Państwowa Uczelnia Zawodowa we Włocławku

Marek Ręklewski

# STATYSTYKA OPISOWA

## Teoria i przykłady

Włocławek 2020

REDAKCJA WYDAWNICTWA
PAŃSTWOWEJ UCZELNI ZAWODOWEJ
WE WŁOCŁAWKU
Statystyka opisowa. Teoria i przykłady

RECENZENT
Dr hab. Maria Chromińska



ISBN 978-83-60607-93-0

Nakład: 100 egz.

Skład, druk i oprawa:
Agencja Reklamowa TOP, ul. Toruńska 148, 87-800 Włocławek
tel. 54 423 20 40, e-mail agencja.top@agencjatop.pl
www.agencjatop.pl

*Celem obliczeń nie są same liczby, lecz ich zrozumienie.*

Richard Wesley Hamming
(1915-1998)

# Spis treści

## Wstęp

Książka ma charakter podręcznika akademickiego i jest przeznaczona dla studentów studiów licencjackich i inżynierskich Państwowej Uczelni Zawodowej we Włocławku. Podręcznik może być przydatny zarówno praktykom, jak i studentom innych uczelni zgłębiających wiedzę ze statystyki opisowej.

Celem proponowanej publikacji pt. „*Statystyka opisowa. Teoria i przykłady*" jest przekazanie studentom podstawowej wiedzy i umiejętności z zakresu stosowanych metod statystycznych. Zawarte treści programowe w książce są niezbędne i pomocne w trakcie studiów podczas przeprowadzanych zaliczeń, projektów czy egzaminów. Popularyzatorski charakter podręcznika ma zachęcić studentów do samodzielnego poznawania bardziej złożonych i wyrafinowanych metod analizy statystycznej.

Na rynku wydawniczym znajduje się wiele interesujących pozycji, które często napisane są językiem wysoce specjalistycznym, niekiedy mało zrozumiałym dla przeciętnego studenta. W niniejszej pracy starano się w sposób przystępny i syntetyczny ująć podstawowe wiadomości z zakresu statystyki opisowej ograniczając się do niezbędnego minimum.

W książce znalazło się 5 rozdziałów z zakresu: podstawowych zagadnień statystycznych, analizy struktury zjawisk masowych (parametry klasyczne i pozycyjne), analizy współzależności (dla cech ilościowych i jakościowych), analizy regresji liniowej dwóch zmiennych oraz statystycznych metod analizy zjawisk w czasie.

Konstrukcja i układ treści w książce został podporządkowany wykładom i ćwiczeniom prowadzonym dla studentów studiów I stopnia. Każdy z wyróżnionych rozdziałów składa się z części teoretycznej i praktycznej. Część teoretyczna składa się z definicji, wzorów, objaśnień do omówionych metod statystycznych. W części praktycznej przedstawiono liczne przykłady z rozwiązaniami. Dane statystyczne prezentowano na licznych wykresach a obliczenia matematyczne zestawiono w dodatkowych tabelach. Otrzymane parametry statystyczne zostały opatrzone właściwą interpretacją opisową.

Podstawowym źródłem danych statystycznych wykorzystanych w podręczniku były informacje pochodzące ze strony internetowej Banku Danych Lokalnych (BDL) Głównego Urzędu Statystycznego, jak również z publikacji książkowych tj.: Badanie Aktywności Ekonomicznej Ludności (BAEL). W zaprezentowanych przykładach wykorzystano w miarę możliwości jak najnowsze dostępne dane statystyczne GUS, dotyczące sytuacji na rynku pracy w Polsce.

# 1. Podstawowe zagadnienia

## 1.1. Przedmiot statystyki

Początki statystyki można odnaleźć już w czasach biblijnych. W starożytnej Grecji, Rzymie były przeprowadzane spisy ludności dające bogate źródło informacji jego władcom o stanie danego społeczeństwa. W Cesarstwie Rzymskim spisy ludności były prowadzone systematycznie co 5 lat (tzn. cenzusy, łac. „*census"*: oszacowanie majątku).

W czasach średniowiecznych zapotrzebowanie na dane statystyczne jeszcze wzrosło. Możnowładzcy, jak i władze kościelne potrzebowały różnorodnych danych dotyczących wielu dziedzin życia społeczno-gospodarczego m.in. ludności, handlu, rolnictwa. Już wówczas dane statystyczne odgrywały istotną rolę we właściwym zarządzaniu i planowaniu posiadanych bogactw. Dane najczęściej zestawiano w postaci tabelarycznej. Taki „tradycyjny" sposób gromadzenia i interpretacji danych liczbowych w miarę upływu czasu był dalece niewystarczający.

**Statystyka** pochodzi od łacińskiego słowa: „*status"*, które oznacza: państwo, stan rzeczy. W literaturze przedmiotu możemy spotkać się z różnymi definicjami „Statystyki". Najczęściej wymienianymi są:

1. Statystyka jest nauką o metodach badania zjawisk masowych.
2. Statystyka jest nauką o metodach zbierania, analizy i interpretacji danych liczbowych.
3. Statystyka jest nauką o ilościowych metodach badania zjawisk (procesów) masowych.
4. Statystyka jest nauką traktującą o specyficznych metodach ilościowych dostosowanych do badania prawidłowości zjawisk masowych.

Z definicji statystyki można zauważyć, że jest to nauka, która posiada swój własny przedmiot badania. W centrum zainteresowania statystyki są zjawiska masowe. **Zjawiska masowe** to takie, które często się powtarzają. Charakteryzują się one dużym natężeniem występowania tworząc tym samym liczne zbiorowości. Zaliczyć można do nich procesy demograficzne np. urodzenia, zgony. W dużej ich masie poddane właściwej procedurze badawczej w zjawiskach masowych można zaobserwować pewne zależności, prawa, reguły czyli tzw. **prawidłowości**. Prawidłowości nie da się wychwycić na podstawie obserwacji pojedynczego przypadku. Aby było to możliwe, należy dysponować dostatecznie liczną zbiorowością. Wówczas możemy zauważyć oddziaływanie dwóch podstawowych przyczyn (czynników): **głównych** i **ubocznych** (przypadkowych o charakterze losowym). W zjawiskach masowych oddziaływanie przyczyn głównych staje się bardziej wyraziste, a tym samym prawidłowości są lepiej zau-

ważalne. Z kolei, różnokierunkowy charakter czynnika przypadkowego powoduje się jego znoszenie. Oznacza to, że minimalizuje się wpływ czynnika ubocznego w danej zbiorowości. W badaniach ilościowych, do których zalicza się metody statystyczne, istniejące prawidłowości możemy poszukiwać badając:

- strukturę (np. płci, wieku, wykształcenia, miejsca zamieszkania),
- szeregi czasowe (np. badania nad identyfikacją determinant kształtujących dany proces),
- dane przestrzenne (np. badania nad zróżnicowaniem danego zjawiska według jednostek administracyjnych).

Rozwój statystyki zawdzięczamy arytmetykom politycznym a zwłaszcza pracy matematyków od momentu pojawienia się tzw. teorii rachunku prawdopodobieństwa w XVII wieku. Za twórców tego nurtu uważa się dwóm francuskim matematyków B. Pascala i P. Fermata. Teoria rachunku prawdopodobieństwa zajmuje się tzw. zdarzeniami losowymi. Definicję prawdopodobieństwa w postaci matematycznej zaprezentował P. S. Laplace'a na początku XVIII wieku. Formuła w zapisie jest następująca:

$$P(A) = \frac{k}{n} \tag{1.1.1}$$

gdzie: $P(A)$ oznacza prawdopodobieństwo wystąpienia zdarzenia $A$, $k$ oznacza liczbę zdarzeń elementarnych sprzyjających zajściu zdarzenia $A$, $n$ oznacza wszystkie zdarzenia elementarne.

Statystykę dzielimy na: **opisową** i **matematyczną**. Teoria rachunku prawdopodobieństwa, która stała się podstawą do powstania **statystyki matematycznej (wnioskowania statystycznego)** umożliwiła rozwój metod statystycznych w zakresie m.in.: testowania parametrycznych i nieparametrycznych hipotez badawczych, estymacji przedziałowej (konstrukcji nieznanego przedziału ufności) czy też schematów losowania próby w badaniach reprezentacyjnych.

Statystyka zajmująca się gromadzeniem, opracowaniem i prezentacją danych liczbowych łącznie z ich opisem, przy wykorzystaniu dostępnych narzędzi statystycznych nosi nazwę **statystyki opisowej**. Na podstawie gotowych wzorów obliczone parametry statystyczne dla badanej cechy nazywane są wówczas statystykami opisowymi. Parametry opisowe charakteryzują badaną zbiorowość. Do statystyk opisowych zaliczamy m.in.: średnią arytmetyczną, dominantę, medianę, wariancję, odchylenie standardowe.

**Zbiorowością statystyczną** nazywamy zbiór elementów (jednostek) objętych badaniem statystycznym (tab. 1.1). **Jednostka statystyczna** stanowi pojedynczy element zbiorowości statystycznej (np. student, pracownik). Zbiorowość statystyczną można podzielić na: general-

ną i próbną. **Zbiorowość generalna** (populacja) składa się ze zbioru elementów objętych badaniem, co do której formułuje się wnioski na podstawie uogólnionych wyników z próby. **Zbiorowość próbna** (próba) losowana jest z populacji generalnej z zastosowaniem odpowiednich metod reprezentacyjnych. Próba dobierana może być także w sposób celowy (nielosowy). Zakłada się, że liczba jednostek w małej próbie wynosi $n \leq 30$, z kolei dużej $n > 30$.

**Tabela 1.1.Przykłady zbiorowości i jednostki statystycznej**

| Zbiorowość statystyczna | Jednostka statystyczna |
|---|---|
| Studenci na kierunku Zarządzanie w PUZ we Włocławku | student |
| Wykładowcy PUZ we Włocławku | wykładowca |
| Pracownicy w firmie X | pracownik |
| Podmioty gospodarcze zatrudniające powyżej 9 pracujących | podmiot gospodarczy |
| Gospodarstwa rolne w miejscowości Y | gospodarstwa rolne |

Źródło: opracowanie własne.

Jednostki statystyczne można opisać zestawem różnorodnych cech. **Cechami statystycznymi** nazywamy właściwości jednostek statystycznych. Wyróżnia się cechy: **stałe** i **zmienne**. **Cechy stałe** umożliwiają wyłącznie zakwalifikowanie danej jednostki do badania statystycznego. Są to cechy wspólne dla wszystkich jednostek badanej zbiorowości (np. tworzymy zbiorowość tylko ze studentów na kierunku Zarządzanie). Spełnienie tego warunku jest kluczowe z punktu widzenia zachowania jednorodności populacji statystycznej. Z kolei, **cechy zmienne** różnicują badane jednostki w zbiorowości pomiędzy sobą. Wówczas te cechy stanowią przedmiot analizy badawczej, które poddaje się dalszej obserwacji statystycznej z wykorzystaniem różnorodnych narzędzi statystycznych. Ważną kwestią w badaniach statystycznych jest określenie zbiorowości pod względem rzeczowym, czasowym i przestrzennym.
Cechy podlegają dalszemu podziałowi na:

- cechy mierzalne (ilościowe):
  – skokowe,
  – ciągłe
  – quasi stałe.
- cechy niemierzalne (jakościowe).

**Cechy mierzalne** (ilościowe) są to cechy (właściwości), które można zmierzyć wyrażając je w ściśle określonych jednostkach np. miary (wzrost w cm), masy (waga w kg), powierzchni ($km^2$, ha), monetarnych (wynagrodzenie w złotych) czy czasu (wiek w latach). Wartości są liczbami mianowanymi.

**Cechy mierzalne skokowe** (dyskretne) przyjmują wartości liczbowe ze skończonych i przeliczalnych zbiorów wartości całkowitych. Cecha zmienia się skokowo bez wartości pośrednich (np. liczba studentów w grupie, liczba osób w rodzinie).

**Cechy mierzalne ciągłe** przyjmują każdą wartość z danego przedziału liczbowego. Jest to zbiór wartości nieskończony i nieprzeliczalny. Cecha przyjmuje wartości pośrednie (np. wzrost, waga, wiek, wynagrodzenia, temperatura).

**Cechy quasi-ilościowe** wyrażają natężenie badanej cechy w sposób opisowy. Taka cecha nazywana jest cechą **quasi-porządkową.** Ze względu na status materialny studentów możemy przyporządkować do odpowiednich wariantów tj.: niski, średni, wysoki.

**Cechy jakościowe (niemierzalne)** to takie, których nie jesteśmy w stanie zmierzyć liczbowo, a jedynie możemy stwierdzić natężenie (liczebność) określonego wariantu, do którego przyporządkowane są jednostki zbiorowości. Przykładem cechy jakościowej jest płeć składająca się dwóch wariantów odpowiedzi tj.: mężczyzna, kobieta, czy miejsce zamieszkania: miasto, wieś.

## 1.2. Skale pomiarowe

Badając zachodzące relacje pomiędzy zjawiskami dokonuje się ich pomiaru. Wyróżnia się 4 podstawowe skale pomiarowe (tab. 1.2):

- nominalną,
- porządkową (rangową),
- przedziałową (interwałową),
- oraz stosunkową (ilorazową).

Skale **nominalna**, **porządkowa** (tzw. skale słabe) dotyczą **cech jakościowych**, z kolei skale **przedziałowa** i **ilorazowa** (tzw. skale mocne) odnoszą się do **cech ilościowych**.

**Tabela 1.2. Skale pomiarowe i ich charakterystyka**

| Rodzaj skali | Opis | Przykłady | Wybrane operacje arytmetyczne |
|---|---|---|---|
| **1. Nominalna** *(nominal scale)*<br><br>Relacja:<br>Równe lub różne | Umożliwia jedynie przyporządkowanie jednostek zbiorowości do właściwych kategorii ze względu na charakteryzujące je cechy. Szczególnymi przypadkami tej skali są:<br>– skale dwudzielne (dychotomiczne) gdy występują dwa warianty cechy (np. płeć: mężczyzna, kobieta, miejsce zamieszkania: miasto, wieś);<br>– wielodzielne politomiczne (np. trychotomiczne), gdy występują trzy warianty cechy (np. odpowiedź na pytanie: tak, nie lub nie wiem). | – płeć, wykształcenie, zawód, stan cywilny, miejsce zamieszkania itp. | – zliczanie jednostek w danych wariantach (określenie częstości występowania),<br>– wskaźniki struktury,<br>– dominanta,<br>– miary korelacji:<br>(np. *C*-Pearsona, *V*-Cramera),<br>– testy nieparametryczne (np. test niezależności, chi-kwadrat). |
| **2. Porządkowa** (*ordinal scale*)<br><br>Relacja:<br>Większy lub mniejszy | Umożliwia przyporządkowanie jednostek zbiorowości w ramach wyróżnionych kategorii ze względu na natężenie badanej cechy. Uporządkowanie może być wykonane w porządku rosnącym lub malejącym. | – wykształcenie, stopnie wojskowe, oceny, skala IQ, itp. | – mediana,<br>– kwartyle, decyle,<br>– rozstęp ćwiartkowy,<br>– rozstęp,<br>– miary korelacji (np. tau-Kendalla, rang Spearmana). |
| **3. Przedziałowa** (*interwal scale*)<br><br>Relacja:<br>Większe o tyle | W skali tej brak jest zera absolutnego (występuje tzw. zero względne). Możliwe jest porównywanie analizowanych jednostek statystycznych przez określenie pomiędzy nimi różnicy. | – temperatura Celsjusza, Farenheita, rok urodzenia itp. | – średnia arytmetyczna,<br>– odchylenie standardowe, współczynnik zmienności, miary asymetrii, koncentracji,<br>– miary korelacji (np. współczynnik korelacji liniowej Pearsona). |
| **4. Stosunkowa** (*relative scale*)<br><br>Relacja:<br>Tyle razy większe | W skali tej występuje naturalny poziom zerowy (zero bezwzględne). Oprócz podania różnicy (odległości) pomiędzy pomiarami można podać krotności wynikające ze stosunku dwóch pomiarów. | – wiek, waga, wzrost, wynagrodzenia, ceny wielkość sprzedaży itp. | – wszystkie operacje i dodatkowo dzielenie. |

Źródło: opracowane na podstawie: M., Walesiak, *Metody analizy danych marketingowych*, PWN, Warszawa, 1996. W., Ignatczyk, M., Chromińska, Statystyka. *Teoria i zastosowanie,* WSB, Poznań 2004. M., Sobczyk, *Statystyka*, PWN, Warszawa 2016.

## 1.3. Badania statystyczne

**Badaniem statystycznym** nazywamy szereg prac związanych m.in. ze zbieraniem, przetwarzaniem i analizowaniem danych statystycznych opisujących badaną zbiorowości pod względem określonych cech zgodnie z wyznaczonym celem badania. Proces badania statystycznego składa się z 4 podstawowych etapów:

1. **Przygotowanie badania** (zawiera szereg prac metodologicznych, do których w szczególności możemy zaliczyć: określenie zakresu podmiotowego/przedmiotowego, opis podstawowych pojęć i definicji, tworzenie algorytmu doboru jednostek do operatu i schematu losowania kartoteki, projektowanie formularza sta-

tystycznego wraz z objaśnieniami, opracowanie i wysyłka powiadomień – monitów – o obowiązku sprawozdawczym, tworzenie harmonogramów zbierania i przetwarzania, podział i przypisanie części kartoteki statystykom itp.).

2. **Zbieranie danych** (obserwacja statystyczna za pośrednictwem określonej metody np. forma papierowa, forma elektroniczna).
3. **Weryfikacja danych** (kontrola kompletności badania i kontrola merytoryczna danych statystycznych, poprzez analizę i poprawę błędów o charakterze przypadkowym lub systematycznym).
4. **Opracowanie i publikacja wyników** (udostępnianie i rozpowszechnianie informacji statystycznych z danego badania w formie drukowanej lub elektronicznej).

Ze względu na zakres obserwacji (liczebność) badania statystyczne dzielimy na **pełne** i **częściowe**. W **badaniu pełnym** wszystkie jednostki zbiorowości podlegają obserwacji statystycznej. Badanie pełne nosi nazwę badania całkowitego, wyczerpującego lub generalnego. Może mieć charakter ciągły, okresowy lub doraźny. Do badań pełnych zalicza się:

- **spisy statystyczne** – dostarczają niezbędnych informacji o stanie i strukturze badanej zbiorowości na dany określony moment (wyróżnia się: Powszechne Spisy Rolne, Narodowe Spisy Powszechne Ludności i Mieszkań, które odbywają się na ogół co 10 lat. Ostatni spis w Polsce miał miejsce w 2011 r., następny planowany jest na 2021 r.),
- **rejestrację bieżącą** – przeprowadzają do tego powołane organy administracji publicznej. Rejestracja bieżąca polega na bieżącej i systematycznej ewidencji (rejestracji) zachodzących zdarzeń (np. urodzeń, zgonów, małżeństw). W Polsce wyróżnia się m.in. rejestry: Krajowy Rejestr Urzędowy Podmiotów Gospodarki Narodowej (REGON), Krajowy Rejestr Urzędowy Podziału Terytorialnego Kraju (TERYT), Krajowy Rejestr Sądowy (KRS),
- **sprawozdawczość statystyczną** – realizują podmioty gospodarcze zobowiązane do przekazywania informacji i danych statystycznych dotyczących prowadzonej działalności właściwym organom administracji rządowej. Udział w badaniu ma charakter obowiązkowy. Dane uzyskane w badaniu objęte są tajemnicą statystyczną.

W **badaniu częściowym** obserwacji statystycznej podlega tylko pewna część zbiorowości generalnej, która nazywana jest próbą. Dobór jednostek do próby ma charakter losowy lub celowy. Przeprowadzenie badań częściowych powoduje wiele korzyści tj.: obniżenie kosztów badania, szybsze zebranie danych, mniejszą pracochłonność nad przygotowaniem danych, szybsze prezentowanie wyników. Badania częściowe dzielą się na:

a) **badanie ankietowe** polega na zbieraniu informacji o zbiorowości za pomocą przygotowanej ankiety (tzw. kwestionariusza ankietowego),
b) **badanie monograficzne** polega na wszechstronnej i wyczerpującej obserwacji typowej jednostki statystycznej (np. jednej wsi, miasta). Takie badanie umożliwia pozyskanie informacji o badanej zbiorowości w zakresie ilościowym i jakościowym. Jednostkę badania zwykle wybiera się celowo.
c) **badanie reprezentacyjne** opiera się na próbie losowej uzyskanej z populacji generalnej (na podstawie wcześniej przygotowanego operatu badania, w którym znajdują zakwalifikowane jednostki zgodnie z zakresem podmiotowym badania) według odpowiednio przyjętego schematu losowania. Otrzymane wyniki z próby losowej uogólnia się z pewnym prawdopodobieństwem na całą populację. Próba losowa powinna być dostatecznie liczna, aby formułowane wnioski miały charakter reprezentatywny.

## 1.4. Źródła i bazy danych

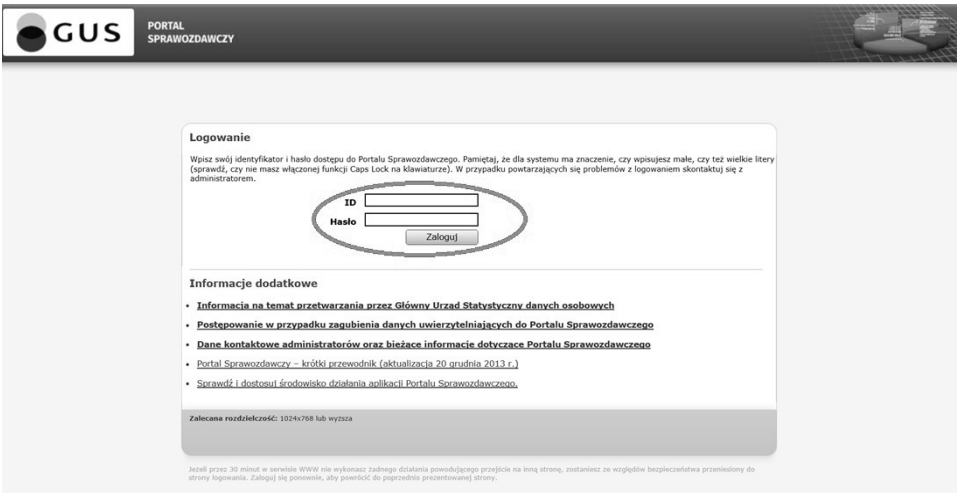
Centralnym urzędem administracji rządowej zajmującym się zbieraniem i udostępnianiem informacji statystycznych jest Główny Urząd Statystyczny (GUS) oraz podległe Urzędy Statystyczne. Do przekazywania danych jednostki gospodarcze obliguje ustawa o statystyce publicznej art. 30 ust. 1 pkt 3 ustawy z dnia 29 czerwca 1995 r. oraz ogłaszany corocznie Program Badań Statystycznych (PBSSP). Oficjalna strona GUS znajduje się pod adresem internetowym: https://stat.gov.pl (rys.1.4.1).


**Rysunek. 1.4.1. Strona internetowa GUS**
Źródło: https://stat.gov.pl/

Od 1 stycznia 2009 roku GUS wprowadził elektroniczną formę przekazywania danych statystycznych poprzez Portal Sprawozdawczy (PS).

**Portal Sprawozdawczy** jest zintegrowanym narzędziem elektronicznym przeznaczonym do obsługi sprawozdawczości GUS. System przeznaczony jest dla sprawozdawców czyli jednostek z zarejestrowaną działalnością gospodarczą, przekazujących dane statystyczne w ramach realizacji obowiązków sprawozdawczych. Obowiązki sprawozdawcze mogą być realizowane za pomocą:

- **aplikacji on-line** – po uprzedniej aktywacji konta i zalogowaniu się na Portalu Sprawozdawczym (rys. 1.4.2),
- **aplikacji off-line** – opracowany w java lub jako „aktywnym pdf.", który wystarczy pobrać ze strony internetowej GUS. Program instalowany jest na komputerze sprawozdawcy, bez konieczności bezpośredniego, stałego połączenia z internetem,
- w szczególnych sytuacjach, dla podmiotów o liczbie pracujących 5 i mniej, dopuszczalna jest forma papierowa.




**Rysunek 1.4.2. Logowanie do Portalu Sprawozdawczego**
Źródło: https://stat.gov.pl/

Możliwości Portalu Sprawozdawczego w zakresie sprawozdawczości są następujące (http://form.stat.gov.pl/formularze/ Portal_Sprawozdawczy_przewodnik_1312.pdf):

- pozwala sprawozdawcom na składanie sprawozdań on-line,
- umożliwia wydruk formularza z wprowadzonymi danymi,
- zawiera informacje o jednostce sprawozdawczej,

- wyświetla jednostce sprawozdawczej listę aktualnych obowiązków sprawozdawczych, a statystykom listę przypisanych badań,
- wyświetla informacje o statusie sprawozdania (nierozpoczęte, wypełniane, zatwierdzone),
- pozwala na korespondencję pomiędzy jednostkami sprawozdawczymi i statystykami,
- pozwala na podgląd danych wprowadzonych przez jednostki sprawozdawcze do formularzy,
- umożliwia kontrolę poprawności danych podczas ich wprowadzania – wyświetla błędy i wskazuje miejsca ich występowania.

W przypadku rynku pracy źródłami danych jednostkowych są następujące badania statystyczne (GUS, 2008; PBSSP na rok 2018):

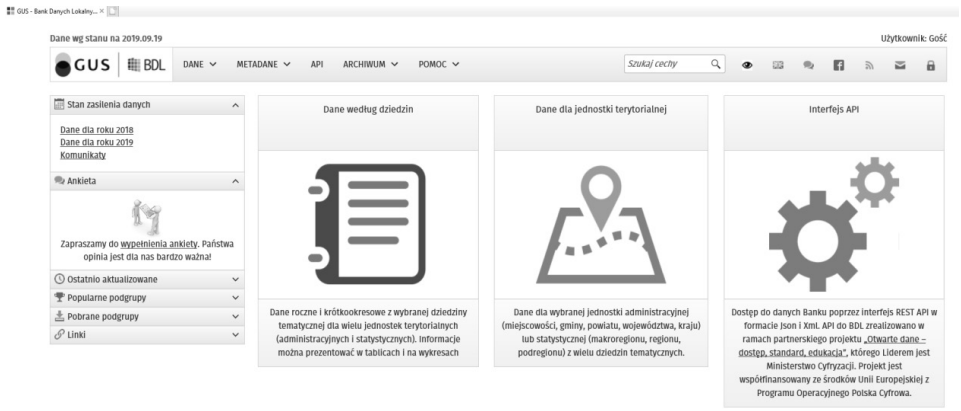
- Z-02 – Sprawozdanie kosztów pracy (cykliczne, co 4 lata reprezentacyjne),
- Z-03 – Sprawozdanie o zatrudnieniu i wynagrodzeniach (kwartalne, pełne),
- Z-05 – Badanie popytu na pracę (kwartalne, reprezentacyjne),
- Z-06 – Sprawozdanie o pracujących, wynagrodzeniach i czasie pracy (roczne, pełne),
- Z-10 – Sprawozdanie o warunkach pracy (roczne, pełne),
- Z-12 – Sprawozdanie o strukturze wynagrodzeń według zawodów (cykliczne, co 2 lata, reprezentacyjne),
- Z-14 – Sprawozdanie o zatrudnieniu i wynagrodzeniach w administracji publicznej (cykliczne, co 2 lata, pełne),
- MRPiPS-01 – Sprawozdanie o rynku pracy (miesięczne i pełne),
- MRPiPS-02 – Sprawozdanie o przychodach i wydatkach Funduszu Pracy (miesięczne i pełne),
- MRPiPS-04 – Sprawozdanie o wydawanych zezwoleniach na pracę cudzoziemcom w RP (półroczne, pełne),
- MRPiPS-07 – Sprawozdanie o osobach niepełnosprawnych bezrobotnych i poszukujących pracy niepozostających w zatrudnieniu (półroczne, pełne),
- Z-KW – Statystyczna Karta Wypadku,
- Z-KS – Karta Statystyczna Strajku,
- DG-1 – Meldunek o działalności gospodarczej (miesięczne, pełne),
- SP-3 – Sprawozdanie o działalności gospodarczej przedsiębiorstw (reprezentacyjne, roczne),

- Badanie Aktywności Ekonomicznej Ludności (BAEL),
- Narodowy Spis Powszechny Ludności i Mieszkań.

Na stronie GUS znajduje się zakładka „Bazy danych" (rys. 1.4.3). Po jej rozwinięciu wyświetli się wykaz aktualnych baz danych statystycznych.


**Rysunek 1.4.3. Bazy danych GUS**
Źródło: https://stat.gov.pl/

Wśród nich znajduje się **Bank Danych Lokalnych** (BDL), który jest największą w Polsce bazą danych o gospodarce, społeczeństwie i środowisku (rys. 1.4.4). BDL oferuje ponad 40 tys. cech statystycznych pogrupowanych tematycznie (GUS, https://stat.gov.pl/).

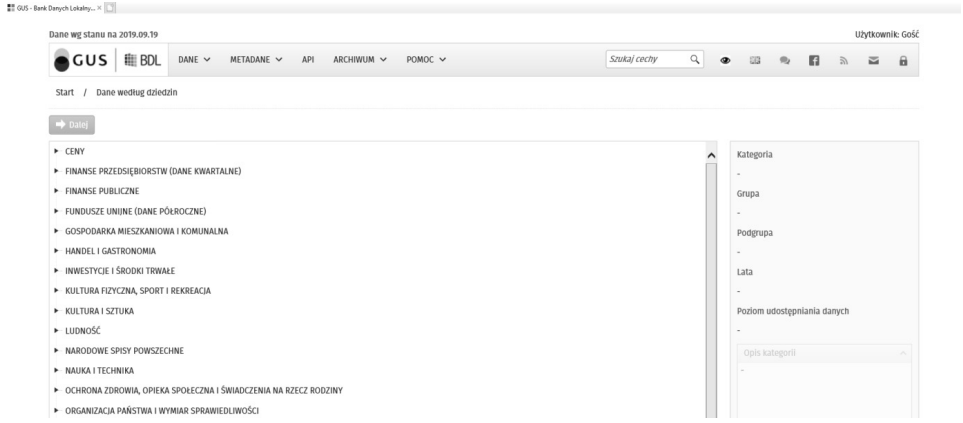



**Rysunek 1.4.4. Strona internetowa Banku Danych Lokalnych**
Źródło: Źródło: https://bdl.stat.gov.pl/

Po otworzeniu okna BDL mamy do wyboru trzy opcje:

- dane według dziedzin,
- dane dla jednostki terytorialnej,
- interfejs API.

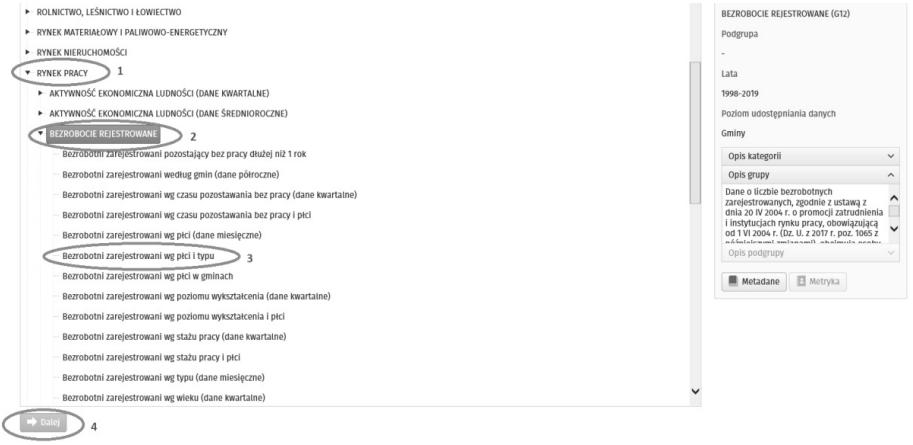
Klikamy kursorem myszki na „Dane według dziedzin" wówczas w oknie pojawią uporządkowane tematycznie do wyboru dziedziny (rys. 1.4.5):




**Rysunek 1.4.5. Dane według dziedzin**
Źródło: https://bdl.stat.gov.pl/

Z listy dziedzin wybieramy tą, która nas interesuje. A więc klikamy np. „RYNEK PRACY" (rys. 1.4.6):

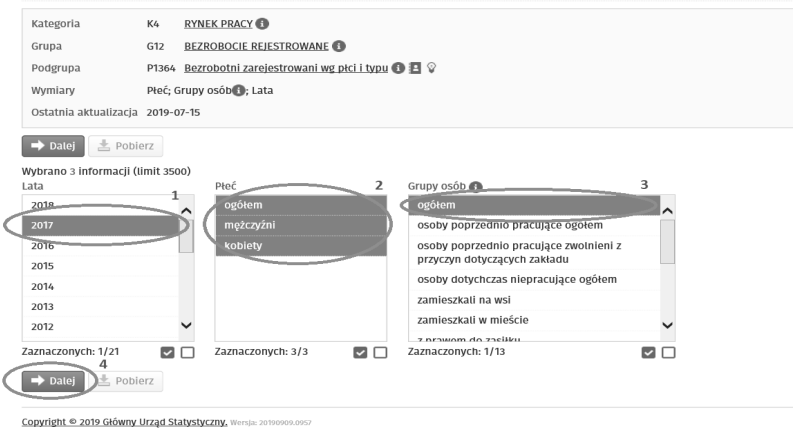
**RYNEK PRACY → BEZROBOCIE REJESTROWANE → Bezrobotni zarejestrowani wg płci i typu → Dalej**




**Rysunek 1.4.6. Wybór dziedziny danych**
Źródło: https://bdl.stat.gov.pl/

Następnie definiujemy pozostałe kryteria: lata, płeć i grupy osób bezrobotnych (rys. 1.4.7):

**2017 → ogółem, mężczyźni, kobiety → ogółem → Dalej**




**Rysunek 1.4.7. Wybór kryteriów danych**
Źródło: https://bdl.stat.gov.pl/

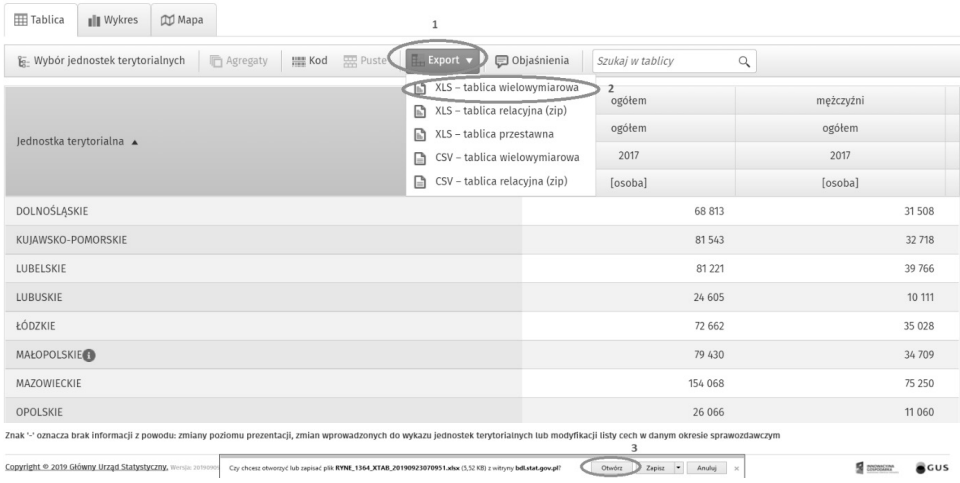
W kolejnym oknie dokonujemy wyboru jednostek terytorialnych. Do wyboru mamy województwa, powiaty i gminy. Wybieramy województwa (rys. 1.4.8):

**Zaznacz → Zaznacz województwa → „→" → Dalej**


**Rysunek 1.4.8. Wybór jednostek terytorialnych**
Źródło: https://bdl.stat.gov.pl/

W ostatnim oknie pojawią się zdefiniowane dane statystyczne, które eksportujemy do pliku w formacie „xlsx" (lub „csv"). Plik z danymi zapisujemy we wskazanym miejscu na dysku własnego komputera (rys. 1.4.9):

**Export → XLS – tablica wielowymiarowa → otwórz → zapisz jako → podać miejsce docelowe → zapisz**




**Rysunek 1.4.9. Eksport danych**
Źródło: https://bdl.stat.gov.pl/

Zapisany plik otwieramy w programie MS Excel. Znajdują się w nim dwie zakładki. W pierwszej zawarte są podstawowe informacje o wygenerowanych danych (tzw. metadane) a w drugiej gotowa tabela z danymi statystycznymi (rys. 1.4.10):

| | A | B | C | D | E |
|---|---|---|---|---|---|
| 1 | Kod | Nazwa | ogółem | mężczyźni | kobiety |
| 2 | | | ogółem | ogółem | ogółem |
| 3 | | | 2017 | 2017 | 2017 |
| 4 | | | [osoba] | [osoba] | [osoba] |
| 5 | 0200000 | DOLNOŚLĄSKIE | 68 813 | 31 508 | 37 305 |
| 6 | 0400000 | KUJAWSKO-POMORSKIE | 81 543 | 32 718 | 48 825 |
| 7 | 0600000 | LUBELSKIE | 81 221 | 39 766 | 41 455 |
| 8 | 0800000 | LUBUSKIE | 24 605 | 10 111 | 14 494 |
| 9 | 1000000 | ŁÓDZKIE | 72 662 | 35 028 | 37 634 |
| 10 | 1200000 | MAŁOPOLSKIE | 79 430 | 34 709 | 44 721 |
| 11 | 1400000 | MAZOWIECKIE | 154 068 | 75 250 | 78 818 |
| 12 | 1600000 | OPOLSKIE | 26 066 | 11 060 | 15 006 |
| 13 | 1800000 | PODKARPACKIE | 90 972 | 42 353 | 48 619 |
| 14 | 2000000 | PODLASKIE | 39 997 | 21 136 | 18 861 |
| 15 | 2200000 | POMORSKIE | 49 653 | 18 744 | 30 909 |
| 16 | 2400000 | ŚLĄSKIE | 94 687 | 40 189 | 54 498 |
| 17 | 2600000 | ŚWIĘTOKRZYSKIE | 46 570 | 22 345 | 24 225 |
| 18 | 2800000 | WARMIŃSKO-MAZURSKIE | 60 003 | 26 186 | 33 817 |
| 19 | 3000000 | WIELKOPOLSKIE | 58 857 | 23 091 | 35 766 |
| 20 | 3200000 | ZACHODNIOPOMORSKIE | 52 599 | 22 022 | 30 577 |

**Rysunek 1.4.10. Dane w pliku xlsx.**
Źródło: opracowanie własne

## 1.5. Prezentacja danych statystycznych

Szeregi statystyczne stanowią formę prezentacji danych statystycznych. Ze względu na formę wyróżnia się następujące szeregi:

- proste, czyli wyliczające (szczegółowe),
- rozdzielcze (jednopunktowe i wielopunktowe).

**Szereg prosty** to uporządkowane (rosnąco lub malejąco) wartości liczbowe badanej cechy $x_i$. Przykładem takiego szeregu mogą stanowić np. wydatki związane z dojazdami do szkoły 10 studentów w ciągu miesiąca (w zł):

65, 73, 79, 80, 82, 95, 100, 109, 112, 120

**Szereg rozdzielczy** to zarówno uporządkowana, jak i pogrupowana badana cecha $x_i$ według określonych wariantów występowania (tab. 1.5.1 – 1.5.2).

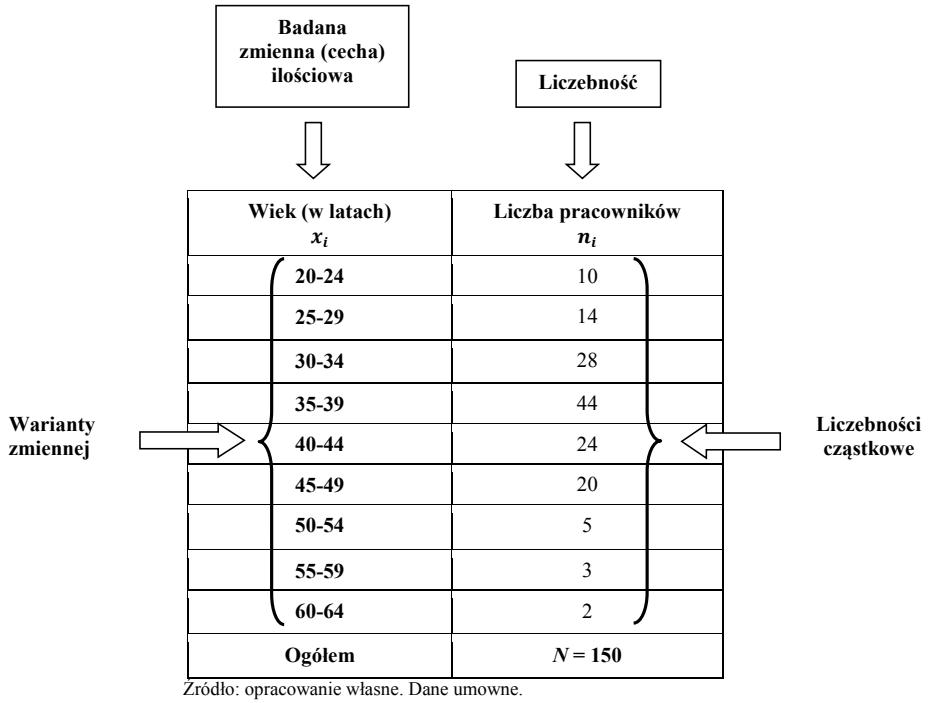
**Tabela 1.5.1. Oceny studentów z zaliczenia ze statystyki na kierunku Zarządzanie – przykład szeregu rozdzielczego jednopunktowego**

| Badana zmienna (cecha) jakościowa | Liczebność |
|---|---|
| **Oceny z zaliczenia ze statystyki $x_i$** | **Liczba studentów $n_i$** |
| **2,0** | 5 |
| **3,0** | 50 |
| **3,5** | 90 |
| **4,0** | 30 |
| **4,5** | 15 |
| **5,0** | 10 |
| **Ogółem** | ***N* = 200** |

**Warianty zmiennej** — **Liczebności cząstkowe**

Źródło: opracowanie własne. Dane umowne.

**Tabela 1.5.2. Pracujący według wieku w firmie X**
**– przykład szeregu rozdzielczego z przedziałami klasowymi**




| Wiek (w latach) $x_i$ | Liczba pracowników $n_i$ |
|---|---|
| **20-24** | 10 |
| **25-29** | 14 |
| **30-34** | 28 |
| **35-39** | 44 |
| **40-44** | 24 |
| **45-49** | 20 |
| **50-54** | 5 |
| **55-59** | 3 |
| **60-64** | 2 |
| **Ogółem** | ***N* = 150** |

Źródło: opracowanie własne. Dane umowne.

**Liczebności cząstkowe** ($\boldsymbol{n_i}$) inaczej absolutne lub bezwzględne przedstawiają liczbę obserwacji (częstości) występowania danego wariantu cechy $x_i$. Liczebności $n_i$ sumuje się do siebie otrzymując liczebność całej badanej zbiorowości *N*:

$$n_1 + n_2 + \ \ldots n_k = \sum_{i=1}^{k} n_i = N$$

gdzie:
$k$ – liczba wariantów cechy $x_i$,
$N$ – liczebność całej zbiorowości.

Oprócz liczebności absolutnych wyróżniamy **liczebności względne (stosunkowe)** tzw. **wskaźniki struktury ($\boldsymbol{w_s}$)** – tab. 1.5.3. Przedstawiają udział części zbiorowości statystycznej do jej całości i najczęściej wyrażone są w procentach:

$$w_s = \frac{n_i}{N} \cdot 100 \tag{1.5.1}$$

gdzie:

$n_i$ – liczebność danej klasy cechy $x_i$.
$N$ – liczebność całej zbiorowości.

Suma wskaźników struktury ($w_s$) powinna wynosić 100%:

$$w_1 + w_2 + \ldots w_k = \sum_{i=1}^{k} w_i = 100\%$$

Wskaźniki struktury są liczbami względnymi (niemianowanymi).

**Tabela 1.5.3. Struktura pracujących według wieku w firmie X**

| Badana zmienna (cecha) jakościowa | Liczebności bezwzględne | Liczebności względne |
|---|---|---|

| Wiek (w latach ) $x_i$ | Liczba pracowników $n_i$ | Wskaźniki Struktury (%) |
|---|---|---|
| **20-24** | 10 | $\frac{n_1}{N} \cdot 100 = \frac{10}{150} \cdot 100 = 6{,}7$ |
| **25-29** | 14 | $\frac{n_2}{N} \cdot 100 = \frac{14}{150} \cdot 100 = 9{,}3$ |
| **30-34** | 28 | $\frac{n_3}{N} \cdot 100 = \frac{28}{150} \cdot 100 = 18{,}7$ |
| **35-39** | 44 | $\frac{n_4}{N} \cdot 100 = \frac{44}{150} \cdot 100 = 29{,}3$ |
| **40-44** | 24 | $\frac{n_5}{N} \cdot 100 = \frac{24}{150} \cdot 100 = 16{,}0$ |
| **45-49** | 20 | $\frac{n_6}{N} \cdot 100 = \frac{20}{150} \cdot 100 = 13{,}3$ |
| **50-54** | 5 | $\frac{n_7}{N} \cdot 100 = \frac{5}{150} \cdot 100 = 3{,}4$ |
| **55-59** | 3 | $\frac{n_8}{N} \cdot 100 = \frac{3}{150} \cdot 100 = 2{,}0$ |
| **60-64** | 2 | $\frac{n_9}{N} \cdot 100 = \frac{2}{150} \cdot 100 = 1{,}3$ |
| **Ogółem** | ***N* = 150** | **100,0** |

Źródło: opracowanie własne. Dane umowne.

W szeregach rozdzielczych możemy wyróżnić dwie sytuacje, kiedy:

a) górna granica przedziału klasowego jest jednocześnie dolną granicą następnego przedziału tzn.: 20-25, 25-30, 30-35, 35-40, 40-45 lat itd. (dla cechy ilościowej ciągłej),
b) górna granica przedziału klasowego **nie pokrywa** się z dolną granicą następnego przedziału tzn.: 20-24, 25-29, 30-34, 35-39, 40-44 lat itd. (dla cechy ilościowej skokowej).

W pkt. a) należy określić zasadę zliczania jednostek statystycznych do danego przedziału a mianowicie: czy do przedziału 20-25 lat kwalifikujemy pracowników w wieku od 20 lat do mniej niż 25 lat. Wówczas przedziały klasowe możemy zapisać w następującej postaci:

[20-25), [25-30), [30-35), [35-40), [40-45)

a może osoby od 20 lat do 25 lat włącznie wtedy przedziały możemy przedstawić:

(20-25], (25-30], (30-35], (35-40], (40-45]

Gdzie nawiasy oznaczają: „( )" przedział niedomknięty, „[ ]" przedział domknięty.
Stosuje się także szeregi statystyczne otwarte dołem i górą np.:

poniżej 20, 20-25, 25-30, 30-35, 35-40, 40-45, 45 lat i więcej

Szeregi klasyfikuje się ponadto ze względu na treść wówczas wyróżnia się szeregi:

- strukturalne,
- przestrzenne
- i czasowe.

Materiał statystyczny poddany badaniu statystycznemu możemy pogrupować w zależności od rodzaju badanej cechy. W przypadku cechy jakościowej mamy do czynienia z tzw. **grupowaniem typologicznym**. Tworzymy wówczas jednorodne podzbiory jednostek statycznych przyporządkowanych do znanych wcześniej wariantów badanej cechy. Z kolei **grupowanie wariancyjne** stosuje się dla cechy ilościowej. Liczbę przedziałów klasowych i ich rozpiętość (rozstęp, interwał), wyznacza się stosując następujące wzory:

- **liczba przedziałów klasowych (*k*):**

$$k = 1 + 3{,}322\, log\, n \tag{1.5.2}$$

$$k \leq 5\, log\, n \tag{1.5.3}$$

$$k \approx \sqrt{n} \tag{1.5.4}$$

gdzie:

*n* – liczebność zbiorowości.

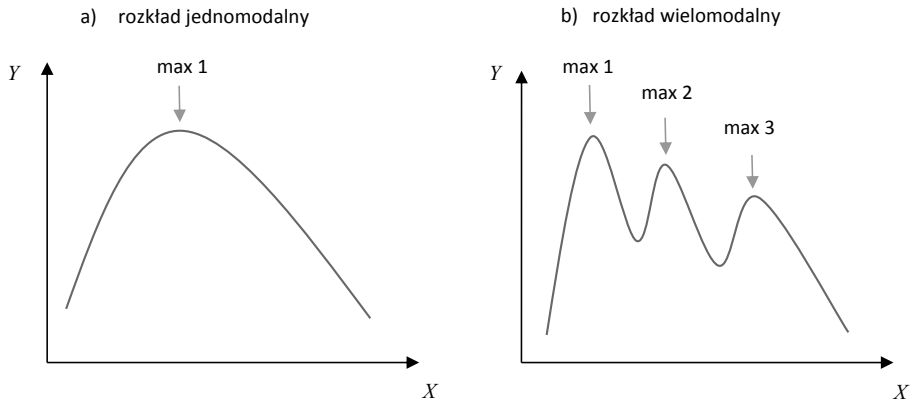
- **rozpiętość przedziału (*h*):**

$$h = \frac{x_{max} - x_{min}}{k} \tag{1.5.5}$$

gdzie:

$x_{min}$ – wartość minimalna uporządkowanego szeregu,
$x_{max}$ – wartość maksymalna uporządkowanego szeregu,
*k* – liczba przedziałów klasowych obliczona według wzorów 1.5.2-1.5.4.

Rozpiętość przedziałów klasowych *h* (wzór 1.5.5) stanowi różnicę pomiędzy wartościami maksymalną i minimalną uporządkowanego szeregu statystycznego dzieloną przez liczbę przedziałów klasowych *k*. Dzięki tej metodzie rozpiętość przedziałów klasowych jest taka sama. Liczba przedziałów dla badanej cechy zależnie od zastosowanego wzoru może się różnić. Ważne jest, aby zagregowana cecha charakteryzowała się tzw. rozkładem jednomodalnym (który posiada tylko jedno maksimum – rys. 1.5.1) oraz brakiem pustych przedziałów klasowych.




**Rysunek 1.5.1. Przykłady rozkładów jednomodalnych i wielomodalnych**
Źródło: opracowanie własne.

Ważną rolę odgrywa sposób przedstawiania danych. Dane statystyczne prezentuje się z wykorzystaniem tablic statystycznych lub w postaci graficznej. Zależnie od stopnia szczegółowości danych wyróżnia się **tablice proste** i **złożone**. **Tablica prosta** przedstawia jedną badaną cechę statystyczną np. poziom wykształcenia (tab. 1.5.4). W tablicy występuje tylko jeden szereg danych liczbowych.

**Tabela 1.5.4. Pracujący według poziomu wykształcenia w firmie X – przykład tablicy prostej**

| Poziom ukończonego wykształcenia | Liczba pracowników w firmie X |
|---|---|
| Wyższe | 20 |
| Policealne i średnie zawodowe | 25 |
| Średnie ogólnokształcące | 34 |
| Zasadnicze zawodowe | 15 |
| Gimnazjalne, podstawowe i niepełne podstawowe | 6 |
| **Ogółem** | **100** |

Źródło: opracowanie własne. Dane umowne.

**Tablica złożona** zestawia dane statystyczne ze względu na dwie lub więcej cech np.: poziom wykształcenia i miejsce zamieszkania (tab. 1.5.5). Wyróżnia się tablice złożone: zbiorcze i kombinowane. W tablicy złożonej nie powinno być prezentowane zbyt dużo danych statystycznych co może wpływać negatywnie na ich analizę.

**Tabela 1.5.5. Pracujący według poziomu wykształcenia i miejsca zamieszkania w firmie X – przykład tablicy złożonej**

| Poziom ukończonego wykształcenia | Miejsce zamieszkania | | |
|---|---|---|---|
| | razem | miasto | wieś |
| Wyższe | 20 | 12 | 8 |
| Policealne i średnie zawodowe | 25 | 20 | 5 |
| Średnie ogólnokształcące | 34 | 30 | 4 |
| Zasadnicze zawodowe | 15 | 10 | 5 |
| Gimnazjalne, podstawowe i niepełne podstawowe | 6 | 3 | 3 |
| **Ogółem** | **100** | **75** | **25** |

Źródło: opracowanie własne. Dane umowne.

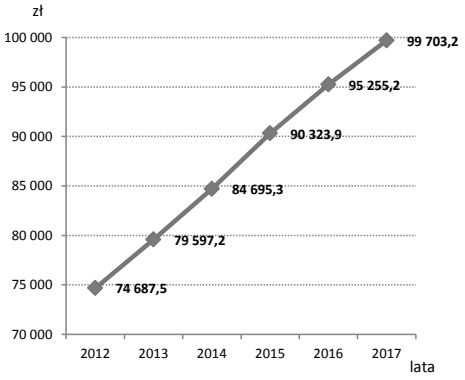
W tablicach statystycznych często stosuje się znaki umowne (tab. 1.5.6).

**Tabela 1.5.6. Zestawienie wybranych znaków umownych stosowany w tablicach statystycznych**

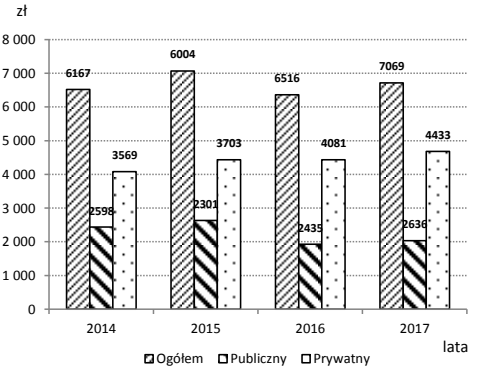
| Oznaczenia znaków | Objaśnienie |
|---|---|
| **Kreska (–)** | dane zjawisko nie wystąpiło |
| **Zero (0)** | dane zjawisko występuje ale nie można go wyrazić w jednostkach miary stosowanych w tablicach |
| **Kropka (.)** | zjawisko istnieje, lecz brak o nim informacji lub wiarygodnych informacji |
| **Krzyżyk (x)** | wypełnienie niecelowe lub niemożliwe ze względu na układ tablicy |
| **Wykrzyknik (!)** | rzadko stosowany znak, oznacza, że liczba podana jest poprawniejsza od poprzedniej podanej |
| **Znak (#)** | oznacza, że dane nie mogą być opublikowane ze względu na konieczność zachowania tajemnicy statystycznej w rozumieniu ustawy o statystyce publicznej |
| **„w tym"** | nie podaje się wszystkich składników sumy |
| **„z tego"** | podaje się wszystkie składniki sumy |

Źródło: *Rocznik Statystyczny Rzeczpospolitej Polski 2018 r.* , GUS, Warszawa, 2018 r., s. 29.

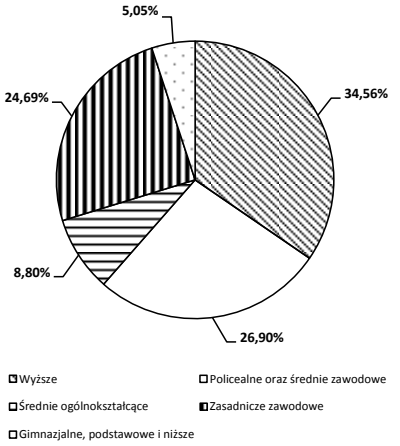
**Wykres** to graficzne narzędzie analizy i prezentacji danych statystycznych. Taki sposób wizualizacji danych z wykorzystaniem różnorodnych barw, kształtów jest bardziej czytelny i przyjazny dla odbiorcy. Wykres poprawnie zbudowany składa się: tytułu wykresu, pola wykresu, skali, legendy, źródła danych statystycznych. Uwzględniając kształt obrazu wyróżnia się wykresy: punktowe, liniowe, słupkowe, kołowe, warstwowe, powierzchniowe, bryłowe, radarowe mapowe, obrazkowe, korelacyjne itd. Przykładową prezentację danych zamieszczono na rys. 1.5.2-1.5.7.



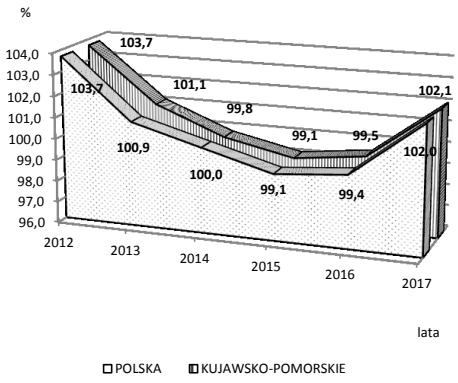

**Rysunek 1.5.2. Wartość brutto środków trwałych w gospodarce narodowej na 1 mieszkańca w Polsce w latach 2012-2017**
Źródło: Bank Danych Lokalnych GUS.



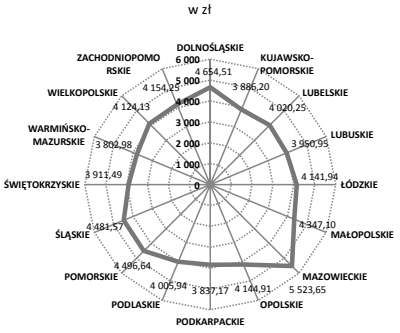

**Rysunek 1.5.3. Nakłady inwestycyjne na 1 mieszkańca w Polsce w latach 2014-2017**
Źródło: Bank Danych Lokalnych GUS.



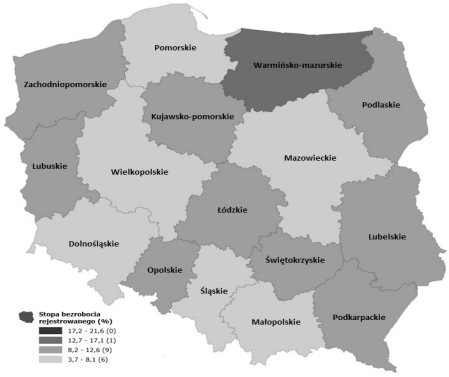

**Rysunek 1.5.4. Struktura pracujących według poziomu wykształcenia na podstawie BAEL w Polsce w 2017 r.**
Źródło: Bank Danych Lokalnych GUS.




**Rysunek 1.5.5. Wskaźnik cen towarów i usług konsumpcyjnych w Polsce i w województwie kujawsko-pomorskim w 2017 r.**
Źródło: Bank Danych Lokalnych GUS.




**Rysunek 1.5.6. Przeciętne miesięczne wynagrodzenie brutto według województw w 2017 r.**
Źródło: Bank Danych Lokalnych GUS.




**Rysunek 1.5.7. Stopa bezrobocia rejestrowanego w Polsce w 2016 r.**
Źródło: Bank Danych Lokalnych GUS.

## Przykłady

**Przykład 1.1.**

Oblicz wskaźniki struktury na podstawie danych przedstawiających liczbę bezrobotnych pozostających bez pracy według płci w Polsce w 2017 r. Proszę zaprezentować otrzymane wyniki metodą graficzną.

| Czas pozostawienia bez pracy | Bezrobotni | |
|---|---|---|
| | mężczyźni | kobiety |
| 3 miesiące i mniej | 171642 | 151579 |
| 3 - 6 miesięcy | 71280 | 88507 |
| 6 - 12 miesięcy | 70285 | 90049 |
| powyżej12 miesięcy | 173009 | 265395 |
| **Ogółem** | **486216** | **595530** |

Źródło: Bank Danych Lokalnych GUS.

**Rozwiązanie**

1. **Obliczam wskaźniki struktury (wzór 1.5.1):**

Obliczenia pomocnicze:

| Mężczyźni | Kobiety |
|---|---|
| $w_s = \frac{n_i}{N} \cdot 100 = \frac{171642}{486216} \cdot 100 = 35{,}3\%$ | $w_s = \frac{n_i}{N} \cdot 100 = \frac{151579}{595530} \cdot 100 = 25{,}5\%$ |
| $w_s = \frac{n_i}{N} \cdot 100 = \frac{71280}{486216} \cdot 100 = 14{,}7\%$ | $w_s = \frac{n_i}{N} \cdot 100 = \frac{885070}{595530} \cdot 100 = 14{,}9\%$ |
| $w_s = \frac{n_i}{N} \cdot 100 = \frac{70285}{486216} \cdot 100 = 14{,}5\%$ | $w_s = \frac{n_i}{N} \cdot 100 = \frac{90049}{595530} \cdot 100 = 15{,}1\%$ |
| $w_s = \frac{n_i}{N} \cdot 100 = \frac{173009}{486216} \cdot 100 = 35{,}6\%$ | $w_s = \frac{n_i}{N} \cdot 100 = \frac{265395}{595530} \cdot 100 = 44{,}6\%$ |

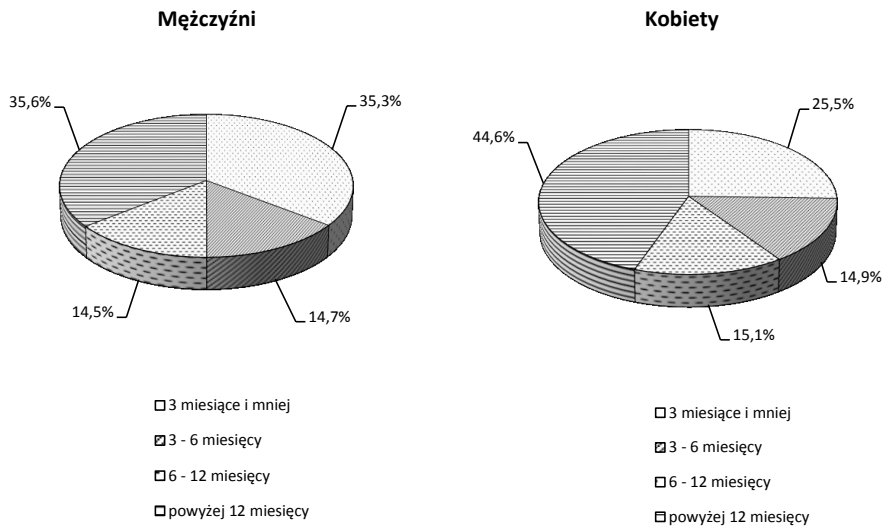
Źródło: opracowanie własne.

| Czas pozostawienia bez pracy | Bezrobotni | | | |
|---|---|---|---|---|
| | mężczyźni | | kobiety | |
| | liczba (w osobach) | wskaźniki struktury (%) | liczba (w osobach) | wskaźniki struktury (%) |
| 3 miesiące i mniej | 171642 | 35,3 | 151579 | 25,5 |
| 3 - 6 miesięcy | 71280 | 14,7 | 88507 | 14,9 |
| 6 - 12 miesięcy | 70285 | 14,5 | 90049 | 15,1 |
| powyżej12 miesięcy | 173009 | 35,6 | 265395 | 44,6 |
| **Ogółem** | **486216** | **100,0** | **595530** | **100,0** |

Źródło: opracowanie własne.

**2. Prezentacja graficzna:**

**Struktura bezrobotnych pozostających bez pracy według płci w Polsce w 2017 r.**




Źródło: opracowanie własne.

**Interpretacja:** Największy udział bezrobotnych w Polsce w 2017 r. zarówno wśród mężczyzn, jak i kobiet stanowiły osoby pozostające bez pracy powyżej 12 miesięcy. Odsetek ten wynosił odpowiednio: mężczyźni – 35,6% i kobiety – 44,6%. Najmniej liczną grupę bezrobotnych wśród mężczyzn stanowiły osoby pozostające bez pracy 6-12 miesięcy (14,5%), a wśród kobiet 3-6 miesięcy (14,9%).

**Przykład 1.2.**

W pewnej firmie „A" z branży elektronicznej zatrudnionych jest 28 pracowników z następującym stażem pracy (cecha ciągła) – dane umowne:

5,0; 1,0; 2,1; 4,0; 5,0; 7,0; 6,2; 10,0; 6,0; 19,0; 6,5; 7,0; 8,4; 7,0;
10,0; 10,0; 1,0; 10,0; 7,0; 10,0; 12,5; 12,0; 6,3; 13,0; 14,0 15,1; 12,3; 16,0

Na podstawie informacji o stażu pracy pracowników:

a) przeprowadź agregację danych poprzez budowę szeregu rozdzielczego (grupowanie wariancyjne),
b) przedstaw otrzymane wyniki agregacji w postaci graficznej.

**Rozwiązanie**

Przed przystąpieniem obliczeń należy uporządkować rosnąco szereg statystyczny.

**a) grupowanie wariancyjne:**

**1,0**; 1,0; 2,1; 4,0; 5,0; 5,0; 6,0; 6,2; 6,3; 6,5; 7,0; 7,0; 7,0; 7,0;
8,4; 10,0; 10,0; 10,0; 10,0; 10,0; 12,0; 12,3; 12,5; 13,0; 14,0; 15,1; 16,0; **19,0**

$x_{min}$ $x_{max}$

$n = 28$

- **grupowanie I (wzór 1.5.2):**

$$k = 1 + 3{,}322 \log n = 1 + 3{,}322 \log 28 = 5{,}81 \approx 6{,}0$$

$$h = \frac{x_{max} - x_{min}}{k} = \frac{19 - 1}{6{,}0} = 3{,}0$$

| Staż pracy (w latach) $x_i$ | Liczba pracowników $n_i$ |
|---|---|
| 1-4 | 4 |
| 4-7 | 10 |
| 7-10 | 6 |
| 10-13 | 4 |
| 13-16 | 3 |
| 16-19 | 1 |
| **Ogółem** | **28** |

**6 przedziałów klasowych**

Źródło: opracowanie własne.

- **grupowanie II (wzór 1.5.3):**

$$k \leq 5 \log n = 5 \log 28 = 7{,}24 \approx 7{,}0$$

$$h = \frac{x_{max} - x_{min}}{k} = \frac{19 - 1}{7{,}0} = 2{,}57 \approx 3{,}0$$

| Staż pracy (w latach) $x_i$ | Liczba pracowników $n_i$ |
|---|---|
| 1-4 | 4 |
| 4-7 | 10 |
| 7-10 | 6 |
| 10-13 | 4 |
| 13-16 | 3 |
| 16-19 | 1 |
| 19-22 | przedział pusty |
| **Ogółem** | **28** |

**7 przedziałów klasowych**

Źródło: opracowanie własne.

- **grupowanie III (wzór 1.5.4):**

$$k \approx \sqrt{n} = \sqrt{28} = 5{,}29 \approx 5{,}0$$

$$h = \frac{x_{max} - x_{min}}{k} = \frac{19 - 1}{5{,}0} = 3{,}6 \approx 4{,}0$$

| Staż pracy (w latach) $x_i$ | Liczba pracowników $n_i$ |
|---|---|
| 1-5 | 6 |
| 5-9 | 9 |
| 9-13 | 9 |
| 13-17 | 3 |
| 17-21 | 1 |
| **Ogółem** | **28** |

**5 przedziałów klasowych**

Źródło: opracowanie własne.

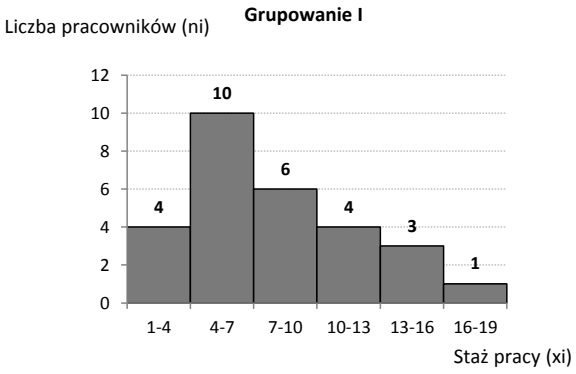
Wyniki agregacji pracowników według stażu pracy zestawiono w tabeli poniżej:

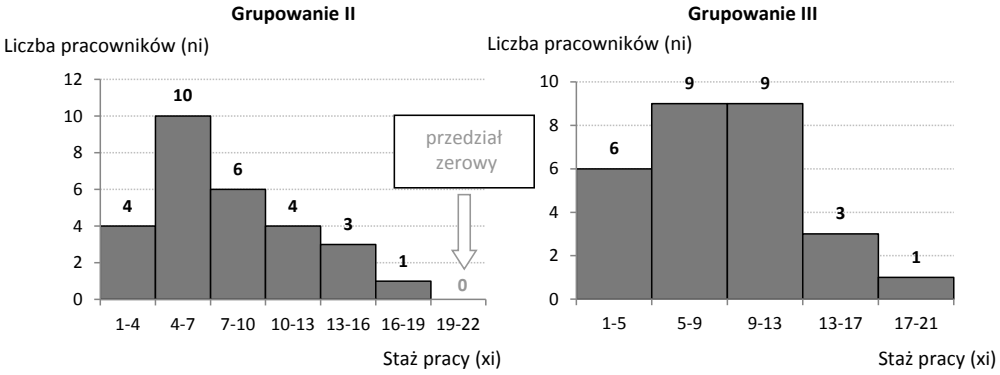
| **Wyszczególnienie** | **Grupowanie** | | |
|---|---|---|---|
| | **I** | **II** | **III** |
| Liczba klas | $k = 6$ | $k = 7$ | $k = 5$ |
| Rozpiętość | $h = 3$ | $h = 3$ | $h = 4$ |

Źródło: opracowanie własne.

**b) prezentacja graficzna wyników agregacji:**

**Histogramy liczebności pracowników według stażu pracy**






Źródło: opracowanie własne.

**Interpretacja:** Na podstawie otrzymanych wyników, najlepsze rezultaty otrzymaliśmy w grupowaniu I. Występuje jedno maksimum przypadające w przedziale 4-7 lat (10 pracowników). Uporządkowany i zagregowany szereg statystyczny posiada cechy rozkładu jednomodalnego. Nie występują puste przedziały klasowe (tak jak w grupowaniu II, gdzie ostatni przedział ma zerową liczebność).

**Przykład 1.3.**

W 20 bankach spółdzielczych w województwie Y odnotowano na koniec grudnia 2019 r. następującą liczbę pracujących (cecha skokowa) – dane umowne:

25; 22; 23; 27; 24; 25; 29; 26; 27; 27; 28; 26; 29; 25; 30; 31; 32; 36; 28; 33

Na podstawie informacji o liczbie pracowników:

a) przeprowadź agregację danych poprzez budowę szeregu rozdzielczego (grupowanie wariancyjne),
b) przedstaw otrzymane wyniki agregacji w postaci graficznej.

**Rozwiązanie**

Przed przystąpieniem obliczeń należy uporządkować rosnąco szereg statystyczny.

**a) grupowanie wariancyjne:**

**22**; 23; 24; 25; 25; 25; 26; 26; 27; 27; 27; 28; 28; 29; 29; 30; 31; 32; 33; **36**

$x_{min}$ $n = 20$ $x_{max}$

- **grupowanie I (wzór 1.5.2):**

$$k = 1 + 3{,}322 \log n = 1 + 3{,}322 \log 20 = 5{,}32 \approx 5{,}0$$

$$h = \frac{x_{max} - x_{min}}{k} = \frac{36 - 22}{5{,}0} = 2{,}8 \approx 3{,}0$$

| Liczba pracowników $x_i$ | Liczba banków $n_i$ |
|---|---|
| 22-24 | 3 |
| 25-27 | 8 |
| 28-30 | 5 |
| 31-33 | 3 |
| 34-36 | 1 |
| **Ogółem** | **20** |

**5 przedziałów klasowych**

Źródło: opracowanie własne.

- **grupowanie II (wzór 1.5.3):**

$$k \leq 5 \log n = 5 \log 20 = 6{,}51 \approx 7{,}0$$

$$h = \frac{x_{max} - x_{min}}{k} = \frac{36 - 22}{7{,}0} = 2{,}0$$

| Liczba pracowników $x_i$ | Liczba banków $n_i$ |
|---|---|
| 22-23 | 2 |
| 24-25 | 4 |
| 26-27 | 5 |
| 28-29 | 4 |
| 30-31 | 2 |
| 32-33 | 2 |
| 34-35 | przedział pusty |
| **Ogółem** | **19** Nie zakwalifikował się 20 bank |

**7 przedziałów klasowych**

Źródło: opracowanie własne.

- **grupowanie III (wzór 1.5.4):**

$$k \approx \sqrt{n} = \sqrt{20} = 4{,}47 \approx 4{,}0$$

$$h = \frac{x_{max} - x_{min}}{k} = \frac{36 - 22}{4{,}0} = 3{,}5 \approx 4{,}0$$

| Liczba pracowników $x_i$ | Liczba banków $n_i$ |
|---|---|
| 22-25 | 6 |
| 26-29 | 9 |
| 30-33 | 4 |
| 34-37 | 1 |
| **Ogółem** | **20** |

**4 przedziałów klasowych**

Źródło: opracowanie własne.

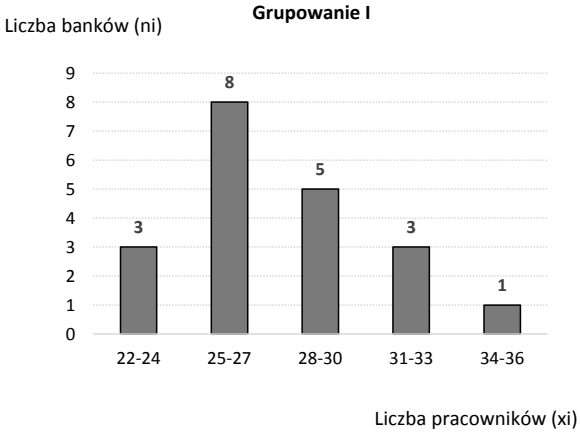
Wyniki agregacji według liczby pracowników zestawiono w tabeli poniżej:

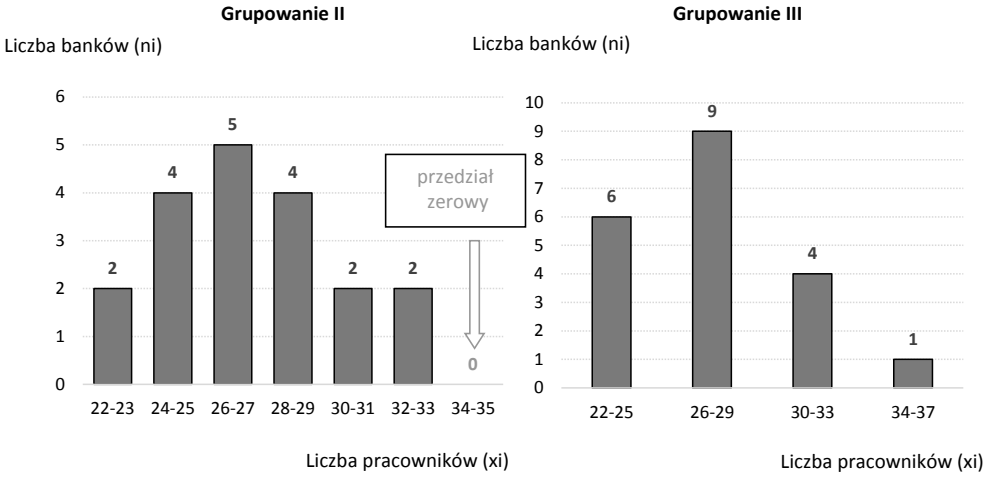
| **Wyszczególnienie** | **Grupowanie** | | |
|---|---|---|---|
| | **I** | **II** | **III** |
| Liczba klas | $k = 5$ | $k = 7$ | $k = 4$ |
| Rozpiętość | $h = 3$ | $h = 2$ | $h = 4$ |

Źródło: opracowanie własne.

**c) prezentacja graficzna wyników agregacji:**

**Histogramy liczebności banków według liczby pracowników**






**Interpretacja:** Na podstawie otrzymanych wyników, najlepsze rezultaty otrzymaliśmy w grupowaniu I.

# 2. Analiza struktury zbiorowości

## 2.1. Miary tendencji centralnej (średnie)

Miary opisu statystycznego stosujemy w odniesieniu do pojedynczych zmiennych, kiedy chcemy obliczyć typowe wartości zmiennej, wewnętrzne rozproszenie badanej zbiorowości, czy określić ocenę rozkładu. Metody analizy struktury charakteryzują badaną zbiorowość statystyczną. Wyróżniamy miary klasyczne i pozycyjne, które dzielimy dodatkowo na miary: przeciętne (średnie), zróżnicowania (zmienności, dyspersji), asymetrii (skośności) i koncentracji (skupienia).

Do podstawowych miar tendencji centralnej zalicza się:

a) miary klasyczne:
   – średnia arytmetyczna,
   – średnia geometryczna,
   – średnia harmoniczna,
   – średnia chronologiczna.
b) miary pozycyjne:
   – dominanta,
   – mediana,
   – kwartyle,
   – decyle, percentyle.

**Średnia arytmetyczna** ($\bar{x}$) to suma wartości zmiennej ($x_i$) wszystkich jednostek badanej zbiorowości podzielona przez liczbę jednostek zbiorowości ($N$). Zależnie od analizowanego szeregu średnia arytmetyczna może być prosta (2.1.1) lub ważona (2.1.2 i 2.1.3).

- dla szeregu prostego (szczegółowego):

$$\bar{x} = \frac{x_1 + x_2 + x_3 + \ \ldots + x_N}{N} = \frac{\sum_{i=1}^{N} x_i}{N} \tag{2.1.1}$$

gdzie: $x_i$ – warianty cechy mierzalnej $X$,
$N$ – liczebność całej zbiorowości statystycznej.

- dla szeregu rozdzielczego jednopunktowego:

$$\bar{x} = \frac{x_1 n_1 + x_2 n_2 + x_3 n_3 + \ \ldots + x_k n_k}{n_1 + n_2 + n_3 + \ \ldots + n_k} = \frac{\sum_{i=1}^{k} x_i n_i}{N} \tag{2.1.2}$$

gdzie: $n_i$ – wagi (liczebności cząstkowe) odpowiadające poszczególnym wariantom zmiennej $X$.

- dla szeregu rozdzielczego z przedziałami klasowymi:

$$\bar{x} = \frac{\dot{x}_i n_1 + \dot{x}_i n_2 + \dot{x}_i n_3 + \ \ldots + \dot{x}_i n_k}{n_1 + n_2 + n_3 + \ \ldots + n_k} = \frac{\sum_{i=1}^{k} \dot{x}_i n_i}{N} \tag{2.1.3}$$

gdzie:

$\dot{x}_i$ – środek przedziału klasowego.

W przypadku szeregu rozdzielczego wielopunktowego środki przedziałów liczbowych oblicza się ze wzorów zamieszczonych w tab. 2.1.1.

**Tabela 2.1.1. Metody obliczania środków przedziałów klasowych**

| *Górna granica pokrywa się z dolną granicą przedziału*<br><br>$x_i$<br>Wiek pracowników (w latach) | Środek przedziału<br><br>$\dot{x}_i = \frac{x_d + x_g}{2}$ | *Górna granica nie pokrywa się z dolną granicą*<br><br>$x_i$<br>Wiek pracowników (w latach) | Środek przedziału<br><br>$\dot{x}_i = \frac{x_d + x_{dn}}{2}$ |
|---|---|---|---|
| $x_d$ **20 - 25** $x_g$ | $\frac{20+25}{2} = 22{,}5$ | $x_d$ **20** - 24 | $\frac{20+25}{2} = 22{,}5$ |
| 25 - 30 | 27,5 | $x_{dn}$ **25** - 29 | 27,5 |
| 30 - 35 | 32,5 | 30 - 34 | 32,5 |
| 35 - 40 | 37,5 | 35 - 39 | 37,5 |

gdzie: $x_d$ – dolna granica przedziału klasowego, $x_{dn}$ – dolna granica następnego przedziału klasowego, $x_g$ – górna granica przedziału klasowego.
Źródło: opracowanie własne.

**Średnia geometryczna** ($\bar{x}_g$) ma zastosowanie do obliczania średniego tempa zmian w analizie szeregów czasowych. Wyróżnia się średnią prostą dla szeregów wyliczających i ważoną dla szeregów rozdzielczych. Średnia geometryczna prosta jest *n*-tym pierwiastkiem z iloczynu wszystkich wartości badanej zmiennej ($x_i$):

$$\bar{x}_g = \sqrt[n]{x_1 \cdot x_2 \cdot x_3 \ldots \cdot x_n} \tag{2.1.4}$$

gdzie: $x_1 \cdot x_2 \cdot x_3 \ldots \cdot x_n$ są iloczynami wartości indeksów łańcuchowych, które można zapisać w postaci $\prod_{i=1}^{n} x_i$.

Z uwagi na trudności obliczeniowe pierwiastków *n*-tego stopnia wzór (2.1.4) można przekształcić do postaci logarytmicznej:

$$\log \bar{x}_g = \frac{1}{n}\left(log\, x_1 + log\, x_2 + log\, x_3 + \cdots + log\, x_n\right) = \frac{1}{n}\sum_{i=1}^{n} \log x_i \tag{2.1.5}$$

Średnią geometryczną ważoną możemy zapisać w postaci wzoru:

$$\bar{x}_g = \sqrt[N]{x_1^{n_1} \cdot x_2^{n_2} \cdot x_3^{n_3} \ldots \cdot x_n^{n_k}} \tag{2.1.6}$$

lub jako rachunek logarytmiczny:

$$log\, \bar{x}_g = \frac{1}{n}\left(n_1 log\, x_1 + n_2 log\, x_2 + n_3 log\, x_3 + \cdots + n_n log\, x_n\right) = \frac{1}{n}\sum_{i=1}^{n} n_i log\, x_i \tag{2.1.7}$$

We wzorach (2.1.5) i (2.1.7) zamiast logarytmu dziesiętnego (log) stosuje się logarytm naturalny (*ln*, dla którego podstawa wynosi $e \approx 2{,}718$ ... ) – patrz przykład 5.4.

**Średnią chronologiczną** ($\bar{x}_{ch}$) oblicza się jako sumę wszystkich wartości zmiennej ($x_i$) dzielonych przez $N - 1$ jednostek zbiorowości, przy czym pierwszy i ostatni wyraz stanowi połowę wartości:

$$\bar{x}_{ch} = \frac{0{,}5 \cdot x_1 + x_2 + x_3 + \ \ldots + 0{,}5 \cdot x_n}{N - 1} \tag{2.1.8}$$

Średnią stosuje się dla określonych momentów czasowych tzn. kiedy dane dotyczą stanu na koniec lub początku dnia, miesiąca, roku.

**Średnia harmoniczna** ($\bar{x}_{har}$) równa jest odwrotności średniej arytmetycznej z odwrotności wartości badanej zmiennej ($x_i$). Zależnie od analizowanego szeregu średnią harmoniczną możemy zapisać:

- dla szeregu prostego:

$$\bar{x}_{har} = \frac{N}{\frac{1}{x_1} + \frac{1}{x_2} + \frac{1}{x_3} + \cdots + \frac{1}{x_n}} = \frac{N}{\sum_{i=1}^{N} \frac{1}{x_i}} \tag{2.1.9}$$

- dla szeregu rozdzielczego:

$$\bar{x}_{har} = \frac{n_1 + n_2 + n_3 + \cdots + n_k}{\frac{n_1}{x_1} + \frac{n_2}{x_2} + \frac{n_3}{x_3} + \cdots + \frac{n_k}{x_k}} = \frac{N}{\sum_{i=1}^{k} \frac{n_i}{x_i}} \tag{2.1.10}$$

**Dominanta** (***D***) jest to wartość zmiennej, która w badanej zbiorowości powtarza się najczęściej (dominująca, typowa). Dominanta nazywana jest wartością modalną lub modą.

W szeregu prostym wartości zmiennej należy uporządkować od wartości najmniejszej do największej tzn. od $x_{min}$ do $x_{max}$ a następnie zliczamy wartości, które się powtarzają. Na poniższym przykładzie analizując oceny studenta ze statystyki:

3, 3, 4, **5, 5, 5** $N = 6$

dominującą oceną jest 5 ($D = 5$), ponieważ wystąpiła ona najczęściej bo, aż 3 krotnie.

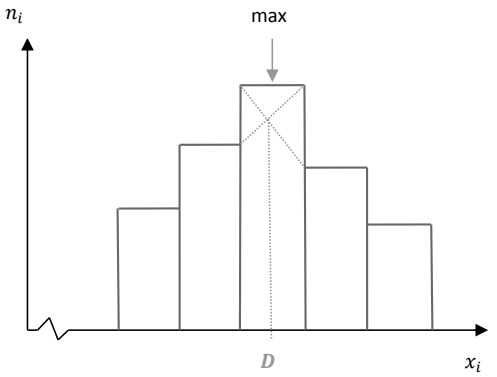
Dominantę w szeregu rozdzielczym wyznaczamy według wzoru interpolacyjnego:

$$D = x_0 + \frac{(n_D - n_{D-1}) \cdot h_0}{(n_D - n_{D-1}) + (n_D - n_{D+1})} \tag{2.1.11}$$

gdzie:

$x_0$ – dolna granica przedziału klasowego dominanty,
$n_D$ – liczebność przedziału klasowego dominanty,
$n_{D-1}$– liczebność przedziału klasowego poprzedzającego przedział dominanty,
$n_{D+1}$– liczebność przedziału klasowego następującego przedział dominanty,
$h_0$ – rozpiętość przedziału klasowego, w którym znajduje się dominanta.

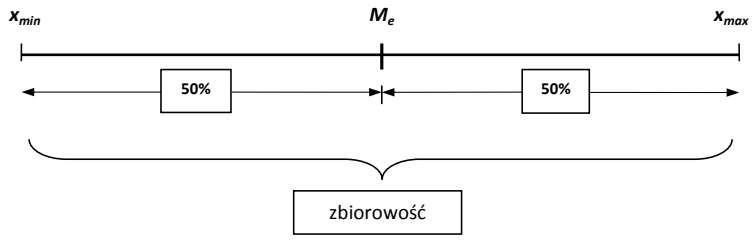
Dominanta występuje w najliczniejszym przedziale klasowym (gdzie liczebność cząstkowa $n_i$ = max). Wartość dominanty można wyznaczyć w sposób graficzny na podstawie histogramu liczebności (rys. 2.1.1).




**Rysunek 2.1.1. Graficzna metoda wyznaczenia dominanty**
Źródło: opracowane na podstawie A. Zeliaś, B. Pawełek, S. Wanat, *Metody statystyczne. Zadania i sprawdziany*, PWE, Warszawa 2002, s. 73.

Wykreślamy dwie przekątne z sąsiednich przedziałów klasowych łącząc je z górnymi wierzchołkami najwyższego prostokąta. Przekątne przecinając się tworząc punkt, z którego wykreślamy prostopadłą względem osi *X* wskazując tym samym wartość dominanty.

**Mediana** ($M_e$) to wartość zmiennej, która dzieli badaną zbiorowość na dwie równe części. Mediana jako miara pozycyjna uzależniona jest od pozycji jaką zajmuje w szeregu statystycznym, tak więc przed jej wyznaczeniem należy uporządkować badaną zbiorowość np. rosnąco od $x_{min}$ do $x_{max}$. Mediana nazywana jest wartością środkową lub kwartylem drugim – $Q_2$. Interpretacja mediany jest następująca: **50% jednostek zbiorowości statystycznej ma wartości cechy niższe lub równe medianie – (*nie większe od* $M_e$) a druga połowa jednostek ma wartości większe lub równe medianie – (*nie mniejsze od* $M_e$).**




Źródło: opracowanie własne.

W szeregu prostym medianę wyznaczamy w zależności czy liczba obserwacji *N* jest parzysta lub nieparzysta:

- *N* – nieparzyste:

$$M_e = x_{\frac{N+1}{2}} \tag{2.1.12}$$

- *N* – parzyste:

$$M_e = \frac{x_{\frac{N}{2}} + x_{\frac{N}{2}+1}}{2} \tag{2.1.13}$$

W szeregu rozdzielczym jednopunktowym (cecha skokowa) medianę wyznacza się na podstawie wartości skumulowanych (*cum*). Mediana jest tą wartością zmiennej, której liczebność skumulowana zawiera numer jednostki $Nr_{M_e}$ wyliczonej według wzorów:

- *N* – parzyste: $$Nr_{M_e} = \frac{N}{2} \tag{2.1.14}$$
- *N* – nieparzyste: $$Nr_{M_e} = \frac{N+1}{2} \tag{2.1.15}$$

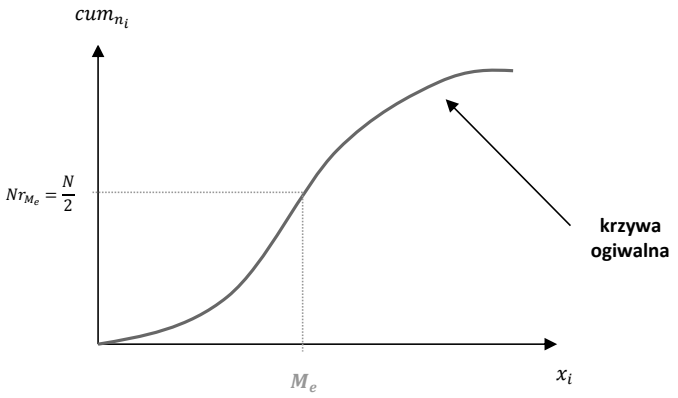
Dla szeregu rozdzielczego z przedziałami klasowymi stosuje się wzór interpolacyjny:

$$M_e = x_0 + \left(\frac{N}{2} - cum_{n_{i-1}}\right)\frac{h_0}{n_{M_e}} \tag{2.1.16}$$

gdzie:

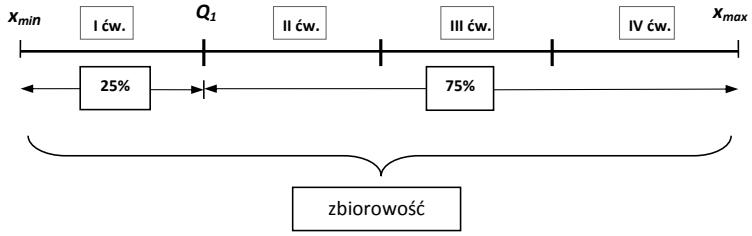
$x_0$ – dolna granica przedziału klasowego mediany,
$\frac{N}{2}$ – połowa liczebności badanej zbiorowości zawarta w przedziale skumulowanym *cum*,
$cum_{n_{i-1}}$ – liczebność skumulowana (kumulacyjna) poprzedzająca przedział klasowy mediany,
$h_0$ – rozpiętość przedziału klasowego, w którym znajduje się mediana,
$n_{M_e}$ – liczebność przedziału klasowego mediany.

Medianę również można wyznaczyć metodą graficzną (rys. 2.1.2). Wykreślamy tzw. **krzywą ogiwalną** na podstawie wartości skumulowanych liczebności (*cum*). Na osi *Y* zaznaczamy numer mediany $Nr_{M_e}$, z którego prowadzimy równolegle do osi *X* prostą łącząc go z grzywą skumulowaną. W punkcie przecięcia prostej z krzywą ogiwalną wykreślamy prostą prostopadłą względem osi *X* wyznaczając w ten sposób wartość mediany.



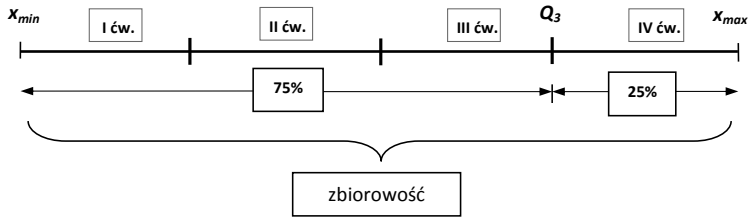

**Rysunek 2.1.2. Graficzna metoda wyznaczenia mediany**
Źródło: opracowane na podstawie W. Ignatczyk, M. Chromińska, *Statystyka. Teoria i zastosowanie*, WSB, Poznań 2004, s. 89.

**Kwartyl pierwszy ($Q_1$)** inaczej kwartyl dolny to wartość zmiennej, która dzieli uporządkowaną zbiorowość na dwie części tzn. na 25% i 75%. Interpretacja kwartyla pierwszego jest następująca: **25% jednostek zbiorowości ma wartości cechy niższe lub równe $Q_1$ – (*nie większe od $Q_1$*), a 75% jednostek ma wartości wyższe lub równe $Q_1$ – (*nie mniejsze od $Q_1$*).**




Źródło: opracowanie własne.

**Kwartyl trzeci ($Q_3$)** inaczej kwartyl górny to wartość zmiennej, która dzieli uporządkowaną zbiorowość na dwie części tzn. na 75% i 25%. Interpretacja kwartyla trzeciego jest następująca: **75% jednostek zbiorowości ma wartości cechy niższe lub równe $Q_3$ – *(nie większe od $Q_3$)*, a 25% jednostek ma wartości wyższe lub równe $Q_3$ – (*nie mniejsze od $Q_3$*).**




Źródło: opracowanie własne.

W szeregu prostym kwartyle $Q_1$ i $Q_3$ oblicza się w dwóch etapach. W pierwszej kolejności badaną zbiorowość dzielimy się na dwie połowy wyznaczając medianę korzystając ze wzorów (2.1.12) i (2.1.13), a następnie każdą połowę dodatkowo dzielimy na dwie równe części. Wówczas otrzymujemy zbiorowość podzieloną na cztery części. Każda część zawiera 25% zbiorowości.

W szeregu rozdzielczym jednopunktowym w pierwszej kolejności oblicza się numery jednostek analizowanej zbiorowości $Nr_{Q_1}$ i $Nr_{Q_3}$ według następujących wzorów:

- $N$ – parzyste: $$Nr_{Q_1} = \frac{N}{4} \quad (2.1.17)$$
- $N$ – nieparzyste: $$Nr_{Q_1} = \frac{N+1}{4} \quad (2.1.18)$$
- $N$ – parzyste: $$Nr_{Q_3} = \frac{3N}{4} \quad (2.1.19)$$

- $N$ – nieparzyste: $$Nr_{Q_3} = \frac{3(N+1)}{4} \quad (2.1.20)$$

Następnie obliczone numery jednostek zbiorowości przyporządkowuje się do przedziałów wartości skumulowanych. W przypadku szeregów rozdzielczych wartości kwartyli wyznacza się na podstawie wzorów:

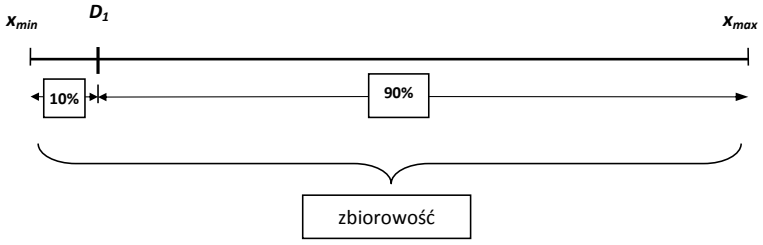
$$Q_1 = x_0 + \left(\frac{N}{4} - cum_{n_{i-1}}\right)\frac{h_0}{n_{Q_1}} \quad (2.1.21)$$

$$Q_3 = x_0 + \left(\frac{3N}{4} - cum_{n_{i-1}}\right)\frac{h_0}{n_{Q_3}} \quad (2.1.22)$$

gdzie:

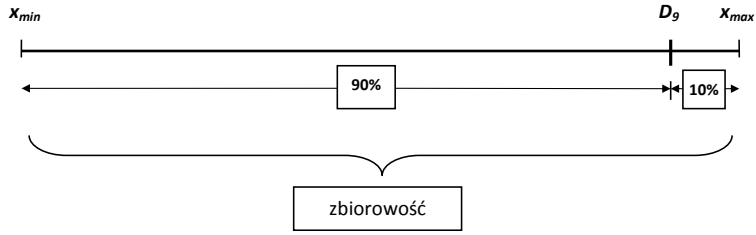
$x_0$ – dolna granica przedziału klasowego $Q_1$ i $Q_3$,
$cum_{n_{i-1}}$ – liczebność skumulowana (kumulacyjna) poprzedzająca przedział klasowy $Q_1$ i $Q_3$,
$h_0$ – rozpiętość przedziału klasowego, w którym znajdują się $Q_1$ i $Q_3$,
$n_{Q_1}$ – liczebność przedziału klasowego $Q_1$,
$n_{Q_3}$ – liczebność przedziału klasowego $Q_3$.

**Decyle** dzielą zbiorowość na 10 równych części. **Decyl pierwszy ($D_1$)** dzieli uporządkowaną zbiorowość na dwie części w ten sposób, że: **10% jednostek zbiorowości ma wartości cechy niższe lub równe $D_1$ – (*nie większe od $D_1$*), a 90% jednostek ma wartości wyższe lub równe $D_1$ – (*nie mniejsze od $D_1$*).**




Źródło: opracowanie własne.

**Decyl dziewiąty ($D_9$)** dzieli uporządkowaną zbiorowość na dwie części w ten sposób, że: **90% jednostek zbiorowości ma wartości cechy niższe lub równe $D_9$ – (*nie większe od $D_9$*), a 10% jednostek ma wartości wyższe lub równe $D_9$ – (*nie mniejsze od $D_9$*).**




Źródło: opracowanie własne.

W szeregu rozdzielczym jednopunktowym obliczamy się numery jednostek analizowanej zbiorowości $Nr_{D_1}$ i $Nr_{D_9}$ według następujących wzorów:

- $N$ – parzyste: $$Nr_{D_1} = \frac{N}{10} \quad (2.1.23)$$
- $N$ – nieparzyste: $$Nr_{D_1} = \frac{N+1}{10} \quad (2.1.24)$$
- $N$ – parzyste: $$Nr_{D_9} = \frac{9N}{10} \quad (2.1.25)$$
- $N$ – nieparzyste: $$Nr_{D_9} = \frac{9(N+1)}{10} \quad (2.1.26)$$

Następnie obliczone numery jednostek zbiorowości przyporządkowuje się do przedziałów skumulowanych.

W szeregu rozdzielczym z przedziałami klasowymi stosuje się wzory:

$$D_1 = x_0 + \left(\frac{N}{10} - cum_{n_{i-1}}\right)\frac{h_0}{n_{D_1}} \quad (2.1.27)$$

$$D_9 = x_0 + \left(\frac{9N}{10} - cum_{n_{i-1}}\right)\frac{h_0}{n_{D_9}} \quad (2.1.28)$$

gdzie:

$x_0$ – dolna granica przedziału klasowego $D_1$ i $D_9$,
$cum_{n_{i-1}}$ – liczebność skumulowana (kumulacyjna) poprzedzająca przedział klasowy $D_1$ i $D_9$,
$h_0$ – rozpiętość przedziału klasowego, w którym znajdują się $D_1$ i $D_9$,
$n_{D_1}$ – liczebność przedziału klasowego $D_1$,
$n_{D_9}$ – liczebność przedziału klasowego $D_9$.

**Percentyle** dzielą zbiorowość na 100 równych części.

Pomiędzy miarami klasycznymi i pozycyjnymi tendencji centralnej występuje szereg różnic wynikających z konstrukcji i sposobu liczenia co ostatecznie przekłada się na interpretację (tab. 2.5.1).

**Tabela 2.5.1. Właściwości matematyczne miar tendencji centralnej**

| Parametr opisowy | Właściwości |
|---|---|
| **Średnia arytmetyczna ($\bar{x}$)** | **Zalety:**<br>– parametr łatwy do obliczenia, prosty w interpretacji,<br>– liczba zawsze mianowana, a więc jest wyrażona w takich samych jednostkach co badana cecha,<br>– dobrze oddaje kształtowanie się poziomu badanego zjawisku, kiedy zbiorowość jest jednorodna o słabym zróżnicowaniu.<br>**Wady:**<br>– na wynik mają wpływ wartości skrajne (ekstremalne), które mogą zniekształcić otrzymany wynik, przy czym silniejszy wpływ na wynik mają wysokie wartości cechy,<br>– średniej arytmetycznej nie należy stosować w przypadku rozkładów bimodalnych, wielomodalnych oraz silnie asymetrycznych.<br>**Dodatkowe:**<br>– średnia arytmetyczna jest wypadkową wszystkich wartości i znajduje się pomiędzy najmniejszą a największą wartością cechy,<br>– średnia arytmetyczna pomnożona przez ogólną liczebność daje sumę wszystkich jednostek zbiorowości,<br>– suma odchyleń poszczególnych wartości cechy od średniej arytmetycznej równa się zero,<br>– suma kwadratów odchyleń poszczególnych wartości cechy od średniej wynosi minimum. |
| **Dominanta ($D$)** | **Zalety:**<br>– dominanta jest miarą najbardziej zrozumiałą wśród miar przeciętnych,<br>– należy do miar tendencji centralnej,<br>– na jej wielkość nie maja wpływu wartości skrajne,<br>– do jej wyznaczenia wystarcza znajomość 3 przedziałów klasowych.<br>**Wady:**<br>– dominanty nie należy stosować, kiedy występuje skrajna asymetria i nierówne rozpiętości przedziałów klasowych,<br>– dokładne wyznaczenie modalnej nie jest możliwe w szeregach rozdzielczych wielostopniowych,<br>– znaczenie dominanty maleje, gdy liczba obserwacji jest mała,<br>– nie nadaje się do przekształceń algebraicznych. |
| **Mediana ($M_e$)** | **Zalety:**<br>– łatwa do obliczenia,<br>– niezależna od wartości krańcowych szeregu,<br>– można ją ustalić w szeregu otwartym,<br>– można ją obliczyć gdy szereg zbudowano na podstawie cechy jakościowej.<br>**Wady:**<br>– można ją wyznaczyć tylko z szeregu uporządkowanego,<br>– nie jest reprezentatywna dla szeregu bardzo nieregularnego,<br>– nie nadaje się do przekształceń algebraicznych. |

Źródło: opracowane na podstawie: W. Ignatczyk, M. Chromińska, *Statystyka. Teoria i zastosowanie*, WSB, Poznań 2004, s. 68-70, 83, 88. M. Sobczyk, *Statystyka*, PWN, Warszawa 216, s. 38-39. E., Dolny, Osińska M., *Statystyka opisowa*, WSG, Bydgoszcz 2009, s. 42, 47.

## 2.2. Miary zróżnicowania (zmienności)

Miary dyspersji służą do oceny wewnętrznego zróżnicowania zbiorowości pod względem badanej cechy. Najczęściej stosowanymi miarami zróżnicowania są:

a) miary klasyczne:
   – wariancja,
   – odchylenie standardowe,
   – klasyczny współczynnik zmienności,
b) miary pozycyjne:
   – odchylenie ćwiartkowe,
   – pozycyjny współczynnik zmienności.

Analizując parametry klasyczne, miarę zróżnicowania oblicza się poprzez wyliczenie różnic pomiędzy wartościami cechy od ich wartości centralnej – najczęściej średniej arytmetycznej. Miary klasyczne odnoszą się do zmiennych wyrażonych w skali interwałowej. W przypadku miar pozycyjnym takim parametrem centralnym zwykle jest mediana.

**Wariancja ($S_x^2$)** jest średnią arytmetyczną kwadratów odchyleń (różnic) poszczególnych wartości jednostek zbiorowości statystycznej od średniej arytmetycznej. Wariancja nie ma interpretacji statystycznej i jest wykorzystywana do budowy innych parametrów tj.: momentów centralnych, odchylenia standardowego. Zależnie od szeregu statystycznego wariancję oblicza się na podstawie wzorów:

- szereg prosty (szczegółowy):

$$S_x^2 = \frac{\sum_{i=1}^{N}(x_i - \bar{x})^2}{N} \tag{2.2.1}$$

- szereg rozdzielczy jednopunktowy:

$$S_x^2 = \frac{\sum_{i=1}^{k}(x_i - \bar{x})^2 n_i}{N} \tag{2.2.2}$$

- szereg rozdzielczy z przedziałami klasowymi:

$$S_x^2 = \frac{\sum_{i=1}^{k}(\dot{x}_i - \bar{x})^2 n_i}{N} \tag{2.2.3}$$

gdzie:

$x_i$ – warianty cechy mierzalnej *X*,
$\dot{x}_i$ – środek przedziału klasowego,
$\bar{x}$ – średnia arytmetyczna,
$n_i$ – wagi (liczebności cząstkowe) odpowiadające poszczególnym wariantom zmiennej *X*,
*N* – liczebność całej zbiorowości statystycznej.

**Odchylenie standardowe ($S_x$)** jest to pierwiastek drugiego stopnia z obliczonej wariancji – $S_x^2$ (wzór: 2.2.1 – 2.2.3). **Odchylenie standardowe informuje, o ile przeciętnie (średnio) wartości badanej zmiennej w zbiorowości statystycznej różnią się (odchylają się) in ± (*in plus lub in minus*) od średniej arytmetycznej.** Odchylenie standardowe oblicza się według wzorów:

- szereg prosty (szczegółowy):

$$S_x = \sqrt{\frac{\sum_{i=1}^{N}(x_i - \bar{x})^2}{N}} \tag{2.2.4}$$

- szereg rozdzielczy jednopunktowy:

$$S_x = \sqrt{\frac{\sum_{i=1}^{k}(x_i - \bar{x})^2 n_i}{N}} \tag{2.2.5}$$

- szereg rozdzielczy z przedziałami klasowymi:

$$S_x = \sqrt{\frac{\sum_{i=1}^{k}(\dot{x}_i - \bar{x})^2 n_i}{N}} \tag{2.2.6}$$

gdzie:

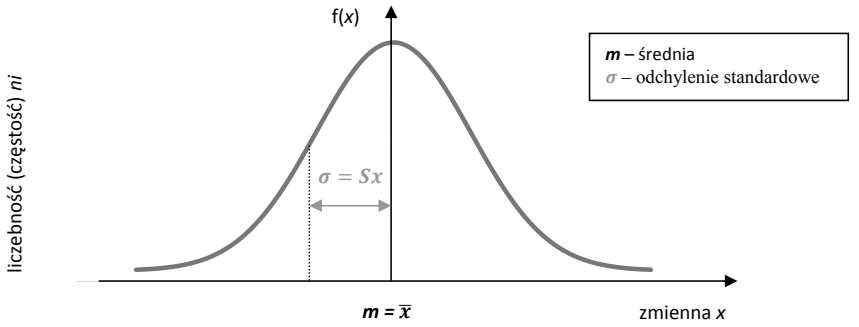
$x_i$ – warianty cechy mierzalnej $X$,
$\dot{x}_i$ – środek przedziału klasowego,
$\bar{x}$ – średnia arytmetyczna,
$n_i$ – wagi (liczebności cząstkowe) odpowiadające poszczególnym wariantom zmiennej $X$,
$N$ – liczebność całej zbiorowości statystycznej.

W tym miejscu należy wspomnieć o rozkładzie normalnym Gaussa (rys. 2.2.1). Średnia arytmetyczna i odchylenie standardowe są parametrami rozkładu normalnego i decydują o jego kształcie. Funkcja gęstości prawdopodobieństwa dla zmiennej losowej $X$, wyrażona jest wzorem:

$$f(x) = \frac{1}{\sigma\sqrt{2\pi}} exp\left(\frac{-(x-m)^2}{2\sigma^2}\right) \tag{2.2.7}$$

gdzie:

$x$ – zmienna losowa,
$m$ – wartość oczekiwana zmiennej losowej (średnia),
$\sigma$ – odchylenie standardowe,
$\pi = 3{,}1416$,
$exp = 2{,}1718$ – podstawa logarytmu naturalnego.




**Rysunek 2.2.1. Rozkład normalny**
Źródło: opracowanie własne.

Standaryzowany rozkład normalny z wartością oczekiwaną 0 i wariancją równą 1, tj. $N(0\,,1)$ możemy zapisać dla zmiennej standaryzowanej *u* w postaci:

$$f(u) = \frac{1}{\sqrt{2\pi}} exp\left(-\frac{u^2}{2}\right) \tag{2.2.8}$$

gdzie:

*u* – zmienna standaryzowana.

Dla obliczenia wartości gęstości zmiennej losowej X o rozkładzie $N(m\,,\sigma)$ stosuje się przekształcenie zwane **standaryzacją**:

$$u = \frac{X - m}{\sigma} \tag{2.2.9}$$

Z odchyleniem standardowym wiąże się tzw. reguła 3 sigm.

**Reguła 3 sigm**

a) około **68,3%** jednostek zbiorowości mieści się w granicach **1** odchylenia standardowego:

$$\bar{x} - S_x < X < \bar{x} - S_x = 0{,}6826,$$

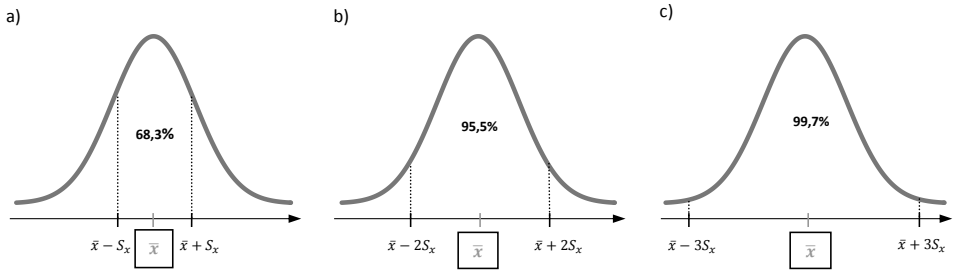
b) około **95,5%** jednostek zbiorowości mieści się w granicach **2** odchyleń standardowych:

$$\bar{x} - 2S_x < X < \bar{x} - 2S_x = 0{,}9545,$$

c) około **99,7%** jednostek zbiorowości mieści się w granicach **3** odchyleń standardowych:

$$\bar{x} - 3S_x < X < \bar{x} - 3S_x = 0{,}9973.$$

Źródło: Piłatowska M., *Repetytorium ze statystyki*, PWN, Warszawa 2016, s. 26.




**Rysunek 2.2.2. Reguła 3 sigm**
Źródło: opracowanie własne.

Wówczas typowy obszar zmienności $\boldsymbol{x_{typ}}$ będzie znajdował się w przedziale:

$$\bar{x} - S_x < \boldsymbol{x_{typ}} < \bar{x} + S_x \tag{2.2.10}$$

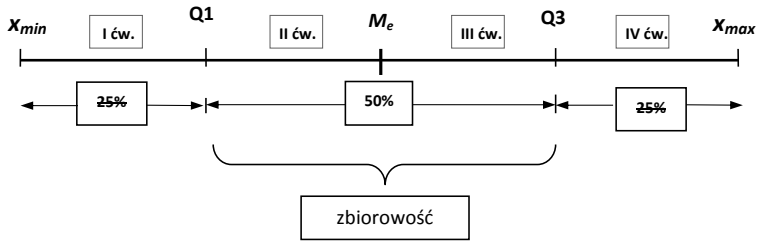
Może zdarzyć się sytuacja, kiedy obliczone odchylenie standardowe w szeregu rozdzielczym o równych przedziałach klasowych ($h$) może być obciążone błędem grupowania, wówczas stosuje się tzw. **poprawkę Shepparda**. Wzór na skorygowane odchylenie standardowe przyjmuje postać:

$$S_{x_{skor}} = \sqrt{S_x^2 - \frac{h^2}{12}} \tag{2.2.11}$$

gdzie:
$S_x$ – odchylenie standardowe.
$h$ – stała rozpiętość przedziałów klasowych.

Poprawkę Shepparda zaleca się w przypadku cech ciągłych, kiedy liczba klas jest mniejsza niż 12 ($k < 12$) a liczebność $N$ jest duża.

**Odchylenie ćwiartkowe ($Q_x$)** jest to połowa obszaru zmienności, w którym znajduje się 50% środkowych jednostek zbiorowości (druga i trzecia ćwiartka). Oznacza to, że 50% jednostek przyjmuje wartości pomiędzy $Q_1$ i $Q_3$ (pomijane jest 25% jednostek z wartościami najniższymi i 25% jednostek z wartościami najwyższymi cechy tzn. z pierwszej i czwartej ćwiartki). Jest to miara pozycyjna, tak więc do opisu tendencji centralnej zastosowano medianę. Odchylenie ćwiartkowe informuje o ile **wartości zmiennej badanych jednostek zbiorowości statystycznej różnią się od mediany** (ale dotyczy to 50% środkowych jednostek zbiorowości).




Źródło: opracowanie własne.

Różnica pomiędzy $Q_3 - Q_1$ nazywa się rozstępem międzykwartylowym. Wzór na odchylenie ćwiartkowe jest następujący:

$$Q_x = \frac{(M_e - Q_1) + (Q_3 - M_e)}{2} = \frac{Q_3 - Q_1}{2} \tag{2.2.12}$$

gdzie:
$Q_1$ – kwartyl pierwszy,
$Q_3$ – kwartyl trzeci,
$M_e$ – mediana.

Obserwacje typowe $\boldsymbol{x_{typ}}$ znajdują się w przedziale:

$$M_e - Q_x < \boldsymbol{x_{typ}} < M_e + Q_x \tag{2.2.13}$$

**Względna miary zróżnicowania (współczynnik zmienności)** stosowana jest do oceny stopnia intensywności zróżnicowania (zmienności) pomiędzy porównywanymi zbiorowościami statystycznymi pod względem badanej cechy**.** Współczynnik zmienności **informuje o procentowym udziale bezwzględnej miary zróżnicowania (odchylenia standardowego lub odchylenia ćwiartkowego) w wartości centralnej (średniej arytmetycznej lub medianie).** Współczynnik zmienności przyjmuje postać:

- **klasyczną ($\boldsymbol{V_x}$)**:

$$V_x = \frac{S_x}{\bar{x}} \cdot 100 \tag{2.2.14}$$

gdzie:
$S_x$ – odchylenie standardowe,
$\bar{x}$ – średnia arytmetyczna.

- **pozycyjną ($\boldsymbol{V_Q}$)**:

$$V_Q = \frac{Q_x}{M_e} \cdot 100 \tag{2.2.15}$$

gdzie:
$Q_x$ – odchylenie ćwiartkowe,
$M_e$ – mediana.

Przyjmuje się, że jeżeli:

- $V_{x/Q} \leq 35\% -$ mała zmienność,
- $35\% < V_{x/Q} \leq 60\% -$ umiarkowana zmienność,
- $60\% < V_{x/Q} \leq 75\% -$ duża zmienność,
- $75\% < V_{x/Q} \leq 100\% -$ bardzo duża zmienność.

Im wyższa wartość współczynnika zmienności tym jednostki statystyczne bardziej różnią się od siebie pod względem wartości cechy i odwrotnie, im niższa wartość współczynnika zmienności tym jednostki statystyczne są bardziej do siebie podobne (mniej różnią się pomiędzy sobą).

## 2.3. Miary asymetrii (skośności)

**Współczynnik asymetrii ($A_s$)** określa zarówno kierunek, jak i siłę asymetrii empirycznego rozkładu. Im wyższa wartość współczynnika asymetrii tym rozkład badanej cechy bardziej różni się od rozkładu symetrycznego.

**Klasyczny współczynnik asymetrii ($A_{s_x}$)** jest to iloraz trzeciego momentu centralnego przez sześcian odchylenia standardowego:

$$A_{s_x} = \frac{\mu_3}{S_x^3} \tag{2.3.1}$$

gdzie:

$\mu_3$ – trzeci moment centralny ($\mu = mi$),
$S_x$ – odchylenie standardowe.

Trzeci moment centralny oblicza się:

- szereg prosty (szczegółowy):

$$\mu_3 = \frac{\sum_{i=1}^{N}(x_i - \bar{x})^3}{N} \tag{2.3.2}$$

- szereg rozdzielczy jednopunktowy:

$$\mu_3 = \frac{\sum_{i=1}^{k}(x_i - \bar{x})^3 n_i}{N} \tag{2.3.3}$$

- szereg rozdzielczy z przedziałami klasowymi:

$$\mu_3 = \frac{\sum_{i=1}^{k}(\dot{x}_i - \bar{x})^3 n_i}{N} \tag{2.3.4}$$

gdzie:

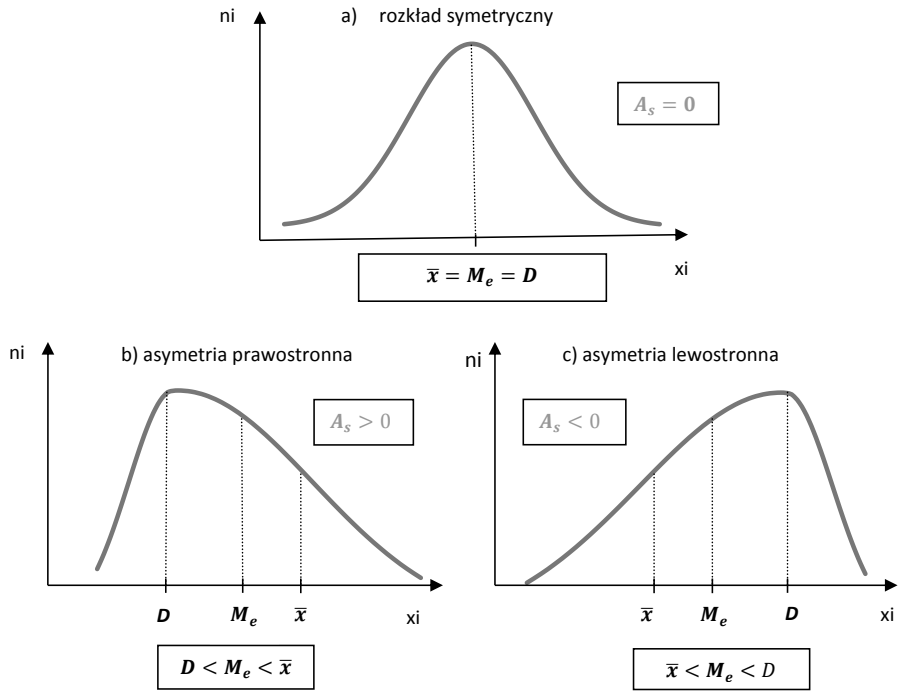
$x_i$ – warianty cechy mierzalnej *X*,
$\dot{x}_i$ – środek przedziału klasowego,
$\bar{x}$ – średnia arytmetyczna,
$n_i$ – wagi (liczebności cząstkowe) odpowiadające poszczególnym wariantom zmiennej *X*,
*N* – liczebność całej zbiorowości statystycznej.

Rozkłady cechy mogą być symetryczne lub asymetryczne (rys. 2.3.1):

a) $A_s = 0$ – **rozkład symetryczny**, tzn. połowa jednostek zbiorowości ma wartości cechy niższe od średniej a druga połowa jednostek ma wartości wyższe od średniej,

b) $A_s > 0$ – **rozkład asymetryczny o asymetrii prawostronnej (dodatnia)**, tzn. dominują jednostki w zbiorowości o wartościach cechy niższych od średniej arytmetycznej,

c) $A_s < 0$ – **rozkład asymetryczny o asymetrii lewostronnej (ujemna)**, tzn. dominują jednostki w zbiorowości o wartościach cechy wyższych od średniej arytmetycznej.




**Rysunek 2.3.1. Charakterystyka rozkładów**
Źródło: opracowanie własne.

Współczynnik asymetrii zwykle znajduje się w przedziale [-1; 1], rzadko przyjmuje wartość spoza przedziału [-2; 2].

**Pozycyjny współczynnik asymetrii ($A_{Q_x}$)** oblicza się według wzoru:

$$A_{Q_x} = \frac{Q_3 + Q_1 - 2M_e}{2Q_x} \tag{2.3.5}$$

gdzie:
$Q_1$ – kwartyl pierwszy,
$Q_3$ – kwartyl trzeci,
$M_e$ – mediana,
$Q_x$ – odchylenie ćwiartkowe.

Pozycyjny współczynnik asymetrii dotyczy jednostek znajdujących się pomiędzy pierwszym i trzecim kwartylem (tj. 50% środkowych jednostek).

**Klasyczno-pozycyjny współczynnik asymetrii:**

$$A_s = \frac{\bar{x} - D}{S_x} \tag{2.3.6}$$

gdzie:
$\bar{x}$ – średnia arytmetyczna,
$D$ – dominanta,
$S_x$ – odchylenie standardowe.

Najczęściej do określenia siły $A_s$ przyjmuje się następujące przedziały:

- $0 < |A_s| \leq 0{,}3$ – słaba,
- $0{,}3 < |A_s| \leq 0{,}6$ – umiarkowana,
- $|A_s| > 0{,}6$ – silna.

Warto podkreślić, że miary asymetrii $A_{s_x}, A_{Q_x}, A_{s_x}$ są nieporównywalne.

## 2.4. Miary koncentracji (skupienia)

**Miara koncentracji** określa stopień skupienia jednostek statystycznych pod względem wartości badanej cechy wokół średniej. Koncentrację nazywa się kurtozą lub ekscesem. Punktem odniesienia do identyfikacji rozkładu badanej cechy jest rozkład normalny.

**Klasyczny współczynnik skupienia ($W_k$)** oblicza się:

$$W_k = \frac{\mu_4}{S_x^4} \tag{2.4.1}$$

gdzie:

$\mu_4$ – czwarty moment centralny ($\mu = mi$),
$S_x$ – odchylenie standardowe.

Czwarty moment centralny oblicza się:

- szereg prosty (szczegółowy):

$$\mu_4 = \frac{\sum_{i=1}^{N}(x_i - \bar{x})^4}{N} \tag{2.4.2}$$

- szereg rozdzielczy jednopunktowy:

$$\mu_4 = \frac{\sum_{i=1}^{k}(x_i - \bar{x})^4 n_i}{N} \tag{2.4.3}$$

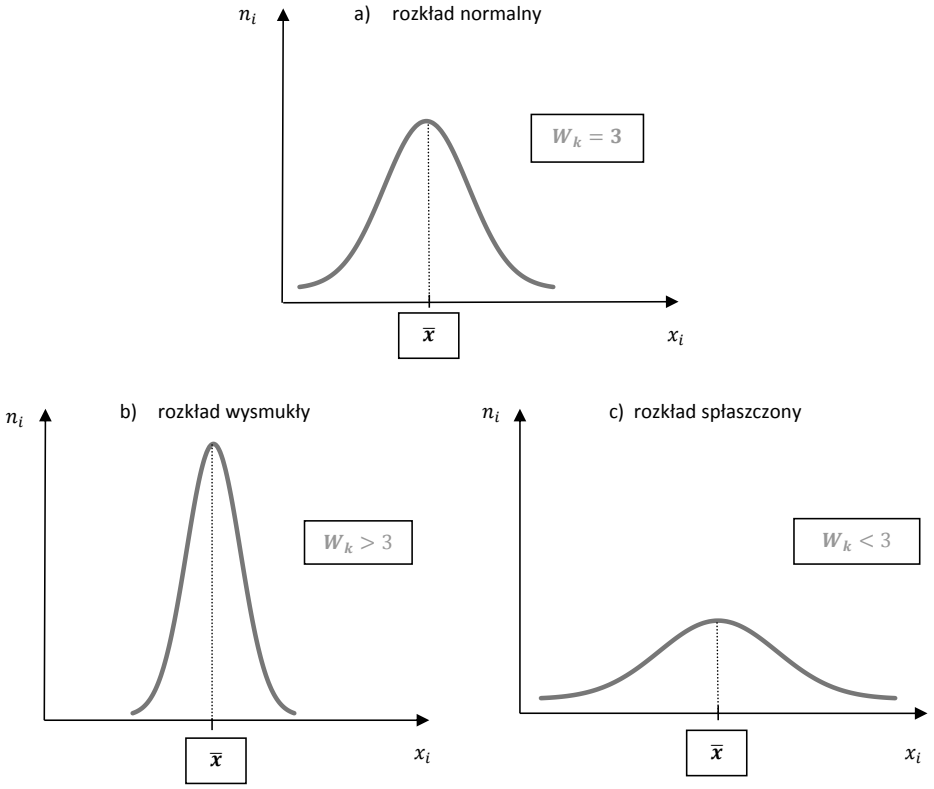
- szereg rozdzielczy z przedziałami klasowymi:

$$\mu_4 = \frac{\sum_{i=1}^{k}(\dot{x}_i - \bar{x})^4 n_i}{N} \tag{2.4.4}$$

gdzie:

$x_i$ – warianty cechy mierzalnej *X*,
$\dot{x}_i$ – środek przedziału klasowego,
$\bar{x}$ – średnia arytmetyczna,
$n_i$ – wagi (liczebności cząstkowe) odpowiadające poszczególnym wariantom zmiennej *X*,
*N* – liczebność całej zbiorowości statystycznej.

Jeżeli **współczynnik koncentracji** spełnia relację (rys. 2.3.2):




**Rysunek 2.3.2. Rozkłady: normalny, wysmukły i spłaszczony**
Źródło: opracowanie własne.

a) $W_k = 3$ rozkład jest **normalny (mezokurtyczny)**, tzn. koncentracja jednostek zbiorowości pod względem badanej cechy jest jak w rozkładzie normalnym,

b) $W_k > 3$ rozkład jest **wysmukły** (**leptokurtyczny**), tzn. koncentracja jednostek zbiorowości pod względem badanej cechy wokół średniej jest silniejsza od rozkładu normalnego,

c) $W_k < 3$ rozkład jest **spłaszczony** (**platokurtyczny**), tzn. koncentracja jednostek zbiorowości pod względem badanej cechy wokół średniej jest słabsza od rozkładu normalnego.

**Pozycyjny współczynnik skupienia ($W_{sk}$)** oblicza się według wzoru:

$$W_{sk} = \frac{D_9 - D_1}{Q_3 - Q_1} \tag{2.4.5}$$

gdzie:

$Q_1$ – kwartyl pierwszy,
$Q_3$ – kwartyl trzeci,
$D_1$ – decyl pierwszy,
$D_9$ – decyl dziewiąty,

Wzór (2.4.5) stosuje się wyłącznie, kiedy mamy do czynienia z rozkładem symetrycznym lub co najwyżej słabo asymetrycznym. Jeżeli **współczynnik koncentracji ($W_{sk}$)** spełnia relację:

a) $W_{sk} = 2$ to rozkład jest **normalny,**

b) $W_{sk} > 2$ to rozkład jest **wysmukły,**

c) $W_{sk} < 2$ to rozkład jest **spłaszczony.**

**Współczynnik koncentracji Lorenza ($W_{KL}$)** jest to udział pola powierzchni ($P_a$) zawartego pomiędzy linią równomiernego rozdziału a krzywą koncentracji do pola powierzchni trójkąta *OBC* lub $P_a + P_b$ (rys. 2.3.3):

$$W_{KL} = \frac{P_a}{P_a + P_b} = \frac{5000 - P_b}{5000} \tag{2.4.6}$$

Pole $P_b$ oblicza się jako sumę pole trójkąta i pól trapezów:

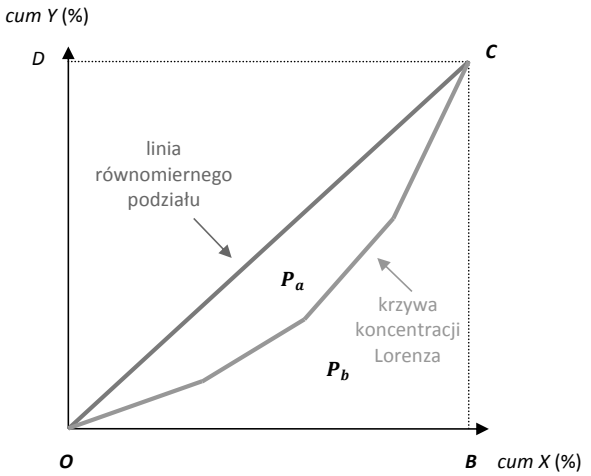
$$P_b = P_\Delta + \sum_{i=2}^{n} P_{trap.} \tag{2.4.7}$$

Bok powstałego kwadratu *OBCD* wynosi 100, a więc pole równe jest 10000. Połowa pola kwadratu stanowi pole trójkąta *OBC*, które wynosi 5000.

Współczynnik koncentracji Lorenza mieści się w przedziale:

$$0 \leq W_{KL} \leq 1.$$

Jeżeli $W_{KL}$= 0 to koncentracja nie występuje, kiedy $W_{KL}$= 1 mamy do czynienia z koncentracją absolutną (zupełną).




**Rysunek 2.3.3. Krzywa koncentracji Lorenza**
Źródło: opracowane na podstawie, *Podstawy Statystyki*, red. naukowa W. Starzyńska, Difin, Warszawa 2015, s. 150.

**Współczynnik Giniego ($W_G$)** określa rozkład badanej cechy ze względu na liczebność zbiorowości:

$$W_G = \sum_{i=1}^{k} \left( \left( cum \frac{n_{i-1}}{N} \right) + \left( cum \frac{n_i}{N} \right) \right) \cdot \frac{x_i n_i}{\sum_{i=1}^{k} x_i \, n_i} - 1 = \sum_{i=1}^{k} z_i - 1 \tag{2.4.8}$$

Współczynnik Giniego przyjmuje wartości z przedziału [0; 1]. Im wyższy poziom współczynnika Giniego tym koncentracja (stopień nierówności) jest większa.

## Przykłady

**Przykład 2.1.**

Wśród pracowników firmy *Z* zatrudniającej 5 pracowników zebrano informacje o wysokości wypłaconych wynagrodzeń brutto za czerwiec, które były następujące:

2601,28, 2705,32, 2705,32, 3091,23, 3136,62 [zł]

Przeprowadź kompleksową analizę wynagrodzeń z wykorzystaniem miar klasycznych i pozycyjnych a następnie zinterpretuj otrzymane parametry.

**Rozwiązanie**

**Typ szeregu:** szereg prosty (szczegółowy).

**Zmienna badana $x_i$:** wynagrodzenia.

**I. Analiza wynagrodzeń z wykorzystaniem miar klasycznych**

1. **Średnia arytmetyczna (wzór 2.1.1):**

$$\bar{x} = \frac{\sum_{i=1}^{N} x_i}{N} = \frac{2601{,}28 + 2705{,}32 \; + 2705{,}32 + \; 3091{,}23 + 3136{,}62}{5} = 2847{,}95 \text{ zł}$$

**Interpretacja:** Wśród pracowników firmy *Z* średnie wypłacone wynagrodzenia za czerwiec wynosiło 2847,95 zł.

2. **Wariancja (wzór 2.2.1):**

Obliczenia pomocnicze:

| Wynagrodzenia (zł) $x_i$ | $x_i - \bar{x}$ | $(x_i - \bar{x})^2$ |
|---|---|---|
| 2601,28 | -246,67 | 60846,09 |
| 2705,32 | -142,63 | 20343,32 |
| 2705,32 | -142,63 | 20343,32 |
| 3091,23 | 243,28 | 59185,16 |
| 3136,62 | 288,67 | 83330,37 |
| **Ogółem** | - | **244048,26** |

Źródło: opracowanie własne.

$$S_x^2 = \frac{\sum_{i=1}^{N}(x_i - \bar{x})^2}{N} = \frac{244048{,}26}{5} = 48809{,}65$$

3. **Odchylenie standardowe (wzór 2.2.4):**

$$S_x = \sqrt{\frac{\sum_{i=1}^{N}(x_i - \bar{x})^2}{N}} = \sqrt{\frac{244048{,}26}{5}} = 220{,}93 \text{ zł}$$

**Interpretacja:** Wynagrodzenia wypłacone pracownikom w czerwcu różnią się (odchylają się) od średniej arytmetycznej ($\bar{x} = 2847{,}95$ zł.) przeciętnie o ± 220,93 zł.

4. **Typowy obszar zmienności (wzór 2.2.10):**

$$\bar{x} - S_x < \boldsymbol{x_{typ}} < \bar{x} + S_x$$
$$2847{,}95 - 220{,}93 < \boldsymbol{x_{typ}} < 2847{,}95 + 220{,}93$$
$$2627{,}02 \text{ zł} < \boldsymbol{x_{typ}} < 3068{,}88 \text{ zł}$$

**Interpretacja:** W obliczonym przedziale wynagrodzeń (2627,02-3068,88 zł) znalazło się około 68% pracowników.

5. **Klasyczny współczynnik zmienności (wzór 2.2.14):**

$$V_x = \frac{S_x}{\bar{x}} \cdot 100 = \frac{220{,}93}{2847{,}95} \cdot 100 = 7{,}8\%$$
$$V_x = 7{,}8\% \leq 35\% - \text{mała zmienność},$$

**Interpretacja:** Zbiorowość pracowników w firmie *Z* charakteryzuje się małym zróżnicowaniem (zmiennością) pod względem wypłaconych wynagrodzeń.

**6. Miara asymetrii (wzory 2.3.2 i 2.3.1):**

Obliczenia pomocnicze:

| $x_i - \bar{x}$ | $(x_i - \bar{x})^3$ |
|---|---|
| -246,67 | -15008904,75 |
| -142,63 | -2901567,29 |
| -142,63 | -2901567,29 |
| 243,28 | 14398565,34 |
| 288,67 | 24054977,59 |
| - | **Suma = 17641503,60** |

Źródło: opracowanie własne.

$$\mu_3 = \frac{\sum_{i=1}^{N}(x_i - \bar{x})^3}{N} = \frac{17641503,60}{5} = 3528300,72$$

$$A_{S_x} = \frac{\mu_3}{S_x^3} = \frac{3528300,72}{220,93^3} = +0,33$$

$A_{S_x} > 0$ – rozkład asymetryczny o asymetrii prawostronnej

$0,3 < |0,33| \leq 0,6$ – umiarkowana asymetria

**Interpretacja:** Rozkład wynagrodzeń wskazuje na umiarkowaną asymetrię prawostronną. Oznacza to, że w badanej zbiorowości dominują pracownicy z wynagrodzeniami niższymi od średniej arytmetycznej ($\bar{x} = 2847,95$ zł.).

**7. Miara koncentracji (wzory 2.4.2 i 2.4.1):**

Obliczenia pomocnicze:

| $x_i - \bar{x}$ | $(x_i - \bar{x})^4$ |
|---|---|
| -246,67 | 3702246534,43 |
| -142,63 | 413850542,49 |
| -142,63 | 413850542,49 |
| 243,28 | 3502882974,83 |
| 288,67 | 6943950381,01 |
| - | **Suma = 14976780975,25** |

Źródło: opracowanie własne.

$$\mu_4 = \frac{\sum_{i=1}^{N}(x_i - \bar{x})^4}{N} = \frac{14976780975,25}{5} = 2995356195,05$$

$$W_k = \frac{\mu_4}{S_x^4} = \frac{2995356195,05}{220,93^4} = 1,26$$

$W_k < 3$

$1,26 < 3$ – rozkład jest spłaszczony (platokurtyczny)

**Interpretacja:** Rozkład wynagrodzeń jest spłaszczony. Oznacza to, że koncentracja zbiorowości pracowników pod względem wynagrodzeń wokół średniej arytmetycznej jest słabsze od rozkładu normalnego.

## II. Analiza wynagrodzeń z wykorzystaniem miar pozycyjnych

### 1. Dominanta:

$$\underbrace{2601{,}28,\ 2705{,}32,\ 2705{,}32}\ ,\ 3091{,}23,\ 3136{,}62\ \text{[zł]}$$

$$D = 2705{,}32\ \text{zł}$$

**Interpretacja:** Wśród badanej zbiorowości dominują pracownicy z wynagrodzeniem 2705,32 zł.

### 2. Mediana (wzór 2.1.12):

*N=5* – szereg nieparzysty

$$M_e = x_{\frac{N+1}{2}} = x_{\frac{5+1}{2}} = x_{\frac{6}{2}} = x_3$$

$$\underset{x_1}{2601{,}28},\ \underset{x_2}{2705{,}32},\ \underbrace{\underset{x_3}{2705{,}32}}\ ,\ \underset{x_4}{3091{,}23},\ \underset{x_5}{3136{,}62}\ \text{[zł]}$$

$$M_e = x_3 = 2705{,}32\ \text{zł}$$

**Interpretacja:** Mediana informuje, że 50% pracowników otrzymała wynagrodzenie 2705,32 zł i niższe, a druga połowa pracowników otrzymała wynagrodzenie 2705,32 zł i wyższe.

### 3. Kwartyle:

*N=5* – szereg nieparzysty

$$M_e = x_{\frac{N+1}{2}} = x_{\frac{5+1}{2}} = x_{\frac{6}{2}} = x_3$$

$$\underset{x_1}{2601{,}28},\ \underset{x_2}{2705{,}32},\ \underbrace{\underset{x_3}{2705{,}32}}\ ,\ \underset{x_4}{3091{,}23},\ \underset{x_5}{3136{,}62}\ \text{[zł]}$$

$$M_e$$

- **kwartyl 1 ($Q_1$):**

*N=3* – szereg nieparzysty

$$Q_1 = x_{\frac{N+1}{2}} = x_{\frac{3+1}{2}} = x_{\frac{4}{2}} = x_2$$

$$\overset{x_1}{2601{,}28},\ \underbrace{\overset{x_2}{2705{,}32},\ \overset{x_3}{2705{,}32}}\ \text{[zł]}$$

$$Q_1 = x_2 = 2705{,}32 \text{ zł}$$

**Interpretacja:** Kwartyl 1 informuje, że 25% pracowników otrzymała wynagrodzenie 2705,32 zł i niższe, a 75% pracowników otrzymała wynagrodzenie 2705,32 zł i wyższe.

- **kwartyl 3 ($Q_3$):**

*N=3* – szereg nieparzysty

$$Q_1 = x_{\frac{N+1}{2}} = x_{\frac{3+1}{2}} = x_{\frac{4}{2}} = x_2 = x_4$$

$$\overset{x_3}{2705{,}32},\ \underbrace{\overset{x_4}{3091{,}23},\ \overset{x_5}{3136{,}62}}\ \text{[zł]}$$

$$Q_3 = x_4 = 3091{,}23 \text{ zł}$$

**Interpretacja:** Kwartyl 3 informuje, że 75% pracowników otrzymała wynagrodzenie 3091,23 zł i niższe, a 25% pracowników otrzymała wynagrodzenie 3091,23 zł i wyższe.

4. **Odchylenie ćwiartkowe (wzór 2.2.12):**

$$Q_x = \frac{Q_3 - Q_1}{2} = \frac{3091{,}23 - 2705{,}32}{2} = 192{,}96 \text{ zł}$$

**Interpretacja:** Wynagrodzenia wypłacone pracownikom w czerwcu różnią się (odchylają się) od mediany ($M_e = 2705{,}32$ zł) przeciętnie o ± 192,96 zł. (ale dotyczy to tylko 50% pracowników znajdujących się w drugiej i trzeciej ćwiartce).

5. **Typowy obszar zmienności (wzór 2.2.13):**

$$M_e - Q_x < \boldsymbol{x_{typ}} < M_e + Q_x$$

$$2705{,}32 - 192{,}96 < \boldsymbol{x_{typ}} < 2705{,}32 + 192{,}96$$

$$2512{,}36 \text{ zł} < \boldsymbol{x_{typ}} < 2898{,}28 \text{ zł}$$

6. **Współczynnik zmienności (wzór 2.2.15):**

$$V_Q = \frac{Q_x}{M_e} \cdot 100 = \frac{192{,}96}{2705{,}32} \cdot 100 = 7{,}1\%$$

$V_Q = 7{,}1\% \leq 35\%$ – mała zmienność,

**Interpretacja:** Zbiorowość pracowników w firmie *Z* charakteryzuje się małym zróżnicowaniem (zmiennością) pod względem wypłaconych wynagrodzeń.

7. **Miara asymetrii (wzór 2.3.5):**

$$A_{Q_x} = \frac{Q_3 + Q_1 - 2M_e}{2Q_x} = \frac{3091{,}23 + 2705{,}32 - 2 \cdot 2705{,}32}{2 \cdot 192{,}96} = 0{,}999 \approx +1{,}0$$

$A_s > 0$ – rozkład asymetryczny o asymetrii prawostronnej

$1{,}0 > 0$

**Interpretacja:** Rozkład wynagrodzeń wskazuje na silną asymetrię prawostronną w obszarze 50% środkowych pracowników tzn. pomiędzy 1 i 3 kwartylem.

Dodatkowo obliczymy klasyczno-pozycyjny współczynnik asymetrii (wzór 2.3.6):

$$A_s = \frac{\bar{x} - D}{S_x} = \frac{2847{,}95 - 2705{,}32}{220{,}93} = +0{,}65$$

$A_s > 0$ – rozkład asymetryczny o asymetrii prawostronnej

$|0{,}65| > 0{,}6$ – silna asymetria

**Zestawienie podstawowych statystyk opisowych wynagrodzeń pracowników w firmie Z**

| Nr | Miary | Wartość |
|---|---|---|
| | **klasyczne** | |
| 1. | Średnia arytmetyczna – $\bar{x}$ | 2847,95 |
| 2. | Wariancja – $S_x^2$ | 48809,65 |
| 3. | Odchylenie standardowe – $S_x$ | 220,93 |
| 4. | Typowy obszar zmienności | 2627,02 < $x_{typ}$ < 3068,88 |
| 5. | Współczynnik zmienności – $V_x$ | 7,8% |
| 6. | Miara asymetrii – $A_{S_x}$ | 0,33 |
| 7. | Miara koncentracji – $W_k$ | 1,26 |
| | **pozycyjne** | |
| 1. | Dominanta – $D$ | 2705,32 |
| 2. | Mediana – $M_e$ | 2705,32 |
| 3. | Kwartyl 1 – $Q_1$ | 2705,32 |
| | Kwartyl 3 – $Q_3$ | 3091,23 |
| 5. | Odchylenie ćwiartkowe – $Q_x$ | 192,96 |
| 6. | Typowy obszar zmienności | 2512,36 < $x_{typ}$ < 2898,28 |
| 7. | Współczynnik zmienności – $V_Q$ | 7,1% |
| 8. | Miara asymetrii – $A_{Q_x}$ | 1,0 |
| | Miara asymetrii (klasyczno-pozycyjna) – $A_s$ | 0,65 |

Źródło: opracowanie własne.

**Przykład 2.2.**

Wśród studentów I roku w grupie pracującej na kierunku Zarządzanie PUZ we Włocławku przeprowadzono obserwację ze względu na otrzymane oceny z zaliczenia ze statystyki. Rezultaty obserwacji pogrupowano w szereg:

| Oceny ze statystyki $x_i$ | Liczba studentów $n_i$ |
|---|---|
| 2 | 4 |
| 3 | 8 |
| 3,5 | 11 |
| 4 | 6 |
| 4,5 | 3 |
| 5 | 2 |
| **Ogółem** | $N$ **= 34** |

Źródło: Dane umowne.

Przeprowadź kompleksową analizę ocen studentów z wykorzystaniem miar klasycznych i pozycyjnych a następnie zinterpretuj otrzymane parametry.

**Rozwiązanie**

**Typ szeregu:** szereg rozdzielczy jednopunktowy.

**Zmienna badana $x_i$:** oceny z zaliczenia ze statystyki.

## I. Analiza ocen ze statystyki z wykorzystaniem miar klasycznych

Obliczenia pomocnicze:

| Oceny ze statystyki $x_i$ | Liczba studentów $n_i$ | $x_i \cdot n_i$ | $x_i - \bar{x}$ | $(x_i - \bar{x})^2$ | $(x_i - \bar{x})^2 n_i$ |
|---|---|---|---|---|---|
| 2 | 4 | 8 | -1,47 | 2,161 | 8,644 |
| 3 | 8 | 24 | -0,47 | 0,221 | 1,768 |
| 3,5 | 11 | 38,5 | 0,03 | 0,001 | 0,011 |
| 4 | 6 | 24 | 0,53 | 0,281 | 1,686 |
| 4,5 | 3 | 13,5 | 1,03 | 1,061 | 3,183 |
| 5 | 2 | 10 | 1,53 | 2,341 | 4,682 |
| **Ogółem** | ***N* = 34** | **118** | - | - | **19,974** |

Źródło: opracowanie własne.

1. **Średnia arytmetyczna (wzór 2.1.2):**

$$\bar{x} = \frac{\sum_{i=1}^{k} x_i n_i}{N} = \frac{118}{34} = 3{,}47$$

**Interpretacja:** Średnia ocen z zaliczenia ze statystyki wśród studentów wynosi 3,47.

2. **Wariancja (wzór 2.2.2):**

$$S_x^2 = \frac{\sum_{i=1}^{k} (x_i - \bar{x})^2 n_i}{N} = \frac{19{,}974}{34} = 0{,}59$$

3. **Odchylenie standardowe (wzór 2.2.5):**

$$S_x = \sqrt{\frac{\sum_{i=1}^{k} (x_i - \bar{x})^2 n_i}{N}} = \sqrt{\frac{19{,}974}{34}} = 0{,}77$$

**Interpretacja:** Oceny z zaliczenia ze statystki różnią się (odchylają się) od średniej arytmetycznej ($\bar{x} = 3{,}47$) przeciętnie o ± 0,77 oceny.

4. **Typowy obszar zmienności (wzór 2.2.10):**

$$\bar{x} - S_x < \boldsymbol{x_{typ}} < \bar{x} + S_x$$

$$3{,}47 - 0{,}77 < \boldsymbol{x_{typ}} < 3{,}47 + 0{,}77$$
$$2{,}7 < \boldsymbol{x_{typ}} < 4{,}24$$

**Interpretacja:** W obliczonym przedziale ocen (2,7-4,24) znajduje się około 68% studentów.

5. **Klasyczny współczynnik zmienności (wzór 2.2.14):**

$$V_x = \frac{S_x}{\bar{x}} \cdot 100 = \frac{0{,}77}{3{,}47} \cdot 100 = 22{,}2\%$$
$$V_x = 22{,}2\% \leq 35\% - \text{mała zmienność},$$

**Interpretacja:** Zbiorowość studentów charakteryzuje się małym zróżnicowaniem (zmiennością) pod względem otrzymanych ocen ze statystyki.

6. **Miara asymetrii (wzory 2.3.3 i 2.3.1):**

Obliczenia pomocnicze:

| $x_i - \bar{x}$ | $(x_i - \bar{x})^3$ | $(x_i - \bar{x})^3 n_i$ |
|---|---|---|
| -1,47 | -3,17652 | -12,70608 |
| -0,47 | -0,10382 | -0,83056 |
| 0,03 | 0,00003 | 0,00033 |
| 0,53 | 0,14888 | 0,89328 |
| 1,03 | 1,09273 | 3,27819 |
| 1,53 | 3,58158 | 7,16316 |
| - | - | **Suma = - 2,20168** |

Źródło: opracowanie własne.

$$\mu_3 = \frac{\sum_{i=1}^{k}(x_i - \bar{x})^3 n_i}{N} = \frac{-2{,}20168}{34} = -0{,}06$$

$$A_{S_x} = \frac{\mu_3}{S_x^3} = \frac{-0{,}06}{0{,}77^3} = -0{,}13$$

$A_{S_x} < 0$ – rozkład asymetryczny o asymetrii lewostronnej

$0 < |-0{,}13| \leq 0{,}3$ – słaba siła asymetrii

**Interpretacja:** Rozkład ocen ze statystyki wskazuje na słabą asymetrię lewostronną. Oznacza to, że w zbiorowości dominują studenci z wyższymi ocenami od średniej arytmetycznej ($\bar{x} = 3{,}47$).

**7. Miara koncentracji (wzory 2.4.3 i 2.4.1):**

Obliczenia pomocnicze:

| $x_i - \bar{x}$ | $(x_i - \bar{x})^4$ | $(x_i - \bar{x})^4 n_i$ |
|---|---|---|
| -1,47 | 4,669489 | 18,67796 |
| -0,47 | 0,048797 | 0,39038 |
| 0,03 | 0,000001 | 0,00001 |
| 0,53 | 0,078905 | 0,47343 |
| 1,03 | 1,125509 | 3,37653 |
| 1,53 | 5,479813 | 10,95963 |
| - | - | **Suma = 33,87794** |

Źródło: opracowanie własne.

$$\mu_4 = \frac{\sum_{i=1}^{k}(x_i - \bar{x})^4 n_i}{N} = \frac{33{,}87794}{34} = 0{,}996 \approx 1{,}0$$

$$W_k = \frac{\mu_4}{S_x^4} = \frac{1{,}0}{0{,}77^4} = 2{,}8$$

$$W_k < 3$$

$2{,}8 < 3$ – rozkład jest spłaszczony (platokurtyczny)

**Interpretacja:** Rozkład ocen ze statystyki jest spłaszczony. Oznacza to, że koncentracja zbiorowości studentów pod względem ocen ze statystyki wokół średniej arytmetycznej jest słabsza od rozkładu normalnego.

## II. Analiza ocen ze statystyki z wykorzystaniem miar pozycyjnych

**1. Dominanta:**

Dominanta ***D* = 3,5** → | Wartość max = 11 ←

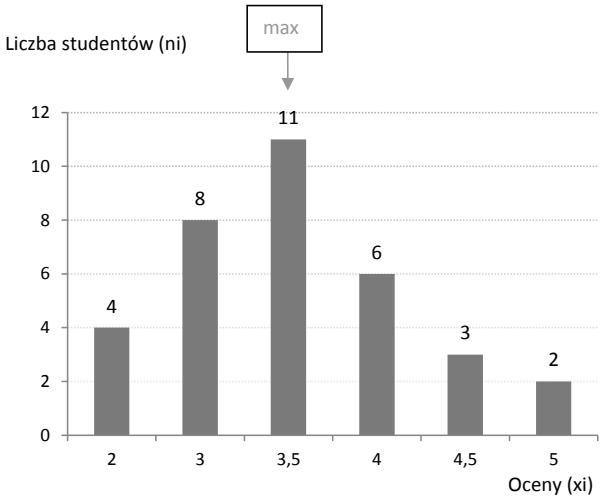
| Oceny ze statystyki $x_i$ | Liczba studentów $n_i$ |
|---|---|
| 2 | 4 |
| 3 | 8 |
| **3,5 *D*** | **11** |
| 4 | 6 |
| 4,5 | 3 |
| 5 | 2 |
| **Ogółem** | ***N* = 34** |

Źródło: opracowanie własne.

$$D = 3{,}5$$

**Interpretacja:** Wśród badanej zbiorowości dominują studenci z oceną 3,5 ($n_i = 11$).

**Rozkład ocen ze statystyki wśród studentów**




Źródło: opracowanie własne.

**2. Mediana:**

Obliczamy liczebności skumulowane *cum*:

| Liczba studentów $n_i$ | Wartości skumulowane *cum* | Przedziały |
|---|---|---|
| 4 | 4 | $n_1 - n_4$ |
| 8 | 4 + 8 = 12 | $n_5 - n_{12}$ |
| 11 | 4 + 8 + 11 = 23 | $n_{13} - n_{23}$ |
| 6 | 4 + 8 + 11 + 6 = 29 | $n_{24} - n_{29}$ |
| 3 | 4 + 8 + 11 + 6 + 3 = 32 | $n_{30} - n_{32}$ |
| 2 | 4 + 8 + 11 + 6 + 3 + 2 = 34 | $n_{33} - n_{34}$ |

Źródło: opracowanie własne.

Jeżeli każdego studenta ponumerujemy od 1 do 34 to otrzymamy:

$$n_1, n_2, n_3, n_4, n_5, n_6 \ldots, n_{34}$$

Wartości skumulowane możemy zapisać w postać przedziałów np. wartość skumulowaną 4 możemy przedstawić jako przedział: $n_1 - n_4$. Oznacza to, że w tym przedziale znajduje się 4 studentów o numerach porządkowych: $n_1$, $n_2$, $n_3$ i $n_4$.

| Oceny ze statystyki $x_i$ | Liczba studentów $n_i$ | Wartości skumulowane *cum* |
|---|---|---|
| 2 | 4 | 4 $n_1 - n_4$ |
| 3 | 8 | 12 $n_5 - n_{12}$ |
| **3,5 $M_e$** | **11** | **23 $n_{13} - n_{23}$** |
| 4 | 6 | 29 $n_{24} - n_{29}$ |
| 4,5 | 3 | 32 $n_{30} - n_{32}$ |
| 5 | 2 | 34 $n_{33} - n_{34}$ |
| **Ogółem** | ***N* = 34** | **-** |

Mediana $\boldsymbol{M_e = 3,5}$

$n_{17}$ student znajduje się w tym przedziale

Źródło: opracowanie własne.

$N = 34$ – szereg parzysty i obliczamy numer mediany (wzór 2.1.14):

$$Nr_{M_e} = \frac{N}{2} = \frac{34}{2} = \mathbf{17} = \boldsymbol{n_{17}}$$

$n_{17}$ – oznacza 17-go studenta

Student $n_{17}$ mieści się w przedziale $n_{13} - n_{23}$. Tak, więc mediana wynosi 3,5.

$$M_e = 3{,}5$$

**Interpretacja:** Mediana informuje, że 50% studentów otrzymała ocenę 3,5 i niższą, a druga połowa studentów otrzymała ocenę 3,5 i wyższą.

### 3. Kwartyle:

| Oceny ze statystyki $x_i$ | Liczba studentów $n_i$ | Wartości skumulowane *cum* |
|---|---|---|
| 2 | 4 | 4 $n_1 - n_4$ |
| **3 $Q_1$** | **8** | **12 $n_5 - n_{12}$** |
| 3,5 | 11 | 23 $n_{13} - n_{23}$ |
| 4 | 6 | 29 $n_{24} - n_{29}$ |
| 4,5 | 3 | 32 $n_{30} - n_{32}$ |
| 5 | 2 | 34 $n_{33} - n_{34}$ |
| **Ogółem** | ***N* = 34** | **-** |

Kwartyl 1 $\boldsymbol{Q_1 = 3,0}$

$n_9$ student znajduje się w tym przedziale

Źródło: opracowanie własne.

$N = 34$ – szereg parzysty i obliczamy numer kwartyla 1 (wzór 2.1.17):

$$Nr_{Q_1} = \frac{N}{4} = \frac{34}{4} = \mathbf{8,5} \approx \mathbf{9} = \boldsymbol{n_9}$$

$n_9$ – oznacza 9-go studenta

Student $n_9$ mieści się w przedziale $n_5 - n_{12}$. Tak, więc kwartyl 1 wynosi 3.

$$Q_1 = 3{,}0$$

**Interpretacja:** Kwartyl 1 informuje, że 25% studentów otrzymała ocenę 3,0 i niższą, a 75% studentów otrzymała ocenę 3,0 i wyższą.

| Oceny ze statystyki $x_i$ | Liczba studentów $n_i$ | Wartości skumulowane *cum* |
|---|---|---|
| 2 | 4 | 4 $n_1 - n_4$ |
| 3 | 8 | 12 $n_5 - n_{12}$ |
| 3,5 | 11 | 23 $n_{13} - n_{23}$ |
| **4 $Q_3$** | **6** | **29 $n_{24} - n_{29}$** |
| 4,5 | 3 | 32 $n_{30} - n_{32}$ |
| 5 | 2 | 34 $n_{33} - n_{34}$ |
| **Ogółem** | ***N* = 34** | **-** |

Kwartyl 3 $\boldsymbol{Q_3 = 4,0}$ → 4 $Q_3$

$n_{26}$ student znajduje się w tym przedziale → 29 $n_{24} - n_{29}$

*N* = 34 – szereg parzysty i obliczamy numer kwartyla 3 (wzór 2.1.19):

$$Nr_{Q_3} = \frac{3N}{4} = \frac{3 \cdot 34}{4} = \mathbf{25,5 \approx 26 = n_{26}}$$

$n_{26}$ – oznacza 26-go studenta

Student $n_{26}$ mieści się w przedziale $n_{24} - n_{29}$. Tak, więc kwartyl 3 wynosi 4.

$$Q_3 = 4{,}0$$

**Interpretacja:** Kwartyl 3 informuje, że 75% studentów otrzymała ocenę 4,0 i niższą, a 25% studentów otrzymała ocenę 4,0 i wyższą.

**4. Decyle:**

| Oceny ze statystyki $x_i$ | Liczba studentów $n_i$ | Wartości skumulowane *cum* |
|---|---|---|
| **2 $D_1$** | **4** | **4 $n_1 - n_4$** |
| 3 | 8 | 12 $n_5 - n_{12}$ |
| 3,5 | 11 | 23 $n_{13} - n_{23}$ |
| 4 | 6 | 29 $n_{24} - n_{29}$ |
| 4,5 | 3 | 32 $n_{30} - n_{32}$ |
| 5 | 2 | 34 $n_{33} - n_{34}$ |
| **Ogółem** | ***N* = 34** | **-** |

Decyl 1 $\boldsymbol{D_1 = 2,0}$ → 2 $D_1$

$n_3$ student znajduje się w tym przedziale → 4 $n_1 - n_4$

Źródło: opracowanie własne.

*N* = 34 – szereg parzysty i obliczamy numer decyla 1 (wzór 2.1.23):

$$Nr_{D_1} = \frac{N}{10} = \frac{34}{10} = \mathbf{3,4 \approx 3 = n_3}$$

$n_3$ – oznacza 3-go studenta

Student $n_3$ mieści się w przedziale $n_1 - n_4$. Tak, więc decyl 1 wynosi 2.

$$D_1 = 2{,}0$$

**Interpretacja:** Decyl 1 informuje, że 10% studentów otrzymała ocenę 2,0 i niższą, a 90% studentów otrzymała ocenę 2,0 i wyższą.

| Oceny ze statystyki $x_i$ | Liczba studentów $n_i$ | Wartości skumulowane *cum* |
|---|---|---|
| 2 | 4 | 4 $n_1 - n_4$ |
| 3 | 8 | 12 $n_5 - n_{12}$ |
| 3,5 | 11 | 23 $n_{13} - n_{23}$ |
| 4 | 6 | 29 $n_{24} - n_{29}$ |
| **4,5 $D_9$** | **3** | **32 $n_{30} - n_{32}$** |
| 5 | 2 | 34 $n_{33} - n_{34}$ |
| **Ogółem** | ***N* = 34** | **-** |

Decyl 9 $\boldsymbol{D_9 = 4,5}$ → **4,5 $D_9$**

**32 $n_{30} - n_{32}$** ← $n_{31}$ student znajduje się w tym przedziale

$N = 34$ – szereg parzysty i obliczamy numer decyla 9 (wzór 2.1.25):

$$Nr_{D_9} = \frac{9N}{10} = \frac{9 \cdot 34}{10} = \boldsymbol{30,6 \approx 31 = n_{31}}$$

$n_{31}$ – oznacza 31-go studenta

Student $n_{31}$ mieści się w przedziale $n_{30} - n_{32}$. Tak, więc decyl 9 wynosi 4,5.

$$D_9 = 4,5$$

**Interpretacja:** Decyl 9 informuje, że 90% studentów otrzymała ocenę 4,5 i niższą, a 10% studentów otrzymała ocenę 4,5 i wyższą.

5. **Odchylenie ćwiartkowe (wzór 2.2.12):**

$$Q_x = \frac{Q_3 - Q_1}{2} = \frac{4,0 - 3,0}{2} = 0,5$$

**Interpretacja:** Oceny z zaliczenia ze statystki różnią się (odchylają się) od mediany ($M_e = 3,5$) przeciętnie o ± 0,5 oceny (ale dotyczy to tylko 50% studentów znajdujących się w drugiej i trzeciej ćwiartce).

6. **Typowy obszar zmienności (wzór 2.2.13):**

$$M_e - Q_x < \boldsymbol{x_{typ}} < M_e + Q_x$$

$$3,5 - 0,5 < \boldsymbol{x_{typ}} < 3,5 + 0,5$$

$$3,0 < \boldsymbol{x_{typ}} < 4,0$$

7. **Współczynnik zmienności (wzór 2.2.15):**

$$V_Q = \frac{Q_x}{M_e} \cdot 100 = \frac{0{,}5}{3{,}5} \cdot 100 = 14{,}3\%$$

$$V_Q = 14{,}3\% \leq 35\% - \text{mała zmienność},$$

**Interpretacja:** Zbiorowość studentów charakteryzuje się małym zróżnicowaniem (zmiennością) pod względem otrzymanych ocen ze statystyki.

8. **Miara asymetrii (wzór 2.3.5):**

$$A_{Q_x} = \frac{Q_3 + Q_1 - 2M_e}{2Q_x} = \frac{4{,}0 + 3{,}0 - 2 \cdot 3{,}5}{2 \cdot 0{,}5} = 0{,}0$$

**Interpretacja:** Rozkład ocen ze statystyki jest symetryczny. Oznacza to, że połowa studentów ma oceny niższe od mediany ($M_e = 3{,}5$), a druga połowa wyższe.

Dodatkowo obliczymy współczynnik asymetrii (wzór 2.3.6):

$$A_s = \frac{\bar{x} - D}{S_x} = \frac{3{,}47 - 3{,}5}{0{,}77} = -0{,}04$$

$A_S < 0$ – rozkład asymetryczny o asymetrii lewostronnej

$0 < |-0{,}04| \leq 0{,}3$ – słaba siła asymetrii

9. **Współczynnik skupienia (wzór 2.4.5):**

$$W_{sk} = \frac{D_9 - D_1}{Q_3 - Q_1} = \frac{4{,}5 - 2{,}0}{4{,}0 - 3{,}0} = 2{,}5$$

$2{,}5 > 2$ to rozkład jest wysmukły

**Interpretacja:** Rozkład ocen ze statystyki jest wysmukły. Koncentracja zbiorowości pod względem badanej cechy wokół mediany jest silniejszy w porównaniu z rozkładem normalnym.

**Zestawienie podstawowych statystyk opisowych ocen studentów ze statystyki**

| Nr | Miary | Wartość |
|---|---|---|
| | **klasyczne** | |
| 1. | Średnia arytmetyczna – $\bar{x}$ | 3,47 |
| 2. | Wariancja – $S_x^2$ | 0,59 |
| 3. | Odchylenie standardowe – $S_x$ | 0,77 |
| 4. | Typowy obszar zmienności | $2,7 < x_{typ} < 4,24$ |
| 5. | Współczynnik zmienności – $V_x$ | 22,2% |
| 6. | Miara asymetrii – $A_{S_x}$ | –0,13 |
| 7. | Miara koncentracji – $W_k$ | 2,8 |
| | **pozycyjne** | |
| 1. | Dominanta – $D$ | 3,5 |
| 2. | Mediana – $M_e$ | 3,5 |
| 3. | Kwartyl 1 – $Q_1$ | 3,0 |
| | Kwartyl 3 – $Q_3$ | 4,0 |
| 4. | Decyl 1 – $D_1$ | 2,0 |
| | Decyl 9 – $D_9$ | 4,5 |
| 5. | Odchylenie ćwiartkowe – $Q_x$ | 0,5 |
| 6. | Typowy obszar zmienności | $3,0 < x_{typ} < 4,0$ |
| 7. | Współczynnik zmienności – $V_Q$ | 14,3% |
| 8. | Miara asymetrii – $A_{Q_x}$ | 0,0 |
| | Miara asymetrii (klasyczno-pozycyjna) – $A_s$ | –0,04 |
| 9. | Miara koncentracji – $W_{sk}$ | 2,5 |

Źródło: opracowanie własne.

**Przykład 2.3.**

Na podstawie zagregowanych danych z przykładu 1.1. (według grupowania I) przeprowadź kompleksową analizę stażu pracy pracowników z wykorzystaniem miar klasycznych i pozycyjnych a następnie zinterpretuj otrzymane parametry:

| Staż pracy (w latach) $x_i$ | Liczba pracowników $n_i$ |
|---|---|
| 1-4 | 4 |
| 4-7 | 10 |
| 7-10 | 6 |
| 10-13 | 4 |
| 13-16 | 3 |
| 16-19 | 1 |
| **Ogółem** | $\boldsymbol{N = 28}$ |

Źródło: Dane umowne.

**Rozwiązanie**

**Typ szeregu:** szereg rozdzielczy z przedziałami klasowymi.

**Zmienna badana $x_i$:** staż pracy pracowników w firmie „A".

## I. Analiza stażu pracy z wykorzystaniem miar klasycznych

Obliczenia pomocnicze:

| Staż pracy (w latach) $x_i$ | Liczba pracowników $n_i$ | Środek przedziału $\dot{x}_i$ | $\dot{x}_i \cdot n_i$ | $\dot{x}_i - \bar{x}$ | $(\dot{x}_i - \bar{x})^2$ | $(\dot{x}_i - \bar{x})^2 n_i$ |
|---|---|---|---|---|---|---|
| 1-4 | 4 | **2,5** | 10,0 | -5,5 | 30,3 | 121,2 |
| 4-7 | 10 | **5,5** | 55,0 | -2,5 | 6,3 | 63,0 |
| 7-10 | 6 | **8,5** | 51,0 | 0,5 | 0,3 | 1,8 |
| 10-13 | 4 | **11,5** | 46,0 | 3,5 | 12,3 | 49,2 |
| 13-16 | 3 | **14,5** | 43,5 | 6,5 | 42,3 | 126,9 |
| 16-19 | 1 | **17,5** | 17,5 | 9,5 | 90,3 | 90,3 |
| **Ogółem** | $N = 28$ | **-** | **223,0** | - | - | **452,4** |

Źródło: opracowanie własne.

1. **Średnia arytmetyczna (wzór 2.1.3):**

$$\bar{x} = \frac{\sum_{i=1}^{k} \dot{x}_i n_i}{N} = \frac{223{,}0}{28} = 7{,}96 \approx 8{,}0 \text{ lat}$$

**Interpretacja:** Średni staż pracy wśród pracowników w firmie „A" wynosi 8 lat.

2. **Wariancja (wzór 2.2.3):**

$$S_x^2 = \frac{\sum_{i=1}^{k} (\dot{x}_i - \bar{x})^2 n_i}{N} = \frac{452{,}4}{28} = 16{,}2$$

3. **Odchylenie standardowe (wzór 2.2.6):**

$$S_x = \sqrt{\frac{\sum_{i=1}^{k} (\dot{x}_i - \bar{x})^2 n_i}{N}} = \sqrt{\frac{452{,}4}{28}} = 4{,}0 \text{ lata}$$

**Interpretacja:** Staż pracy wśród pracowników w firmie „A" różni się (odchyla się) od średniej arytmetycznej ($\bar{x} = 8$ lat) przeciętnie o ± 4 lata.

4. **Typowy obszar zmienności (wzór 2.2.10):**

$$\bar{x} - S_x < \boldsymbol{x_{typ}} < \bar{x} + S_x$$
$$8{,}0 - 4{,}0 < \boldsymbol{x_{typ}} < 8{,}0 + 4{,}0$$
$$4{,}0 < \boldsymbol{x_{typ}} < 12{,}0$$

**Interpretacja:** W obliczonym przedziale stażu pracy (4–12 lat) znajduje się około 68% pracowników z firmy „A”.

5. **Klasyczny współczynnik zmienności (wzór 2.2.14):**

$$V_x = \frac{S_x}{\bar{x}} \cdot 100 = \frac{4{,}0}{8{,}0} \cdot 100 = 50{,}0\%$$

$35\% < V_x = 50\% \leq 60\% -$ umiarkowana zmienność

**Interpretacja:** Zbiorowość pracowników w firmie „A” charakteryzuje się umiarkowanym zróżnicowaniem (zmiennością) pod względem stażu pracy.

6. **Miara asymetrii (wzory 2.3.4 i 2.3.1):**

Obliczenia pomocnicze:

| $\dot{x}_i - \bar{x}$ | $(\dot{x}_i - \bar{x})^3$ | $(\dot{x}_i - \bar{x})^3 n_i$ |
|---|---|---|
| -5,5 | -166,4 | -665,6 |
| -2,5 | -15,6 | -156,0 |
| 0,5 | 0,1 | 0,6 |
| 3,5 | 42,9 | 171,6 |
| 6,5 | 274,6 | 823,8 |
| 9,5 | 857,4 | 857,4 |
| - | - | **Suma = 1031,8** |

Źródło: opracowanie własne.

$$\mu_3 = \frac{\sum_{i=1}^{k}(\dot{x}_i - \bar{x})^3 n_i}{N} = \frac{1031{,}8}{28} = 36{,}9$$

$$A_{S_x} = \frac{\mu_3}{S_x^3} = \frac{36{,}9}{4{,}0^3} = +0{,}58$$

$A_{S_x} > 0$ – rozkład asymetryczny o asymetrii prawostronnej

$0{,}3 < |+0{,}58| \leq 0{,}6$ – umiarkowana siła asymetrii

**Interpretacja:** Rozkład stażu pracy wskazuje na umiarkowaną asymetrię prawostronną. Oznacza to, że w zbiorowości dominują pracownicy o stażu pracy niższym od średniej arytmetycznej ($\bar{x} = 8{,}0$).

**7. Miara koncentracji (wzory 2.4.4 i 2.4.1):**

Obliczenia pomocnicze:

| $\dot{x}_i - \bar{x}$ | $(\dot{x}_i - \bar{x})^4$ | $(\dot{x}_i - \bar{x})^4 n_i$ |
|---|---|---|
| -5,5 | 915,1 | 3660,4 |
| -2,5 | 39,1 | 391,0 |
| 0,5 | 0,1 | 0,6 |
| 3,5 | 150,1 | 600,4 |
| 6,5 | 1785,1 | 5355,3 |
| 9,5 | 8145,1 | 8145,1 |
| - | - | **Suma = 18152,8** |

Źródło: opracowanie własne.

$$\mu_4 = \frac{\sum_{i=1}^{k}(\dot{x}_i - \bar{x})^4 n_i}{N} = \frac{18152{,}8}{28} = 648{,}3$$

$$W_k = \frac{\mu_4}{S_x^4} = \frac{648{,}3}{4{,}0^4} = 2{,}5$$

$$W_k < 3$$

$2{,}5 < 3$ – rozkład jest spłaszczony (platokurtyczny)

**Interpretacja:** Rozkład stażu pracy jest spłaszczony. Oznacza to, że koncentracja zbiorowości pracujących pod względem stażu pracy wokół średniej arytmetycznej jest słabsza od rozkładu normalnego.

## II. Analiza stażu pracy z wykorzystaniem miar pozycyjnych

**1. Dominanta (wzór 2.1.11):**

W tym przedziale znajduje się dominanta →

| Staż pracy (w latach) $x_i$ | Liczba pracowników $n_i$ |
|---|---|
| 1-4 | 4 $n_{D-1}$ |
| **4-7 *D*** | **10 $n_D$** ← Wartość max = 10 |
| 7-10 | 6 $n_{D+1}$ |
| 10-13 | 4 |
| 13-16 | 3 |
| 16-19 | 1 |
| **Ogółem** | ***N* = 28** |

Źródło: opracowanie własne.

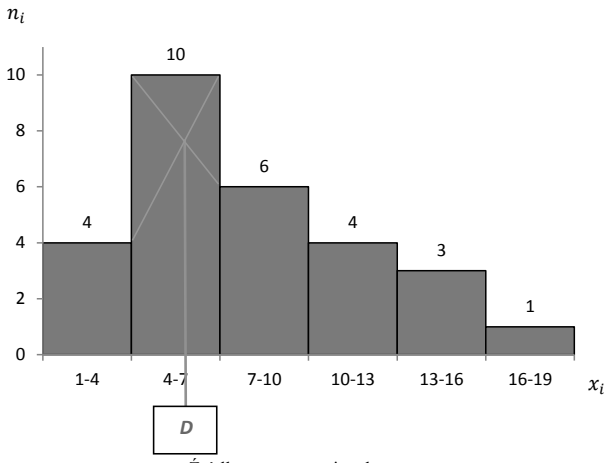
Rozpiętość przedziału: $h_0 = x_{max} - x_{min} = 7 - 4 = 3$

Początek przedziału: $x_0 = 4$

$$D = x_0 + \frac{(n_D - n_{D-1}) \cdot h_0}{(n_D - n_{D-1}) + (n_D - n_{D+1})} = 4 + \frac{(10-4) \cdot 3}{(10-4) + (10-6)} = 5{,}8 \text{ lat}$$

**Interpretacja:** W firmie „A" wśród badanej zbiorowości dominują pracownicy ze stażem pracy równym 5,8 lat.

**Wyznaczenie dominanty *D* metodą graficzną**
**Histogram liczebności cechy $x_i$.**




Źródło: opracowanie własne.

**2. Mediana (wzór 2.1.16):**

Obliczamy liczebności skumulowane *cum*:

| Liczba pracowników $n_i$ | Wartości skumulowane *cum* | Przedziały |
|---|---|---|
| 4 | 4 | $n_1 - n_4$ |
| 10 | 4 + 10 = 14 | $n_5 - n_{14}$ |
| 6 | 4 + 10 + 6 = 20 | $n_{15} - n_{20}$ |
| 4 | 4 + 10 + 6 + 4 = 24 | $n_{21} - n_{24}$ |
| 3 | 4 + 10 + 6 + 4 + 3 = 27 | $n_{25} - n_{27}$ |
| 1 | 4 + 10 + 6 + 4 + 3 + 1 = 28 | $n_{28}$ |

Źródło: opracowanie własne.

Jeżeli każdego pracownika ponumerujemy od 1 do 28 to otrzymamy:

$$n_1, n_2, n_3, n_4, n_5, n_6 \ldots, n_{28}$$

Wartości skumulowane możemy zapisać w postać przedziałów np. wartość skumulowaną 4 możemy przedstawić jako przedział: $n_1 - n_4$. Oznacza to, że w tym przedziale znajduje się 4 pracowników o numerach porządkowych: $n_1$, $n_2$, $n_3$ i $n_4$.

W tym przedziale znajduje się mediana →

| Staż pracy (w latach) $x_i$ | Liczba pracowników $n_i$ | Wartości skumulowane *cum* |
|---|---|---|
| 1-4 | 4 | 4 $n_1 - n_4$ ← $cum_{n_{i-1}}$ |
| **4-7 $M_e$** | **10 $n_{M_e}$** | **14 $n_5 - n_{14}$** |
| 7-10 | 6 | 20 $n_{15} - n_{20}$ |
| 10-13 | 4 | 24 $n_{21} - n_{24}$ |
| 13-16 | 3 | 27 $n_{25} - n_{27}$ |
| 16-19 | 1 | 28 $n_{28}$ |
| **Ogółem** | ***N* = 28** | **-** |

Źródło: opracowanie własne.

Obliczamy numer pracownika:

$$n_{M_e} = \frac{N}{2} = \frac{28}{2} = \mathbf{14} = \boldsymbol{n_{14}}$$

$n_{14}$ – oznacza 14-go pracownika

Pracownik $n_{14}$ mieści się w przedziale wartości skumulowanych $n_5 - n_{14}$. Mediana stażu pracy znajduje się w przedziale klasowym 4 – 7 lat.

Rozpiętość przedziału: $h_0 = x_{max} - x_{min} = 7 - 4 = 3$

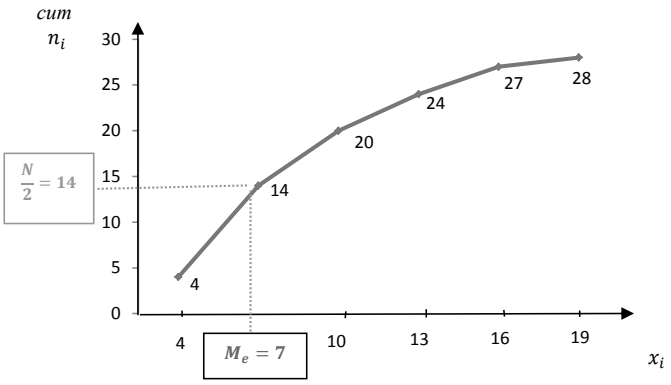
Początek przedziału: $x_0 = 4$

$$n_{M_e} = 10$$

$$M_e = x_0 + \left(\frac{N}{2} - cum_{n_{i-1}}\right)\frac{h_0}{n_{M_e}} = 4 + (\mathbf{14} - 4)\frac{3}{10} = 7{,}0 \text{ lat}$$

**Interpretacja:** Mediana informuje, że 50% pracowników ma staż pracy 7 lat i niższy, a druga połowa pracowników ma staż pracy 7 lat i wyższy.

**Wyznaczenie mediany $M_e$ metodą graficzną**




Źródło: opracowanie własne.

### 3. Kwartyle (wzory 2.1.21 i 2.1.22):

W tym przedziale znajduje się kwartyl 1 →

| Staż pracy (w latach) $x_i$ | Liczba pracowników $n_i$ | Wartości skumulowane *cum* |
|---|---|---|
| 1-4 | 4 | 4 $n_1 - n_4$ ← $cum_{n_{i-1}}$ |
| **4-7 $Q_1$** | **10 $n_{Q_1}$** | **14 $n_5 - n_{14}$** |
| 7-10 | 6 | 20 $n_{15} - n_{20}$ |
| 10-13 | 4 | 24 $n_{21} - n_{24}$ |
| 13-16 | 3 | 27 $n_{25} - n_{27}$ |
| 16-19 | 1 | 28 $n_{28}$ |
| **Ogółem** | ***N* = 28** | **-** |

Źródło: opracowanie własne.

Obliczamy numer pracownika:

$$n_{Q_1} = \frac{N}{4} = \frac{28}{4} = \mathbf{7} = \boldsymbol{n_7}$$

$n_7$ – oznacza 7-go pracownika

Pracownik $n_7$ mieści się w przedziale wartości skumulowanych $n_5 - n_{14}$.

Rozpiętość przedziału: $h_0 = x_{max} - x_{min} = 7 - 4 = 3$

Początek przedziału: $x_0 = 4$

$$n_{Q_1} = 10$$

$$Q_1 = x_0 + \left(\frac{N}{4} - cum_{n_{i-1}}\right)\frac{h_0}{n_{Q_1}} = 4 + (\mathbf{7} - 4)\frac{3}{10} = 4{,}9 \text{ lat}$$

**Interpretacja:** Kwartyl 1 informuje, że 25% pracowników ma staż pracy 4,9 lata i niższy, a 75% pracowników ma staż pracy 4,9 lata i wyższy.

| Staż pracy (w latach) $x_i$ | Liczba pracowników $n_i$ | Wartości skumulowane *cum* |
|---|---|---|
| 1-4 | 4 | 4 $n_1 - n_4$ |
| 4-7 | 10 | 14 $n_5 - n_{14}$ |
| 7-10 | 6 | 20 $n_{15} - n_{20}$ ← $cum_{n_{i-1}}$ |
| **10-13 $Q_3$** | **4 $n_{Q_1}$** | **24 $n_{21} - n_{24}$** |
| 13-16 | 3 | 27 $n_{25} - n_{27}$ |
| 16-19 | 1 | 28 $n_{28}$ |
| **Ogółem** | ***N* = 28** | **-** |

W tym przedziale znajduje się kwartyl 3 →

Źródło: opracowanie własne.

Obliczamy numer pracownika:

$$n_{Q_3} = \frac{3N}{4} = \frac{3 \cdot 28}{4} = \mathbf{21} = \boldsymbol{n_{21}}$$

$n_{21}$ – oznacza 21-go pracownika

Pracownik $n_{21}$ mieści się w przedziale wartości skumulowanych $n_{21} - n_{24}$.

Rozpiętość przedziału: $h_0 = x_{max} - x_{min} = 13 - 10 = 3$

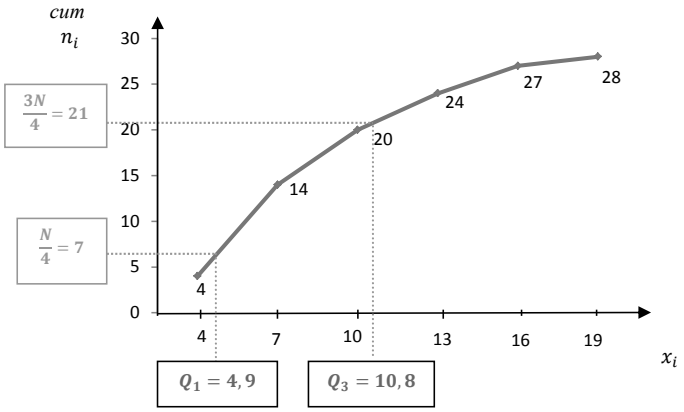
Początek przedziału: $x_0 = 10$

$$n_{Q_3} = 4$$

$$Q_3 = x_0 + \left(\frac{3N}{4} - cum_{n_{i-1}}\right)\frac{h_0}{n_{Q_3}} = 10 + (\mathbf{21} - 20)\frac{3}{4} = 10{,}75 \approx 10{,}8 \text{ lat}$$

**Interpretacja:** Kwartyl 3 informuje, że 75% pracowników ma staż pracy 10,8 lat i niższy, a 25% pracowników ma staż pracy 10,8 lat i wyższy.

**Wyznaczenie kwartyli $Q_1$ i $Q_3$ metodą graficzną**




Źródło: opracowanie własne.

**4. Decyle (wzory 2.1.27 i 2.1.28):**

W tym przedziale znajduje się decyl 1 →

| Staż pracy (w latach) $x_i$ | Liczba pracowników $n_i$ | Wartości skumulowane *cum* |
|---|---|---|
| **1-4 $D_1$** | **4 $n_{D_1}$** | **4 $n_1 - n_4$** |
| 4-7 | 10 | 14 $n_5 - n_{14}$ |
| 7-10 | 6 | 20 $n_{15} - n_{20}$ |
| 10-13 | 4 | 24 $n_{21} - n_{24}$ |
| 13-16 | 3 | 27 $n_{25} - n_{27}$ |
| 16-19 | 1 | 28 $n_{28}$ |
| **Ogółem** | ***N* = 28** | **-** |

Źródło: opracowanie własne.

Obliczamy numer pracownika:

$$n_{D_1} = \frac{N}{10} = \frac{28}{10} = \mathbf{2,8} = \boldsymbol{n_3}$$

$n_3$ – oznacza 3-go pracownika

Pracownik $n_3$ mieści się w przedziale wartości skumulowanych $n_1 - n_4$.

Rozpiętość przedziału: $h_0 = x_{max} - x_{min} = 4 - 1 = 3$

Początek przedziału: $x_0 = 1$

$$n_{D_1} = 4$$

$$D_1 = x_0 + \left(\frac{N}{10} - cum_{n_{i-1}}\right)\frac{h_0}{n_{D_1}} = 1 + (\mathbf{2,8} - 0)\frac{3}{4} = 3{,}1 \text{ lat}$$

**Interpretacja:** Decyl 1 informuje, że 10% pracowników ma staż pracy 3,1 lata i niższy, a 90% pracowników ma staż pracy 3,1 lata i wyższy.

| Staż pracy (w latach) $x_i$ | Liczba pracowników $n_i$ | Wartości skumulowane *cum* |
|---|---|---|
| 1-4 | 4 | 4 $n_1 - n_4$ |
| 4-7 | 10 | 14 $n_5 - n_{14}$ |
| 7-10 | 6 | 20 $n_{15} - n_{20}$ |
| 10-13 | 4 | 24 $n_{21} - n_{24}$ |
| **13-16 $D_9$** | **3 $n_{D_9}$** | **27 $n_{25} - n_{27}$** |
| 16-19 | 1 | 28 $n_{28}$ |
| **Ogółem** | ***N* = 28** | **-** |

W tym przedziale znajduje się decyl 9 → ← $cum_{n_{i-1}}$

Źródło: opracowanie własne.

Obliczamy numer pracownika:

$$n_{D_9} = \frac{9N}{10} = \frac{9 \cdot 28}{10} = \mathbf{25,2} = \boldsymbol{n_{25}}$$

$n_{25}$ – oznacza 25-go pracownika

Pracownik $n_{25}$ mieści się w przedziale wartości skumulowanych $n_{25} - n_{27}$.

Rozpiętość przedziału: $h_0 = x_{max} - x_{min} = 16 - 13 = 3$

Początek przedziału: $x_0 = 13$

$$n_{D_9} = 3$$

$$D_9 = x_0 + \left(\frac{9N}{10} - cum_{n_{i-1}}\right)\frac{h_0}{n_{D_9}} = 13 + (\mathbf{25,2} - 24)\frac{3}{3} = 14{,}2 \text{ lat}$$

**Interpretacja:** Decyl 9 informuje, że 90% pracowników ma staż pracy 14,2 lata i niższy, a 10% pracowników ma staż pracy 14,2 lata i wyższy.

**5. Odchylenie ćwiartkowe (wzór 2.2.12):**

$$Q_x = \frac{Q_3 - Q_1}{2} = \frac{10{,}8 - 4{,}9}{2} = 2{,}95 \approx 3{,}0 \text{ lata}$$

**Interpretacja:** Staż pracy wśród pracowników w firmie „A" różni się (odchyla się) od mediany ( $M_e$ = 7 lat) przeciętnie o ± 3 lata (ale dotyczy to tylko 50% zbiorowości znajdującej się w drugiej i trzeciej ćwiartce).

6. **Typowy obszar zmienności (wzór 2.2.13):**

$$M_e - Q_x < \boldsymbol{x_{typ}} < M_e + Q_x$$

$$7{,}0 - 3{,}0 < \boldsymbol{x_{typ}} < 7{,}0 + 3{,}0$$

$$4{,}0 < \boldsymbol{x_{typ}} < 10{,}0$$

7. **Współczynnik zmienności (wzór 2.2.15):**

$$V_Q = \frac{Q_x}{M_e} \cdot 100 = \frac{3{,}0}{7{,}0} \cdot 100 = 42{,}9\%$$

$$35\% < V_Q = 42{,}5\% \leq 60\% - \text{umiarkowana zmienność}$$

**Interpretacja:** Zbiorowość pracowników w firmie „A" charakteryzuje się umiarkowanym zróżnicowaniem (zmiennością) pod względem stażu pracy.

8. **Miara asymetrii (wzór 2.3.5):**

$$A_{Q_x} = \frac{Q_3 + Q_1 - 2M_e}{2Q_x} = \frac{10{,}8 + 4{,}9 - 2 \cdot 7{,}0}{2 \cdot 3{,}0} = +0{,}28$$

$A_{Q_x} > 0$ – rozkład asymetryczny o asymetrii prawostronnej

$0 < |+0{,}28| \leq 0{,}3$ – słaba siła asymetrii

**Interpretacja:** Rozkład stażu pracy wskazuje na słabą asymetrię prawostronną w obszarze 50% środkowych jednostek tzn. pomiędzy kwartylem 1 i 3.

Dodatkowo obliczymy współczynnik asymetrii (wzór 2.3.6):

$$A_s = \frac{\bar{x} - D}{S_x} = \frac{8{,}0 - 5{,}8}{4{,}0} = +0{,}55$$

$A_S > 0$ – rozkład asymetryczny o asymetrii prawostronnej

$0{,}3 < |+0{,}55| \leq 0{,}6$ – umiarkowana siła asymetrii

**9. Współczynnik skupienia (wzór 2.4.5):**

Rozkład jest słabo symetryczny $A_{Q_x} = +0{,}28$, tak więc możemy obliczyć współczynnik skupienia:

$$W_{sk} = \frac{D_9 - D_1}{Q_3 - Q_1} = \frac{14{,}2 - 3{,}1}{10{,}8 - 4{,}9} = 1{,}88 \approx 1{,}9$$

$1{,}9 < 2$ – rozkład jest spłaszczony (platokurtyczny)

**Interpretacja:** Rozkład stażu pracy jest spłaszczony. Oznacza to, że koncentracja zbiorowości pracujących pod względem stażu pracy wokół mediany jest słabsza od rozkładu normalnego.

**Zestawienie podstawowych statystyk opisowych stażu pracy pracowników**

| Nr | Miary | Wartość |
|---|---|---|
| | **klasyczne** | |
| 1. | Średnia arytmetyczna – $\bar{x}$ | 8,0 |
| 2. | Wariancja– $S_x^2$ | 16,2 |
| 3. | Odchylenie standardowe – $S_x$ | 4,0 |
| 4. | Typowy obszar zmienności | 4,0 < $x_{typ}$ < 12,0 |
| 5. | Współczynnik zmienności – $V_x$ | 50,0% |
| 6. | Miara asymetrii – $A_{S_x}$ | +0,58 |
| 7. | Miara koncentracji – $W_k$ | 2,5 |
| | **pozycyjne** | |
| 1. | Dominanta – $D$ | 5,8 |
| 2. | Mediana – $M_e$ | 7,0 |
| 3. | Kwartyl 1 – $Q_1$ | 4,9 |
| | Kwartyl 3 – $Q_3$ | 10,8 |
| 4. | Decyl 1 – $D_1$ | 3,1 |
| | Decyl 9 – $D_9$ | 14,2 |
| 5. | Odchylenie ćwiartkowe – $Q_x$ | 3,0 |
| 6. | Typowy obszar zmienności | 4,0 < $x_{typ}$ < 10,0 |
| 7. | Współczynnik zmienności – $V_Q$ | 42,9% |
| 8. | Miara asymetrii – $A_{Q_x}$ | +0,28 |
| | Miara asymetrii (klasyczno-pozycyjna) – $A_s$ | +0,55 |
| 9. | Miara koncentracji – $W_{sk}$ | 1,9 |

Źródło: opracowanie własne.

**Przykład 2.4.**

Liczba ofert pracy w Polsce według miesięcy w 2018 r. była następująca:

| Miesiące | Liczba ofert pracy (stan w końcu miesiąca) |
|---|---|
| styczeń | 103341 |
| luty | 102401 |
| marzec | 99699 |
| kwiecień | 113686 |
| maj | 112086 |
| czerwiec | 110601 |
| lipiec | 106021 |
| sierpień | 107592 |
| wrzesień | 98042 |
| październik | 99128 |
| listopad | 90314 |
| grudzień | 62711 |

Źródło: Bank Danych Lokalnych GUS.

Oblicz średnią chronologiczną.

**Rozwiązanie**

**Obliczam średnią chronologiczną (wzór 2.1.8):**

$$\bar{x}_{ch} = \frac{0{,}5 \cdot x_1 + x_2 + x_3 + \ \ldots + 0{,}5 \cdot x_n}{N - 1}$$

$$\bar{x}_{ch} = \frac{0{,}5 \cdot 103341 + 102401 + 99699 + \ 113686 \ + 112086 + 110601 +}{12 - 1}$$

$$\frac{+106021 + 107592 + 98042 + 99128 + 90314 + 0{,}5 \cdot 62711}{12 - 1} = 102054{,}2$$

**Interpretacja:** Średnia liczba ofert pracy w 2018 r. wynosiła 102054,2.

**Przykład 2.5.**

W pewnym zakładzie pracy w ciągu 1 godziny poddano obserwacji 10 pracowników ze względu na liczbę produkowanych detali. Dane zestawiono w tabeli:

| Produkcja (w szt.) $x_i$ | Liczba pracowników $n_i$ |
|---|---|
| 2 | 3 |
| 4 | 5 |
| 6 | 2 |
| **Suma** | $\boldsymbol{N = 10}$ |

Źródło: dane umowne.

Oblicz średnią harmoniczną.

**Rozwiązanie**

**Obliczam średnią harmoniczną (wzór 2.1.10):**

$$\bar{x}_{\square ar} = \frac{n_1 + n_2 + n_3 + \cdots + n_k}{\frac{n_1}{x_1} + \frac{n_2}{x_2} + \frac{n_3}{x_3} + \cdots + \frac{n_k}{x_k}} = \frac{3 + 5 + 2}{\frac{3}{2} + \frac{5}{4} + \frac{2}{6}} = 3{,}2 \text{ szt.}$$

**Interpretacja:** Średnio w ciągu godziny każdy pracownik produkował 3,2 szt.

**Przykład 2.6.**

Wysokość godzinowych stawek pracowników zestawiono w tabeli:

| Stawka godzinowa (w zł) | Liczba pracowników |
|---|---|
| 15-20 | 50 |
| 20-25 | 30 |
| 25-30 | 20 |
| 30-35 | 10 |
| 35-40 | 5 |

Źródło: dane umowne.

Oblicz poprawkę Shepparda.

**Rozwiązanie**

Obliczenia pomocnicze:

| Stawka godzinowa (w zł) $x_i$ | Liczba pracowników $n_i$ | Środek przedziału $\dot{x}_i$ | $\dot{x}_i \cdot n_i$ | $\dot{x}_i - \bar{x}$ | $(\dot{x}_i - \bar{x})^2$ | $(\dot{x}_i - \bar{x})^2 n_i$ |
|---|---|---|---|---|---|---|
| 15-20 | 50 | **17,5** | 875,0 | -5,22 | 27,25 | 1362,50 |
| 20-25 | 30 | **22,5** | 675,0 | -0,22 | 0,05 | 1,50 |
| 25-30 | 20 | **27,5** | 550,0 | 4,78 | 22,85 | 457,00 |
| 30-35 | 10 | **32,5** | 325,0 | 9,78 | 95,65 | 956,50 |
| 35-40 | 5 | **37,5** | 187,5 | 14,78 | 218,45 | 1092,25 |
| **Ogółem** | ***N* = 115** | **-** | **2612,5** | - | - | **3869,75** |

Źródło: opracowanie własne.

1. **Średnia arytmetyczna (wzór 2.1.3):**

$$\bar{x} = \frac{\sum_{i=1}^{k} \dot{x}_i n_i}{N} = \frac{2612{,}5}{115} = 22{,}72 \text{ zł}$$

**Interpretacja:** Średni stawka godzinowa wynagrodzenia wśród pracowników wynosi 22,72 zł.

2. **Odchylenie standardowe (wzór 2.2.6):**

$$S_x = \sqrt{\frac{\sum_{i=1}^{k} (\dot{x}_i - \bar{x})^2 n_i}{N}} = \sqrt{\frac{3869{,}75}{115}} = 5{,}80 \text{ zł}$$

**Interpretacja:** Stawki godzinowe wśród pracowników różnią się (odchylają się) od średniej arytmetycznej ($\bar{x} = 22{,}72$ zł) przeciętnie o ± 5,80 zł.

3. **Poprawka Shepparda (wzór 2.2.11):**

$$S_{x_{skor}} = \sqrt{S_x^2 - \frac{h^2}{12}} = \sqrt{5{,}80^2 - \frac{5^2}{12}} = 5{,}62$$

**Interpretacja:** Wartość skorygowanego odchylenia wynosi 5,62 zł i jest niższa o 0,18 zł, od odchylenia standardowego $S_x = 5{,}80$ zł.

**Przykład 2.7.**

Oblicz współczynnik koncentracji Lorenza liczby pracowników w firmach:

| Wielkość firmy według liczby pracowników | Liczba firm | Liczba pracowników w firmach |
|---|---|---|
| poniżej 9 | 25 | 125 |
| 10 - 49 | 20 | 600 |
| 50 - 249 | 16 | 2400 |
| 250 i więcej | 9 | 3500 |
| **Suma** | **70** | **6625** |

Dane umowne.

## Rozwiązanie

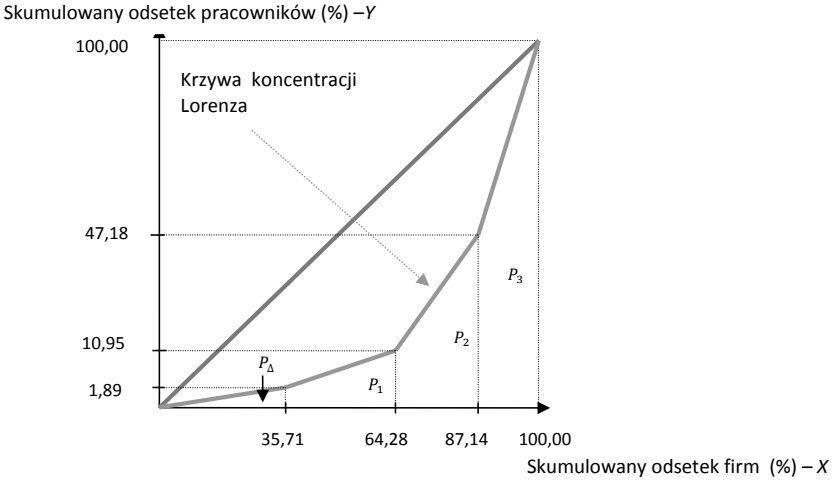
Obliczenia pomocnicze:

| Wielkość firmy według liczby pracowników | Liczba firm | Liczba pracowników w firmach | Odsetek firm (%) | Odsetek pracowników (%) | Skumulowane odsetki firm ($X$) | Skumulowane odsetki pracowników ($Y$) |
|---|---|---|---|---|---|---|
| poniżej 9 | 25 | 125 | 35,71 | 1,89 | 35,71 | 1,89 |
| 10 - 49 | 20 | 600 | 28,57 | 9,06 | 64,28 | 10,95 |
| 50 - 249 | 16 | 2400 | 22,86 | 36,23 | 87,14 | 47,18 |
| 250 i więcej | 9 | 3500 | 12,86 | 52,82 | 100,00 | 100,00 |
| **Suma** | **70** | **6625** | **100,00** | **100,00** | **x** | **x** |

Źródło: opracowanie własne.

### 1. Obliczam pola trójkąta i trapezów (wzór 2.4.7):

**Graficzne wyznaczenie krzywej koncentracji Lorenza**




Źródło: opracowanie własne.

- **pole trójkąta:**

$$P_{\Delta} = \frac{1{,}89 \cdot 35{,}71}{2} = 33{,}7$$

- **pole trapezów:**

$$P_1 = \frac{1{,}89 + 10{,}95}{2} \cdot 28{,}57 = 183{,}4$$

$$P_2 = \frac{10{,}95 + 47{,}18}{2} \cdot 22{,}86 = 664{,}4$$

$$P_3 = \frac{47{,}18 + 100}{2} \cdot 12{,}86 = 946{,}4$$

$$P_b = P_{\Delta} + \sum_{i=2}^{n} P_{trap.} = 33{,}7 + 183{,}4 + 664{,}4 + 946{,}4 = 1827{,}9$$

2. **Obliczam współczynnik koncentracji Lorenza (wzór 2.4.6):**

$$W_{KL} = \frac{5000 - P_b}{5000} = \frac{5000 - 1827{,}9}{5000} = 0{,}63$$

**Interpretacja:** Obliczona wartość ($W_{KL} = 0{,}63$) wskazuje na duży stopień koncentracji pracowników w firmach.

**Przykład 2.8.**

Na podstawie danych przedstawiających liczbę nowo utworzonych miejsc pracy według województw w 2018 r:

| Województwo | Liczba nowo utworzonych miejsc pracy (w tys.) |
|---|---|
| Lubuskie | 11,2 |
| Podlaskie | 11,6 |
| Opolskie | 14,0 |
| Warmińsko-mazurskie | 16,8 |
| Świętokrzyskie | 18,5 |
| Lubelskie | 28,2 |
| Zachodniopomorskie | 29,1 |
| Podkarpackie | 30,5 |
| Kujawsko-pomorskie | 31,2 |
| Łódzkie | 39,5 |
| Pomorskie | 43,5 |
| Dolnośląskie | 47,6 |
| Małopolskie | 68,9 |
| Wielkopolskie | 79,4 |
| Śląskie | 96,1 |
| Mazowieckie | 151,7 |

Źródło: Bank Danych Lokalnych GUS.

Oblicz współczynnik Giniego.

**Rozwiązanie**

Obliczenia pomocnicze:

| Województwo | Liczba nowo utworzonych miejsc pracy (w tys.) $x_i$ | Liczba województw $n_i$ | $\frac{n_i}{N}$ | $\frac{x_i n_i}{\sum_{i=1}^{k} x_i n_i}$ | $cum\frac{n_i}{N}$ | $z_i$ |
|---|---|---|---|---|---|---|
| Lubuskie | 11,2 | 1 | 0,0625 | 0,016 | 0,063 | 0,001 |
| Podlaskie | 11,6 | 1 | 0,0625 | 0,016 | 0,125 | 0,003 |
| Opolskie | 14,0 | 1 | 0,0625 | 0,020 | 0,188 | 0,006 |
| Warmińsko-mazurskie | 16,8 | 1 | 0,0625 | 0,023 | 0,250 | 0,010 |
| Świętokrzyskie | 18,5 | 1 | 0,0625 | 0,026 | 0,313 | 0,014 |
| Lubelskie | 28,2 | 1 | 0,0625 | 0,039 | 0,375 | 0,027 |
| Zachodniopomorskie | 29,1 | 1 | 0,0625 | 0,041 | 0,438 | 0,033 |
| Podkarpackie | 30,5 | 1 | 0,0625 | 0,042 | 0,500 | 0,040 |
| Kujawsko-pomorskie | 31,2 | 1 | 0,0625 | 0,043 | 0,563 | 0,046 |
| Łódzkie | 39,5 | 1 | 0,0625 | 0,055 | 0,625 | 0,065 |
| Pomorskie | 43,5 | 1 | 0,0625 | 0,061 | 0,688 | 0,080 |
| Dolnośląskie | 47,6 | 1 | 0,0625 | 0,066 | 0,750 | 0,095 |
| Małopolskie | 68,9 | 1 | 0,0625 | 0,096 | 0,813 | 0,150 |
| Wielkopolskie | 79,4 | 1 | 0,0625 | 0,111 | 0,875 | 0,187 |
| Śląskie | 96,1 | 1 | 0,0625 | 0,134 | 0,938 | 0,243 |
| Mazowieckie | 151,7 | 1 | 0,0625 | 0,211 | 1,000 | 0,410 |
| **Ogółem** | **$N$ = 717,8** | **$N$ = 16** | **1,000** | **1,000** | **x** | **1,410** |

Źródło: opracowanie własne.

**Obliczam współczynnik Giniego (wzór 2.4.8):**

$$W_G = \sum_{i=1}^{k} \left( \left( cum \frac{n_{i-1}}{N} \right) + \left( cum \frac{n_i}{N} \right) \right) \cdot \frac{x_i n_i}{\sum_{i=1}^{k} x_i \, n_i} - 1 =$$

$$= \sum_{i=1}^{k} z_i - 1 = 1{,}41 - 1 = 0{,}41$$

**Interpretacja:** Obliczona wartość ($W_G = 0{,}41$) wskazuje na umiarkowany stopień koncentracji (nierówności) liczby nowo utworzonych miejsc pracy według województw w Polsce w 2018 r.

# 3. Analiza współzależności

## 3.1. Podstawowe pojęcia

Badaniem związków pomiędzy cechami statystycznymi zajmuje się dział statystyki nazywanej teorią współzależności. Celem analizy współzależności jest poznawanie występujących związków przyczynowo - skutkowych pomiędzy cechami. Zależność korelacyjna (szczególny przypadek zależności stochastycznej) występuje wtedy, gdy wartościom jednej cechy przyporządkowane są średnie drugiej cechy. Istnienie wiele różnorodnych miar korelacji zarówno dla cech ilościowych, jak i jakościowych. Do najważniejszych miar można zaliczyć:

a) współczynnik korelacji liniowej Pearsona,
b) współczynnik zbieżności Czuprowa,
c) współczynnik kolejnościowy rang Spearmana,
d) współczynnik korelacji rang Kendalla,
e) współczynnik kontyngencji *C*-Pearsona.

Analiza korelacji w sposób liczbowy zajmuje się badaniem siły i kierunku wpływu, pomiędzy cechą *X* i *Y*. Każda z tych cech ma swoją określoną nazwę a mianowicie:

- cecha *X* – jest przyczyną, zmienną objaśniającą (niezależną),
- cecha *Y* – jest to skutkiem, zmienną objaśnianą (zależną).

Ze względu na kierunek zależności wyróżnia się następujące korelacje (rys. 3.1.1):

a) liniową dodatnią (+), tzn. wzrostowi wartości cechy *X* towarzyszy jednoczesny wzrost wartości cechy *Y*, są to zmiany jednokierunkowe,
b) liniową ujemną (–), tzn. wzrostowi wartości cechy *X* towarzyszy jednoczesny spadek wartości cechy *Y*, są to zmiany różnokierunkowe,
c) krzywoliniową (nieliniowa), tzn. tutaj występuje zarówno zależność korelacyjna dodatnia, jak i ujemna,
d) brak korelacji, tzn. nie występuje zależność korelacyjna cech.

Zależność dodatnią i ujemną pomiędzy cechami *X* i *Y* można przedstawić za pomocą dwóch szeregów danych zestawionych obok siebie:

**Tabela 3.1.1. Przykład korelacji dodatniej (+)**

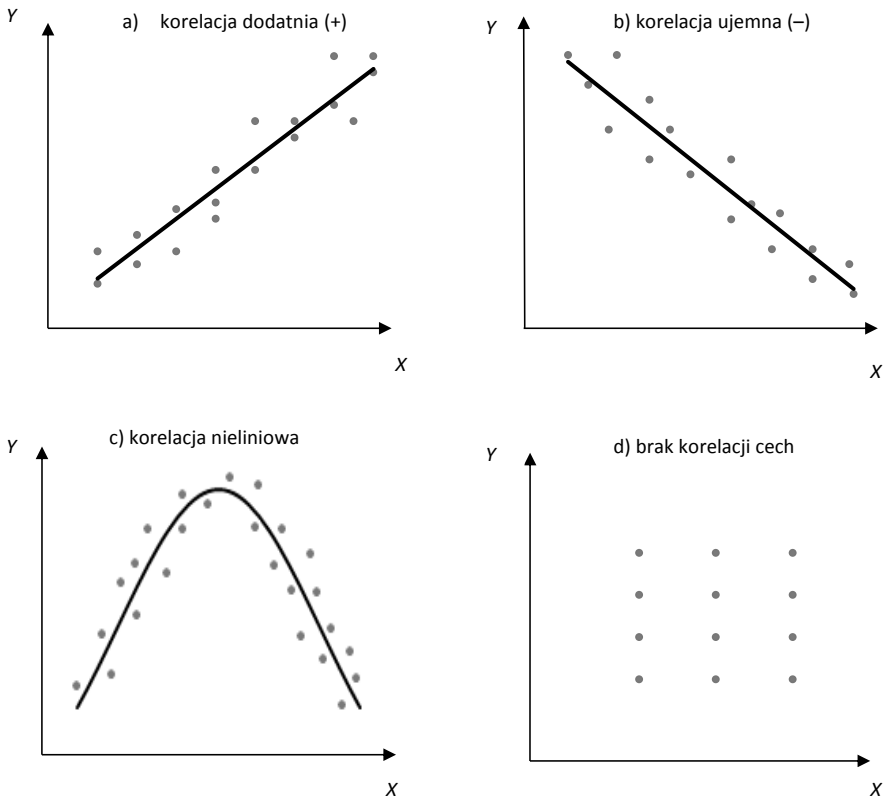
| | | | | | | |
|---|---|---|---|---|---|---|
| Cecha *X* – rośnie | 4 | 7 | 10 | 12 | 16 | 20 |
| Cecha *Y* – rośnie | 2 | 5 | 8 | 10 | 14 | 18 |

Źródło: opracowanie własne.

**Tabela 3.1.2. Przykład korelacji ujemnej (–)**

| | | | | | | |
|---|---|---|---|---|---|---|
| Cecha *X* – rośnie | 4 | 7 | 10 | 12 | 16 | 20 |
| Cecha *Y* – maleje | 23 | 20 | 15 | 11 | 9 | 5 |

Źródło: opracowanie własne.




**Rysunek 3.1.1. Zależności korelacyjne pomiędzy cechami *X* i *Y***
Źródło: opracowanie własne.

## 3.2. Współczynnik korelacji Pearsona

**Współczynnik korelacji liniowej Pearsona ($r_{xy}$)** jest miarą opisową, która określa zarówno siłę, jak i kierunek zależności korelacyjnej pomiędzy dwoma cechami ilościowymi *X* i *Y* gdy związek pomiędzy nimi jest liniowy. Współczynnik korelacji Pearsona oparty jest na tzw. kowariancji cech – *cov*(*xy*), którą możemy zapisać w postaci:

$$cov(xy) = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{N} \tag{3.2.1}$$

gdzie:

$x_i$ – wartość cechy *X*,
$y_i$ – wartość cechy *Y*,
$\bar{x}$ – średnia arytmetyczna cechy *X*,
$\bar{y}$ – średnia arytmetyczna cechy *Y*,
*N* – liczba obserwacji.

Współczynnik korelacji Pearsona oblicza się według wzoru:

$$r_{xy} = r_{yx} = \frac{cov(xy)}{S(x)S(y)} = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_{i=1}^{n}(x_i - \bar{x})^2 \sum_{i=1}^{n}(y_i - \bar{y})^2}} \tag{3.2.2}$$

gdzie:

$S(x)$ – odchylenie standardowe cechy *X*,
$S(y)$ – odchylenie standardowe cechy *Y*.

Współczynnik korelacji liniowej Pearsona jest miarą unormowaną, która przyjmuje wartość z przedziału:

$$-1 \leq r_{xy} \leq +1.$$

Kierunek współzależności określa znak współczynnika korelacji: znak (–) oznacza korelację ujemną, z kolei znak (+) korelację dodatnią. Im wartość współczynnika jest bliższa 1 (lub -1) tym zależność korelacyjna cech jest silniejsza, natomiast im wartość współczynnika zbliża się do 0 wówczas siła korelacji jest coraz słabsza. W przypadku braku zależności współczynnik $r_{xy} = 0$. W sytuacji, kiedy pomiędzy cechami $r_{xy} = 1$ lub $-1$ mówimy o tzw. zależności doskonałej. Orientacyjnie siłę korelacji pomiędzy cechami *X* i *Y* można zdefiniować, jako:

- $0{,}0 < |r_{xy}| \leq 0{,}2$ – bardzo słaba,
- $0{,}2 < |r_{xy}| \leq 0{,}4$ – słaba,
- $0{,}4 < |r_{xy}| \leq 0{,}6$ – umiarkowana,
- $0{,}6 < |r_{xy}| \leq 0{,}8$ – silna,
- $0{,}8 < |r_{xy}| \leq 1{,}0$ – bardzo silna.

Kwadrat współczynnika korelacji liniowej Pearsona $r_{xy}^2$ – nazywamy **współczynnikiem determinacji**. Współczynnik determinacji informuje o tym jaka część zmienności cechy *X* została wyjaśniona przez zmienność cechy *Y*. Z kolei $100\% \cdot r_{xy}^2$ umożliwia określenie zmienności cech w ujęciu procentowym.

## 3.3. Współczynnik zbieżności Czuprowa

**Współczynnik zbieżności Czuprowa ($T_{xy}$)** stosuje się w badaniu zależności cech ilościowych, jak i jakościowych. Miara ta wykorzystuje statystykę *chi-kwadrat* ($\chi^2$) o postaci:

$$\chi^2 = \sum_{i=1}^{k}\sum_{j=1}^{r}\frac{\left(n_{ij} - \hat{n}_{ij}\right)^2}{\hat{n}_{ij}} = \sum_{i=1}^{k}\sum_{j=1}^{r}\left(\frac{\hat{n}_{ij}^2}{\hat{n}_{ij}}\right) - n \tag{3.3.1}$$

gdzie:

$$\hat{n}_{ij} = \frac{n_{i.} \cdot n_{.j}}{n} \tag{3.3.2}$$

$n_{ij}$ – liczebności rzeczywiste (empiryczne),
$\hat{n}_{ij}$ – liczebności teoretyczne (oczekiwane),
$n_{i.}$ – suma liczebności *i*-tego wiersza,
$n_{.j}$ – suma liczebności *j*-tej kolumny,
*n* – łączna suma liczebności,
*r* – liczba wierszy,
*k* – liczba kolumn.

Statystykę $\chi^2$ według wzoru 3.3.1 stosujemy w tablicach prostokątnych. W tablica kwadratowych 2 x 2 jeżeli N > 40 i wszystkie liczebności oczekiwane większe od 10 stosujemy wzór:

$$\chi^2 = \frac{(ad + bc)^2 \cdot n}{(a+b)(c+d)(a+c)(b+d)} \tag{3.3.3}$$

Jeżeli 20 < N < 40 i którakolwiek z częstości jest mniejsza od 5 to stosujemy poprawkę Yatesa:

$$\chi^2 = \frac{\left(|ad - bc| - \frac{n}{2}\right)^2 \cdot n}{(a+b)(c+d)(a+c)(b+d)} \tag{3.3.4}$$

Współczynnik zbieżności oblicza się według wzoru:

$$T_{xy} = \sqrt{\frac{\chi^2}{n\sqrt{(r-1)(k-1)}}} \tag{3.3.5}$$

Wartości współczynnika korelacji Czuprowa znajdują się w przedziale [0, 1]. Ten współczynnik nie wskazuje kierunku korelacji (jest zawsze dodatni) co jest jego największą wadą. Jeżeli $T_{xy}$= 0, wówczas zmienne są stochastycznie niezależne. Im współczynnik zbieżności jest bliższy zeru tym zależność między cechami jest słabsza.

## 3.4. Współczynnik rang Spearmana

**Współczynnik korelacji rang Spearmana ($r_s$)** służy do opisu siły korelacji zarówno cech ilościowych, jak i jakościowych w sytuacji, kiedy mamy możliwość uporządkowania ich wariantów. Wzór na obliczenie współczynnika kolejnościowego rang Spearmana możemy zapisać w postaci:

$$r_s = 1 - \frac{6\sum_{i=1}^{n} d_i^2}{n(n^2-1)} \tag{3.4.1}$$

gdzie:

$d_i$ – różnica pomiędzy rangami cechy $X$ i $Y$ tzn. $d_i = x_i - y_i$,
$n$ – liczba par cech $X$ i $Y$.

Rangowanie polega na uporządkowaniu wartości cech $X$ i $Y$ w kolejności rosnącej lub malejącej. Następnie nadaje się im tzw. rangi: 1, 2, 3 …, $n$. W sytuacji, kiedy w uporządkowanym szeregu wystąpią identyczne wartości wówczas obliczamy średnią arytmetyczną z ich kolejnych numerów i przyporządkowujemy im otrzymaną wartość. Współczynnik korelacji rang przyjmuje wartości z przedziału:

$$-1 \leq r_s \leq +1.$$

Im współczynnik $r_s$ jest bliższy 1 (lub -1) związek jest silniejszy, z kolei, kiedy jest bliższy 0 związek korelacyjny jest coraz słabszy. Współczynnik korelacji rang Spearmana stosuje się, kiedy liczba obserwacji wynosi $n < 30$.

## 3.5. Współczynnik rang Kendalla

**Wartość współczynnika korelacji rang Kendalla ($r_k$)** oblicza się według wzoru:

$$r_k = \frac{S}{\frac{n(n-1)}{2}} \tag{3.5.1}$$

gdzie:
*S* – różnica wartości *W* i *M,*
*W* – suma liczb większych na prawo od danej rangi,
*M* – suma liczb mniejszych na prawo od danej rangi,
*n* – liczba par.

Alternatywnie wartość współczynnika rang Kendalla można obliczyć:

$$r_k = \frac{2R}{\frac{n(n-1)}{2}} - 1 \tag{3.5.2}$$

gdzie:
*R* – suma not +1,
*n* – liczba par.

Współczynnik rang Kendalla (wzór 3.5.1) oblicza się w ten sposób, że w pierwszej kolejności porządkujemy nadane rangi wartościom cechy *X* w kolejności rosnącej, natomiast rangi dla cechy *Y* pozostawiamy w kolejności naturalnej. Wówczas analizując rangi dla cechy *Y* ustalamy ile liczb jest większych (*W*) i mniejszych (*M*) na prawo od danej rangi dokonując ich zsumowania. Współczynnik korelacji rang Kendalla przyjmuje wartości z przedziału:

$$-1 \leq r_k \leq +1.$$

Procedura obliczenia rang Kendalla według wzoru (3.5.2) polega na nadaniu wartościom cech *X* i *Y* zależnie od natężenia odpowiednich rang. Rangi cechy *X* porządkuje się w kolejności rosnącej, natomiast rangi cechy *Y* pozostawia się w kolejności naturalnej. Rangi cechy *Y* łączy się w pary. Jeżeli wartość rangi poprzedzającej w parze jest wyższa od wartości rangi następnej nadaje się notę (-1) a w odwrotnej sytuacji notę +1. Ostatnim krokiem jest zsumowanie not +1.

## 3.6. Korelacja cech niemierzalnych

Badanie zależności cech jakościowych odbywa się za pomocą tzw. tablic wielodzielnych (kontyngencji, krzyżowych, asocjacji). Tablice wielodzielne mogą być kwadratowe lub prostokątne zależnie od liczby pól. Wartości liczbowe badanych cech *X* i *Y* zestawia się w następujący sposób:

| X \ Y | $Y^1$ | $Y^2$ | **Ogółem** |
|---|---|---|---|
| $X^1$ | *a* | *b* | ***a + b*** |
| $X^2$ | *c* | *d* | ***c + d*** |
| **Ogółem** | ***a + c*** | ***b + d*** | ***n = a + b +c + d*** |

gdzie: $X^1$ *i* $X^2$ – warianty cechy *X;* $Y^1$ *i* $Y^2$ – warianty cechy *X; a* – liczba jednostek występująca w wariancie $X^1$ *i* $Y^1$ dla cech *X i Y; b* – liczba jednostek występująca w wariancie $X^1$ *i* $Y^2$ dla cech *X i Y; c* – liczba jednostek występująca w wariancie $X^2$ *i* $Y^1$ dla cech *X i Y; d* – liczba jednostek występująca w wariancie $X^2$ *i* $Y^2$ dla cech *X i Y; n* – liczebność próby.

**Rysunek 3.1.1. Przykład tablicy wielodzielnej kwadratowej – czteropolowej (2 x 2)**
Źródło: opracowane na podstawie: M., Sobczyk, *Statystyka*, PWN, Warszawa 2016, s. 245;
H., Augustyniak, *Statystyka opisowa z elementami demografii*, Poznań 2002, s. 115.

Dla tablic dwudzielnych (2 x 2 – czteropolowych) mają zastosowanie następujące współczynniki korelacji cech jakościowych:

- **współczynnik Pearsona ($\varphi$):**

$$\varphi = \frac{ad - bc}{\sqrt{(a+b)\cdot(c+d)\cdot(a+c)\cdot(b+d)}} \tag{3.6.1}$$

- **współczynnik Bykowskiego (*W*):**

$$W = \frac{(a+d) - (b+c)}{a+b+c+d} \tag{3.6.2}$$

- **współczynnik Yule'a-Kendalla (*Q*):**

$$Q = \frac{ad - bc}{ad + bc} \tag{3.6.3}$$

Współczynniki te przyjmują wartości z przedziału – 1 do + 1. Znak współczynnika cech jakościowych nie informuje o kierunku zależności – nie ma on znaczenia. Znak współczynnika zależy od układu wartości liczbowych w tabeli wielodzielnej.

**Współczynnik kontyngencji (zbieżności) *C*-Pearsona ( $C_{xy}$)** może być stosowany w tablicach wielodzielnych o dowolnej liczbie pól. Współczynnik kontyngencji oblicza się w sytuacji, kiedy zależność pomiędzy badanymi cechami jest istotna statystycznie. Wzór jest następujący:

$$C_{xy} = \sqrt{\frac{\chi^2}{\chi^2 + n}} \tag{3.6.4}$$

gdzie:

*n* – liczebność próby,

$\chi^2$ – oblicza się ze wzoru (3.3.1) dla tablicy prostokątnej lub ze wzorów (3.3.3 - 3.3.4) dla tablicy kwadratowej.

Ze względu na nieskończony wzrost wartości $C_{xy}$ wraz ze wzrostem liczy kolumn i wierszy ustala się tzw. kres górny ($C_{max}$). W przypadku tablicy kwadratowej maksymalna wartość współczynnika C wynosi:

$$C_{max} = \sqrt{\frac{k-1}{k}} \tag{3.6.5}$$

gdzie:

*k* – liczba kolumn.

W przypadku tablicy prostokątnej $C_{max}$ wynosi:

$$C_{max} = \frac{\sqrt{\frac{k-1}{k}} + \sqrt{\frac{r-1}{r}}}{2} \tag{3.6.6}$$

gdzie:

*k* – liczba kolumn,

*r* – liczba wierszy.

Skorygowana wartość współczynnika $C_{kor_{xy}}$ oblicza się jako iloraz wartości współczynnik kontyngencji *C*-Pearsona $C_{xy}$ przez wartość $C_{max}$:

$$C_{kor_{xy}} = \frac{C_{xy}}{C_{max}} \tag{3.6.7}$$

Wartości współczynnika $C_{kor}$ mieszczą się teoretyczne w przedziale:

$$0 \leq C_{kor_{xy}} \leq 1$$

Zwykle jest tak, że współczynnik $C_{kor}$ przyjmuje wartości z przedziału:

$$0 \leq C_{kor_{xy}} \leq C_{max}$$

## Przykłady

**Przykład 3.1.**

Na podstawie danych dotyczących liczby pracujących na 1000 ludności i przeciętnego miesięcznym wynagrodzenia w gospodarce narodowej według województw w 2017 r.:

| Województwa | Liczba pracujących na 1000 ludności (w ‰) $y_i$ | Przeciętne miesięczne wynagrodzenia brutto w gospodarce narodowej (w zł) $x_i$ |
|---|---|---|
| | 2017 r. | |
| Dolnośląskie | 391,2 | 4400,05 |
| Kujawsko-pomorskie | 354,5 | 3717,21 |
| Lubelskie | 396,3 | 3824,28 |
| lubuskie | 345,9 | 3754,54 |
| Łódzkie | 405,8 | 3925,97 |
| Małopolskie | 417,0 | 4097,35 |
| Mazowieckie | 486,2 | 5219,09 |
| Opolskie | 335,8 | 3923,58 |
| Podkarpackie | 400,3 | 3684,71 |
| Podlaskie | 364,8 | 3815,23 |
| Pomorskie | 376,2 | 4211,69 |
| Śląskie | 383,6 | 4247,44 |
| Świętokrzyskie | 385,7 | 3705,65 |
| Warmińsko-mazurskie | 315,8 | 3641,32 |
| Wielkopolskie | 441,9 | 3937,81 |
| Zachodniopomorskie | 330,4 | 3890,86 |

Źródło: Bank Danych Lokalnych GUS.

Określ siłę i kierunek zależności pomiędzy badanymi cechami z wykorzystaniem współczynnika korelacji Pearsona.

**Rozwiązanie**

Średnie arytmetyczne cech $x_i$ i $y_i$ wynoszą odpowiednio:

$$\bar{x} = 3999{,}80 \text{ zł}$$

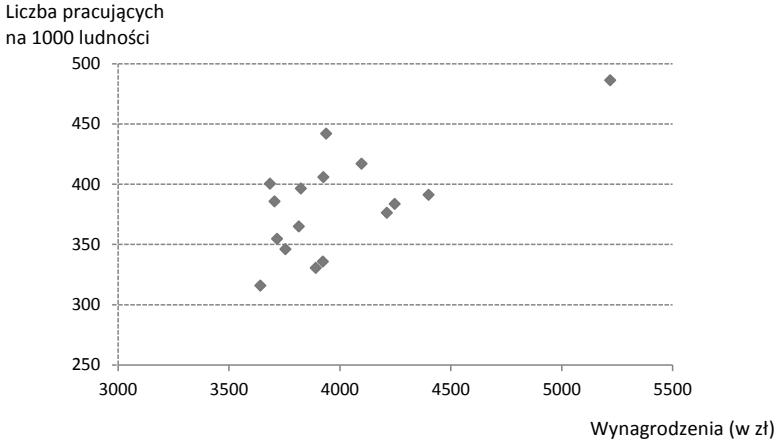
$$\bar{y} = 383{,}2‰.$$

Obliczenia pomocnicze:

| $x_i$ | $y_i$ | $x_i - \bar{x}$ | $y_i - \bar{y}$ | $(x_i - \bar{x})(y_i - \bar{y})$ | $(x_i - \bar{x})^2$ | $(y_i - \bar{y})^2$ |
|---|---|---|---|---|---|---|
| 4400,05 | 391,2 | 400,25 | 8,0 | 3202,0 | 160200,1 | 64,0 |
| 3717,21 | 354,5 | -282,59 | -28,7 | 8110,3 | 79857,1 | 823,7 |
| 3824,28 | 396,3 | -175,52 | 13,1 | -2299,3 | 30807,3 | 171,6 |
| 3754,54 | 345,9 | -245,26 | -37,3 | 9148,2 | 60152,5 | 1391,3 |
| 3925,97 | 405,8 | -73,83 | 22,6 | -1668,6 | 5450,9 | 510,8 |
| 4097,35 | 417,0 | 97,55 | 33,8 | 3297,2 | 9516,0 | 1142,4 |
| 5219,09 | 486,2 | 1219,29 | 103,0 | 125586,9 | 1486668,1 | 10609,0 |
| 3923,58 | 335,8 | -76,22 | -47,4 | 3612,8 | 5809,5 | 2246,8 |
| 3684,71 | 400,3 | -315,09 | 17,1 | -5388,0 | 99281,7 | 292,4 |
| 3815,23 | 364,8 | -184,57 | -18,4 | 3396,1 | 34066,1 | 338,6 |
| 4211,69 | 376,2 | 211,89 | -7,0 | -1483,2 | 44897,4 | 49,0 |
| 4247,44 | 383,6 | 247,64 | 0,4 | 99,1 | 61325,6 | 0,2 |
| 3705,65 | 385,7 | -294,15 | 2,5 | -735,4 | 86524,2 | 6,3 |
| 3641,32 | 315,8 | -358,48 | -67,4 | 24161,6 | 128507,9 | 4542,8 |
| 3937,81 | 441,9 | -61,99 | 58,7 | -3638,8 | 3842,8 | 3445,7 |
| 3890,86 | 330,4 | -108,94 | -52,8 | 5752,0 | 11867,9 | 2787,8 |
| - | - | - | - | **171152,9** | **2308775,1** | **28422,4** |

Źródło: opracowanie własne.

**Diagram korelacyjny (rozrzutu) cech $x_i$ i $y_i$**




Źródło: opracowanie własne.

**1. Obliczam współczynnik korelacji liniowej Pearsona (wzór 3.2.2):**

$$r_{xy} = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_{i=1}^{n}(x_i - \bar{x})^2 \sum_{i=1}^{n}(y_i - \bar{y})^2}} = \frac{171152,9}{\sqrt{2308775,1 \cdot 28422,4}} = 0,67$$

$0,6 < |0,67| \leq 0,8$ – silna zależność

**Interpretacja:** Pomiędzy badanymi cechami zachodzi dodatnia i silna zależność korelacyjna ($r_{xy} = 0{,}67$). Oznacza to, że wzrost przeciętnego miesięcznego wynagrodzenia brutto w gospodarce narodowej powoduje wzrost liczby pracujących na 1000 ludności według województw w 2017 r.

2. **Obliczam współczynnik determinacji:**

$$r_{xy}^2 \cdot 100\% = 0{,}67^2 \cdot 100\% = 44{,}89\%$$

**Interpretacja:** Współczynnik determinacji wynosi $r_{xy}^2 = 44{,}89\%$. Oznacza to, że około 45% zmienności cechy $y_i$ zostało wyjaśnione zmiennością cechy $x_i$.

**Przykład 3.2.**

Na podstawie liczby pracujących w Polsce za IV kwartał w 2017 r., zbadaj czy występuje zależność pomiędzy płcią pracujących a poziomem wykształcenia. Do obliczeń zastosuj współczynnik zbieżności Czuprowa.

| Płeć | Poziom wykształcenia ( w tys.) | | | | | | **Razem** |
|---|---|---|---|---|---|---|---|
| | wyższe | policealne | średnie zawodowe | średnie ogólnokształcące | zasadnicze zawodowe | gimnazjalne i niepełne podstawowe | |
| mężczyźni | 2466 | 169 | 2319 | 743 | 2864 | 529 | **9090** |
| kobiety | 3273 | 377 | 1493 | 716 | 1177 | 278 | **7314** |
| **Razem** | **5739** | **546** | **3812** | **1459** | **4041** | **807** | **16404** |

Źródło: *Aktywność ekonomiczna ludności Polski IV kwartał 2017 r.*, GUS, Warszawa 2018, s.115.

**Rozwiązanie**

1. **Obliczam wartości teoretyczne $\hat{n}_{ij}$ (wzór 3.3.2):**

| Płeć | Poziom wykształcenia ( w tys.) | | | | | | $n_{i.}$ |
|---|---|---|---|---|---|---|---|
| | wyższe | policealne | średnie zawodowe | średnie ogólnokształcące | zasadnicze zawodowe | gimnazjalne i niepełne podstawowe | |
| mężczyźni | $\hat{n}_{11}$ | $\hat{n}_{12}$ | $\hat{n}_{13}$ | $\hat{n}_{14}$ | $\hat{n}_{15}$ | $\hat{n}_{16}$ | **9090** |
| kobiety | $\hat{n}_{21}$ | $\hat{n}_{22}$ | $\hat{n}_{23}$ | $\hat{n}_{24}$ | $\hat{n}_{25}$ | $\hat{n}_{26}$ | **7314** |
| $n_{.j}$ | **5739** | **546** | **3812** | **1459** | **4041** | **807** | **16404** |

Źródło: opracowanie własne.

| Mężczyźni | |
|---|---|
| $\hat{n}_{11} = \frac{n_{i.} \cdot n_{.j}}{N} = \frac{9090 \cdot 5739}{16404} = 3180$ | $\hat{n}_{12} = \frac{n_{i.} \cdot n_{.j}}{N} = \frac{9090 \cdot 546}{16404} = 303$ |
| $\hat{n}_{13} = \frac{n_{i.} \cdot n_{.j}}{N} = \frac{9090 \cdot 3812}{16404} = 2112$ | $\hat{n}_{14} = \frac{n_{i.} \cdot n_{.j}}{N} = \frac{9090 \cdot 1459}{16404} = 808$ |
| $\hat{n}_{15} = \frac{n_{i.} \cdot n_{.j}}{N} = \frac{9090 \cdot 4041}{16404} = 2239$ | $\hat{n}_{16} = \frac{n_{i.} \cdot n_{.j}}{N} = \frac{9090 \cdot 807}{16404} = 447$ |
| Kobiety | |
| $\hat{n}_{21} = \frac{n_{i.} \cdot n_{.j}}{N} = \frac{7314 \cdot 5739}{16404} = 2559$ | $\hat{n}_{22} = \frac{n_{i.} \cdot n_{.j}}{N} = \frac{7314 \cdot 546}{16404} = 243$ |
| $\hat{n}_{23} = \frac{n_{i.} \cdot n_{.j}}{N} = \frac{7314 \cdot 3812}{16404} = 1700$ | $\hat{n}_{24} = \frac{n_{i.} \cdot n_{.j}}{N} = \frac{7314 \cdot 1459}{16404} = 651$ |
| $\hat{n}_{25} = \frac{n_{i.} \cdot n_{.j}}{N} = \frac{7314 \cdot 4041}{16404} = 1802$ | $\hat{n}_{26} = \frac{n_{i.} \cdot n_{.j}}{N} = \frac{7314 \cdot 807}{16404} = 360$ |

Źródło: opracowanie własne.

| Płeć | Poziom wykształcenia ( w tys.) | | | | | | **Razem** |
|---|---|---|---|---|---|---|---|
| | wyższe | policealne | średnie zawodowe | średnie ogólnokształcące | zasadnicze zawodowe | gimnazjalne i niepełne podstawowe | |
| mężczyźni | 3180 | 303 | 2112 | 808 | 2239 | 447 | **9089** |
| kobiety | 2559 | 243 | 1700 | 651 | 1802 | 360 | **7315** |
| **Razem** | **5739** | **546** | **3812** | **1459** | **4041** | **807** | **16404** |

Źródło: opracowanie własne.

## 2. Obliczam statystykę *chi*-kwadrat (wzór 3.3.1):

Obliczenia pomocnicze:

| $n_{ij}$ | $\hat{n}_{ij}$ | $n_{ij} - \hat{n}_{ij}$ | $\left(n_{ij} - \hat{n}_{ij}\right)^2$ | $\frac{\left(n_{ij} - \hat{n}_{ij}\right)^2}{\hat{n}_{ij}}$ |
|---|---|---|---|---|
| 2466 | 3180 | -714 | 509796 | 160,3 |
| 169 | 303 | -134 | 17956 | 59,3 |
| 2319 | 2112 | 207 | 42849 | 20,3 |
| 743 | 808 | -65 | 4225 | 5,2 |
| 2864 | 2239 | 625 | 390625 | 174,5 |
| 529 | 447 | 82 | 6724 | 15,0 |
| 3273 | 2559 | 714 | 509796 | 199,2 |
| 377 | 243 | 134 | 17956 | 73,9 |
| 1493 | 1700 | -207 | 42849 | 25,2 |
| 716 | 651 | 65 | 4225 | 6,5 |
| 1177 | 1802 | -625 | 390625 | 216,8 |
| 278 | 360 | -82 | 6724 | 18,7 |
| **1604** | **1604** | **x** | **x** | **974,9** |

Źródło: opracowanie własne.

$$\chi^2 = \sum_{i=1}^{k}\sum_{j=1}^{r}\frac{\left(n_{ij} - \hat{n}_{ij}\right)^2}{\hat{n}_{ij}} = 974{,}9$$

**3. Obliczam współczynnik zbieżności Czuprowa (wzór 3.3.5):**

$$T_{xy} = \sqrt{\frac{\chi^2}{n\sqrt{(r-1)(k-1)}}} = \sqrt{\frac{974{,}9}{16404\sqrt{(2-1)(6-1)}}} = 0{,}16$$

$|0{,}16| \leq 0{,}2$ – bardzo słaba zależność

**Interpretacja:** Wśród pracujących pomiędzy płcią i poziomem wykształcenia zachodzi bardzo słaba zależność korelacyjna ($T_{xy} = 0{,}16$).

**Przykład 3.3.**

Określ za pomocą współczynnika korelacji rang Spearmana siłę i kierunek zależności pomiędzy liczbą ofert pracy na 1000 ludności a stopą bezrobocia, jeżeli uzyskano następujące dane:

| Województwa | Liczba ofert pracy na 1000 ludności (w ‰) $x_i$ | Stopa bezrobocia rejestrowanego (w %) $y_i$ |
|---|---|---|
| | 2017 r. | |
| Dolnośląskie | 2,8 | 5,7 |
| Kujawsko-pomorskie | 1,5 | 9,9 |
| Lubelskie | 0,9 | 8,8 |
| lubuskie | 2,6 | 6,5 |
| Łódzkie | 2,2 | 6,7 |
| Małopolskie | 1,5 | 5,3 |
| Mazowieckie | 1,5 | 5,6 |
| Opolskie | 3,4 | 7,3 |
| Podkarpackie | 0,9 | 9,6 |
| Podlaskie | 1,1 | 8,5 |
| Pomorskie | 1,8 | 5,4 |
| Śląskie | 2,5 | 5,1 |
| Świętokrzyskie | 1,0 | 8,8 |
| Warmińsko-mazurskie | 1,3 | 11,7 |
| Wielkopolskie | 1,2 | 3,7 |
| Zachodniopomorskie | 2,0 | 8,5 |

Źródło: Bank Danych Lokalnych GUS.

**Rozwiązanie**

Obliczenia pomocnicze:

| Województwa | $x_i$ | $y_i$ | Rangi $x_i$ | Rangi $y_i$ | $d_i = x_i - y_i$ | $d_i^2$ |
|---|---|---|---|---|---|---|
| Dolnośląskie | 2,8 | 5,7 | 15 | 6 | 9 | 81,00 |
| Kujawsko-pomorskie | 1,5 | 9,9 | 8 | 15 | -7 | 49,00 |
| Lubelskie | 0,9 | 8,8 | 1,5 | 12,5 | -11 | 121,00 |
| lubuskie | 2,6 | 6,5 | 14 | 7 | 7 | 49,00 |
| Łódzkie | 2,2 | 6,7 | 12 | 8 | 4 | 16,00 |
| Małopolskie | 1,5 | 5,3 | 8 | 3 | 5 | 25,00 |
| Mazowieckie | 1,5 | 5,6 | 8 | 5 | 3 | 9,00 |
| Opolskie | 3,4 | 7,3 | 16 | 9 | 7 | 49,00 |
| Podkarpackie | 0,9 | 9,6 | 1,5 | 14 | -12,5 | 156,25 |
| Podlaskie | 1,1 | 8,5 | 4 | 10,5 | -6,5 | 42,25 |
| Pomorskie | 1,8 | 5,4 | 10 | 4 | 6 | 36,00 |
| Śląskie | 2,5 | 5,1 | 13 | 2 | 11 | 121,00 |
| Świętokrzyskie | 1,0 | 8,8 | 3 | 12,5 | -9,5 | 90,25 |
| Warmińsko-mazurskie | 1,3 | 11,7 | 6 | 16 | -10 | 100,00 |
| Wielkopolskie | 1,2 | 3,7 | 5 | 1 | 4 | 16,00 |
| Zachodniopomorskie | 2,0 | 8,5 | 11 | 10,5 | 0,5 | 0,25 |
| - | | - | - | - | - | **961,00** |

Źródło: opracowanie własne.

**Obliczam współczynnik korelacji rang Spearmana (wzór 3.4.1):**

$$r_s = 1 - \frac{6\sum_{i=1}^{n} d_i^2}{n(n^2-1)} = 1 - \frac{6 \cdot 961}{16(16^2-1)} = -0{,}41$$

$0{,}4 < |-0{,}41| \leq 0{,}6$ – umiarkowana zależność

**Interpretacja:** Występuje umiarkowany i ujemny związek korelacyjny pomiędzy liczbą ofert pracy na 1000 ludności a stopą bezrobocia ($r_s = -0{,}41$). Oznacza to, że wzrost liczby ofert pracy na 1000 ludności powoduje spadek stopy bezrobocia.

**Przykład 3.4.**

Za pomocą współczynnika korelacji rang Kendalla, należy zbadać czy istnieje zależność korelacyjna pomiędzy przeciętnym miesięcznym wynagrodzeniem brutto a wskaźnikiem zatrudnienia według poziomu wykształcenia na podstawie następujących danych:

| Poziom wykształcenia | Przeciętne wynagrodzenie brutto[1] (w zł) $x_i$ | Wskaźnik zatrudnienia według BAEL[2] (w %) $y_i$ |
|---|---|---|
| Wyższe | 4325,21 | 78,2 |
| Policealne i średnie zawodowe | 4125,36 | 60,6 |
| Średnie ogólnokształcące | 3945,68 | 47,5 |
| Zasadnicze zawodowe | 4289,11 | 54,7 |
| Gimnazjalne i niepełne podstawowe | 3421,59 | 14,8 |

[1] – Dane umowne, [2] – Dane średnioroczne
Źródło: Bank Danych Lokalnych GUS.

**Rozwiązanie**

**Obliczam wartość współczynnika korelacji rang Kendalla (wzór 3.5.1):**

- **nadajemy rangi dla cech:**

| Poziom wykształcenia | $x_i$ | $y_i$ | Rangi $x_i$ | Rangi $y_i$ |
|---|---|---|---|---|
| Wyższe | 4325,21 | 78,2 | 5 | 5 |
| Policealne i średnie zawodowe | 4125,36 | 60,6 | 3 | 4 |
| Średnie ogólnokształcące | 3945,68 | 47,5 | 2 | 2 |
| Zasadnicze zawodowe | 4289,11 | 54,7 | 4 | 3 |
| Gimnazjalne i niepełne podstawowe | 3421,59 | 14,8 | 1 | 1 |

Źródło: opracowanie własne.

- **porządkujemy rosnąco cechę $x_i$ według rang:**

| | | | | | |
|---|---|---|---|---|---|
| Rangi $x_i$ | 1 | 2 | 3 | 4 | 5 |
| Rangi $y_i$ | 1 | 2 | 4 | 3 | 5 |

Źródło: opracowanie własne.

Rangi większe od liczb: 1, 2, 4, 3, 5 po prawej stronie cechy $y_i$:

$$W = 4 + 3 + 1 + 1 + 0 = 9$$

Rangi mniejsze od liczb: 1, 2, 4, 3, 5 po prawej stronie cechy $y_i$:

$$M = 0 + 0 + 1 + 0 + 0 = 1$$

$$S = W - M = 9 - 1 = 8$$

$$r_k = \frac{S}{\frac{n(n-1)}{2}} = \frac{8}{\frac{5(5-1)}{2}} = 0{,}8$$

$0{,}6 < |0{,}8| \leq 0{,}8$ – silna zależność

1. **Obliczam wartość współczynnika korelacji rang Kendalla (wzór 3.5.2):**

- **tworzymy pary z rang dla cechy $y_i$:**

| | | | | | |
|---|---|---|---|---|---|
| Rangi $y_i$ | 1 | 2 | 4 | 3 | 5 |

Pary dla rang:

1: (1, 2), (1, 4), (1, 3) , (1, 5),

2: (2, 4), (2, 3), (2, 5),

4: (4, 3), (4, 5),

3: (3, 5).

- **wstawiamy noty dla poszczególnych par:**

Jeżeli w danej parze pierwsza ranga jest mniejsza od drugiej to dla tej pary wstawiamy notę +1 i odwrotnie, jeżeli pierwsza ranga jest większa od drugiej to wówczas wstawiamy notę -1.

1: (1, 2), (1, 4), (1, 3) , (1, 5),

**Nota** +1 +1 +1 +1

2: (2, 4), (2, 3), (2, 5),

**Nota** +1 +1 +1

4: (4, 3), (4, 5),

**Nota** - 1 +1

3: (3, 5).

**Nota** +1

Suma dodatnich not równa się $R = 9$.

$$r_k = \frac{2R}{\frac{n(n-1)}{2}} - 1 = \frac{2 \cdot 9}{\frac{5(5-1)}{2}} - 1 = 0{,}8$$

**Interpretacja:** Występuje silna i dodatnia zależność korelacyjna pomiędzy przeciętnym miesięcznym wynagrodzeniem brutto a wskaźnikiem zatrudnienia ($r_k = 0{,}8$). Oznacza to, że wzrost przeciętnego miesięcznego wynagrodzenia powoduje wzrost wskaźnika zatrudnienia według poziomu wykształcenia.

**Przykład 3.5.**

Na podstawie danych oblicz związek korelacyjny pomiędzy płcią a miejscem zamieszkania z wykorzystaniem współczynnika Pearsona, Bykowskiego i Yule'a-Kendalla.

| Miejsce zamieszkania | Płeć | | **Ogółem** |
|---|---|---|---|
| | mężczyźni | kobiety | |
| miasto | 35 a | 38 b | **73** |
| wieś | 45 c | 22 d | **67** |
| **Ogółem** | **80** | **60** | **140** |

Dane umowne.

**Rozwiązanie**

- **współczynnik Pearsona (wzór 3.6.1):**

$$\varphi = \frac{ad - bc}{\sqrt{(a+b)\cdot(c+d)\cdot(a+c)\cdot(b+d)}} = \\ = \frac{35 \cdot 22 - 38 \cdot 45}{\sqrt{(35+38)\cdot(45+22)\cdot(35+45)\cdot(38+22)}} = -0{,}19$$

**Interpretacja:** Pomiędzy płcią a miejscem zamieszkania zachodzi bardzo słaba zależność korelacyjna ($W = -0{,}19$).

- **współczynnik Bykowskiego (wzór 3.6.2):**

$$W = \frac{(a+d)-(b+c)}{a+b+c+d} = \frac{(35+22)-(38+45)}{35+38+45+22} = -0{,}19$$

**Interpretacja:** Pomiędzy płcią a miejscem zamieszkania zachodzi bardzo słaba zależność korelacyjna ($W = -0{,}19$).

- **współczynnik Yule'a-Kendalla (wzór 3.6.3):**

$$Q = \frac{ad - bc}{ad + bc} = \frac{35 \cdot 22 - 38 \cdot 45}{35 \cdot 22 + 38 \cdot 45} = -0{,}38$$

**Interpretacja:** Pomiędzy płcią a miejscem zamieszkania zachodzi słaba zależność korelacyjna ($Q = -0{,}38$).

**Przykład 3.6.**

Na podstawie danych (Przykład 3.2) oblicz współczynnik kontyngencji *C*-Pearsona.

**Rozwiązanie**

$\chi^2$ wynosi:

$$\chi^2 = \sum_{i=1}^{k}\sum_{j=1}^{r}\frac{\left(n_{ij} - \hat{n}_{ij}\right)^2}{\hat{n}_{ij}} = 974{,}9$$

1. **Obliczam współczynnik kontyngencji *C*-Pearsona (wzór 3.6.4):**

$$C_{xy} = \sqrt{\frac{\chi^2}{\chi^2 + n}} = \sqrt{\frac{974{,}9}{974{,}9 + 16404}} = 0{,}24$$

2. **Obliczam $C_{max}$ dla tablicy wielodzielnej prostokątnej o wymiarach 2 x 6 (wzór 3.6.6):**

$$C_{max} = \frac{\sqrt{\frac{k-1}{k}} + \sqrt{\frac{r-1}{r}}}{2} = \frac{\sqrt{\frac{6-1}{6}} + \sqrt{\frac{2-1}{2}}}{2} = 0{,}81$$

3. **Obliczam skorygowaną wartość współczynnika $C_{kor_{xy}}$ (wzór 3.6.7):**

$$C_{kor_{xy}} = \frac{C_{xy}}{C_{max}} = \frac{0{,}24}{0{,}81} = 0{,}30$$

$0{,}2 < |0{,}30| \leq 0{,}4$ – słaba zależność,

**Interpretacja:** Pomiędzy płcią i poziomem wykształcenia zachodzi słaba zależność korelacyjna ($C_{kor_{xy}} = 0{,}30$).

# 4. Analiza regresji liniowej

## 4.1. Podstawowe pojęcia

Badanie zależności korelacyjnej pozwala na określenie kierunku i siły pomiędzy badanymi cechami. Celem regresji jest obliczenie wartości średnich **zmiennej zależnej** (*Y*) na podstawie kształtowania się średnich wielkości **zmiennych niezależnych** (*X*). Najpopularniejszym równaniem matematycznym stosowanym w analizie regresji jest funkcja liniowa o postaci:

$$Y_i = \alpha + \beta X_i + \varepsilon_i, \qquad i = (1,2,\dots,n) \qquad (4.1.1)$$

gdzie:
$Y_i$ – zmienna zależna (objaśniana),
$\alpha$ – wyraz wolny,
$\beta$ – współczynnik regresji (parametr strukturalny),
$x_i$ – zmienna niezależna (objaśniająca),
$\varepsilon_i$ – składnik losowy (przypadkowy, resztowy).

W badaniach statystycznych parametry $\beta$ stojące przy zmiennych *X* są nieznane i należy je oszacować. Parametry $\beta$ określają o ile przeciętnie zmieni się wartość zmiennej zależnej *Y*, gdy zmienna niezależna *X* wzrośnie o jednostkę przy założeniu stałości pozostałych zmiennych objaśniających. Parametr $\alpha$, jest interpretowany jako średni poziom zmiennej *Y* gdy pozostałe zmienne objaśniające *X* przyjmują wartość zero. Graficzną prezentację funkcji regresji zaprezentowano na rys. 4.1.1.

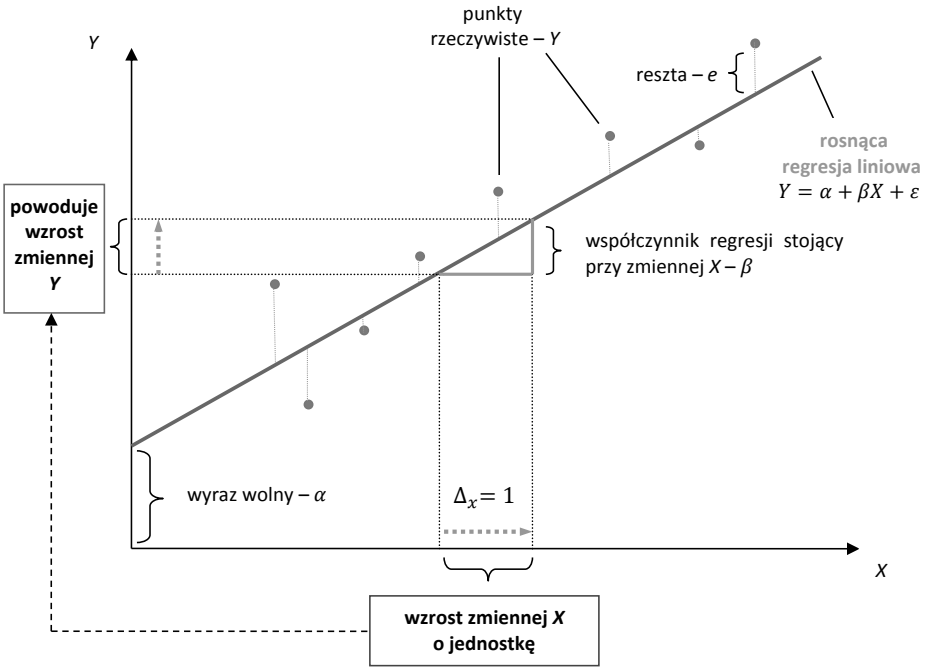
Składnik losowy $\varepsilon$ pełni rolę błędu przypadkowego, zakłócającego. Składnik losowy w regresji występuje ze względu na zbyt skomplikowane zależności występujące w rzeczywistym świecie. Nie jesteśmy w stanie uwzględnić w równaniu wszystkich zmiennych objaśniających *X*. Wynika to, często z braku dostępności danych statystycznych, jak również ze względu na nieprzewidywalny charakter różnorodnych zjawisk np.: warunki pogodowe.

W analizie regresji składnik losowy jest wyrażony poprzez przypadkowe odchylenia – *e* (reszty), pomiędzy wartościami empirycznymi (wartościami rzeczywistymi) – *Y*, od wartości teoretycznych obliczonych na podstawie regresji – $\hat{Y}$.

Procedura postępowania w analizie regresji liniowej jest następująca:

a) badanie zależności korelacyjnej pomiędzy *X* i *Y*,
b) oszacowanie nieznanych parametrów równania regresji,

c) weryfikacja regresji (badanie stopnia dopasowania regresji do danych rzeczywistych).




**Rysunek 4.1.1. Interpretacja graficzna regresji liniowej**
Źródło: opracowane na podstawie M. Piłatowska., *Repetytorium ze statystyki*, PWN, Warszawa 2016, s. 108.

## 4.2. Estymacja parametrów

Parametry regresji liniowej szacuje się **klasyczną metoda najmniejszych kwadratów** (KMNK). W metodzie KMNK zakłada się, że suma kwadratów odchyleń wartości zmiennej objaśnianej – $y_i$ od jej wartości teoretycznych – $\hat{y}_i$ obliczonych na podstawie regresji była jak najmniejsza, a więc dąży do minimum:

$$\sum_{i=1}^{n}(y_i - \hat{y}_i)^2 = \sum_{i=1}^{n} e_i^2 \rightarrow min \tag{4.2.1}$$

gdzie:

$y_i$ – wartości empiryczne (rzeczywiste) zmiennej objaśnianej $Y$,
$\hat{y}_i$ – wartości teoretyczne (wyznaczone na podstawie regresji) zmiennej objaśnianej $Y$,
$e_i$ – reszty regresji obliczone według wzoru:

$$e_i = y_i - \hat{y}_i \tag{4.2.2}$$

Dla funkcji liniowej o postaci:

$$y = \boldsymbol{a} + \boldsymbol{b}x \tag{4.2.3}$$

Warunek (wzór 4.2.1) można zapisać w postaci funkcji kryterium:

$$\sum_{i=1}^{n}(y_i - a - bx)^2 \rightarrow min \tag{4.2.4}$$

Znalezienie minimum dla funkcji kwadratowej (wzór 4.2.4) wymaga obliczenia pochodnych cząstkowych otrzymując tym samym następujący układ równań:

$$\begin{cases} \sum_{i=1}^{n} y_i = na + b\sum_{i=1}^{n} x_i, \\ \sum_{i=1}^{n} x_i y_i = a\sum_{i=1}^{n} x_i + b\sum_{i=1}^{n} {x_i}^2 \end{cases} \tag{4.2.5}$$

Po dokonaniu odpowiednich przekształceń otrzymujemy **oceny parametrów regresji *b*** i ***a***:

$$\boldsymbol{b} = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sum_{i=1}^{n}(x_i - \bar{x})^2} \tag{4.2.6}$$

$$\boldsymbol{a} = \bar{y} - \boldsymbol{b}\bar{x} \tag{4.2.7}$$

gdzie:

$x_i$ – wartość zmiennej *X*,
$y_i$ – wartość zmiennej *Y*,
$\bar{x}$ – średnia arytmetyczna zmiennej *X*,
$\bar{y}$ – średnia arytmetyczna zmiennej *Y*.

Ocenę parametru ***b*** regresji liniowej można obliczyć również metodą pośrednią, wykorzystując relację współczynnika korelacji i odchyleń standardowych dla zmiennych *X* i *Y*:

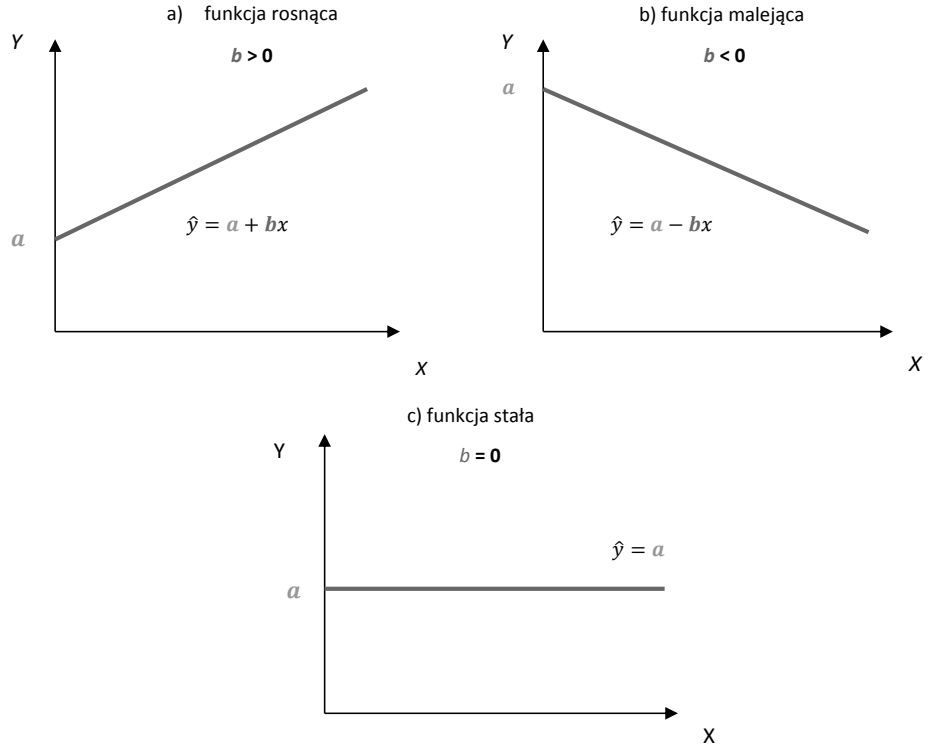
$$\boldsymbol{b} = r_{xy} \cdot \frac{S_y}{S_x} \tag{4.2.8}$$

Współczynniki regresji ***b*** może przyjmować dowolne wartości, wówczas zmienia się jego interpretacja (rys. 4.2.1):

- dodatni (+) współczynnik regresji ***b*** informuje, że **wzrost** zmiennej *X* o jednostkę, powoduje średni **wzrost** zmiennej *Y (o tyle ile wynosi wartość bezwzględna parametru b)*, jest to funkcja rosnąca ($b > 0$),

- ujemny (–) współczynnik regresji ***b*** informuje, że **wzrost** zmiennej *X* o jednostkę, powoduje średni **spadek** zmiennej *Y (o tyle ile wynosi wartość bezwzględna parametru b)*, jest to funkcja malejąca ($b < 0$),
- współczynnik ***b*** równy zero informuje, że zmienna *X* nie ma żadnego wpływu na zmienną *Y*, jest to funkcja stała ($b = 0$).

Dodatnia korelacja liniowa Pearsona między zmiennymi *X* i *Y* powinna być odzwierciedlona w dodatnim współczynniku regresji. I odwrotnie, ujemna zależność ujemnym współczynnikiem.




**Rysunek 4.2.1. Regresja liniowa**
Źródło: opracowanie własne.

**Błędy standardowe parametrów regresji** obliczamy z następujących wzorów:

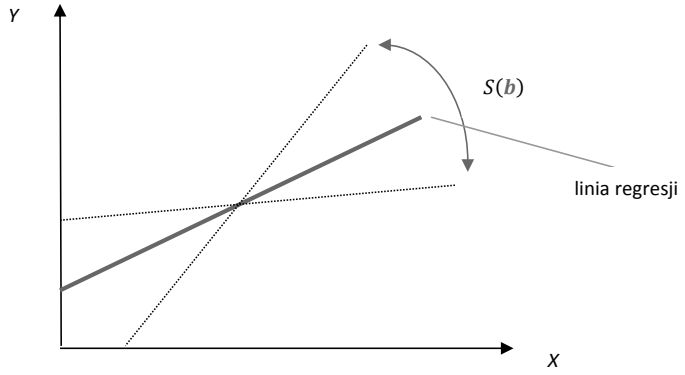
$$S(\boldsymbol{a}) = \sqrt{\frac{S_e^2 \cdot \sum_{i=1}^{n} x_i^2}{n \cdot \sum_{i=1}^{n} (x_i - \bar{x}^2)}} \tag{4.2.9}$$

$$S(\boldsymbol{b}) = \frac{S_e}{\sqrt{\sum_{i=1}^{n} x_i^2 - n\bar{x}^2}} \tag{4.2.10}$$

gdzie:

$n$ – liczebność próby losowej,
$S_e$– błąd standardowy reszt (wzór 4.3.1),
$x_i$ – wartość zmiennej $X$,
$\bar{x}$ – średnia arytmetyczna zmiennej $X$,

Średnie błędy ocen informują, **o ile średnio (±) wahają się (odchylają się) oceny parametrów strukturalnych regresji od ich wartości prawdziwych** (rys. 4.2.2).




**Rysunek 4.2.2. Graficzna interpretacja średniego błędu oceny – $S(\boldsymbol{b})$**
Źródło: opracowane na podstawie D. Strahl, E. Sobczak, M. Markowska, B. Bal-Domańska, *Modelowanie ekonometryczne z Excelem. Materiały pomocnicze do laboratoriów z ekonometrii*, UE, Wrocław 2015, s. 112.

Zmiana wartości oceny parametru $\boldsymbol{b}$ decyduje o kącie nachylenia linii regresji względem osi $X$. Często wyraz wolny $\boldsymbol{a}$ przyjmuje wartości ujemne, co rzadko daje się merytorycznie zinterpretować.

## 4.3. Miary dopasowania

Po oszacowaniu regresji liniowej należy poddać ją ocenie dopasowania (dokładności) do danych rzeczywistych. W tym celu obliczamy parametry struktury stochastycznej:

a) odchylenie standardowe składnika resztowego – $S_e$,
b) współczynnik zmienności resztowej – $V_e$,

c) współczynnik zbieżności – $\varphi^2$,

d) współczynnik determinacji – $R^2$.

**Odchylenie standardowe składnika resztowego ($S_e$)** jest pierwiastkiem kwadratowym ze średniej arytmetycznej kwadratów odchyleń danych empirycznych od teoretycznych:

$$S_e = \sqrt{\frac{\sum_{t=1}^{n} e_i^2}{n-2}} \tag{4.3.1}$$

gdzie:

$n$ – liczebność próby losowej,

$e_i$ – reszty regresji (wzór 4.2.2).

Wartość odchylenia $S_e$ można również obliczyć według wzoru:

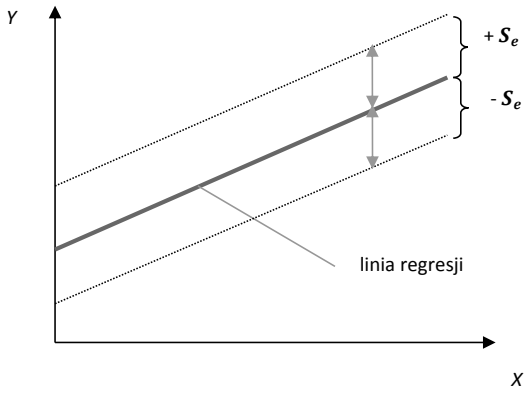
$$S_e = S_y\sqrt{1 - r_{xy}^2} \tag{4.3.2}$$

gdzie:

$S_y$ – odchylenie standardowe zmiennej $Y$,

$r_{xy}^2$ – kwadrat współczynnika korelacji zmiennych $X$ i $Y$.

Odchylenie standardowe składnika resztowego **informuje, o ile średnio (±) odchylają się wartości rzeczywiste zmiennej objaśnianej od teoretycznych obliczonych na podstawie regresji** (rys. 4.3.1). Im mniejszy jest błąd składnika resztowego, tym lepsze dopasowanie modelu do danych rzeczywistych.




**Rysunek 4.3.1. Graficzna interpretacja odchylenia standardowego składnika resztowego – $S_e$**
Źródło: opracowane na podstawie H. Augustyniak, *Statystyka opisowa z elementami demografii*, Poznań 2002 s. 89.

**Współczynnik zmienności resztowej ($V_e$)** jest to udział odchylenie standardowe składnika resztowego $S_e$ w wartości średniej zmiennej objaśnianej $\bar{y}$ :

$$V_e = \frac{S_e}{\bar{y}} \cdot 100 \tag{4.3.3}$$

Współczynnik zmienności resztowej **informuje o tym, jaką część średniej wartości zmiennej objaśnianej stanowi jej odchylenie standardowe reszt.** W praktyce zależy nam, aby współczynnik zmienności był możliwie jak najmniejszy. Regresję przyjmuje się za dopuszczalną, jeśli $V_e \leq V_g$. Wartość graniczna $V_g$ ustalana jest w sposób umowny. Najczęściej przyjmuje się poziom 10% lub 15%.

**Współczynnik zbieżności – zgodności ($\varphi^2$)** można zapisać w postaci:

$$\varphi^2 = 1 - R^2 \tag{4.3.4}$$

lub:

$$\varphi^2 = \frac{\sum_{t=1}^{n}(y_i - \hat{y}_i)^2}{\sum_{t=1}^{n}(y_i - \bar{y})^2} \tag{4.3.5}$$

gdzie:

$R^2$– współczynnik determinacji (wzory 4.3.6 lub 4.3.7),
$y_i$ – wartość rzeczywista zmiennej *Y*,
$\hat{y}_i$ – wartość teoretyczna zmiennej *Y*,
$\bar{y}$ – średnia arytmetyczna zmiennej *Y*.

Współczynnik zbieżności $\varphi^2 \cdot 100\%$ **informuje ile % zmienności zmiennej *Y* nie została wyjaśniona przez zmienne regresji czyli jaką część zmienności zmiennej *Y* stanowi zmienność odchyleń losowych.** Im wartość współczynnika zbieżności jest bliższa 0, tym szacowana regresja jest lepiej dopasowana do wartości rzeczywistych zmiennej *Y*. Przyjmuje on wartość z przedziału [0, 1] lub [0%, 100%].

**Współczynnik determinacji ($R^2$) jest najważniejszą miarą dopasowania regresji** i oblicza się go następująco:

$$R^2 = 1 - \varphi^2 \tag{4.3.6}$$

lub:

$$R^2 = \frac{\sum_{i=1}^{n}(\hat{y}_i - \bar{y})^2}{\sum_{i=1}^{n}(y_i - \bar{y})^2} \tag{4.3.7}$$

gdzie:

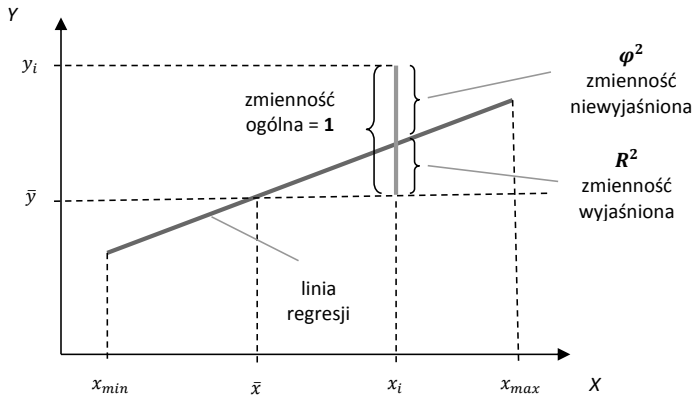
$\varphi^2$ – współczynnik zbieżności (wzory 4.3.4 lub 4.3.5),
$y_i$ – wartość rzeczywista zmiennej *Y*,
$\hat{y}_i$ – wartość teoretyczna zmiennej *Y*,
$\bar{y}$ – średnia arytmetyczna zmiennej *Y*.

Współczynnik determinacji $R^2 \cdot 100\%$ **informuje ile % zmienności zmiennej *Y* została wyjaśniona przez zmienne regresji.** Im jego wartość bliższa 1 tym dopasowanie regresji jest lepsze. Przyjmuje on wartość z przedziału [0, 1] lub [0%, 100%]. Kwadrat współczynnika korelacji $r_{xy}^2$ daje nam wartość współczynnika determinacji $R^2$:

$$R^2 = r_{xy}^2 \tag{4.3.8}$$

Całkowita zmienność zmiennej *Y* stanowi sumę zmienności niewyjaśnionej – $\varphi^2$ i wyjaśnionej – $R^2$ na podstawie oszacowanej regresji liniowej i możemy ją zapisać w postaci (rys. 4.3.1):

$$\varphi^2 + R^2 = 1 \tag{4.3.9}$$




**Rysunek 4.3.1. Graficzna interpretacja zmienności ogólnej zmiennej *Y* oraz zmienności wyjaśnionej i niewyjaśnionej funkcją regresji liniowej *Y* od *X***
Źródło: opracowane na podstawie M., Sobczyk, *Statystyka*, PWN, Warszawa 2016, s. 265.

## Przykłady

**Przykład 4.1.**

Wyznacz współczynniki liniowej funkcji regresji na podstawie danych z przykładu 3.1. Oblicz i zinterpretuj miary dopasowania „dobroci" do danych rzeczywistych.

**Rozwiązanie**

$y_i$ – liczba pracujących na 1000 ludności (w ‰),

$x_i$ – przeciętne miesięczne wynagrodzenia brutto w gospodarce narodowej (w zł).

Średnie arytmetyczne cech $x_i$ i $y_i$ wynoszą odpowiednio:

$$\bar{x}_i = 3999{,}8 \text{ zł},$$

$$\bar{y}_i = 383{,}2‰.$$

Obliczenia pomocnicze:

| Zmienna objaśniana $y_i$ | Zmienna objaśniająca $x_i$ | $x_i - \bar{x}$ | $y_i - \bar{y}$ | $(x_i - \bar{x})(y_i - \bar{y})$ | $(x_i - \bar{x})^2$ |
|---|---|---|---|---|---|
| 391,2 | 4400,05 | 400,25 | 8,0 | 3202,0 | 160200,1 |
| 354,5 | 3717,21 | -282,59 | -28,7 | 8110,3 | 79857,1 |
| 396,3 | 3824,28 | -175,52 | 13,1 | -2299,3 | 30807,3 |
| 345,9 | 3754,54 | -245,26 | -37,3 | 9148,2 | 60152,5 |
| 405,8 | 3925,97 | -73,83 | 22,6 | -1668,6 | 5450,9 |
| 417,0 | 4097,35 | 97,55 | 33,8 | 3297,2 | 9516,0 |
| 486,2 | 5219,09 | 1219,29 | 103,0 | 125586,9 | 1486668,1 |
| 335,8 | 3923,58 | -76,22 | -47,4 | 3612,8 | 5809,5 |
| 400,3 | 3684,71 | -315,09 | 17,1 | -5388,0 | 99281,7 |
| 364,8 | 3815,23 | -184,57 | -18,4 | 3396,1 | 34066,1 |
| 376,2 | 4211,69 | 211,89 | -7,0 | -1483,2 | 44897,4 |
| 383,6 | 4247,44 | 247,64 | 0,4 | 99,1 | 61325,6 |
| 385,7 | 3705,65 | -294,15 | 2,5 | -735,4 | 86524,2 |
| 315,8 | 3641,32 | -358,48 | -67,4 | 24161,6 | 128507,9 |
| 441,9 | 3937,81 | -61,99 | 58,7 | -3638,8 | 3842,8 |
| 330,4 | 3890,86 | -108,94 | -52,8 | 5752,0 | 11867,9 |
| - | - | - | - | **171152,9** | **2308775,1** |

Źródło: opracowanie własne.

1. **Obliczam współczynniki regresji liniowej (wzory 4.2.6 i 4.2.7):**

$$\hat{y}_i = a + bx_i$$

$$b = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sum_{i=1}^{n}(x_i - \bar{x})^2} = \frac{171152{,}9}{2308775{,}1} = 0{,}074$$

$$a = \bar{y} - b\bar{x} = 383{,}2 - 0{,}074 \cdot 3999{,}8 = 87{,}215$$

Ocenę parametru ***b*** regresji liniowej można obliczyć również metodą pośrednią (wzór 4.2.8). Współczynnik korelacji Pearsona pomiędzy zmiennymi *X* i *Y* wynosi $r_{xy} = 0{,}67$ (przykład 3.1). Odchylenia standardowe $S_x$ i $S_y$ obliczymy poniżej:

Obliczenia pomocnicze:

| $x_i$ | $y_i$ | $x_i - \bar{x}$ | $y_i - \bar{y}$ | $(x_i - \bar{x})^2$ | $(y_i - \bar{y})^2$ |
|---|---|---|---|---|---|
| 4400,05 | 391,2 | 400,25 | 8,0 | 160200,1 | 64,0 |
| 3717,21 | 354,5 | -282,59 | -28,7 | 79857,1 | 823,7 |
| 3824,28 | 396,3 | -175,52 | 13,1 | 30807,3 | 171,6 |
| 3754,54 | 345,9 | -245,26 | -37,3 | 60152,5 | 1391,3 |
| 3925,97 | 405,8 | -73,83 | 22,6 | 5450,9 | 510,8 |
| 4097,35 | 417,0 | 97,55 | 33,8 | 9516,0 | 1142,4 |
| 5219,09 | 486,2 | 1219,29 | 103,0 | 1486668,1 | 10609,0 |
| 3923,58 | 335,8 | -76,22 | -47,4 | 5809,5 | 2246,8 |
| 3684,71 | 400,3 | -315,09 | 17,1 | 99281,7 | 292,4 |
| 3815,23 | 364,8 | -184,57 | -18,4 | 34066,1 | 338,6 |
| 4211,69 | 376,2 | 211,89 | -7,0 | 44897,4 | 49,0 |
| 4247,44 | 383,6 | 247,64 | 0,4 | 61325,6 | 0,2 |
| 3705,65 | 385,7 | -294,15 | 2,5 | 86524,2 | 6,3 |
| 3641,32 | 315,8 | -358,48 | -67,4 | 128507,9 | 4542,8 |
| 3937,81 | 441,9 | -61,99 | 58,7 | 3842,8 | 3445,7 |
| 3890,86 | 330,4 | -108,94 | -52,8 | 11867,9 | 2787,8 |
| - | - | - | - | **2308775,1** | **28422,4** |

Źródło: opracowanie własne.

$$S_x = \sqrt{\frac{\sum_{i=1}^{N}(x_i - \bar{x})^2}{N}} = \sqrt{\frac{2308775{,}1}{16}} = 379{,}87$$

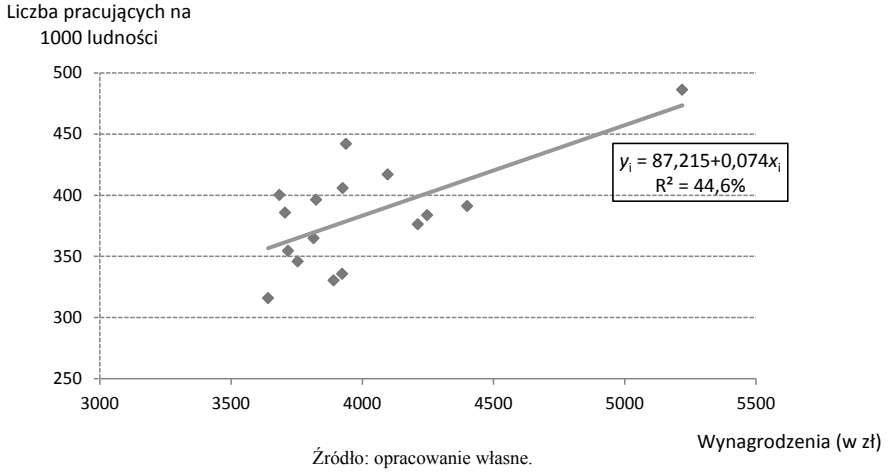
$$S_y = \sqrt{\frac{\sum_{i=1}^{N}(y_i - \bar{y})^2}{N}} = \sqrt{\frac{28422{,}4}{16}} = 42{,}15$$

$$b = r_{xy} \cdot \frac{S_y}{S_x} = 0{,}67 \cdot \frac{42{,}15}{379{,}87} = 0{,}074$$

Regresja liniowa przyjmuje postać:

$$\hat{y}_i = 87{,}215 + 0{,}074x_i$$

**Zależność między liczbą pracujących na 1000 ludności a przeciętnymi miesięcznymi wynagrodzeniami brutto w gospodarce narodowej według województw w 2017 r.**




Źródło: opracowanie własne.

**Interpretacja współczynników regresji *a* i *b***:

- $a = 87{,}215$ – współczynnik $a$ informuje o średnim poziomie liczby pracujących na 1000 ludności w 2017 r. pod warunkiem, że zmienna objaśniająca $x_i$ przyjmuje wartość zero,
- $b = 0{,}074$ – współczynnik $b$ jest dodatni oznacza to, że jeżeli zmienna $x_i$ wzrośnie o jednostkę (o 1 zł), to nastąpi wzrost zmiennej $y_i$ średnio o 0,074‰.

Współczynnik $b$ stojący przy zmiennej $x_i$ potwierdza dodatnią zależność korelacyjną pomiędzy cechami $x_i$ i $y_i$.

**2. Obliczam błędy standardowe parametrów regresji (wzory 4.2.9 i 4.2.10):**

$S_e = 33{,}5$ – wartość obliczona w części miary dopasowania regresji pkt a).

$$S_e^2 = 33{,}5^2 = 1122{,}25$$

$$S(a) = \sqrt{\frac{S_e^2 \cdot \sum_{i=1}^{n} x_i^2}{n \cdot \sum_{i=1}^{n}(x_i - \bar{x}^2)}} = \sqrt{\frac{1122{,}25 \cdot 258283015{,}5}{16 \cdot 2308775{,}1}} = 88{,}6$$

$$S(b) = \frac{S_e}{\sqrt{\sum_{i=1}^{n} x_i^2 - n\bar{x}^2}} = \frac{33{,}5}{\sqrt{258283015{,}5 - 16 \cdot 3999{,}8^2}} = 0{,}02$$

Obliczenia pomocnicze:

| $x_i$ | $x_i^2$ | $x_i - \bar{x}$ | $(x_i - \bar{x})^2$ |
|---|---|---|---|
| 4400,05 | 19360440,0 | 400,25 | 160200,1 |
| 3717,21 | 13817650,2 | -282,59 | 79857,1 |
| 3824,28 | 14625117,5 | -175,52 | 30807,3 |
| 3754,54 | 14096570,6 | -245,26 | 60152,5 |
| 3925,97 | 15413240,4 | -73,83 | 5450,9 |
| 4097,35 | 16788277,0 | 97,55 | 9516,0 |
| 5219,09 | 27238900,4 | 1219,29 | 1486668,1 |
| 3923,58 | 15394480,0 | -76,22 | 5809,5 |
| 3684,71 | 13577087,8 | -315,09 | 99281,7 |
| 3815,23 | 14555980,0 | -184,57 | 34066,1 |
| 4211,69 | 17738332,7 | 211,89 | 44897,4 |
| 4247,44 | 18040746,6 | 247,64 | 61325,6 |
| 3705,65 | 13731841,9 | -294,15 | 86524,2 |
| 3641,32 | 13259211,3 | -358,48 | 128507,9 |
| 3937,81 | 15506347,6 | -61,99 | 3842,8 |
| 3890,86 | 15138791,5 | -108,94 | 11867,9 |
| - | **258283015,5** | - | **2308775,1** |

Źródło: opracowanie własne.

Błędy standardowe parametrów pokazujemy w nawiasach pod parametrami:

$$\hat{y}_i = \underset{(\pm 88,6)}{87,215} + \underset{(\pm 0,02)}{0,074 x_i}$$

**Interpretacja błędów standardowych ocen współczynników regresji *a* i *b*:**

- $S(a) = 88,6$ – współczynnik $a = 87,215$ odchyla się średnio o ± 88,6,
- $S(b) = 0,02$ – współczynnik $b = 0,074$ odchyla się średnio o ± 0,02.

**Miary dopasowania regresji liniowej:**

**a) odchylenie standardowe składnika resztowego (wzór 4.3.1):**

$$S_e = \sqrt{\frac{\sum_{t=1}^{n} e_i^2}{n-2}} = \sqrt{\frac{15743,7}{16-2}} = 33,5$$

Obliczenia pomocnicze:

| Wartości rzeczywiste $y_i$ | Zmienna objaśniająca $x_i$ | Wartości teoretyczne: $\hat{y}_i = 87{,}215 + 0{,}074x_i$ | Reszty $e_i = y_i - \hat{y}_i$ | $e_i^2$ |
|---|---|---|---|---|
| 391,2 | 4400,05 | 87,215+(0,074 · 4400,05) = 412,8 | -21,6 | 466,6 |
| 354,5 | 3717,21 | 87,215+(0,074 · 3717,21) = 362,3 | -7,8 | 60,8 |
| 396,3 | 3824,28 | 87,215+(0,074 · 3824,28) = 370,2 | 26,1 | 681,2 |
| 345,9 | 3754,54 | 87,215+(0,074 · 3754,54) = 365,1 | -19,2 | 368,6 |
| 405,8 | 3925,97 | 87,215+(0,074 · 3925,97) = 377,7 | 28,1 | 789,6 |
| 417,0 | 4097,35 | 87,215+(0,074 · 4097,35) = 390,4 | 26,6 | 707,6 |
| 486,2 | 5219,09 | 87,215+(0,074 · 5219,09) = 473,4 | 12,8 | 163,8 |
| 335,8 | 3923,58 | 87,215+(0,074 · 3923,58) = 377,6 | -41,8 | 1747,2 |
| 400,3 | 3684,71 | 87,215+(0,074 · 3684,71) = 359,9 | 40,4 | 1632,2 |
| 364,8 | 3815,23 | 87,215+(0,074 · 3815,23) = 369,5 | -4,7 | 22,1 |
| 376,2 | 4211,69 | 87,215+(0,074 · 4211,69) = 398,9 | -22,7 | 515,3 |
| 383,6 | 4247,44 | 87,215+(0,074 · 4247,44) = 401,5 | -17,9 | 320,4 |
| 385,7 | 3705,65 | 87,215+(0,074 · 3705,65) = 361,4 | 24,3 | 590,5 |
| 315,8 | 3641,32 | 87,215+(0,074 · 3641,32) = 356,7 | -40,9 | 1672,8 |
| 441,9 | 3937,81 | 87,215+(0,074 · 3937,81) = 378,6 | 63,3 | 4006,9 |
| 330,4 | 3890,86 | 87,215+(0,074 · 3890,86) = 375,1 | -44,7 | 1998,1 |
| - | - | - | - | **15743,7** |

Źródło: opracowanie własne.

**Interpretacja:** Odchylenie standardowe składnika resztowego informuje, że liczba pracujących na 1000 ludności według województw odchyla się od wartości teoretycznych obliczonych na podstawie regresji średnio o ± 33,5‰.

**b) współczynnik zmienności resztowej (wzór 4.3.3):**

$$V_e = \frac{S_e}{\bar{y}} \cdot 100 = \frac{33{,}5}{383{,}2} \cdot 100 = 8{,}7\%$$

$$8{,}7\% \leq 10{,}0\%$$

$$V_e \leq V_g$$

**Interpretacja:** Odchylenie standardowe składnika resztowego stanowi 8,7% średniej arytmetycznej liczby pracujących na 1000 ludności co świadczy o dobrym dopasowaniu regresji do danych rzeczywistych.

**c) współczynnik zbieżności (wzór 4.3.5):**

Obliczenia pomocnicze:

| $y_i$ | $\hat{y}_i$ | $y_i - \hat{y}_i$ | $(y_i - \hat{y}_i)^2$ | $y_i - \bar{y}$ | $(y_i - \bar{y})^2$ |
|---|---|---|---|---|---|
| 391,2 | 412,8 | -21,6 | 466,6 | 8,0 | 64,0 |
| 354,5 | 362,3 | -7,8 | 60,8 | -28,7 | 823,7 |
| 396,3 | 370,2 | 26,1 | 681,2 | 13,1 | 171,6 |
| 345,9 | 365,1 | -19,2 | 368,6 | -37,3 | 1391,3 |
| 405,8 | 377,7 | 28,1 | 789,6 | 22,6 | 510,8 |
| 417,0 | 390,4 | 26,6 | 707,6 | 33,8 | 1142,4 |
| 486,2 | 473,4 | 12,8 | 163,8 | 103,0 | 10609,0 |
| 335,8 | 377,6 | -41,8 | 1747,2 | -47,4 | 2246,8 |
| 400,3 | 359,9 | 40,4 | 1632,2 | 17,1 | 292,4 |
| 364,8 | 369,5 | -4,7 | 22,1 | -18,4 | 338,6 |
| 376,2 | 398,9 | -22,7 | 515,3 | -7,0 | 49,0 |
| 383,6 | 401,5 | -17,9 | 320,4 | 0,4 | 0,2 |
| 385,7 | 361,4 | 24,3 | 590,5 | 2,5 | 6,3 |
| 315,8 | 356,7 | -40,9 | 1672,8 | -67,4 | 4542,8 |
| 441,9 | 378,6 | 63,3 | 4006,9 | 58,7 | 3445,7 |
| 330,4 | 375,1 | -44,7 | 1998,1 | -52,8 | 2787,8 |
| - | - | - | **15743,7** | - | **28422,4** |

Źródło: opracowanie własne.

$$\varphi^2 = \frac{\sum_{t=1}^{n}(y_i - \hat{y}_i)^2}{\sum_{t=1}^{n}(y_i - \bar{y})^2} = \frac{15743,7}{28422,4} = 0,554$$

$$\varphi^2 \cdot 100\% = 0,554 \cdot 100\% = 55,4\%$$

**Interpretacja:** Około 55,0% zmienności liczby pracujących na 1000 ludności nie zostało wyjaśnione na podstawie regresji (przez zmienną $x_i$). Powyższa zmienność jest efektem pozostałych czynników o charakterze losowymi i nielosowym.

**d) współczynnik determinacji (wzór 4.3.6):**

$$R^2 = 1 - \varphi^2 = 1 - 0,554 = 0,446$$
$$R^2 \cdot 100\% = 0,446 \cdot 100\% = 44,6\%$$

**Interpretacja:** Około 45,0% zmienności liczby pracujących na 1000 ludności zostało wyjaśnione na podstawie regresji (przez zmienną $x_i$).

**Przykład 4.2.**

Wyznacz współczynniki liniowej funkcji regresji na podstawie danych przedstawiających zależność pomiędzy przeciętnymi miesięcznymi wynagrodzeniami brutto a stopą bezrobocia rejestrowanego według województw w 2017 r.

| Województwa | Stopa bezrobocia rejestrowanego (w %) $y_i$ | Przeciętne miesięczne wynagrodzenia brutto w gospodarce narodowej (w zł) $x_i$ |
|---|---|---|
| | 2017 r. | |
| Dolnośląskie | 5,7 | 4400,05 |
| Kujawsko-pomorskie | 9,9 | 3717,21 |
| Lubelskie | 8,8 | 3824,28 |
| lubuskie | 6,5 | 3754,54 |
| Łódzkie | 6,7 | 3925,97 |
| Małopolskie | 5,3 | 4097,35 |
| Mazowieckie | 5,6 | 5219,09 |
| Opolskie | 7,3 | 3923,58 |
| Podkarpackie | 9,6 | 3684,71 |
| Podlaskie | 8,5 | 3815,23 |
| Pomorskie | 5,4 | 4211,69 |
| Śląskie | 5,1 | 4247,44 |
| Świętokrzyskie | 8,8 | 3705,65 |
| Warmińsko-mazurskie | 11,7 | 3641,32 |
| Wielkopolskie | 3,7 | 3937,81 |
| Zachodniopomorskie | 8,5 | 3890,86 |

Źródło: Bank Danych Lokalnych GUS.

Oblicz i zinterpretuj miary dopasowania „dobroci" do danych rzeczywistych.

**Rozwiązanie**

$y_i$ – stopa bezrobocia rejestrowanego (w %),

$x_i$ – przeciętne miesięczne wynagrodzenia brutto w gospodarce narodowej (w zł).

Średnie arytmetyczne cech $x_i$ i $y_i$ wynoszą odpowiednio:

$$\bar{x}_i = 3999{,}8 \text{ zł},$$

$$\bar{y}_i = 7{,}3\%.$$

1. **Obliczam współczynniki regresji liniowej (wzory 4.2.6 i 4.2.7):**

$$\hat{y}_i = a + bx_i$$

$$b = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sum_{i=1}^{n}(x_i - \bar{x})^2} = \frac{-7485{,}4}{2308775{,}1} = -0{,}003242$$

$$a = \bar{y} - b\bar{x} = 7{,}3 - (-0{,}003242 \cdot 3999{,}8) = 20{,}26735$$

Obliczenia pomocnicze:

| Zmienna objaśniana $y_i$ | Zmienna objaśniająca $x_i$ | $x_i - \bar{x}$ | $y_i - \bar{y}$ | $(x_i - \bar{x})(y_i - \bar{y})$ | $(x_i - \bar{x})^2$ |
|---|---|---|---|---|---|
| 5,7 | 4400,05 | 400,25 | -1,6 | -640,4 | 160200,1 |
| 9,9 | 3717,21 | -282,59 | 2,6 | -734,7 | 79857,1 |
| 8,8 | 3824,28 | -175,52 | 1,5 | -263,3 | 30807,3 |
| 6,5 | 3754,54 | -245,26 | -0,8 | 196,2 | 60152,5 |
| 6,7 | 3925,97 | -73,83 | -0,6 | 44,3 | 5450,9 |
| 5,3 | 4097,35 | 97,55 | -2,0 | -195,1 | 9516,0 |
| 5,6 | 5219,09 | 1219,29 | -1,7 | -2072,8 | 1486668,1 |
| 7,3 | 3923,58 | -76,22 | 0,0 | 0,0 | 5809,5 |
| 9,6 | 3684,71 | -315,09 | 2,3 | -724,7 | 99281,7 |
| 8,5 | 3815,23 | -184,57 | 1,2 | -221,5 | 34066,1 |
| 5,4 | 4211,69 | 211,89 | -1,9 | -402,6 | 44897,4 |
| 5,1 | 4247,44 | 247,64 | -2,2 | -544,8 | 61325,6 |
| 8,8 | 3705,65 | -294,15 | 1,5 | -441,2 | 86524,2 |
| 11,7 | 3641,32 | -358,48 | 4,4 | -1577,3 | 128507,9 |
| 3,7 | 3937,81 | -61,99 | -3,6 | 223,2 | 3842,8 |
| 8,5 | 3890,86 | -108,94 | 1,2 | -130,7 | 11867,9 |
| - | - | - | - | **-7485,4** | **2308775,1** |

Źródło: opracowanie własne.

Ocenę parametru ***b*** regresji liniowej można obliczyć również metodą pośrednią (wzór 4.2.8) ale najpierw należy obliczyć współczynnik korelacji Pearsona pomiędzy zmiennymi *X* i *Y* oraz odchylenia standardowe $S_x$ i $S_y$.

$$r_{xy} = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_{i=1}^{n}(x_i - \bar{x})^2 \sum_{i=1}^{n}(y_i - \bar{y})^2}} = \frac{-7485{,}4}{\sqrt{2308775{,}1 \cdot 70{,}8}} = -0{,}5855$$

$0{,}4 < |-0{,}5855| \leq 0{,}6$ – zależność umiarkowana

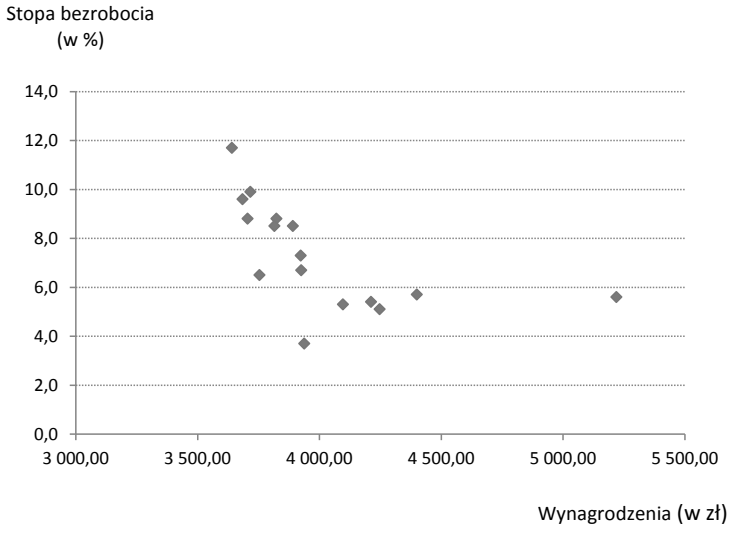
**Interpretacja:** Pomiędzy badanymi cechami zachodzi umiarkowana i ujemna zależność korelacyjna ($r_{xy} = -0{,}5855$). Oznacza to, że wzrost przeciętnego miesięcznego wynagrodzenia brutto w gospodarce narodowej powoduje spadek stopy bezrobocia w województwach w 2017 r.

Obliczenia pomocnicze:

| $x_i$ | $y_i$ | $x_i - \bar{x}$ | $y_i - \bar{y}$ | $(x_i - \bar{x})(y_i - \bar{y})$ | $(x_i - \bar{x})^2$ | $(y_i - \bar{y})^2$ |
|---|---|---|---|---|---|---|
| 4400,05 | 5,7 | 400,25 | -1,6 | -640,4 | 160200,1 | 2,6 |
| 3717,21 | 9,9 | -282,59 | 2,6 | -734,7 | 79857,1 | 6,8 |
| 3824,28 | 8,8 | -175,52 | 1,5 | -263,3 | 30807,3 | 2,3 |
| 3754,54 | 6,5 | -245,26 | -0,8 | 196,2 | 60152,5 | 0,6 |
| 3925,97 | 6,7 | -73,83 | -0,6 | 44,3 | 5450,9 | 0,4 |
| 4097,35 | 5,3 | 97,55 | -2,0 | -195,1 | 9516,0 | 4,0 |
| 5219,09 | 5,6 | 1219,29 | -1,7 | -2072,8 | 1486668,1 | 2,9 |
| 3923,58 | 7,3 | -76,22 | 0,0 | 0,0 | 5809,5 | 0,0 |
| 3684,71 | 9,6 | -315,09 | 2,3 | -724,7 | 99281,7 | 5,3 |
| 3815,23 | 8,5 | -184,57 | 1,2 | -221,5 | 34066,1 | 1,4 |
| 4211,69 | 5,4 | 211,89 | -1,9 | -402,6 | 44897,4 | 3,6 |
| 4247,44 | 5,1 | 247,64 | -2,2 | -544,8 | 61325,6 | 4,8 |
| 3705,65 | 8,8 | -294,15 | 1,5 | -441,2 | 86524,2 | 2,3 |
| 3641,32 | 11,7 | -358,48 | 4,4 | -1577,3 | 128507,9 | 19,4 |
| 3937,81 | 3,7 | -61,99 | -3,6 | 223,2 | 3842,8 | 13,0 |
| 3890,86 | 8,5 | -108,94 | 1,2 | -130,7 | 11867,9 | 1,4 |
| - | - | - | - | **-7485,4** | **2308775,1** | **70,8** |

Źródło: opracowanie własne.

**Diagram korelacyjny (rozrzutu) cech $x_i$ i $y_i$**




Źródło: opracowanie własne.

$$S_x = \sqrt{\frac{\sum_{i=1}^{N}(x_i - \bar{x})^2}{N}} = \sqrt{\frac{2308775{,}1}{16}} = 379{,}87$$

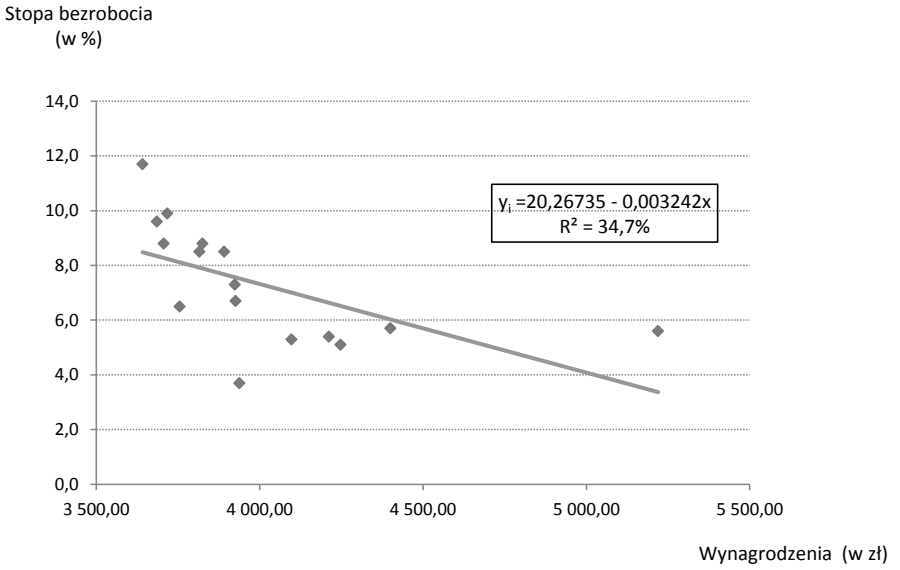
$$S_y = \sqrt{\frac{\sum_{i=1}^{N}(y_i - \bar{y})^2}{N}} = \sqrt{\frac{70{,}8}{16}} = 2{,}1$$

$$b = r_{xy} \cdot \frac{S_y}{S_x} = -0{,}5855 \cdot \frac{2{,}1}{379{,}87} = -0{,}003242$$

Regresja liniowa przyjmuje postać:

$$\hat{y}_i = 20{,}26735 - 0{,}003242x_i$$

**Zależność między stopą bezrobocia rejestrowanego a przeciętnymi miesięcznymi wynagrodzeniami brutto w gospodarce narodowej według województw w 2017 r.**




Źródło: opracowanie własne.

**Interpretacja współczynników regresji *a* i *b***:

- $a = 20{,}26735$ – współczynnik $a$ informuje o średnim poziomie stopy bezrobocia w 2017 r. pod warunkiem, że zmienna objaśniająca $x_i$ przyjmuje wartość zero,
- $b = -0{,}003242$ – współczynnik $b$ jest ujemny oznacza to, że jeżeli zmienna $x_i$ wzrośnie o jednostkę (o 1 zł), to nastąpi spadek zmiennej $y_i$ średnio o 0,003242%.

Współczynnik $b$ stojący przy zmiennej $x_i$ potwierdza ujemną zależność korelacyjną pomiędzy cechami $x_i$ i $y_i$.

**3. Obliczam błędy standardowe parametrów regresji (wzory 4.2.9 i 4.2.10):**

$S_e = 1{,}8$ – wartość obliczona w części miary dopasowania regresji pkt a).

$$S_e^2 = 1{,}8^2 = 3{,}24$$

Obliczenia pomocnicze:

| $x_i$ | $x_i^2$ | $x_i - \bar{x}$ | $(x_i - \bar{x})^2$ |
|---|---|---|---|
| 4400,05 | 19360440,0 | 400,25 | 160200,1 |
| 3717,21 | 13817650,2 | -282,59 | 79857,1 |
| 3824,28 | 14625117,5 | -175,52 | 30807,3 |
| 3754,54 | 14096570,6 | -245,26 | 60152,5 |
| 3925,97 | 15413240,4 | -73,83 | 5450,9 |
| 4097,35 | 16788277,0 | 97,55 | 9516,0 |
| 5219,09 | 27238900,4 | 1219,29 | 1486668,1 |
| 3923,58 | 15394480,0 | -76,22 | 5809,5 |
| 3684,71 | 13577087,8 | -315,09 | 99281,7 |
| 3815,23 | 14555980,0 | -184,57 | 34066,1 |
| 4211,69 | 17738332,7 | 211,89 | 44897,4 |
| 4247,44 | 18040746,6 | 247,64 | 61325,6 |
| 3705,65 | 13731841,9 | -294,15 | 86524,2 |
| 3641,32 | 13259211,3 | -358,48 | 128507,9 |
| 3937,81 | 15506347,6 | -61,99 | 3842,8 |
| 3890,86 | 15138791,5 | -108,94 | 11867,9 |
| - | **258283015,5** | - | **2308775,1** |

Źródło: opracowanie własne.

$$S(a) = \sqrt{\frac{S_e^2 \cdot \sum_{i=1}^{n} x_i^2}{n \cdot \sum_{i=1}^{n} (x_i - \bar{x}^2)}} = \sqrt{\frac{3{,}24 \cdot 258283015{,}5}{16 \cdot 2308775{,}1}} = 4{,}7596$$

$$S(b) = \frac{S_e}{\sqrt{\sum_{i=1}^{n} x_i^2 - n\bar{x}^2}} = \frac{1{,}8}{\sqrt{258283015{,}5 - 16 \cdot 3999{,}8^2}} = 0{,}0012$$

Błędy standardowe parametrów pokazujemy w nawiasach pod parametrami:

$$\hat{y}_i = \underset{(\pm 4{,}7596)}{20{,}26735} - \underset{(\pm 0{,}0012)}{0{,}003242} x_i$$

**Interpretacja błędów standardowych ocen współczynników regresji *a* i *b*:**

- $S(a) = 4{,}7596$ – współczynnik $a = 20{,}26735$ odchyla się średnio o ± 4,7596,
- $S(b) = 0{,}0012$ – współczynnik $b = 0{,}03242$ odchyla się średnio o ± 0,0012.

**Miary dopasowania regresji liniowej:**

**a) odchylenie standardowe składnika resztowego (wzór 4.3.1):**

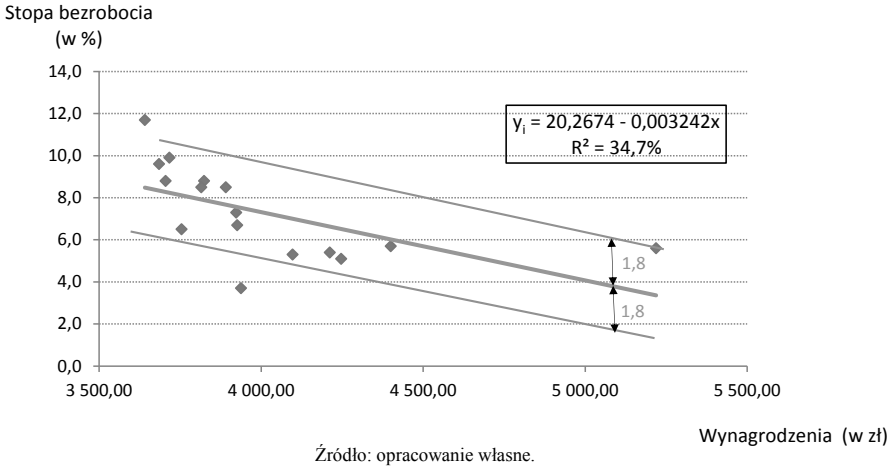
$$S_e = \sqrt{\frac{\sum_{t=1}^{n} e_i^2}{n-2}} = \sqrt{\frac{46{,}23}{16-2}} = 1{,}8$$

Obliczenia pomocnicze:

| Wartości rzeczywiste $y_i$ | Zmienna objaśniająca $x_i$ | Wartości teoretyczne: $\hat{y}_i = 20{,}26735 - 0{,}003242x_i$ | Reszty $e_i = y_i - \hat{y}_i$ | $e_i^2$ |
|---|---|---|---|---|
| 5,7 | 4400,05 | 20,26735-(0,003242 · 4400,05) = 6,0 | -0,3 | 0,09 |
| 9,9 | 3717,21 | 20,26735-(0,003242 · 3717,21) = 8,2 | 1,7 | 2,89 |
| 8,8 | 3824,28 | 20,26735-(0,003242 · 3824,28) = 7,9 | 0,9 | 0,81 |
| 6,5 | 3754,54 | 20,26735-(0,003242 · 3754,54) = 8,1 | -1,6 | 2,56 |
| 6,7 | 3925,97 | 20,26735-(0,003242 · 3925,97) = 7,5 | -0,8 | 0,64 |
| 5,3 | 4097,35 | 20,26735-(0,003242 · 4097,35) = 7,0 | -1,7 | 2,89 |
| 5,6 | 5219,09 | 20,26735-(0,003242 · 5219,09) = 3,3 | 2,3 | 5,29 |
| 7,3 | 3923,58 | 20,26735-(0,003242 · 3923,58) = 7,5 | -0,2 | 0,04 |
| 9,6 | 3684,71 | 20,26735-(0,003242 · 3684,71) = 8,3 | 1,3 | 1,69 |
| 8,5 | 3815,23 | 20,26735-(0,003242 · 3815,23) = 7,9 | 0,6 | 0,36 |
| 5,4 | 4211,69 | 20,26735-(0,003242 · 4211,69) = 6,6 | -1,2 | 1,44 |
| 5,1 | 4247,44 | 20,26735-(0,003242 · 4247,44) = 6,5 | -1,4 | 1,96 |
| 8,8 | 3705,65 | 20,26735-(0,003242 · 3705,65) = 8,3 | 0,5 | 0,25 |
| 11,7 | 3641,32 | 20,26735-(0,003242 · 3641,32) = 8,5 | 3,2 | 10,24 |
| 3,7 | 3937,81 | 20,26735-(0,003242 · 3937,81) = 7,5 | -3,8 | 14,44 |
| 8,5 | 3890,86 | 20,26735-(0,003242 · 3890,86) = 7,7 | 0,8 | 0,64 |
| - | - | - | **-** | **46,23** |

Źródło: opracowanie własne.

**Graficzna prezentacja odchylenia standardowego składnika resztowego**




Źródło: opracowanie własne.

**Interpretacja:** Odchylenie standardowe składnika resztowego informuje, że stopa bezrobocia rejestrowanego według województw odchyla się od wartości teoretycznych obliczonych na podstawie regresji średnio o ± 1,8%.

**a) współczynnik zmienności resztowej (wzór 4.3.3):**

$$V_e = \frac{S_e}{\bar{y}} \cdot 100 = \frac{1{,}8}{7{,}3} \cdot 100 = 24{,}7\%$$

$$24{,}7\% > 10{,}0\%$$

$$V_e > V_g$$

**Interpretacja:** Odchylenie standardowe składnika resztowego stanowi 24,7% średniej arytmetycznej stopy bezrobocia rejestrowanego oznacza to, że model należy uznać za źle dopasowany do danych rzeczywistych.

**b) współczynnik zbieżności (wzór 4.3.5):**

Obliczenia pomocnicze:

| $y_i$ | $\hat{y}_i$ | $y_i - \hat{y}_i$ | $(y_i - \hat{y}_i)^2$ | $y_i - \bar{y}$ | $(y_i - \bar{y})^2$ |
|---|---|---|---|---|---|
| 5,7 | 6,0 | -0,3 | 0,09 | -1,6 | 2,6 |
| 9,9 | 8,2 | 1,7 | 2,89 | 2,6 | 6,8 |
| 8,8 | 7,9 | 0,9 | 0,81 | 1,5 | 2,3 |
| 6,5 | 8,1 | -1,6 | 2,56 | -0,8 | 0,6 |
| 6,7 | 7,5 | -0,8 | 0,64 | -0,6 | 0,4 |
| 5,3 | 7,0 | -1,7 | 2,89 | -2,0 | 4,0 |
| 5,6 | 3,3 | 2,3 | 5,29 | -1,7 | 2,9 |
| 7,3 | 7,5 | -0,2 | 0,04 | 0,0 | 0,0 |
| 9,6 | 8,3 | 1,3 | 1,69 | 2,3 | 5,3 |
| 8,5 | 7,9 | 0,6 | 0,36 | 1,2 | 1,4 |
| 5,4 | 6,6 | -1,2 | 1,44 | -1,9 | 3,6 |
| 5,1 | 6,5 | -1,4 | 1,96 | -2,2 | 4,8 |
| 8,8 | 8,3 | 0,5 | 0,25 | 1,5 | 2,3 |
| 11,7 | 8,5 | 3,2 | 10,24 | 4,4 | 19,4 |
| 3,7 | 7,5 | -3,8 | 14,44 | -3,6 | 13,0 |
| 8,5 | 7,7 | 0,8 | 0,64 | 1,2 | 1,4 |
| - | - | - | **46,23** | - | **70,8** |

Źródło: opracowanie własne.

$$\varphi^2 = \frac{\sum_{t=1}^{n}(y_i - \hat{y}_i)^2}{\sum_{t=1}^{n}(y_i - \bar{y})^2} = \frac{46{,}23}{70{,}8} = 0{,}653$$

$$\varphi^2 \cdot 100\% = 0{,}653 \cdot 100\% = 65{,}3\%$$

**Interpretacja:** Ponad 65,0% zmienności stopy bezrobocia rejestrowanego nie zostało wyjaśnione na podstawie regresji (przez zmienną $x_i$). Powyższa zmienność jest efektem pozostałych czynników o charakterze losowymi i nielosowym.

**c) współczynnik determinacji (wzór 4.3.6):**

$$R^2 = 1 - \varphi^2 = 1 - 0{,}653 = 0{,}347$$
$$R^2 \cdot 100\% = 0{,}446 \cdot 100\% = 34{,}7\%$$

**Interpretacja:** Tylko około 35,0% zmienności stopy bezrobocia rejestrowanego zostało wyjaśnione na podstawie regresji (przez zmienną $x_i$).

# 5. Analiza szeregów czasowych

## 5.1. Przyrosty absolutne i względne

Podstawowymi miarami wykorzystywanymi w analizie dynamiki są: **przyrosty** (absolutne, względne) **Przyrosty absolutne – bezwzględne (Δ)** dzielą się na przyrosty: jednopodstawowe i zmiennopodstawowe (łańcuchowe). **Przyrosty absolutne jednopodstawowe** ($\Delta_{t+1/t}$) wskazują różnicę pomiędzy wartością bieżącą w badanym okresie $y_{t+n}$ a wartością z okresu podstawowego $y_t$:

$$\Delta_{2/1} = y_2 - y_1, \quad \Delta_{3/1} = y_3 - y_1, \ldots \quad \Delta_{t+n/t} = y_{t+n} - y_t \tag{5.1.1}$$

Przyrosty absolutne o podstawie stałej **informują o ile wzrosło, zmalało lub pozostało bez zmian badane zjawisko w bieżącym okresie w porównaniu do okresu podstawowego (bazowego).**

**Przyrosty absolutne łańcuchowe** ($\Delta_{t/t-1}$) wskazują różnicę pomiędzy wartością bieżącą w badanym okresie $y_t$ a wartością z okresu poprzedniego $y_{t-1}$:

$$\Delta_{2/1} = y_2 - y_1, \quad \Delta_{3/2} = y_3 - y_2, \ldots \quad \Delta_{t/t-1} = y_t - y_{t-1} \tag{5.1.2}$$

Przyrosty absolutne o podstawie łańcuchowej **informują o ile wzrosło, zmalało lub pozostało bez zmian badane zjawisko w bieżącym okresie w porównaniu do okresu poprzedzającego.**

**Przyrosty względne – wskaźniki tempa ($w$)** dzielą się na przyrosty**:** jednopodstawowe i łańcuchowe. Mogą być one wyrażone ułamkowo lub procentowo. **Przyrosty względne jednopodstawowe** ($w^s_{t+1/t}$) stanowią stosunek przyrostu absolutnego jednopodstawowego ($\Delta_{t+1/t}$) do wartości z okresu podstawowego $y_t$:

$$w^s_{2/1} = \frac{y_2 - y_1}{y_1} \cdot 100, \qquad w^s_{3/1} = \frac{y_3 - y_1}{y_1} \cdot 100, \cdots \; w^s_{t+1/t} = \frac{y_{t+n} - y_t}{y_t} \cdot 100 \tag{5.1.3}$$

Zatem można zapisać:

$$w_{2/1}^{s} = \frac{\Delta_{2/1}}{y_1} \cdot 100, \quad w_{3/1}^{s} = \frac{\Delta_{3/1}}{y_1} \cdot 100, \cdots \; w_{t+1/t}^{s} = \frac{\Delta_{t+n/t}}{y_t} \cdot 100 \tag{5.1.4}$$

Przyrosty względne jednopodstawowe **informują o ile procent wzrosło, spadło lub pozostało bez zmian poziom badanego zjawiska w danym okresie w porównaniu z okresem podstawowym.**

**Przyrosty względne łańcuchowe** ($\boldsymbol{w_{t/t-1}^{z}}$) stanowią stosunek przyrostu absolutnego łańcuchowego ($\Delta_{t/t-1}$) do wartości z okresu poprzedniego $y_{t-1}$:

$$w_{2/1}^{z} = \frac{y_2 - y_1}{y_1} \cdot 100, \; w_{3/2}^{z} = \frac{y_3 - y_2}{y_2} \cdot 100, \cdots \quad w_{t/t-1}^{z} = \frac{y_t - y_{t-1}}{y_{t-1}} \cdot 100 \tag{5.1.5}$$

Wówczas możemy zapisać:

$$w_{2/1}^{z} = \frac{\Delta_{2/1}}{y_1} \cdot 100, \; w_{3/2}^{z} = \frac{\Delta_{3/2}}{y_2} \cdot 100, \cdots \quad w_{t/t-1}^{z} = \frac{\Delta_{t/t-1}}{y_{t-1}} \cdot 100 \tag{5.1.6}$$

Przyrosty względne jednopodstawowe **informują o ile procent wzrosło, spadło lub pozostało bez zmian poziom badanego zjawiska w danym okresie w porównaniu z okresem poprzednim.**

## 5.2. Indeksy indywidualne (wskaźniki dynamiki)

**Indeksy indywidualne (*I*)** dzielą się na jednopodstawowe i łańcuchowe. Mogą być one wyrażone ułamkowo lub procentowo. **Indeks indywidualny o podstawie stałej ($\boldsymbol{I_{t+1/t}^{s}}$)** przedstawia udział badanego zjawiska w danym okresie w poziomie zjawiska w okresie bazowym:

$$I_{2/1}^{s} = \frac{y_2}{y_1} \cdot 100, \quad I_{3/1}^{s} = \frac{y_3}{y_1} \cdot 100, \cdots \; I_{t+1/t}^{s} = \frac{y_{t+n}}{y_t} \cdot 100 \tag{5.2.1}$$

Indeks indywidualny jednopodstawowy **informuje jak zmienił się poziom badanego zjawisko w okresie badanym się (czy wzrosło, spadło lub pozostało bez zmian) w porównaniu z okresem podstawowym.**

**Indeksy indywidualne o podstawie zmiennej ($I^z_{t/t-1}$)** przedstawia udział badanego zjawiska w danym okresie w poziomie zjawiska w okresie poprzednim:

$$I^z_{2/1} = \frac{y_2}{y_1} \cdot 100, \quad I^z_{3/2} = \frac{y_3}{y_2} \cdot 100, \cdots \quad I^z_{t/t-1} = \frac{y_t}{y_{t-1}} \cdot 100 \tag{5.2.2}$$

Indeks indywidualny zmiennopodstawowy **informuje jak zmienił się poziom badanego zjawiska w okresie badanym się (czy wzrosło, spadło lub pozostało bez zmian) w porównaniu z okresem poprzednim.** Jeśli indeks indywidualny jest mniejszy od 100%, świadczy to o spadku zjawiska, jeśli indeks jest większy lub 100%, świadczy to o wzroście poziomu zjawiska w badanym okresie.

## 5.3. Średnie tempo wzrostu

Często zdarza się, że chcemy zbadać zmiany w wielkości danego zjawiska w całym analizowanym okresie. Przyrosty i indeksy indywidualne przedstawiają wyłącznie zmiany z okresu na okres, a więc tylko pomiędzy dwoma momentami czasowymi. W tym celu wykorzystuje się średnią geometryczną jako iloczyn łańcuchowych indeksów dynamiki (wzór 2.1.4). Wówczas możemy obliczyć **średnie tempo wzrostu badanego zjawiska** w analizowanym czasie stosując następujący wzór:

$$\bar{I}_g = \sqrt[t-1]{\frac{y_2}{y_1} \cdot \frac{y_3}{y_2} \cdot \frac{y_4}{y_3} \dots \cdot \frac{y_t}{y_{t-1}}} = \sqrt[t-1]{\prod_{t=2}^{t} \frac{y_t}{y_{t-1}}} \tag{5.3.1}$$

Warunkiem wyznaczenia tego parametru są zaobserwowane w badanym szeregu jednokierunkowe zmiany (wyłącznie wzrostu lub spadku). Średni indeks dynamiki (wzór 5.3.1) możemy zapisać w postaci uproszczonej:

$$\bar{I}_g = \sqrt[t-1]{\frac{y_t}{y_1}} \cdot 100 \tag{5.3.2}$$

gdzie:

$y_t$ – poziom zjawiska w **ostatnim** okresie badanym,
$y_1$ – poziom zjawiska w **pierwszym** okresie badanym,
$t$ – liczba badanych okresów.

W związku z trudnościami obliczania pierwiastka *n*-tego stopnia można zastosować przekształcenie logarytmiczne:

$$log\bar{I}_g = \frac{1}{t-1}(logy_t - logy_1) \tag{5.3.3}$$

## 5.4. Indeksy agregatowe (zespołowe)

W analizie szeregów czasowych często zachodzi konieczność przeprowadzenia analizy dynamiki dla całego agregatu (zbioru, zespołu elementów) zjawisk niejednorodnych. W tym celu stosuje się zespołowe indeksy dla wielkości absolutnych do, których zalicza się:

- indeks wartości (*w*),
- indeks cen (*p*),
- indeks ilości (*q*).

**Indeks wartości (*w*),** oblicze się według wzoru:

$$I_w = \frac{\sum q_t p_t}{\sum q_0 p_0} \cdot 100 \tag{5.4.1}$$

gdzie:

$p_t$ – cena *i*-tego elementu w okresie badanym,
$p_0$ – cena *i*-tego elementu w okresie podstawowym,
$p_t$ – ilość *i*-tego elementu w okresie badanym,
$p_0$ – ilość *i*-tego elementu w okresie podstawowym,
$\sum q_t p_t$ – suma wartość agregatu w okresie badanym,
$\sum q_0 p_0$– suma wartość agregatu w okresie podstawowym.

Indeks wartości **informuje o ile % zmieni się wartość agregatu w okresie badanym, w porównaniu do okresu podstawowego.**

Indeksy ilości cen oblicza się na podstawie formuł standaryzacyjnych: Laspeyresa, Paaschego i Fishera.

- **Indeks cen Laspeyresa ($I_p^L$)** można zapisać:

$$I_p^L = \frac{\sum q_0 p_t}{\sum q_0 p_0} \cdot 100 \tag{5.4.2}$$

Indeks cen Laspeyresa **informuje o ile % zmieni się cena w okresie badanym w porównaniu do okresu podstawowego przy założeniu stałości ilości dla okresu podstawowego.**

- **Indeks ilości Laspeyresa ($I_q^L$)** ma postać:

$$I_q^L = \frac{\sum q_t p_0}{\sum q_0 p_0} \cdot 100 \tag{5.4.3}$$

Indeks ilości Laspeyresa **informuje o ile % zmieni się ilość w okresie badanym w porównaniu do okresu podstawowego przy założeniu stałości ceny dla okresu podstawowego.**

- **Indeks cen Paaschego ($I_p^P$)** oblicza się:

$$I_p^P = \frac{\sum q_t p_t}{\sum q_t p_0} \cdot 100 \tag{5.4.4}$$

Indeks cen Paaschego **informuje o ile % zmieni się cena w okresie badanym w porównaniu do okresu podstawowego przy założeniu stałości ilości dla okresu badanego.**

- **Indeks ilości Paaschego ($I_q^P$)** ma postać:

$$I_q^P = \frac{\sum q_t p_t}{\sum q_0 p_t} \cdot 100 \tag{5.4.5}$$

Indeks ilości Paaschego **informuje o ile % zmieni się ilość w okresie badanym w porównaniu do okresu podstawowego przy założeniu stałości ceny dla okresu badanego.**

- **Indeks cen Fishera ($I_p^F$)** jest średnią geometryczną obliczonych indeksów cen według formuły Laspeyresa i Paaschego:

$$I_p^F = \sqrt{I_p^L \cdot I_p^P} \tag{5.4.6}$$

Indeks cen Fishera **informuje o średnich zmianach ceny w okresie badanym w porównaniu do okresu podstawowego.**

- **Indeks ilości Fishera ($I_q^F$)** jest średnią geometryczną obliczonych indeksów ilości według formuły Laspeyresa i Paaschego:

$$I_q^F = \sqrt{I_q^L \cdot I_q^P} \tag{5.4.7}$$

Indeks cen Fishera **informuje o średnich zmianach ilości w okresie badanym w porównaniu do okresu podstawowego.**

## 5.5. Wyodrębnienie tendencji rozwojowej

**Wygładzenie (wyrównanie)** szeregu czasowego polega na identyfikacji tendencji rozwojowej eliminując oddziaływania wahań o charakterze **przypadkowym i okresowym.** Tendencja rozwojowa uwypukla w badanym zjawisku działanie przyczyn głównych.

**Tendencja rozwojowa (trend)** określa ogólny kierunek (rosnący, malejący lub stabilizujący) rozwoju zjawiska w badanym czasie. Trend wyodrębnia się metodami:

- mechaniczną (średnie ruchome *k* – okresowe),
- analityczną (identyfikacja funkcji trendu).

**Metoda mechaniczna** wygładzenia szeregu czasowego polega na wyliczeniu tzw. średnich ruchomych. **Średnie ruchome** oblicza się na podstawie stałej liczbie wyrazów w szeregu czasowym przesuwając się stopniowo się o jedną obserwację do przodu odrzucając tym samym pierwszą z nich. Poszczególne wartości szeregu czasowego możemy opisać jako:

$$y_1, y_2, y_3, \dots\ y_{n-3}, y_{n-2}, y_{n-1}, y_n$$

Gdy występuje nieparzysta liczba okresów (dla *k* = 3) średnie ruchome obliczamy w następujący sposób:

$$\bar{y}_2 = \frac{y_1 + y_2 + y_3}{3}, \qquad \bar{y}_3 = \frac{y_2 + y_3 + y_4}{3}, \qquad \bar{y}_4 = \frac{y_3 + y_4 + y_5}{3} \tag{5.5.1}$$

Mając wstępnie 5 wyrazów w szeregu czasowym $y_1, \dots\ y_5$ otrzymujemy tylko 3 średnie ruchome $\bar{y}_2, \bar{y}_3\ i\ \bar{y}_4$ , a więc tracimy część informacji o badanym zjawisku. W ten sposób utraciliśmy informację dla 2 wyrazów a mianowicie dla pierwszego i ostatniego. Jeżeli obliczamy średnie ruchome z większej liczby wyrazów *k* straty informacji są jeszcze większe. Dla średniej ruchomej *k* = 5 stracimy informację dla 4 wyrazów itd.

Kiedy występuje parzysta liczba wyrazów (dla *k* = 4) stosuje się **scentrowaną średnią ruchomą,** biorąc do obliczeń połowę wartości z pierwszego i ostatniego wyrazu. Wówczas wzory są następujące:

$$\bar{y}_2 = \frac{0{,}5 \cdot y_1 + y_2 + y_3 + 0{,}5 \cdot y_4}{4}, \qquad \bar{y}_3 = \frac{0{,}5 \cdot y_2 + y_3 + y_4 + 0{,}5 \cdot y_5}{4}, \tag{5.5.2}$$

$$\bar{y}_3 = \frac{0{,}5 \cdot y_3 + y_4 + y_5 + 0{,}5 \cdot y_6}{4}$$

**Metoda analityczna** polega na identyfikacji funkcji matematycznej do danych rzeczywistych. Gdzie zmienną objaśniającą w funkcji stanowi tzw. zmienna czasowa *t*. Wygładzenie szeregu dynamicznego polega na oszacowaniu parametrów trendu liniowego o postaci:

$$y_t = a + bt + e_t \tag{5.5.3}$$

gdzie:

$y_t$ – zmienna objaśniana,
$t$ – zmienna czasowa ($t$ = 1,2,3, …, $n$),
$a\ i\ b$ – parametry funkcji,
$e_t$ – składnik losowy.

Współczynnik *b* funkcji trendu liniowego informuje o ile średnio zmieni się poziom badanego zjawiska *y* (np. średni wzrost z miesiąca na miesiąc, średni spadek z roku na rok) w anali-

zowanym okresie. Wyraz wolny *a* informuje o teoretycznym poziomie badanego zjawiska w okresie wyjściowym (dla $t = 0$). Metodą wyznaczenia tendencji jest **klasyczna metoda najmniejszych kwadratów (KMNK).** Funkcję kryterium minimalizacji KMNK określa się jako:

$$\sum_{t=1}^{n}(y_t - a - bt)^2 \rightarrow min \tag{5.5.4}$$

Po rozwiązaniu układu równań:

$$\begin{cases} \sum_{t=1}^{n} y_t = na + b\sum_{t=1}^{n} t \\ \sum_{t=1}^{n} y_t\, t = a\sum_{t=1}^{n} t + b\sum_{t=1}^{n} t^2 \end{cases} \tag{5.5.5}$$

Otrzymujemy gotowe wzory na obliczenie współczynników *a* i *b*:

$$a = \frac{\sum_{t=1}^{n}(y_t - \bar{y}_t)(t - \bar{t})}{\sum_{t=1}^{n}(t - \bar{t})^2} \tag{5.5.6}$$

$$b = \bar{y}_t - a\bar{t} \tag{5.5.7}$$

gdzie:

$y_t$ – wartości rzeczywiste,
$t$ – zmienna czasowa ($t = 1,2,3, \ldots, n$),
$\bar{y}_t$ – średnia arytmetyczna zmiennej $y_t$.
$\bar{t}$ – średnia arytmetyczna zmiennej *t*.

**Błędy standardowe parametrów** funkcji trendu liniowego obliczamy z następujących wzorów:

$$S(a) = \sqrt{\frac{S^2_{e(t)} \cdot \sum_{i=1}^{t} t_i^2}{n \cdot (\sum_{t=1}^{n} t_i^2 - n\bar{t}^2)}} \tag{5.5.8}$$

$$S(b) = \frac{S_{e_{(t)}}}{\sqrt{\sum_{t=1}^{n} t_i^2 - n\bar{t}^2}} \tag{5.5.9}$$

gdzie:

$n$ – liczebność próby losowej,
$S_{e_{(t)}}$ – błąd standardowy reszt (wzór 5.5.10),
$t$ – zmienna czasowa ($t = 1,2,3, \ldots, n$),
$\bar{t}$ – średnia arytmetyczna zmiennej *t*.

Ocenę dopasowania funkcji trendu mierzy się w taki sam sposób jak w analizie regresji stosując identyczne mierniki dokładności (podrozdział 4.4.3).

**Odchylenie standardowe składnika resztowego ($S_{e_{(t)}}$)** oblicza się:

$$S_{e_{(t)}} = \sqrt{\frac{\sum_{t=1}^{n}(y_t - \hat{y}_t)^2}{n-k}} \tag{5.5.10}$$

gdzie:
$y_t$ – wartości rzeczywiste zmiennej objaśnianej,
$\hat{y}_t$ – wartości teoretyczne zmiennej objaśnianej,
$n$ – liczebność próby losowej,
$k$ – liczba oszacowanych parametrów ($k = 2$)

**Współczynnik zmienności resztowej ($V_{e_{(t)}}$):**

$$V_{e_{(t)}} = \frac{S_{e_{(t)}}}{\bar{y}_t} \cdot 100 \tag{5.5.11}$$

Funkcję trendu przyjmuje się za dopuszczalną jeśli $V_{e_{(t)}} \leq V_g$. Wartość graniczną $V_g$ ustala się w sposób arbitralny. Najczęściej przyjmowany poziom to 10% lub 15%.

**Współczynnik zbieżności – zgodności ($\varphi^2$)** można zapisać w postaci:

$$\varphi^2 = \frac{\sum_{t=1}^{n}(y_t - \hat{y}_t)^2}{\sum_{t=1}^{n}(y_t - \bar{y}_t)^2} \tag{5.5.12}$$

gdzie:
$y_t$ – wartości rzeczywiste,
$\hat{y}_t$ – wartości teoretyczne,
$\bar{y}_t$ – średnia arytmetyczna.

Współczynnik zbieżności $\varphi^2 \cdot 100\%$ **informuje ile % zmienności zmiennej $y_t$ nie została wyjaśniona przez zmienną czasową *t*.** Im wartość współczynnika zbieżności jest bliższa 0, tym szacowana funkcja trendu jest lepiej dopasowana do wartości rzeczywistych zmiennej $y_t$. Współczynnik zbieżności mieści się w przedziale:

$$0 \leq \varphi^2 \leq 1$$

**Współczynnik determinacji ($R^2$):**

$$R^2 = 1 - \varphi^2 \tag{5.5.13}$$

Współczynnik determinacji $R^2 \cdot 100\%$ **informuje ile % zmienności zmiennej $y_t$ została wyjaśniona przez zmienną czasową *t*.** Im jego wartość jest bliższa 1 tym dopasowanie trendu jest lepsze. Współczynnik determinacji mieści się w przedziale:

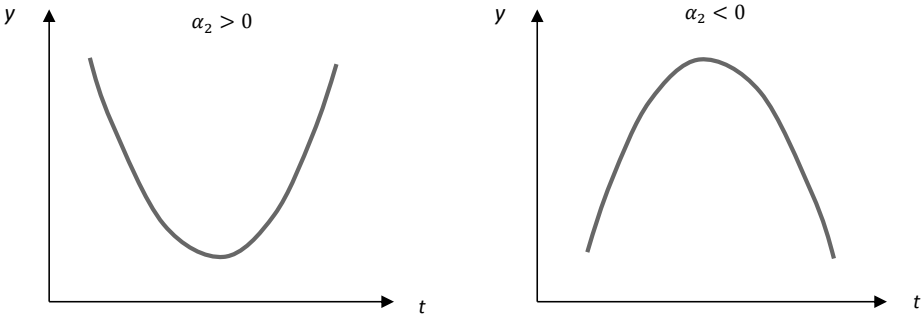
$$0 \leq R^2 \leq 1$$

Wówczas $\varphi^2 + R^2 = 1$.

Kształt tendencji rozwojowej badanego zjawiska w czasie może przyjmować również postać funkcji nieliniowej (rys. 5.5.14 – 5.5.18):

**a) wielomianowa stopnia 2 (kwadratowa):**

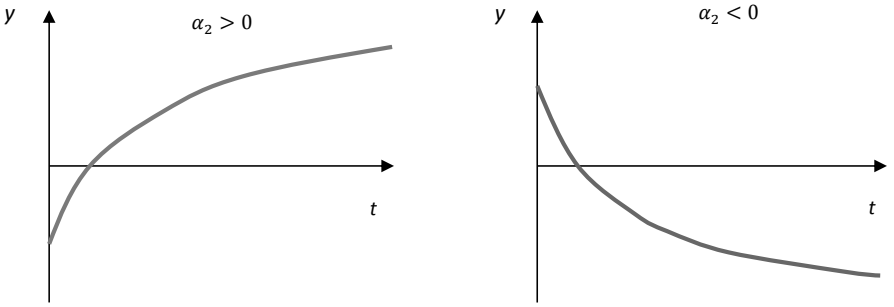
$$y_t = \alpha_0 + \alpha_1 t + \alpha_1 t^2 \tag{5.5.14}$$




Źródło: opracowanie własne.

b) **logarytmiczna**:

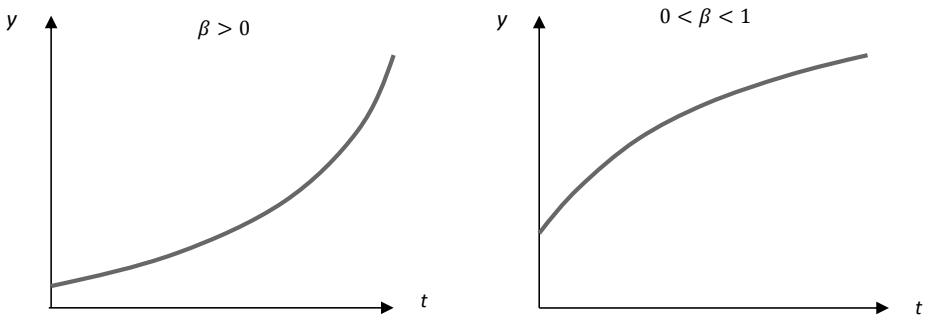
$$y_t = \alpha + \beta \ln t, \tag{5.5.15}$$




Źródło: opracowanie własne.

c) **potęgowa**:

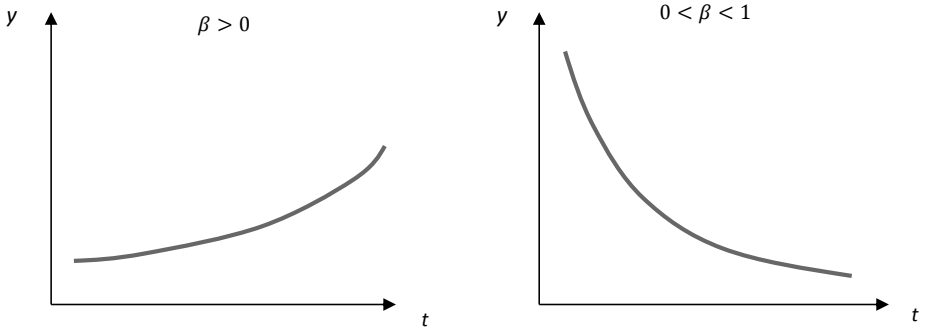
$$y_t = \alpha t^{\beta}, \tag{5.5.16}$$




Źródło: opracowanie własne.

d) **wykładnicza**:

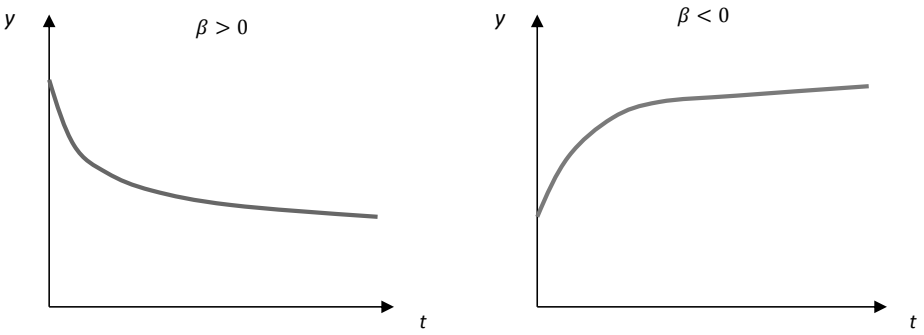
$$y_t = \alpha \beta^{t}, \tag{5.5.17}$$




Źródło: opracowanie własne.

e) **hiperboliczna**:

$$y_t = \alpha + \frac{\beta}{t}. \tag{5.5.18}$$




Źródło: opracowanie własne.

## 5.6. Wyodrębnienie wahań sezonowych

**Wahania sezonowe** są to wahania, które powtarzają się z pewną regularnością w ściśle określonych podokresach w danym roku (rys. 5.6.1.). Wyróżnia się podokresy ($k$) w ujęciu miesięcznym ($k$ = 12), kwartalnym ($k$ = 4) i półrocznym ($k$ = 2). Wahania sezonowe można wyodrębnić w wartościach bezwzględnych (model addytywny[1]) i procentach (model multiplikatywny[2]). W przypadku wyodrębnienia **bezwzględnych wahań sezonowych** w **modelu addytywnym** procedura postępowania jest następująca:

- eliminujemy trend w szeregu czasowym poprzez wyliczenie różnic ($d_t$):

$$d_t = y_t - \hat{y}_t \tag{5.6.1}$$

gdzie:
$y_t$ – wartości rzeczywiste szeregu czasowego,
$\hat{y}_t$ – wartości teoretyczne obliczone na podstawie trendu (tzw. szereg wygładzony).

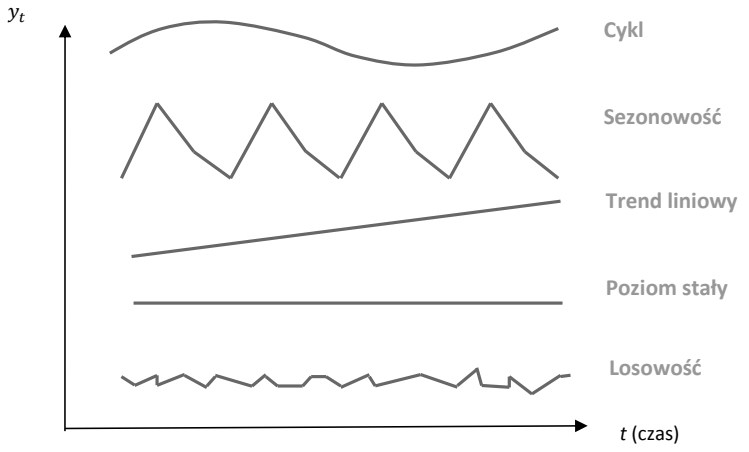
- obliczamy **surowe bezwzględne wahania sezonowe ($W_{s_k}$)** dla jednoimiennego podokresu $k$:

$$W_{s_k} = \frac{\sum_t (y_t - \hat{y}_t)}{l} = \frac{\sum_t d_t}{l} \tag{5.6.2}$$

[1] **Model addytywny** – zakłada, że wartości zmiennej $y$ stanowią sumę wszystkich (bądź niektórych) składowych szeregu czasowego (tzn. trendu, wahań sezonowych, wahań cyklicznych i składnika losowego).
[2] **Model multiplikatywny** – zakłada, że wartości zmiennej $y$ stanowią iloczyn wszystkich (bądź niektórych) składowych szeregu czasowego (tzn. trendu, wahań sezonowych, wahań cyklicznych i składnika losowego).

gdzie:
$l$ – liczba cykli (lat).

Surowe składniki sezonowe stanowią średnią odchyleń wartości rzeczywistych od wartości wyznaczonych na podstawie trendu dzięki temu eliminuje się wahania sezonowe.




**Rysunek 5.6.1. Elementy składowe szeregu czasowego**
Źródło: opracowane na podstawie M., Rabiej, *Statystyka z programem Statistica*, Helion, Gliwice 2012, s. 282.

- jeśli suma surowych składników sezonowości dla badanego podokresu będzie różna od zera np. dla danych kwartalnych $\sum_{k=1}^{m} W_{s_k} = W_{s_I} + W_{s_{II}} + W_{s_{III}} + W_{s_{IV}} \neq 0$ to wówczas obliczamy **współczynnik korygujący ($W_k$)**:

$$W_k = \frac{\sum_{k=1}^{m} W_{s_k}}{m} \tag{5.6.3}$$

gdzie:
$m$ – liczba kwartałów, miesięcy.

- obliczamy **czyste bezwzględne wahania sezonowości ($W_{c_k}$):**

$$W_{c_k} = W_{s_k} - W_k \tag{5.6.4}$$

Wówczas suma $W_{c_k}$ będzie równa zero. Czyste wskaźniki sezonowości informują o ile badane zjawisko średnio wzrosło (znak +) lub spadło (znak –) w porównaniu z wartościami wyznaczonymi przez trend gdyby nie występowała sezonowość.

W **modelu multiplikatywnym** wyodrębnienie **wskaźników sezonowości** należy przeprowadzić według następującej procedury:

- uwalniamy szereg czasowy od trendu w szeregu czasowym ($\tilde{d}$):

$$\tilde{d}_t = \frac{y_t}{\hat{y}_t} \tag{5.6.5}$$

gdzie:
$y_t$ – wartości rzeczywiste szeregu czasowego,
$\hat{y}_t$ – wartości teoretyczne obliczone na podstawie trendu (tzw. szereg wygładzony).

- obliczamy **surowe wskaźniki sezonowe ($\widetilde{W}_{s_k}$)** dla jednoimiennego podokresu *k*:

$$\widetilde{W}_{s_k} = \frac{\sum_t \frac{y_t}{\hat{y}_t}}{l} = \frac{\sum_t \tilde{d}_t}{l} \tag{5.6.6}$$

gdzie:
$l$ – liczba cykli (lat).

- obliczamy **czyste wskaźniki sezonowości ($\widetilde{W}_{c_k}$)** poprzez skorygowanie surowych wskaźników sezonowości dzieląc je przez ich średnią:

$$\widetilde{W}_{c_k} = \frac{\widetilde{W}_{s_k}}{\bar{\widetilde{W}}_{sk}} \tag{5.6.7}$$

W badaniach nad sezonowością w szeregach czasowych wykorzystuje się również **analizę harmoniczną.** W tej metodzie występują tzw. harmoniki składające się z funkcji sinusa i cosinusa, które się sumuje. Liczba harmonik równa jest $\frac{n}{2}$. Pierwszą harmonikę wylicza się z całego szeregu, drugą z połowy szeregu, trzecią z jednej trzeciej, czwartą z jednej czwartej itd. Ostatecznie uwzględnia się tylko te harmoniki, które w najwyższym stopniu wyjaśniają zmiany zmiennej objaśnianej $y_t$. Analiza harmoniczna opiera się na badaniu wahań wokół średniej. Model przedstawia się następująco:

$$y_t = \alpha_0 + \sum_{i=1}^{\frac{n}{2}} \left[ \alpha_i sin\left(\frac{2\pi}{n} it\right) + \beta_i cos\left(\frac{2\pi}{n} it\right) \right] \tag{5.6.8}$$

gdzie:

$\alpha_0$– średni poziom zmiennej$y_t$,
$\alpha_i, \beta_i$ – oceny parametrów funkcji sinus i cosinus,
$i$– numer harmoniki.

Możemy również uwzględnić trend wówczas model możemy zapisać w postaci:

$$y_t = f(t) + \sum_{i=1}^{\frac{n}{2}} \left[ \alpha_i sin\left(\frac{2\pi}{n} it\right) + \beta_i cos\left(\frac{2\pi}{n} it\right) \right] \tag{5.6.9}$$

gdzie:

$f(t)$ – funkcja trendu.

Parametry $a_0, a_i, b_i$ oblicza się według wzoru:

$$a_0 = \frac{\sum_{i=1}^{n} y_t}{n} \tag{5.6.10}$$

$$a_i = \frac{2}{n} \sum_{t=1}^{n} y_t sin\left(\frac{2\pi}{n} it\right) \tag{5.6.11}$$

$$b_i = \frac{2}{n} \sum_{t=1}^{n} y_t cos\left(\frac{2\pi}{n} it\right) \tag{5.6.12}$$

gdzie:

$a_0, a_i, b_i$ – oceny parametrów $\alpha_0, \alpha_i, \beta_i$.

W ostatniej harmonice o numerze $\frac{n}{2}$ parametr $\alpha_{\frac{2}{n}}$ wynosi zero a $\beta_i$ jest równy:

$$\beta_{\frac{2}{n}} = \frac{1}{n} \sum_{t=1}^{n} y_t cos(\pi t) \tag{5.6.13}$$

Amplitudę harmonik oblicza się według wzoru:

$$A_i = \sqrt{a_i^2 + b_i^2} \tag{5.6.14}$$

W celu zlokalizowania amplitud w czasie obliczamy tzw. przesunięcie fazowe:

$$p_{f_i} = \frac{\varepsilon_i}{\theta_i} \tag{5.6.15}$$

Wartości $\varepsilon_i$ i $\theta_i$ obliczamy:

$$\varepsilon_i = arc\ tg\left(\frac{a_i}{b_i}\right) \tag{5.6.16}$$

$$\theta_i = \frac{2\pi i}{n} \tag{5.6.17}$$

Udział wariancji poszczególnej harmoniki do ogółem wariancji wyznacza się według wzoru:

$$A_i = \frac{0{,}5(a_i^2 + b_i^2)}{S_y^2} \cdot 100 \tag{5.6.18}$$

gdzie:
$S_y^2$ – wariancja zmiennej objaśnianej.

Z kolei dla ostatniej harmoniki udział wariancji ustala się według formuły:

$$A_i = \frac{(a_i^2 + b_i^2)}{S_y^2} \cdot 100 \tag{5.6.19}$$

## 5.7. Wyodrębnienie wahań przypadkowych

**Wahania przypadkowe** są to wahania o charakterze losowym. Wahania przypadkowe (składniki resztowe) w **modelu addytywnym** oblicza się według wzoru:

$$e_t = y_t - \hat{y}_t - W_{c_k} \tag{5.7.1}$$

gdzie:
$y_t$ – wartości rzeczywiste szeregu czasowego,
$\hat{y}_t$ – wartości teoretyczne obliczone na podstawie trendu,
$W_{c_k}$ – czyste wskaźniki sezonowości (wzór 5.6.4).

Składniki resztowe ( $e_t$) informują o wpływie na badane zjawisko tych czynników, które nie zostały wyjaśnione przez tendencję rozwojową i wahaniami sezonowymi.

Następnie obliczamy **odchylenie składnika resztowego $S(e_t)$** według wzoru:

$$S(e_t) = \sqrt{\frac{\sum_{t=1}^{n}\left(y_t - \hat{y}_t - W_{c_k}\right)^2}{n-k}} = \sqrt{\frac{\sum_{t=1}^{n} e_t^2}{n-k}} \tag{5.7.2}$$

oraz **współczynnik zmienności resztowej** $\boldsymbol{V(e_t)}$:

$$V(e_t) = \frac{S(e_t)}{\bar{y}_t} \cdot 100 \tag{5.7.3}$$

## 5.8. Predykcja na podstawie trendu

Predykcja (prognozowanie) nazywa się procesem szacowania wartości zmiennej prognozowanej w przyszłość na podstawie funkcji matematycznej. Zakłada się, że horyzont prognozy nie powinien przekraczać 10-20% długości analizowanego szeregu czasowego. Prognozy wyróżnia się jako:

- punktowe (jest to określona wartość zmiennej prognozowanej),
- przedziałowe (jest to przedział liczbowy z góry nadanym prawdopodobieństwem).

**Prognozowanie na podstawie trendu** przeprowadza się na podstawie modelu tendencji rozwojowej drogą prostej jego ekstrapolacji. **Przedział prognozy** wyznacza się dla założonej wiarygodności prognozy α tak, że:

$$P\{dy_T^* < y_t < gy_T^*\} = 1 - \alpha \tag{5.8.1}$$

gdzie:

$dy_T^* = y_T^* - u_\alpha S_{pT}$ jest dolną granicą przedziału,
$gy_T^* = y_T^* + u_\alpha S_{pT}$ jest górną granicą przedziału.

Dla ustalonej wiarygodności prognozy $\alpha$ i $n-k-1$ stopni swobody dla małej próby ($n \leq 30$) wartości graniczne $t_\alpha$ odczytujemy z tablic *t*-Studenta. Z kolei, dla dużej próby ($n > 30$) wartość $u_\alpha$ odczytujemy z dystrybuanty rozkładu normalnego. Dokładność prognozy mierzy się następującymi miernikami:

a) **średni błąd prognozy *ex ante*** $(\boldsymbol{S_{pT}})$ – oblicza się go przed realizacją prognozy**,** który określa, o ile przeciętnie prognozy będą się różnić od rzeczywistych wartości zmiennej prognozowanej w okresie prognozowania:

(5.8.2)

$$S_{pT} = \sqrt{\left[\frac{(T-\bar{t})^2}{\sum_{t=1}^{n}(t-\bar{t})^2} + \frac{1}{n} + 1\right] \cdot S^2_{e_{(t)}}}$$

gdzie:

$S_e^2$ – jest wariancją resztową trendu.

b) **względny błąd *ex ante* ($V_T^*$)** oblicza się go przed realizacją prognozy**:**

$$V_T^* = \frac{S_{pT}}{y_T^*} \cdot 100\% \tag{5.8.3}$$

Określa ile procent prognozy stanowi średni błąd predykcji $S_{pT}$. Jeżeli:

$$V_T^* \leq V_G \tag{5.8.4}$$

to prognoza jest dopuszczalna. Wartość Graniczną błędu $V_G$ wyznacza się najczęściej na poziomie 5% lub 10%.

Jeżeli:

$$V_T^* > V_G \tag{5.8.5}$$

wówczas prognoza jest traktowana jako niedopuszczalna.

c) **błąd prognozy *ex post* ($\delta_T$)** – oblicza się po realizacji prognozy stanowi różnicę między realizacją zmiennej prognozowanej a prognozą, co zapisuje się w postaci:

$$\delta_T = y_T - y_T^* \tag{5.8.6}$$

d) **względny błąd prognozy *ex post* ($\delta_T^*$)** – oblicza się po realizacji prognozy:

$$\delta_T^* = \frac{\delta_T}{y_T} \cdot 100\% \tag{5.8.7}$$

Jeżeli:

$$|\delta_T^*| \leq \delta_G \tag{5.8.8}$$

to prognoza jest trafna.

Jeśli:

$$|\delta_T^*| > \delta_G \tag{5.8.9}$$

wówczas prognoza jest nietrafna.

Granicę błędu $\delta_G$ wyznacza się arbitralnie na poziomie 5% lub 10%.

**Przykłady**

**Przykład 5.1.**

Przeciętne miesięczne wynagrodzenia brutto w Polsce w latach 2008-2018 wynosiły odpowiednio:

| Lata | Wynagrodzenia (w zł) |
|---|---|
| 2008 | 3158,48 |
| 2009 | 3315,38 |
| 2010 | 3435,00 |
| 2011 | 3625,21 |
| 2012 | 3744,38 |
| 2013 | 3877,43 |
| 2014 | 4003,99 |
| 2015 | 4150,86 |
| 2016 | 4290,52 |
| 2017 | 4527,89 |
| 2018 | 4834,76 |

Źródło: Bank Danych Lokalnych GUS.

Na podstawie danych przeprowadź analizę szeregu czasowego obliczając:

a) jednopodstawowe przyrosty absolutne, przyjmując rok 2008=100,

b) zmiennopodstawowe przyrosty absolutne,

c) zinterpretować wyniki dla 2018 r.

**Rozwiązanie**

Do obliczenia przyrostów absolutnych **(Δ)** skorzystamy ze wzorów 5.1.1-5.1.2.

| Lata | Wynagrodzenia (w zł) | Przyrosty absolutne (Δ) | |
|---|---|---|---|
| | | jednopodstawowe (2008=100) | zmiennopodstawowe (rok poprzedni=100) |
| 2008 | 3158,48 | 3158,48 - 3158,48 = 0 | – |
| 2009 | 3315,38 | 3315,38 - 3158,48 = 156,90 | 3315,38 - 3158,48 = 156,90 |
| 2010 | 3435,00 | 3435,00 - 3158,48 = 276,52 | 3435,00 - 3315,38 = 119,62 |
| 2011 | 3625,21 | 3625,21 - 3158,48 = 466,73 | 3625,21 - 3435,00 = 190,21 |
| 2012 | 3744,38 | 3744,38 - 3158,48 = 585,90 | 3744,38 - 3625,21 = 119,17 |
| 2013 | 3877,43 | 3877,43 - 3158,48 = 718,95 | 3877,43 - 3744,38 = 133,05 |
| 2014 | 4003,99 | 4003,99 - 3158,48 = 845,51 | 4003,99 - 3877,43 = 126,56 |
| 2015 | 4150,86 | 4150,86 - 3158,48 = 992,38 | 4150,86 - 4003,99 = 146,87 |
| 2016 | 4290,52 | 4290,52 - 3158,48 = 1132,04 | 4290,52 - 4150,86 = 139,66 |
| 2017 | 4527,89 | 4527,89 - 3158,48 = 1369,41 | 4527,89 - 4290,52 = 237,37 |
| 2018 | 4834,76 | 4834,76 - 3158,48 = **1676,28** | 4834,76 - 4527,89 = **306,87** |

Źródło: opracowanie własne.

**Interpretacja:**

- **o podstawie stałej dla 2018 r.:**

Przeciętne miesięczne wynagrodzenie brutto w Polsce w 2018 r. wzrosło o **1676,28 zł** w porównaniu do 2008 r.

- **o podstawie zmiennej dla 2018 r.:**

Przeciętne miesięczne wynagrodzenie brutto w Polsce w 2018 r. wzrosło o **306,87 zł** w porównaniu do 2017 r.

**Przykład 5.2.**

Na podstawie danych z przykładu 5.1. przeprowadź analizę szeregu czasowego obliczając:

a) jednopodstawowe przyrosty względne, przyjmując rok 2008=100,
b) zmiennopodstawowe przyrosty względne,
c) zinterpretować wyniki dla 2018 r.

**Rozwiązanie**

Do obliczenia przyrostów względnych **($w$)** skorzystamy ze wzorów 5.1.3-5.1.5.

| Lata | Wynagrodzenia (w zł) | Przyrosty względne ($w$) | |
|---|---|---|---|
| | | jednopodstawowe (2008=100) | zmiennopodstawowe (rok poprzedni=100) |
| | | (w %) | |
| 2008 | 3158,48 | $\frac{3158,48 - 3158,48}{3158,48} \cdot 100 = 0,0$ | – |
| 2009 | 3315,38 | $\frac{3315,38 - 3158,48}{3158,48} \cdot 100 = 5,0$ | $\frac{3315,38 - 3158,48}{3158,48} \cdot 100 = 5,0$ |
| 2010 | 3435,00 | $\frac{3435,00 - 3158,48}{3158,48} \cdot 100 = 8,8$ | $\frac{3435,00 - 3315,38}{3315,38} \cdot 100 = 3,6$ |
| 2011 | 3625,21 | $\frac{3625,21 - 3158,48}{3158,48} \cdot 100 = 14,8$ | $\frac{3625,21 - 3435,00}{3435,00} \cdot 100 = 5,5$ |
| 2012 | 3744,38 | $\frac{3744,38 - 3158,48}{3158,48} \cdot 100 = 18,6$ | $\frac{3744,38 - 3625,2}{3625,2} \cdot 100 = 3,3$ |
| 2013 | 3877,43 | $\frac{3877,43 - 3158,48}{3158,48} \cdot 100 = 22,8$ | $\frac{3877,43 - 3744,38}{3744,38} \cdot 100 = 3,6$ |
| 2014 | 4003,99 | $\frac{4003,99 - 3158,48}{3158,48} \cdot 100 = 26,8$ | $\frac{4003,99 - 3877,43}{3877,43} \cdot 100 = 3,3$ |
| 2015 | 4150,86 | $\frac{4150,86 - 3158,48}{3158,48} \cdot 100 = 31,4$ | $\frac{4150,86 - 4003,99}{4003,99} \cdot 100 = 3,7$ |
| 2016 | 4290,52 | $\frac{4290,52 - 3158,48}{3158,48} \cdot 100 = 35,8$ | $\frac{4290,52 - 4150,86}{4150,86} \cdot 100 = 3,4$ |
| 2017 | 4527,89 | $\frac{4527,89 - 3158,48}{3158,48} \cdot 100 = 43,4$ | $\frac{4527,89 - 4290,52}{4290,52} \cdot 100 = 5,5$ |
| 2018 | 4834,76 | $\frac{4834,76 - 3158,48}{3158,48} \cdot 100 = \mathbf{53,1}$ | $\frac{4834,76 - 4527,89}{4527,89} \cdot 100 = \mathbf{6,8}$ |

Źródło: opracowanie własne.

**Interpretacja:**

- **o podstawie stałej dla 2018 r.:**

Przeciętne miesięczne wynagrodzenie brutto w Polsce w 2018 r. wzrosło o **53,1%** w porównaniu do 2008 r.

- **o podstawie zmiennej dla 2018 r.:**

Przeciętne miesięczne wynagrodzenie brutto w Polsce w 2018 r. wzrosło o **6,8%** w porównaniu do 2017 r.

**Przykład 5.3.**

Na podstawie danych z przykładu 5.1. oblicz indeksy dynamiki:

a) jednopodstawowe, przyjmując 2008=100,
b) zmiennopodstawowe,
c) zinterpretować wyniki dla 2018 r.

**Rozwiązanie**

Do obliczenia indeksów dynamiki **(*I*)** skorzystamy ze wzorów 5.2.1-5.2.2.

| Lata | Wynagrodzenia (w zł) | Indeksy dynamiki (*I*) | |
|---|---|---|---|
| | | jednopodstawowe (2008=100) | zmiennopodstawowe (rok poprzedni=100) |
| | | (w %) | |
| 2008 | 3158,48 | $\frac{y_{2008}}{y_{2008}} \cdot 100 = \frac{3158,48}{3158,48} \cdot 100 = 100,0$ | – |
| 2009 | 3315,38 | $\frac{y_{2009}}{y_{2008}} \cdot 100 = \frac{3315,38}{3158,48} \cdot 100 = 105,0$ | $\frac{y_{2009}}{y_{2008}} \cdot 100 = \frac{3315,38}{3158,48} \cdot 100 = 105,0$ |
| 2010 | 3435,00 | $\frac{y_{2010}}{y_{2008}} \cdot 100 = \frac{3435,00}{3158,48} \cdot 100 = 108,8$ | $\frac{y_{2010}}{y_{2009}} \cdot 100 = \frac{3435,00}{3315,38} \cdot 100 = 103,6$ |
| 2011 | 3625,21 | $\frac{y_{2011}}{y_{2008}} \cdot 100 = \frac{3625,21}{3158,48} \cdot 100 = 114,8$ | $\frac{y_{2011}}{y_{2010}} \cdot 100 = \frac{3625,21}{3435,00} \cdot 100 = 105,5$ |
| 2012 | 3744,38 | $\frac{y_{2012}}{y_{2008}} \cdot 100 = \frac{3744,38}{3158,48} \cdot 100 = 118,6$ | $\frac{y_{2012}}{y_{2011}} \cdot 100 = \frac{3744,38}{3625,2} \cdot 100 = 103,3$ |
| 2013 | 3877,43 | $\frac{y_{2013}}{y_{2008}} \cdot 100 = \frac{3877,43}{3158,48} \cdot 100 = 122,8$ | $\frac{y_{2013}}{y_{2012}} \cdot 100 = \frac{3877,43}{3744,38} \cdot 100 = 103,6$ |
| 2014 | 4003,99 | $\frac{y_{2014}}{y_{2008}} \cdot 100 = \frac{4003,99}{3158,48} \cdot 100 = 126,8$ | $\frac{y_{2014}}{y_{2013}} \cdot 100 = \frac{4003,99}{3877,43} \cdot 100 = 103,3$ |
| 2015 | 4150,86 | $\frac{y_{2015}}{y_{2008}} \cdot 100 = \frac{4150,86}{3158,48} \cdot 100 = 131,4$ | $\frac{y_{2015}}{y_{2014}} \cdot 100 = \frac{4150,86}{4003,99} \cdot 100 = 103,7$ |
| 2016 | 4290,52 | $\frac{y_{2016}}{y_{2008}} \cdot 100 = \frac{4290,52}{3158,48} \cdot 100 = 135,8$ | $\frac{y_{2016}}{y_{2015}} \cdot 100 = \frac{4290,52}{4150,86} \cdot 100 = 103,4$ |
| 2017 | 4527,89 | $\frac{y_{2017}}{y_{2008}} \cdot 100 = \frac{4527,89}{3158,48} \cdot 100 = 143,4$ | $\frac{y_{2017}}{y_{2016}} \cdot 100 = \frac{4527,89}{4290,52} \cdot 100 = 105,5$ |
| 2018 | 4834,76 | $\frac{y_{2018}}{y_{2008}} \cdot 100 = \frac{4834,76}{3158,48} \cdot 100 = \mathbf{153,1}$ | $\frac{y_{2018}}{y_{2017}} \cdot 100 = \frac{4834,76}{4527,89} \cdot 100 = \mathbf{106,8}$ |

Źródło: opracowanie własne.

**Interpretacja:**

- **o podstawie stałej dla 2018 r.:**

Przeciętne miesięczne wynagrodzenie brutto w Polsce w 2018 r. wzrosło o **53,1%** (153,1% – 100,0% = 53,1%) w porównaniu do 2008 r.

- **o podstawie zmiennej dla 2018 r.:**

Przeciętne miesięczne wynagrodzenie brutto w Polsce w 2018 r. wzrosło o **6,8%** (106,8% – 100,0% = 6,8%) w porównaniu do 2017 r.

**Przykład 5.4.**

Oblicz i zinterpretuj średnie roczne tempo wzrostu przeciętnych miesięcznych wynagrodzeń brutto w Polsce (dane z przykładu 5.1.).

**Rozwiązanie**

Do obliczenia średniego tempa wzrostu $\bar{I}_g$ skorzystamy ze wzorów 5.3.1-5.3.3:

- **na podstawie łańcuchowych indeksów dynamiki:**

$$t = 11$$

$$\bar{I}_g = \sqrt[t-1]{\frac{y_2}{y_1}\cdot\frac{y_3}{y_2}\cdot\frac{y_4}{y_3}\ldots\cdot\frac{y_t}{y_{t-1}}} = \sqrt[11-1]{\frac{y_{2009}}{y_{2008}}\cdot\frac{y_{2010}}{y_{2009}}\cdot\frac{y_{2011}}{y_{2010}}\ldots\cdot\frac{y_{2018}}{y_{2017}}}$$

$$= \sqrt[10]{105{,}0\cdot 103{,}6\cdot 105{,}5\ldots\cdot 106{,}8} = 104{,}3\%$$

- **na podstawie metody uproszczonej (ostatni i pierwszy okres):**

$$y_t = y_{2018} = 4834{,}76$$

$$y_1 = y_{2008} = 3158{,}48$$

$$\bar{I}_g = \sqrt[t-1]{\frac{y_t}{y_1}}\cdot 100 = \sqrt[11-1]{\frac{y_{2018}}{y_{2008}}}\cdot 100 = \sqrt[10]{\frac{4834{,}76}{3158{,}48}}\cdot 100 = 104{,}3\%$$

- **na podstawie zlogarytmowanych danych (ostatni i pierwszy okres):**

$$logy_t = logy_{2018} = log_{10}4834{,}76 = 3{,}6844$$

$$logy_1 = logy_{2008} = log_{10}3158{,}48 = 3{,}4995$$

$$log\bar{I}_g = \frac{1}{t-1}(logy_t - logy_1) = \frac{1}{11-1}(log4834{,}76 - log3158{,}48) =$$

$$= \frac{1}{10}(3{,}6844 - 3{,}4995) = 0{,}01849$$

$$10^{0{,}01849} = 1{,}043$$

$$1{,}043 \cdot 100 = 104{,}3\%$$

**Interpretacja:** Przeciętne miesięczne wynagrodzenie brutto w Polsce w latach 2008-2018 średniorocznie (z roku na rok) rosło o 4,3%.

**Przykład 5.5.**

W hurtowni X wielkość sprzedanych produktów i ich ceny w 2015 i 2018 r. kształtowała się następująco:

| Produkt | 2015 | | 2018 | |
|---|---|---|---|---|
| | Cena (zł) | Ilość (szt.) | Cena (zł) | Ilość (szt.) |
| A | 20 | 8 | 25 | 9 |
| B | 30 | 10 | 35 | 12 |

Dane umowne.

Na podstawie danych oblicz agregatowy indeksy:

a) wartości,
b) cen według formuły Laspeyresa i Paaschego,
a) ilości według formuły Laspeyresa i Paaschego,
b) ilości i cen według Fishera.

**Rozwiązanie**

| Produkt | 2015 | | 2018 | |
|---|---|---|---|---|
| | Cena (zł) $p_0$ | Ilość (szt.) $q_0$ | Cena (zł) $p_t$ | Ilość (szt.) $q_t$ |
| A | 20 | 8 | 25 | 9 |
| B | 30 | 10 | 35 | 12 |

Źródło: opracowanie własne.

Obliczenia pomocnicze:

| Produkt | $q_0p_0$ | $q_tp_t$ | $q_0p_t$ | $q_tp_0$ |
|---|---|---|---|---|
| A | 160 | 225 | 200 | 180 |
| B | 300 | 420 | 350 | 360 |
| **Ogółem** | **460** | **645** | **550** | **540** |

Źródło: opracowanie własne.

**a) Indeks wartości (wzór 5.4.1):**

$$I_w = \frac{\sum q_t p_t}{\sum q_0 p_0} \cdot 100 = \frac{\sum q_{2018} p_{2018}}{\sum q_{2015} p_{2015}} \cdot 100 = \frac{645}{460} \cdot 100 = 140{,}2\%$$

**Interpretacja:** Łączna wartość sprzedanych produktów A i B w 2018 r. jest o 40,2% wyższa w porównaniu do wartość z 2015 r.

**b) Indeks cen (wzory 5.4.2 i 5.4.4):**

- **Laspeyresa:**

$$I_p^L = \frac{\sum q_0 p_t}{\sum q_0 p_0} \cdot 100 = \frac{\sum q_{2015} p_{2018}}{\sum q_{2015} p_{2015}} \cdot 100 = \frac{550}{460} \cdot 100 = 119{,}6\%$$

**Interpretacja:** Cena sprzedanych produktów w 2018 r. wzrosła o 19,6% w porównaniu do 2015 r. przy założeniu, że wielkość sprzedanych produktów w 2018 była taka sama jak w 2015 r.

- **Paaschego:**

$$I_p^P = \frac{\sum q_t p_t}{\sum q_t p_0} \cdot 100 = \frac{\sum q_{2018} p_{2018}}{\sum q_{2018} p_{2015}} \cdot 100 = \frac{645}{540} \cdot 100 = 119{,}4\%$$

**Interpretacja:** Cena sprzedanych produktów w 2018 r. wzrosła o 19,4% w porównaniu do 2015 r. przy założeniu, że wielkość sprzedanych produktów w 2015 była taka sama jak w 2018 r.

**c) Indeks ilości (wzory 5.4.3 i 5.4.5):**

- **Laspeyresa:**

$$I_q^L = \frac{\sum q_t p_0}{\sum q_0 p_0} \cdot 100 = \frac{\sum q_{2018} p_{2015}}{\sum q_{2015} p_{2015}} \cdot 100 = \frac{540}{460} \cdot 100 = 117{,}4\%$$

**Interpretacja:** Wielkość sprzedaży produktów w 2018 r. wzrosła o 17,4% w porównaniu do 2015 r. przy założeniu że ceny w 2018 r. były takie same jak w 2015 r.

- **Paaschego:**

$$I_q^P = \frac{\sum q_t p_t}{\sum q_0 p_t} \cdot 100 = \frac{\sum q_{2018} p_{2018}}{\sum q_{2015} p_{2018}} \cdot 100 = \frac{645}{550} \cdot 100 = 117{,}3\%$$

**Interpretacja:** Wielkość sprzedaży produktów w 2018 r. wzrosła o 17,3% w porównaniu do 2015 r. przy założeniu że ceny w 2015 r. były takie same jak w 2018 r.

**d) Indeks cen i ilości Fishera (wzory 5.4.6 i 5.4.7):**

- **indeks cen:**

$$I_p^F = \sqrt{I_p^L \cdot I_p^P} = \sqrt{119{,}6 \cdot 119{,}4} = 119{,}5\%$$

**Interpretacja:** Średni wzrost cen sprzedaży produktów A i B w 2018 r. w porównaniu do 2015 r. wynosił 19,5%.

- **indeks ilości:**

$$I_q^F = \sqrt{I_q^L \cdot I_q^P} = \sqrt{117{,}4 \cdot 117{,}3} = 117{,}3\%$$

**Interpretacja:** Średni wzrost wielkości sprzedaży produktów A i B w 2018 r. w porównaniu do 2015 r. wynosił 17,3%.

**Przykład 5.6.**

Wyznacz tendencję rozwojową w sposób mechaniczny, wykorzystując średnią ruchomą trzy i pięcioletnią (dane z przykładu 5.1).

**Rozwiązanie**

- **średnia ruchoma trzyletnia (wzór 5.5.1):**

| Lata | Wynagrodzenia (w zł) | Średnia ruchoma 3 okresowa |
|---|---|---|
| 2008 | 3158,48 | – |
| 2009 | 3315,38 | $\frac{y_1+y_2+y_3}{3} = \frac{3158{,}48+3315{,}38+3435{,}00}{3} = 3302{,}95$ |
| 2010 | 3435,00 | $\frac{y_2+y_3+y_4}{3} = \frac{3315{,}38+3435{,}00+3625{,}21}{3} = 2458{,}53$ |
| 2011 | 3625,21 | $\frac{y_3+y_4+y_5}{3} = \frac{3435{,}00+3625{,}21+3744{,}38}{3} = 3601{,}53$ |
| 2012 | 3744,38 | $\frac{y_4+y_5+y_6}{3} = \frac{3625{,}21+3744{,}38+3877{,}43}{3} = 3749{,}01$ |
| 2013 | 3877,43 | $\frac{y_5+y_6+y_7}{3} = \frac{3744{,}38+3877{,}43+4003{,}99}{3} = 3875{,}27$ |
| 2014 | 4003,99 | $\frac{y_6+y_7+y_8}{3} = \frac{3877{,}43+4003{,}99+4150{,}86}{3} = 4010{,}76$ |
| 2015 | 4150,86 | $\frac{y_7+y_8+y_9}{3} = \frac{4003{,}99+4150{,}86+4290{,}52}{3} = 4148{,}46$ |
| 2016 | 4290,52 | $\frac{y_8+y_9+y_{10}}{3} = \frac{4150{,}86+4290{,}52+4527{,}89}{3} = 4323{,}09$ |
| 2017 | 4527,89 | $\frac{y_9+y_{10}+y_{11}}{3} = \frac{4290{,}52+4527{,}89+4834{,}76}{3} = 4551{,}06$ |
| 2018 | 4834,76 | – |

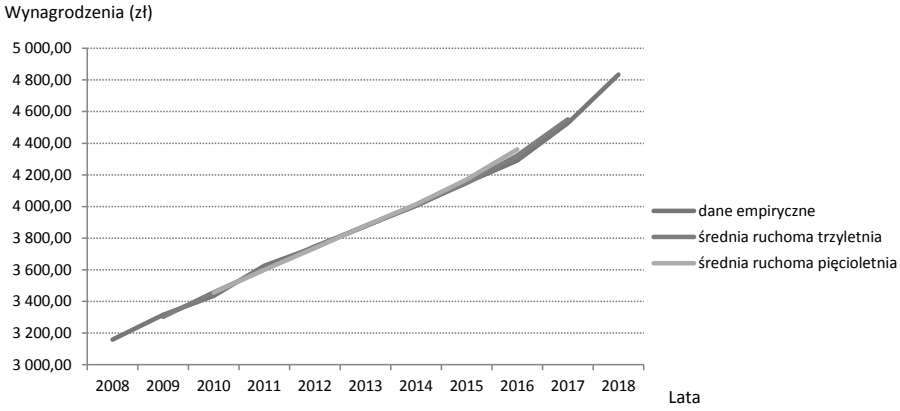
Źródło: opracowanie własne.

- **średnia ruchoma pięcioletnia:**

| Lata | Wynagrodzenia (w zł) | Średnia ruchoma 5 okresowa |
|---|---|---|
| 2008 | 3158,48 | – |
| 2009 | 3315,38 | – |
| 2010 | 3435,00 | $\frac{y_1+y_2+y_3+y_4+y_5}{5}=\frac{3158,48+3315,38+3435,00+3625,21+3744,38}{5}=3455,69$ |
| 2011 | 3625,21 | $\frac{y_2+y_3+y_4+y_5+y_6}{5}=\frac{3315,38+3435,00+3625,21+3744,38+3877,43}{5}=3599,48$ |
| 2012 | 3744,38 | $\frac{y_3+y_4+y_5+y_6+y_7}{5}=\frac{3435,00+3625,21+3744,38+3877,43+4003,99}{5}=3737,20$ |
| 2013 | 3877,43 | $\frac{y_4+y_5+y_6+y_7+y_8}{5}=\frac{3625,21+3744,38+3877,43+4003,99+4150,86}{5}=3880,37$ |
| 2014 | 4003,99 | $\frac{y_5+y_6+y_7+y_8+y_9}{5}=\frac{3744,38+3877,43+4003,99+4150,86+4290,52}{5}=4013,44$ |
| 2015 | 4150,86 | $\frac{y_6+y_7+y_8+y_9+y_{10}}{5}=\frac{3877,43+4003,99+4150,86+4290,52+4527,89}{5}=4170,14$ |
| 2016 | 4290,52 | $\frac{y_7+y_8+y_9+y_{10}+y_{11}}{5}=\frac{4003,99+4150,86+4290,52+4527,89+4834,76}{5}=4361,60$ |
| 2017 | 4527,89 | – |
| 2018 | 4834,76 | – |

Źródło: opracowanie własne.

**Przeciętne miesięczne wynagrodzenie brutto w Polsce w latach 2008-2018**




Źródło: opracowanie własne.

**Przykład 5.7.**

Oszacuj i dokonaj oceny stopnia dopasowania funkcji trendu. Wyznacz prognozę przeciętnych miesięcznych wynagrodzeń na 2019 r. (dane z przykładu 5.1).

**Rozwiązanie**

$\bar{y}_t = 3905{,}81$ zł, $\quad \bar{t} = 6$ lat

Obliczenia pomocnicze:

| Zmienna czasowa $t$ | Wynagrodzenia (w zł) $y_t$ | $y_t - \bar{y}_t$ | $t - \bar{t}$ | $(y_t - \bar{y}_t)(t - \bar{t})$ | $(t - \bar{t})^2$ |
|---|---|---|---|---|---|
| 1 | 3158,48 | -747,33 | -5 | 3736,65 | 25 |
| 2 | 3315,38 | -590,43 | -4 | 2361,72 | 16 |
| 3 | 3435,00 | -470,81 | -3 | 1412,43 | 9 |
| 4 | 3625,21 | -280,60 | -2 | 561,20 | 4 |
| 5 | 3744,38 | -161,43 | -1 | 161,43 | 1 |
| 6 | 3877,43 | -28,38 | 0 | 0,00 | 0 |
| 7 | 4003,99 | 98,18 | 1 | 98,18 | 1 |
| 8 | 4150,86 | 245,05 | 2 | 490,10 | 4 |
| 9 | 4290,52 | 384,71 | 3 | 1154,13 | 9 |
| 10 | 4527,89 | 622,08 | 4 | 2488,32 | 16 |
| 11 | 4834,76 | 928,95 | 5 | 4644,75 | 25 |
| - | - | - | - | **17108,91** | **110** |

Źródło: opracowanie własne.

1. **Obliczam współczynniki regresji liniowej (wzory 5.5.6 i 5.5.7):**

$$\hat{y}_t = a + bt$$

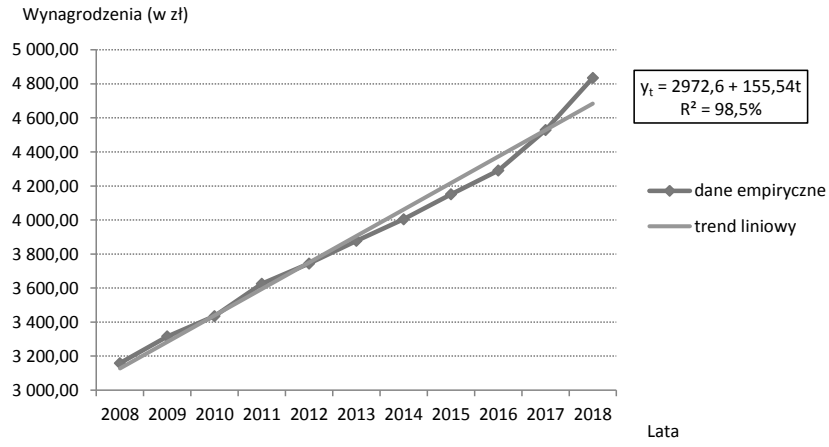
$$a = \frac{\sum_{t=1}^{n}(y_t - \bar{y}_t)(t - \bar{t})}{\sum_{t=1}^{n}(t - \bar{t})^2} = \frac{17108{,}91}{110} = 155{,}54$$

$$b = \bar{y}_t - a\bar{t} = 3905{,}81 - 155{,}54 \cdot 6 = 2972{,}57$$

Trend liniowy przyjmuje postać:

$$\hat{y}_t = 2972{,}57 + 155{,}54t$$

**Trend liniowy przeciętnych miesięcznych wynagrodzeń brutto w Polsce w latach 2008-2018**




Źródło: opracowanie własne.

**Interpretacja parametrów *a* i *b*:**

- $a = 2972{,}57$ – wyraz wolny *a* informuje o teoretycznym poziomie przeciętnych miesięcznych wynagrodzeń brutto w Polsce w okresie wyjściowym (dla *t* = 0), a więc dla *t* = 2007 r.,
- $b = 155{,}54$ – współczynnik kierunkowy *b* jest dodatni (*b* > 0) oznacza to, że średnioroczny wzrost przeciętnych miesięcznych wynagrodzeń brutto w Polsce w latach 2008-2018 wynosił 155,54 zł.

**2. Obliczam błędy standardowe parametrów trendu liniowego (wzory 5.5.8 i 5.5.9):**

Obliczenia pomocnicze:

| Zmienna czasowa $t$ | Wynagrodzenia (w zł) $y_t$ | $t^2$ |
|---|---|---|
| 1 | 3158,48 | 1 |
| 2 | 3315,38 | 4 |
| 3 | 3435,00 | 9 |
| 4 | 3625,21 | 16 |
| 5 | 3744,38 | 25 |
| 6 | 3877,43 | 36 |
| 7 | 4003,99 | 49 |
| 8 | 4150,86 | 64 |
| 9 | 4290,52 | 81 |
| 10 | 4527,89 | 100 |
| 11 | 4834,76 | 121 |
| - | - | **506** |

Źródło: opracowanie własne.

gdzie:

$n$ = 16 (województw)

$S_{e_{(t)}} = 67{,}5$ – wartość obliczona w części miary dopasowania pkt a).

$$S^2_{e_{(t)}} = 67{,}5^2 = 4556{,}25$$

$$S(a) = \sqrt{\frac{S^2_{e_{(t)}} \cdot \sum_{i=1}^{t} t_i^2}{n \cdot (\sum_{t=1}^{n} t_i^2 - n\bar{t}^2)}} = \sqrt{\frac{4556{,}25 \cdot 506}{11 \cdot (506 - 11 \cdot 6^2)}} = 43{,}65$$

$$S(b) = \frac{S_{e_{(t)}}}{\sqrt{\sum_{t=1}^{n} t_i^2 - n\bar{t}^2}} = \frac{67{,}5}{\sqrt{506 - 11 \cdot 6^2}} = 6{,}44$$

Błędy standardowe parametrów pokazujemy w nawiasach pod parametrami:

$$\hat{y}_t = \underset{(\pm 43{,}65)}{2972{,}57} + \underset{(\pm 6{,}44)}{155{,}54}t$$

**Interpretacja błędów standardowych ocen współczynników regresji *a* i *b*:**

- $S(a) = 43{,}65$ – współczynnik $a$ = 2972,57 odchyla się średnio o ± 43,65,
- $S(b) = 6{,}44$ – współczynnik $b$ = 155,54 odchyla się średnio o ± 6,44.

**Miary dopasowania trendu liniowego:**

**a) odchylenie standardowe składnika resztowego (wzór 5.5.10):**

Obliczenia pomocnicze:

| Wartości rzeczywiste $y_t$ | Zmienna czasowa $t$ | Wartości teoretyczne: $\hat{y}_t = 2972{,}57 + 155{,}54t$ | Reszty $y_t - \hat{y}_t$ | $(y_t - \hat{y}_t)^2$ |
|---|---|---|---|---|
| 3158,48 | 1 | 2972,57+(155,54 · 1) = 3128,11 | 30,37 | 922,34 |
| 3315,38 | 2 | 2972,57+(155,54 · 2) = 3283,65 | 31,73 | 1006,79 |
| 3435,00 | 3 | 2972,57+(155,54 · 3) = 3439,19 | -4,19 | 17,56 |
| 3625,21 | 4 | 2972,57+(155,54 · 4) = 3594,73 | 30,48 | 929,03 |
| 3744,38 | 5 | 2972,57+(155,54 · 5) = 3750,27 | -5,89 | 34,69 |
| 3877,43 | 6 | 2972,57+(155,54 · 6) = 3905,81 | -28,38 | 805,42 |
| 4003,99 | 7 | 2972,57+(155,54 · 7) = 4061,35 | -57,36 | 3290,17 |
| 4150,86 | 8 | 2972,57+(155,54 · 8) = 4216,89 | -66,03 | 4359,96 |
| 4290,52 | 9 | 2972,57+(155,54 · 9) = 4372,43 | -81,91 | 6709,25 |
| 4527,89 | 10 | 2972,57+(155,54 · 10) = 4527,97 | -0,08 | 0,01 |
| 4834,76 | 11 | 2972,57+(155,54 · 11) = 4683,51 | 151,25 | 22876,56 |
| - | - | - | - | **40951,78** |

Źródło: opracowanie własne.

$$S_{e_{(t)}} = \sqrt{\frac{\sum_{t=1}^{n}(y_t - \hat{y}_t)^2}{n-k}} = \sqrt{\frac{40951{,}78}{11-2}} = 67{,}5$$

**Interpretacja:** Odchylenie standardowe składnika resztowego informuje, że rzeczywiste wartości przeciętnych miesięcznych wynagrodzeń brutto odchylają się (różnią się) od wartości teoretycznych obliczonych na podstawie trendu średnio o ± 67,5 zł.

**b) współczynnik zmienności resztowej (wzór 5.5.11):**

$$V_{e_{(t)}} = \frac{S_{e_{(t)}}}{\bar{y}_t} \cdot 100 = \frac{67{,}5}{3905{,}81} \cdot 100 = 1{,}7\%$$

$$1{,}7\% \leq 10{,}0\%$$

$$V_{e_{(t)}} \leq V_g$$

**Interpretacja:** Odchylenie standardowe składnika resztowego stanowi tylko 1,7% średniej arytmetycznej przeciętnych miesięcznych wynagrodzeń brutto w Polsce w latach 2008-2018 co świadczy o bardzo dobrym dopasowaniu trendu liniowego do danych rzeczywistych.

c) **współczynnik zbieżności (wzór 5.5.12):**

Obliczenia pomocnicze:

| Wartości rzeczywiste $y_t$ | Wartości teoretyczne $\hat{y}_t$ | $y_t - \hat{y}_t$ | $(y_t - \hat{y}_t)^2$ | $y_t - \bar{y}_t$ | $(y_t - \bar{y}_t)^2$ |
|---|---|---|---|---|---|
| 3158,48 | 3128,11 | 30,37 | 922,34 | -747,33 | 558502,13 |
| 3315,38 | 3283,65 | 31,73 | 1006,79 | -590,43 | 348607,58 |
| 3435,00 | 3439,19 | -4,19 | 17,56 | -470,81 | 221662,06 |
| 3625,21 | 3594,73 | 30,48 | 929,03 | -280,60 | 78736,36 |
| 3744,38 | 3750,27 | -5,89 | 34,69 | -161,43 | 26059,64 |
| 3877,43 | 3905,81 | -28,38 | 805,42 | -28,38 | 805,42 |
| 4003,99 | 4061,35 | -57,36 | 3290,17 | 98,18 | 9639,31 |
| 4150,86 | 4216,89 | -66,03 | 4359,96 | 245,05 | 60049,50 |
| 4290,52 | 4372,43 | -81,91 | 6709,25 | 384,71 | 148001,78 |
| 4527,89 | 4527,97 | -0,08 | 0,01 | 622,08 | 386983,53 |
| 4834,76 | 4683,51 | 151,25 | 22876,56 | 928,95 | 862948,10 |
| - | - | - | **40951,78** | - | **2701995,41** |

Źródło: opracowanie własne.

$$\varphi^2 = \frac{\sum_{t=1}^{n}(y_t - \hat{y}_t)^2}{\sum_{t=1}^{n}(y_t - \bar{y}_t)^2} = \frac{40951{,}78}{2701995{,}41} = 0{,}015$$

$$\varphi^2 \cdot 100\% = 0{,}015 \cdot 100\% = 1{,}5\%$$

**Interpretacja:** Tylko 1,5% zmienności przeciętnego miesięcznego wynagrodzenia brutto w Polsce w latach 2008-2018 nie zostało wyjaśnione przez liniową funkcję trendu.

d) **współczynnik determinacji (wzór 5.5.13):**

$$R^2 = 1 - \varphi^2 = 1 - 0{,}015 = 0{,}985$$

$$R^2 \cdot 100\% = 0{,}985 \cdot 100\% = 98{,}5\%$$

**Interpretacja:** Zmienność przeciętnego miesięcznego wynagrodzenia brutto w Polsce w latach 2008-2018 została wyjaśniona w 98,5% przez liniową funkcję trendu.

**Prognoza punktowa wynagrodzeń na 2019 r. (dla $t = 12$):**

$$\hat{y}_{2019} = y^*_{2019} = 2972{,}57 + 155{,}54t = 2972{,}57 + 155{,}54 \cdot 12 = 4839{,}05 \text{ zł}$$

**Interpretacja:** Szacuje się, że przeciętne miesięczne wynagrodzenie brutto w 2019 r. może wynieść około 4839,05 zł.

**Prognoza przedziałowa (wzór 5.8.1)**

Dla poziomu istotności (testu dwustronnego) α = 0,05 i $df = 11 - 2 = 9$ stopni swobody dla małej próby ($n \leq 30$) odczytujemy z tablic rozkładu *t*-Studenta wartość krytyczną $t_\alpha = 2{,}2622$.

$$P\{y^*_{2019} - t_\alpha S_{pT} < y_t < y^*_{2019} + t_\alpha S_{pT}\} = 1 - \alpha$$
$$P\{4839{,}05 - 2{,}2622 \cdot 9{,}8 < y_t < 4839{,}05 + 2{,}2622 \cdot 9{,}8\} = 0{,}95$$
$$4816{,}88 < y_t < 4861{,}22$$

**Interpretacja:** Z 95% prawdopodobieństwem można przypuszczać że wielkość wynagrodzeń w 2019 r. będzie kształtować się w granicach wyznaczonego przedziału [4816,88 zł; 4861,22 zł].

**Wartości krytyczne dla testu t-Studenta**

| *df* | Poziom istotności dla testu jednostronnego | | | | | |
|---|---|---|---|---|---|---|
| | 0,1 | 0,05 | 0,025 | 0,01 | 0,005 | 0,0005 |
| | Poziom istotności dla testu dwustronnego | | | | | |
| | 0,2 | 0,1 | 0,05 | 0,02 | 0,01 | 0,001 |
| 1 | 3,077684 | 6,313752 | 12,706205 | 31,820516 | 63,656741 | 636,619249 |
| 2 | 1,885618 | 2,919986 | 4,302653 | 6,964557 | 9,924843 | 31,599055 |
| 3 | 1,637744 | 2,353363 | 3,182446 | 4,540703 | 5,840909 | 12,923979 |
| 4 | 1,533206 | 2,131847 | 2,776445 | 3,746947 | 4,604095 | 8,610302 |
| 5 | 1,475884 | 2,015048 | 2,570582 | 3,364930 | 4,032143 | 6,868827 |
| 6 | 1,439756 | 1,943180 | 2,446912 | 3,142668 | 3,707428 | 5,958816 |
| 7 | 1,414924 | 1,894579 | 2,364624 | 2,997952 | 3,499483 | 5,407883 |
| 8 | 1,396815 | 1,859548 | 2,306004 | 2,896459 | 3,355387 | 5,041305 |
| 9 | 1,383029 | 1,833113 | 2,262157 | 2,821438 | 3,249836 | 4,780913 |
| 10 | 1,372184 | 1,812461 | 2,228139 | 2,763769 | 3,169273 | 4,586894 |
| 11 | 1,363430 | 1,795885 | 2,200985 | 2,718079 | 3,105807 | 4,436979 |

Źródło: opracowane na podstawie: W. Artichowicz, *Statystyka. Korzystanie z podstawowych rozkładów prawdopodobieństwa*, Gdańsk, PG, 2014/2015, s. 17, https://artichowiczdotnet.files.wordpress.com/2016/03/statystykacwiczenia5.pdf

**Mierniki dokładności prognozy**

a) **średni błąd prognozy *ex ante* (wzór 5.8.2):**

$$S_{pT} = \sqrt{\left[\frac{(T-\bar{t})^2}{\sum_{t=1}^{n}(t-\bar{t})^2} + \frac{1}{n} + 1\right] \cdot S^2_{\hat{e}_{(t)}}} = \sqrt{\left[\frac{(12-6)^2}{110} + \frac{1}{11} + 1\right] \cdot 67{,}5} = 9{,}8$$

**Interpretacja:** Wartość odchyleń przeciętnych miesięcznych wynagrodzeń brutto od prognozy w 2019 r. wynosi 9,8 zł.

b) **względny błąd *ex ante* (wzór 5.8.3):**

$$V_T^* = \frac{S_{p2019}}{y_{2019}^*} \cdot 100\% = \frac{9{,}8}{4839{,}05} \cdot 100\% = 0{,}21\%$$

$$V_T^* \leq V_G$$

$$0{,}21\% \leq 5\%$$

**Interpretacja:** Ponieważ dla przyjętego poziomu błędu prognozy $V_G = 5\%$ spełniony jest warunek $V_T^* \leq V_G$ stwierdza się, że wartość prognozy na 2019 r. jest na poziomie dopuszczalnym.

c) **błąd prognozy *ex post* (wzór 5.8.6):**

$$y_{2019} = 5023{,}4 \text{ – wartość umowna}$$
$$\delta_T = y_{2019} - y_{2019}^* = 5023{,}4 - 4839{,}05 = 184{,}35 \text{ zł}$$

**Interpretacja:** Wartość rzeczywista różni się od wartości prognozowanej o 184,35 zł.

d) **względny błąd prognozy *ex post* (wzór 5.8.7):**

$$\delta_T^* = \frac{\delta_{2019}}{y_{2019}} \cdot 100\% = \frac{184{,}35}{5023{,}4} \cdot 100\% = 3{,}67\%$$

$$|\delta_T^*| \leq \delta_G$$

$$|3{,}67\%| \leq 5\%$$

**Interpretacja:** Ponieważ dla przyjętego poziomu błędu prognozy $\delta_G = 5\%$ spełniony jest warunek $\delta_T^* \leq \delta_G$ oznacza to, że prognozę na 2019 r. uznaje się za trafną.

**Przykład 5.8.**

Na podstawie danych kwartalnych liczby pracujących w latach 2014-2017 w Polsce wyznacz addytywne i multiplikatywne składniki sezonowości:

| Lata | Kwartały $Q$ | Liczba pracujących (w tys.) |
|---|---|---|
| 2014 | I | 15573 |
| 2014 | II | 15793 |
| 2014 | III | 16063 |
| 2014 | IV | 16018 |
| 2015 | I | 15837 |
| 2015 | II | 15986 |
| 2015 | III | 16234 |
| 2015 | IV | 16280 |
| 2016 | I | 16012 |
| 2016 | II | 16182 |
| 2016 | III | 16266 |
| 2016 | IV | 16328 |
| 2017 | I | 16281 |
| 2017 | II | 16496 |
| 2017 | III | 16510 |
| 2017 | IV | 16404 |

Źródło:, *Aktywność ekonomiczna ludności Polski IV kwartał 2017 r.*, GUS Warszawa 2018, s. 25.

Zgodnie z danymi:

$t = 16$ – liczba okresów,

$l = 4$ lata – liczba cykli,

$Q = 4$ kwartały.

**Rozwiązanie**

**I. Addytywne bezwzględne wahania sezonowe**

Parametry trendu liniowego liczby pracujących mają postać:

$$\hat{y}_t = \underset{(\pm 64{,}60)}{15724{,}83} + \underset{(\pm 6{,}68)}{49{,}01}t$$

Mierniki dopasowania trendu są następujące:

$$R^2 = 79{,}4\%,\ \ S_{e_{(t)}} = 123{,}2,\ \ V_{e_{(t)}} = 0{,}76\%$$

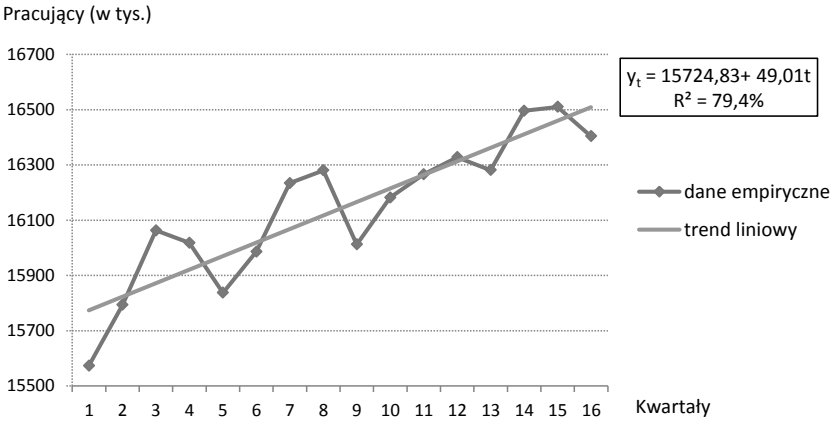
- **uwalniamy szereg czasowy od trendu (wzór 5.6.1):**

Obliczenia pomocnicze:

| $t$ | Lata | Kwartały $Q$ | $y_t$ | $\hat{y}_t$ | $d = y_t - \hat{y}_t$ |
|---|---|---|---|---|---|
| 1 | 2014 | I | 15573 | 15774 | -201 |
| 2 | 2014 | II | 15793 | 15823 | -30 |
| 3 | 2014 | III | 16063 | 15872 | 191 |
| 4 | 2014 | IV | 16018 | 15921 | 97 |
| 5 | 2015 | I | 15837 | 15970 | -133 |
| 6 | 2015 | II | 15986 | 16019 | -33 |
| 7 | 2015 | III | 16234 | 16068 | 166 |
| 8 | 2015 | IV | 16280 | 16117 | 163 |
| 9 | 2016 | I | 16012 | 16166 | -154 |
| 10 | 2016 | II | 16182 | 16215 | -33 |
| 11 | 2016 | III | 16266 | 16264 | 2 |
| 12 | 2016 | IV | 16328 | 16313 | 15 |
| 13 | 2017 | I | 16281 | 16362 | -81 |
| 14 | 2017 | II | 16496 | 16411 | 85 |
| 15 | 2017 | III | 16510 | 16460 | 50 |
| 16 | 2017 | IV | 16404 | 16509 | -105 |

Źródło: opracowanie własne.

**Trend liniowy liczby pracujących w Polsce w latach 2014-2017**




Źródło: opracowanie własne.

- **obliczam sumę odchyleń wartości rzeczywistych i empirycznych okresów jednoimiennych (*d*):**

| Lata | I kwartał | II kwartał | III kwartał | IV kwartał |
|---|---|---|---|---|
| 2014 | -201 | -30 | 191 | 97 |
| 2015 | -133 | -33 | 166 | 163 |
| 2016 | -154 | -33 | 2 | 15 |
| 2017 | -81 | 85 | 50 | -105 |
| **Suma** | **-569** | **-11** | **409** | **170** |

Źródło: opracowanie własne.

- **obliczam surowe wahania sezonowe (wzór 5.6.2)**:

| | |
|---|---|
| – dla I kwartału<br>$W_{s_I} = \frac{\sum_t (y_t - \hat{y}_t)}{l} = \frac{-569}{4} = -142{,}25$ | – dla II kwartału<br>$W_{s_{II}} = \frac{\sum_t (y_t - \hat{y}_t)}{l} = \frac{-11}{4} = -2{,}75$ |
| – dla III kwartału<br>$W_{s_{III}} = \frac{\sum_t (y_t - \hat{y}_t)}{l} = \frac{409}{4} = 102{,}25$ | – dla IV kwartału<br>$W_{s_{IV}} = \frac{\sum_t (y_t - \hat{y}_t)}{l} = \frac{170}{4} = 42{,}5$ |

Źródło: opracowanie własne.

$$\sum_{k=1}^{m} W_{s_k} = W_I + W_{II} + W_{III} + W_{IV} = -142{,}25 - 2{,}75 + 102{,}25 + 42{,}5 = -0{,}25$$

Suma surowych składników sezonowych wynosi -0,25, tak więc należy obliczyć współczynnik korygujący ($W_k$) według wzoru 5.6.3:

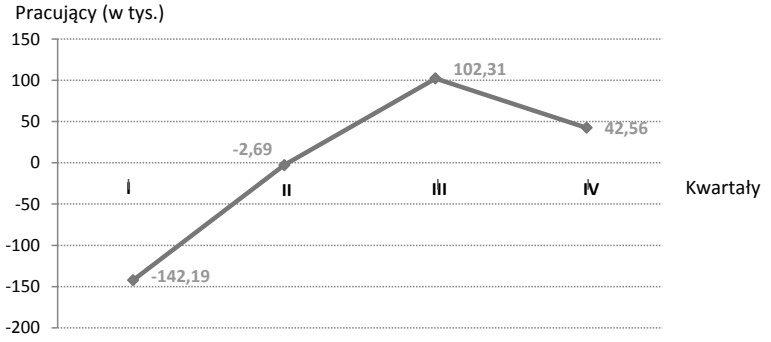
$$W_k = \frac{\sum_{k=1}^{m} W_{s_k}}{m} = \frac{-0{,}25}{4} = -0{,}06$$

- **obliczam czyste wahania sezonowości (wzór 5.6.4):**

| | |
|---|---|
| – dla I kwartału<br>$W_{c_I} = W_{s_k} - W_k = -142{,}25 - (-0{,}06 = -142{,}19$ | – dla II kwartału<br>$W_{c_{II}} = W_{s_k} - W_k = -2{,}75 - (-0{,}06) = -2{,}69$ |
| – dla III kwartału<br>$W_{c_{III}} = W_{s_k} - W_k = 102{,}25 - (-0{,}06) = 102{,}31$ | – dla IV kwartału<br>$W_{c_{IV}} = W_{s_k} - W_k = 42{,}5 - (-0{,}06) = 42{,}56$ |

Źródło: opracowanie własne.

**Kwartalne bezwzględne wahania sezonowości liczby pracujących w Polsce w latach 2014-2017**




Źródło: opracowanie własne.

**Interpretacja:** W każdym pierwszym i drugim kwartale w Polsce w latach 2014-2017 na skutek występowanie wahań sezonowych liczba pracujących była niższa średnio o: 142,19 tys. i 2,69 tys. osób w porównaniu do wartości wyznaczonych na podstawie trendu liniowego. Z kolei w trzecim i czwartym kwartale wpływ sezonowości spowodował wzrost liczby pracujących średnio o 102,31 tys. i 42,56 tys. osób.

## II. Multiplikatywne wskaźniki sezonowe

- **uwalniamy szereg czasowy od trendu (wzór 5.6.5):**

Obliczenia pomocnicze:

| $t$ | Lata | Kwartały $Q$ | $y_t$ | $\hat{y}_t$ | $\tilde{d}_t = \frac{y_t}{\hat{y}_t}$ |
|---|---|---|---|---|---|
| 1 | 2014 | I | 15573 | 15774 | 0,9873 |
| 2 | 2014 | II | 15793 | 15823 | 0,9981 |
| 3 | 2014 | III | 16063 | 15872 | 1,0120 |
| 4 | 2014 | IV | 16018 | 15921 | 1,0061 |
| 5 | 2015 | I | 15837 | 15970 | 0,9917 |
| 6 | 2015 | II | 15986 | 16019 | 0,9979 |
| 7 | 2015 | III | 16234 | 16068 | 1,0103 |
| 8 | 2015 | IV | 16280 | 16117 | 1,0101 |
| 9 | 2016 | I | 16012 | 16166 | 0,9905 |
| 10 | 2016 | II | 16182 | 16215 | 0,9980 |
| 11 | 2016 | III | 16266 | 16264 | 1,0001 |
| 12 | 2016 | IV | 16328 | 16313 | 1,0009 |
| 13 | 2017 | I | 16281 | 16362 | 0,9950 |
| 14 | 2017 | II | 16496 | 16411 | 1,0052 |
| 15 | 2017 | III | 16510 | 16460 | 1,0030 |
| 16 | 2017 | IV | 16404 | 16509 | 0,9936 |

Źródło: opracowanie własne.

- **obliczam sumę średnich wartości $\tilde{d}_t$ dla kwartałów I-IV:**

| Lata | I kwartał | II kwartał | III kwartał | IV kwartał |
|---|---|---|---|---|
| 2014 | 0,9873 | 0,9981 | 1,0120 | 1,0061 |
| 2015 | 0,9917 | 0,9979 | 1,0103 | 1,0101 |
| 2016 | 0,9905 | 0,9980 | 1,0001 | 1,0009 |
| 2017 | 0,9950 | 1,0052 | 1,0030 | 0,9936 |
| **Suma** | **3,9645** | **3,9992** | **4,0254** | **4,0107** |

Źródło: opracowanie własne.

- **obliczam surowe wskaźniki sezonowości (wzór 5.6.6)**:

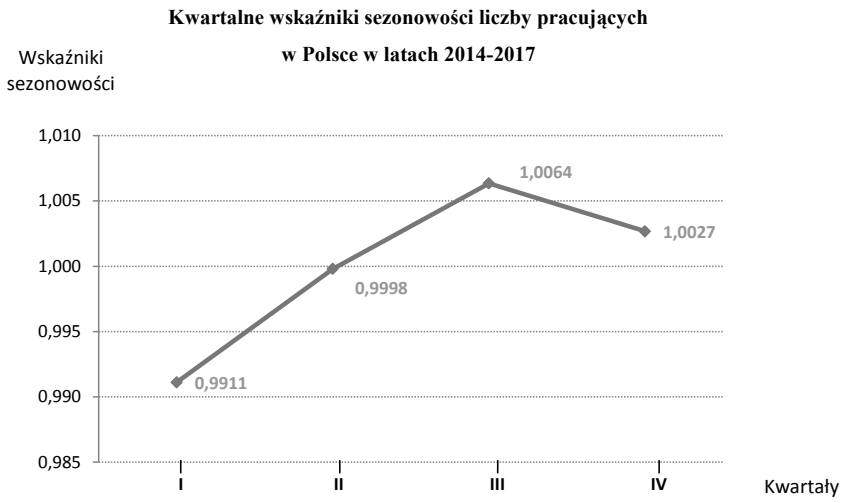
| | |
|---|---|
| – dla I kwartału $\widetilde{W}_{s_I} = \frac{\sum_t \frac{y_t}{\hat{y}_t}}{l} = \frac{3{,}9645}{4} = 0{,}9911$ | – dla II kwartału $\widetilde{W}_{s_{II}} = \frac{\sum_t \frac{y_t}{\hat{y}_t}}{l} = \frac{3{,}9992}{4} = 0{,}9998$ |
| – dla III kwartału $\widetilde{W}_{s_{III}} = \frac{\sum_t \frac{y_t}{\hat{y}_t}}{l} = \frac{4{,}0254}{4} = 1{,}0064$ | – dla IV kwartału $\widetilde{W}_{s_{IV}} = \frac{\sum_t \frac{y_t}{\hat{y}_t}}{l} = \frac{4{,}0107}{4} = 1{,}0027$ |

Źródło: opracowanie własne.

- **obliczam czyste wskaźniki sezonowości (wzór 5.6.7)**:

$$\widetilde{\widetilde{W}}_{sk} = \frac{0{,}9911 + 0{,}9998 + 1{,}0064 + 1{,}0027}{4} = 1{,}0$$

| | |
|---|---|
| – dla I kwartału $\widetilde{W}_{c_I} = \frac{\widetilde{W}_{s_k}}{\widetilde{\widetilde{W}}_{sk}} = \frac{0{,}9911}{1{,}0} = 0{,}9911$ | – dla II kwartału $\widetilde{W}_{c_{II}} = \frac{\widetilde{W}_{s_k}}{\widetilde{\widetilde{W}}_{sk}} = \frac{0{,}9998}{1{,}0} = 0{,}9998$ |
| – dla III kwartału $\widetilde{W}_{c_{III}} = \frac{\widetilde{W}_{s_k}}{\widetilde{\widetilde{W}}_{sk}} = \frac{1{,}0064}{1{,}0} = 1{,}0064$ | – dla IV kwartału $\widetilde{W}_{c_{IV}} = \frac{\widetilde{W}_{s_k}}{\widetilde{\widetilde{W}}_{sk}} = \frac{1{,}0027}{1{,}0} = 1{,}0027$ |




Źródło: opracowanie własne.

- **kwartał I:**

$$\widetilde{W}_{c_I} = (0{,}9911 \cdot 100\%) - 100\% = 99{,}11\% - 100\% = -0{,}89\%$$

- **kwartał II:**

$$\widetilde{W}_{c_{II}} = (0{,}9998 \cdot 100\%) - 100\% = 99{,}98\% - 100\% = -0{,}02\%$$

- **kwartał III:**

$$\widetilde{W}_{c_{III}} = (1{,}0064 \cdot 100\%) - 100\% = 100{,}64\% - 100\% = 0{,}64\%$$

- **kwartał IV:**

$$\widetilde{W}_{c_{IV}} = (1{,}0027 \cdot 100\%) - 100\% = 100{,}27\% - 100\% = 0{,}27\%$$

**Interpretacja:** W każdym pierwszym i drugim kwartale w Polsce w latach 2014-2017 na skutek występowanie sezonowości liczba pracujących była niższa odpowiednio o: 0,89% i 0,02% od przeciętnej kwartalnej równej 100%. Z kolei w trzecim i czwartym kwartale wpływ sezonowości powodował wzrost liczby pracujących odpowiednio o: 0,64% i 0,27%.

**Przykład 5.9.**

Przeprowadź analizę harmoniczną liczby pracujących w Polsce za lata 2014-2017 na podstawie danych z przykładu 5.8. W przykładzie ograniczono się do wyznaczenia tylko pierwszej i drugiej harmoniki ze względu na obszerność obliczeń.

Obliczenia w tym przykładzie najlepiej wykonywać w programie Excel.

**Składnia funkcji trygonometrycznych w Microsoft Excel**

Sinus **=SIN(liczba)**

Cosinus **=COS(liczba)**

arc tg (arcus tangens) **=ATAN(liczba)**

Funkcja **ATAN()** zwraca kąt w radianach z przedziału $\left[-\frac{\pi}{2},\frac{\pi}{2}\right]$, które zamieniamy na stopnie:

$$\alpha° = \frac{180^{\circ}\cdot \text{rad}}{\pi}$$

Możemy skorzystać z gotowych komend:

**=ATAN(liczba)*180/PI()**

**=STOPNIE(ATAN(liczba))**

**Pamiętaj, że:**

$$2\pi = 360°$$

$$\pi = 180°$$

Źródło: opracowane na podstawie https://support.office.com/pl-pl/article/stopnie-funkcji

## Rozwiązanie

Obliczenia pomocnicze harmoniki pierwszej:

| $t$ | $y_t$ | $x = \frac{2\pi}{n}it$ | $sinx$ | $cosx$ | $y_t sinx$ | $y_t cosx$ |
|---|---|---|---|---|---|---|
| 1 | 15573 | 0,3927 | 0,3827 | 0,9239 | 5959,5 | 14387,6 |
| 2 | 15793 | 0,7854 | 0,7071 | 0,7071 | 11167,3 | 11167,3 |
| 3 | 16063 | 1,1781 | 0,9239 | 0,3827 | 14840,3 | 6147,0 |
| 4 | 16018 | 1,5708 | 1,0000 | 0,0000 | 16018,0 | 0,0 |
| 5 | 15837 | 1,9635 | 0,9239 | -0,3827 | 14631,5 | -6060,6 |
| 6 | 15986 | 2,3562 | 0,7071 | -0,7071 | 11303,8 | -11303,8 |
| 7 | 16234 | 2,7489 | 0,3827 | -0,9239 | 6212,5 | -14998,3 |
| 8 | 16280 | 3,1416 | 0,0000 | -1,0000 | 0,0 | -16280,0 |
| 9 | 16012 | 3,5343 | -0,3827 | -0,9239 | -6127,5 | -14793,2 |
| 10 | 16182 | 3,9270 | -0,7071 | -0,7071 | -11442,4 | -11442,4 |
| 11 | 16266 | 4,3197 | -0,9239 | -0,3827 | -15027,8 | -6224,7 |
| 12 | 16328 | 4,7124 | -1,0000 | 0,0000 | -16328,0 | 0,0 |
| 13 | 16281 | 5,1051 | -0,9239 | 0,3827 | -15041,7 | 6230,5 |
| 14 | 16496 | 5,4978 | -0,7071 | 0,7071 | -11664,4 | 11664,4 |
| 15 | 16510 | 5,8905 | -0,3827 | 0,9239 | -6318,1 | 15253,3 |
| 16 | 16404 | 6,2832 | 0,0000 | 1,0000 | 0,0 | 16404,0 |
| ***n*=16** | **258263** | - | - | - | **-1817,1** | **151,2** |

Źródło: opracowanie własne.

- **liczba harmonik wynosi:**

$$\frac{n}{2} = \frac{16}{2} = 8$$

Obliczenia pomocnicze:

| Numer harmoniki | Liczba kwartałów | Liczba lat |
|---|---|---|
| 1 | 16 | 4 |
| 2 | 8 | 2 |
| 3 | 5,3 | 1,3 |
| 4 | 4 | 1 |
| 5 | 3,2 | 0,8 |
| 6 | 2.67 | 0,67 |
| 7 | 2,29 | 0,57 |
| 8 | 2 | 0,5 |

Źródło: opracowanie własne.

- **obliczam parametr $a_0$ (wzór 5.6.10):**

$$a_0 = \frac{\sum_{i=1}^{n} y_t}{n} = \frac{258263}{16} = 16141{,}44$$

- obliczam **wariancję (wzór 2.2.1):**

Obliczenia pomocnicze:

| $t$ | $y_t$ | $y_i - \bar{y}$ | $(y_i - \bar{y})^2$ |
|---|---|---|---|
| 1 | 15573 | -568,4 | 323124,0 |
| 2 | 15793 | -348,4 | 121410,4 |
| 3 | 16063 | -78,4 | 6152,8 |
| 4 | 16018 | -123,4 | 15237,4 |
| 5 | 15837 | -304,4 | 92683,7 |
| 6 | 15986 | -155,4 | 24161,6 |
| 7 | 16234 | 92,6 | 8567,4 |
| 8 | 16280 | 138,6 | 19198,9 |
| 9 | 16012 | -129,4 | 16754,7 |
| 10 | 16182 | 40,6 | 1645,1 |
| 11 | 16266 | 124,6 | 15515,2 |
| 12 | 16328 | 186,6 | 34804,6 |
| 13 | 16281 | 139,6 | 19477,0 |
| 14 | 16496 | 354,6 | 125712,8 |
| 15 | 16510 | 368,6 | 135836,5 |
| 16 | 16404 | 262,6 | 68937,8 |
| ***n*=16** | **258263** | **-** | **1029219,9** |

Źródło: opracowanie własne.

$$S_y^2 = \frac{\sum_{i=1}^{N} (y_i - \bar{y})^2}{n} = \frac{1029219{,}9}{16} = 64326{,}24$$

- **obliczam parametry $a_1$ i $b_1$ (wzory 5.6.11 i 5.6.12):**

$$a_1 = \frac{2}{n}\sum_{t=1}^{n} y_t \, sin\left(\frac{2\pi}{n} it\right) = \frac{-1817{,}1}{8} = -227{,}1$$

$$b_1 = \frac{2}{n}\sum_{t=1}^{n} y_t \, cos\left(\frac{2\pi}{n} it\right) = \frac{151{,}2}{8} = 18{,}9$$

- **obliczam procent wyjaśnionej wariancji (wzór 5.6.18):**

$$A_1 = \frac{0{,}5 \cdot (a_1^2 + b_1^2)}{S_y^2} \cdot 100 = \frac{0{,}5 \cdot (-227{,}1^2 + 18{,}9^2)}{64326{,}24} \cdot 100 = 40{,}4\%$$

**Interpretacja:** Pierwsza harmonika (w okresie 4 letnim) wyjaśnia 40,4% ogólnej zmienności zmiennej $y_t$.

- **obliczam amplitudę harmoniki pierwszej (wzór 5.6.14):**

$$A_1 = \sqrt{a_1^2 + b_1^2} = \sqrt{-227{,}1^2 + 18{,}9^2} = 227{,}9$$

Amplituda wahań wynosi 227,9.

- **przesunięcie fazowe amplitudy pierwszej na osi czasu (wzory 5.6.15-5.6.17):**

$$\varepsilon_1 = arc\ tg\left(\frac{a_1}{b_1}\right) = arc\ tg\left(\frac{-227{,}1}{18{,}9}\right) = -1{,}4878$$

$$\varepsilon_1 = \alpha° = \frac{180^{\circ} \cdot \text{rad}}{\pi} = \frac{180^{\circ} \cdot (-1{,}4878)}{\pi} \approx -85{,}24$$

$$\theta_1 = \frac{2\pi i}{n} = \frac{360° \cdot 1}{16} = 22{,}5$$

$$p_{f_1} = \frac{\varepsilon_1}{\theta_1} = \frac{-85{,}24}{22{,}5} = -3{,}8$$

**Model możemy zapisać w postaci:**

$$\hat{y}_t = 16141{,}44 - 227{,}1 \, sin\left(\frac{\pi}{2} t\right) + 18{,}9 \, cos\left(\frac{\pi}{2} t\right)$$

Obliczamy parametry drugiej harmoniki dzieląc szereg czasowy na dwie równe części:

Obliczenia pomocnicze harmoniki drugiej:

| $t$ | $y_t$ | $x = \frac{2\pi}{n} it$ | $sinx$ | $cosx$ | $y_t sinx$ | $y_t cosx$ |
|---|---|---|---|---|---|---|
| 1 | 15573 | 0,7854 | 0,7071 | 0,7071 | 11011,77 | 11011,77 |
| 2 | 15793 | 1,5708 | 1,0000 | 0,0000 | 15793,00 | 0,00 |
| 3 | 16063 | 2,3562 | 0,7071 | -0,7071 | 11358,26 | -11358,26 |
| 4 | 16018 | 3,1416 | 0,0000 | -1,0000 | 0,00 | -16018,00 |
| 5 | 15837 | 3,9270 | -0,7071 | -0,7071 | -11198,45 | -11198,45 |
| 6 | 15986 | 4,7124 | -1,0000 | 0,0000 | -15986,00 | 0,00 |
| 7 | 16234 | 5,4978 | -0,7071 | 0,7071 | -11479,17 | 11479,17 |
| 8 | 16280 | 6,2832 | 0,0000 | 1,0000 | 0,00 | 16280,00 |
| ***n*=8** | **127784** | - | - | - | **-500,59** | **196,24** |
| 9 | 16012 | 0,7854 | 0,7071 | 0,7071 | 11322,19 | 11322,19 |
| 10 | 16182 | 1,5708 | 1,0000 | 0,0000 | 16182,00 | 0,00 |
| 11 | 16266 | 2,3562 | 0,7071 | -0,7071 | 11501,80 | -11501,80 |
| 12 | 16328 | 3,1416 | 0,0000 | -1,0000 | 0,00 | -16328,00 |
| 13 | 16281 | 3,9270 | -0,7071 | -0,7071 | -11512,41 | -11512,41 |
| 14 | 16496 | 4,7124 | -1,0000 | 0,0000 | -16496,00 | 0,00 |
| 15 | 16510 | 5,4978 | -0,7071 | 0,7071 | -11674,33 | 11674,33 |
| 16 | 16404 | 6,2832 | 0,0000 | 1,0000 | 0,00 | 16404,00 |
| ***n*=8** | **130479** | - | - | - | **-676,75** | **58,32** |
| **Suma** | **258263** | - | - | - | **-1177,34** | **254,56** |

Źródło: opracowanie własne.

- **obliczam parametry $a_2$ *i* $b_2$ (wzory 5.6.11 i 5.6.12):**

$$a_2 = \frac{2}{n}\sum_{t=1}^{n} y_t \, sin\left(\frac{2\pi}{n} it\right) = \frac{-1177{,}34}{8} = -147{,}17$$

$$b_2 = \frac{2}{n}\sum_{t=1}^{n} y_t \, cos\left(\frac{2\pi}{n} it\right) = \frac{254{,}56}{8} = 31{,}82$$

- **obliczam procent wyjaśnionej wariancji (wzór 5.6.18):**

$$A_2 = \frac{0{,}5 \cdot (a_1^2 + b_1^2)}{S_y^2} \cdot 100 = \frac{0{,}5 \cdot (-147{,}17^2 + 31{,}82^2)}{64326{,}24} \cdot 100 = 17{,}62\%$$

**Interpretacja:** Druga harmonika (w okresie 2 letnim) wyjaśnia 17,62% ogólnej zmienności zmiennej $y_t$.

- **obliczam amplitudę harmoniki drugiej (wzór 5.6.14):**

$$A_2 = \sqrt{a_1^2 + b_1^2} = \sqrt{-147{,}17^2 + 31{,}82^2} = 150{,}57$$

Amplituda wahań wynosi 150,57.

- **przesunięcie fazowe amplitudy drugiej na osi czasu (wzory 5.6.15-5.6.17):**

$$\varepsilon_2 = arc\ tg\left(\frac{a_1}{b_1}\right) = arc\ tg\left(\frac{-147{,}17}{31{,}82}\right) = -1{,}3579$$

$$\varepsilon_2 = \alpha° = \frac{180^\circ \cdot \text{rad}}{\pi} = \frac{180^\circ \cdot (-1{,}3579)}{\pi} \approx -77{,}80$$

$$\theta_2 = \frac{2\pi i}{n} = \frac{360° \cdot 2}{16} = 45{,}0$$

$$p_{f_2} = \frac{\varepsilon_2}{\theta_2} = \frac{-77{,}80}{45} = -1{,}7$$

**Wtedy model możemy zapisać w postaci:**

$$\hat{y}_t = 16141{,}44 - 147{,}17\ sin\left(\frac{\pi}{2}t\right) + 31{,}82\ cos\left(\frac{\pi}{2}t\right)$$

Harmoniki pierwsza i druga wyjaśniają 58,02% wariancji zmiennej $y_t$.

Analizę harmoniczną ze średnią stosuje się w szeregach czasowych, w których nie występuje tendencja rozwojowa. W analizowanym szeregu liczby pracujących w Polsce zidentyfikowano trend liniowy o postaci:

$$\hat{y}_t = 15724{,}83 + 49{,}01t$$

Stąd też, zastosujemy analizę harmoniczną z trendem (wzór 5.6.9). Analizowany szereg składa się z reszt $\tilde{x}_t$ obliczonych pomiędzy wartościami rzeczywistymi a teoretycznymi.

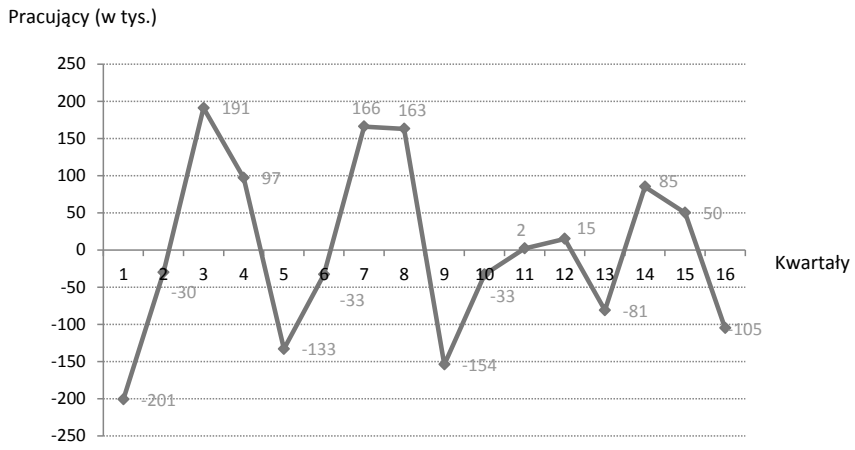
## Rozwiązanie

Obliczenia pomocnicze:

| $t$ | *Reszty* $\tilde{x}_t = y_t - f(t)$ | $x = \frac{2\pi}{n} it$ | $sinx$ | $cosx$ | $\tilde{x}_t sinx$ | $\tilde{x}_t cosx$ |
|---|---|---|---|---|---|---|
| 1 | -201 | 0,3927 | 0,3827 | 0,9239 | -76,92 | -185,70 |
| 2 | -30 | 0,7854 | 0,7071 | 0,7071 | -21,21 | -21,21 |
| 3 | 191 | 1,1781 | 0,9239 | 0,3827 | 176,46 | 73,09 |
| 4 | 97 | 1,5708 | 1,0000 | 0,0000 | 97,00 | 0,00 |
| 5 | -133 | 1,9635 | 0,9239 | -0,3827 | -122,88 | 50,90 |
| 6 | -33 | 2,3562 | 0,7071 | -0,7071 | -23,33 | 23,33 |
| 7 | 166 | 2,7489 | 0,3827 | -0,9239 | 63,53 | -153,36 |
| 8 | 163 | 3,1416 | 0,0000 | -1,0000 | 0,00 | -163,00 |
| 9 | -154 | 3,5343 | -0,3827 | -0,9239 | 58,93 | 142,28 |
| 10 | -33 | 3,9270 | -0,7071 | -0,7071 | 23,33 | 23,33 |
| 11 | 2 | 4,3197 | -0,9239 | -0,3827 | -1,85 | -0,77 |
| 12 | 15 | 4,7124 | -1,0000 | 0,0000 | -15,00 | 0,00 |
| 13 | -81 | 5,1051 | -0,9239 | 0,3827 | 74,83 | -31,00 |
| 14 | 85 | 5,4978 | -0,7071 | 0,7071 | -60,10 | 60,10 |
| 15 | 50 | 5,8905 | -0,3827 | 0,9239 | -19,13 | 46,19 |
| 16 | -105 | 6,2832 | 0,0000 | 1,0000 | 0,00 | -105,00 |
| ***n*=16** | **1,00** | **-** | **-** | **-** | **153,66** | **-240,81** |

Źródło: opracowanie własne.

**Liczba pracujących w Polsce w latach 2014-2017 (wykres reszt $\tilde{x}_t$)**




Źródło: opracowanie własne.

- **obliczam wariancję (wzór 2.2.1):**

$$\bar{x} = -0{,}0625$$

$$S_y^2 = \frac{\sum_{i=1}^{N}(\tilde{x}_t - \bar{x})^2}{n} = \frac{212438{,}94}{16} = 13277{,}43$$

Obliczenia pomocnicze:

| $t$ | $\tilde{x}_t$ | $\tilde{x}_t - \bar{x}$ | $(\tilde{x}_t - \bar{x})^2$ |
|---|---|---|---|
| 1 | -201 | -200,94 | 40375,88 |
| 2 | -30 | -29,94 | 896,25 |
| 3 | 191 | 191,06 | 36504,88 |
| 4 | 97 | 97,06 | 9421,13 |
| 5 | -133 | -132,94 | 17672,38 |
| 6 | -33 | -32,94 | 1084,88 |
| 7 | 166 | 166,06 | 27576,75 |
| 8 | 163 | 163,06 | 26589,38 |
| 9 | -154 | -153,94 | 23696,75 |
| 10 | -33 | -32,94 | 1084,88 |
| 11 | 2 | 2,06 | 4,25 |
| 12 | 15 | 15,06 | 226,88 |
| 13 | -81 | -80,94 | 6550,88 |
| 14 | 85 | 85,06 | 7235,63 |
| 15 | 50 | 50,06 | 2506,25 |
| 16 | -105 | -104,94 | 11011,88 |
| ***n*=16** | **1,00** | **-** | **212438,94** |

Źródło: opracowanie własne.

- **obliczam parametry $a_1$ *i* $b_1$ (wzory 5.6.11 i 5.6.12):**

$$a_1 = \frac{2}{n}\sum_{t=1}^{n} y_t \, sin\left(\frac{2\pi}{n} it\right) = \frac{153{,}66}{8} = 19{,}21$$

$$b_1 = \frac{2}{n}\sum_{t=1}^{n} y_t \, cos\left(\frac{2\pi}{n} it\right) = \frac{-240{,}81}{8} = -30{,}1$$

- **obliczam procent wyjaśnionej wariancji (wzór 5.6.18):**

$$A_1 = \frac{0{,}5 \cdot (a_1^2 + b_1^2)}{S_{\tilde{y}}^2} \cdot 100 = \frac{0{,}5 \cdot ((19{,}21^2 + (-30{,}1^2))}{13277{,}43} \cdot 100 = 4{,}80\%$$

**Interpretacja:** Pierwsza harmonika (w okresie 4 letnim) wyjaśnia zaledwie 4,8% ogólnej zmienności zmiennej $y_t$.

- **obliczam amplitudę harmoniki pierwszej (wzór 5.6.14):**

$$A_1 = \sqrt{a_1^2 + b_1^2} = \sqrt{19{,}21^2 + (-30{,}1)^2} = 35{,}71$$

Amplituda wahań wynosi 35,71.

- **przesunięcie fazowe amplitudy pierwszej na osi czasu (wzory 5.6.15-5.6.17):**

$$\varepsilon_1 = arc\ tg\left(\frac{a_1}{b_1}\right) = arc\ tg\left(\frac{19{,}21}{-30{,}1}\right) = -0{,}5680$$

$$\varepsilon_1 = \alpha° = \frac{180^\circ \cdot \text{rad}}{\pi} = \frac{180^\circ \cdot (-0{,}5680)}{\pi} \approx -32{,}55$$

$$\theta_1 = \frac{2\pi i}{n} = \frac{360° \cdot 1}{16} = 22{,}5$$

$$p_{f_1} = \frac{\varepsilon_1}{\theta_1} = \frac{-32{,}55}{22{,}5} = -1{,}4$$

**Model z trendem możemy zapisać w postaci:**

$$\hat{y}_t = 15724{,}83 + 49{,}01t + 19{,}21\ sin\left(\frac{\pi}{2}t\right) - 30{,}1\ cos\left(\frac{\pi}{2}t\right)$$

Obliczamy parametry drugiej harmoniki dzieląc szereg czasowy na dwie równe części:

Obliczenia pomocnicze harmoniki drugiej:

| $t$ | $\tilde{x}_t = y_t - f(t)$ | $x = \frac{2\pi}{n}it$ | $sinx$ | $cosx$ | $\tilde{x}_t sinx$ | $\tilde{x}_t cosx$ |
|---|---|---|---|---|---|---|
| 1 | -201 | 0,7854 | 0,7071 | 0,7071 | -142,13 | -142,13 |
| 2 | -30 | 1,5708 | 1,0000 | 0,0000 | -30,00 | 0,00 |
| 3 | 191 | 2,3562 | 0,7071 | -0,7071 | 135,06 | -135,06 |
| 4 | 97 | 3,1416 | 0,0000 | -1,0000 | 0,00 | -97,00 |
| 5 | -133 | 3,9270 | -0,7071 | -0,7071 | 94,05 | 94,05 |
| 6 | -33 | 4,7124 | -1,0000 | 0,0000 | 33,00 | 0,00 |
| 7 | 166 | 5,4978 | -0,7071 | 0,7071 | -117,38 | 117,38 |
| 8 | 163 | 6,2832 | 0,0000 | 1,0000 | 0,00 | 163,00 |
| ***n*=8** | **220** | - | - | - | **-27,41** | **0,24** |
| 9 | -154 | 0,7854 | 0,7071 | 0,7071 | -108,89 | -108,89 |
| 10 | -33 | 1,5708 | 1,0000 | 0,0000 | -33,00 | 0,00 |
| 11 | 2 | 2,3562 | 0,7071 | -0,7071 | 1,41 | -1,41 |
| 12 | 15 | 3,1416 | 0,0000 | -1,0000 | 0,00 | -15,00 |
| 13 | -81 | 3,9270 | -0,7071 | -0,7071 | 57,28 | 57,28 |
| 14 | 85 | 4,7124 | -1,0000 | 0,0000 | -85,00 | 0,00 |
| 15 | 50 | 5,4978 | -0,7071 | 0,7071 | -35,36 | 35,36 |
| 16 | -105 | 6,2832 | 0,0000 | 1,0000 | 0,00 | -105,00 |
| ***n*=8** | **-221** | - | - | - | **-203,56** | **-137,68** |
| **Suma** | **1,00** | - | - | - | **-230,97** | **-137,44** |

Źródło: opracowanie własne.

- **obliczam parametry $a_2$ i $b_2$ (wzory 5.6.11 i 5.6.12):**

$$a_2 = \frac{2}{n}\sum_{t=1}^{n} y_t\ sin\left(\frac{2\pi}{n}it\right) = \frac{-230{,}97}{8} = -28{,}87$$

$$b_2 = \frac{2}{n}\sum_{t=1}^{n} y_t\ cos\left(\frac{2\pi}{n}it\right) = \frac{-137{,}44}{8} = -17{,}18$$

- **obliczam procent wyjaśnionej wariancji (wzór 5.6.18):**

$$A_2 = \frac{0{,}5 \cdot (a_1^2 + b_1^2)}{S_y^2} \cdot 100 = \frac{0{,}5 \cdot (-28{,}87^2) + (-17{,}18^2)}{13277{,}43} \cdot 100 = 4{,}25\%$$

**Interpretacja:** Druga harmonika (w okresie 2 letnim) wyjaśnia tylko 4,25% ogólnej zmienności zmiennej $y_t$.

- **obliczam amplitudę harmoniki drugiej (wzór 5.6.14):**

$$A_2 = \sqrt{a_1^2 + b_1^2} = \sqrt{(-28{,}87^2) + (-17{,}18^2)} = 33{,}6$$

Amplituda wahań wynosi 33,6.

- **przesunięcie fazowe amplitudy drugiej na osi czasu (wzory 5.6.15-5.6.17):**

$$\varepsilon_2 = arc\ tg\left(\frac{a_1}{b_1}\right) = arc\ tg\left(\frac{-28{,}87}{-17{,}18}\right) = 1{,}034$$

$$\varepsilon_2 = \alpha° = \frac{180^{\circ} \cdot \text{ rad}}{\pi} = \frac{180^{\circ} \cdot 1{,}034}{\pi} \approx 59{,}24$$

$$\theta_2 = \frac{2\pi i}{n} = \frac{360° \cdot 2}{16} = 45{,}0$$

$$p_{f_2} = \frac{\varepsilon_2}{\theta_2} = \frac{59{,}24}{45} = 1{,}3$$

**Model z trendem możemy zapisać w postaci:**

$$\hat{y}_t = 15724{,}83 + 49{,}01t - 28{,}87\ sin\left(\frac{\pi}{2}t\right) - 17{,}18\ cos\left(\frac{\pi}{2}t\right)$$

Pierwsza (4,8%) i druga (4.25%) harmonika wyjaśniają tylko 9,05% całkowitej wariancji zmiennej prognozowanej. Stąd, też należy obliczyć i poszukać harmoniki, które będą miały najwyższe udziały w wyjaśnianiu zmienności zmiennej $y_t$.

**Przykład 5.10.**

Wykorzystując dane z przykładu 5.8 oblicz addytywne wahania przypadkowe.

**Rozwiązanie**

Dekompozycja szeregu czasowego liczby pracujących przedstawia poniższa tabela:

Obliczenia pomocnicze:

| $t$ | Lata | Kwartały $Q$ | Liczba pracujących $y_t$ | Trend liniowy $\hat{y}_t$ | Wahania sezonowe $W_{c_k}$ | Wahania przypadkowe $e_t = y_t - \hat{y}_t - W_{c_k}$ | $e_t^2$ |
|---|---|---|---|---|---|---|---|
| 1 | 2014 | I | 15573 | 15774 | -142,19 | -58,8 | 3457,4 |
| 2 | 2014 | II | 15793 | 15823 | -2,69 | -27,3 | 745,3 |
| 3 | 2014 | III | 16063 | 15872 | 102,31 | 88,7 | 7867,7 |
| 4 | 2014 | IV | 16018 | 15921 | 42,56 | 54,4 | 2959,4 |
| 5 | 2015 | I | 15837 | 15970 | -142,19 | 9,2 | 84,6 |
| 6 | 2015 | II | 15986 | 16019 | -2,69 | -30,3 | 918,1 |
| 7 | 2015 | III | 16234 | 16068 | 102,31 | 63,7 | 4057,7 |
| 8 | 2015 | IV | 16280 | 16117 | 42,56 | 120,4 | 14496,2 |
| 9 | 2016 | I | 16012 | 16166 | -142,19 | -11,8 | 139,2 |
| 10 | 2016 | II | 16182 | 16215 | -2,69 | -30,3 | 918,1 |
| 11 | 2016 | III | 16266 | 16264 | 102,31 | -100,3 | 10060,1 |
| 12 | 2016 | IV | 16328 | 16313 | 42,56 | -27,6 | 761,8 |
| 13 | 2017 | I | 16281 | 16362 | -142,19 | 61,2 | 3745,4 |
| 14 | 2017 | II | 16496 | 16411 | -2,69 | 87,7 | 7691,3 |
| 15 | 2017 | III | 16510 | 16460 | 102,31 | -52,3 | 2735,3 |
| 16 | 2017 | IV | 16404 | 16509 | 42,56 | -147,6 | 21785,8 |
| **Suma** | - | - | - | - | - | - | **82423,4** |

Źródło: opracowanie własne.

**Interpretacja:**

Na liczbę pracujących np. z 2017 r. w IV kwartale – 16404 tys. składają się następujące składniki:

- liczba pracujących będąca wynikiem czynników głównych – trend (16509 tys.),
- liczba pracujących będąca wynikiem działania wahań sezonowych (42,56 tys.),
- liczba pracujących będąca wynikiem wahań przypadkowych (-147,6 tys.).

Oznacza to, że:

16509 tys. + 42,56 tys. – 147,6 tys. = 16404 tys. pracujących

- **obliczamy odchylenie składnika resztowego (wzór 5.7.2):**

$$S(e_t) = \sqrt{\frac{\sum_{t=1}^{n} e_t^2}{n-k}} = \sqrt{\frac{82423{,}4}{12-2}} = 90{,}8$$

**Interpretacja:** Przeciętna siła wpływu wahań przypadkowych w kwartale w badanym szeregu czasowym wynosi ±90,8 tys. pracujących.

- **obliczamy współczynnik zmienności resztowej (wzór 5.7.3):**

$$V(e_t) = \frac{S(e_t)}{\bar{y}_t} \cdot 100 = \frac{90{,}8}{16141{,}4} \cdot 100 = 0{,}56\%$$

**Interpretacja:** Odchylenia przypadkowe stanowią tylko 0,56% przeciętnego poziomu liczby pracujących.

## Literatura

**Aczel A.D., Sounderpandian J.,** *Statystyka w zarządzaniu,* PWN, Warszawa 2017.

**Artichowicz W.,** *Statystyka. Korzystanie z podstawowych rozkładów prawdopodobieństwa*, PG Gdańsk, 2014/2015, s. 17. Materiały do ćwiczeń.

**Augustyniak H.,** *Statystyka opisowa z elementami demografii*, Poznań 2002.

**Cieślak M., (red.),** *Prognozowanie gospodarcze. Metody i zastosowania*, PWN, Warszawa 2005.

**Czyżycki R., Klóska R., (red),** *Ekonometria i prognozowanie zjawisk ekonomicznych w przykładach i zadaniach*, Economicus, Szczecin 2011.

**Dolny E., Osińska M.,** *Statystyka opisowa*, WSG, Bydgoszcz 2009.

**Gatnar E., Walesiak M. (red.)**, *Statystyczna analiza danych z wykorzystaniem programu R,* PWN, Warszawa 2009.

GUS, *Rocznik Statystyczny Rzeczpospolitej Polski 2018 r*., Warszawa 2018.

GUS, *Aktywność ekonomiczna ludności Polski IV kwartał 2017 r.*, Warszawa 2018.

GUS, *Zasady Metodyczne Statystyki Rynku Pracy i Wynagrodzeń*, Warszawa 2008.

**Ignatczyk W., Chromińska M.,** *Statystyka. Teoria i zastosowanie*, WSB, Poznań 2004.

**Jóźwiak J., Podgórski J.,** *Statystyka od podstaw*, PWE, Warszawa 2012.

**Kukuła K.,** *Elementy statystyki w zadaniach*, PWN, Warszawa 2016.

**Lemańczyk L.,** *Zbiór zadań ze statystyki medycznej*, UM, Poznań 2008.

**Maksimowicz-Ajchel A.,** *Wstęp do statystyki. Metody opisu statystycznego*, UW, Warszawa 2017.

**Osińska M. (red.),** *Ekonometria współczesna*, Dom Organizatora, Toruń 2007.

**Paradysz J. (red.),** *Statystyka*, red. naukowa AE, Poznań 2005.

**Parlińska M., Parliński J.,** *Statystyczna analiza danych z Excelem*, SGGW, Warszawa 2011.

**Piłatowska M.,** *Repetytorium ze statystyki*, PWN, Warszawa 2016.

**Rabiej M.,** *Statystyka z programem Statistica*, Helion, Gliwice 2012.

**Regel W.,** *Podstawy statystyki w Excelu*, PWN, Warszawa 2019.

**Rubacha K.,** *Metodologia badań nad edukacją,* WAiP, Warszawa 2008.

**Starzyńska W.,** *Statystyka praktyczna*, PWN, Warszawa 2019.

**Starzyńska W. (red.),** *Podstawy Statystyki,* Difin, Warszawa 2015.

**Sobczyk M.,** *Statystyka,* PWN, Warszawa 2016.

**Sobczyk M.,** *Statystyka, aspekty praktyczne i teoretyczne,* UMCS, Warszawa 2006.

**Strahl D., Sobczak E., Markowska M., Bal-Domańska B.,** *Modelowanie ekonometryczne z Excelem*, Materiały pomocnicze do laboratoriów z ekonometrii, UE, Wrocław 2015.

**Zając K.,** *Zarys metod statystycznych*, PWE, Warszawa 1994.

**Zeliaś A., Pawełek B., Wanat S.,** *Metody statystyczne. Zadania i sprawdziany*, PWE, Warszawa 2002.

**Źródła internetowe**

https://artichowiczdotnet.files.wordpress.com/2016/03/statystykacwiczenia5.pdf
https://stat.gov.pl
https://bdl.stat.gov.pl/BDL/start
http://form.stat.gov.pl/formularze/przewodnik/Portal_Sprawozdawczy_przewodnik_1312.pdf
https://support.office.com/pl-pl/article/stopnie-funkcja

**Akty prawne**

Rozporządzenie Rady Ministrów z dnia 19 grudnia 2017 r. w sprawie programu badań statystycznych statystyki publicznej na rok 2018 (Dz. U. poz. 2471).