika Czeska	677	235,6	55 507,36
8.	Słowacja	399	-42,4	1 797,76
9.	Słowenia	758	316,6	100 235,56
10.	Węgry	488	46,6	2 171,56
Razem		4414	×	281 854,40

Jest to szereg szczegółowy. Wzory do obliczenia:

$$\bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i = \frac{1}{10} \cdot 4414 = 441{,}4$$

$$S^2(x) = \frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})^2 = \frac{1}{10} \cdot 281854{,}4 = 28185{,}44$$

$$S = \sqrt{S^2} = \sqrt{28185{,}44} = 167{,}9$$

Liczba abonentów telefonii komórkowej na 1000 ludności w krajach wstępujących do UE różni się od średniej arytmetycznej przeciętnie o 167,9.

Odchylenie przeciętne (d) jest średnią arytmetyczną wartości bezwzględnych (modułów) odchyleń wartości od jej średniej arytmetycznej:

$$d = \frac{1}{n}\sum_{i=1}^{n}\left|x_i - \bar{x}\right| \qquad \text{dla szeregu szczegółowego}$$

$$d = \frac{1}{n}\sum_{i=1}^{k}\left|x_i - \bar{x}\right| \cdot n_i \qquad \text{dla szeregu rozdzielczego}$$

Przykład 2.3.2

Liczba braków wyprodukowanych przez pewną brygadę robotników w ciągu tygodnia jest następująca:

Dzień tygodnia	Liczba braków (szt.)
Poniedziałek	4
Wtorek	6
Środa	5
Czwartek	7
Piątek	6

Wyznaczyć odchylenie przeciętne.

Rozwiązanie

Obliczamy średnią arytmetyczną prostą:

$$\bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i = \frac{4+6+5+7+6}{5} = \frac{28}{5} = 5{,}6$$

Tworzymy tablicę absolutnych różnic

x_i	$\lvert x_i - \bar{x}\rvert$
4	1,6
6	0,4
5	0,6
7	1,4
6	0,4
Suma	4,4

$$d = \frac{\sum \lvert x_i - \bar{x}\rvert}{n} = \frac{4{,}4}{5} = 0{,}88$$

Wartości poszczególnych obserwacji odchylają się średnio od średniej arytmetycznej o około 1 sztukę.

Jeżeli istnieje potrzeba dokonania porównań kilku zbiorowości ze względu na zmienność do oceny dyspersji, stosuje się ***współczynnik zmienności*** V_s.

Współczynnik zmienności jest to względna miara dyspersji, wyrażona w procentach w następujący sposób:

$$V_s = \frac{S}{\bar{x}} \cdot 100\% \qquad \text{lub rzadziej} \qquad V_s = \frac{d}{\bar{x}} \cdot 100\%$$

Współczynnik zmienności jest ilorazem bezwzględnej miary dyspersji i odpowiednich wartości średnich. Pozwala ocenić natężenie zróżnicowania badanej cechy w zbiorowości. Jego wartość bliska zero świadczy o tym, że

badana zbiorowość jest jednorodna, a im bardziej zróżnicowana jest zbiorowość, tym większy jest współczynnik zmienności.

Przykład 2.3.3

Analizując liczbę wyprodukowanych sztuk detali pewnej brygady zanotowano dane, które przedstawia szereg rozdzielczy przedziałowy:

Liczba detali (x_i)	12 – 14	14 – 16	16 – 18	18 – 20
Liczba pracowników (n_i)	6	7	11	6

Obliczyć odchylenie standardowe, współczynnik zmienności V_s i określić typowy przedział zmienności.

Rozwiązanie

x_i	n_i	$\dot{x}_i$	$\dot{x}_i \cdot n_i$	$\dot{x}_i - \bar{x}$	$(\dot{x}_i - \bar{x})^2$	$(\dot{x}_i - \bar{x})^2 \cdot n_i$
12 – 14	6	13	78	–3,1	9,61	57,66
14 – 16	7	15	105	–1,1	1,21	8,47
16 – 18	11	17	187	0,9	0,81	8,91
18 – 20	6	19	114	2,9	8,41	50,46
Razem	30	x	484	x	x	125,5

$$\bar{x} = \frac{1}{n}\sum_{i=1}^{k} x_i n_i = \frac{1}{30} \cdot 484 = 16,1 \text{ sztuk}$$

$$S^2 = \frac{1}{n}\sum_{i} (\dot{x}_i - \bar{x})^2 n_i = \frac{1}{30} \cdot 125,5 = 4,2$$

$$S = \sqrt{S^2} = \sqrt{4,2} = 2,0 \text{ sztuki}$$

Typowy przedział zmienności: $\bar{x} - S < x_{typ} < \bar{x} + S$

$$16,1 - 2,0 < x_{typ} < 16,1 + 2,0$$

$$14,1 < x_{typ} < 18,1$$

Współczynnik zmienności:

$$V_s = \frac{S}{\bar{x}} \cdot 100\% = \frac{2,0}{16,1} \cdot 100\% = 12,4\%$$

Liczba wyprodukowanych detali badanej brygady odchyla się od średniej arytmetycznej przeciętnie o 2,0 sztuki, obszar zmienności wynosi (14,1; 18,1), zaś odchylenie standardowe stanowi 12,4% średniej arytmetycznej.

Przykład 2.3.4

Średnie miesięczne wpływy za świadczenie usług noclegowych w trzech losowo wybranych hotelach A, B, C były równe [20]:

$$\bar{x}_A = 600 \text{ zł} \qquad \bar{x}_B = 300 \text{ zł} \qquad \bar{x}_C = 500 \text{ zł}$$

Odchylenia standardowe wartości sprzedanych usług wynosiły:

$$S_A = 110 \text{ zł} \qquad S_B = 90 \text{ zł} \qquad S_C = 120 \text{ zł}$$

W którym hotelu występuje największa dyspersja miesięcznych wpływów za świadczenie usług noclegowych?

Rozwiązanie

Odchylenia standardowe nie mogą być podstawą do wyciągnięcia wniosków o sile dyspersji ze względu na znaczne różnice w średnim poziomie wpływów w poszczególnych hotelach. Należy zastosować współczynnik zmienności.

Dla hotelu A: $V_S = \frac{S}{\bar{x}} \cdot 100\% = \frac{110}{600} \cdot 100\% = 18{,}3\%$

Dla hotelu B: $V_S == \frac{90}{300} \cdot 100\% = 30{,}0\%$

Dla hotelu C: $V_S == \frac{120}{500} \cdot 100\% = 24{,}0\%$

Największe względne zróżnicowanie miesięcznych wpływów za świadczenie usług noclegowych miało miejsce w hotelu B, najmniejsze zaś w hotelu A.

2.3.2. Pozycyjne miary dyspersji

Empiryczny obszar zmienność (rozstęp)(*R*) jest różnicą między największą i najmniejszą wartością cechy.

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

Jest to miara bardzo ogólna. Obszar zmienności możemy określić ściśle dla szeregu szczegółowego i dla szeregu punktowego, a dla przedziałowego podać jedynie przybliżoną wartość. W przypadku otwartych przedziałów klasowych nawet przybliżone określenie obszaru zmienności jest niemożliwe. Rozstęp oblicza się w celu wstępnej orientacji o zmienności badanej cechy.

Odchylenie ćwiartkowe (Q) opiera się na wartościach Q_1 i Q_3:

$$Q = \frac{Q_3 - Q_1}{2}$$

Interpretuje się go jako połowę obszaru zmienności środkowych 50% jednostek zbiorowości. Jest to miara bezwzględna.

Typowy obszar zmienności za pomocą miar pozycyjnych możemy określić następująco:

$$M_e - Q < x_{typ} < M_e + Q$$

Współczynnik zmienności zdefiniowany za pomocą miar pozycyjnych wyraża wzór:

$$V_Q = \frac{Q}{M_e} \cdot 100\% \text{ (miara względna)}$$

Pomiędzy odchyleniami: ćwiartkowym, przeciętnym i standardowym obliczonych do tego samego szeregu zachodzi relacja:

$$Q < d < S$$

Przykład 2.3.5

Wiek wybranej grupy przedszkolaków (w latach) jest następujący: 3,1; 3,2; 3,5; 3,6; 4,1; 5,0; 3,0; 2,9; 3,4; 2,8. Określić empiryczny obszar zmienności:

Rozwiązanie

$R = x_{max} - x_{min} = 5,0 - 2,8 = 2,2$ *lat*

Przykład 2.3.6

Rozkład gospodarstw domowych (dane umowne) według liczby osób w gospodarstwie przedstawia tablica, w której zawarto również liczebności skumulowane.

Tablica obliczeniowa

	Liczba osób w gospodarstwie	Liczba gospodarstw (w tys.)	Liczebności skumulowane	
	1	2 188	2 188	
→	2	2 673	4 861	przedział Q_1
→	3	2 427	7 288	przedział Q_2
→	4	2 632	9 920	przedział Q_3
	5	1 171	11 091	
	6	514	11 605	
	7 i więcej	365	11 970	
	Ogółem	11 970	×	

Obliczyć odchylenie ćwiartkowe i współczynnik zmienności, typowy obszar zmienności.

Rozwiązanie

Pozycja $Q_1 = \frac{n}{4} = 2992{,}5$ mieści się w przedziale drugim, czyli wartość $Q_1 = 2$ osoby, tzn. 25% gospodarstw domowych liczy co najwyżej 2 osoby, 75% gospodarstw liczy co najmniej 2 osoby.

Pozycja $M_e = \frac{n}{2} = 5985$ mieści się w trzecim przedziale, czyli wartość $M_e = 3$ osoby, tzn. 50% gospodarstw domowych liczy co najwyżej 3 osoby, a drugie 50% liczy co najmniej 3 osoby.

Pozycja $Q_3 = \frac{n}{4} = 8977{,}5$ mieści się w przedziale czwartym, czyli wartość $Q_3 = 4$ osoby, tzn. 75% gospodarstw domowych liczy co najwyżej 4 osoby, 25% gospodarstw liczy co najmniej 3 osoby.

Odchylenie ćwiartkowe: $Q = \frac{Q_3 - Q_1}{2} = \frac{4-2}{2} = 1\,osoba$

Współczynnik zmienności: $V_Q = \frac{Q}{M_e} \cdot 100\% = \frac{1}{3} \cdot 100\% = 33{,}3\%$

Średnio biorąc, liczba osób w gospodarstwie różni się od mediany o ±1 osobę. Wartość pozycyjnego współczynnika zmienności informuje, że odchylenie ćwiartkowe stanowi 33,3% mediany.

Typowy obszar zmienności: $M_e - Q < x_{typ} < M_e + Q$, czyli $2 < x_{typ} < 4$.

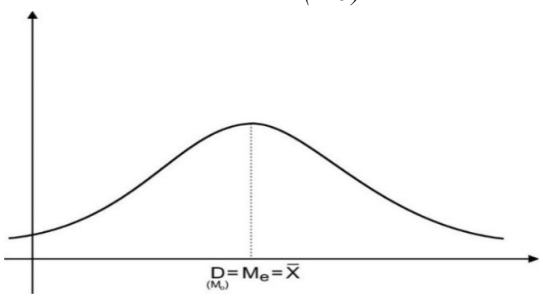
2.4. Miary asymetrii

Miary asymetrii służą do badania skośności rozkładu badanej cechy. Pomiar dotyczy sposobu rozmieszczania liczebności przy wartościach cechy. Oceniając asymetrię zwracamy uwagę na to, w jakim miejscu na osi OX znajduje się punkt skupienia obserwacji, czyli obserwujemy, gdzie znajduje się modalna (dominanta). Jeżeli średnia ta mieści się w środku rozkładu, mamy do czynienia z rozkładem symetrycznym. Jeżeli obserwujemy przesunięcie modalnej (dominanty) w kierunku krańców rozkładu, wtedy rozkład jest asymetryczny. Im większe jest to przesunięcie, tym asymetria jest większa. Jeżeli punkt skupienia znajduje się przy niskich wartościach cechy, to mamy do czynienia z asymetrią dodatnią. Jeśli obserwacje wykazują tendencję do skupiania się przy wyższych wartościach cechy, wtedy

rozkład ma asymetrię ujemną. Kierunek asymetrii można łatwo ocenić na wykresie, który ilustruje obserwacje zamieszczone w szeregu rozdzielczym; wyraźnie widać punkt skupienia. Omówimy trzy przypadki.

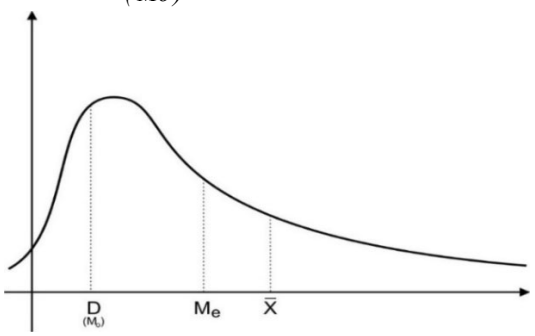
1. ***W szeregu symetrycznym*** relacja między średnimi jest następująca:

$$\bar{x} = M_e = \underset{(Mo)}{D}$$



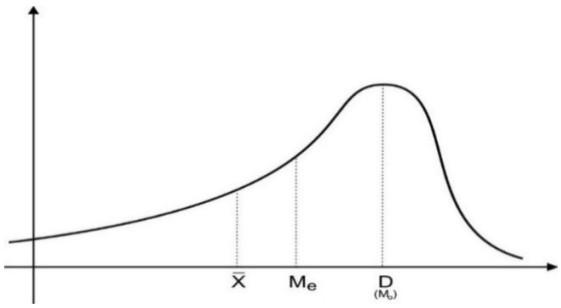
2. W rozkładzie o ***asymetrii dodatniej*** (prawostronnej) obserwacje skupiają się przy wartościach cechy niższych od średniej arytmetycznej. Relacja średnich jest następująca:

$$\underset{(Mo)}{D} < M_e < \bar{x}$$



3. W rozkładzie o ***asymetrii ujemnej*** (lewostronnej) relatywnie liczne są jednostki posiadające wartości cechy wyższe od średniej arytmetycznej. Relacja średnich jest następująca:

$$\bar{x} < M_e < \underset{(Mo)}{D}$$



Oprócz oceny kierunku asymetrii badamy natężenie (siłę) asymetrii. Miarą klasyczną jest współczynnik:

$$A_\mu = \frac{\mu_3}{S^3}$$

gdzie μ_3 jest to trzeci element centralny i wynosi:

$$\mu_3 = \frac{1}{n}\sum_i (x_i - \bar{x})^3 n_i \qquad i = 1, 2, \ldots, k$$

Przyjmuje on wartości (na ogół) z przedziału od –1 do +1. Im bliżej zera, tym asymetria jest mniejsza. Ujemne wartości wskazują na asymetrię lewostronną, zaś dodatnie na asymetrię prawostronną. Wzór ten stosujemy zarówno w analizie szeregów rozdzielczych punktowych, jak i przedziałowych.

Najprostszą miarą asymetrii jest wskaźnik skośności:

$$W_s = \bar{x} - M_o$$

Chociaż jest to miara mało przydatna, określa ona kierunek asymetrii.

Miarą określającą zarówno kierunek, jak i siłę asymetrii jest współczynnik asymetrii (skośności):

$$A_S = \frac{\bar{x} - M_o}{S}$$

Jest to miara niemianowana, unormowana, co umożliwia porównywanie asymetrii różnych rozkładów.

Jeżeli rozkład empiryczny nie spełnia warunków niezbędnych do obliczenia średniej arytmetycznej czy dominanty, wówczas stosujemy współczynnik zbudowany na podstawie kwartyli:

$$A_Q = \frac{(Q_3 - M_e) - (M_e - Q_1)}{Q_3 - Q_1} = \frac{Q_3 + Q_1 - 2M_e}{2Q}$$

Przyjmuje wartości (na ogół) od –1 do +1.

Uwagi:

1. Współczynniki asymetrii przy bardzo silnej asymetrii mogą przekroczyć wartość ±1.
2. Każda z tych miar jest skonstruowana na innych zasadach i dlatego mogą być różne wyniki. Do interpretacji należy podchodzić ostrożnie. Znak przy wyraźnej asymetrii jest tak sam, ale wartości bezwzględne są na ogół różne.
3. Dla rozkładu symetrycznego $A_S = A_Q = A_\mu = 0$ .Im większa jest wartość bezwzględna, tym silniejsza asymetria.
4. W przypadku rozkładu umiarkowanie asymetrycznego zachodzi równość: $\bar{x} - D = 3(\bar{x} - M_e)$, (wzór Pearsona).

Przykład 2.4.1

Odchylenie standardowe kosztu jednostkowego produkcji wyrobu wynosi 3 zł, a najczęściej spotykany koszt to 43 zł. Rozkład jednostkowych kosztów produkcji wyrobów jest prawostronnie asymetryczny i wynosi 1. Obliczyć przeciętny koszt jednostkowy wyrobu.

Rozwiązanie

Obliczamy współczynnik asymetrii:

$$A_s = \frac{\bar{x} - M_o}{S}$$

$$1 = \frac{\bar{x} - 43}{3}$$

stąd $\bar{x} = 46$.

Przeciętny koszt jednostkowy wyrobu wynosi 46 zł.

Przykład 2.4.2

Analizując zbiorowość mieszkań w pewnym mieście rozpatrywaną według pomieszczeń w m^2, otrzymano parametry: $\bar{x} = 84\,m^2$, $M_o = 76{,}2\ m^2$, $S = 28{,}5\ m^2$. Obliczyć współczynnik asymetrii.

Rozwiązanie

$$A_s = \frac{\bar{x} - M_o}{S} = \frac{84 - 76{,}2}{28{,}5} = 0{,}27$$

Współczynnik wskazuje na dodatnią (prawostronną) umiarkowaną asymetrię rozkładu powierzchni mieszkań.

Przykład 2.4.3

Sprawdź, czy posiadając informacje o średniej arytmetycznej równej 55, modalnej równej 53 oraz klasycznym współczynniku zmienności wynoszącym 25%, prawidłowo obliczono wartość klasyczno-pozycyjnego współczynnika asymetrii równą 0,5?

Rozwiązanie

Obliczamy współczynnik asymetrii

$$A_s = \frac{\bar{x} - M_o}{S}$$

$$V_s = \frac{S}{\bar{x}} \cdot 100\%$$

$$25\% = \frac{S}{55} \cdot 100\%$$

$$S = 0{,}25 \cdot 55 = 13{,}75$$

$$A_s = \frac{55 - 53}{13{,}75} = 0{,}145$$

Wartość klasyczno-pozycyjnego współczynnika asymetrii obliczono nieprawidłowo, ponieważ $A_s = 0{,}145$, a nie 0,5, jak podano w treści zadania.

Przykład 2.4.4

Odchylenie standardowe kosztu jednostkowego produkcji wyrobu wynosi 3 zł, a najczęściej spotykany koszt 43 zł. Rozkład jednostkowych kosztów produkcji wyrobów jest prawostronnie asymetryczny i wynosi 1. Obliczyć przeciętny koszt jednostkowy wyrobu.

Rozwiązanie

Obliczamy współczynnik asymetrii

$$A_s = \frac{\bar{x} - M_o}{S}$$

$$1 = \frac{\bar{x} - 43}{3}$$

stąd $\bar{x} = 46$.

Przeciętny koszt jednostkowy wyrobu wynosi 46.

Przykład 2.4.5

Analizując zbiorowość mieszkań w pewnym mieście rozpatrywaną według pomieszczeń w m^2 otrzymano parametry: $\bar{x} = 84\ m^2$, $M_o = 76{,}2\ m^2$, $S = 28{,}5\ m^2$. Obliczyć współczynnik asymetrii.

Rozwiązanie

$$A_S = \frac{\bar{x} - M_o}{S} = \frac{84 - 76{,}2}{28{,}5} = 0{,}27$$

Współczynnik wskazuje na dodatnią (prawostronną) umiarkowaną asymetrię rozkładu powierzchni mieszkań.

Przykład 2.4.6

W ramach przeprowadzonego badania ludności zapytano o wykonywanie pracy dodatkowej poza podstawowym miejscem pracy. Dla 100 losowo wybranych osób otrzymano następujące parametry rozkładu według wieku: M_o = 38,7 lat, V_S = 25,6%, A_S = 0,04. Określić przeciętny wiek badanych osób.

Rozwiązanie

Przeciętny wiek możemy określić stosując współczynnik asymetrii:

$$A_S = \frac{\bar{x} - M_o}{S}$$

Aby obliczyć ten współczynnik należy rozwiązań układ równań:

$$\begin{cases} 0{,}04 = \dfrac{\bar{x} - 38{,}7}{S} \\ 25{,}6\% = \dfrac{S}{\bar{x}} \cdot 100\% \end{cases}$$

Korzystając z drugiego równania wyznaczmy $S = 0{,}256 \cdot \bar{x}$. Podstawimy obliczone S do pierwszego równania otrzymując:

$$0{,}04 = \frac{\bar{x} - 38{,}7}{0{,}256\,\bar{x}}.$$

Wykonujemy przekształcenia:

$$0{,}04 \cdot 0{,}256\bar{x} = \bar{x} - 38{,}7$$
$$0{,}01\bar{x} = \bar{x} - 38{,}7$$
$$0{,}01\bar{x} - \bar{x} = -38{,}7$$
$$-0{,}99\bar{x} = -38{,}7$$
$$\bar{x} \approx 39$$

Przeciętny wiek badanych osób wynosi około 39 lat.

Przykład 2.4.7

Zapytano 30 studentów o średnią ocen ze statystyki i otrzymano informacje:

Średnia ocen x_i	< 3,0	3,0–3,25	3,25–3,5	3,5–3,75	3,75–4,0	4,5–5,0
Liczba studentów n_i	4	10	8	4	2	2

Zbadać asymetrię rozkładu ocen studentów.

Rozwiązanie

Zauważamy brak ocen w przedziale 4,0 – 4,5 i otwarty pierwszy przedział klasowy, dlatego zastosujemy parametry pozycyjne (obserwacje są nietypowe). W oparciu o podany szereg wyznaczymy kwartyle.

Tablica obliczeniowa

x_i	n_i	Liczebność skumulowana	
poniżej 3,0	4	4	
3,0 – 3,25	10	14	Q_1
3,25 – 3,5	8	22	Q_2
3,5 – 3,75	4	26	Q_3
3,75 – 4,0	2	28	
4,0 – 5,0	2	30	

$$N_{Q1} = \frac{30}{4} = 7{,}5; \quad N_{Q2} = \frac{30}{2} = 15; \quad N_{Q3} = 22{,}5$$

$$Q_1 = x_{02} + \frac{N_{01} - \sum_{i-1}^{2-1} n_i}{n_2} \cdot h_2 = 3 + \frac{7{,}5 - 4}{10} \cdot 0{,}25 = 3{,}09$$

$$Q_2 = x_{03} + \frac{N_{02} - \sum_{i-1}^{3-1} n_i}{n_3} \cdot h_3 = 3{,}25 + \frac{15 - 14}{8} \cdot 0{,}25 = 3{,}28$$

$$Q_3 = x_{04} + \frac{N_{03} - \sum_{i-1}^{4-1} n_i}{n_4} \cdot h_4 = 3{,}5 + \frac{22{,}5 - 22}{4} \cdot 0{,}25 = 3{,}53$$

Obliczamy pozycyjny współczynnik asymetrii:

$$A_Q = \frac{(Q_3 - Q_2) - (Q_2 - Q_1)}{Q_3 - Q_1} = \frac{(3{,}53 - 3{,}28) - (3{,}28 - 3{,}09)}{3{,}53 - 3{,}09} = 0{,}14$$

Wynik oznacza bardzo niewielką asymetrię dodatnią.

2.5. Miary koncentracji

Spłaszczenie rozkładu, zwane kurtozą, wynika ze stopnia skupienia obserwacji, czyli koncentracji wokół wartości średniej arytmetycznej. W związku z tym, ocena stopnia spłaszczenia odnosi się do szeregów symetrycznych lub umiarkowanie symetrycznych (charakteryzują się tendencją centralną). Kształt krzywej liczebności zależy od tego, jak duże jest to skupienie. Gdy skupienie jest silne, otrzymamy rozkład wysmukły, zaś przy słabym będzie to rozkład spłaszczony.

Koncentracja zbiorowości wokół wartości średniej jest związana z rozproszeniem wartości cechy. Jednak zdarzają się sytuacje, gdy dwa szeregi o podobnym odchyleniu standardowym różnią się pod względem koncentracji. Klasyczną miara spłaszczenia jest moment czwarty centralny dany wzorem:

$$\mu_4 = \frac{1}{n}\sum_i (x_i - \bar{x})^4 \cdot n_i \qquad \text{i} = 1, 2, \ldots, \text{k}$$

Po podzieleniu μ_4 przez odchylenie standardowe podniesione do potęgi czwartej otrzymujemy wzór, służący do oceny koncentracji rozkładu wokół średniej. A zatem, miarą natężenia koncentracji poszczególnych wartości cechy wokół średniej arytmetycznej jest współczynnik koncentracji, który można obliczyć następująco:

$$K = \frac{\mu_4}{S^4}$$

gdzie μ_4 jest to czwarty moment centralny.

Skupienie wartości wokół średniej w znacznym stopniu jest uzależnione od poziomu dyspersji i obszaru zmienności cechy. Na ogół przyjmujemy, że jeżeli:

- $K = 3$, to rozkład jest normalny,
- $K > 3$, to rozkład jest wysmukły, o skupieniu silniejszym od normalnego,
- $K < 3$, to rozkład jest spłaszczony, o skupieniu słabszym od normalnego.

Przykład 2.5.1

Badano liczbę błędów w kodzie źródłowym 30 programistów. Otrzymano następujące wyniki: 3, 2, 5, 3, 4, 5, 3, 1, 0, 2, 6, 3, 4, 5, 3, 1, 5, 3, 0, 1, 2, 2, 4, 3, 4, 4, 3, 2, 6, 5.

Wyznaczyć klasyczny współczynnik asymetrii A_μ.

Rozwiązanie

Dane prezentujemy w statystycznym szeregu punktowym:

Liczba błędów	0	1	2	3	4	5	6
Liczebność	2	3	5	8	5	5	2

Tablica obliczeniowa jest następująca:

x_i	n_i	$x_i \cdot n_i$	$x_i - \bar{x}$	$(x_i - \bar{x})^2$	$(x_i - \bar{x})^2 \cdot n_i$	$(x_i - \bar{x})^3$	$(x_i - \bar{x})^3 \cdot n_i$
0	2	0	-3,1	9,61	0	-29,791	-59,852
1	3	3	-2,1	4,41	13,23	-9,261	-27,783
2	5	10	-1,1	1,21	6,05	-1,331	-6,655
3	8	24	-0,1	0,01	0,08	-0,001	-0,008
4	5	20	0,9	0,81	4,05	0,729	3,645
5	5	25	1,9	3,61	18,05	6,859	34,295
6	2	12	2,9	8,41	16,82	24,389	48,778
Razem	30	94	×	×	58,28	×	-7,31

$$\bar{x} = \frac{1}{n}\sum_{i=1}^{k} x_i \cdot n_i = \frac{1}{30} \cdot 94 \approx 3{,}1$$

$$S^2 = \frac{1}{n}\sum_{i=1}^{k} (x_i - \bar{x})^2 \cdot n_i = \frac{1}{30} \cdot 58{,}28 = 1{,}94$$

$$S = \sqrt{S^2} = \sqrt{1{,}94} = 1{,}39$$

$$\mu_3 = \frac{1}{n}\sum_{i=1}^{k} (x_i - \bar{x})^3 \cdot n_i = \frac{1}{30} \cdot (-7{,}31) = -0{,}24$$

$$A_\mu = \frac{\mu_3}{S^3} = \frac{-0{,}24}{(1{,}39)^3} \approx -0{,}09$$

Rozkład błędów programistów wykazuje bardzo niewielką asymetrię ujemną, lewostronną.

Przykład 2.5.2

Badanie 56 losowo wybranych czteroosobowych gospodarstw domowych pod względem miesięcznych wydatków w zł na kulturę (teatr, kino, książki) dostarczyło danych, zawartych w tabeli.

Miesięczne wydatki w zł *(x_i)*	20-40	40-60	60-80	80-100	100-120	120-140
Liczba gospodarstw *(n_i)*	1	4	17	25	8	1

Ocenić asymetrię i koncentrację rozkładu.

Rozwiązanie

Do oceny asymetrii zastosujemy współczynnik $A_\mu = \frac{\mu_3}{S^3}$, zaś do oceny koncentracji wykorzystamy współczynnik $K = \frac{\mu_4}{S^4}$.Do wykonania obliczeń zbudowano tablicę pomocniczą.

Tablica obliczeniowa pomocnicza

$[x_i, x_{i+1})$	n_i	$\dot{x}_i$	$\dot{x}_i \cdot n_i$	$\dot{x}_i - \bar{x}$	$(\dot{x}_i - \bar{x})^2$	$(\dot{x}_i - \bar{x})^2 \cdot n_i$	$(\dot{x}_i - \bar{x})^3$	$(\dot{x}_i - \bar{x})^3 \cdot n_i$	$(\dot{x}_i - \bar{x})^4$	$(\dot{x}_i - \bar{x})^4 \cdot n_i$
20-40	1	30	30	-54	2916	2916	-157464	-157464	8503056	8503056
40-60	4	50	200	-34	1156	4624	-39304	-157216	1336336	5345344
60-80	17	70	1190	-14	196	3332	-2744	-46648	38416	653072
80-100	25	90	2250	6	36	900	216	5400	1296	32400
100-120	8	110	880	26	676	5408	17576	140608	456976	3655808
120-140	1	130	130	46	2116	2116	97336	97336	4477456	4477456
Razem	56	×	4680	×	×	19296	×	-117984	×	22667136

Źródło: Obliczenia własne

$$\bar{x} = \frac{1}{n}\sum_{i=1}^{k} \dot{x}_i \cdot n_i = \frac{1}{56} \cdot 4680 = 84 \text{ zł}$$

$$S^2 = \frac{1}{n}(\dot{x}_i - \bar{x})^2 \cdot n_i = \frac{1}{56} \cdot 19296 = 345$$

$$S = \sqrt{S^2} = \sqrt{1345} = 18{,}57 \text{ zł}$$

$$V_S = \frac{S}{\bar{x}} \cdot 100\% = \frac{18{,}57}{84} \cdot 100\% = 22{,}1\%$$

$$\mu_3 = \frac{1}{n}\sum_{i=1}^{k}(x_i - \bar{x})^3 \cdot n_i = \frac{1}{56} \cdot (-117984) = -2106{,}86$$

$$A_\mu = \frac{\mu_3}{S^3} = \frac{-2106{,}86}{(18{,}57)^3} = -0{,}329$$

$$\mu_4 = \sum_{i=1}^{k}(\dot{x}_i - \bar{x})^4 \cdot n_i = \frac{1}{56} \cdot 22667236 = 404770{,}29$$

$$K = \frac{\mu_4}{S^4} = \frac{404770{,}29}{(18{,}57)^4} = 3{,}4$$

Interpretując uzyskane wyniki możemy stwierdzić, że:

- średnia wartość wydatków na kulturę w badanej grupie wynosi 84 zł.
- dyspersja wyników w stosunku do średniej arytmetycznej wynosiła przeciętnie biorąc ± 18,57 zł.,
- odchylenie standardowe stanowiło 22,1% średniej arytmetycznej,
- współczynnik asymetrii wskazuje na rozkład lewostronny, umiarkowany, co oznacza, że w badanej grupie przeważają kieszonkowe wyższe od średniego,
- rozkład jest wysmukły, o skupieniu silniejszym od normalnego.

Przykład 2.5.3

W 100-osobowej grupie losowo wybranych studentów przeprowadzono test sprawnościowy, który oceniono punktami. Wyniki testu przedstawia tabela (dane umowne).

Liczba punktów (x_i)	0-20	20-40	40-60	60-80	80-100
Liczba studentów (n_i)	4	12	25	35	24

Ocenić asymetrię i koncentrację rozkładu.

Rozwiązanie

Do oceny asymetrii zastosujemy współczynnik $A_\mu = \frac{\mu_3}{S^3}$, zaś do oceny koncentracji wykorzystamy współczynnik $K = \frac{\mu_4}{S^4}$. Wszelkie niezbędne obliczenia zawiera tablica pomocnicza.

Tablica obliczeniowa pomocnicza

Liczba Punktów x_i	Liczba studentów n_i	$\dot{x}_i$	$\dot{x}_i \cdot n_i$	$\dot{x}_i - \bar{x}$	$(\dot{x}_i - \bar{x})^2$	$(\dot{x}_i - \bar{x})^2 \cdot n_i$	$(\dot{x}_i - \bar{x})^3$	$(\dot{x}_i - \bar{x})^3 \cdot n_i$	$(\dot{x}_i - \bar{x})^4$	$(\dot{x}_i - \bar{x})^4 \cdot n_i$
0 – 20	4	10	40	-52,6	2766,76	11067,04	-145531,58	-582126,32	7654960,90	30619843,59
20 – 40	12	30	360	-32,6	1062,76	12753,12	-34645,98	-415751,76	1129458,82	13553505,81
40 – 60	25	50	1250	-12,6	158,76	3969,00	-2000,38	-50009,50	25204,74	630118,44
60 –80	35	70	2450	7,4	54,76	1916,60	405,22	14182,70	2998,66	104953,02
80 – 100	24	90	2160	27,4	750,76	18018,24	20570,82	493699,68	563640,58	13527373,86
Razem	100	×	6260	×	×	47724,00	×	-540005,20	×	58435794,72

$$\bar{x} = \frac{1}{n}\sum_i x_i \cdot n_i = \frac{1}{100} \cdot 6260 = 62{,}6\ pkt.$$

$$S^2 = \frac{1}{n}\sum_i (x_i - \bar{x})^2 \cdot n_i = \frac{1}{100} \cdot 47724\ \ pkt.$$

$$S = \sqrt{S^2} = 21{,}8\ pkt.$$

$$V_S = \frac{S}{\bar{x}} \cdot 100\% = \frac{21{,}8}{62{,}6} \cdot 100\% = 34{,}8\%$$

$$\mu_3 = \frac{1}{100} \cdot (-540005{,}2) = -5400{,}05$$

$$A_S = \frac{u_3}{S^3} = \frac{-5400{,}05}{10360{,}23} = -0{,}52$$

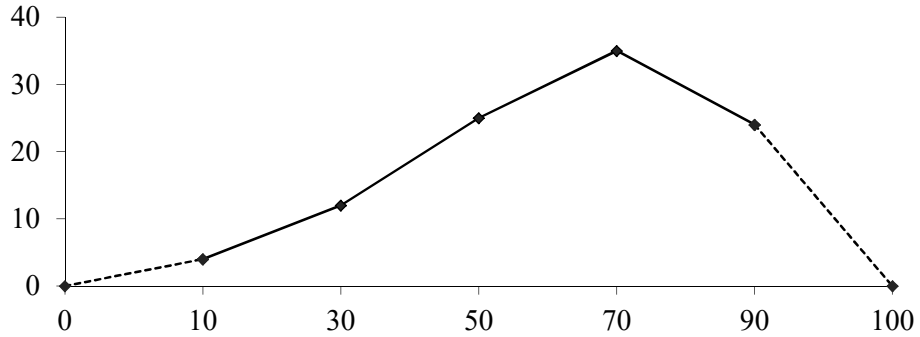
$$\mu_4 = \frac{1}{100} \cdot 58435794{,}72 = 584357{,}95$$

$$K = \frac{u_4}{S^4} = \frac{584357{,}95}{225853{,}0} = 2{,}59$$

Interpretując uzyskane wyniki możemy stwierdzić, że:

- średnia liczba punktów uzyskanych przez badaną grupę studentów wynosiła 62,6,
- dyspersja wyników w stosunku do średniej arytmetycznej wynosiła przeciętnie ±21,8 pkt.,
- odchylenie standardowe stanowiło 34,8% średniej arytmetycznej,
- współczynnik asymetrii wskazuje na rozkład lewostronny, umiarkowany, co oznacza, że wśród otrzymanych ocen przeważały oceny wyższe niż średnia,
- rozkład uzyskanych wyników ma koncentrację słabszą od normalnej.

Wykres na rysunku 2.3 pokazuje wykaz studentów wg uzyskanych ocen z testu sprawnościowego.



Wykres 2.3. Studenci według uzyskanych ocen z testu sprawnościowego
Źródło: Opracowanie własne

2.6. Przykłady praktyczne

Przykład 2.6.1
Dwaj studenci *X, Y* razem przygotowywali się do sesji egzaminacyjnej i obaj chcieli uzyskać stypendium naukowe. Który z nich miał większą szansę zdobycia stypendium, jeżeli pierwszy otrzymał oceny: 3; 3; 4; 5 zaś drugi: 3; 4; 5; 5.

Rozwiązanie
Obliczymy średnią arytmetyczną ocen studentów:

$$\bar{x} = \frac{1}{4}(3+3+4+5) = 3{,}75$$

$$\bar{y} = \frac{1}{4}(3+4+5+5) = 4{,}25$$

Student Y miał większą szansę otrzymania stypendium.

Przykład 2.6.2

Obliczyć wskaźniki struktury eksportu w wybranym województwie oraz średni obrót.

Wyszczególnienie	Eksport (mln zł)	w_i	$x_i \cdot w_i$
Żywność i zwierzęta żywe	3295,5	0,076	250,458
Napoje i tytoń	117,0	0,003	0,351
Surowce mineralne z wyjątkiem paliw	1203,0	0,027	32,481
Paliwa mineralne, smary i materiały pochodne	2160,0	0,050	108,000
Oleje, tłuszcze, woski zwierzęce i roślinne	17,7	0,000	0,000
Chemikalia i produkty pokrewne	2827,4	0,065	183,781
Towary przemysłowe sklasyfikowane głównie według surowca	10746,1	0,247	2654,287
Maszyny, urządzenia i sprzęt transportowy	15142,5	0,347	5254,448
Różne wyroby przemysłowe	8047,1	0,185	1488,714
Razem	43556,3	1,000	9972,519

Źródło: Dane umowne

Rozwiązanie

Jeżeli w miejscu liczebności n_i występują wskaźniki struktury w_i, to średnia arytmetyczna wyraża się wzorem:

$$\overline{x} = \sum_{i=1}^{k} x_i w_i$$

Średnie obroty eksportu w wybranym do analizy okresie w badanym województwie wynosiły 997 2,52 mln zł.

Przykład 2.6.3

Analizując kwartalne wydatki na reklamę (w tys. zł) w 100 zakładach usługowych otrzymano następujący szereg rozdzielczy przedziałowy:

Wydatki na reklamę(w tys. zł)	0 – 5	5 – 10	10 – 15	15 – 20	20 – 25
Liczba zakładów	10	20	40	25	5

Wyznaczyć odchylenie standardowe, współczynnik zmienności V_Si określić typowy przedział zmienności.

Rozwiązanie

Tablica obliczeniowa

$[x_i, x_{i+1})$	n_i	$\dot{x}_i$	$\dot{x}_i \cdot n_i$	$\dot{x}_i - \bar{x}$	$(\dot{x}_i - \bar{x})^2$	$(\dot{x}_i - \bar{x})^2 \cdot n_i$
0 – 5	10	2,5	25	-9,8	96,04	960,4
5 – 10	20	7,5	150	-4,8	23,4	460,8
10 – 15	40	12,5	500	0,2	0,04	1,6
15 – 20	25	17,5	437,5	5,2	27,4	676,0
20 – 25	5	22,5	112,5	10,2	104,04	520,2
Razem	100	×	1225	×	×	2619,0

$$\bar{x} = \frac{1}{n}\sum_i x_i \cdot n_i = \frac{1}{100} \cdot 1225 = 12{,}25 \approx 12{,}3 \text{ tys. zł.}$$

$$S^2 = \frac{1}{n}\sum_i (x_i - \bar{x})^2 \cdot n_i = \frac{1}{100} \cdot 2619 = 26{,}19 \approx 26{,}2$$

$$S = \sqrt{S^2} = \sqrt{26{,}2} = 5 \text{ tys. zł}$$

Typowy przedział zmienności:

$$\bar{x} - S < x_{typ} < \bar{x} + S$$

$$12{,}3 - 5 < x_{typ} < 12{,}3 + 5$$

$$7{,}3 < x_{typ} < 17{,}3$$

Współczynnik zmienności:

$$V = \frac{S}{\bar{x}} \cdot 100\% = \frac{5}{12{,}4} \cdot 100\% = 40{,}7\%$$

Kwartalne wydatki na reklamę odchylają się od średniej arytmetycznej przeciętnie o 5 tys. złotych, obszar zmienności jest (7,3; 17,3), zaś odchylenie standardowe stanowi 40,7% średniej arytmetycznej.

Przykład 2.6.4

W pewnej grupie 36 palaczy uzyskano następujące dane o liczbie wypalanych dziennie papierosów:

15	9	14	15	18	17
14	12	16	14	12	12
12	11	15	14	15	9
15	15	15	14	12	10
12	9	7	10	7	14
9	14	14	12	10	11

Wyznaczyć dominującą liczbę wypalanych dziennie papierosów w tej grupie.

Rozwiązanie

Należy zbudować szereg rozdzielczy punktowy.

Liczba wypalanych papierosów	Liczba osób	
7	2	
9	4	
10	3	
11	2	
12	7	
→ 14	8	M_o
15	7	

Patrząc na liczebności zauważamy, że wartość najwyższa jest 8, a zatem dominującą liczbą wypalanych dziennie papierosów jest 14.

Przykład 2.6.5

Miesięczna wartość stypendium socjalnego w pewnej grupie studentów kształtowała się następująco:

Lp.	Wartość stypendium (w zł)	Liczba osób	Liczebność skumulowana
1	<800, 1000)	5	5
2	<1000, 1200)	10	15
3	<1200, 1400)	20	35
4	<1400, 1600)	15	50
5	<1600, 1800)	10	60
Razem		60	×

Obliczyć kwartyle i modalną.

Rozwiązanie

Po skumulowaniu liczebności wyznaczamy pozycje poszczególnych kwartyli:

$$N_{Q1} = \frac{n}{4} = \frac{60}{4} = 15 \text{ – pozycja kwartyla pierwszego}$$

$$N_{Q2} = \frac{n}{2} = \frac{60}{2} = 30 \text{ – pozycja kwartyla drugiego}$$

$$N_{Q3} = \frac{3n}{4} = \frac{3 \cdot 60}{4} = 45 \text{ – pozycja kwartyla trzeciego}$$

Odszukujemy w liczebnościach skumulowanych przedziały, w których wyznaczamy kwartyle:

$$Q_1 = x_{02} + \frac{N_{Q1} - \sum_{i=1}^{2-1} n_i}{4} \cdot h_2 = 1000 + \frac{15-5}{10} \cdot (1200-1000) = 1200\,zł$$

$$Q_2 = x_{03} + \frac{N_{Q2} - \sum_{i=1}^{3-1} n_i}{4} \cdot h_3 = 1200 + \frac{30-15}{20} \cdot (1400-1200) = 1350\,zł$$

$$Q_3 = x_{04} + \frac{N_{Q3} - \sum_{i=1}^{4-1} n_i}{4} \cdot h_4 = 1400 + \frac{45-35}{15} \cdot (1600-1400) = 1533{,}33\,zł$$

Otrzymane wyniki oznaczają, że 25% studentów otrzymało stypendium socjalne nie większe niż 1200 zł, a pozostałe 75% nie mniejsze niż 1200 zł. Połowa studentów otrzymała nie więcej niż 1350 zł, a druga połowa nie mniej niż 1350zł. Natomiast 75% studentów otrzymało nie więcej niż 1533,33 zł, a pozostałe 25% nie mniej niż 1533,33 zł.

W tym szeregu modalna występuje w trzeciej klasie, ponieważ liczebność w tym przedziale jest największa. Zatem:

$$M_o = x_{03} + \frac{n_3 - n_2}{(n_3 - n_2) + (n_3 - n_4)} \cdot h_3 = 1200 + \frac{20-10}{(20-10)+(20-15)} \cdot 200 =$$

$$= 1200 + 133{,}33 = 1333{,}33\,zł$$

Interpolacyjnie wyznaczona modalna wynosi 1333,33 zł, co oznacza dominującą wartość stypendium socjalnego w tej grupie studentów.

Przykład 2.6.6

Liczba sklepów w sektorze prywatnym w gminach przygranicznych pewnego województwa była następująca: 25, 183, 19, 141, 40, 47, 80, 29, 214, 82. Na podstawie danych obliczyć średnią, medianę, odchylenie standardowe, przeciętne odchylenie od średniej.

Rozwiązanie

Średnią arytmetyczną obliczamy ze wzoru

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

Zatem mamy:

$$\bar{x} = \frac{25 + 183 + 19 + 141 + 40 + 47 + 80 + 29 + 214 + 82}{10} = \frac{860}{10} = 86$$

W celu wyznaczenia mediany należy uporządkować dane:

19, 25, 29, 40, 47, 80, 82, 141, 183, 214.

Ponieważ jest ich parzysta liczba (n=10), korzystamy ze wzoru

$$M_e = \frac{x_{\frac{n}{2}} + x_{\frac{n}{2}+1}}{2}, \text{ czyli}$$

$$M_e = \frac{x_5 + x_6}{2} = \frac{47 + 80}{2} = 63{,}5$$

Przed obliczeniem odchylenia standardowego obliczamy wariancję:

$$S^2 = \frac{1}{n}\sum_i (x_i - \bar{x})^2$$

$$S^2 = \frac{1}{10}[(19-86)^2 + (25-86)^2 + (29-86)^2 + (40-86)^2 + (47-86)^2 + (80-86)^2 +$$
$$+(82-86)^2 + (141-86)^2 + (183-86)^2 + (214-86)^2] =$$

$$= \frac{1}{10}[4489 + 3721 + 3249 + 2116 + 1521 + 36 + 16 + 3025 + 9409 + 16384] =$$
$$= \frac{1}{10} \cdot 43966 = 4396{,}6$$

Zatem odchylenie standardowe $S = \sqrt{4396{,}6} = 66{,}31$.

Pozostaje obliczyć przeciętne odchylenie od średniej $d = \frac{1}{n}\sum_i |x_i - \bar{x}|$.

$$d = \frac{1}{10}[|-67| + |-61| + |-57| + |-46| + |-39| + |-6| + |-4| + |55| + |97| + |128|]$$
$$|-67| + |-61| + |-57| + |-46| + |-39| + |-6|$$

Przykład 2.6.7

Oceny końcowe z matematyki w I semestrze siódmej klasy przedstawia tablica.

Ocena w I semestrze	Liczba dzieci
niedostateczna	423
dopuszczająca	5596
dostateczna	7595
dobra	5587
bardzo dobra	391
celująca	37

Obliczyć odchylenie ćwiartkowe, typowy obszar zmienności i współczynniki zmienności V_Q.

Rozwiązanie

Tablica obliczeniowa

Ocena w I semestrze	Liczba dzieci	Liczebności skumulowane
1	423	423
2	5596	6019
3	7595	13614
4	5587	19201
5	391	19592
6	37	19629
Razem	19629	×

Pozycja kwartyla pierwszego $N_{Q1} = \frac{n}{4} = \frac{19629}{4} = 4907,25$ mieści się w przedziale drugim, zatem $Q_1 = 2$, tzn. 25% dzieci otrzymało co najwyżej ocenę dopuszczającą, a 75% dzieci otrzymało co najmniej ocenę dopuszczającą.

Pozycja kwartyla drugiego $N_{Q2} = \frac{n}{2} = \frac{19629}{2} = 9814,5$ mieści się w trzecim przedziale, zatem $Q_2 = 3$, tzn. 50% dzieci otrzymało co najwyżej ocenę dostateczną, a 75% dzieci otrzymało co najmniej ocenę dostateczną.

Pozycja kwartyla trzeciego $N_{Q3} = \frac{3n}{4} = \frac{3 \cdot 19629}{4} = 14721,75$ mieści się w czwartym przedziale, zatem $Q_3 = 4$, tzn. 75% dzieci otrzymało co najwyżej ocenę dobrą, a 25% dzieci otrzymało co najmniej ocenę dobrą.

Odchylenie ćwiartkowe $Q = \frac{Q_3 - Q_1}{2} = 1$

Typowy przedział zmienności:

$$M_e - Q < x_{typ} < M_e + Q$$
$$3 - 1 < x_{typ} < 3 + 1$$
$$2 < x_{typ} < 4$$

Współczynnik zmienności zdefiniowany za pomocą miar pozycyjnych wyraża się wzorem:

$$V_Q = \frac{Q}{M_e} \cdot 100\% = \frac{1}{3} \cdot 100\% = 33,3\%$$

Wartość pozycyjnego współczynnika zmienności informuje, że odchylenie ćwiartkowe stanowi 33,3% mediany.

Przykład 2.6.8

W zakładzie przeprowadzono kontrolę jakości żarówek, otrzymane wyniki zapisano w tabeli.

Wyniki badania	Liczba żarówek
Żarówki wadliwe	21
Żarówki dobre	479

Wyznacz średnią awaryjność oraz jej dyspersję.

Rozwiązanie

Szereg z cechą jakościową (niemierzalną) jest szczególnym przypadkiem szeregu strukturalnego. Badając zbiorowość ze względu na cechę jakościową możemy przyjąć, że awaryjność przybiera wartość 1, gdy żarówka jest wadliwa, a wartość 0 – gdy jest dobra. Wobec powyższego otrzymamy szereg szczegółowy ważony.

Wartość badanej cechy x_i	Liczebność n_i
1	21
0	479

Budujemy tablicę obliczeniową.

x_i	n_i	$x_i \cdot n_i$	$x_i - \bar{x}$	$(x_i - \bar{x})^2$	$(x_i - \bar{x})^2 \cdot n_i$
1	21	21	0,958	0,917764	0,917764
0	479	0	-0,042	0,001764	0,844956
Razem	500	21	×	×	1,76272

$$\bar{x} = \frac{1}{n}\sum_i x_i \cdot n_i = \frac{1}{500} \cdot 21 = 0{,}042$$

$$S^2 = \frac{1}{n}\sum_i (x_i - \bar{x})^2 \cdot n_i = \frac{1}{500} \cdot 1{,}76272 = 0{,}00353$$

$$S = \sqrt{S^2} = \sqrt{0{,}00353} = 0{,}0594$$

$$V_S = \frac{S}{\bar{x}} \cdot 100\% = \frac{0{,}0594}{0{,}042} \cdot 100\% = 141{,}4\%$$

Odchylenie standardowe stanowi 141,4% wartości średniej arytmetycznej badanej zbiorowości.

Przykład 2.6.9

Na podstawie danych zawartych w tablicy obliczyć wybrane parametry i dokonać analizy porównawczej struktury wieku nowożeńców obu płci.

Parametry	Miano	Kobiety	Mężczyźni
Średnia arytmetyczna	lata	24	27
Mediana	lata	22	26
Moda	lata	20	24
Wariancja		16	36

Rozwiązanie

Zauważmy, że ***w grupie kobiet*** mamy:

$$\bar{x} = 24$$
$$M_e = 22$$
$$M_o = 20$$

Ponieważ $\bar{x} > M_e > M_o$ rozkład jest o asymetrii dodatniej (prawostronnej), a obserwacje skupiają się przy wartościach cechy niższych od średniej arytmetycznej.

$$S^2 = 16 \Rightarrow S = 4$$
$$V_S(k) = \frac{S}{\bar{x}} \cdot 100\% = \frac{4}{24} \cdot 100\% = 16{,}7\%$$
$$A_S = \frac{\bar{x} - M_o}{S} = \frac{24 - 20}{4} = 1$$

Współczynnik skośności równy 1 oznacza bardzo silną skośność.

W grupie mężczyzn parametry przedstawiają się następująco:

$$\bar{x} = 27$$
$$M_e = 262$$
$$M_o = 24$$

Ponieważ $\bar{x} > M_e > M_o$ rozkład jest o asymetrii dodatniej (prawostronnej), a obserwacje skupiają się przy wartościach cechy niższych od średniej arytmetycznej.

$$S^2 = 36 \Rightarrow S = 6$$

$$V_S(k) = \frac{S}{\bar{x}} \cdot 100\% = \frac{6}{27} \cdot 100\% = 22{,}2\%$$
$$A_S = \frac{\bar{x} - M_o}{S} = \frac{27 - 24}{6} = 0{,}5$$

Rozkład jest umiarkowanie asymetryczny, bowiem zachodzi równość $\bar{x} - M_o = 3(\bar{x} - M_e)$ zwana wzorem Pearsona.

Ponieważ $V_s(m) > V_s(k)$ wnioskujemy, że znacznie bardziej zróżnicowana wewnętrznie jest grupa mężczyzna.

Przykład 2.6.10

W czasie epidemii koronawirusa wybrano losowo 100 osób i przeprowadzono badania na obecność tej choroby, otrzymując wyniki:

Wynik badania	Liczba osób
pozytywny	60
negatywny	40

Ocenić średnią zachorowalność oraz jej dyspersję.

Rozwiązanie

Szereg z cechą jakościową (niemierzalną) jest szczególnym przypadkiem szeregu strukturalnego. Badając zbiorowość ze względu na cechę jakościową możemy przyjąć, że cecha ta przybiera wartość 1, gdy jednostka posiada tę cechę, a wartość 0 – gdy jej nie posiada. Wobec powyższego otrzymujemy szereg szczegółowy ważony.

Wartość badanej cechy x_i	Liczebność n_i
1	60
0	40

Budujemy tablicę obliczeniową.

x_i	n_i	$x_i \cdot n_i$	$x_i - \bar{x}$	$(x_i - \bar{x})^2$	$(x_i - \bar{x})^2 \cdot n_i$
1	60	60	0,4	0,16	9,6
0	40	0	-0,6	0,36	14,4
Razem	100	60	×	×	24,0

$$\bar{x} = \frac{1}{n}\sum_i x_i \cdot n_i = \frac{1}{100} \cdot 60 = 0{,}6$$

$$S^2 = \frac{1}{n}\sum_i (x_i - \bar{x})^2 \cdot n_i = \frac{1}{100} \cdot 24 = 0{,}24$$

$$S = \sqrt{S^2} = \sqrt{0{,}24} = 0{,}49$$

$$V_S = \frac{S}{\bar{x}} \cdot 100\% = \frac{0{,}49}{0{,}6} \cdot 100\% = 81{,}7\%$$

Średnią wartością cechy jakościowej jest częstość względna, którą można zinterpretować jako częstość występowania cechy jakościowej w tej zbiorowości, w tym przypadku chodzi o częstość zachorowań. Odchylenie standardowe stanowi 81,7% wartości średniej arytmetycznej badanej zbiorowości.

Przykład 2.6.11

W grupie 100 losowo wybranych dzieci zapytano o wartość dziennego kieszonkowego. Na podstawie uzyskanych informacji sporządzono tabelę.

Wartość dziennego kieszonkowego w zł x_i	Liczba dzieci n_i
0-6	10
7-13	20
14-20	25
21-27	35
28-34	10

Przeprowadzić analizę struktury z wykorzystaniem momentów rozkładu.

Rozwiązanie

Metodę momentów stosuje się głównie dla szeregów rozdzielczych, gdy badany szereg statystyczny ma równe i domknięte przedziały losowe. Ponieważ warunki te analizowany szereg spełnia, możemy zbudować tablicę obliczeniową i wyznaczyć momenty zwykłe.

Tablica obliczeniowa

x_i	n_i	$\dot{x}_i$	$\dot{x}_i^2$	$\dot{x}_i^3$	$\dot{x}_i^4$	$\dot{x}_i \cdot n_i$	$\dot{x}_i^2 \cdot n_i$	$\dot{x}_i^3 \cdot n_i$	$\dot{x}_i^4 \cdot n_i$
0-6	10	3	9	27	81	30	90	270	810
7-13	20	10	100	1000	10000	200	2000	20000	200000
14-20	25	17	289	4913	83521	425	7225	122825	2088025
21-27	35	24	576	13824	331776	840	20160	483840	11612160
28-34	10	31	961	29791	923521	310	9610	297910	9235210
Razem	100	×	×	×	×	1805	39085	924845	23136205

$$m_1 = \frac{1}{n}\sum_{i=1}^{k} \dot{x}_i \cdot n_i = \frac{1}{100} \cdot 1805 = 18{,}05$$

$$m_2 = \frac{1}{n}\sum_{i=1}^{k} \dot{x}_i^2 \cdot n_i = \frac{1}{100} \cdot 39085 = 390{,}85$$

$$m_3 = \frac{1}{n}\sum_{i=1}^{k} \dot{x}_i^3 \cdot n_i = \frac{1}{100} \cdot 924845 = 9248{,}45$$

$$m_4 = \frac{1}{n}\sum_{i=1}^{k} \dot{x}_i^4 \cdot n_i = \frac{1}{100} \cdot 23136205 = 231362{,}05$$

Po obliczeniu momentów zwykłych możemy wyznaczyć parametry statystyczne, charakteryzujące badaną zbiorowość:

$$\bar{x} = m_1 = 18{,}05 \text{ zł}$$

$$S = \sqrt{m_2 - m_1^2} = \sqrt{390{,}85 - 18{,}05^2} = \sqrt{65{,}0475} = 8{,}07 \text{ zł}$$

$$V_S = \frac{S}{\bar{x}} \cdot 100\% = \frac{8{,}07}{18{,}05} \cdot 100\% = 44{,}7\%$$

$$A_S = \frac{m_3 - 3m_2 \cdot m_1 + 2m_1^3}{S^3} = \frac{-154{,}61}{525{,}56} = -0{,}29$$

$$K = \frac{m_4 - 4m_3 \cdot m_1 + 6m_2 \cdot m_1^2 - 3m_1^4}{S^4} = \frac{9221{,}9958}{4241{,}2526} = 2{,}17$$

Oceniając wyniki stwierdzamy, że średnia wysokość dziennego kieszonkowego w badanej grupie dzieci wynosiła 18,05 zł. Przeciętne zróżnicowanie wysokości kieszonkowego wynosiło 8,07, co stanowiło 44,7% średniej arytmetycznej. Współczynnik asymetrii jest ujemny, co świadczy o lewostronnym rozkładzie, a jego wartość bezwzględna oznacza słabą asymetrię. Współczynnik koncentracji 2,17 skłania do wniosku, że rozkład jest spłaszczony, a koncentracja wysokości kieszonkowego jest mniejsza niż normalna.

Przykład 2.6.12

Rozkład powierzchni użytków rolnych w 100 wybranych losowo gospodarstwach indywidualnych podaje tabela (dane umowne).

Grupy obszarowe gospodarstw w ha (x_i)	Liczba gospodarstw (n_i)
do 2	12
2 – 5	35
5 – 10	23
10 – 15	20
15 i powyżej	10

Przeprowadzić analizę struktury zbiorowości.

Rozwiązanie

Analizę przeprowadzimy za pomocą miar pozycyjnych ponieważ nie można obliczyć średniej arytmetycznej, że względu na otwarte przedziały. Nie możemy też wyznaczyć dominanty, bowiem przedziały sąsiadujące z przedziałem najliczniejszym nie są równe.

Tablica obliczeniowa

x_i	n_i	Liczebność skumulowana	
do 2	12	12	
2 – 5	35	47	Q_1
5 – 10	23	70	Q_2
10 – 15	20	90	Q_3
15 i powyżej	10	100	
Razem	100	×	

Wyznaczamy pozycje kwartyli

$$N_{Q1} = \frac{100}{4} = 25;\ N_{Q2} = \frac{100}{2} = 50;\ N_{Q3} = \frac{3 \cdot 100}{4} = 75$$

$$Q_1 = x_{02} + \frac{N_{Q1} - \sum_{i=1}^{2-1} n_i}{n_2} \cdot h_2 = 2 + \frac{25 - 12}{35} \cdot 3 = 3{,}11\ ha$$

$$Q_2 = x_{03} + \frac{N_{Q2} - \sum_{i=1}^{3-1} n_i}{n_3} \cdot h_3 = 5 + \frac{50 - 47}{23} \cdot 5 = 5{,}65\ ha$$

$$Q_3 = x_{04} + \frac{N_{Q3} - \sum_{i=1}^{4-1} n_i}{n_4} \cdot h_4 = 10 + \frac{75 - 70}{20} \cdot 5 = 11{,}25\ ha$$

$$Q = \frac{Q_3 - Q_1}{2} = \frac{11{,}25 - 3{,}11}{2} = \frac{8{,}14}{2} = 4{,}07\ ha$$

$$V_Q = \frac{Q}{M_e} \cdot 100\% = \frac{4{,}07}{5{,}65} \cdot 100\% = 72\%$$

$$A_Q = \frac{(Q_3 - Q_2) - (Q_2 - Q_1)}{Q_3 - Q_1} = \frac{(11{,}25 - 5{,}65) - (5{,}65 - 3{,}11)}{11025 - 3{,}11} = 0{,}38$$

Średnia powierzchnia użytków rolnych w badanych gospodarstwach wynosi 5,65 ha. Odchylenie od mediany wynosi ±4,07 ha, przeciętnie biorąc. Pozycyjny współczynnik zmienności informuje, że odchylenie ćwiartkowe stanowi 72% mediany i jest bardzo wysoki. Asymetria rozkładu jest prawostronna, umiarkowana.

Przykład 2.6.13

W pewnym biurze turystycznym wylosowano w sposób niezależny 100 turystów, którzy wybierali się na wycieczkę jednodniową do Berlina i zapytano

o planowane wydatki. Na podstawie zebranych informacji sporządzono tabelę (dane umowne).

Planowane wydatki w euro (x_i)	0-20	20-40	40-60	60-80	80-100
Liczba turystów (n_i)	10	15	25	35	15

Przeprowadzić analizę struktury z wykorzystaniem momentów rozkładu.

Rozwiązanie

Metodę momentów stosuje się głównie do szeregów rozdzielczych, gdy badany szereg statystyczny ma równe i domknięte przedziały losowe. Ponieważ warunki te analizowany szereg spełnia, możemy zbudować tablicę obliczeniową i wyznaczyć momenty zwykłe.

Tablica obliczeniowa

x_i	n_i	$\dot{x}_i$	$\dot{x}_i^2$	$\dot{x}_i^3$	$\dot{x}_i^4$	$\dot{x}_i \cdot n_i$	$\dot{x}_i^2 \cdot n_i$	$\dot{x}_i^3 \cdot n_i$	$\dot{x}_i^4 \cdot n_i$
0 – 20	10	10	100	1000	10000	100	1000	10000	100000
20 – 40	15	30	900	27000	810000	450	13500	405000	12150000
40 – 60	25	50	2500	125000	6250000	1250	62500	3125000	156250000
60 –80	35	70	4900	343000	24010000	2450	171500	12005000	840350000
80 – 100	15	90	8100	729000	65610000	1350	121500	10935000	984150000
Razem	100	×	×	×	×	5600	370000	26480000	1993000000

$$m_1 = \frac{1}{n}\sum_{i=1}^{k} \dot{x}_i \cdot n_i = \frac{1}{100} \cdot 5600 = 56$$

$$m_2 = \frac{1}{n}\sum_{i=1}^{k} \dot{x}_i^2 \cdot n_i = \frac{1}{100} \cdot 370000 = 3700$$

$$m_3 = \frac{1}{n}\sum_{i=1}^{k} \dot{x}_i^3 \cdot n_i = \frac{1}{100} \cdot 264800000 = 264800$$

$$m_4 = \frac{1}{n}\sum_{i=1}^{k} \dot{x}_i^4 \cdot n_i = \frac{1}{100} \cdot 1993000000 = 19930000$$

Po obliczeniu momentów zwykłych możemy wyznaczyć parametry statystyczne, charakteryzujące badaną zbiorowość.

$$\bar{x} = m_1 = 56 \ €$$

$$S = \sqrt{m_2 - m_1^2} = \sqrt{3700 - 56^2} = \sqrt{564} = 23{,}7 \ €$$

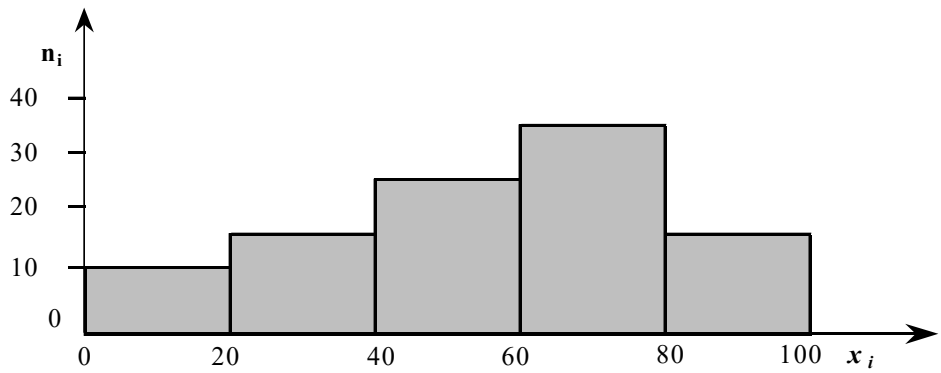
$$V_S = \frac{S}{\bar{x}} \cdot 100\% = \frac{23{,}7}{56} \cdot 100\% = 42{,}3\%$$

$$A_S = \frac{m_3 - 3m_2 \cdot m_1 + 2m_1^3}{S^3} = \frac{-5568}{13312{,}05} = -0{,}242$$

$$K = \frac{m_4 - 4m_3 \cdot m_1 + 6m_2 \cdot m_1^2 - 3m_1^4}{S^4} = \frac{730512}{315945,7} = 2,32$$

Oceniając wyniki stwierdzamy, że średnia wysokość planowanych wydatków w badanej grupie turystów wynosiła 56 €. Przeciętne zróżnicowanie wysokości wydatków wynosiło 23,7€, co stanowiło 42,3% średniej arytmetycznej. Współczynnik asymetrii jest ujemny, co świadczy o lewostronnym rozkładzie, a jego wartość bezwzględna oznacza asymetrię umiarkowaną. Współczynnik koncentracji 2,32 skłania do wniosku, że rozkład jest spłaszczony, a koncentracja wysokości wydatków jest mniejsza niż normalna.

Wysokość wydatków w euro przedstawia wykres na rysunku 2.4.



Wykres 2.4. Wysokość wydatków w euro

Źródło: Opracowanie własne

Przykład 2.6.14

W wyniku sondażu uzyskano informacje dotyczące rocznych zbiorów chmielu w gospodarstwach rolnych indywidualnych w pewnym regionie kraju w okresie dziewięciu lat. Odnotowane zbiory chmielu w tys. ton zamieszczono w tabeli (dane umowne).

Lata	Zbiory chmielu w (tys. ton)
1	2,0
2	1,6
3	1,4
4	1,9
5	1,9
6	1,9
7	2,6
8	2,6
9	2,5

Przeprowadzić analizę struktury badanych zbiorów chmielu stosując poznane miary klasyczne.

Rozwiązanie

Jest to cecha mierzalna, skokowa. Zbudujemy szereg rozdzielczy punktowy.

Zbiory chmielu w (tys. ton) x_i	Częstości absolutne n_i	
1,40	1	
1,60	1	
1,90	3	M_o
2,00	1	
2,50	1	
2,60	2	
Razem	9	

Wartość dominująca zbiorów chmielu wyniosła 1,9 tys. ton. Aby wyznaczyć średnią arytmetyczną i odchylenie standardowe należy utworzyć tablicę obliczeniową.

x_i	n_i	$x_i \cdot n_i$	$x_i - \bar{x}$	$(x_i - \bar{x})^2$	$(x_i - \bar{x})^2 \cdot n_i$
1,40	1	1,40	- 0,64	0,41	0,41
1,60	1	1,60	- 0,44	0,19	0,19
1,90	3	5,70	- 0,14	0,02	0,06
2,00	1	2,00	- 0,04	0,00	0,00
2,50	1	2,50	0,46	0,21	0,21
2,60	2	5,20	0,56	0,31	0,62
Razem	9	18,40	×	×	1,49

$$\bar{x} = \frac{1}{n}\sum_i x_i n_i = \frac{1}{9} \cdot 18{,}4 = 2{,}04 \text{ tys. ton}$$

$$S^2(x) = \frac{1}{n}\sum_i (x_i - \bar{x})^2 n_i = \frac{1}{9} \cdot 1{,}49 = 0{,}166$$

$$S(x) = \sqrt{S^2(x)} = \sqrt{0{,}167} = 0{,}41 \text{ tys. ton}$$

$$V_s = \frac{S(x)}{\bar{x}} = \frac{0{,}41}{2{,}04} \cdot 100\% = 20{,}1\%$$

$$A_s = \frac{\bar{x} - M_o}{S(x)} = \frac{2{,}04 - 1{,}90}{0{,}41} = 0{,}34$$

Przeciętne zbiory chmielu w badanych gospodarstwach indywidualnych wynosiły 2,04 tys. ton, zaś odchylenie od średniej 0,41 tys. ton. Odchylenie standardowe stanowiło 20,1% średniej arytmetycznej, zaś asymetria 0,34 była prawostronna, umiarkowana.

ROZDZIAŁ 3
ANALIZA SZEREGÓW CZASOWYCH

Podstawą analizy dynamiki jest szereg czasowy, zawierający informacje na temat danego zjawiska masowego w czasie.

Szeregiem czasowym nazywamy ciąg wyników obserwacji uporządkowanych w czasie, tzn. $\{t, y_t\}$. Przez t oznaczamy numery kolejnych jednostek czasu, a przez y_t wielkość badanej cechy (zjawiska) w momencie t.

W szeregach czasowych zmienną niezależną jest czas, natomiast zmienną zależną są wartości liczbowe badanego zjawiska

Celem analizy szeregów czasowych jest wykrycie i opis prawidłowości, jakim podlegają zjawiska w czasie. Wyróżnia się następujące czynniki tych zmian zjawiska: trend, wahania sezonowe, wahania cykliczne, czy wahania przypadkowe.

Trendem nazywamy długookresowe, systematyczne zmiany, jakim podlega dane zjawisko. Jednak analizując szereg czasowy nie wnika się w zależności przyczynowo-skutkowe od innych zjawisk. Wahania sezonowe określane są jako regularne odchylenia od tendencji rozwojowej, czyli trendu, które wynikają z warunków klimatycznych (np. pory roku). Wahania cykliczne związane są z cyklem koniunkturalnym, zaś wszystkie nieregularne zmiany traktuje się jako przypadkowe.

W analizie szeregów czasowych dąży się do wyodrębnienia i pomiaru wyróżnionych czynników, co nazywamy dekompozycją szeregu czasowego. Rozłożenie szeregu czasowego na poszczególne czynniki wiąże się z możliwością zdobycia i wykorzystania wiedzy o schemacie każdego z tych czynników do prognozowania zjawiska.

W badaniach statystycznych wyróżniamy dwie grupy metod analizy szeregów czasowych:

- metody indeksowe (służą do liczbowego określenia tempa i intensywności zmian zjawiska w czasie),
- metody wyodrębnienia tendencji rozwojowej (trendu, wahań okresowych i wahań przypadkowych); szczególnym przypadkiem wahań okresowych są wahania sezonowe, powtarzające się najczęściej w cyklu rocznym.

Analiza szeregów czasowych wiąże się z określeniem dynamiki badanego zjawiska oraz czynników wywołujących jego zmienność. Możemy wyodrębnić dwie grupy miar statystycznych, które znajdują zastosowanie w odniesieniu do szeregów czasowych:

- miary średnie,
- miary dynamiki.

3.1. Szeregi czasowe momentów lub okresów

Miary analizy struktury rzadko znajdują zastosowanie w szeregach czasowych, ale wyjątkiem są miary średnie. Zjawiska, które nas otaczają są przedmiotem oceny statystycznej, ponieważ, jak wiadomo, podlegają ciągłym zmianom. W zależności od charakteru zbiorowości statystycznej, zmiany w czasie mogą być w różny sposób rejestrowane.

Niektóre ze zjawisk mogą być badane tylko w określonych momentach, np.: zasoby pieniężne, zapasy wyrobów gotowych, liczba ludności. Mamy wtedy do czynienia z „zasobami", których wielkość możemy określić w danym momencie. Ale są też zjawiska, obserwowane w jednostce czasu obejmującej pewien okres (rok, kwartał, miesiąc), np.: liczba urodzeń, liczba mieszkań oddanych do użytku, produkcja. Te zjawiska mają charakter „strumienia", którego wielkość mierzymy w pewnym okresie.

Przykład 3.1.1

Tabela przedstawia cenę cukru białego kryształ w pewnym sklepie w latach 2012-2018.

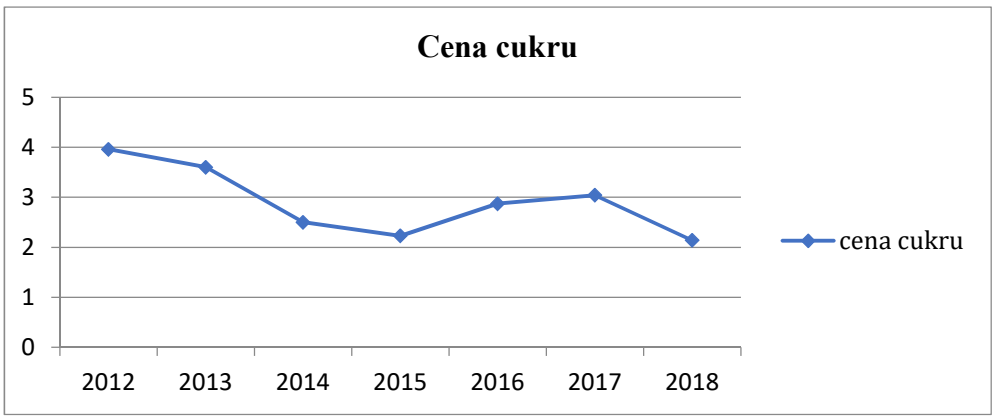
Lata	2012	2013	2014	2015	2016	2017	2018
Cena cukru (w zł za 1 kg) stan na 31 XII	3,96	3,60	2,50	2,23	2,87	3,04	2,14

Źródło: Opracowanie własne

Rozpoznać, czy jest to szereg czasowy momentów, czy okresów i obliczyć średnią cenę cukru.

Rozwiązanie

Zilustrujemy zamieszczone w tabeli dane na wykresie 3.1.



Rys. 3.1. Cena cukru (w zł za kg) w latach 2012-2018

Jest to przykład szeregu czasowego momentów, w którym podano cenę cukru białego kryształ (w zł za kg) według stanu w określonym momencie (31 XII) poszczególnych lat. Dla szeregu momentów obliczamy średnią chronologiczną:

$$\bar{y}_{Ch} = \frac{\frac{1}{2}y_1 + y_2 + \cdots + \frac{1}{2}y_n}{n-1}.$$

Zatem mamy:

$$\bar{y}_{Ch} = \frac{\frac{1}{2}\cdot 3{,}96 + 3{,}60 + 2{,}50 + 2{,}23 + 2{,}87 + 3{,}04 + \frac{1}{2}2{,}14}{7-1}.$$

$$\bar{y}_{Ch} = 2{,}88 \text{ zł}$$

Przeciętna cena cukru w ostatnim z badanych momentów (31 XII 2018) wynosiła 2,88 zł.

Przykład 3.1.2

Tabela przedstawia sprzedaż aut marki Mercedes-Benz w Polsce w latach 2014-2020.

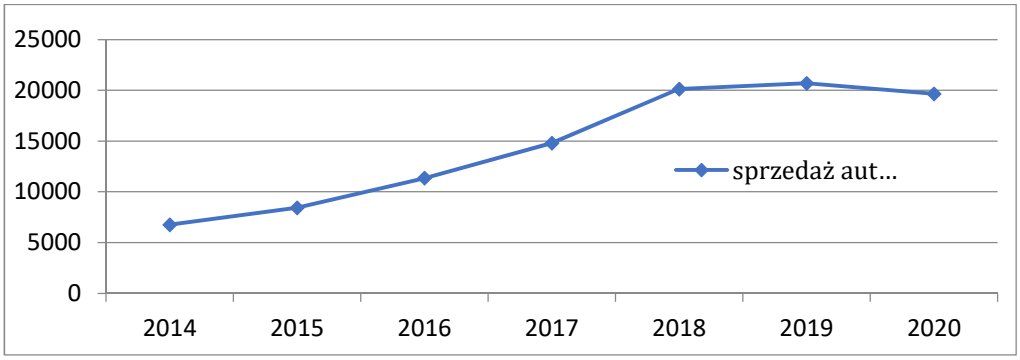
Lata	2014	2015	2016	2017	2018	2019	2020
Liczba (szt.)	6762	8429	11353	14805	20129	20699	19653

Źródło: Opracowanie własne

Rozpoznać, czy jest to szereg czasowy momentów, czy okresów i obliczyć średni roczny poziom sprzedaży aut marki Mercedes-Benz.

Rozwiązanie

Zilustrujemy zamieszczone w tabeli dane na wykresie 3.2.



Rys. 3.2. Sprzedaż aut marki Mercedes-Benz w latach 2014-2020

Ten przykład pokazuje szereg czasowy okresów, ponieważ sprzedaż aut marki Mercedes-Benz w Polsce jest sumą wszystkich samochodów sprzedanych w ciągu całego roku. Dla szeregu okresów obliczamy średnią arytmetyczną prostą:

$$\bar{y} = \frac{y_1 + y_2 + \cdots + y_n}{n}.$$

Zatem mamy:

$$\bar{y} = \frac{6762 + 8429 + 11353 + 14805 + 20129 + 20699 + 19653}{7}.$$

$$\bar{y} = 14547 \text{ sztuki}$$

Średni roczny poziom sprzedaży Mercedes-Benz w badanym okresie wynosi 14547 sztuk.

Przykład 3.1.3

Tabela przedstawia dane dotyczące zatrudnienia kobiet w Wojsku Polskim w latach 2017-2020.

Lata	31 XII 2017	31 XII 2018	31 XII 2019	31 XII 2020
Liczba kobiet (w tys.)	26,4	27,6	28,3	29,7

Źródło: Rocznik Statystyczny Rzeczypospolitej Polskiej 2019 i 2020

Wyznaczyć średni stan zatrudnienia kobiet w Wojsku Polskim w wyróżnionych latach.

Rozwiązanie

Aby uzyskać odpowiedź na pytanie, jaki był średni stan zatrudnienia kobiet w Wojsku Polskim w ostatnim dniu badanych lat, należy do podanych informacji zastosować średnią chronologiczną, tak więc:

$$\bar{y}_{Ch} = \frac{\frac{1}{2}y_1 + y_2 + \cdots + \frac{1}{2}y_n}{n-1}.$$

$$\bar{y}_{Ch} = \frac{\frac{1}{2}\cdot 26{,}4 + 27{,}6 + 28{,}3 + \frac{1}{2}29{,}7}{4-1}.$$

$$\bar{y}_{Ch} = 27{,}98 \text{ tys.}$$

Jeżeli chcemy poznać średni roczny poziom zatrudnienia kobiet w Wojsku Polskim w danym okresie, to obliczymy średnią arytmetyczną następująco:

$$\bar{y} = \frac{y_1 + y_2 + \cdots + y_n}{n}.$$

$$\bar{y} = \frac{26{,}4 + 27{,}6 + 28{,}3 + 29{,}7}{4}.$$

$$\bar{y} = 28 \text{ tys.}$$

Średni stan zatrudnienia kobiet w Wojsku Polskim w wyróżnionych latach wynosił 28 tysięcy.

Przykład 3.1.4

Wyznaczyć średni miesięczny zapas wyprodukowanych detali (dane umowne) w pewnym przedsiębiorstwie mając dane:

Dzień, miesiąc rok	Zapasy wyprodukowanych detali (w tonach)
30 IX 2020 r.	12,7
31 X 2020 r.	13,1
30 XI 2020 r.	13,3
31 XII 2020 r.	12,5

Rozwiązanie

Aby wyznaczyć przeciętny miesięczny zapas w ostatnim kwartale 2020 roku, obliczamy średnią chronologiczną z szeregu momentów, stosując formułę:

$$\bar{y}_{Ch} = \frac{\frac{1}{2}y_1 + y_2 + \ldots + \frac{1}{2}y_n}{n-1},$$

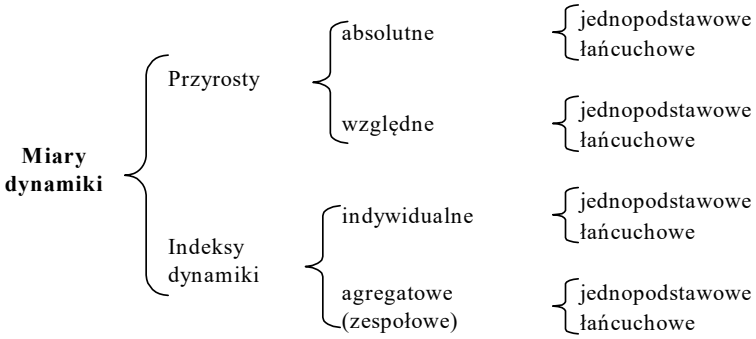
a zatem:

$$\bar{y}_{Ch} = \frac{\frac{1}{2}\cdot 12{,}7 + 13{,}1 + 13{,}3 + \frac{1}{2}\cdot 12{,}5}{4-1} = 13\ ton$$

Przeciętny stan zapasów w ostatnim kwartale 2020 roku wynosił 13 ton.

3.2. Metody indeksowe

Miary dynamiki szeregu czasowego przedstawia schemat:



3.2.1. Przyrosty absolutne i względne

Najprostszym sposobem porównywania zmian zjawiska w czasie jest analiza przyrostów absolutnych i względnych.

Przyrosty absolutne informują, o ile wzrósł (zmalał) poziom zjawiska w okresie badanym w porównaniu z jego poziomem w okresie bazowym (podstawowym). Wielkości te są mianowane.

Przyrosty absolutne mogą być obliczone w stosunku do:

a) jednego okresu (***jednopodstawowe***):

$$y_2 - y_1, \ldots, y_{t-1} - y_1, y_t - y_1 \qquad t = 1, \ldots, n$$

gdzie t = 1 oznacza, że pierwszy okres szeregu statystycznego jest przyjęty jako stała podstawa porównań,

b) dowolnego k-tego okresu przyjętego za stałą podstawę porównań (k≠1):

$$y_1 - y_k, \ldots, y_{t-1} - y_k, y_t - y_k$$

gdzie t = k oznacza, że dowolny k-ty okres przyjęto za stałą podstawę porównań,

c) stale zmieniającego się okresu bazowego (***łańcuchowe***):

$$y_2 - y_1, \ldots, y_{t-1} - y_{t-2}, y_t - y_{t-1}, \text{ tutaj } t = 2, 3, \ldots, n$$

Przyrostem względnym nazywamy stosunek przyrostu absolutnego zjawiska do jego poziomu w okresie bazowym (inna nazwa: wskaźnik tempa przyrostu).

Przyrosty względne jednopodstawowe:

$$\frac{\Delta_{t/k}}{y_k} = \frac{y_t - y_k}{y_k}, \qquad t = 1, 2, \ldots, n$$

Przyrosty względne łańcuchowe:

$$\frac{\Delta_{t/t-1}}{y_{t-1}} = \frac{y_t - y_{t-1}}{y_{t-1}}, \quad t = 2, 3, \ldots, n$$

Mnożąc przez 100 przyrosty względne otrzymamy procentowe przyrosty względne (tempo przyrostu lub obniżki).

Przykład 3.2.1

Ceny usługi dotyczącej makijażu permanentnego powiek zmieniały się w salonie kosmetycznym w ciągu pierwszych siedmiu miesięcy 2021 roku. Dla podanych w tabeli danych obliczyć przyrosty absolutne i względne.

Lp.	Cena usługi w zł	Przyrosty absolutne		Przyrosty względne	
		jedno-podstawowe	łańcuchowe	jedno-podstawowe	łańcuchowe
1	322	0,0	-	0	-
2	320	-2,0	-2,0	-0,62	-0,62
3	329	7,0	9,0	2,17	2,81
4	346	24,0	24,0	7,45	7,29
5	380	58,0	17,0	18,01	4,91
6	418	96,0	38,0	29,81	10,0
7	380	58,0	-38,0	11,80	9,09

3.2.2. Indywidualne indeksy dynamiki

Relatywne zmiany w szeregach czasowych mierzymy za pomocą wskaźników dynamiki zwanych indeksami.

Indeksem nazywamy iloraz poziomu zjawiska w okresie badanym y_t do poziomu zjawiska w okresie przyjętym za podstawę porównań. Indeksy najczęściej wyrażamy w procentach.

Indeksy jednopodstawowe:

$$i_{t/0} = \frac{y_t}{y_0} \cdot 100\%.$$

Indeksy łańcuchowe:

$$i_{t/t\text{-}1} = \frac{y_t}{y_{t-1}} \cdot 100.$$

Tempo przyrostu (względny przyrost łańcuchowy):

$$T = \frac{y_t - y_{t-1}}{y_{t-1}} \cdot 100 = \left(\frac{y_t}{y_{t-1}} - 1 \right) \cdot 100 \cdot$$

Chcąc ustalić średnie względne zmiany, obliczamy średni łańcuchowy wskaźnik dynamiki badanego zjawiska w pewnym okresie czasu, stosując średnią geometryczną z indeksów łańcuchowych:

$$\bar{i} = \sqrt[n-1]{\frac{y_1}{y_0} \cdot \frac{y_2}{y_1} \cdot \frac{y_3}{y_2} \cdot \ldots \cdot \frac{y_n}{y_{n-1}}} = \sqrt[n-1]{\frac{y_n}{y_0}}$$

Średnie tempo obliczamy wykorzystując średni indeks:

$$\bar{T} = (\bar{i} - 1) \cdot 100.$$

Miernik powyższy interpretujemy jako średnią względną zmianę (wzrost lub spadek – w zależności od znaku) badanego zjawiska z okresu na okres.

Uwagi:

1. Jeżeli dane są wskaźniki dynamiki o podstawie łańcuchowej (zmiennej), to wskaźniki dynamiki o podstawie stałej można uzyskać przez kolejne mnożenie wskaźników łańcuchowych.
1. Jeżeli mamy dane wskaźniki dynamiki o podstawie stałej, to możemy obliczyć wskaźniki o podstawie łańcuchowej, dzieląc kolejno (łańcuchowo) wartości wskaźników jednopodstawowych.
2. Jeżeli chcemy zmienić podstawę porównań, dzielimy kolejno wyrazy szeregu indeksów przez wskaźnik dla okresu, który ma być podstawą.
3. Jeżeli tempo zmian jest w miarę równomierne, możemy wykorzystać średni indeks dla oszacowania poziomu zjawiska w niezbyt odległej

przyszłości, np. w okresie t = n + p (n – okres miniony, zaś p – okres prognozowany).

$$\hat{y}_{n+p} = y_0 \cdot (\bar{i})^{n+p}$$

Zakładamy, że obliczone tempo zmian nie ulegnie zmianom w okresie prognozowanym.

4. Równość indeksowa dla indeksów indywidualnych: $i_w = i_q \cdot i_p$,

gdzie: w – wartość, q – ilość, p – cena.

Przykład 3.2.2

Ustalić, czy większa była dynamika wartości aktywów subfunduszu Allianz Akcji Małych i Średnich Spółek (w mln zł) w latach 2014- 2017, czy w latach 2017- 2020.

Lata	2014	2015	2016	2017	2018	2019	2020
Aktywa (mln zł)	120,79	131,85	120,78	146,22	164,04	159,28	175,56

Rozwiązanie

Obliczenia pomocnicze

t	2014	2015	2016	2017	2018	2019	2020
$i_{t/t-1}$	–	1,092	0,916	1,211	1,122	0,971	1,102

Średnie tempo zmian wartości aktywów w latach 2014 – 2017 wynosi:

$$\bar{i}_1 = \sqrt[3]{1{,}092 \cdot 0{,}916 \cdot 1{,}211} = \sqrt[3]{1{,}211} = 1{,}0659 = 106{,}59\%$$

Średnie tempo zmian wartości aktywów w latach 2017 – 2020 wynosi:

$$\bar{i}_2 = \sqrt[3]{1{,}122 \cdot 0{,}971 \cdot 1{,}102} = \sqrt[3]{1{,}201} = 1{,}0630 = 106{,}30\%$$

W pierwszym okresie dynamika wartości aktywów była większa i wynosiła 6,59%, zaś w drugim nieznacznie mniejsza, bo 6,30%.

Sprawdzimy, czy stosując formułę $\bar{i} = \sqrt[n-1]{\frac{y_n}{y_0}}$ uzyskamy identyczne odpowiedzi.

$$\bar{i}_1 = \sqrt[3]{\frac{146{,}22}{120{,}79}} = \sqrt[3]{1{,}211} = 1{,}0659 = 106{,}59\%$$

$$\bar{i}_1 = \sqrt[3]{\frac{175,56}{146,22}} = \sqrt[3]{1,201} = 1,0630 = 106,30\%$$

A zatem wyniki niewiele różnią się.

Przykład 3.2.3

Na podstawie danych dotyczących liczby absolwentów uczelni wyższych na kierunku weterynarii obliczyć: przyrosty absolutne i względne, indeksy indywidualne oraz ustalić średnie względne zmiany.

Rozwiązanie

Lata	Absolwenci weterynarii	Przyrosty absolutne		Przyrosty względne		Indeksy indywidualne	
		jednopodstawowe $y_n - y_0$	łańcuchowe $y_n - y_{n-1}$	jednopodstawowe $\frac{y_n - y_0}{y_0} \cdot 100$	łańcuchowe $\frac{y_n - y_{n-1}}{y_{n-1}} \cdot 100$	jednopodstawowe $\frac{y_n}{y_0} \cdot 100$	łańcuchowe $\frac{y_n}{y_{n-1}} \cdot 100$
2015	617	–	–	–	–	100	–
2016	685	68	68	11,0	11,0	111,0	111,0
2017	686	69	1	11,2	0,1	111,2	100,1
2018	780	163	94	26,4	13,7	126,4	113,7
2019	918	301	138	48,8	17,7	148,8	117,7
2020	935	318	17	51,5	1,9	151,5	101,9

Źródło: Opracowanie własne na podstawie Rocznika Statystycznego RP, GUS 2021

Obserwując indeksy jednopodstawowe zauważamy ogólną tendencję zmian w stosunku do stanu z 2015 roku, który został przyjęty za podstawę porównań. Liczba absolwentów weterynarii relatywnie wzrastała w odniesieniu do okresu podstawowego i w 2020 roku stanowiła 151,5% stanu z 2015, była zatem większa o 51,5%. Indeksy łańcuchowe pokazują względne zmiany z roku na rok. Obliczenia wskazują, że największy przyrost względny miał miejsce w roku 2019, a najmniejszy w 2017.

Aby ustalić średnie względne zmiany, obliczamy średni łańcuchowy wskaźnik dynamiki badanego zjawiska w pewnym okresie, stosując średnią geometryczną z indeksów łańcuchowych.

$$\bar{i} = \sqrt[n-1]{\frac{y_1}{y_0} \cdot \frac{y_2}{y_1} \cdot \frac{y_3}{y_2} \cdot \ldots \cdot \frac{y_n}{y_{n-1}}}$$

$$\bar{i} = \sqrt[5]{1,11 \cdot 1,001 \cdot 1,137 \cdot 1,177 \cdot 1,019} = \sqrt[5]{1,515} = 1,087 = 108,7\%$$

Średnie tempo zmian możemy określić również następująco:

$$\overline{T} = (\bar{i} - 1) \cdot 100\% = 8{,}7\%$$

Przykład 3.2.4

Dynamika liczby widzów w kinach w pewnym mieście w latach 2017-2020 w stosunku do roku 2016 (2016−100%) przedstawiała się następująco: 101%, 98,8%, 85,1%, 78%. Wyznaczyć, w którym roku spadek liczby widzów był największy i o ile? Jaki był średni spadek dynamiki i średnie tempo zmian w latach 2019-2020?

Rozwiązanie

Ustalmy indeksy jednopodstawowe (w stosunku do 2016) i łańcuchowe w badanym okresie.

t	2016	2017	2018	2019	2020
$i_{t/0}$	1,00	1,01	0,988	0,851	0,780
$i_{t/t-1}$	–	1,01	0,978	0,861	0,917

Spadki między poszczególnymi latami wynoszą odpowiednio:

$$i_{2018/0} - i_{2017/0} = 1.01 - 0{,}988 = 0{,}112 = 11{,}2\%$$

$$i_{2019/0} - i_{2018/0} = 0{,}988 - 0{,}851 = 10{,}137 = 13{,}7\%$$

$$i_{2020/0} - i_{2019/0} = 0{,}851 - 0{,}78 = 0{,}071 = 7{,}1\%$$

Największy spadek liczby widzów nastąpił między rokiem 2018, a 2019, zaś średni spadek dynamiki w latach 2019-2020 obliczamy następująco:

$$\bar{i} = \sqrt{0{,}861 \cdot 0{,}917} = 0{,}888$$

Średnie tempo spadku liczby widzów w latach 2019-2020 wynosi:

$$\overline{T} = (\bar{i} - 1) \cdot 100\% = -0{,}112 \cdot 100\% = -11{,}2\%$$

Przykład 3.2.5

Zatrudnienie na podstawie stosunku pracy (w tys.) w pewnym mieście, stan na 31 XII (dane umowne) przedstawia tabela:

Lata	2017	2018	2019	2020
Zatrudnienie (w tys.)	10247	10187	9947	9505

1. Obliczyć indeksy jednopodstawowe i łańcuchowe.
2. Wyznaczyć wielkość zatrudnienia w 2021 roku, zakładając, że indeks dynamiki zatrudnienia w 2021 r. (w stosunku do 2017 r.) wyniósł 95%.
3. Ustalić przeciętne tempo zmian w badanym okresie.

Rozwiązanie

Obliczenia pomocnicze

Lata	2017	2018	2019	2020
Indeksy jednopodstawowe	1,000	0,994	0,971	0,928
Indeksy łańcuchowe	–	0,994	0,976	0,956

Zatrudnienie w 2021 r. będzie wynosiło:

$$y_{2021}\text{-}y_{2017}\ i_{2021/2017} = 10247 \cdot 0{,}95 - 9734{,}6 \text{ tys. osób}$$

Przeciętne tempo zmian w tym okresie wynosi:

$$\bar{i} = \sqrt[3]{0{,}994 \cdot 0{,}976 \cdot 0{,}956} = \sqrt[3]{0{,}927} = 0{,}975 = 97{,}5\%$$

Wynik oznacza, że zatrudnienie w latach 2017 – 2020 zmniejszyło się średnio o 2,5% w badanym okresie.

Przykład 3.2.6

W pewnej spółce zysk kwartalny w latach 2019 – 2020 przedstawia tabela.

Lata	2019				2020			
Kwartały	I	II	III	IV	I	II	III	IV
Zysk w tys. PLN	27,3	29,2	31,0	34,2	37,1	38,8	37,5	39,4

Wyznaczyć indeksy łańcuchowe, średnie kwartalne tempo przyrostu zysku oraz oszacować dochód w I kwartale 2021 roku.

Rozwiązanie

Obliczenia pomocnicze

t	0	1	2	3	4	5	6	7
$i_{t/t-1}$	–	1,07	1,06	1,10	1,08	1,05	0,97	1,05

$$\bar{i} = \sqrt[7]{\frac{39{,}4}{27{,}3}} = \sqrt[7]{1{,}44} = 1{,}053$$

Średnie tempo zmian wynosi:

$$\bar{T} = (\bar{i} - 1) \cdot 100 = (1{,}053 - 1) \cdot 100 = 5{,}3\%$$

Dochód w badanym przedsiębiorstwie zwiększał się z kwartału na kwartał średnio o 5,3%. Zauważmy, że tempo zmian jest w miarę równomierne. Zakładając, że nie ulegnie ono zmianie możemy oszacować dochód przedsiębiorstwa w I kwartale 2021 roku., t = *8,* więc:

$$\hat{y}_8 = 27{,}3 \cdot 1{,}053^8 = 41{,}2\,tys.PLN$$

Przykład 3.2.7

Liczbę okresowych kontroli samochodów ogólnego przeznaczenia przeprowadzonych przez uprawnionego diagnostę w latach 2016-2020 przedstawia szereg czasowy.

Lata	2016	2017	2018	2019	2020
Samochody ogólnego przeznaczenia poddane kontroli okresowej	441	520	592	647	532

1. Obliczyć indeksy jednopodstawowe.
2. Korzystając z obliczonych indeksów jednopodstawowych wyznaczyć indeksy łańcuchowe.

Rozwiązanie

Obliczenia pomocnicze

Lata	2016	2017	2018	2019	2020
Indeksy jednopodstawowe 2016 = 100	1,000	1,179	1,342	1,467	1,206

Aby otrzymać wskaźniki o podstawie zmiennej dzielimy kolejno (łańcuchowo) wartości wskaźników jednopodstawowych:

1,342 : 1,179 = 1,138
1,467 : 1,342 = 1,093
1,206 : 1,467 = 0,822

Indeksy łańcuchowe: 1,179; 1,138; 1,093; 0,822.

Przykład 3.2.8

Lata	1017	2018	2019	2020	2021
Apteki (stan na 31 XII)	6948	7342	7484	7875	8318

Źródło: Rocznik Statystyczny GUS, 2001

1. Obliczyć indeksy łańcuchowe.
2. Korzystając z obliczonych indeksów łańcuchowych wyznaczyć indeksy jednopodstawowe.

Rozwiązanie

Obliczenia pomocnicze

Lata	1017	2018	2019	2020	2021
Apteki (stan na 31 XII)	—	1,057	1,019	1,052	1,056

Aby otrzymać indeksy jednopodstawowe należy pomnożyć kolejno indeksy łańcuchowe:

$$1{,}057 \cdot 1{,}019 = 1{,}077$$
$$1{,}077 \cdot 1{,}052 = 1{,}133$$
$$1{,}133 \cdot 1{,}056 = 1{,}196$$

Indeksy jednopodstawowe: 1,000; 1,057; 1,077; 1,133; 1,196

Przykład 3.2.9

Dynamika produkcji pewnego wyrobu (dane umowne) przedstawia tabela:

Lata	2017	2018	2019	2020	2021
Indeks jednopodstawowy1996 = 100	1,00	1,20	1,38	1,49	1,56

Przekształcić ciąg indeksów jednopodstawowych w ciąg indeksów o podstawie z 2019 r.

Rozwiązanie

Aby przekształcić ciąg indeksów jednopodstawowych w ciąg indeksów o podstawie z 2019 r. dzielimy kolejne wyrazy szeregu tych indeksów przez wskaźnik z 2019 r., wynoszący 1,38

$$\frac{1{,}00}{1{,}38} = 0{,}73 \qquad \frac{1{,}20}{1{,}38} = 0{,}87 \qquad \frac{1{,}38}{1{,}38} = 1{,}00 \qquad \frac{1{,}49}{1{,}38} = 1{,}08$$
$$\frac{1{,}56}{1{,}38} = 1{,}13$$

Indeksy o podstawie z 2019 roku są następujące: 0,73; 0,87; 1,00; 1,08; 1,13.

Przykład 3.2.10

Noclegi udzielone w turystycznych obiektach noclegowych (w tys.) w wybranych krajach Unii Europejskiej prezentuje poniższa tabela. Obliczyć indeksy jednopodstawowe i łańcuchowe.

Lata	Kraje				
	Austria	Cypr	Grecja	Włochy	Polska
2014	110	14	95	378	67
2015	113	13	99	393	71
2016	117	15	99	403	79
2017	121	17	102	421	84

Źródło: Opracowanie własne w oparciu o *Turystyka w Unii Europejskiej*, GUS 2017

Rozwiązanie

Obliczenia pomocnicze

Indeks	Lata	Kraje				
		Austria	Cypr	Grecja	Włochy	Polska
$i_{t/0}$	2014	1,000	1,000	1,000	1,000	1,000
	2015	1,027	0,929	1,042	1,040	1,060
	2016	1,064	1,071	1,042	1,066	1,179
	2017	1,1	1,214	1,074	1,114	1,254
$i_{t/t-1}$	2014	–	–	–	–	–
	2015	1,027	0,929	1,042	1,040	1,060
	2016	1,035	1,154	1,000	1,025	1,113
	2017	1,034	1,133	1,030	1,045	1,063

3.2.3. Indeksy agregatowe

W statystycznych badaniach dynamiki procesów ekonomicznych często istnieje konieczność badania zmian pewnej cechy nie u jednej jednostki, ale u pewnego powiązanego ze sobą zbioru tych jednostek, czyli agregatu.

Indeksy agregatowe wyrażają zmiany wartości pewnej zmiennej w niejednorodnej zbiorowości jednostek statystycznych za pomocą jednej liczby. Aby to było możliwe trzeba dokonać ich agregacji, czyli znaleźć dla nich pewien wspólny czynnik agregujący, zwany inaczej wagą. Wymaga się, aby wagi były stałe. Stałość wag eliminuje wpływ czynnika przyjętego jako wagę.

Niech dany będzie agregat n jednostek statystycznych w dwóch okresach czasu, badanych ze względu na pewną cechę y:

$$y_{01}, y_{02}, \ldots, y_{0n}$$

$$y_{11}, y_{12}, \ldots, y_{1n}$$

Ogólnie indeks agregatowy przedstawia formuła:

$$I_y = \frac{\sum_{i=1}^{n} y_{1i} \cdot w_i}{\sum_{i=1}^{n} y_{0i} \cdot w_i}$$

w_i – waga i-tej jednostki statystycznej

Istnieją dwie podstawowe formuły stałych wag: formuła Laspeyresa oraz formuła Paaschego.

Według formuły Laspeyresa przyjmuje się system wag z okresu podstawowego:

$$I_y^L = \frac{\sum_{i=1}^{n} y_{1i} \cdot w_{0i}}{\sum_{i=1}^{n} y_{0i} \cdot w_{0i}}.$$

Według formuły Paaschego przyjmuje się system wag z okresu badanego:

$$I_y^P = \frac{\sum_{i=1}^{n} y_{1i} \cdot w_{1i}}{\sum_{i=1}^{n} y_{0i} \cdot w_{1i}}.$$

Klasycznymi przykładami indeksów agregatowych są indeksy cen oraz indeksy masy towarowej.

Indeks cen wg formuły Laspeyresa bada zmiany cen pewnego agregatu towarów. Wagami są ilości towarów poszczególnego rodzaju, jakie znajdowały się na rynku lub zostały wyprodukowane w okresie podstawowym.

$$I_p^L = \frac{\sum p_1 \cdot q_0}{\sum q_0 \cdot p_0}$$

p – cena jednostkowa (i-tego towaru), q – ilość towaru

W indeksie cen wg formuły Paaschego wagami są ilości towarów z okresu badanego.

$$I_p^P = \frac{\sum p_1 \cdot q_1}{\sum q_1 \cdot p_0}$$

Indeks masy towarowej (indeks produkcji, ***ilości***) pokazuje zmiany ilości sprzedanego lub wyprodukowanego agregatu towarów, przy czym wagami są ceny jednostkowe. Wzory są następujące:

$$I_q^L = \frac{\sum q_1 \cdot p_0}{\sum q_0 \cdot p_0} \qquad I_q^P = \frac{\sum q_1 \cdot p_1}{\sum q_0 \cdot p_1}$$

Przykład 3.2.11

Pewna spółka eksportuje dwa typy urządzeń. Informacje o wielkości i cenach jednostkowych eksportu w latach 2018 – 2021 podaje tabela (dane umowne). Przeprowadzić analizę dynamiki wartości eksportu, wielkości fizycznych i cen dla obu typów urządzeń łącznie, stosując poznane indeksy agregatowe.

Typ urządzenia	Wielkość eksportu (tys. szt.)		Cena jednostkowa (tys. PLN)	
	2018 (q_0)	2021 (q_1)	2018 (p_0)	2021 (p_1)
I	20	30	90	80
II	10	20	110	140

Rozwiązanie

Obliczenia pomocnicze

Typ urządzenia	q_0p_0	q_1p_1	q_0p_1	q_1p_0
I	1800	2400	1600	2700
II	1100	2800	1400	2200
Razem	2900	5200	3000	4900

$$I_w = \frac{\sum q_1 \cdot p_1}{\sum q_0 \cdot p_0} \cdot 100\% = \frac{5200}{2900} \cdot 100\% = 179{,}3\%$$

Wartość eksportu wzrosła w 2021 roku w porównaniu z 2018 o 79,3% z powodu zmian cen oraz ilości.

$$I_p^L = \frac{\sum p_1 \cdot q_0}{\sum q_0 \cdot p_0} \cdot 100\% = \frac{3000}{2900} \cdot 100\% = 103{,}4\%$$

Gdyby rozmiary eksportu w jednostkach fizycznych były w obu okresach stałe na poziomie 2018 roku to w 2021 nastąpiłby wzrost cen o 3,4% w porównaniu z 2018 rokiem.

$$I_p^P = \frac{\sum p_1 \cdot q_1}{\sum p_0 \cdot q_1} \cdot 100\% = \frac{5200}{4900} \cdot 100\% = 106{,}1\%$$

Gdyby przyjąć niezmienne rozmiary eksportu w jednostkach w 2021 roku wówczas wzrost cen wyniósłby 6,1% w stosunku do 2018.

$$I_q^L = \frac{\sum q_1 \cdot p_0}{\sum q_0 \cdot p_0} \cdot 100\% = \frac{4900}{2900} \cdot 100\% = 169{,}0\%$$

Gdyby przyjąć stałe ceny na poziomie 2018 roku nastąpiłby wzrost wielkości eksportu w 2021 o 69,0% w porównaniu z 2018 rokiem.

$$I_q^P = \frac{\sum q_1 \cdot p_1}{\sum q_0 \cdot p_1} \cdot 100\% = \frac{5200}{3000} \cdot 100\% = 173{,}3\%$$

Gdyby założyć stałe ceny z 2021 roku, nastąpiłby wzrost wielkości eksportu w 2021 o 73,3% w stosunku do 2018 roku.

Sprawdzamy równości indeksowe:

$$I_w = I_p^L \cdot I_q^P = 1{,}034 \cdot 1{,}733 = 1{,}79$$

$$I_w = I_p^P \cdot I_q^L = 1{,}061 \cdot 1{,}689 = 1{,}79$$

Przykład 3.2.12

Spółka BUT produkuje damskie szpilki w trzech kolorach. Wartość sprzedaży oraz zmiany cen w latach 2019 – 2021 podaje tabela (dane umowne).

Kolor butów	Wartość sprzedaży (tys. zł)		Zmiany cen w 2021 r. w stosunku do 2019 r. (w %)
	2019 (q_0p_0)	2021 (q_1p_1)	
Czarne	50	60	spadek o 5%
Beżowe	40	30	bez zmian
Czerwone	20	25	spadek do 2%

Zbadać dynamikę wartości obrotów tymi butami łącznie oraz określić, w jakim stopniu zmiana cen wpłynęła na dynamikę wartości obrotów.

Rozwiązanie

Zauważmy, że $\sum q_0 \cdot p_0 = 110$, a $\sum q_1 \cdot p_1 = 115$. Możemy więc obliczyć indeks wartości:

$$I_w = \frac{\sum q_1 \cdot p_1}{\sum q_0 \cdot p_0} \cdot 100\% = \frac{115}{110} \cdot 100\% = 104{,}5\%$$

Niech $i_p = \frac{p_1}{p_0}$ oznacza zmiany cen. Wykonamy obliczenia pomocnicze.

Obliczenia pomocnicze

Kolor butów	$i_p = \frac{p_1}{p_0}$	$q_0 p_0 \cdot i_p$	$\frac{q_1 p_1}{i_p}$
Czarne	0,05	47,5	63,2
Beżowe	1,00	40,0	30,0
Czerwone	0,98	19,6	25,5
Razem	×	107,1	118,7

Korzystając z tablicy obliczeniowej możemy wyznaczyć indeksy agregatowe cen.

$$I_p^L = \frac{\sum p_0 \cdot q_0 \cdot i_p}{\sum q_0 \cdot p_0} \cdot 100\% = \frac{107,1}{110} \cdot 100\% = 97,4\%$$

$$I_q^L = \frac{\sum p_1 \cdot q_1}{\sum \frac{p_1 \cdot q_1}{i_p}} \cdot 100\% = \frac{115}{118,7} \cdot 100\% = 96,9\%$$

Oceniając dynamikę wartości obrotów tymi butami łącznie możemy stwierdzić, że ogólnie wzrosły one o 4,5%. Na wzory te miała wpływ obniżka cen, która wynosiła od 2,6% do 3,1% w stosunku do 2019 roku.

Przykład 3.2.13

W pewnym sklepie w latach 2019- 2020 zanotowano wyniki sprzedaży piwa i soków. Przedstawiono je w tabeli.

Nazwa napoju	Wartość sprzedaży (w tys. PLN) w 2020 r	Zmiany ilości
Piwo	1400	wzrost o 2%
Soki	85	spadek o 10%

Wiedząc, że łączna wartość sprzedaży w 2019 roku wynosiła 900 tys. PLN scharakteryzować jej dynamikę oraz ocenić wpływ na zmiany wartości za pomocą odpowiednich miar.

Rozwiązanie

Z danych tabeli wynika, że łączna wartość sprzedaży piwa i soków w 2020 roku wyniosła 1485 tys. PLN. Obliczamy indeks wartości.

$$I_w = \frac{\sum q_1 \cdot p_1}{\sum q_0 \cdot p_0} \cdot 100\% = \frac{1485}{900} \cdot 100\% = 165\%$$

Zauważamy wzrost łącznej wartości sprzedaży napojów o 65%.

Wykorzystując dane możemy wyznaczyć agregatowy indeks ilości Paaschego, który określi stopień średnich zmian sprzedaży badanych napojów. Zauważmy, że w 2020 r. nastąpił wzrost ilości sprzedaży piwa o 2%, a więc $i_q = 1,02$, sprzedaż soku zaś spadła o 10%, zatem $i_p = 0,9$.

Obliczamy indeks ilości Paaschego:

$$I_q^P = \frac{\sum p_1 \cdot q_1}{\sum \frac{p_1 \cdot q_1}{i_p}} \cdot 100\% = \frac{1485}{1400 : 1,02 + 85 : 0,9} \cdot 100\% = 101,3\% \cdot$$

Oceniamy, że łączna sprzedaż napojów w 2020 roku wzrosła o 1,3% w stosunku do 2019 r. przy stałych cenach z 2020 roku.

Stosując równość indeksową obliczamy agregatowy indeks cen Laspeyresa:

$$I_p^L = \frac{I_w}{I_q^P} \cdot = \frac{1,65}{1,013} = 1,63\,(163\%)$$

Uzyskaliśmy informację, że cena wzrosła w 2020 r. w porównaniu z 2019 o 63%, przy założeniu, że w 2020 sprzedano te same ilości napojów co w 2019 roku.

Przykład 3.2.14

W salonie mody zanotowano obroty w tys. PLN, które przedstawia tabela (dane umowne).

Towary	Wartość obrotów		Podwyżka cen w t_2 w stosunku do t_1
	t_1	t_2	
Ubrania	500	600	15
Dodatki	400	400	10

Ocenić, w jakim stopniu dynamika masy fizycznej wpłynęła na dynamikę wartości.

Rozwiązanie

Sumując odpowiednio zauważamy, że:

$$\sum q_0 \cdot p_0 = 900,\ a\ \sum q_1 \cdot p_1 = 1000.$$

Wykorzystamy równość indeksową dla indeksów indywidualnych: $i_w = i_p \cdot i_q$

Obliczenia pomocnicze

Towary	$i_p = \frac{p_1}{p_0}$	$i_w = \frac{w_1}{w_0}$	$i_q = \frac{q_1}{q_0}$	$q_0 p_0 \cdot i_q$	$\frac{q_1 p_1}{i_q}$
Ubrania	1,15	1,2	1,04	520	577
Dodatki	1,10	1,0	0,91	364	440
Razem	×	×	×	884	1017

$$I_w = \frac{\sum q_1 \cdot p_1}{\sum q_0 \cdot p_0} \cdot 100\% = \frac{1000}{900} \cdot 100\% = 111{,}1\%$$

$$I_q^L = \frac{\sum p_0 \cdot q_0 \cdot i_q}{\sum q_0 \cdot p_0} \cdot 100\% = \frac{884}{900} \cdot 100\% = 98{,}2\%$$

$$I_q^P = \frac{\sum p_1 \cdot q_1}{\sum \frac{p_1 \cdot q_1}{i_p}} \cdot 100\% = \frac{1000}{1017} \cdot 100\% = 98{,}3\% \cdot$$

Wynik oznacza, że nastąpił wzrost obrotów w t_2 w stosunku do t_1 o 11,1%. Na wzrost ten wpłynęła obniżka ilości o około 1,7% do 1,8% w t_2 w stosunku do t_1.

3.3. Dekompozycja szeregu czasowego

W przypadku zbiorowości statystycznych w *ujęciu statycznym* posługujemy się szeregami rozdzielczymi. W odniesieniu do zbiorowości *ujętych dynamicznie* podstawowym narzędziem, jakie stosuje się jest szereg czasowy.

3.3.1. Metody wyodrębniania tendencji rozwojowej

Szereg czasowy zawiera w porządku chronologicznym wartości badanej cechy prezentowane w określonych momentach bądź w okresach. Zmiany zachodzące w szeregach czasowych mają z reguły różny charakter i mogą je wywoływać różne przyczyny. Podstawowe typy zmian to:

a) wahania ***nieregularne***, wywoływane działaniem ubocznych przyczyn o charakterze losowym,
b) wahania ***okresowe***, które charakteryzują się prawidłowością spowodowaną trwałym działaniem określonych przyczyn głównych, polegających na okresowym wzroście lub spadku rozmiarów zjawiska,
c) wahania ***sezonowe***, które są wahaniami okresowymi o cyklu rocznym.

Analiza szeregów czasowych spełnia dwa podstawowe cele. Z jednej strony pozwala poznać rozwój zjawisk, a drugiej zaś umożliwia przewidywanie przyszłości.

Ocena charakteru zmian w czasie wymaga wyodrębnienia trzech podstawowych składników kształtujących poziom zjawiska. Poziom zjawiska rozpatrywanego w czasie *(Y)* jest funkcją trendu $(\hat{Y})$, wahań sezonowych *(S)*, wahań przypadkowych *(E)*.

$$Y = f(\hat{Y}, S, E)$$

Wyodrębnienie powyższych składników nosi nazwę ***dekompozycji*** szeregu czasowego. Poszczególne składniki szeregu czasowego oceniane są za pomocą odpowiednich charakterystyk liczbowych:

- trend jest opisywany za pomocą średnich ruchomych oraz funkcji analitycznych,
- wahania sezonowe są charakteryzowane za pomocą wskaźników sezonowości,
- miarą wahań przypadkowych jest wariancja resztowa i współczynnik zbieżności.

Średnie ruchome

Zastosowanie ***metody średnich ruchomych*** doprowadza do wygładzenia szeregu czasowego przez częściowe eliminowanie wahań okresowych, jak też przypadkowych. Średnia ruchoma jest średnią określonej liczby ***k*** kolejnych wartości szeregu czasowego.

Średnią ruchomą nieparzystookresową (3-letnią, 5-letnią, 7-letnią) obliczamy głównie wtedy, gdy chcemy pozbyć się wahań przypadkowych. Zazwyczaj dotyczy to rocznych jednostek czasowych.

Średnią ruchomą parzystookresową obliczamy, kiedy w szeregu czasowym występują wahania sezonowe.

Niech $y_1, y_2, \ldots, y_n$ oznaczają kolejne wartości szeregu czasowego. Średnie ruchome liczby okresów wyznaczamy następująco:

dla k = 3 trzyletnia średnia ruchoma

$$\bar{y}_2 = \frac{y_1 + y_2 + y_3}{3}$$

$$\bar{y}_3 = \frac{y_2 + y_3 + y_4}{3}$$

$$\ldots\ldots\ldots\ldots\ldots\ldots$$

$$\bar{y}_{n-1} = \frac{y_{n-2} + y_{n-1} + y_n}{3}$$

dla k = 5 pięcioletnia średnia ruchoma

$$\bar{y}_3 = \frac{y_1 + y_2 + y_3 + y_4 + y_5}{5}$$

$$\bar{y}_4 = \frac{y_2 + y_3 + y_4 + y_5 + y_6}{5}$$

$$\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots$$

$$\bar{y}_{n-2} = \frac{y_{n-4} + y_{n-3} + y_{n-2} + y_{n-1} + y_n}{5}$$

Analogicznie dla k = 7.

Przykład 3.3.1

W poniższej tabeli zamieszczono dane dotyczące bezrobotnych absolwentów (w tys.) (y_t) w okresie od II 2020 do IV 2021 w pewnym regionie. Korzystając z tabeli wyznaczyć tendencję rozwojową w sposób mechaniczny, stosując średnią ruchomą 3-letnią i 5-letnią oraz sporządzić wykres.

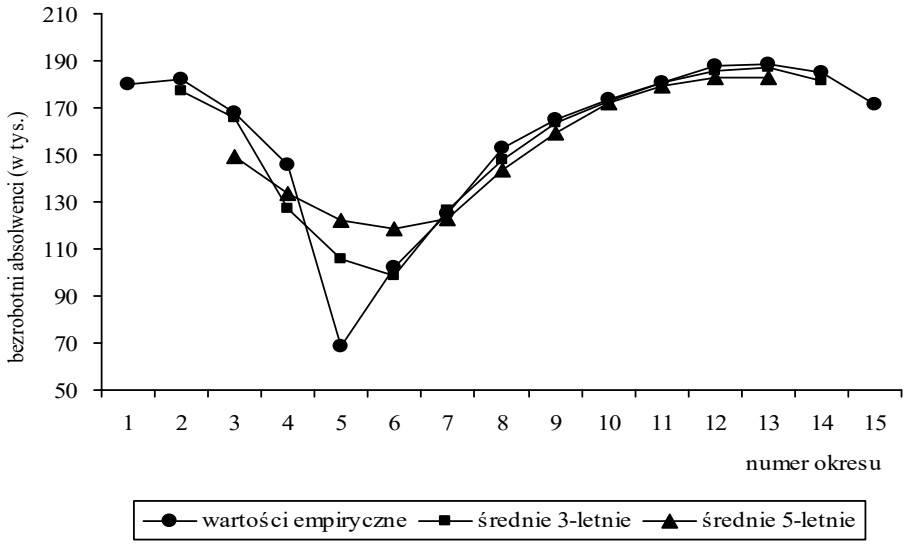
Okres czasu		Bezrobotni absolwenci (w tys.)
2020	II	180,0
	III	182,4
	IV	168,0
	V	146,0
	VI	68,3
	VII	102,3
	VIII	125,0
	IX	153,0
	X	165,0
	XI	173,5
	XII	180,7
2021	I	187,6
	II	188,8
	III	184,7
	IV	171,5

Źródło: Dane umowne

Rozwiązanie

Okres czasu		t	y_t	Średnie ruchome	
				3-letnie	5-letnie
2020	II	1	180,0	–	–
	III	2	182,4	176,8	–
	IV	3	168,0	165,5	148,9
	V	4	146,0	127,4	133,4
	VI	5	68,3	105,5	121,9
	VII	6	102,3	98,5	118,9
	VIII	7	125,0	126,8	122,7
	IX	8	153,0	147,7	143,8
	X	9	165,0	163,8	159,4
	XI	10	173,5	173,1	172,0
	XII	11	180,7	180,6	179,1
2021	I	12	187,6	185,7	183,1
	II	13	188,8	187,0	182,7
	III	14	184,7	181,7	–
	IV	15	171,5	–	–

Wartości empiryczne oraz średnie 3-letnie i 5-letnie bezrobotnych absolwentów przedstawia wykres na ryzunku 3.3.



Rys. 3.3. Wartości empiryczne, średnie 3-letnie oraz średnie 5-letnie bezrobotnych absolwentów

Z powyższego przykładu wynika, że średnie ruchome o większej liczbie okresów lepiej wygładzają szereg. W szeregu empirycznym różnica między y_{max} i y_{min} wynosi: 188,8 – 68,3 = 120,5 tys. osób, w szeregu wygładzonym średnia trzyletnia: 187 – 98,5 = 88,5, zaś w szeregu wygładzonym średnia 5-letnia 183,1 – 118,9 = 64,2.

Przy obliczaniu średnich ruchomych przystookresowych, dokonujemy tzw. „centrowania". Technikę obliczania pokażemy dla k = 4:

$$\bar{y}_3 = \frac{\frac{1}{2}y_1 + y_2 + y_3 + y_4 + \frac{1}{2}y_5}{4}$$

$$\bar{y}_4 = \frac{\frac{1}{2}y_2 + y_3 + y_4 + y_5 + \frac{1}{2}y_6}{4}$$

$$\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots$$

$$\bar{y}_{n-2} = \frac{\frac{1}{2}y_{n-4} + y_{n-3} + y_{n-2} + y_{n-1} + \frac{1}{2}y_n}{4}$$

Przykład 3.3.2

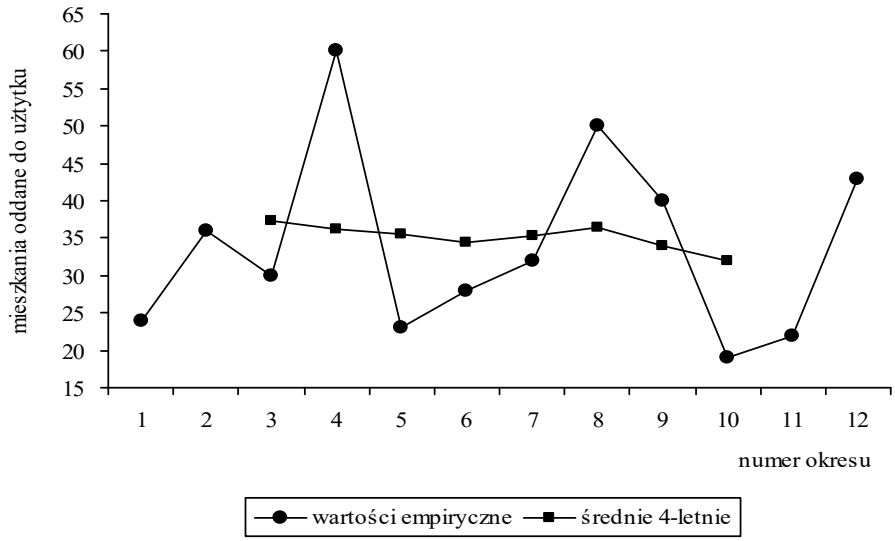
Liczbę mieszkań oddanych do użytku w poszczególnych kwartałach w latach 2019 – 2021 (dane umowne) przedstawia poniższa tabela. Obliczyć średnią scentrowaną z czterech okresów i sporządzić wykres.

Kwartały	2019	2020	2021
I	24	23	40
II	36	28	19
III	30	32	22
IV	60	50	43

Obliczenia pomocnicze

t	y_t	$\overline{y}_t$
1	24	–
2	36	–
3	30	37,4
4	60	36,2
5	23	35,5
6	28	34,5
7	32	35,4
8	50	36,4
9	40	34,0
10	19	31,9
11	22	–
12	43	–

Wartości empiryczne oraz średnie 4-letnie liczby oddanych do użytku mieszkań pokazuje wykres na rysunku 3.4.



Rys. 3.3.2. Wartości empiryczne oraz średnie 4-letnie liczby oddanych do użytku mieszkań

W szeregu empirycznym różnica między y_{max} i y_{min} wynosi: 60 – 19 = 41, a w szeregu wygładzonym średnią scentrowana z czterech okresów wynosi: 37,4 – 31,9 = 5,5.

Uwagi:

1. Szereg średnich ruchomych jest krótszy od szeregu empirycznego (pierwotnego), przy trzyletniej średniej o 2, przy pięcioletniej o 4, ogólnie ($k–1$) obserwacji, a obliczając średnie ruchome scentrowane tracimy k obserwacji.
2. Zasadniczą wadą tej metody jest skracanie szeregu pierwotnego oraz problemy z wykorzystaniem jej do celów predykacji, a zaletą prostota obliczeń.
3. Metodę średnich ruchomych stosuje się zwykle do analizy szeregów, które charakteryzują się dużymi różnokierunkowymi zmianami poziomu zjawiska, co utrudnia dobór odpowiedniej funkcji analitycznej.

3.3.2. Liniowa funkcja trendu

Wyodrębnienie tendencji rozwojowej metodą analityczną polega na znalezieniu takiej postaci funkcji, która byłaby najbardziej zbliżona do funkcji empirycznej (funkcja trendu). Do wyznaczenia parametrów funkcji stosuje się między innymi klasyczną metodę najmniejszych kwadratów,

która zakłada minimalizację sumy kwadratów odchyleń wartości empirycznych y_t od oszacowanych za pomocą funkcji $\hat{y}_t$ w sposób następujący:

$$\sum_{t=1}^{n}(y_t - \hat{y}_t)^2 \Rightarrow min$$

Postać funkcji trendu ustala się na podstawie obserwacji zmian poziomu zjawiska w badanym okresie. Bardzo przydatna jest analiza graficzna szeregu empirycznego (wzrokowa), która ułatwia wybór właściwej postaci funkcji trendu.

Jedną z najczęściej stosowanych funkcji jest funkcja liniowa postaci:

$$\hat{y}_t = at + b$$

gdzie: t – zmienna czasowa,
a, b – parametry funkcji trendu.

Do wyznaczenia parametrów liniowej funkcji trendu stosujemy następujące wzory:

$$a = \frac{\sum_{t=1}^{n}(t - \bar{t}) \cdot y_t}{\sum_{t=1}^{n}(t - \bar{t})^2} \qquad b = \bar{y} - a \cdot \bar{t}\,; \quad \bar{t} = \frac{1}{n}\sum_{t=1}^{n} t \qquad \text{lub} \qquad \bar{t} = \frac{n+1}{2}$$

$$\bar{y} = \frac{1}{n}\sum_{t=1}^{n} y_t$$

Parametr *a* oznacza okresowe tempo wzrostu ($a > 0$) lub spadku ($a < 0$) poziomu badanego zjawiska, zaś parametr b interpretujemy jako stan zjawiska w okresie wyjściowym (tj. $t = 0$).

„Dobroć dopasowania" funkcji trendu do danych empirycznych ocenia się za pomocą następujących miar:

1) ***odchylenia standardowego*** składnika resztowego, które pokazuje średnią różnicę pomiędzy zaobserwowanymi wartościami w szeregu czasowym a wartościami wyznaczonymi funkcją trendu:

$$S_y = \sqrt{\frac{\sum_{t=1}^{n}(y_t - \hat{y}_t)^2}{n-k}} \qquad k \text{ – liczba parametrów}$$

2) ***współczynnika zmienności przypadkowej***, który pokazuje natężenie wahań przypadkowych w stosunku do średniego poziomu zjawiska:

$$V_y = \frac{S_y}{\bar{y}} \cdot 100$$

3) ***współczynnika zbieżności***, który jest stosunkiem zmienności przypadkowej do zmienności całkowitej, a pokazuje, jaka część zmienności w czasie badanego zjawiska jest spowodowana czynnikami przypadkowymi:

$$\varphi^2 = \frac{\sum_{t=1}^{n} (y_t - \hat{y}_t)^2}{\sum_{t=1}^{n} (y_t - \bar{y}_t)^2}$$

4) ***współczynnika determinacji***, który wskazuje, jaka część zmienności w czasie badanego zjawiska jest wyjaśniona funkcją trendu:

$$R^2 = 1 - \varphi^2$$

Poznanie tendencji rozwojowej jest podstawą do predykcji. Na ogół zasady i metody wnioskowania na przyszłość na podstawie odpowiedniego modelu statystyczno-ekonometrycznego nazywa się „predykcją", a wynik procesu wnioskowania nosi miano „prognozy".

Przykład 3.3.3

Oszacować parametry liniowej funkcji trendu dla szeregu podanego w tabeli w badanym okresie oraz ocenić dobroć dopasowania funkcji trendu do danych empirycznych. Sporządzić wykres.

Rok 2021	miesiące											
	I	II	III	IV	V	VI	VII	VIII	IX	X	XI	XII
Liczba badanych osób	41,2	42,1	43,3	43,7	44,1	44,6	45,6	46,3	46,9	47,7	49,0	50,1

Źródło: Dane umowne

Rozwiązanie

Obliczenia pomocnicze

Miesiące 2021 r.	y_t	t	$t-\bar{t}$	$(t-\bar{t})y_t$	$(t-\bar{t})^2$	$\hat{y}_t$	$(y_t-\hat{y}_t)^2$	$(y_t-\bar{y})^2$
I	41,2	1	-5,5	-226,60	30,25	41,26	0,00	17,50
II	42,1	2	-4,5	-189,45	20,25	42,01	0,01	10,78
III	43,3	3	-3,5	-151,55	12,25	42,76	0,29	4,34
IV	43,7	4	-2,5	-109,25	6,25	43,51	0,04	2,83
V	44,1	5	-1,5	-66,15	2,25	44,26	0,03	1,65
VI	44,6	6	-0,5	-22,30	0,25	45,01	0,17	0,61
VII	45,6	7	0,5	22,80	0,25	45,76	0,03	0,05
VII	46,3	8	1,5	69,45	2,25	46,51	0,04	0,84
IX	46,9	9	2,5	117,25	6,25	47,26	0,13	2,30
X	47,7	10	3,5	166,95	12,25	48,01	0,09	5,37
XI	49,0	11	4,5	220,50	20,25	48,76	0,06	13,08
XII	50,1	12	5,5	275,55	30,25	49,51	0,35	22,25
Razem	544,6	78	X	107,20	143,00	544,60	1,23	81,60

$$\bar{t} = \frac{1}{n}\sum t = \frac{78}{12} = 6{,}5 \qquad \bar{y} = \frac{1}{n}\sum y_t = \frac{544{,}6}{12} = 45{,}38$$

$$a = \frac{\sum_{t=1}^{n}(t-\bar{t})\cdot y_t}{\sum_{t=1}^{n}(t-\bar{t})^2} = \frac{107{,}20}{143{,}00} = 0{,}75 \qquad b = \bar{y} - a\cdot\bar{t} = 45{,}38 - 7{,}5\cdot 6{,}5 = 40{,}5$$

a zatem: $\hat{y}_t = 0{,}75\cdot t + 40{,}5$

Wiadomo, że n = 12, k = 2

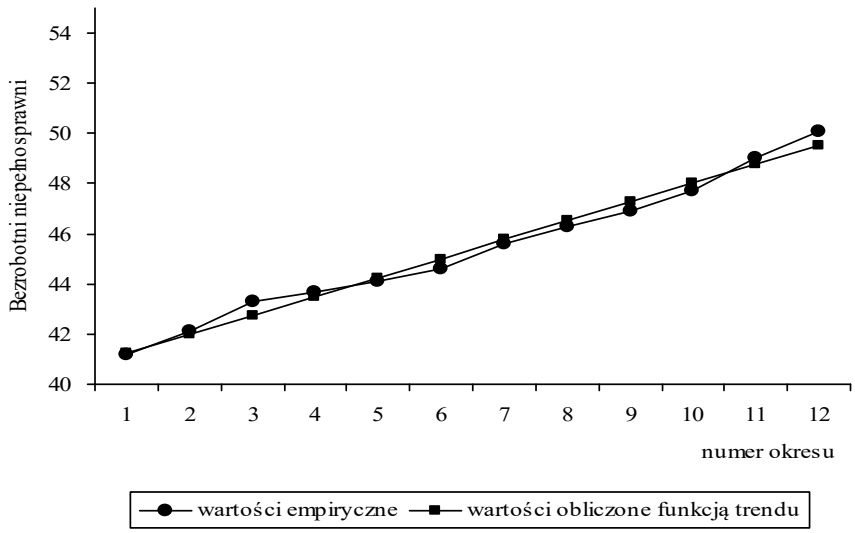
$$S_y = \sqrt{\frac{\sum_{t=1}^{n}(y_t-\hat{y}_t)^2}{n-k}} = \sqrt{\frac{1{,}23}{10}} = \sqrt{0{,}123} = 0{,}35$$

$$V_y = \frac{S_y}{\bar{y}}\cdot 100 = \frac{0{,}35}{45{,}38}\cdot 100 = 0{,}8\%$$

$$\varphi^2 = \frac{\sum_{t=1}^{n}(y_t - \hat{y}_t)^2}{\sum_{t=1}^{n}(y_t - \bar{y}_t)^2} \cdot 100 = \frac{1,23}{81,60} \cdot 100 = 1,5\%$$

$$R^2 = 1 - \varphi^2 = 1 - 1,5\% = 98,5\%$$

Na wykresie na rysunku 3.5 pokazano wartości empiryczne oraz wartości obliczone funkcją trendu dla badanych.



Rys. 3.5. Wartości empiryczne oraz wartości obliczone funkcją trendu dla badanych

Funkcja trendu ma postać $\hat{y}_t = 0,75 \cdot t + 40,51$. Parametr a oznacza, że liczba badanych osób (w tys.) wzrastała z miesiąca na miesiąc w badanym okresie średnio o 0,75 tys. osób. Parametr b interpretujemy teoretycznie dla $t = 0$ (XII 2020 r.).

Dopasowanie funkcji do danych empirycznych jest dobre. Obliczone miary świadczą o niewielkim wpływie wahań przypadkowych. Zaobserwowano różnicę od oszacowanej funkcji trendu średnio o 0,35 tys. osób, co stanowi 0,8% średniego poziomu miesięcznego. Współczynnik zbieżności pokazuje, że zaledwie 1,5% zmienności liczby badanych wywołały czynniki przypadkowe, tak więc 98,5% zmienności zostało wyjaśnione funkcją trendu.

Jeżeli założymy, że tendencja rozwojowa zjawiska nie ulegnie zmianie do końca kwietnia 2022 roku to możemy dokonać ekstrapolacji szeregu czasowego. Kwiecień 2022 roku jest 16. wyrazem szeregu. Podstawiając do równania trendu $t = 16$ otrzymujemy:

$$\hat{y}_{(n+p)} = \hat{y}_{16} = 0{,}75 \cdot 16 + 40{,}50 = 52{,}50$$

Określamy przedział prognozy

$$\hat{y}_{(n+p)} - S_y \langle y_{(n+p)} \langle \hat{y}_{(n+p)} + S_y$$

$$52{,}50 - 0{,}35 \langle y_{IV\ 2002} \langle 52{,}50 + 0{,}35$$

$$52{,}15 \langle y_{IV\ 2002} \langle 52{,}85$$

Przewidywana liczba osób badanych w kwietniu 2022 roku będzie w granicach od 52,15 tys. osób do 52,85 tys. osób przy założeniu, że tendencja rozwojowa zjawiska nie ulegnie zmianie.

3.4. Przykłady praktyczne

Przykład 3.4.1

Obliczyć indeksy łańcuchowe w obu okresach i ustalić, czy większa była dynamika liczby zapotrzebowania na stal w miesiącach 1-8, czy w miesiącach 8-15?

Rozwiązanie

Wyznaczmy indeksy łańcuchowe w pierwszym okresie (tj. w miesiącach 1-8)

t	1	2	3	4	5	6	7	8
$i_{t/t-1}$	–	1,13	1,066	1,092	0,901	1,016	0,938	0,984

Wyznaczmy indeksy łańcuchowe w drugim okresie (tj. w miesiącach 8-15)

t	8	9	10	11	12	13	14	15
$i_{t/t-1}$	–	0,917	1,109	1,148	0,871	1,082	0,909	0,883

Średnie tempo zmian zapotrzebowania na stal w miesiącach 1-8 obliczamy ze wzoru:

$$\bar{i} = \sqrt[n-1]{\frac{y_n}{y_0}}$$

$$\bar{i}_1 = \sqrt[7]{\frac{60}{54}} = \sqrt[7]{1{,}111} = 1{,}0152 = 101{,}52\%$$

Podobnie obliczając średnie tempo zmian zapotrzebowania na stal w miesiącach 8–15 mamy:

$$\bar{i}_2 = \sqrt[7]{\frac{53}{60}} = \sqrt[7]{0{,}883} = 0{,}9823 = 98{,}23\%$$

W pierwszym badanym okresie dynamika liczby zapotrzebowania na stal była wyższa i wynosiła 1,52%, zaś w drugim była niższa o 1,77%.

Przykład 3.4.2

Dynamika liczby widzów w kinach w pewnym mieście w latach 2017-2020 w stosunku do roku 2016 (2016–100%) przedstawia się następująco: 101%, 98,8%, 85,1%, 78%. Wyznaczyć, w którym roku spadek liczby widzów był największy i o ile? Jaki był średni spadek dynamiki i średnie tempo zmian w latach 2019-2020?

Rozwiązanie

Ustalmy indeksy jednopodstawowe(w stosunku do 2016) i łańcuchowe w badanym okresie.

t	2016	2017	2018	2019	2020
$i_{t/0}$	1,00	1,01	0,988	0,851	0,780
$i_{t/t-1}$	–	1,01	0,978	0,861	0,917

Spadki między poszczególnymi latami wynoszą odpowiednio:

$$i_{2018/0} - i_{2017/0} = 1{,}01 - 0{,}988 = 0{,}112 = 11{,}2\%$$

$$i_{2019/0} - i_{2018/0} = 0{,}988 - 0{,}851 = 0{,}137 = 13{,}7\%$$

$$i_{2020/0} - i_{2019/0} = 0{,}851 - 0{,}78 = 0{,}071 = 7{,}1\%$$

Największy spadek liczby widzów nastąpił między rokiem 2018 a 2019, zaś średni spadek dynamiki w latach 2019-2020 obliczamy następująco:

$$\bar{i} = \sqrt{0{,}861 \cdot 0{,}917} = 0{,}888$$

Średnie tempo spadku liczby widzów w latach 2019-2020 wynosi:

$$\bar{T} = (\bar{i} - 1) \cdot 100\% = -0{,}112 \cdot 100\% = 11{,}2\%$$

Przykład 3.4.3

Obliczyć indeksy jednopodstawowe dynamiki produkcji czekolady w kolejnych kwartałach 2021 roku (dane umowne), jeżeli znane są indeksy łańcuchowe w badanym okresie.

Kwartały	I	II	III	IV
Indeks łańcuchowy Kwartał poprzedni=100	1,01	1,05	0,98	1,04

Rozwiązanie

$$1{,}01 \cdot 1{,}05 = 1{,}06$$
$$1{,}06 \cdot 0{,}98 = 1{,}04$$
$$1{,}04 \cdot 1{,}04 = 1{,}08$$

Zatem indeksy jednopodstawowe: 1,01: 1,06; 1,04; 1,08.

Przykład 3.4.4

Obliczyć łańcuchowe wskaźniki tempa wzrostu (spadku) liczby godzin przeznaczonych na granie w gry komputerowe przez dzieci w wieku 10-11 lat w okresie od sierpnia do grudnia 2020 roku (dane umowne), jeżeli znane są indeksy jednopodstawowe dla badanych okresów.

Miesiąc	VIII	IX	X	XI	XII
Indeks jednopodstawowy lipiec 2020 = 100	1,09	1,23	1,36	1,50	1,61

Rozwiązanie

$$1{,}23 : 1{,}09 = 1{,}13$$
$$1{,}36 : 1{,}23 = 1{,}11$$
$$1{,}50 : 1{,}36 = 1{,}10$$
$$1{,}61 : 1{,}50 = 1{,}07$$

Indeksy łańcuchowe: 1,09; 1,13; 1,11; 1,10; 1,07.

Przykład 3.4.5

Dynamikę przyjęć studentów na pierwszy rok teologii na pewnej uczelni (dane umowne) charakteryzuje ciąg indeksów:

Lata	2017	2018	2019	2020	2021
Wskaźnik Rok poprzedni=100%	–	1,01	0,98	0,85	0,78

Przekształcić ciąg indeksów jednopodstawowych w ciąg indeksów o podstawie z 2019 roku.

Rozwiązanie

Jeżeli dane są indeksy o podstawie łańcuchowej, to przekształcamy je w indeksy jednopodstawowe: 1,00; 1,01; 0,99; 0,84; 0,66.

W celu wyznaczenia indeksów jednopodstawowych o podstawie z 2019 roku, dzielimy kolejne wyrazy szeregu tych indeksów przez wskaźnik dla 2019, równy 0,99:

$$\frac{1,00}{0,99} = 1,01$$

$$\frac{0,99}{0,99} = 1,00$$

$$\frac{0,84}{0,99} = 0,85$$

$$\frac{0,66}{0,99} = 0,67$$

Indeksy o podstawie z 2019 roku to: 1,01; 1,02; 1,00; 0,85; 0,67.

Przykład 3.4.6

Zatrudnienie na podstawie stosunku pracy (w tys.), stan na 31 XII przedstawia tabela:

Lata	2016	2017	2018	2019
Zatrudnienie (w tys.)	15293,3	15710,8	15949,7	16120,6

Źródło: Rocznik Statystyczny GUS, 2020

1. Obliczyć indeksy jednopodstawowe i łańcuchowe.
2. Wyznaczyć wielkość zatrudnienia w 2020 roku, zakładając, że indeks dynamiki zatrudnienia w 2020 r. (w stosunku do 2016 r.) wyniósł 95%.
3. Ustalić przeciętne tempo zmian w badanym okresie.

Rozwiązanie

Obliczenia pomocnicze

Lata	2016	2017	2018	2019
Indeksy jednopodstawowe	1,000	1,027	1,042	1,054
Indeksy łańcuchowe	–	1,027	1,015	1,011

Zatrudnienie w 2001 roku będzie wynosiło:

$$y_{2001} - y_{1997} \cdot i_{2001/1997} = 10247 \cdot 0,95 = 9734,6 \text{ tys. osób}$$

$$\bar{i} = \sqrt[3]{i_{2017/2016} \cdot i_{2018/2017} \cdot i_{2019/2018}} = \sqrt[3]{1{,}027 \cdot 1{,}015 \cdot 1{,}011} =$$

$$= \sqrt[3]{1{,}054} = 1{,}018 = 101{,}8\%$$

Wynik oznacza, że zatrudnienie w badanym okresie (tj. w latach 2016-2019) wzrosło średnio o 1,8%.

Przykład 3.4.7

W pewnej spółce zysk kwartalny w latach 2019 – 2020 przedstawia tabela (dane umowne)

Lata	2019				2020			
Kwartały	I	II	III	IV	I	II	III	IV
Zysk w tys. PLN	27,3	29,2	31,0	34,2	37,1	38,8	37,5	39,4

Wyznaczyć indeksy łańcuchowe, średnie kwartalne tempo przyrostu zysku oraz oszacować dochód w I kwartale 2021 roku.

Rozwiązanie

Obliczenia pomocnicze

t	0	1	2	3	4	5	6	7
$i_{t/t-1}$	–	1,07	1,06	1,10	1,08	1,05	0,97	1,05

$$\bar{i} = \sqrt[7]{\frac{39{,}4}{27{,}3}} = \sqrt[7]{1{,}44} = 1{,}053$$

Średnie tempo zmian wynosi: $\bar{T} = (\bar{i} - 1) \cdot 100 = (1{,}053 - 1) \cdot 100 = 5{,}3\%$

Dochód w badanym przedsiębiorstwie zwiększał się z kwartału na kwartał średnio o 5,3%. Zauważmy, że tempo zmian jest w miarę równomierne. Zakładając, że nie ulegnie ono zmianie możemy oszacować dochód przedsiębiorstwa w I kwartale 2021 roku, t = 8, więc:

$$y_{2021-Ikw.} = y_{2029-Ikw.} \cdot \bar{i}^{\,8} = 27{,}3 \cdot 1{,}053^{8} = 41{,}2\,tys.zł$$

Przykład 3.4.8

Noclegi udzielone w turystycznych obiektach noclegowych (w tys.) w wybranych krajach Unii Europejskiej prezentuje poniższa tabela. Obliczyć indeksy jednopodstawowe i łańcuchowe.

Lata	Kraje				
	Austria	Cypr	Grecja	Włochy	Polska
2014	110	14	95	378	67
2015	113	13	99	393	71
2016	117	15	99	403	79
2017	121	17	102	421	84

Źródło: Opracowanie własne w oparciu o *Turystyka w Unii Europejskiej*, GUS 2017

Rozwiązanie

Obliczenia pomocnicze

Indeks	Lata	Kraje				
		Austria	Cypr	Grecja	Włochy	Polska
$i_{t/0}$	2014	1,000	1,000	1,000	1,000	1,000
	2015	1,027	0,929	1,042	1,040	1,060
	2016	1,064	1,071	1,042	1,066	1,179
	2017	1,1	1,214	1,074	1,114	1,254
$i_{t/t-1}$	2014	–	–	–	–	–
	2015	1,027	0,929	1,042	1,040	1,060
	2016	1,035	1,154	1,000	1,025	1,113
	2017	1,034	1,133	1,030	1,045	1,063

Przykład 3.4.9

Pewna firma produkuje trzy wyroby I, II, III. Informację o wielkości i cenach jednostkowych eksportu tych wyrobów w latach 2017-2020 pokazuje tabela (dane umowne).

Wyrób	Wielkość eksportu (tys. szt.)		Cena jednostkowa (tys. zł)	
	2017 (q_0)	2020 (q_1)	2017 (p_0)	2020 (p_1)
I	100	20	5	20
II	100	100	10	10
III	200	400	20	6

Przeprowadzić analizę dynamiki eksportu, wielkości fizycznych i cen dla wyrobów łącznie stosując indeksy agregatowe.

Rozwiązanie

Tabela obliczeniowa

Wyrób	q_0p_0	q_1p_1	q_0p_1	q_1p_0
I	500	400	2000	100
II	1000	1000	1000	1000
III	4000	2400	1200	8000
Razem	5500	3800	4200	9100

$$I_w = \frac{\sum q_1 \cdot p_1}{\sum q_0 \cdot p_0} \cdot 100\% = \frac{3800}{5500} \cdot 100\% = 69{,}1\%$$

Wartość eksportu zmalała w 2020 roku w porównaniu z 2017 o 30,9% z powodu zmian cen oraz ilości.

$$I_P^L = \frac{\sum q_0 \cdot p_1}{\sum q_0 \cdot p_0} \cdot 100\% = \frac{4200}{5500} \cdot 100\% = 76{,}36\%$$

Gdyby rozmiary eksportu w jednostkach fizycznych były w obu okresach stałe na poziomie 2017 roku to w 2020 nastąpiłby spadek cen o 23,64% w porównaniu z 2017 rokiem.

$$I_P^P = \frac{\sum q_1 \cdot p_1}{\sum q_1 \cdot p_0} \cdot 100\% = \frac{3800}{9100} \cdot 100\% = 41{,}76\%$$

Gdyby przyjąć niezmienne rozmiary eksportu w jednostkach w 2020 wówczas spadek cen wyniósłby 58,24% w stosunku do 2017 roku.

$$I_Q^L = \frac{\sum q_1 \cdot p_0}{\sum q_0 \cdot p_0} \cdot 100\% = \frac{9100}{5500} \cdot 100\% = 165{,}45\%$$

Gdyby przyjąć stałe ceny na poziomie 2017 roku nastąpiłby wzrost wielkości eksportu w 2020 r. o 65,45% w porównaniu z 2017 rokiem.

$$I_Q^P = \frac{\sum q_1 \cdot p_1}{\sum q_0 \cdot p_1} \cdot 100\% = \frac{3800}{4200} \cdot 100\% = 90{,}48\%$$

Gdyby założyć stałe ceny z 2020 roku nastąpiłby spadek wielkości eksportu w 2020 o 9,52% w stosunku do 2017 roku.

Sprawdzamy równości indeksowe:

$$I_W = I_P^L \cdot I_Q^P = 0{,}7636 \cdot 0{,}9048 = 0{,}6909$$

$$I_W = I_Q^L \cdot I_P^P = 1{,}6545 \cdot 0{,}4176 = 0{,}6909$$

Przykład 3.4.10

Ceny i ilości trzech produktów A, B, C kupowanych przez ludność w dwóch okresach przedstawiają się następująco (dane umowne):

Produkt	Cena (w zł/kg) w okresie		Zakupione ilości (w kg))	
	podstawowym (p_0)	badanym (p_1)	podstawowym (q_0)	badanym (q_1)
A	2,0	2,1	50	75
B	1,0	1,3	20	20
C	10,0	10,1	10	5

Wyznaczyć indeks dynamiki cen przeciętnych, indeksy cen Laspeyresa i Paaschego.

Rozwiązanie

Wyznaczmy średnią cenę w okresie podstawowym:

$$\overline{p}_0 = \frac{2{,}0 \cdot 50 + 1{,}0 \cdot 20 + 10{,}0 \cdot 10}{50 + 20 + 10} = \frac{220}{80} = 2{,}75$$

A teraz średnią cenę w okresie badanym:

$$\overline{p}_1 = \frac{2{,}1 \cdot 75 + 1{,}3 \cdot 20 + 10{,}1 \cdot 5}{75 + 20 + 5} = \frac{234}{100} = 2{,}34$$

Indeks dynamiki cen obliczamy ze wzoru:

$$i_p = \frac{\overline{p}_1}{\overline{p}_0} \cdot 100\% = \frac{2{,}34}{2{,}72} \cdot 100\% = 85\%$$

Zanim wyznaczymy indeksy cen Laspeyresa i Paaschego wykonamy obliczenia pomocnicze w tabeli:

Produkt	p_0q_0	p_0q_1	p_1q_0	p_1q
A	100	150	105	157,5
B	20	20	26	26,0
C	100	50	101	50,5
Razem	220	220	232	234

$$I_P^L = \frac{\sum p_1 \cdot q_0}{\sum p_0 \cdot q_0} \cdot 100\% = \frac{232}{220} \cdot 100\% = 105{,}45\%$$

Gdyby zakupione ilości były w obu okresach stałe na poziomie podstawowym, to nastąpiłby wzrost cen o 5,45% w porównaniu z okresem podstawowym.

$$I_P^P = \frac{\sum p_1 \cdot q_1}{\sum p_0 \cdot q_1} \cdot 100\% = \frac{234}{220} \cdot 100\% = 106{,}36\%$$

Gdyby przyjąć, że zakupione ilości są na poziomie badanym, wówczas wzrost cen wynosiłby 6,36%.

Przykład 3.4.11

W poniższej tabeli zamieszczono dane dotyczące liczby osób osadzonych w zakładach karnych od 31 X 2019 do 31 XII 2020 roku. Korzystając z tabeli obliczyć tendencję rozwojową w sposób mechaniczny stosując średnią ruchomą 3-letnią i 5-letnią oraz sporządzić wykres.

Okres czasu	Liczba osadzonych
31 X 2019	74732
30 XI 2019	74819
31 XII 2019	74130
31 I 2020	75104
29 II 2020	75664
31 III 2020	74154
30 IV 2020	71578
31 V 2020	70345
30 VI.2020	69894
31 VII.2020	69375
31 VIII 2020	68812
30 IX 2020	69065
31 X 2020	68518
30 XI 2020	68180
31 XII 2020	67894

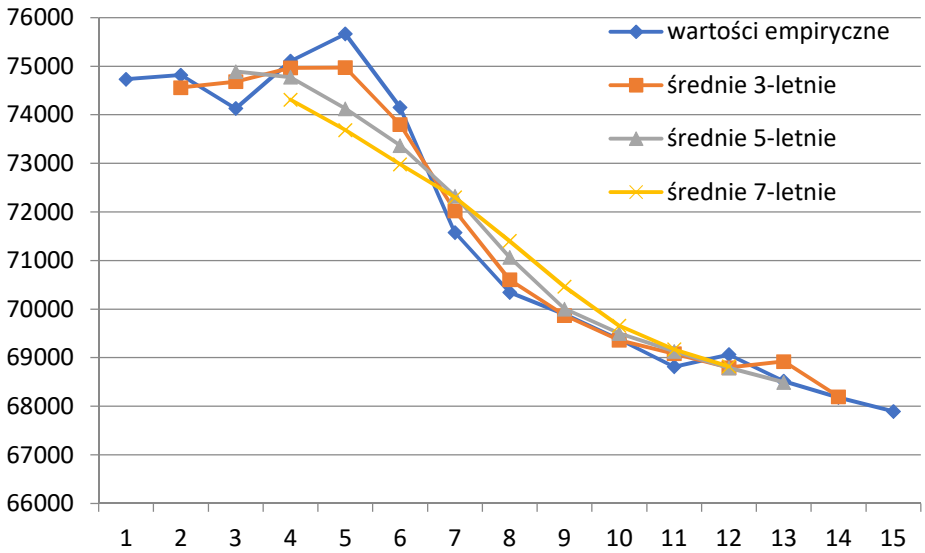
Źródło: Ministerstwo Sprawiedliwości, Roczna informacja statystyczna za rok 2020

Rozwiązanie

Średnie ruchome nieparzystookresowe (3-letnia,5-letnia,7-letnia) są średnimi arytmetycznymi z badanych okresów.

Okres czasu		t	y_t	Średnie ruchome		
				3-letnie	5-letnie	7-letnie
2019	X	1	74732	–	–	–
	XI	2	74819	74560	–	–
	XII	3	74130	74684	74890	–
2020	I	4	75104	74966	74774	74312
	II	5	75664	74974	74126	73685
	III	6	74154	73799	73369	72981
	IV	7	71578	72026	72327	72302
	V	8	70345	70606	71069	71403
	VI	9	69894	69871	70001	70460
	VII	10	69375	69360	69498	69655
	VIII	11	68812	69084	69133	69170
	IX	12	69065	68798	68790	68820
	X	13	68518	68921	68494	–
	XI	14	68180	68197	–	–
	XII	15	67894	–		

Wartości empiryczne oraz średnie: 3-letnie, 5-letnie i 7-letnie pokazuje poniższy wykres (rys. 3.6).



Rys. 3.6. Wartości empiryczne, średnie 3-letnie, 5-letnie, 7-letnie

Zauważmy, że średnie ruchome o większej liczbie okresów lepiej wygładzają szereg. W szeregu empirycznym różnica między y_{max} i y_{min} wynosi: 75664–67894=7770 osadzonych, natomiast w szeregu wygładzonym średnia trzyletnia: 74974–68197=6777 osób, zaś w szeregu wygładzonym średnią pięcioletnia: 74890–68494=6396, a siedmioletnia: 74312–68821 = 5491.

Przykład 3.4.12

Liczbę kasacji w sprawach karnych skierowanych przez Rzecznika Praw Obywatelskich w poszczególnych kwartałach w latach 2019-2021 przedstawia poniższą tabela. Obliczyć średnią scentrowaną z czterech okresów i sporządzić wykres.

Kwartały	2019	2020	2021
I	11	7	31
II	17	5	24
III	14	11	30
IV	7	20	22

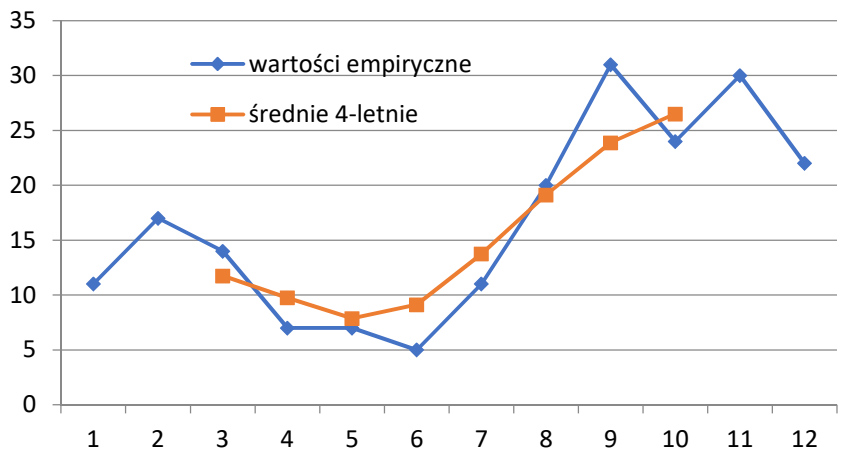
Źródło: Biuletyn Informacji Publicznej RPO

Rozwiązanie

Średnie ruchome parzystookresowe (4-kwartalna, 4-miesięczna, 4-letnia) są średnimi chronologicznymi z badanych okresów.

Obliczenia pomocnicze

t	y_t	$\bar{y}_t$
1	11	–
2	17	–
3	14	11,75
4	7	9,75
5	7	7,875
6	5	9,125
7	11	13,75
8	20	19,125
9	31	23,875
10	24	26,50
11	30	–
12	22	–



Rys. 3.7. Wartości empiryczne oraz średnie 4-letnie liczby kasacji.

W szeregu empirycznym różnica między y_{max} i y_{min} wynosi: 31-5=26 spraw, a w szeregu wygładzonym średnia scentrowana z 4 okresów wynosi: 26,5-7,875=18,6 spraw (p. rys. 3.7)

Przykład 3.4.13

Oszacować parametry liniowej funkcji trendu dla szeregu podanego w tabeli w badanym okresie oraz ocenić dobroć dopasowania funkcji trendu do danych empirycznych. Sporządzić wykres.

Bezrobotni niepełnosprawni (w tys.) w 2020 roku Polsce

Rok	miesiące											
	I	II	III	IV	V	VI	VII	VIII	IX	X	XI	XII
Liczba bezrobotnych	60,2	59,8	58,6	59,1	59,5	58,8	57,9	57,0	56,3	55,5	55,3	55,7

Źródło: niepelnosprawni.gov.pl

Rozwiązanie

Obliczenia pomocnicze

Miesiąc 2020 r.	y_t	t	$t-\bar{t}$	$(t-\bar{t})y_t$	$(t-\bar{t})^2$	$\hat{y}_t$	$(y_t-\hat{y}_t)^2$	$(y_t-\bar{\hat{y}}_t)^2$
I	60,2	1	–5,5	–331,10	30,25	60,45	0,06	5,71
II	59,8	2	–4,5	–269,10	20,25	59,97	0,03	3,96
III	58,6	3	–3,5	–205,10	12,25	59,49	0,79	0,62
IV	59,1	4	–2,5	–147,75	6,25	59,01	0,01	1,66
V	59,5	5	–1,5	–89,25	2,25	58,53	0,94	2,86
VI	58,8	6	–0,5	–29,40	0,25	58,05	0,56	0,98
VII	57,9	7	0,5	28,95	0,25	57,57	0,11	0,01
VIII	57,0	8	1,5	85,50	2,25	57,09	0,01	0,66
IX	56,3	9	2,5	140,75	6,25	56,61	0,10	2,28
X	55,5	10	3,5	194,25	12,25	56,13	0,40	5,34
XI	55,3	11	4,5	248,85	20,25	55,65	0,12	6,30
XII	55,7	12	5,5	304,15	30,25	55,17	0,28	4,45
Razem	693,7	78	X	–69,25	143,00	693,72	3,41	34,83

$$\bar{t} = \frac{1}{n}\sum_{i=1}^{n} t_i = \frac{78}{12} = 6{,}5$$

$$\bar{y} = \frac{1}{n}\sum_{i=1}^{n} y_i = \frac{693{,}7}{12} = 57{,}81$$

$$a = \frac{\sum_{i=1}^{n}(t-\bar{t})y_t}{\sum_{i=1}^{n}(t-\bar{t})^2} = \frac{-69,25}{143} = -0,48$$

$$b = \bar{y} - a \cdot \bar{t} = 57,81 + 0.48 \cdot 6,5 = 60,93$$

Zatem

$$\hat{y}_t = -0,48t + 60,93$$

Wiadomo, że n=12, k=2

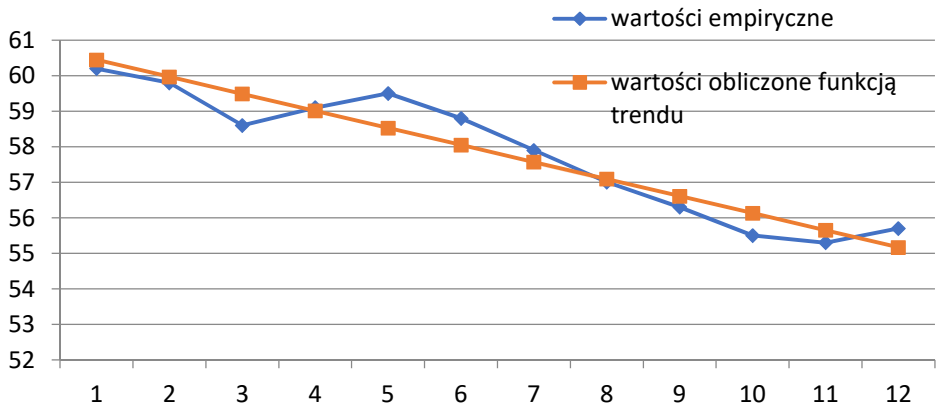
$$S_y = \sqrt{\frac{\sum_{i=1}^{n}(y_t - \hat{y}_t)^2}{n-k}} = \sqrt{\frac{34,83}{10}} = 0,58$$

$$V_y = \frac{S_y}{\bar{y}} \cdot 100\% = \frac{0,58}{57,81} \cdot 100\% = 1\%$$

$$\varphi^2 = \frac{\sum_{i=1}^{n}(y_t - \hat{y}_t)^2}{\sum_{i=1}^{n}(y_t - \bar{\hat{y}}_t)^2} \cdot 100\% = \frac{3,41}{34,83} \cdot 100\% = 9,8\%$$

$$R^2 = 1 - \varphi^2 = 100\% - 9,8\% = 90,2\%$$

Wykres na rysunku 3.8 pokazuje wartości empiryczne oraz wartości obliczone funkcją trendu dla bezrobotnych niepełnosprawnych.



Rys. 3.8. Wartości empiryczne oraz wartości obliczone funkcją trendu dla bezrobotnych niepełnosprawnych

Funkcja trendu ma postać $\hat{y}_t = -0{,}48t + 60{,}93$. Parametr a oznacza, że liczba bezrobotnych niepełnosprawnych (w tys.) malała z miesiąca na miesiąc w badanym okresie średnio o 0,48 tys. osób. Parametr b interpretujemy jako teoretyczne bezrobocie dla $t = 0$ (XII 2020 r.).

Dopasowanie funkcji do danych empirycznych jest dobre. Obliczone miary świadczą o niewielkim wpływie wahań przypadkowych. Zaobserwowane bezrobocie różni się od oszacowanej funkcji trendu średnio o 0,58 tys. osób, co stanowi 1% średniego poziomu miesięcznego. Współczynnik zbieżności pokazuje, że 9,8% zmienności liczby bezrobotnych wywołały czynniki przypadkowe, tak więc 90,2% zmienności zostało wyjaśnione funkcją trendu.

Jeżeli założymy, że tendencja rozwojowa zjawiska nie ulegnie zmianie do końca kwietnia 2021 roku, to możemy dokonać ekstrapolacji szeregu czasowego. Kwiecień 2021 roku jest 16. wyrazem szeregu. Podstawiając do równania trendu $t = 16$ otrzymujemy:

$$\hat{y}_{(n+p)} = y_{16} = -0{,}48 \cdot 16 + 60{,}93 = 53{,}25$$

Określamy przedział prognozy:

$$\hat{y}_{(n+p)} - S_y < y_{(n+p)} < \hat{y}_{(n+p)} + S_y$$

$$53{,}25 - 0{,}58 < \hat{y}_{(n+p)} < 53{,}25 + 058$$

$$52{,}67 < \hat{y}_{IV\ 2021} < 53{,}83$$

Przewidywana liczba bezrobotnych niepełnosprawnych w kwietniu 2021 roku będzie w granicach od 52,67 tys. osób do 53,83 tys. osób, przy założeniu, że tendencja rozwojowa zjawiska nie ulegnie zmianie.

ROZDZIAŁ 4
ANALIZA WSPÓŁZALEŻNOŚCI ZJAWISK

Celem tego rodzaju analizy jest stwierdzenie, czy między badanymi zmiennymi zachodzą jakieś zależności? Jaka jest ich siła, kształt i kierunek? Współzależność między zmiennymi może być dwojakiego rodzaju: funkcyjna, stochastyczna (probabilistyczna).

Istota zależności funkcyjnej jest znana z matematyki. Zależność stochastyczna występuje wtedy, gdy wraz ze zmianą jednej zmiennej zmienia się rozkład prawdopodobieństwa drugiej zmiennej. Szczególnym przypadkiem zależności stochastycznej jest zależność korelacyjna.

4.1. Tablica korelacyjna

Badanie związków korelacyjnych ma sens tylko wtedy, gdy między nimi istnieje więź przyczynowo-skutkowa, dająca się logicznie wytłumaczyć. Gdy obserwacje statystyczne dotyczące badanych zmiennych są liczne, w celu stwierdzenia istnienia lub braku związku korelacyjnego konstruuje się tablicę korelacyjną.

Tablica korelacyjna składa się z dwóch szeregów statystycznych, podzielonych na kolumny i wiersze. Na skrzyżowaniu kolumn z wierszami wpisywane są liczebności jednostek, u których zaobserwowano występowanie określonej wartości cech x_i oraz y_j.

Tablica korelacyjna

y / x	y_1	y_2	...	y_j	...	y_r	$\sum_j$
x_1	n_{11}	n_{12}	...	n_{1j}	...	n_{1r}	$n_{1\bullet}$
x_2	n_{21}	n_{22}	...	n_{2j}	...	n_{2r}	$n_{2\bullet}$
⋮	⋮	⋮	...	⋮	...	⋮	⋮
x_i	n_{i1}	n_{i2}	...	n_{ij}	...	n_{ir}	$n_{i\bullet}$
⋮	⋮	⋮	...	⋮	...	⋮	⋮
x_k	n_{k1}	n_{k2}	...	n_{kj}	...	n_{kr}	$n_{k\bullet}$
$\sum_i$	$n_{\bullet 1}$	$n_{\bullet 2}$	...	$n_{\bullet j}$	...	$n_{\bullet r}$	n

W tablicy korelacyjnej zawarte są dwa rodzaje rozkładów:
- brzegowe,
- warunkowe.

Rozkład brzegowy prezentuje strukturę wartości jednej zmiennej (cechy) X lub Y bez względu na kształtowanie się wartości drugiej zmiennej. Wynika stąd, że w tablicy korelacyjnej są dwa rozkłady brzegowe. Rozkład brzegowy zmiennej X tworzy pierwsza i ostatnia kolumna tej tabeli, natomiast rozkład brzegowy zmiennej Y tworzy pierwszy i ostatni wiersz.

Rozkład warunkowy przedstawia strukturę wartości jednej zmiennej (X lub Y) pod warunkiem, że druga zmienna przyjęła określoną wartość. Rozkład warunkowy zmiennej X zapisujemy $X / Y = y_i$, natomiast rozkład warunkowy zmiennej Y zapisujemy $Y / X = x_i$. Rozkładów warunkowych zmiennej X jest tyle ile jest wariantów zmiennej Y i na odwrót.

Przykład 4.1.1

Wydajność pracy Y (w tys. sztuk wyrobów na osobę) oraz staż pracy X (w latach) pracowników w pewnym zakładzie podano w postaci tablicy korelacyjnej.

y / x	1-3	3-5	5-7	7-9	Razem
0-2	6	4	–	–	10
2-4	2	10	–	–	12
4-6	–	8	12	12	36
6-8	–	4	20	20	42
Razem	8	26	34	32	100

Rozkład brzegowy zmiennej X podaje strukturę wszystkich pracowników wg stażu pracy, niezależnie od wydajności.

Staż pracy w latach	Liczba pracowników
0-2	10
2-4	12
4-6	36
6-8	42
Razem	100

Rozkład brzegowy zmiennej *Y* przedstawia strukturę pracowników wg wydajności niezależnie od stażu pracy.

Wydajność w tys. szt/osobę	Liczba pracowników
1-3	8
3-5	26
5-7	34
7-9	32
Razem	100

Rozkładów warunkowych zmiennej *X* jest 4, gdyż tyle jest wariantów zmiennej *Y*. Rozkładów warunkowych zmiennej *Y* jest 4, gdyż tyle jest wariantów zmiennej *X*.

Wydajność w tys. szt/osobę	Liczba pracowników
1-3	6
3-5	4
5-7	–
7-9	–
Razem	10

Rozkład warunkowy zmiennej *X* dla stażu pracy 0-2 lata wg wydajności.

Staż pracy w latach	Liczba pracowników
0-2	4
2-4	10
4-6	8
6-8	4
Razem	26

Rozkład warunkowy zmiennej *Y* dla wydajności 3-5 tys. szt./osobę wg stażu pracy.

4.2. Miary korelacji

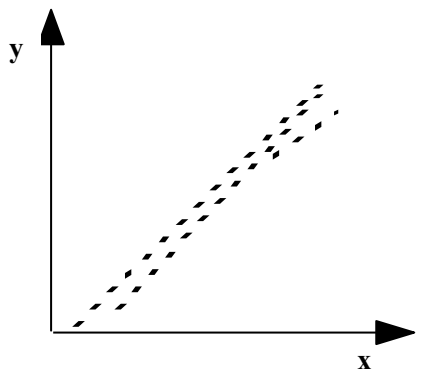
W tej części podręcznika zostaną scharakteryzowane współczynniki korelacji liniowej Pearsona i współczynnik korelacji rang Spearmana.

4.2.1.Współczynnik korelacji liniowej Pearsona

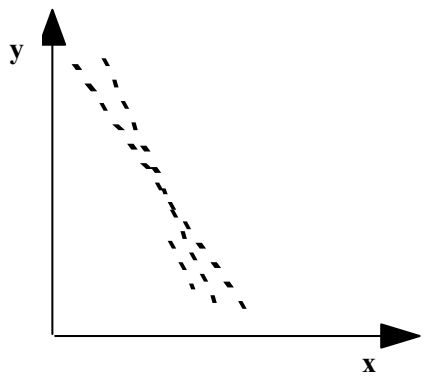
Zależność korelacyjna charakteryzuje się tym, że określonym wartościom jednej zmiennej przyporządkowane są ściśle określone średnie wartości drugiej zmiennej. Stopień zależności liniowej pomiędzy badanymi ce-

chami mierzalnymi określany jest za pomocą współczynnika korelacji liniowej r_{xy}.

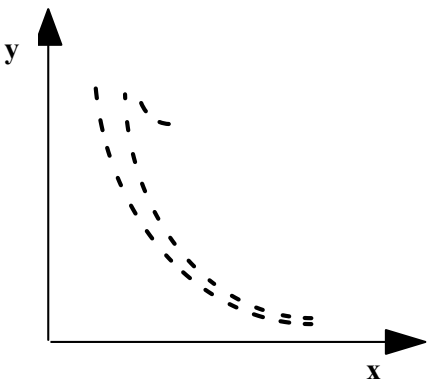
Zakładamy, że zbiorowość jest badana ze względu na dwie zmienne (cechy) X oraz Y a realizacje tych zmiennych w populacji lub próbie są zestawione w postaci dwóch szeregów szczegółowych. Najprostszą metodą określania siły i rodzaju zależności jest ocena wzrokowa. Na płaszczyźnie realizacjom zmiennych X i Y odpowiadają punkty o współrzędnych (x_i, y_i) $i=1,2,...,n$. Punkty odpowiadające poszczególnym wartościom cech tworzą korelacyjny wykres rozrzutu.



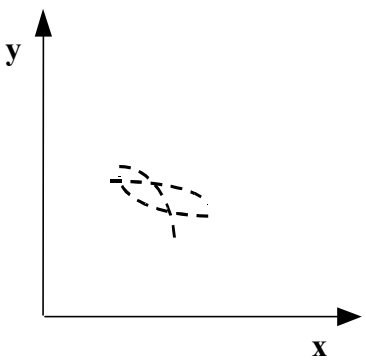
Rys. 4.1. Korelacja liniowa dodatnia $(r_{xy} > 0)$



Rys. 4.2. Korelacja liniowa ujemna $(r_{xy} < 0)$



Rys.4 3. Korelacja nieliniowa $(r_{xy} = 0)$



Rys. 4.4. Brak korelacji liniowej $(r_{xy} = 0)$

Korelacja dodatnia występuje wtedy, gdy wzrostowi jednej cechy odpowiada wzrost średnich wartości drugiej cechy.

Korelacja ujemna występuje wtedy, gdy wzrostowi jednej cechy odpowiada spadek średnich wartości drugiej cechy.

Współczynnik korelacji Pearsona, przyjmujący wartości z przedziału [−1,+1], jest miarą siły **związku liniowego** między cechami. Współczynnik ten wyznacza się z następującej zależności:

$$r_{xy} = \frac{\sum_{i=1}^{n}(x_i - \bar{x}) \cdot (y_i - \bar{y})}{\sqrt{\sum_{i=1}^{n}(x_i - \bar{x})^2 \cdot \sum_{i=1}^{n}(y_i - \bar{y})^2}} \qquad \text{lub} \qquad r_{xy} = \frac{C(X,Y)}{S_x \cdot S_y}$$

gdzie: $\bar{x}, \bar{y}$ – wartości średnie,

S_x, S_y – odpowiednie odchylenia standardowe,

$C(X,Y)$ – kowariancja między cechami

$$C(X,Y) = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{n}.$$

Znak współczynnika korelacji informuje o kierunku korelacji, jego bezwzględna wartość o sile związku. Jeżeli $r_{xy} = -1$ albo $r_{xy} = 1$ to oznacza, że między zmiennymi (cechami) zachodzi zależność w postaci funkcji liniowej. Gdy $r_{xy} = 0$ cechy nie są skorelowane, nie ma pomiędzy nimi zależności liniowej.

W analizie statystycznej oceniamy siłę związku pomiędzy cechami za pomocą współczynnika r_{xy} następująco:

- $0 < r_{xy} \leq 0{,}2$ – brak związku liniowego pomiędzy cechami,
- $0{,}2 < r_{xy} \leq 0{,}4$ – zależność liniowa pomiędzy cechami wyraźna, lecz niska,
- $0{,}4 < r_{xy} \leq 0{,}7$ – zależność liniowa pomiędzy cechami umiarkowana,
- $0{,}7 < r_{xy} \leq 0{,}9$ – zależność liniowa pomiędzy cechami znacząca, silna,
- $r_{xy} > 0{,}9$ – zależność liniowa pomiędzy cechami bardzo silna.

Uwagi:

a) r_{xy} bliski zeru oznacza brak zależności liniowej (może być inna),

a) wartość współczynnika korelacji r_{xy} zależy od zakresu zmienności badanych cech,

b) na podstawie małej liczby obserwacji nie należy obliczać r_{xy} (wynik może być błędny),

c) r_{xy} podlega wpływom wartości skrajnych, podobnie jak średnia arytmetyczna.

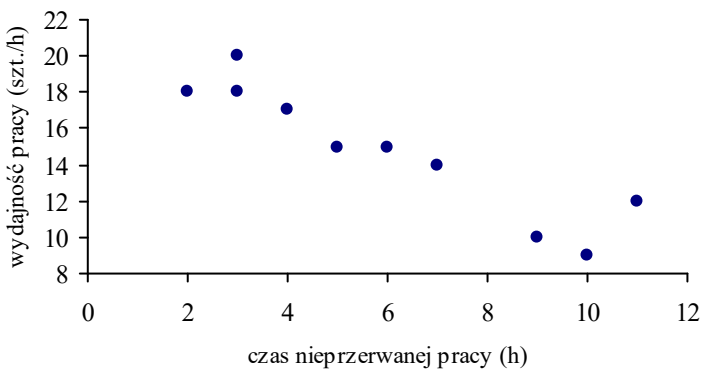
Przykład 4.2.1

Postanowiono dowiedzieć się, czy istnieje korelacja między wydajnością pracy robotników (Y) a czasem ich nieprzerwanej pracy (X). W celu sprawdzenia tego przypuszczenia pobrano próbę losową liczącą 10 robotników i uzyskano informacje:

Czas nieprzerwanej pracy x_i [h]	2	3	3	4	5	6	7	11	9	10
Wydajność pracy y_i [szt/h]	18	20	18	17	15	15	14	12	10	9

Stosując współczynnik korelacji liniowej Pearsona ocenić siłę i kierunek związku.

Rozwiązanie



Rys. 4.5. Wydajność pracy w zależności od czasu nieprzerwanej pracy

Tablica obliczeniowa

i	y_i	x_i	$y_i - \bar{y}$	$x_i - \bar{x}$	$(x_i - \bar{x})(y_i - \bar{y})$	$(y_i - \bar{y})^2$	$(x_i - \bar{x})^2$
1	18	2	3,2	-4	-12,80	10,24	16,00
2	20	3	5,2	-3	-15,60	27,04	9,00
3	18	3	3,2	-3	-9,60	10,24	9,00
4	17	4	2,2	-2	-4,40	4,84	4,00
5	15	5	0,2	-1	-0,20	0,04	1,00
6	15	6	0,2	0	0,00	0,04	0,00
7	14	7	-0,8	1	-0,80	0,64	1,00
8	12	11	-2,8	5	-14,00	7,84	25,00
9	10	9	-4,8	3	-14,40	23,04	9,00
10	9	10	-5,8	4	-23,20	33,64	16,00
Σ	148	60	X	X	-95,00	117,60	90,00

$$\bar{y} = \frac{\sum_{i=1}^{n} y_i}{n} = \frac{148}{10} = 14{,}8 \qquad \bar{x} = \frac{\sum_{i=1}^{n} x_i}{n} = \frac{60}{10} = 6{,}0$$

$$r_{xy} = \frac{\sum_{i=1}^{n} (x_i - \bar{x}) \cdot (y_i - \bar{y})}{\sqrt{\sum_{i=1}^{n} (x_i - \bar{x})^2 \cdot \sum_{i=1}^{n} (y_i - \bar{y})^2}} = \frac{-95{,}00}{\sqrt{90{,}00 \cdot 117{,}6}} = \frac{-95{,}00}{102{,}88} = -0{,}92$$

Pomiędzy wydajnością pracy robotników a czasem nieprzerwanej pracy istnieje silna ujemna zależność korelacyjna (współczynnik Pearsona równy –0,92).

4.2.2.Współczynnik korelacji rang Spearmana

Do opisu siły korelacji dwóch cech, wtedy gdy przynajmniej jedna ma charakter jakościowy i istnieje możliwość uporządkowania obserwacji empirycznych w określonej kolejności, służy współczynnik Spearmana. Miarę ta można stosować do badania zależności między cechami ilościowymi, ale w przypadku niewielkiej liczby obserwacji.

$$r_s = 1 - \frac{6 \sum_{i=1}^{n} d_i^2}{n(n^2 - 1)}$$

gdzie: d_i – różnice między rangami odpowiadających sobie wartości cechy x_i oraz y_i.

Sposób obliczania współczynnika rang Spearmana

1. Porządkujemy wyjściowe informacje według rosnących lub malejących wariantów jednej z cech. Uporządkowanym wartościom zmiennych nadajemy numery kolejnych liczb naturalnych (rangujemy). Sposób rangowania musi być jednakowy dla obu zmiennych:
   – rangujemy od największej do najmniejszej wartości lub odwrotnie,
   – gdy występują jednakowe wartości realizacji zmiennych, przyporządkowujemy im średnią arytmetyczną obliczoną z kolejnych numerów, mówimy wówczas, że wystąpiły węzły.
1. Jednakowe rangi wartości badanych zmiennych świadczą o dodatniej korelacji, przeciwstawna numeracja sugeruje istnienie korelacji ujemnej.

$$-1 \le r_s \le +1$$

2. Interpretacja identyczna jak dla współczynnika Pearsona r_{xy}. Im współczynnik jest bliższy +1 lub -1, tym silniejsza jest badana zależność.

Przykład 4.2.2

Ustalić natężenie współzależności między opiniami o nauczycielach: dyrektora szkoły i wizytatora. Opinie te zostały wydane na podstawie kontroli całokształtu pracy zawodowej i kwalifikacji nauczycieli. Wyniki kontroli ujęto w punktach.

Nauczyciele		A	B	C	D	E	F	G	H	I	J	K
Punkty	Dyrektora	41	27	35	33	25	47	38	53	43	35	36
	Wizytatora	38	24	34	29	27	47	43	52	39	31	29

Rozwiązanie

Punktowym wynikom oceny nauczycieli nadajemy rangi, przy czym największej liczbie punktów przypisujemy rangę 1.

Rangi ocen	Dyrektor	4	10	7,5	9	11	2	5	1	3	7,5	6
	Wizytator	5	11	6	8,5	10	2	3	1	4	7	8,5
Różnice rang	d_i	-1	-1	1,5	0,5	1	0	2	0	-1	0,5	-2,5
	d_i^2	1	1	2,25	0,25	1	0	4	0	1	0,25	6,25

Korzystając z wyników zamieszczonych w tabeli i wzoru otrzymujemy wartość badanego współczynnika Spearmana:

$$\sum_{i=1}^{11} d_i^2 = 17 \qquad r_s = 1 - \frac{6\sum_{i=1}^{n} d_i^2}{n(n^2-1)} = 1 - \frac{6 \cdot 17}{11(11^2-1)} = 0{,}92$$

Wynik $r_s = 0{,}92$ wskazuje, że współzależność opinii dyrektora i wizytatora jest bardzo silna. Oceniając nauczycieli zarówno dyrektor, jak też wizytator kierowali się podobnymi kryteriami.

4.3. Regresja liniowa

Badając związki zachodzące między zjawiskami lub cechami chcemy określić wpływ, jaki wywiera zmienna, będąca „przyczyną" na zmienną, która jest „skutkiem". Formalnym zapisem tego wpływu są funkcje regresji, które określają sposób przyporządkowania wartości zmiennej zależnej określonym wartościom zmiennej niezależnej.

Analizę regresji można wykorzystać do:

a) rozpoznania wielkości wpływu jednej z cech na drugą w związku przyczynowo-skutkowym,
b) objaśniania zmienności jednej cechy zmiennością drugiej, co ma szczególne znaczenie przy badaniu współwystępowania zjawisk,
c) szacowania nieznanych wartości jednej cechy na podstawie znanych lub założonych wartości drugiej cechy.

Funkcja regresji jest to funkcja matematyczna określonego typu, która jest przybliżeniem (aproksymantą) funkcyjnej zależności między zmiennymi. Postać funkcji określamy na podstawie zaobserwowanych wartości (x_i, y_i).

Należy zauważyć, że zaobserwowane wartości zmiennej zależnej będą się odchylały od funkcji także pod wpływem zmiennych nie uwzględnionych w badaniu oraz na skutek działania czynników przypadkowych.

W zależności od rodzaju związku pomiędzy zmiennymi, funkcje regresji mogą przyjmować postać liniową lub nieliniową (funkcja kwadratowa, wykładnicza, potęgowa, hiperboliczna, logarytmiczna).

4.3.1. Szacowanie parametrów liniowej funkcji regresji jednej zmiennej

Oszacowaniem funkcji regresji Y względem X w populacji generalnej jest funkcja regresji y względem x w próbie losowej (zwana aproksymantą).

$$\hat{y}_i = a_0 + a_x \cdot x_i + z_i$$

gdzie: i – numery cech,

a_x – określa o ile jednostek przeciętnie wzrośnie ($a_x > 0$) lub zmaleje ($a_x < 0$) wartość zmiennej zależnej gdy zmiennej niezależnej wzrośnie o jedną jednostkę,

a_0 – wolny wyraz w równaniu (nie ma najczęściej interpretacji ekonomicznej),

z_i – składnik resztowy służący do oceny dopasowania funkcji regresji do punktów empirycznych.

Oszacowaniem funkcji regresji X względem Y w populacji generalnej jest funkcja regresji x względem y w próbie losowej:

$$\hat{x}_i = a_0 + a_y \cdot x_i + z_i$$

Funkcje regresji są dobrymi aproksymantami funkcji liniowych, jeżeli spełnione są dwa warunki:

1. odchylenia wartości empirycznych y_i, x_i od wartości teoretycznych $\hat{y}_i, \hat{x}_i$ mają nieistotny charakter losowy,
2. suma kwadratów odchyleń wartości empirycznych od teoretycznych stanowi minimum.

Parametry odpowiedniej funkcji regresji najczęściej wyznacza się metodą najmniejszych kwadratów. Metoda ta opiera się na założeniu, że suma kwadratów odchyleń zaobserwowanych wartości zmiennej zależnej od wartości teoretycznych, obliczonych na podstawie wybranej funkcji, jest najmniejsza. Założenie to zapisuje się w postaci:

$$\sum_i (y_i - \hat{y}_i)^2 = min \text{ dla } \hat{y}_i = f(x_i)$$

oraz

$$\sum_i (x_i - \hat{x}_i)^2 = min \text{ dla } \hat{x}_i = f(y_i)$$

Analiza obu funkcji regresji jest uzasadniona wtedy, gdy między cechami występuje związek dwustronny, np.: między wielkością majątku trwałego i zatrudnieniem w pewnej branży przemysłu. Parametry tylko jednej funkcji regresji szacuje się wtedy, gdy związek ma wyraźnie charakter przyczynowo-skutkowy, np.: wielkość opadów i plony ziemniaka.

Linię regresji określa się jako miejsce geometryczne średnich wartości zmiennej zależnej przy ustalonych wartościach zmiennej niezależnej.

Niech funkcja regresji zmiennej zależnej (objaśnianej) Y przy danych wartościach zmiennej niezależnej (objaśniającej) X będzie oznaczona następująco:

$$\hat{y} = a_y \cdot x + b_y .$$

Metoda najmniejszych kwadratów (MNK) polega na takim oszacowaniu parametrów funkcji $\hat{y}$, aby dla danych z próby spełniony był warunek:

$$\sum_i (y_i - \hat{y}_i)^2 = \sum_i (y_i - a_y \cdot x - b_y)^2 \Rightarrow min$$

gdzie: y_i – wartości empiryczne cechy Y,

$\hat{y}_i$ - wartości teoretyczne cechy Y wyznaczone na podstawie funkcji regresji.

Obliczając miejsca zerowe pierwszych pochodnych cząstkowych względem odpowiednich parametrów funkcji otrzymujemy:

$$a_y = \frac{\sum_i (x_i - \bar{x}) \cdot (y_i - \bar{y})}{\sum_i (x_i - \bar{x})^2} \quad \text{lub} \quad a_y = \frac{C(X,Y)}{S_x^2}$$

$$b_y = \bar{y} - a_y \cdot \bar{x}$$

Analogicznie postępujemy w przypadku funkcji regresji zmiennej X względem Y

$$\hat{x} = a_x \cdot y + b_x$$

$$a_x = \frac{\sum_i (x_i - \bar{x}) \cdot (y_i - \bar{y})}{\sum_i (y_i - \bar{y})^2} \quad \text{lub} \quad a_x = \frac{C(X,Y)}{S_y^2}$$

$$b_x = \bar{x} - a_x \cdot \bar{y}$$

gdzie: S_x^2, S_y^2–odpowiednie wariancje,
$C(X,Y)$ – kowariancja.

Uwagi:

1. Parametry a_y, a_x noszą nazwę współczynników regresji.
2. Wartości współczynników regresji a_y, a_x określają o ile jednostek przeciętnie wzrośnie (zmaleje) wartość zmiennej zależnej, gdy wartość zmiennej niezależnej wzrośnie o jedną jednostkę.
3. Parametry b_x, b_y tylko niekiedy mają interpretację ekonomiczną.
4. Kowariancja $C(X,Y) = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{n}$.

4.3.2. Dopasowanie funkcji regresji do danych empirycznych

Do oceny dopasowania funkcji regresji do punktów empirycznych wykorzystuje się tzw. reszty, które stanowią różnicę pomiędzy wartościami empirycznymi a teoretycznymi funkcji regresji.

Dla regresji Y względem X reszty przedstawia wzór: $z_i = y_i - \hat{y}_i$, i = 1,2,...,n
gdzie: y_i – wartości empiryczne,
$\hat{y}_i$ – wartości teoretyczne cechy Y (wyznaczone z funkcji $\hat{y} = a_y \cdot x + b_y$).

Analogicznie wyznacza się reszty dla regresji X względem Y: $u_i = x_i - \hat{x}_i$.

Funkcja regresji jest poprawnie oszacowana, jeżeli wartości reszt są niewielkie i mają charakter losowy.

Wariancję składnika resztowego dla regresji Y względem X określa wzór:

$$S^2(z_i) = \frac{\sum_i (y_i - \hat{y}_i)^2}{n-k}$$

gdzie: k – liczba szacowanych parametrów (dla funkcji liniowej k=2)
n – liczba obserwacji.

Dla regresji X względem Y mamy: $S^2(u_i) = \dfrac{\sum_i (x_i - \hat{x}_i)^2}{n-k}$

Odchylenie standardowe reszt $s(z_i)$ lub $s(u_i)$, zwane też średnim błędem szacunku określa, o ile (średnio biorąc) wartości empiryczne odchylają się od wartości teoretycznych. Wraz ze wzrostem odchylenia standardowego reszt maleje „dobroć" oszacowania funkcji regresji.

W analizie regresji do oceny dopasowania funkcji regresji często stosowaną miarą jest ***współczynnik zbieżności*** φ^2 :

$$\varphi^2_{yx} = \frac{\sum_i (y_i - \hat{y}_i)^2}{\sum_i (y_i - \bar{y})^2}$$

Współczynnik zbieżności φ^2 przyjmuje wartości z przedziału domkniętego [0,1]. Im mniejszą wartość przyjmuje współczynnik zbieżności, tym lepsze jest dopasowanie funkcji regresji do punktów empirycznych.

Analogicznie dla regresji X względem Y:

$$\varphi^2_{xy} = \frac{\sum_i (x_i - \hat{x}_i)^2}{\sum_i (x_i - \bar{x})^2}$$

Współczynnikiem determinacji R^2 nazywa się wyrażenie: $R^2 = 1 - \varphi^2$.

W przypadku zależności liniowej współczynnik ten jest równy współczynnikowi korelacji liniowej, a więc:

$$R^2 = r_{yx}{}^2 = r_{xy}{}^2 = 1 - \varphi^2$$

Im bliżej jedności, tym „dobroć" dopasowania funkcji regresji do danych empirycznych jest lepsza.

Uwagi:

1. Współczynnik korelacji r_{xy} jest średnią geometryczną współczynników regresji: $r_{xy} = \pm\sqrt{a_x \cdot a_y}$. Znak r_{xy} jest taki, jak współczynników a_x, b_y.
2. Współczynniki regresji funkcji $\hat{y}, \hat{x}$ można wyznaczyć ze wzorów:

$$a_y = r_{yx} \cdot \frac{S_y}{S_x}$$

$$a_x = r_{xy} \cdot \frac{S_x}{S_y}$$

3. Funkcja regresji może służyć do przewidywania (prognozowania) wartości jednej cechy, przy ustalonym poziomie drugiej z nich.

Przykład 4.3.1

Przeprowadzono badanie dotyczące wytrzymałości na złamanie w kg (Y) spawanych prętów o różnej średnicy wyrażonej w mm (X) i otrzymano następujące wyniki:

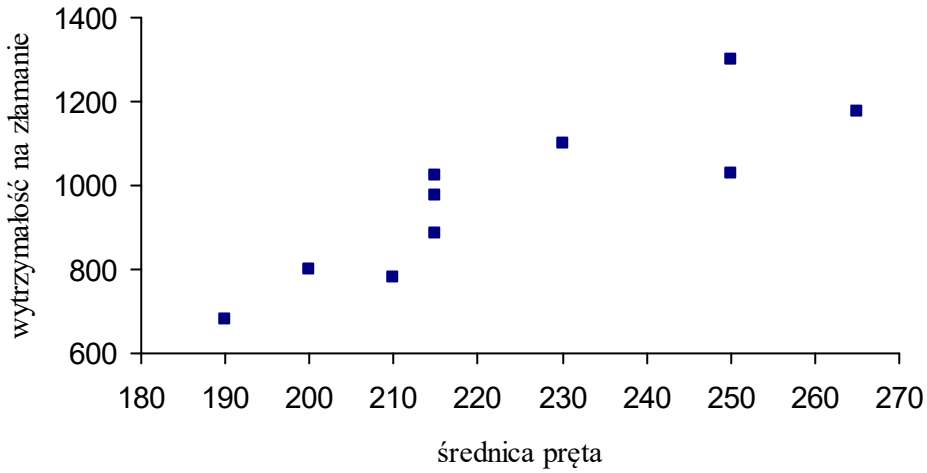
X	190	200	210	215	215	215	230	250	265	250
Y	680	800	780	885	975	1025	1100	1030	1175	1300

Na podstawie powyższych informacji:

1. Ocenić, czy istnieje współzależność między zmiennymi.
2. Ustalić siłę i kierunek nadanego związku.
3. Wyznaczyć teoretyczne linie regresji.
4. Sporządzić wykres linii regresji

Rozwiązanie

1. Sporządzamy wykres korelacyjny (p. rys. 4.6).



Rys. 4.6. Wytrzymałość na złamanie w zależności od średnicy pręta

Oceniając rozrzut punktów empirycznych na korelacyjnym wykresie rozrzutu możemy stwierdzić, że:

- zależność między badanymi zmiennymi występuje,
- jest to zależność o kierunku dodatnim,
- można oczekiwać, że jest związek korelacyjny silny, zależność o kształcie liniowym.

2. Założyliśmy, że zależność jest liniowa, zatem do oceny siły i kierunku tej zależności wykorzystamy wzór Pearsona:

$$r_{xy} = \frac{\sum_{i=1}^{n}(x_i - \bar{x}) \cdot (y_i - \bar{y})}{\sqrt{\sum_{i=1}^{n}(x_i - \bar{x})^2 \cdot \sum_{i=1}^{n}(y_i - \bar{y})^2}}$$

Tablica obliczeniowa

i	x_i	y_i	$x_i - \bar{x}$	$y_i - \bar{y}$	$(x_i - \bar{x})(y_i - \bar{y})$	$(x_i - \bar{x})^2$	$(y_i - \bar{y})^2$
1	190	680	-34	-295	10030	1156	87025
2	200	800	-24	-175	4200	576	30625
3	210	780	-14	-195	2730	196	38025
4	215	885	-9	-90	810	81	8100
5	215	975	-9	0	0	81	0
6	215	1025	-9	50	-450	81	2500
7	230	1100	6	125	750	36	15625
8	250	1030	26	55	1430	676	3025
9	265	1175	41	200	8200	1681	40000
10	250	1300	26	325	8450	676	105625
Σ	2240	9750	x	x	36150	5240	330550

$$\bar{x} = \frac{\sum_{i=1}^{n} x_i}{n} = \frac{2240}{10} = 224 \qquad \bar{y} = \frac{\sum_{i=1}^{n} y_i}{n} = \frac{9750}{10} = 975$$

$$r_{xy} = \frac{36150}{\sqrt{5240 \cdot 330550}} \approx 0{,}87$$

Pomiędzy wytrzymałością na złamanie, a średnicą pręta występuje silny związek korelacyjny o kierunku dodatnim. Zwiększenie średnicy pręta powoduje wzrost wytrzymałości na złamanie.

3. Wyznaczmy teoretycznie linie regresji

$$\hat{y} = a_y \cdot x + b_y$$

$$a_y = \frac{36150}{5240} = 6{,}899$$

$$b_y = 975 - 6{,}899 \cdot 224 = -570{,}376$$

$$\hat{y} = 6{,}899 \cdot x - 570{,}376$$

Jeżeli średnicę pręta zwiększymy o jednostkę, czyli o 1 mm to wzrośnie wytrzymałość na złamanie (w kg) średnio o 6,899 kg.

$$\hat{x} = a_x \cdot y + b_x$$

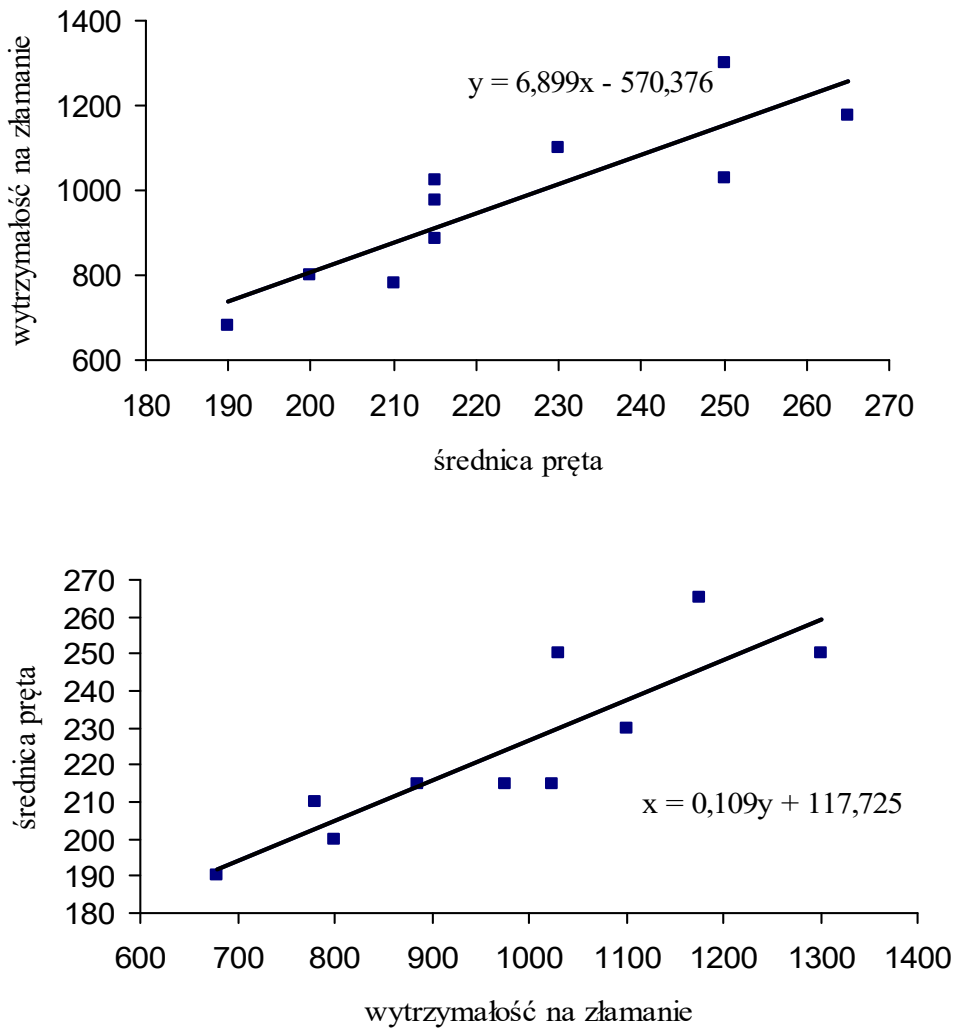
$$a_x = \frac{36150}{330550} = 0{,}109$$

$$b_x = 224 - 0{,}109 \cdot 975 = 117{,}725$$

$$\hat{x} = 0{,}109 \cdot y + 117{,}725$$

Jeżeli wytrzymałość na złamanie (w kg)zwiększymy o jednostkę, czyli o 1 kg to średnica pręta wzrośnie średnio 0,109 mm.

4. Wyznaczmy graficznie linie regresji (rys. 4.7).



Rys. 4.7. Wyznaczone graficznie linie regresji

Uwaga!

Sprawdź, czy $r_{xy} = \pm\sqrt{a_x \cdot a_y}$.

Przykład 4.3.2

Przeprowadzono badania wydatków na żywność w przeliczeniu na osobę (Y) w wybranych losowo rodzinach i dochodach (X) i uzyskano następujące równania regresji:

$$\hat{y} = 0{,}51 \cdot x + 17{,}5 \qquad \hat{x} = 1{,}53 \cdot y - 24{,}2$$

Ustalić siłę i kierunek badanego związku.

Rozwiązanie

Obliczamy współczynnik korelacji liniowej Pearsona, stosując wzór $r_{xy} = \pm\sqrt{a_x \cdot a_y}$.

$$r_{xy} = \pm\sqrt{0{,}51 \cdot 1{,}53} = \sqrt{0{,}78} \approx 0{,}88$$

Wynik oznacza, że między badanymi zmiennymi występuje silny związek korelacyjny, dodatni.

Przykład 4.3.3

Rozkład empiryczny czasu oczekiwania (Y) 100 losowo wybranych klientów na dostawę kanapy wykonywanej na zamówienie w prywatnej firmie podaje tabela.

Czas oczekiwania w dniach	Liczba klientów
10 – 20	10
20 – 30	30
30 – 40	40
40 – 50	20

Średni czas wykonania kanap (X) wynosił 15 dni, a jego względna dyspersja wynosiła 20%. Pomiędzy czasem oczekiwania a czasem wykonania występuje zależność liniowa, przy czym wydłużenie czasu wykonania o jeden dzień powoduje przedłużenie czasu oczekiwania średnio o dwa dni. Określić siłę i kierunek badanego związku.

Rozwiązanie

Y – czas oczekiwania na dostawę kanapy; X – czas wykonania kanapy.

Obliczamy współczynnik korelacji liniowej Pearsona, ponieważ założyliśmy, że badany związek jest liniowy.

$$a_y = r_{xy} \cdot \frac{S_y}{S_x} \quad \text{stąd} \qquad r_{xy} = a_y \cdot \frac{S_x}{S_y}$$

Z treści zadania wynika, że: $a_y = 2;\ \bar{x} = 15\ ;\ V_x = 20\%.$

Obliczamy S_x korzystając ze wzoru: $V_x = \frac{S_x}{\bar{x}} \cdot 100\%$

$$20\% = \frac{S_x}{15} \cdot 100\% \qquad \text{stąd} \qquad S_x = 0{,}2 \cdot 15 = 3$$

W celu obliczenia S_y wykonamy obliczenia pomocnicze.

Tablica obliczeniowa

y_i	n_i	$\dot{y}_i$	$\dot{y}_i n_i$	$(\dot{y}_i - \bar{y})^2$	$(\dot{y}_i - \bar{y})^2 n_i$
10 – 20	10	15	150	289	2890
20 – 30	30	25	750	49	1470
30 – 40	40	35	1400	9	360
40 – 50	20	45	900	169	3380
Razem	100	x	3200	x	8100

$$\bar{y} = \frac{1}{n} \sum_{i=1}^{n} \dot{y}_i \cdot n_i = \frac{3200}{100} = 32$$

$$S_y = \sqrt{\frac{1}{n} \sum_{i=1}^{n} (\dot{y}_i - \bar{y})^2 \cdot n_i} = \sqrt{\frac{1}{100} \cdot 8100} = \sqrt{81} = 9$$

stąd

$$r_{xy} = a_y \cdot \frac{S_x}{S_y} = 2 \cdot \frac{3}{9} = \frac{2}{3} \approx 0{,}67$$

Pomiędzy czasem oczekiwania na kanapę, a czasem wykonania istnieje umiarkowany związek korelacyjny o kierunku dodatnim. Wydłużonemu czasowi oczekiwania towarzyszy dłuższy czas wykonania (i odwrotnie).

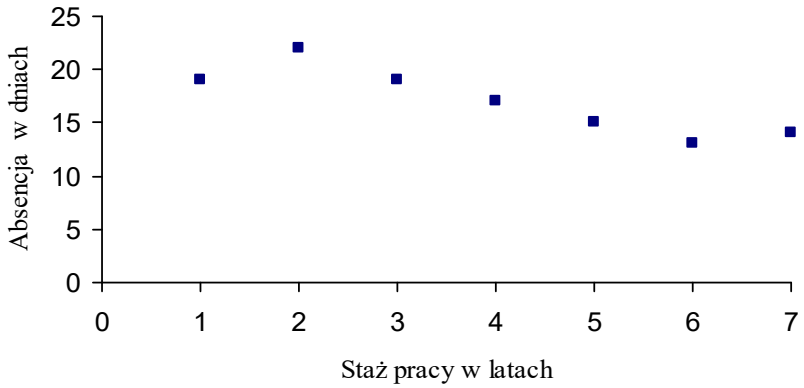
Przykład 4.3.4

Absencja pracowników w zależności od stażu pracy przedstawia się następująco:

Staż pracy w latach (X)	1	2	3	4	5	6	7
Absencja w dniach (Y)	19	22	19	17	15	13	14

1. Określić rodzaj badanej zależności na podstawie korelacyjnego wykresu rozrzutu oraz obliczyć współczynnik korelacji.
2. Wykonać wykres zależności Y od X.
3. Obliczyć r_{xy}.
4. Wyznaczyć rachunkowo i graficznie teoretyczną linię regresji.
5. Ocenić dopasowanie otrzymanej funkcji do danych empirycznych.

Rozwiązanie



Rys. 4.8. Zależność absencji od stażu pracy

Rysunek 4.8. rozpoczyna rozwiązanie zadania od analizy wzrokowej, widać współzależność, którą należy ocenić. Obliczamy na początku współczynnik korelacji liniowej Pearsona.

$$\bar{x} = \frac{\sum_{i=1}^{n} x_i}{n} = \frac{28}{7} = 4 \qquad \bar{y} = \frac{\sum_{i=1}^{n} y_i}{n} = \frac{119}{7} = 17$$

Tablica obliczeniowa

i	x_i	y_i	$x_i - \bar{x}$	$y_i - \bar{y}$	$(x_i - \bar{x})(y_i - \bar{y})$	$(y_i - \bar{y})^2$	$(x_i - \bar{x})^2$
1	1	19	-3	2	-6	4	9
2	2	22	-2	5	-10	25	4
3	3	19	-1	2	-2	4	1
4	4	17	0	0	0	0	0
5	5	15	1	-2	-2	4	1
6	6	13	2	-4	-8	16	4
7	7	14	3	-3	-9	9	9
Razem	28	119	x	x	-37	62	28

$$r_{xy} = \frac{\sum_{i=1}^{n}(x_i - \bar{x}) \cdot (y_i - \bar{y})}{\sqrt{\sum_{i=1}^{n}(x_i - \bar{x})^2 \cdot \sum_{i=1}^{n}(y_i - \bar{y})^2}} = \frac{-37}{\sqrt{28 \cdot 62}} = -0{,}89$$

Współczynnik korelacji liniowej Pearsona wynosi –0,89, co oznacza ujemną korelację i jest zgodne z rysunkiem 4.8. W dalszej kolejności należy wyznaczyć teoretycznie i graficznie linie regresji, a następnie dokonać oceny otrzymanej funkcji do danych empirycznych.

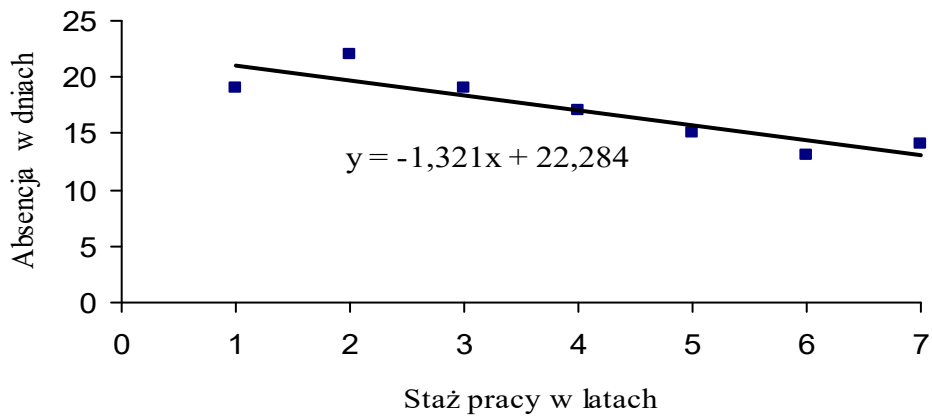
$$\hat{y} = a_y x + b_y$$

$$a_y = \frac{\sum_{i}(x_i - \bar{x}) \cdot (y_i - \bar{y})}{\sum_{1}(x_i - \bar{x})^2} = \frac{-37}{28} = -1{,}321$$

$$b_y = \bar{y} - a_y \bar{x} = -17 - (-1{,}32) \cdot 4 = 17 + 5{,}28 = 22{,}284$$

$$\hat{y} = -1{,}321x + 22{,}284$$

Jeżeli staż pracy wzrośnie o jeden rok, to wówczas absencja zmaleje średnio o 1,321 dnia. Poniżej zamieszczono rysunek 4.9 przedstawiający wyznaczoną graficznie linię regresji.



Rys. 4.9. Wyznaczona graficznie linia regresji

Pomiędzy absencją pracowników a stażem pracy istnieje silna, ujemna zależność korelacyjna, co oznacza, że wraz ze wzrostem stażu pracy maleje absencja.

Aby ocenić „dobroć" dopasowania funkcji do danych empirycznych obliczymy odchylenie standardowe reszt $s(z_i)$, współczynnik zbieżności φ^2 oraz współczynnik determinacji R^2. Do dalszych obliczeń sporządzimy tablicę pomocniczą.

Tablica obliczeniowa

i	x_i	y_i	$\hat{y}_i$	$(y_i - \hat{y}_i)^2$	$(y_i - \bar{y})^2$
1	1	19	20,963	3,853	4
2	2	22	19,642	5,560	25
3	3	19	18,321	0,461	4
4	4	17	17,000	0,000	0
5	5	15	15,679	0,461	4
6	6	13	14,358	1,844	16
7	7	14	13,037	0,927	9
Razem	28	119	x	13,107	62

$$s(z_i) = \sqrt{\frac{\sum_i (y_i - \hat{y}_i)^2}{n-k}} = \sqrt{\frac{13,107}{7-2}} = \sqrt{2,6214} = 1,619$$

Odchylenie standardowe składnika resztowego dla funkcji regresji Y względem X wskazuje, że przeciętne odchylenie wartości empirycznych absencji w dniach od wartości teoretycznych wynosi 1,619.

$$\varphi^2_{yx} = \frac{\sum_i (y_i - \hat{y}_i)^2}{\sum_i (y_i - \bar{y}_i)^2} = \frac{13,107}{62} = 0,211$$

Współczynnik zbieżności świadczy o dobrym dopasowaniu funkcji regresji do danych empirycznych. Tylko 21,1% zmiennej objaśnianej nie zostało wyjaśnione przez zmienną objaśniającą.

$$R^2 = 1 - \varphi^2 = 1 - 0,211 = 0,789$$

Współczynnik determinacji oznacza, że 78,9% zmian wartości absencji (stażu pracy) zostało wyjaśnione zmianami stażu pracy (absencji).

Uwaga!

Wyznacz funkcję regresji X względem Y. Sprawdź, czy proste regresji obliczone na podstawie danych z tego przykładu przecinają się w punkcie $P(\bar{x},\bar{y})$. Sporządź wykres.

4.4 Przykłady praktyczne

Przykład 4.4.1

Tabela korelacyjna przedstawia zależność ceny Y (w tys.zł za 1m^2) względem powierzchni użytkowej mieszkania X.

Cena Y / Powierzchnia X	4,5–5,5	5,5–6,5	6,5–7,5	7,5–8,5	Razem
45–65	9	3	–	–	12
65–85	2	2	1	1	6
85–105	–	1	2	–	3
105–125	–	–	1	-	1
Razem	11	6	4	1	22

Rozkład brzegowy zmiennej X podaje strukturę wszystkich mieszkań według powierzchni użytkowej, niezależnie od ceny.

Powierzchnia w m^2	Liczba mieszkań
45-65	12
65-85	6
85–105	3
105–125	1
Razem	22

Rozkład brzegowy zmiennej Y przedstawia strukturę ceny metra kwadratowego mieszkania, niezależnie od powierzchni.

Cena za m^2	Liczba mieszkań
4,5–5,5	11
5,5–6,5	6
6,5–7,5	4
7,5–8,5	1
Razem	22

Rozkładów warunkowych zmiennej X jest 4, gdyż tyle jest wariantów zmiennej Y. Podobnie rozkładów warunkowych zmiennej Y jest 4, gdyż tyle jest wariantów zmiennej X.

Rozkład warunkowy zmiennej X dla powierzchni użytkowej 45–65 m^2 przedstawia się następująco:

Cena za m^2	Liczba mieszkań
4,5–5,5	9
5,5–6,5	3
6,5–7,5	–
7,5–8,5	–
Razem	12

Rozkład warunkowy zmiennej *Y* dla ceny 5,5–6,5 tys. zł za m^2według powierzchni przedstawia się następująco:

Powierzchnia w m^2	Liczba mieszkań
45–65	3
65–85	2
85–105	1
105–125	–
Razem	6

Przykład 4.4.2

W pewnym zakładzie zakupiono urządzenie i obserwowano jego pracę przez 100 dni. Tabela korelacyjna przedstawia zależność liczby produkowanych braków *Y* (liczba sztuk/dzień) względem liczby awarii urządzenia *X* (liczba awarii/dzień).

Liczba braków *Y* / Liczba awarii *X*	0–4	5–9	10–14	Razem
0–5	7	12	17	36
6-11	12	44	8	64
Razem	19	56	25	100

Wyznaczyć mierniki rozkładów brzegowych zmiennych *X* i *Y* oraz mierniki rozkładów warunkowych tych zmiennych.

Rozwiązanie

1. Mierniki rozkładów brzegowych.

Tabela obliczeniowa dla zmiennej *X*

x_i	$n_{i\bullet}$	$\dot{x}_i$	$\dot{x}_i \cdot n_{i\bullet}$	$\dot{x}_i - \bar{x}$	$(\dot{x}_i - \bar{x})^2$	$(\dot{x}_i - \bar{x})^2 \cdot n_{i\bullet}$
0–5	36	2,5	90	-3,84	14,75	531
6–11	64	8,5	544	2,16	4,67	298,88
Razem	100	x	634	x	x	829,88

$$\bar{x} = \frac{1}{n}\sum_{i=1}^{k} \dot{x}_i \cdot n_{i\bullet} = \frac{634}{100} = 6{,}34 \text{ awarie/dziennie}$$

$$s^2 = \frac{1}{n}\sum_{i=1}^{k} (\dot{x}_i - \bar{x})^2 \cdot n_{i\bullet} = \frac{829{,}88}{100} = 8{,}3$$

Tabela obliczeniowa dla zmiennej Y

$\bar{y}_i$	$n_{\bullet j}$	$\dot{y}_i$	$\dot{y}_i \cdot n_{\bullet j}$	$\dot{y}_i - \bar{y}$	$(\dot{y}_i - \bar{y})^2$	$(\dot{y}_i - \bar{y})^2 \cdot n_{\bullet j}$
0-4	19	2	38	-5,3	28,09	533,71
5-9	56	7	392	-0,3	0,09	5,04
10-14	25	12	300	4,7	22,09	552,25
Razem	100	x	730	x	x	1091

$$\bar{y} = \frac{1}{n}\sum_{j=1}^{r} \dot{y}_i \cdot n_{\bullet j} = \frac{730}{100} = 7{,}3 \text{ braki/dziennie}$$

$$s^2 = \frac{1}{n}\sum_{j=1}^{r} (\dot{y}_j - \bar{y})^2 \cdot n_{\bullet j} = \frac{1091}{100} = 10{,}9$$

2. Mierniki rozkładów warunkowych.

Średnie warunkowe zmiennej X obliczamy ze wzoru:

$$\bar{x}_j = \frac{1}{n_{\bullet j}}\sum_{i=1}^{k} \dot{x}_i \cdot n_{ij}$$

Zatem:

$$\bar{x}_{1|\dot{y}=2} = \frac{2{,}5 \cdot 7 + 8{,}5 \cdot 12}{19} = 6{,}29$$

$$\bar{x}_{2|\dot{y}=7} = \frac{2{,}5 \cdot 12 + 8{,}5 \cdot 44}{56} = 7{,}21$$

$$\bar{x}_{3|\dot{y}=12} = \frac{2{,}5 \cdot 17 + 8{,}5 \cdot 8}{25} = 4{,}42$$

Liczba produkowanych braków y_j	$\dot{y}_i$	Średnia liczba awarii urządzenia $\bar{x}_j$	Liczba dni $n_{\bullet j}$
0-4	2	6,29	19
5-9	7	7,21	56
10-14	12	4,42	25

Średnie warunkowe zmiennej *Y* obliczamy ze wzoru:

$$\bar{y}_j = \frac{1}{n_i} \sum_{i=1}^{k} \dot{y}_i \cdot n_{ij}$$

Zatem:

$$\bar{y}_{1|\dot{x}=2,5} = \frac{2 \cdot 7 + 7 \cdot 12 + 12 \cdot 17}{36} = 8,39$$

$$\bar{y}_{2|\dot{x}=8,5} = \frac{2 \cdot 12 + 7 \cdot 44 + 12 \cdot 8}{64} = 6,69$$

Liczba awarii urządzenia x_i	$\dot{x}_i$	Średnia liczba produkowanych braków $\bar{y}_i$	Liczba dni $n_{i\bullet}$
0-5	2,5	8,39	36
6-11	8,5	6,69	64

W celu wyznaczenia wariancji średnich warunkowych wykonujemy obliczenia pomocnicze w tabeli obliczeniowej:

$\bar{x}_j$	$n_{\bullet j}$	$\bar{x}_j - \bar{x}$	$(\bar{x}_j - \bar{x})^2$	$(\bar{x}_j - \bar{x})^2 \cdot n_{\bullet j}$
6,29	19	-0,05	0,0025	0,05
7,21	56	0,87	0,7569	42,39
4,42	25	-1,92	3,6864	92,16
Razem	100	x	x	134,6

$$s^2(\bar{x}_j) = \frac{1}{n} \sum_{j=1}^{r} (\bar{x}_j - \bar{x})^2 \cdot n_{\bullet j} = \frac{134,6}{100} = 1,35$$

Podobnie dla zmiennej *Y* mamy:

$\bar{y}_i$	$n_{i\bullet}$	$\bar{y}_i - \bar{y}$	$(\bar{y}_i - \bar{y})^2$	$(\bar{y}_i - \bar{y})^2 \cdot n_{i\bullet}$
8,39	36	1,09	1,1881	42,77
6,69	64	-0,61	0,3721	23,81
Razem	100	x	x	66,58

$$s^2(\bar{y}_i) = \frac{1}{n} \sum_{i=1}^{r} (\bar{y}_i - \bar{y})^2 \cdot n_{i\bullet} = \frac{66,58}{100} = 0,67$$

Stosunek korelacji liniowej Pearsona:

$$e_{Y,X} = \sqrt{\frac{S^{2(\bar{y}_i)}}{S^2(y)}} = \sqrt{\frac{0,67}{10,61}} = 0,06$$

świadczy o bardzo niskiej zależności liczby produkowanych braków z awaryjnością urządzenia.

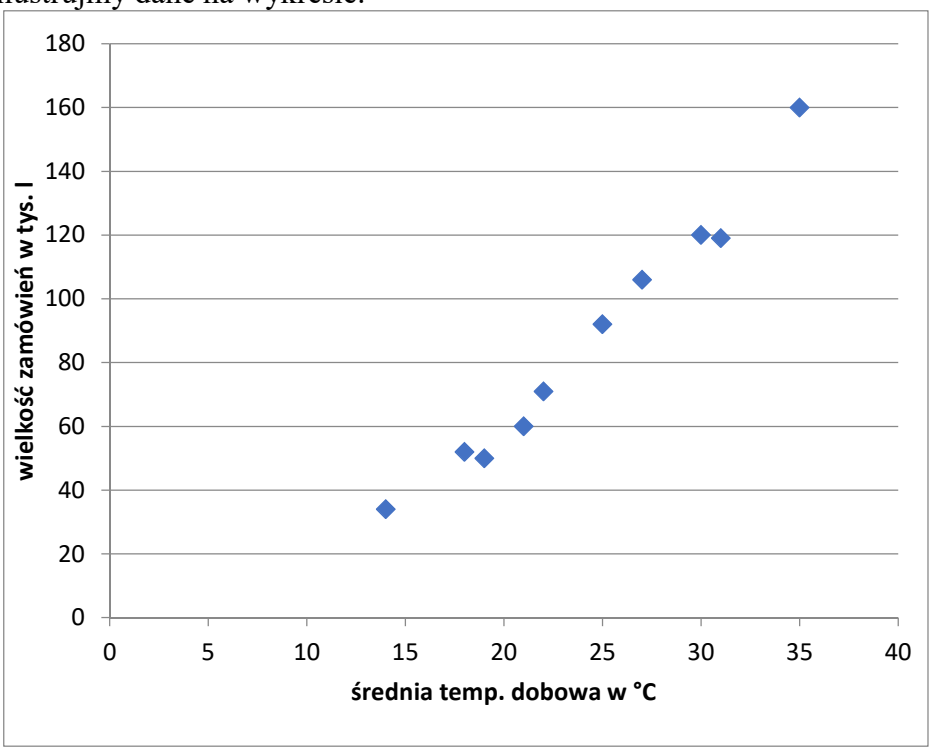
Przykład 4.4.3

Stosując współczynnik korelacji liniowej Pearsona ocenić siłę i kierunek związku między wielkością zamówień napojów chłodzących Y (w tys. litrów) i średnią temperaturą dobową X (w stopniach C) w okresie 10 losowo wybranych dni lipca.

Średnia temperatura dobowa w $°C$	19	25	31	21	35	18	14	27	30	22
Wielkość zamówień w tys. litrów	50	92	119	60	160	52	34	106	120	71

Rozwiązanie

Zilustrujmy dane na wykresie.



Rys. 4.10. Wielkość zamówień w zależności od średniej temperatury

Tablica obliczeniowa

i	y_i	x_i	$y_i - \bar{y}$	$x_i - \bar{x}$	$(x_i - \bar{x}) \cdot (y_i - \bar{y})$	$(y_i - \bar{y})^2$	$(x_i - \bar{x})^2$
1	50	19	-36,4	-5,2	189,28	1324,96	27,04
2	92	25	5,6	0,8	4,48	31,36	0,64
3	119	31	32,6	6,8	221,68	1062,76	46,24
4	60	21	-26,4	-3,2	84,48	696,96	10,24
5	160	35	73,6	10,8	794,88	5416,96	116,64
6	52	18	-34,4	-6,2	213,28	1183,36	38,44
7	34	14	-52,4	-10,2	534,48	2745,76	104,04
8	106	27	19,6	2,8	54,88	384,16	7,84
9	120	30	33,6	5,8	194,88	1128,96	33,64
10	71	22	-15,4	-2,2	33,88	237,16	4,84
Σ	864	242	x	x	2326,2	14212,4	389,6

$$\bar{y} = \frac{1}{n}\sum_{i=1}^{n} y_i = \frac{864}{10} = 86{,}4 \text{ tys. litrów}$$

$$\bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i = \frac{242}{10} = 24{,}2\ ^{\circ}\text{C}$$

$$r_{y,x} = \frac{\sum_{i=1}^{n}(x_i - \bar{x}) \cdot (y_i - \bar{y})}{\sqrt{\sum_{i=1}^{n}(y_i - \bar{y})^2 \cdot \sum_{i=1}^{n}(x_i - \bar{x})^2}} = \frac{2326{,}2}{\sqrt{14212{,}4 \cdot 389{,}6}} = 0{,}99$$

Między wielkością zamówień napojów chłodzących a średnią dobową temperaturą istnieje bardzo silna dodatnia zależność korelacyjna. Oznacza to, że wraz ze wzrostem temperatury rośnie wielkość zamówień napojów chłodzących, natomiast gdy maleje temperatura, wielkość zamówień napojów chłodzących maleje.

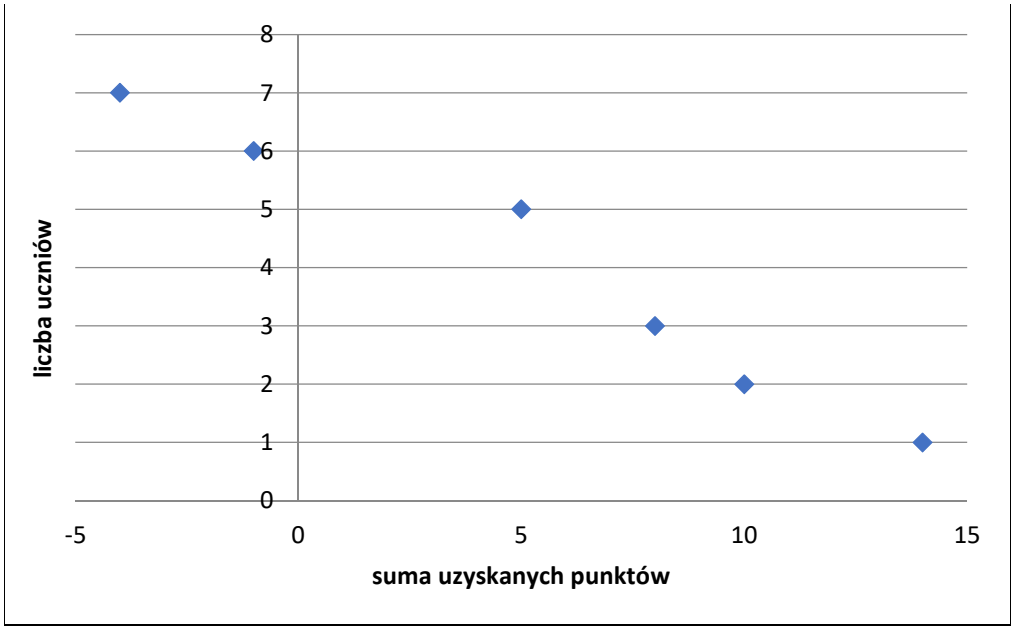
Przykład 4.4.4

W konkursie matematycznym uczniowie otrzymywali 1 punkt za prawidłową odpowiedź, 0 przy braku odpowiedzi, -1 punkt w przypadku błędnej odpowiedzi. Wyznaczyć współczynnik korelacji liniowej Pearsona dla danych zamieszczonych w tabeli poniżej.

Suma uzyskanych punktów x_i	8	5	-1	-4	10	14
Liczba uczniów y_i	3	5	6	7	2	1

Rozwiązanie

Zilustrujmy dane na wykresie (rys. 4.11).



Rys. 4.11. Liczba uczniów w zależności od sumy uzyskanych punktów

Budujemy tablicę obliczeniową

i	y_i	x_i	$y_i - \bar{y}$	$x_i - \bar{x}$	$(x_i - \bar{x}) \cdot (y_i - \bar{y})$	$(y_i - \bar{y})^2$	$(x_i - \bar{x})^2$
1	8	3	2,7	-1	-2,7	7,29	1
2	5	5	-0,3	1	-0,3	0,09	1
3	-1	6	-6,3	2	-12,6	39,69	4
4	-4	7	-9,3	3	-27,9	86,49	9
5	10	2	4,7	-2	-9,4	22,09	4
6	14	1	8,7	-3	-26,1	75,69	9
Σ	32	24	x	x	-79	231,34	28

$$\bar{y} = \frac{1}{n}\sum_{i=1}^{n} y_i = \frac{24}{6} = 5{,}3 \text{ uczniów}$$

$$\bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i = \frac{22}{6} = 24{,}2 \text{ punktów}$$

$$r_{y,x} = \frac{\sum_{i=1}^{n}(x_i - \bar{x}) \cdot (y_i - \bar{y})}{\sqrt{\sum_{i=1}^{n}(y_i - \bar{y})^2 \cdot \sum_{i=1}^{n}(x_i - \bar{x})^2}} = \frac{-79}{\sqrt{231{,}34 \cdot 28}} = \frac{-79}{80{,}48} = -0{,}98$$

Między liczbą uczniów a liczbą uzyskanych punktów istnieje bardzo silna liniowa zależność korelacyjna. Ujemny znak współczynnika wskazuje, że ze wzrostem sumy uzyskanych punktów maleje liczba uczniów, którzy je uzyskali.

Przykład 4.4.5

Walory wypoczynkowe województw zostały ocenione w skali 1-80 punktów. Zależność między walorami wypoczynkowymi Y a liczbą miejsc noclegowych X przedstawia poniższa tabela:

Walory wypoczynkowe *Y* Miejsca noclegowe w tys. *X*	1–40	41–80	Razem
0–25	7	2	9
26–51	2	3	5
52–77	-	2	2
Razem	9	7	16

Zbadać siłę i kierunek zależności.

Rozwiązanie

Siłę i kierunek zależności pomiędzy walorami wypoczynkowymi a liczbą miejsc noclegowych zbadamy za pomocą współczynnika korelacji Pearsona w postaci:

$$r_{x,y} = \frac{\sum_{i=1}^{n}(x_i - \bar{x}) \cdot (y_i - \bar{y})}{\sqrt{\sum_{i=1}^{n}(x_i - \bar{x})^2 \cdot \sum_{i=1}^{n}(y_i - \bar{y})^2}} = \frac{cov(X,Y)}{S_x \cdot S_y}$$

Obliczamy $\bar{x}$ *i* $\bar{y}$:

$\bar{x} = \frac{12{,}5 \cdot 9 + 38{,}5 \cdot 5 + 64{,}5 \cdot 2}{16} = 27{,}1$ tys. – średnia liczba miejsc noclegowych

$\bar{y} = \frac{20,5 \cdot 9 + 60,5 \cdot 7}{16} = 38$ punktów.– średnia liczba punktów walorów wypoczynkowych

Kowariancję zmiennych *X, Y* obliczymy korzystając z tabeli pomocniczej:

$\dot{x}_i - \bar{x}$	$\dot{y}_i - \bar{y}$		Razem
	20,5-38 = -17,5	60,5-38=22,5	
12,5-27,1=-14,6	7	2	9
38,5-27,1=11,4	2	3	5
64,5-27,1=37,4	–	2	2
Razem	9	7	16

$$cov(X,Y) = \frac{1}{16}[(-17,5)\cdot(-14,6)\cdot 7 + (-14,6)\cdot 22,5\cdot 2 + (-17,5)\cdot 11,4\cdot 2 + 11,4\cdot 22,5\cdot 3 +$$
$$+ 37,4\cdot 22,5\cdot 2] = \frac{1}{16}\cdot 3185 = 199,1$$

Obliczamy S_x i S_y:

$$S_x = \sqrt{\frac{1}{n}\sum_{i=1}^{k}(\dot{x}_i - \bar{x})^2 \cdot n_{i\bullet}}$$

$$S_x = \sqrt{\frac{(-14,6)^2 \cdot 9 + 11,4^2 \cdot 5 + 37,4^2 \cdot 2}{16}} = \sqrt{\frac{5365,76}{16}} = \sqrt{335,36} = 18,3$$

$$S_y = \sqrt{\frac{1}{n}\sum_{i=1}^{r}(\dot{y}_i - \bar{y})^2 \cdot n_{\bullet j}}$$

$$S_y = \sqrt{\frac{(-17,5)^2 \cdot 9 + 22,5^2 \cdot 7}{16}} = \sqrt{\frac{6300}{16}} = \sqrt{393,75} = 19,9$$

$$r_{x,y} = \frac{cov(X,Y)}{S_x \cdot S_y} = \frac{199,1}{18,3 \cdot 19,9} = \frac{199,1}{364,2} = 0,55$$

Pomiędzy walorami wypoczynkowymi a liczbą miejsc noclegowych istnieje umiarkowana dodatnia zależność liniowa, co oznacza, że wraz ze wzrostem liczby miejsc noclegowych rosną walory wypoczynkowe województwa i odwrotnie.

Przykład 4.4.6

Nadać rangi następującym wartościom: 3,1,10,7,8,3,1,9,7,3,7,3.

Rozwiązanie

Porządkujemy dane według rosnących (malejących) wartości:

1, 1, 3, 3, 3,3, 7, 7,7, 8, 9, 10

Uporządkowanym wartościom zmiennych nadajemy numery kolejnych liczb naturalnych (rangujemy).

1	2	3	4	5	6	7	8	9	10	11	12
1	1	3	3	3	3	7	7	7	8	9	10

Gdy występują jednakowe wartości zmiennych, to przyporządkowujemy im średnią arytmetyczną obliczoną z kolejnych numerów. Mówimy wówczas o wystąpieniu tzw. Węzłów.

W powyższym przykładzie węzłami są: 1, 3 i 7.

Ranga dla 1 jest średnią arytmetyczną numerów 1 i 2, czyli $\frac{1+2}{2} = 1,5$

Zatem ranga dla 3 wynosi $\frac{3+4+5+6}{4} = 4,5$

Natomiast dla 7 mamy: $\frac{7+9+8}{3} = 8$

Ostatecznie otrzymujemy:

Wartości	1	1	3	3	3	3	7	7	7	8	9	10
Rangi	1,5	1,5	4,5	4,5	4,5	4,5	8	8	8	10	11	12

Przykład 4.4.7

Dziesięciu wykładowców pewnej uczelni zostało ocenionych w skali 0-10 za jakość swojej pracy przez władze uczelni i przez organizację studencką. Wyniki oceny przedstawia tabela:

Wykładowcy	1	2	3	4	5	6	7	8	9	10
Punktacja władz	7	6	6	8	8	6	9	1	7	5
Punktacja studentów	5	4	5	4	7	6	9	2	5	4

Wyznaczyć współczynnik korelacji rang Spearmana i podać jego interpretację.

Rozwiązanie

Tablica obliczeniowa

Wykładowcy	Punktacja władz	Punktacja studentów	Rangi władz	Rangi studentów	d_i	d_i^2
1	7	5	6,5	6	0,5	0,25
2	6	4	4	3	1	1
3	6	5	4	6	-2	4
4	8	4	8,5	3	5,5	30,25
5	8	7	8,5	9	-0,5	0,25
6	6	6	4	8	-4	16
7	9	9	10	10	0	0
8	1	2	1	1	0	0
9	7	5	6,5	6	0,5	0,25
10	5	4	2	3	-1	1
Razem	x	x	x	x	x	53

Współczynnik korelacji rang Spearmana obliczamy ze wzoru:

$$r_s = 1 - \frac{6 \cdot \sum_{i=1}^{n} d_i^2}{n(n^2 - 1)}$$

Mamy: $r_s = 1 - \frac{6 \cdot 53}{10 \cdot 99} \approx 1 - 0{,}32 \approx 0{,}68$

Wynik powyższy oznacza, że pomiędzy oceną władz a oceną studentów zachodzi umiarkowana korelacja dodatnia.

Przykład 4.4.8

Istnieje przypuszczenie, że odległość między miejscem zamieszkania studenta a uczelnią ma wpływ na wyniki w nauce. Zebrano dane o 8 studentach nadając im odpowiednie rangi.

Studenci	1	2	3	4	5	6	7	8
Rangi dla odległości	1	2	3	4	5	6	7	8
Rangi dla wyników w nauce	6	4	8	2	1	7	3	5

Wyznaczyć współczynnik korelacji rang Spearmana i podać jego interpretację.

Rozwiązanie

Studenci	Rangi dla odległości	Rangi dla wyników w nauce	d_i	d_i^2
1	1	6	-5	25
2	2	4	-2	4
3	3	8	-5	25
4	4	2	2	4
5	5	1	4	16
6	6	7	-1	1
7	7	3	4	16
8	8	5	3	9
Razem	x	x	x	100

$$r_s = 1 - \frac{6 \cdot \sum_{i=1}^{n} d_i^2}{n(n^2 - 1)} = 1 - \frac{6 \cdot 100}{8 \cdot (64 - 1)} = \frac{600}{8 \cdot 83} = 1 - \frac{75}{63} \approx 1 - 1{,}19 \approx -0{,}19$$

Wartość współczynnika świadczy o braku związku liniowego między badanymi cechami, a jego ujemność oznacza, że im wyższa ranga studenta na liście odległości zamieszkania od uczelni, tym niższa jest jego ranga na liście wyników w nauce. Brak zależności między cechami wskazuje na nieistnienie merytorycznego związku pomiędzy cechami.

Przykład 4.4.9

W pewnym serwisie samochodowym prowadzono obserwację 100 aut. Analizowano, jak wygląda zależność wieku pojazdu Y od liczby napraw serwisowych X.

Wiek pojazdu w latach Y / Liczba napraw X	0–2	2–4	4–6	6–8	Razem
1	1	2	–	–	3
2	9	7	15	–	31
3	6	11	10	10	37
4	–	–	5	10	15
5	–	–	–	7	7
6	–	–	–	7	7
Razem	16	20	30	34	100

Na podstawie powyższych informacji sporządzić wykres empirycznych linii regresji.

Rozwiązanie

W celu wyznaczenia empirycznej linii regresji zmiennej X względem zmiennej Y obliczamy średnie warunkowe zmiennej X przy ustalonych wartościach zmiennej Y:

$$\bar{x}_{1|\dot{y}=1} = \frac{1\cdot 1+2\cdot 9+3\cdot 6}{16} = \frac{1+18+18}{16} = \frac{37}{16} = 2{,}31$$

$$\bar{x}_{2|\dot{y}=3} = \frac{1\cdot 2+2\cdot 7+3\cdot 11}{20} = \frac{2+14+33}{20} = \frac{49}{20} = 2{,}45$$

$$\bar{x}_{3|\dot{y}=5} = \frac{2\cdot 15+3\cdot 10+4\cdot 5}{30} = \frac{30+30+20}{30} = \frac{80}{30} = 2{,}67$$

$$\bar{x}_{4|\dot{y}=7} = \frac{3\cdot 10+4\cdot 10+5\cdot 7+6\cdot 7}{34} = \frac{30+40+35+42}{34} = \frac{147}{34} = 4{,}32$$

Współrzędne punktów empirycznej linii regresji zmiennej X względem zmiennej Y wynoszą:(2,31; 1), (2,45; 3), (2,67; 5), (4,32; 7).

W celu wyznaczenia empirycznej linii regresji zmiennej Y, względem zmiennej X, obliczamy średnie warunkowe zmiennej Y przy ustalonych wartościach zmiennej X:

$$\bar{y}_{1|\dot{x}=1} = \frac{1\cdot 1+2\cdot 3}{3} = \frac{1+6}{3} = \frac{7}{3} = 2{,}33$$

$$\bar{y}_{2|\dot{x}=2} = \frac{9\cdot 1+3\cdot 7+5\cdot 15}{31} = \frac{9+21+75}{31} = \frac{105}{31} = 3{,}39$$

$$\bar{y}_{3|\dot{x}=3} = \frac{6\cdot 1+11\cdot 3+10\cdot 5+10\cdot 7}{37} = \frac{6+33+50+70}{37} = \frac{159}{37} = 4{,}30$$

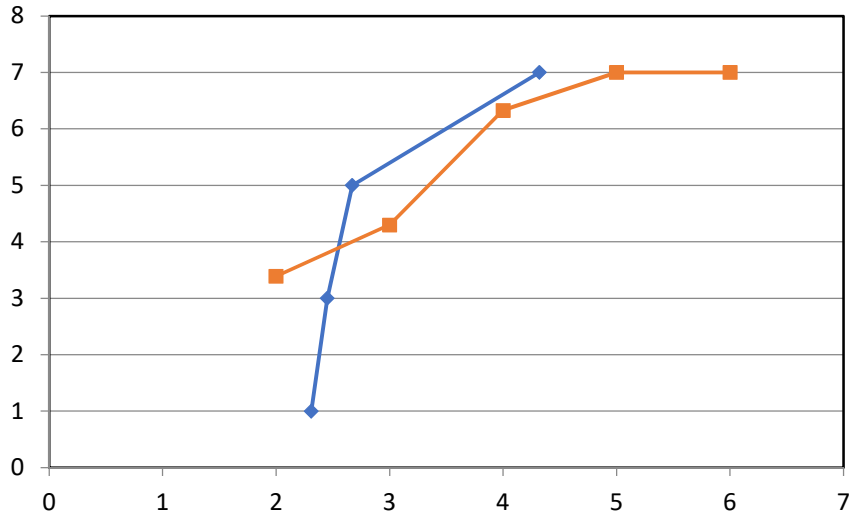
$$\bar{y}_{4|\dot{x}=4} = \frac{5\cdot 5+10\cdot 7}{15} = \frac{25+70}{15} = \frac{95}{15} = 6{,}33$$

$$\bar{y}_{5|\dot{x}=5} = \frac{7\cdot 7}{7} = \frac{49}{7} = 7$$

$$\bar{y}_{6|\dot{x}=6} = \frac{7\cdot 7}{7} = \frac{49}{7} = 7$$

Empiryczna linia regresji zmiennej Y względem zmiennej X ma współrzędne: (1; 2,33), (2; 3,39), (3; 4,30), (4; 6,33), (5; 7), (6; 7).

Empiryczne linie regresji przedstawia poniższy rysunek 4.12.



Rys. 4.12. Empiryczne linie regresji

Przykład 4.4.10

Tabela przedstawia wagę dziesięciu zwierząt (w kg) oraz odpowiadającą im wagę mózgu (w g)

Zwierzę	Koń	Słoń	Żyrafa	Krowa	Tygrys	Goryl	Krokodyl	Delfin	Orka	Szym-pans
Masa ciała x_i [kg]	300	6000	1300	800	180	200	350	80	4000	60
Masa mózgu y_i [g]	605	4800	680	425	275	540	175	1600	5600	420

Na podstawie powyższych danych oszacować parametry regresji zmiennej Y względem zmiennej X oraz zmiennej X względem zmiennej Y. Obliczyć siłę związku.

Rozwiązanie

Obliczenia pomocnicze wykonamy w poniższej tabeli:

Lp.	Masa ciała x_i	Masa mózgu y_i	$x_i - \bar{x}$	$y_i - \bar{y}$	$(x_i - \bar{x}) \cdot (y_i - \bar{y})$	$(x_i - \bar{x})^2$	$(y_i - \bar{y})^2$
1	300	605	-1027	-907	931489	1054729	822649
2	6000	4800	4673	3288	15364824	21836929	10810944
3	1300	680	-27	-832	22464	729	692224
4	800	425	-527	-1087	572849	277729	1181569
5	180	275	-1147	-1237	1418839	1315609	1530169
6	200	540	-1127	-972	1095444	1270129	944784
7	350	175	-977	-1337	1306249	954529	1787569
8	80	1600	-1247	88	-109736	1555009	7744
9	4000	5600	2673	4088	10927224	7144929	16711744
10	60	420	-1267	-1092	1383564	1605289	1192464
Razem	13270	15120	X	X	32913210	37015610	35681860

$$\bar{x} = \frac{13270}{10} = 1327\,\text{kg} \qquad \bar{y} = \frac{15120}{10} = 1512\,\text{g}$$

$$a_1 = \frac{\sum_{i=1}^{n}(x_i - \bar{x}) \cdot (y_i - \bar{y})}{\sum_{i=1}^{n}(x_i - \bar{x})^2} = \frac{32913210}{37015610} = 0{,}89$$

$$a_0 = \bar{y} - a_1 \cdot \bar{x} = 1512 - 0{,}89 \cdot 1327 = 1512 - 1181{,}03 = 330{,}97$$

$$\hat{y} = 330{,}97 + 0{,}89x$$

Jeżeli zwiększymy wagę zwierzęcia o 1 kilogram, wówczas waga mózgu wzrośnie relatywnie o 0,89 grama.

$$b_1 = \frac{\sum_{i=1}^{n}(x_i - \bar{x}) \cdot (y_i - \bar{y})}{\sum_{i=1}^{n}(y_i - \bar{y})^2} = \frac{32913210}{35681860} = 0{,}92$$

$$b_0 = \bar{x} - b_1 \cdot \bar{y} = 1327 - 0{,}92 \cdot 1512 = 1327 - 1391{,}04 = -64{,}04$$

$$\hat{x} = -64{,}04 + 0{,}92y$$

Jeżeli masa mózgu wzrośnie o 1 gram, wówczas masa ciała wzrośnie o 0,92 kilograma.

Współczynnik korelacji liniowej Pearsona wynosi:

$$r_{x,y} = \frac{\sum_{i=1}^{n}(x_i - \bar{x}) \cdot (y_i - \bar{y})}{\sqrt{\sum_{i=1}^{n}(x_i - \bar{x})^2 \cdot \sum_{i=1}^{n}(y_i - \bar{y})^2}} = \frac{32913210}{\sqrt{3701561035681860}} = \frac{32913210}{36342647} = 0{,}906$$

Pomiędzy masą ciała zwierzęcia a masą jego mózgu istnieje bardzo silna zależność korelacyjna.

Parametry funkcji regresji mogą być szacowane różnymi metodami, między innymi metodą wykorzystującą współczynnik korelacji liniowej Pearsona oraz średnie arytmetyczne i odchylenia standardowe. W wyniku tej metody otrzymamy następujące wzory:

$$a_1 = r_{x,y} \cdot \frac{S_y}{S_x} \qquad a_0 = \bar{y} - a_1 \cdot \bar{x}$$

$$b_1 = r_{x,y} \cdot \frac{S_x}{S_y} \qquad b_0 = \bar{x} - b_1 \cdot \bar{y}$$

Wykorzystując obliczenia z tabeli roboczej otrzymujemy:

$$S_x = \sqrt{\frac{\sum_{i=1}^{n}(x_i - \bar{x})^2}{n}} = \sqrt{\frac{37015610}{10}} = 1923{,}9$$

$$S_y = \sqrt{\frac{\sum_{i=1}^{n}(y_i - \bar{y})^2}{n}} = \sqrt{\frac{356818600}{10}} = 1888{,}96$$

$$a_1 = r_{x,y} \cdot \frac{S_y}{S_x} = 0{,}906 \cdot \frac{1888{,}96}{1923{,}9} = 0{,}89$$

$$a_0 = \bar{y} - a_1 \cdot \bar{x} = 1512 - 0{,}89 \cdot 1327 = 330{,}97$$

$$\hat{y} = 330{,}97 + 0{,}89x$$

$$b_1 = r_{x,y} \cdot \frac{S_x}{S_y} = 0{,}906 \cdot \frac{1923{,}9}{1888{,}96} = 0{,}92$$

$$b_0 = \bar{x} - b_1 \cdot \bar{y} = 1327 - 0{,}92 \cdot 11512 = -64{,}04$$

$$\hat{x} = 330{,}97 + 0{,}89y$$

Współczynnik korelacji liniowej Pearsona można również obliczyć korzystając z parametrów regresji:

$$r_{x,y} = \pm \cdot \sqrt{a_1 \cdot b_1}$$

Znak współczynnika jest taki sam, jak znak współczynników regresji.

Przykład 4.4.11

Na podstawie danych i obliczeń z przykładu 10 ocenić dopasowanie funkcji regresji do danych empirycznych.

Rozwiązanie

Obliczamy teoretyczne wartości: $\hat{y}_i$

$$\hat{y} = 330{,}97 + 0{,}89x$$

$$\hat{y}_1 = 330{,}97 + 0{,}89 \cdot 300 = 597{,}97 \approx 598$$
$$\hat{y}_2 = 330{,}97 + 0{,}89 \cdot 6000 = 5670{,}97 \approx 5671$$
$$\hat{y}_3 = 330{,}97 + 0{,}89 \cdot 1300 = 1487{,}97 \approx 1488$$
$$\hat{y}_4 = 330{,}97 + 0{,}89 \cdot 800 = 1042{,}97 \approx 1043$$
$$\hat{y}_5 = 330{,}97 + 0{,}89 \cdot 180 = 491{,}17 \approx 491$$
$$\hat{y}_6 = 330{,}97 + 0{,}89 \cdot 200 = 508{,}97 \approx 509$$
$$\hat{y}_7 = 330{,}97 + 0{,}89 \cdot 350 = 642{,}47 \approx 642$$
$$\hat{y}_8 = 330{,}97 + 0{,}89 \cdot 80 = 402{,}17 \approx 402$$
$$\hat{y}_9 = 330{,}97 + 0{,}89 \cdot 4000 = 3890{,}97 \approx 3891$$
$$\hat{y}_{10} = 330{,}97 + 0{,}89 \cdot 60 = 384{,}37 \approx 384$$

Obliczamy teoretyczne wartości $\hat{x}_i$:

$$\hat{x} = -64{,}04 + 0{,}92y$$

$$\hat{x}_1 = -64{,}04 + 0{,}92 \cdot 605 = 492{,}56 \approx 493$$
$$\hat{x}_2 = -64{,}04 + 0{,}92 \cdot 4800 = 4351{,}96 \approx 4352$$
$$\hat{x}_3 = -64{,}04 + 0{,}92 \cdot 680 = 561{,}56 \approx 562$$
$$\hat{x}_4 = -64{,}04 + 0{,}92 \cdot 425 = 326{,}96 \approx 327$$
$$\hat{x}_5 = -64{,}04 + 0{,}92 \cdot 275 = 188{,}96 \approx 189$$
$$\hat{x}_6 = -64{,}04 + 0{,}92 \cdot 540 = 432{,}76 \approx 433$$
$$\hat{x}_7 = -64{,}04 + 0{,}92 \cdot 175 = 96{,}96 \approx 97$$

$$\hat{x}_8 = -64{,}04 + 0{,}92 \cdot 1600 = 1407{,}96 \approx 1408$$
$$\hat{x}_9 = -64{,}04 + 0{,}92 \cdot 5600 = 5087{,}96 \approx 5088$$
$$\hat{x}_{10} = -64{,}04 + 0{,}92 \cdot 420 = 322{,}36 \approx 322$$

Kolejne obliczenia wykonamy w tablicy obliczeniowej.

y_i	$\hat{y}_i$	$y_i - \hat{y}_i$	$(y_i - \hat{y}_i)^2$	x_i	$\hat{x}_i$	$x_i - \hat{x}_i$	$(x_i - \hat{x}_i)^2$
605	598	7	49	300	493	-193	37249
4800	5671	-871	758641	6000	4352	1648	2715904
680	1488	-808	652864	1300	562	738	544644
425	1043	-618	381924	800	327	473	223729
275	491	-216	46656	180	189	-9	81
540	509	31	961	200	433	-233	54289
175	642	-467	218089	350	97	253	64009
1600	402	1198	1435204	80	1408	-1328	1763584
5600	3891	1709	2920681	4000	5088	-1088	1183744
420	384	36	1296	60	322	-262	68644
×	×	×	6416365	×	×	×	6655877

Obliczamy odchylenie standardowe składnika resztowego dla funkcji regresji Y względem X:

$$S_z = \sqrt{\frac{\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}{n-k}} = \sqrt{\frac{6416365}{10-2}} = \sqrt{802045{,}625} = 896$$

Odchylenie standardowe składnika resztowego dla funkcji regresji Y względem X wskazuje, że przeciętne odchylenie wartości empirycznych wagi mózgu od wartości teoretycznych wyznaczonych z funkcji regresji wynosi 896 gramów.

Obliczamy odchylenie standardowe składnika resztowego dla funkcji regresji X względem Y:

$$S_u = \sqrt{\frac{\sum_{i=1}^{n}(x_i - \hat{x}_i)^2}{n-k}} = \sqrt{\frac{6655877}{10-2}} = \sqrt{831984{,}625} = 912{,}13 \approx 912$$

Odchylenie standardowe składnika resztowego dla funkcji regresji X względem Y wskazuje, że przeciętne odchylenie wartości empirycznych wagi zwierzęcia od wartości teoretycznych wyznaczonych z funkcji regresji wynosi 912 kilogramów.

Teraz obliczamy współczynnik zbieżności dla funkcji Y względem X:

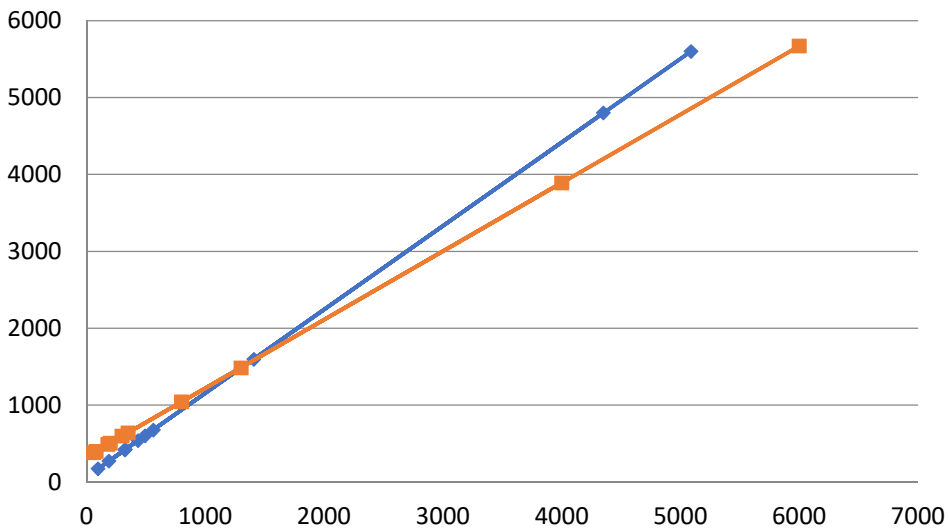
$$\varphi^2_{y,x} = \frac{\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}{\sum_{i=1}^{n}(y_i - \bar{y})^2} = \frac{6416365}{35681860} = 0{,}18$$

Dla funkcji X względem Y współczynnik zbieżności również wynosi 0,18, gdyż można go obliczyć wykorzystując współczynnik korelacji liniowej Pearsona $r_{x,y}$, który jest symetryczny, tzn.: $r_{x,y} = r_{y,x}$.

$$\varphi^2_{x,r} = 1 - (r_{x,y})^2 = 1 - (0{,}906)^2 = 0{,}18$$

Współczynnik zbieżności świadczy o dobrym dopasowaniu funkcji regresji do danych empirycznych. Tylko 18% informacji o zmiennej objaśnianej nie zostało wyjaśnione przez zmienną objaśniającą.

Graficznym obrazem równań regresji są teoretyczne linie regresji, które otrzymujemy nanosząc otrzymane punkty $(\hat{y}_i, x_i)$ dla równania $\hat{y} = 330{,}97 + 0{,}89x$ i punkty $(\hat{x}_i, y_i)$ dla $\hat{x} = -64{,}04 + 0{,}92y$ na układ współrzędnych (p. rys. 4.13).



Rys. 4.13.. Empiryczne linie regresji

Punkt przecięcia teoretycznych linii regresji ma współrzędne $(\bar{x}, \bar{y})$. Im bliżej siebie położone są obie linie regresji, tym zależność korelacyjna jest silniejsza.

Przykład 4.4.12

W pewnym Urzędzie Stanu Cywilnego przeanalizowano dane dotyczące wieku kobiety *X* i wieku mężczyzny *Y* w momencie zawierania małżeństwa. Uzyskano następujące informacje: średni wiek kobiety wyniósł 24 lata, średni wiek mężczyzny 26 lat, zmienność wieku kobiety 20,4%, zmienność wieku mężczyzny 24,3%, a współczynnik korelacji liniowej Pearsona $r_{x,y} = 0,95$.

a) wyznaczyć krzywą regresji wieku kobiety i oszacować wiek kobiety, gdy wiek mężczyzny wynosi 50 lat.
b) wyznaczyć krzywą regresji wieku mężczyzny i oszacować wiek mężczyzny przy wieku kobiety 30 lat.

Rozwiązanie

Z treści zadania wiemy, że:

$$\bar{x} = 24 \qquad \bar{y} = 26$$

$$V_x = 20,4\% \quad V_y = 24,3\% \quad r_{x,y} = 0,95$$

Mając dane współczynniki zmienności V_X i V_Y obliczymy odchylenia standardowe wieku kobiety i wieku mężczyzny.

$$V_x = \frac{S_x}{\bar{x}} \cdot 100\% \text{ czyli } s(x) = \bar{x} \cdot V_x = 24 \cdot 0,204 = 4,896$$

$$V_y = \frac{S_y}{\bar{y}} \cdot 100\% \text{ czyli } s(y) = \bar{y} \cdot V_y = 24 \cdot 0,243 = 6,318$$

Obliczamy parametry równania regresji wieku kobiety:

$$b_1 = r_{x,y} \cdot \frac{S_x}{S_y} = 0,95 \cdot \frac{4,896}{6,318} = 0,74 \quad b_0 = \bar{x} - b_1\bar{y} = 24 - 0,74 \cdot 26 = 4,76$$

Zatem krzywa regresji wieku kobiety ma postać: $\hat{x} = 4,76 + 0,74y$.

Jeśli wiek mężczyzny jest równy 50 lat, czyli $y_i = 50$,to szacowany wiek kobiety wynosi $\hat{x}_i = 4,76 + 0,74 \cdot 50 = 42$ lata.

Obliczamy parametry równania regresji wieku mężczyzny:

$$a_1 = r_{y,x} \cdot \frac{S_y}{S_x} = 0,95 \cdot \frac{6,318}{4,896} = 1,23 \quad a_0 = \bar{y} - a_1\bar{x} = 26 - 0,74 \cdot 24 = -3,52$$

Zatem krzywa regresji wieku mężczyzny ma postać: $\hat{y} = -3,52 + 1,23x$

Jeśli wiek kobiety jest równy 30 lat, czyli $x_i = 30$,to szacowany wiek mężczyzny wynosi:

$\hat{y}_i = -3,52 + 1,23 \cdot 30 = 33$ lata.

Przykład 4.4.13

Dokonano 7 obserwacji zmiennej X i Y. Dysponujemy tylko częściowo przetworzonymi informacjami o tych zmiennych: $\bar{x} = 4$, $\bar{y} = 17$, $V_x = 73{,}2\%$, $V_y = 11{,}8\%$, $\sum_{i=1}^{7} x_i y_i = 439$. Wyznaczyć współczynnik korelacji liniowej Pearsona, a następnie krzywe regresji Y względem X i X względem Y.

Rozwiązanie

Kowariancję zmiennych X i Y obliczamy ze wzoru:

$$C(X,Y) = \frac{1}{n}\sum_{i=1}^{n} x_i y_i - \bar{x} \cdot \bar{y}$$

Zatem $C(X,Y) = \frac{1}{n}\sum_{i=1}^{n} x_i y_i - \bar{x} \cdot \bar{y} = \frac{439}{7} = 4 \cdot 17 = 62{,}7 - 68 = -5{,}3$

Ze wzorów na współczynniki zmienności obliczamy odchylenia standardowe zmiennych X i Y.

$$V_x = \frac{S_x}{\bar{x}} \cdot 100\% \text{ czyli } S_x = \bar{x} \cdot V_x = 4 \cdot 0{,}732 = 6{,}318$$

$$V_y = \frac{S_y}{\bar{y}} \cdot 100\% \text{ czyli } S_y = \bar{y} \cdot V_y = 17 \cdot 0{,}118 = 2{,}006$$

Ostatecznie współczynnik korelacji liniowej Pearsona jest równy:

$$r_{x,y} = \frac{C(X,Y)}{S_x \cdot S_y} = \frac{-5{,}3}{2{,}928 \cdot 2{,}006} = \frac{-5{,}3}{5{,}874} = -0{,}90$$

Pomiędzy zmiennymi X i Y zachodzi silna ujemna zależność. Oznacza to, że wraz ze wzrostem zmiennej X maleje wartość zmiennej Y i odwrotnie.

Obliczamy parametry równania regresji Y względem X:

$$a_1 = r_{x,y} \cdot \frac{S_y}{S_x} = -0{,}90 \cdot \frac{2{,}006}{2{,}928} = -0{,}62$$

$$a_0 = \bar{y} - a_1 \bar{x} = 17 - (-0{,}62) \cdot 4 = 17 + 2{,}48 = 19{,}48$$

$$\hat{y} = 19{,}48 - 0{,}62x$$

Jeżeli wartość zmiennej X wzrośnie o 1, wówczas wartość zmiennej Y zmaleje o 0,62.

Obliczamy parametry równania regresji X względem Y:

$$b_1 = r_{xy} \cdot \frac{S_x}{S_y} = -0{,}90 \cdot \frac{2{,}928}{2{,}006} = -1{,}31$$

$$b_0 = \bar{x} - b_1 \bar{y} = 4 - (-1{,}31) \cdot 17 = 4 + 22{,}27 = 26{,}27$$

$$\hat{x} = 26{,}27 - 1{,}31y$$

Jeżeli wartość zmiennej *Y* wzrośnie o 1, wówczas wartość zmiennej *X* zmaleje o 1,31.

ROZDZIAŁ 5
ZADANIA DO SAMODZIELNEGO ROZWIĄZANIA

Zadanie 1
Oblicz ocenę ukończenia studiów pewnego studenta, który uzyskał średnią z egzaminów i zaliczeń 3,5; z pracy dyplomowej 4,5; z egzaminu końcowego 4, wiedząc że wagi poszczególnych ocen są równe odpowiednio: 0,6; 0,2; 0,2.

Zadanie 2
Roczny procentowy przyrost przychodów pewnej firmy farmaceutycznej w kolejnych czterech latach wynosił 10%, 20%. 5%, 15%. Jaki był średni przyrost dochodów w tym okresie?

Zadanie 3
Badano wydajność 20 serwisantów. Otrzymane dane, dotyczące czasu usuwania określonej awarii, uporządkowano niemalejąco: 48, 52, 53, 54, 56, 64, 65, 68, 68, 68, 70, 72, 72, 73, 74, 76, 83, 87, 89, 120. Oblicz średnią, wariancję, odchylenie standardowe, wartość modalną, kwantyle, typowy przedział zmienności i współczynnik zmienności czasu usuwania awarii.

Zadanie 4
Dla podanego szeregu przedziałowego uzupełnij tabelę i wyznacz: średnią, wariancję, odchylenie standardowe, wartość modalną, kwantyle, typowy przedział zmienności i współczynnik zmienności.

$< x_i, x_{i+1})$	n_i	$\dot{x}_i$	$\dot{x}_i \cdot n_i$	$\dot{x}_i - \bar{x}$	$(\dot{x}_i - \bar{x})^2$	$(\dot{x}_i - \bar{x})^2 \cdot n_i$
9-11	1					
11-13	2					
13-15	17					
15-17	25					
17-19	35					
19-21	16					
21-23	3					
23-25	1					
Razem						

Zadanie 5

W 10-osobowej grupie studentów przeprowadzono badanie w celu zrozumienia, czy ilość czasu poświęconego na naukę ma wpływ na wynik egzaminu. Oto dane dotyczące czasu nauki w godzinach x_i oraz wyniku egzaminu w procentach.

x_i	5	3	6	4	7	8	2	5	4	6
y_i	78	65	89	72	96	98	55	80	68	85

Wyznacz siłę związku liniowego obu cech, obliczając współczynnik korelacji tych cech.

Zadanie 6

Dziesięciu uczniów rozwiązywało dwa testy psychologiczne. Wyniki testów podane w punktach przedstawiały się następująco:

X	6	7	9	2	9	8	5	6	5	3
Y	3	6	9	4	9	8	4	6	5	4

Wykorzystując współczynnik korelacji rang Spearmana oblicz, jak silna zależność występuje między wynikami testów?

Zadanie7

W 15 osobowej grupie studentów zarządzania przeprowadzono badanie ze względu na parę cech (X, Y), gdzie X – ocena końcowa z matematyki, Y – ocena końcowa ze statystyki. Otrzymano wyniki: (3,4), (4,4), (5,5), (5,4), (2,2), (2,3), (2,2), (3,4), (3,3), (3,2), (2,3), (4,5), (3,3), (2,2), (4,4). Oblicz współczynnik korelacji Pearsona i wyznacz równania regresji między tymi cechami.

Zadanie 8

Na podstawie danych z zadania 7 skonstruuj tablicę korelacyjną i wyznacz rozkłady brzegowe cechy X i Y.

Zadanie 9

Korzystając z tablicy korelacyjnej z zadania 8 wyznacz szeregi warunkowe $X / Y = v_J$ oraz szeregi warunkowe $Y / X = w_J$.

Zadanie 10

Poniższe dane opisują dochód obywateli Stanów Zjednoczonych (w mld $) dla lat 2005-2014.

Rok	2005	2006	2007	2008	2009	2010	2011	2012	2013	2014
Dochód	1265	1391	1540	1733	1951	2165	2429	2585	2744	3012

Wyznacz indeks prosty przyjmując za rok bazowy 2009.

Zadanie 11

Poniższe dane to ceny kawy brazylijskiej (w $ za jednostkę hurtową) w latach 2001-2006.

Rok	2001	2002	2003	2004	2005	2006
Cena	25,8	21,2	33,1	52,4	21,3	13,3

Skonstruuj indeks prosty ceny hurtowej dla tego towaru, przyjmując za rok bazowy 2002.

Zadanie 12

Poniższe dane to ceny jednostki hurtowej ryżu, pszenicy i orzeszków ziemnych w latach 2011-2015.

Lata	2011	2012	2013	2014	2015
Ryż	316	252	323	326	266
Pszenica	332	275	345	345	285
Orzeszki	124	108	87	87	84

Skonstruuj indeks agregatowy dla tych trzech dóbr, przyjmując za rok bazowy 2011.

Zadanie 13

Uzupełnij tabelę i wyznacz $\bar{x}, s^2, s, v, \mu_3, A_S, K$. Zinterpretuj uzyskane wyniki.

$< x_i, x_{i+1})$	n_i	$\dot{x}$	$\dot{x} \cdot n_i$	$\dot{x} - \bar{x}$	$(\dot{x} - \bar{x})^2$	$(\dot{x} - \bar{x})^2 \cdot n_i$	$(\dot{x} - \bar{x})^3$	$(\dot{x} - \bar{x})^3 \cdot n_i$
11-13	2							
13-15	7							
15-17	20							
17-19	30							
19-21	8							
21-23	3							
Razem								

Zadanie 14
Badano wysokości kredytów w tysiącach złotych udzielone przez pewien bank w ciągu lutego 2020 roku. Otrzymane dane są przedstawione w szeregu rozdzielczym.

Wysokość kredytu	0-10	10-20	20-30	30-40	40-50	50-60	60-70	Razem
Liczba kredytów	32	88	73	58	25	16	8	300

Wyznacz średnią arytmetyczną, medianę, dominantę, wariancję, odchylenie standardowe, rozstęp, współczynnik zmienności, typowy przedział wielkości kredytów i wskaźnik asymetrii.

Zadanie 15
Badano dodatek do wynagrodzenia (w zł) 40 pracowników pewnego przedsiębiorstwa. Otrzymano następujące dane: 305, 320, 311, 327, 379, 340, 378, 368, 337, 352,321, 314, 302, 322, 362, 331, 314, 337, 305, 390, 325, 325, 300, 332, 347, 385, 319, 300, 325, 358, 339, 360, 305, 369, 306, 331, 312, 387,316, 315.
Przedstaw powyższe dane w szeregu rozdzielczym przedziałowym.

Rozwiązania

Zadanie 1
$\overline{X}_w = 3,8$

Zadanie 2
$\overline{X}_G = 12,36\%$

Zadanie 3
$\bar{x} = 70,6$; $s^2 = 250,14$; $s = 15,8$; $M_o = 68$; $Q_1 = 60$; $Q_2 = 69$; $Q_3 = 75$; $v_S = 22,4\%$; $54,8 < x_{typ} < 86,4$

Zadanie 4
$\bar{x} = 17,1$; $s^2 = 5,8$; $s = 2,4$; $M_o = 17,7$; $Q_1 = 15,4$; $Q_2 = 17,3$; $Q_3 = 18,7$; $v_S = 14,04\%$; $15,7 < x_{typ} < 19,5$

Zadanie 5

$r_{x,y} = 0,99$

Zadanie 6

$r_S = 0,82$

Zadanie 7

$r_{x,y} = 0,79$; $\hat{y} = 0,78x + 0,88$; $\hat{x} = 0,80y + 0,45$

Zadanie 8

Tablica korelacyjna

y / x	2	3	4	5
2	3	2		
3	1	2	2	
4			2	1
5			1	1

Rozkład brzegowy cechy X

Ocena z matematyki	Liczebność ocen
2	5
3	5
4	3
5	2
Razem	15

Rozkład brzegowy cechy Y

Ocena ze statystyki	Liczebność ocen
2	4
3	4
4	5
5	2
Razem	15

Zadanie 9

Szereg warunkowy $X/Y = 2$

Ocena z matematyki	Liczebność ocen
2	3
3	1
Razem	4

Szereg warunkowy $X/Y = 3$

Ocena z matematyki	Liczebność ocen
2	2
3	2
Razem	4

Szereg warunkowy $X/Y = 4$

Ocena z matematyki	Liczebność ocen
3	2
4	2
5	1
Razem	5

Szereg warunkowy $X/Y = 5$

Ocena z matematyki	Liczebność ocen
4	1
5	1
Razem	2

Rozkład warunkowy $Y/X = 2$

Ocena ze statystyki	Liczebność ocen
2	3
3	2
Razem	5

Rozkład warunkowy $Y/X = 3$

Ocena ze statystyki	Liczebność ocen
2	1
3	2
4	2
Razem	5

Rozkład warunkowy $Y/X = 4$

Ocena ze statystyki	Liczebność ocen
4	2
5	1
Razem	3

Rozkład warunkowy *Y/X = 5*

Ocena ze statystyki	Liczebność ocen
4	1
5	1
Razem	2

Zadanie 10
64,84; 71,30; 78,93; 88,83; 100; 110,97; 124,50; 132,50; 140,65; 154,38

Zadanie 11
121,70; 100; 156,13; 247,17; 100,47; 62,74

Zadanie 12
772; 635; 755; 758; 635; 100; 82,25; 97,80; 98,19; 82,25

Zadanie 13
$\bar{x} = 17$; $s^2 = 4,43$; $s = 2,1$; $v = 12,35\%$; $\mu_3 = 2,31$; $A_S = 0,25$; K = 3,22

Zadanie 14
$\bar{x} = 26,2$ tys. zł; $s^2 = 213,89$ tys. zł; $s = 14,63$ tys. zł; $v = 56\%$; $A_S = 0,57$; $M_o = 17,89$ tys. zł; $Q_2 = 24,11$ tys. zł; $R = 70$ tys. zł; $11,6 \text{ tys. zł} < x_{typ} < 40,8 \text{ tys. zł}$

Zadanie 15

$< x_i, x_{i+1})$	<291,309)	<309,327)	<327,345)	<345,363)	<363,381)	<381,399)
n_i	7	13	8	5	4	3

BIBLIOGRAFIA

1. Aczel A.D., Sounderpandian J., *Statystyka w zarządzaniu*, PWN, Warszawa 2017.
2. Bąk I., Markowicz I., Mojsiewicz M., Wawrzyniak K., *Statystyka opisowa. Przykłady i zadania*, CeDeWu, Warszawa 2020.
3. Buga J.(red.), *Statystyka opisowa w przykładach*, Skrypt Politechniki Radomskiej 2000.
4. Gdakowicz A., Hozer-Kocmel M., Markowicz I., *Zastosowanie metod opisu statystycznego do badania zjawisk społeczno-ekonomicznych,* CeDeWu, Warszawa 2022.
5. Iwasiewicz A., Paszek Z., *Statystyka z elementami statystycznych metod sterowania jakością*, Wydawnictwo AE, Kraków 2000.
6. Jóźwiak J., Podgórski J., *Statystyka od podstaw*, PWE, Warszawa 2022.
7. Kassyk-Rokicka H., *Statystyka. Zbiór zadań*, PWE, Warszawa 2014.
8. King B., Minium E., *Statystyka dla psychologów i pedagogów*, PWN, Warszawa 2020.
9. Kukuła K., *Elementy statystyki w zadaniach*, PWN, Warszawa 2016.
10. Luszniewicz A., Słaby T., *Statystyka* z pakietem *komputerowym STATISTICA PL*, Wydawnictwo C.H.Beck, Warszawa 2008.
11. Makać W., Urbanek-Krzysztofiak D., *Metody opisu statystycznego*, UG, Gdańsk 2020.
12. Maksimowicz-Ajchel A., *Wstęp do statystyki. Metody opisu statystycznego*, Wydawnictwa UW, Warszawa 2007.
13. Maksimowicz-Ajchel A., *Zarys statystyki. Podręcznik do nauki zawodu. Branża ekonomiczna*, WSiP, Warszawa 2013.
14. Nielsen A., *Szeregi czasowe*, Helion, Gliwice 2020.
15. Olbrych B., Sagan T., *Statystyka. Podstawy teorii i zastosowania,* Wydawnictwo WSH, Radom 2003.
16. Ostasiewicz S., Rusnak Z., Siedlecka U., *Statystyka. Elementy teorii i zadania*, wyd. 6, Wydawnictwo AE we Wrocławiu, Wrocław 2006.
17. Panek T., *Statystyka społeczna*, PWE, Warszawa 2007.
18. Pilatowska M., *Repetytorium ze statystyki*, PWN, Warszawa 2016.
19. Podgórski J., *Statystyka dla studiów licencjackich*, PWE, Warszawa 2022.
20. Ręklewski M., *Statystyka opisowa. Teoria i przykłady*, Wydawnictwo Państwowej Uczelni Zawodowej we Włocławku, Włocławek 2020.
21. Rumsey D.J., *Statystyka dla bystrzaków*, Helion, Gliwice 2023.

22. Sobczak M., *Statystyka,* PWN, Warszawa 2023.
23. Sobczak M., *Statystyka opisowa*, C.H.Beck, Warszawa 2010.
24. Starzyńska W., *Statystyka praktyczna*, PWN, Warszawa 2023.
25. Szymański W., *Podstawy statystyki dla psychologii*, Difin, Warszawa 2020.