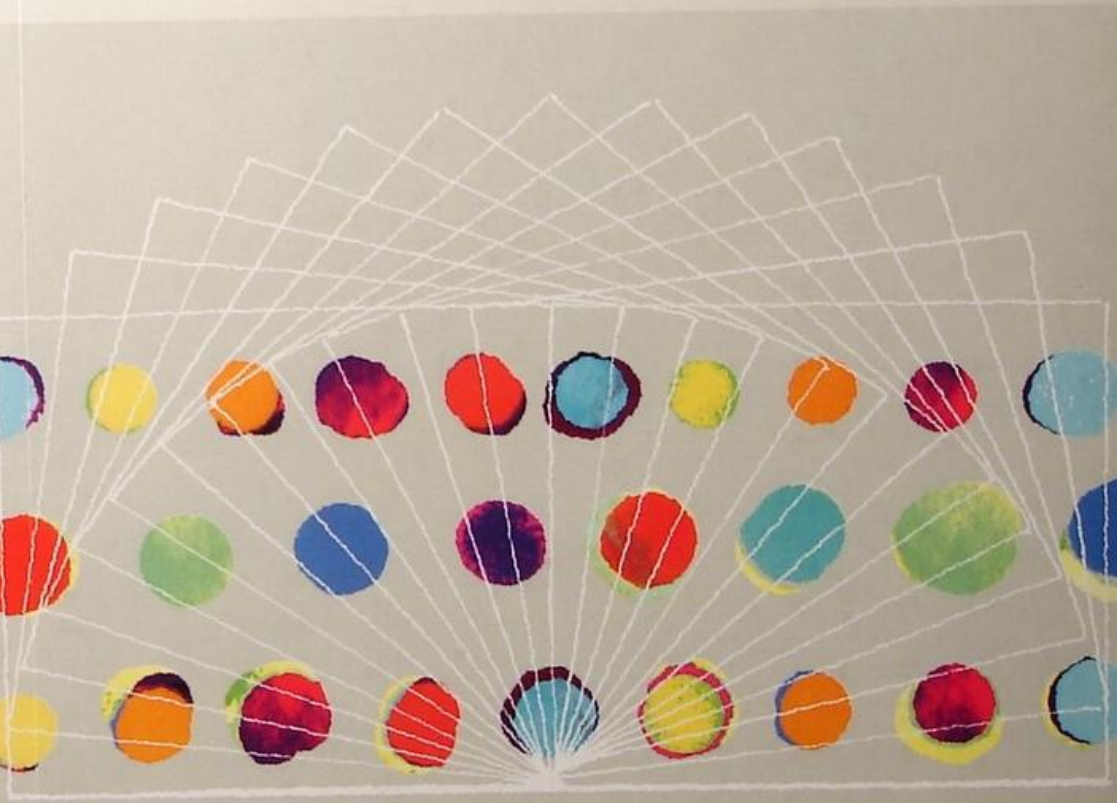
Mieczysław Sobczyk

STATYSTYKA

aspekty praktyczne i teoretyczne



WYDAWNICTWO UMCS

Mieczysław Sobczyk

STATYSTYKA

aspekty praktyczne i teoretyczne

WYDAWNICTWO UNIWERSYTETU MARII CURIE-SKŁODOWSKIEJ
LUBLIN 2006

Recenzent
Prof. dr hab. Jan Paradysz

Redakcja
Marek Jędrych

Projekt okładki i stron tytułowych
Marta i Zdzisław Kwiatkowscy

ISBN 83-227-2423-3

WYDAWNICTWO UNIWERSYTETU MARII CURIE-SKŁODOWSKIEJ
20-031 Lublin, pl. Marii Curie-Skłodowskiej 5
tel. (0-81) 537-53-04
www.press.umcs.lublin.pl
Dział Handlowy: tel./faks (0-81) 537-53-02
e-mail: press@ramzes.umcs.lublin.pl

SPIS TREŚCI

WSTĘP

Na stu ludzi
wiedzących wszystko lepiej
– pięćdziesięciu dwóch;
niepewnych każdego kroku
– prawie cała reszta;
[...]
skulonych, obolałych
i bez latarki w ciemności
– osiemdziesięciu trzech
prędzej czy później;
godnych współczucia
– dziewięćdziesięciu dziewięciu;
śmiertelnych
– stu na stu
Liczba, która jak dotąd nie ulega zmianie.

(Wisława Szymborska,
Przyczynek do statystyki)

W warunkach gospodarki rynkowej obserwuje się wzrost znaczenia metod statystycznych w podejmowaniu różnorodnych decyzji. Proces zarządzania staje się bowiem coraz bardziej złożony i nie może opierać się wyłącznie na doświadczeniu i intuicji, ale powinien bazować na profesjonalnie zbieranych i przetwarzanych informacjach liczbowych. Umiejętność wykorzystywania i wyciągania wniosków z wyników badań statystycznych staje się – w obecnej dobie – cenną zaletą każdego ekonomisty, menedżera, właściciela firmy, gracza giełdowego, urzędnika itp.

Treść niniejszej pracy została ujęta w ośmiu merytorycznych rozdziałach. Pierwszy z nich wprowadza Czytelnika w problematykę badań statystycznych. W szczególności omówiono tu przedmiot i zadania statystyki, podstawowe pojęcia statystyczne, rodzaje i etapy badań statystycznych oraz system informacyjny statystyki publicznej.

Rozdział II poświęcono prezentacji opisowych metod analizy struktury jednowymiarowych rozkładów empirycznych. Obejmują one miary średnie (przeciętne, położenia), zmienności (zróżnicowania, rozproszenia,

dyspersji), asymetrii (skośności) oraz koncentracji. Miary te pozwalają na wszechstronną kompleksową analizę struktury zjawisk masowych.

W rozdziale III zaprezentowano probabilistyczne podstawy wnioskowania statystycznego. Rachunek prawdopodobieństwa stanowi teoretyczną podstawę statystyki matematycznej (indukcyjnej), zajmującej się zasadami i metodami uogólniania wyników otrzymanych z próby losowej na całą populację generalną, z której ona pochodzi. Ten proces uogólniania wyników nosi nazwę wnioskowania statystycznego. W rozdziale tym omówiono losowy dobór próby, zmienne losowe skokowe (dyskretne) i ciągłe oraz przedstawiono podstawowe rozkłady teoretyczne wyróżnionych zmiennych losowych i statystyk z próby.

Rozważania zawarte w rozdziale IV dotyczą estymacji (szacowania) podstawowych parametrów rozkłady jednej zmiennej, tzn. wartości oczekiwanej, wariancji i odchylenia standardowego oraz wskaźnika struktury (odsetka, frakcji, prawdopodobieństwa). W końcowej części tego rozdziału skoncentrowano uwagę na zagadnieniu minimalnej liczebności próby.

Wnioskowanie statystyczne obejmuje – oprócz estymacji – również weryfikację (testowanie) hipotez. Wykorzystanie parametrycznych i nieparametrycznych testów istotności do weryfikacji hipotez statystycznych przedstawiono w rozdziale V.

Rozdział VI dotyczy metod analizy rozkładów wielowymiarowych. Metody te obejmują zagadnienia badania siły związku (korelację) oraz analizę mechanizmu powiązań między zmiennymi (regresję). W analizie korelacji przedstawiono różne miary określone rodzajem zmiennych (ilościowe, jakościowe), sposobem prezentacji danych (szeregi, tablice korelacyjne) oraz kształtem związku (prostoliniowy, krzywoliniowy). W badaniu mechanizmu powiązań między zmiennymi skoncentrowano uwagę na liniowych funkcjach regresji. Prezentowane w tym rozdziale metody analizy korelacji i regresji wykorzystano zarówno w badaniach wyczerpujących, jak też opartych na próbach losowych.

Przedostatni, VII rozdział podręcznika, poświęcono omówieniu metod analizy szeregów czasowych (dynamicznych, chronologicznych). W pierwszej kolejności przedstawiono tu metody dekompozycji szeregów czasowych (wyodrębnianie tendencji rozwojowej, wahań sezonowych i przypadkowych), a następnie zaprezentowano możliwości wykorzystania wnioskowania statystycznego w analizie dynamiki zjawisk masowych.

W ostatnim, VIII rozdziale, skoncentrowano uwagę na prezentacji metod analizy dynamiki zjawisk masowych za pomocą indeksów statystycznych. W szczególności, przedstawiono tu proste metody badania zmian szeregu czasowego (przyrosty absolutne, względne, indeksy indywidualne, średnie tempo zmian) oraz indeksy zespołowe dla wielkości absolutnych (cen, ilości i wartości) i stosunkowych (ilorazowych).

Prezentowane w podręczniku zagadnienia podawane są od podstaw i nie są nazbyt skomplikowane pod względem matematycznym. Cenną zaletą podręcznika jest – jak sądzę – zamieszczenie wielu rozwiązanych przykładów, które ilustrują istotę omawianych procedur statystycznych oraz pokazują możliwości interpretacji otrzymanych wyników. Ponadto praca zawiera bogaty zestaw zadań do samodzielnego rozwiązywania. Kontrolę stopnia opanowania wiedzy ułatwiają zamieszczone na końcu podręcznika odpowiedzi do wszystkich zadań.

W podręczniku zamieszczono obszerną bibliografię, zawierającą publikacje książkowe z zakresy statystyki. Pozwoli ona na pogłębienie znajomości zagadnień prezentowanych w niniejszej pracy, jak również na poszerzenie wiedzy w zakresie tematów pominiętych.

Podręcznik przeznaczony jest dla studentów różnych kierunków, a zwłaszcza akademii ekonomicznych i uniwersyteckich wydziałów ekonomicznych. Sądzę, że ze względu na zawarty w nim zakres przedmiotowy i zrozumiały sposób ujęcia rozpatrywanych zagadnień będzie on przyjazny także dla studentów innych kierunków (socjologii, psychologii, pedagogiki itp.), jak również dla szerokiego grona praktyków prowadzących analizy statystyczne.

Rozdział I

PRZEDMIOT STATYSTYKI I ORGANIZACJA BADAŃ STATYSTYCZNYCH

1.1. Przedmiot i zadania statystyki

Termin statystyka wywodzi się od łacińskiego słowa *status* – państwo. W piśmiennictwie słowo *statystyka* zostało użyte po raz pierwszy w połowie XVIII w. przez Gottfrieda Achenwalla (1719–1772), profesora uniwersytetów w Marburgu i Getyndze. Pod pojęciem **statystyki** G. Achenwall rozumiał zbiór szeroko ujmowanych wiadomości o państwie (warunkach fizjograficznych, ludności, ustroju społecznym, gospodarce itp.). Przedmiot zainteresowań statystyki jako nauki o państwie wyłonił się z potrzeb ówczesnej administracji państwowej, na użytek której zaczęto gromadzić dane liczbowe. W opracowaniach państwoznawczych z tamtego czasu posługiwano się bogatymi danymi liczbowymi, ujmowanymi w postaci zestawień tabelarycznych. Autorów tego rodzaju tablic nazywano **tabelarystami**. Za twórcę kierunku tabelarystycznego uważa się Duńczyka J. P. Anchersena.

Dynamiczny rozwój statystyki jako narzędzia wykorzystywanego do opisu zjawisk gospodarczych (tablice bilansowe, warunki bytu ludności itp.) datuje się od czasów A. Quetleta (1796–1874). W okresie tym statystykę zaczęto traktować jako naukę, czego wyrazem było powołanie w 1834 r. w Anglii Królewskiego Towarzystwa Statystycznego oraz zwołanie I Międzynarodowego Kongresu Statystycznego w Brukseli (1854 r.).

W dobie współczesnej terminu **statystyka** używa się w różnych znaczeniach, a mianowicie:

1) jako nazwy zbioru informacji liczbowych dotyczących określonych zjawisk i procesów (urodzeń, rolnictwa, wypadków drogowych itp.);

2) jako nazwy wszelkich prac związanych z gromadzeniem i opracowywaniem danych liczbowych;

3) jako nazwy pewnych charakterystyk liczbowych opisujących właściwości jednostek tworzących zbiorowości próbne (np. średnia arytmetyczna z próby, odchylenie standardowe z próby itp.);

4) jako nazwy dyscypliny naukowej zajmującej się metodami analizy zbiorów danych liczbowych, odnoszących się do powtarzalnych zjawisk masowych lub wyników eksperymentów.

W niniejszym opracowaniu pojęcie **statystyki** będziemy rozpatrywać w ostatnim z przytoczonych wyżej znaczeń, jakkolwiek przy omawianiu metod wnioskowania statystycznego – terminu statystyka będziemy używać w trzecim z wymienionych określeń.

Celem analizy statystycznej jest wykrycie **prawidłowości** rządzących badanymi zjawiskami masowymi. Prawidłowości te są kształtowane przez **przyczyny główne** (systematyczne) i **uboczne** (przypadkowe). Przyczyny główne działają na każde zjawisko w sposób jednakowy, mają wewnętrzny charakter i są wspólne dla wszystkich jednostek badanej zbiorowości. Przyczyny uboczne działają na każde zjawisko w odmienny sposób i wyrażają osobnicze zróżnicowanie. Skutki działania przyczyn głównych uwidaczniają się tym wyraźniej, im liczniejsza jest poddana badaniu zbiorowość. Mówimy wówczas, że działa **prawo wielkich liczb**.

Statystyczne metody analizy mają charakter uniwersalny i są wykorzystywane niemal we wszystkich dyscyplinach naukowych (psychologii, socjologii, demografii, historii, medycynie, politologii itp.). Wśród dyscyplin związanych ze sferą biznesu, metody statystyczne znajdują zastosowanie w marketingu, ubezpieczeniach, rachunkowości, zarządzaniu itp. Wykorzystanie metod ilościowych, a wśród nich ściśle statystycznych, do analizy złożonych problemów otaczającej nas rzeczywistości jest doniosłym usprawnieniem. Metody te zwielokrotniają bowiem siłę poznawczą danej dyscypliny naukowej, pozwalają na wygodne, ścisłe i jednoznaczne oddawanie myśli za pomocą liczb. Językiem statystyki nieporównanie dokładniej niż słowami (werbalnie) można opisać różnorodne zjawiska masowe. Korzyści wynikające ze statystycznego sposobu myślenia oraz z zastosowania metod statystycznych do badań można określić następująco[1]:

- „statystyka umożliwia dokładniejszy sposób opisu interesującej nas rzeczywistości,
- statystyka zmusza nas do dokładności i śmiałości w działaniu i rozumowaniu,
- statystyka umożliwia formułowanie uogólnień na podstawie uzyskanych wyników analizy,
- statystyka pozwala na przewidywanie rozwoju zjawisk w przyszłości, czyli na budowanie prognoz,
- statystyka dostarcza narzędzi do porządkowania informacji o zjawiskach – a przez to pozwala na budowę ich ogólnego obrazu,

[1] *Statystyka ogólna*, red. M. Woźniak, AE, Kraków 1994, s. 9.

- statystyka dostarcza narzędzi do prowadzenia analizy przyczyn kształtujących badane zjawiska i procesy, a więc umożliwia dokonanie ich klasyfikacji na czynniki systematyczne i przypadkowe".

Za podstawowe zadanie statystyki uznaje się dostarczanie wiarygodnych informacji niezbędnych do podejmowania różnego rodzaju decyzji. Każda decyzja – jeśli ma być skuteczna i efektywna – powinna bazować na bogatym, poprawnie zgromadzonym i wszechstronnie przeanalizowanym zbiorze informacji. W szczególności na wyróżnienie zasługują trzy podstawowe funkcje spełniane przez statystykę: **informacyjna** (dająca pełny i obiektywny obraz badanych zjawisk), **analityczna** (dzięki której możliwe jest określenie czynników kształtujących konkretne procesy i zjawiska) oraz **prognostyczna** (pozwalająca na przewidywanie kierunku rozwoju analizowanych zjawisk).

W zależności od rodzaju posiadanych informacji i preferencji stawianych przez użytkowników, wykorzystanie metod statystycznych może służyć różnym celom. W związku z tym wyróżnia się **statystykę opisową, matematyczną** oraz **statystyki branżowe** (np. statystyka przemysłu, rolnictwa itp.).

Statystyka opisowa zajmuje się problemami programowania badań statystycznych oraz metodami obserwacji statystycznej, sposobami opracowywania i prezentacji materiału statystycznego oraz syntetyzującym (sumarycznym) opisem właściwości zbioru danych.

Przedmiotem zainteresowań **statystyki matematycznej** są metody wnioskowania (estymacji i weryfikacji hipotez) o całej zbiorowości generalnej na podstawie zbadania wybranej w sposób losowy pewnej jej części zwanej próbą.

Statystyki branżowe (zwane też stosowanymi) zajmują się badaniem poszczególnych sfer działalności (np. transportu, handlu, usług, produkcji przemysłowej). Wykorzystuje się tu zarówno statystyczne metody opisu, jak również wnioskowania.

1.2. Podstawowe pojęcia statystyczne

Statystyka, jak każda dyscyplina naukowa, operuje swoistym słownictwem. Do podstawowych pojęć należą: **zbiorowość (populacja) statystyczna, jednostka statystyczna** oraz **cecha statystyczna**.

Mianem **zbiorowości** (populacji) **statystycznej** określa się zbiory dowolnych elementów (osób, przedmiotów, faktów) podobnych do siebie pod względem określonych właściwości. Kompletny zbiór elementów nosi nazwę **zbiorowości** (populacji) **generalnej**. Indywidualne składowe zbiorowości nazywamy **jednostkami statystycznymi** lub **jednostkami badania**. Zbiorowości powinny być ściśle określone pod względem **rzeczowym** (kto lub co jest przedmiotem badania), **przestrzennym** (gdzie znajduje się zbio-

rowość) oraz **czasowym** (jakiego momentu lub okresu dotyczy badanie). Podkreślić należy, że w tej samej zbiorowości można – w zależności od celu badania – wyróżnić różne jednostki. Na przykład w zbiorowości osób jednostkami statystycznymi mogą być: poszczególne osoby, rodziny lub gospodarstwa domowe.

Zbiorowości można wyodrębniać według różnych kryteriów. I tak, ze względu na kryterium czasu wyróżnia się zbiorowości **statyczne** (np. zbiorowość ludności Polski wg stanu na dzień 31.12.2004 r., na 30.6.2005 r. itd.) i **dynamiczne** (np. samochody zarejestrowane w Polsce w 2004 r.). **Zbiorowości statyczne** tworzą więc jednostki, które istniały, istnieją lub będą istniały w ściśle określonym momencie. **Zbiorowości dynamiczne** składają się z jednostek obserwowanych w pewnym przedziale czasu.

Przyjmując za kryterium wyodrębnienia liczebności, wyróżniamy zbiorowości **skończenie liczne** i **nieskończenie liczne**. Pierwsze z nich składają się z określonej przeliczalnej liczby jednostek (np. studenci UMCS w roku akademickim 2004/2005), drugie zaś – tworzą elementy o nieprzeliczalnej liczbie (np. zbiorowość organizmów żywych na kuli ziemskiej).

Jednostki statystyczne będące elementami składowymi zbiorowości charakteryzują się określonymi właściwościami, które nazywamy **cechami statystycznymi**. Najogólniej rzecz ujmując, cechy statystyczne można podzielić na **stałe** i **zmienne**. **Cechy stałe** (rzeczowe, przestrzenne i czasowe) są wspólne dla wszystkich jednostek zbiorowości. Taką zbiorowość nazywamy wówczas **jednorodną**. Cechy stałe nie podlegają badaniu statystycznemu, a jedynie umożliwiają zaliczanie jednostek do określonych zbiorowości. Właściwości, które różnicują jednostki statystyczne między sobą, nazywamy **cechami zmiennymi**. Przedmiotem badania statystycznego są zbiorowości składające się z jednostek posiadających jedną lub kilka cech wspólnych (stałych) oraz jedną lub wiele cech (zmiennych) je różnicujących. Cechy zmienne dzielimy na **ilościowe** (mierzalne) oraz **jakościowe** (niemierzalne).

Każda z cech zmiennych może występować w różnych odmianach zwanych **wariantami cechy**. Warianty mogą być określane słownie (np. płeć, zawód, wykształcenie) lub też przyjmować różne wartości liczbowe będące wynikiem zliczania (liczba osób w rodzinie, liczba uczestników kursu itp.) bądź też rezultatem dokonanego pomiaru (np. wzrost w cm, dochód w zł/osobę). Cechy, których warianty podawane są w sposób opisowy, nazywamy **cechami niemierzalnymi** (jakościowymi). Cechy, których warianty są wyrażane za pomocą liczb, określa się mianem **cech mierzalnych** (ilościowych).

Cechy jakościowe, których natężenie określone jest przez przymiotniki (np. ocena wiadomości ucznia: bardzo dobra, dobra, dostateczna, niedostateczna), nazywane są **cechami porządkowymi**. Następującym po sobie wariantom takich cech można przyporządkować kolejne liczby (rangi), odpowiadające zmieniającemu się natężeniu cechy. Cechy jakościowe, których

warianty są określone w sposób opisowy i nie jest możliwe ustalenie dla nich hierarchii ważności (np. zawody: ekonomiczne, techniczne, humanistyczne), nazywane są **cechami nominalnymi**.

Wśród cech ilościowych wyróżnia się **cechy skokowe** i **ciągłe**. Pierwsze z nich przyjmują warianty zmieniające się skokowo (bez wartości pośrednich), w postaci liczb całkowitych (np. ilość wyprodukowanych braków w ciągu zmiany roboczej). Cechy ilościowe ciągłe mogą przyjmować każdą wartość z różną dokładnością (np. wiek, wzrost, waga). Podkreślić należy, że dokonując podziału cech na skokowe i ciągłe nie bierzemy pod uwagę faktycznego sposobu pomiaru cechy, lecz jedynie teoretyczną możliwość. Dlatego też wiek, mimo że jest zazwyczaj podawany w pełnych latach, jest przykładem cechy ciągłej, gdyż między dwa kolejne warianty wieku, np. 30 i 31 lat, można teoretycznie wstawić wiele wariantów różniących się o ułamki roku (np. miesiące, dni).

Podział cech na mierzalne i niemierzalne wiąże się ze **skalą pomiarową**. Wyróżnia się cztery skale pomiarowe: **nominalną**, **porządkową** (rangową), **przedziałową** (interwałową) oraz **stosunkową** (ilorazową). Cechy niemierzalne mierzone są w **skalach słabych** (tj. nominalnej i porządkowej), cechy ilościowe – w skalach **mocnych** (interwałowej i ilorazowej).

Skala nominalna stanowi najmniej precyzyjny sposób pomiaru. Liczby pełnią tu jedynie rolę umownych symboli służących do identyfikacji jednostek statystycznych i ich zaklasyfikowania do wyróżnionych kategorii. Rezultatem pomiaru w tej skali jest więc jedynie stwierdzenie, że n_1 jednostek należy do kategorii pierwszej, n_2 jednostek należy do kategorii drugiej, ..., n_k jednostek należy do k-tej kategorii. Na cechach mierzonych w skali nominalnej można wykonywać tylko niektóre operacje arytmetyczne, takie jak: zliczanie jednostek w danych kategoriach, ustalenie klasy najliczniejszej, określenie częstości występowania. Przykładami liczb typu nominalnego są: numery na koszulkach zawodników drużyny piłki nożnej, trasy autobusów, numery pokoi hotelowych, grupy krwi itp.

Skala porządkowa (rangowa) posiada wszystkie cechy skali nominalnej, a dodatkowo pozwala na porządkowanie jednostek w ramach wyróżnionych kategorii, gdyż mierzone jest tu natężenie badanej cechy (np. wykształcenie: podstawowe, średnie, licencjackie, wyższe). Każdej kategorii można przypisać liczbę (**rangę**) zgodnie z natężeniem cechy, czyli ustalić kolejność jednostek statystycznych. W skali porządkowej określone są np. gatunki owoców i warzyw (I, II, III), stopnie wojskowe, preferencje konsumentów. Na cechach mierzonych w skali porządkowej można wykonywać te operacje, które są właściwe dla skali nominalnej oraz dodatkowo – obliczać charakterystyki porządkowe i miary korelacji.

Skala przedziałowa (interwałowa) występuje wtedy, gdy uporządkowany zbiór wartości badanej cechy składa się z liczb rzeczywistych. W skali tej brak jest jednak zera absolutnego, gdyż punkt zerowy jest tu ustalany

umownie. Przykładowo, w skali temperatury Celsjusza zero ustalone jest jako punkt, w którym zamarza woda. Niższe temperatury zapisywane są ze znakiem minus, a wyższe – ze znakiem plus. W skali przedziałowej wyrażane są np. lata kalendarzowe, indeksy cen, skale temperatur Celsjusza i Fahrenheita. Wszystkie operacje na skalach słabych (tj. nominalnej i porządkowej) są dopuszczalne na skali przedziałowej, a ponadto można tu obliczać średnie i miary zróżnicowania jednostek. Nie można natomiast wykonywać dzielenia (z uwagi na brak absolutnego punktu zerowego).

Skala stosunkowa (ilorazowa) ma własności trzech poprzednich skal, a jej cechą charakterystyczną jest posiadanie naturalnego punktu zerowego. Punkt ten oznacza brak danej cechy (fizyczny poziom zerowy). Umożliwia to dokonywanie na liczbach w tej skali wszystkich operacji matematycznych (również dzielenia). W skali stosunkowej mierzy się wiek, masę, odległość, dochody, wielkość sprzedaży, ceny towarów, koszty itp.

Identyfikacja skali pomiaru zmiennej ma istotne znaczenie; z nią związane są bowiem dopuszczalne działania matematyczne. Im silniejsza jest skala pomiaru, tym większy zasób metod statystycznych może być wykorzystany.

1.3. Rodzaje badań statystycznych

Pod pojęciem **badania statystycznego** rozumie się zespół czynności zmierzających do uzyskania – za pomocą metod statystycznych – informacji charakteryzujących badaną zbiorowość. Badania statystyczne mogą być klasyfikowane według różnych kryteriów. I tak, z punktu widzenia liczby jednostek objętych badaniem, wyróżnia się **badania pełne** (kompletne, całkowite, wyczerpujące) i **badania częściowe** (niekompletne, niepełne). **Badania pełne** to takie, w których obserwacji podlegają wszystkie jednostki zbiorowości (populacji) generalnej. Do analizy wykorzystuje się tu metody statystyki opisowej, a jej wyniki stanowią charakterystykę badanej zbiorowości.

W **badaniach częściowych** obserwowane są tylko niektóre wybrane lub wylosowane ze zbiorowości generalnej jednostki. Wylosowane ze zbiorowości generalnej jednostki tworzą próbę losową. Na podstawie prób losowych wnioskujemy o zbiorowościach generalnych na określonym poziomie prawdopodobieństwa. Do tego celu wykorzystuje się metody statystyki matematycznej. Obecnie coraz częściej stosuje się metody badań częściowych (szczególnie z wykorzystaniem metody reprezentacyjnej). Badania te obniżają koszty, skracają czas badania, angażują mniejszy zespół ludzi, a jeśli są przeprowadzane profesjonalnie – dają wyniki zbliżone do tych, jakie uzyskano by w badaniach pełnych.

W badaniach częściowych dobór jednostek do próby może być losowy lub celowy. Stąd też wśród badań częściowych wyróżnia się badania **reprezentacyjne**, **monograficzne** oraz **ankietowe**. **Badanie reprezentacyjne** polega na tym, że z całej zbiorowości generalnej pobiera się próbę losową, przeprowadza się dokładne badanie jednostek wchodzących w jej skład, a następnie przenosi się wyniki badania próby na całą populację. Uogólnianie wyników z badania próby na całą zbiorowość generalną odbywa się przy wykorzystaniu rachunku prawdopodobieństwa i nosi nazwę **wnioskowania statystycznego**. Przykładem badania reprezentacyjnego jest prowadzone od 1992 r. przez Główny Urząd Statystyczny badanie aktywności ekonomicznej ludności (BAEL). Jego celem jest uzyskanie informacji o wielkości i strukturze zasobów siły roboczej w Polsce oraz śledzenie zmian w poziomie aktywności zawodowej ludności. Próba obejmuje tu ok. 25 tys. wylosowanych gospodarstw domowych, a obserwacji statystycznej podlegają członkowie tych gospodarstw w wieku 15 lat i więcej (ok. 60 tys. osób). W badaniu tym obserwuje się bieżącą aktywność zawodową wylosowanych osób w połączeniu z cechami demograficznymi (wiek, płeć, stan cywilny) i społeczno-zawodowymi (zawód, wykształcenie, miejsce zamieszkania). Wyniki badania są uogólniane na cały kraj i na makroregiony.

W **celowym** (subiektywnym) doborze jednostek do próby dużą rolę odgrywa osoba prowadząca badanie. Od jej wiedzy i prezentowanych poglądów zależy to, które z jednostek zostaną zakwalifikowane do próby. Przykładem badań częściowych z celowym doborem jednostek są **badania monograficzne** i **ankietowe**. Metoda **monograficzna** polega na wszechstronnym (ilościowym i jakościowym) opisie celowo wybranej do badania jednostki. Wybrana świadomie jednostka powinna być typową lub wyróżniającą się (wzorzec). **Badanie ankietowe** polega na badaniu opinii dotyczących określonego problemu lub zjawiska według ustalonego zestawu pytań, skierowanego do ściśle zdefiniowanego grona osób bądź bliżej nieznanych respondentów. Istotnym czynnikiem rzutującym na wyniki badań ankietowych jest właściwe opracowanie kwestionariusza zawierającego pytania. Powinny być one tak skonstruowane, aby stwarzały ankietowanemu możliwość formułowania własnych opinii, wniosków i spostrzeżeń.

W przypadku braku możliwości przeprowadzenia zarówno badania pełnego, jak i częściowego stosuje się postępowanie zwane **szacunkiem statystycznym**. Szczególnymi metodami szacunku statystycznego są rachunki **interpolacyjne** i **ekstrapolacyjne**. Interpolacja polega na szacowaniu nieznanych wartości badanej cechy na podstawie znajomości jej wcześniejszych i późniejszych realizacji. Ekstrapolacja pozwala na ustalenie nieznanych wartości znajdujących się poza przedziałem wartości znanych.

1.4. Etapy badania statystycznego

W badaniach statystycznych – niezależnie od ich rodzaju – można wyróżnić cztery etapy:

- przygotowanie (programowanie),

- obserwację statystyczną,
- opracowanie i prezentację materiału statystycznego,
- opis (lub wnioskowanie) statystyczny.

Programowanie oznacza szczegółowo rozpisaną koncepcję zamierzonego badania. Etap ten obejmuje czynności przygotowawcze, takie jak: ustalenie celu, przedmiotu i zakresu badania. Z punktu widzenia poprawności wyników badania istotne jest również właściwe określenie jednostki sprawozdawczej, czyli osób fizycznych lub prawnych (przedsiębiorstw, instytucji, organizacji itp.), dysponujących niezbędnymi źródłami informacji. Ważny jest również wybór metody badania, jak również ustalenie harmonogramu prac oraz kosztorysu.

Obserwacja statystyczna polega na ustaleniu wartości cech ilościowych lub odmian cech jakościowych u jednostek tworzących badaną zbiorowość. Może to odbywać się poprzez bezpośredni pomiar lub zbieranie informacji. W przypadku badań pełnych można wykorzystać wyniki **spisów powszechnych**, **rejestracji bieżącej** czy **sprawozdawczości statystycznej**. W przypadku badań częściowych korzysta się z rezultatów uzyskanych metodami: reprezentacyjną, monograficzną czy ankietową (por. punkt 1.3).

Spis powszechny jest jednorazowym lub cyklicznie powtarzanym badaniem statystycznym, którego celem jest ustalenie wielkości i struktury analizowanego zjawiska w określonym momencie czasu (np. powszechne spisy ludności, spisy gospodarstw rolnych).

Rejestracja bieżąca polega na systematycznym ewidencjonowaniu ściśle określonych faktów (np. urodzeń, zgonów, małżeństw, bezrobotnych), które są przedmiotem badania. W wyniku rejestracji bieżącej powstają **rejestry administracyjne** (REGON, PESEL, rejestry ZUS itp.). Stanowią one cenne źródło informacji wykorzystywanych w badaniach statystycznych.

Sprawozdawczość statystyczna – to wszelkiego rodzaju sprawozdania sporządzone na jednolitych formularzach i z określoną częstotliwością przez ustawowo zobowiązane jednostki gospodarki narodowej. Dzięki stosowaniu jednolitych formularzy zebrane dane można agregować w dowolnych przekrojach.

Uzyskany w wyniku obserwacji zbiór danych tworzy **materiał statystyczny**. Jest on zazwyczaj obciążony **błędami**. Ze względu na przyczynę powodującą błędy, dzielimy je na **systematyczne** (tendencyjne) oraz **przypadkowe** (niesystematyczne). Błędy systematyczne wynikają ze świadomej tendencji do zniekształcania badanej rzeczywistości. Błędy przypadkowe popełniane są nieumyślnie i wynikają z nieuwagi, z nieumiejętności podania prawidłowych informacji bądź ze zwykłego niedbalstwa. W celu zapewnienia wiarygodności badań, wszelkiego rodzaju błędy powinny zostać wyeliminowane.

W celu sensownego wykorzystania zebranego materiału statystycznego należy poddać go odpowiedniemu opracowaniu i prezentacji. Opracowania materiału statystycznego dokonuje się wykorzystując **metodę grupowania**. Grupowanie polega na wyodrębnianiu jednorodnych lub względnie jednorodnych części w ramach badanej zbiorowości. W niektórych przypadkach grupowanie ma charakter naturalny (np. pracowników wg płci można podzielić na dwie rozłączne grupy: kobiety i mężczyźni). Z reguły jednak przeprowadzający badanie sam decyduje o tym, na ile i na jakie grupy podzielić daną zbiorowość. Decyduje tu głównie cel badania. Przykładowo, ludność według stanu cywilnego można podzielić na dwie grupy: wolnych (wolne) oraz żonatych (zamężne) lub na cztery grupy: panny (kawalerów), zamężne (żonatych), wdowy (wdowców), rozwiedzione (rozwiedzeni). Po wydzieleniu grup następuje zliczanie jednostek wchodzących do wydzielonych części.

Grupowanie statystyczne może być związane z cechą mierzalną lub niemierzalną. Grupowanie według cechy niemierzalnej nosi nazwę **typologicznego**, a według cechy mierzalnej – **wariancyjnego**. Grupowaniem typologicznym jest np. podział pracowników według cechy „wykształcenie" na trzy grupy: posiadających wykształcenie wyższe, średnie oraz zawodowe. W grupowaniu wariancyjnym należy uprzednio określić liczbę klas dla wyróżnionej cechy mierzalnej. Bierze się tu pod uwagę cel badania, charakter zabranego materiału statystycznego oraz obszar zmienności (tj. różnicę między maksymalną i minimalną wartością zmiennej) interesującej nas cechy.

Wyróżnia się dwa główne źródła danych statystycznych: **pierwotne** (źródłowe) i **wtórne**. Pierwsze z nich uzyskuje się poprzez obserwację bezpośrednią (np. ankiety bezpośredniej lub w wyniku przeprowadzonego wywiadu telefonicznego). Dane wtórne powstają w rezultacie uporządkowania i przetworzenia danych pierwotnych (np. dane zawarte w zasobach internetowych, w rocznikach statystycznych, w publikacjach prasowych).

Pochodzący z obserwacji statystycznej materiał liczbowy jest z reguły dość obszerny. Utrudnia to dokonanie analizy statystycznej. Materiał ten należy odpowiednio zaprezentować. Celowi temu służą **szeregi statystyczne, tablice statystyczne** oraz **wykresy statystyczne**.

Szeregiem statystycznym nazywamy ciąg wartości liczbowych badanej cechy, uporządkowanych według określonych kryteriów (np. rosnąco lub malejąco). Wyróżnia się następujące rodzaje szeregów:

- proste (wyliczające, szczegółowe),
- rozdzielcze z cechą jakościową,
- rozdzielcze z cechą ilościową (punktowe i przedziałowe),
- kumulacyjne szeregi rozdzielcze,
- geograficzne (przestrzenne),
- czasowe (dynamiczne, chronologiczne) momentów i okresów.

Szereg prosty przedstawia materiał statystyczny uporządkowany wyłącznie według wartości badanej cechy. Porządkowanie polega tu na wypisy-

waniu wartości liczbowych w kolejności rosnącej lub malejącej. Szeregi proste zawierają zwykle wartości powtarzające się.

Szeregiem rozdzielczym (strukturalnym) nazywamy ciąg wartości liczbowych uporządkowanych według wariantów badanej cechy mierzalnej lub niemierzalnej. Poszczególnym wariantom zmiennej przyporządkowane są odpowiadające im liczebności. Szereg rozdzielczy cechy niemierzalnej jest zestawieniem poszczególnych wariantów danej cechy i odpowiadających im liczebności (por. tab. 1.1).

Tab. 1.1. Pracownicy przedsiębiorstwa „X" według poziomu wykształcenia w lipcu 2005 r.

Poziom wykształcenia	Liczba pracowników
Podstawowe	13
Średnie	62
Wyższe	25
Ogółem	100

Źródło: dane umowne.

Tworząc szeregi rozdzielcze w oparciu o cechę mierzalną, jej warianty można określić **punktowo** lub **przedziałowo**. Szeregi rozdzielcze **punktowe** buduje się dla zmiennej skokowej. Przykładem tego rodzaju szeregu jest poniższe zestawienie:

Liczba stomatologów	0	1	2	3	4
Liczba gmin	12	18	31	19	16

Szeregi rozdzielcze **przedziałowe** dotyczą cechy ciągłej. W tym przypadku należy najpierw ustalić liczbę **przedziałów klasowych** (klas), ich **rozpiętości** oraz sposób oznaczania **granic przedziałów**. Przy ustalaniu liczby klas zaleca się wykorzystywać następujące wzory:

$$k \leq 5 \log N, \tag{1.1}$$

$$k \leq \sqrt{\frac{N}{2}}, \tag{1.2}$$

$$k = 1 + 3{,}222 \log N, \tag{1.3}$$

gdzie: k jest liczbą przedziałów klasowych, N – liczebnością badanej zbiorowości.

Rozpiętość przedziału (zwana też **interwałem**) – to różnica między górną a dolną wartością przedziału klasowego. Rozpiętości przedziałów mogą być równe lub różne. W analizie strukturalnej większą przydatność mają szeregi o równych rozpiętościach. Przy ustalaniu rozpiętości wykorzystuje się relacje:

$$c = \frac{R}{k}, \tag{1.4}$$

lub

$$c = \frac{R}{1 + 3{,}222 \log N}, \tag{1.5}$$

gdzie: c jest rozpiętością przedziału klasowego, R – obszarem zmienności cechy (tj. różnicą między największą i najmniejszą wartością cechy).

Podanych wyżej sposobów ustalania liczby przedziałów klasowych i ich interwałów nie należy traktować jako bezwzględnie obowiązujących. Ogólnie można stwierdzić, że im większa jest liczebność zbiorowości (N) i dłuższy obszar zmienności – tym większa powinna być liczba przedziałów klasowych. W praktyce statystycznej niekiedy celowo tworzy się szeregi rozdzielcze o nierównych interwałach. Przykładowo, w szeregu przedstawiającym strukturę indywidualnych gospodarstw rolnych według powierzchni w *ha* zazwyczaj wyodrębnia się klasy: 1–2 *ha*, 2–5 *ha*, 5–7 *ha*, 7–10 *ha*, 10–15 *ha* oraz powyżej 15 *ha*. W tak skonstruowanym szeregu dla małych gospodarstw tworzy się małe przedziały, a dla dużych – większe. W ten sposób otrzymuje się przejrzyste przedstawienie liczebności gospodarstw. W każdym bądź razie, rozkład liczby obserwacji (liczebności) powinien ukazywać prawidłowości charakteryzujące strukturę zbiorowości według badanej cechy. Szereg powinien mieć jeden wyraźny punkt skupienia obserwacji, nie może mieć przedziałów pustych. Ponadto, w jednym przedziale klasowym nie może koncentrować się nazbyt duża część zbiorowości. Zarówno nadmierne rozdrobnienie jednostek, jak też zbytnie ich skupianie się utrudnia analizę struktury zbiorowości.

Granice przedziałów klasowych mogą być oznaczane dwojako: dolna granica następnego przedziału równa jest górnej granicy przedziału poprzedniego (np. 2–4; 4–6; 6–8 itd.) lub też granice te różnią się (np. 2–4; 5–7; 8–10 itd.). W sensie formalnym zapis granic przedziałów nie ma większego znaczenia. Chodzi jednak o to, by w trakcie budowy szeregu rozdzielczego przedziałowego, jednostki o tych samych wartościach cechy zaliczać zawsze do tego samego przedziału.

Szeregi rozdzielcze przedziałowe mogą mieć zamknięte wszystkie przedziały lub też otwarty pierwszy, ostatni lub pierwszy i ostatni przedział. Otwarte przedziały stosuje się wówczas, gdy w badanej zbiorowości występują wartości skrajne, znacznie różniące się od pozostałych.

Kumulacyjne szeregi rozdzielcze otrzymujemy w drodze łączenia kolejnych przedziałów klasowych i dodawania odpowiadających im liczebności. Liczebność skumulowana dla ostatniego przedziału klasowego jest wówczas równa ogólnej liczebności badanej zbiorowości (por. tab. 1.2).

Kumulacyjne szeregi rozdzielcze informują o tym, ile jest w badanej zbiorowości statystycznej jednostek przyjmujących co najwyżej określoną wartość cechy.

Szeregi geograficzne (zwane też przestrzennymi lub terytorialnymi) przedstawiają rozmieszczenie wielkości statystycznych według jednostek administracyjnych kraju (gmin, powiatów, województw), części świata, regionów gospodarczych itp.

Tab. 1.2. Gminy województwa L według powierzchni (w km^2)

Powierzchnia w km^2	Liczba gmin	Szereg rozdzielczy skumulowany	
		powierzchnia w km^2	skumulowane liczebności
50–85	2	poniżej 85	2
85–120	4	poniżej 120	6
120–155	8	poniżej 155	14
155–190	12	poniżej 190	26
190–225	16	poniżej 225	42
225–260	7	poniżej 260	49
260–295	3	poniżej 295	52
Razem	52	X	X

Źródło: dane umowne.

Szeregi czasowe (dynamiczne, chronologiczne) prezentują rozwój zjawisk w czasie, przy czym może być tu uwzględniony ściśle określony moment, np. 31 grudnia każdego roku (szeregi czasowe **momentów**) lub pewien przedział, jak lata czy miesiące (szeregi czasowe **okresów**).

Szeregi statystyczne powstają w wyniku grupowania jednostek według jednej cechy. Jeśli grupowanie przeprowadzone jest według dwóch lub większej liczby cech – to wyniki przedstawione są w formie tablicy. **Tablice statystyczne** wykorzystywane są do prezentacji danych uporządkowanych według określonego kryterium. Aby tablice statystyczne były użyteczne, powinny spełniać określone warunki dotyczące formalnej (zewnętrznej) budowy oraz merytorycznej (wewnętrznej) spoistości. Z formalnego punktu widzenia każda tablica statystyczna powinna zawierać tytuł, słowne oznaczenie kolumn i wierszy, źródło danych statystycznych oraz – w razie potrzeby – dodatkowe objaśnienia. W tytule tablicy podaje się w sposób jasny i zwięzły treść tablicy, określającą zbiorowość statystyczną pod względem rzeczowym, czasowym i przestrzennym. Obowiązuje zasada bezwzględnego wypełniania wszystkich kolumn i wierszy w tablicy. Jeżeli z różnych przyczyn nie wszystkie pola tablicy mogą być wypełnione liczbami, należy stosować odpowiednie **znaki umowne**. W polskiej praktyce statystycznej stosuje się następujące znaki umowne:

kreska (—) – zjawisko nie występuje;

zero (0) – zjawisko występuje, jednakże w ilościach mniejszych od liczb, które mogły być wyrażone uwzględnionymi w tablicy znakami cyfrowymi;

kropka (•) – zupełny brak informacji lub brak informacji wiarygodnych;

krzyżyk (×) – wypełnienie pozycji, ze względu na układ tablicy, jest niemożliwe lub niecelowe;

„w tym" – oznacza, że nie podaje się wszystkich składników sumy.

Do opisu zbiorowości statystycznych wykorzystuje się różne rodzaje tablic. Ze względów konstrukcyjnych i poznawczych tablice statystyczne można podzielić na **proste** i **złożone** (kombinowane, wielodzielne). **Tablice proste** charakteryzują badaną zbiorowość ze względu na jedną cechę, przy czym może to być zarówno cecha ilościowa, jak i jakościowa. Wynika stąd, że tablica prosta może być utożsamiana z szeregiem statystycznym. **Tablice złożone** opisują jedną zbiorowość statystyczną według kilku cech (co najmniej dwóch), dwie lub kilka zbiorowości według dwóch lub kilku cech. Tego rodzaju tablice prezentują więc zespół szeregów statystycznych, a stopień ich złożoności zależny jest od liczby badanych cech lub liczby zbiorowości (badanych zjawisk).

W celu ułatwienia interpretacji i analizy zebranego materiału statystycznego dość często wykorzystuje się **prezentację graficzną (wykresy)**. **Wykresy statystyczne** są narzędziem analizy i formą rejestracji informacji statystycznych. Każdy wykres składa się z części tekstowej oraz pola wykresu. W części tekstowej zamieszcza się tytuł wykresu oraz opis źródła, które stanowiło podstawę do jego sporządzenia. Pole wykresu jest natomiast graficznym obrazem danej cechy statystycznej, przedstawionym za pomocą odpowiednich znaków graficznych (punktów, linii, figur itp.). Ponadto, do każdego wykresu należy podać legendę, czyli wyjaśnienie dotyczące zastosowanych barw, symboli, znaków itd.

W graficznej prezentacji danych statystycznych stosuje się wiele różnorodnych **wykresów**[2]. Ogólnie można je podzielić na: liniowe proste (przedstawiające tylko jedno zjawisko) i złożone słupkowe, powierzchniowe (kołowe, prostokątne, kwadratowe itp.), przestrzenne (kuliste, stożkowe itp.), punktowe, obrazkowe, mapowe (kartogramy). W porównaniu z tabelaryczną prezentacją danych, wykresy statystyczne są mniej precyzyjne i szczegółowe, ale za to bardziej sugestywne. Stąd też są one często używane do celów popularyzatorskich.

Do celów analitycznych wykresy statystyczne sporządza się w **prostokątnym układzie współrzędnych**. W przypadku cechy ciągłej do tego rodzaju wykresów zalicza się m.in. **histogram, diagram** (wielobok liczebności) i **krzywą liczebności** (ogiwę). **Histogram** jest wykresem złożonym z prostokątów, których podstawy (równe interwałom klasowym) spoczywają na osi odciętych, natomiast wysokości są określone na osi rzędnych przez liczebności (absolutne lub względne).

Pola prostokątów tworzących histogram są proporcjonalne do liczebności poszczególnych klas, a łączna powierzchnia wszystkich prostokątów reprezentuje ogólną liczebność badanej zbiorowości.

---

[2] Por. np.: A. Maksimowicz-Ajchel, *Funkcjonowanie przedsiębiorstwa. Wybrane zagadnienie statystyki*, WSiP, Warszawa 2004.

Szczególną postacią histogramu jest **histogram kumulacyjny**. W tym przypadku na osi rzędnych odkłada się skumulowane liczebności absolutne lub względne.

Diagram można otrzymać z histogramu przez połączenie odcinkami kolejnych środków przedziałów klasowych, reprezentujących punkty środkowe górnych boków poszczególnych prostokątów. Otrzymana w ten sposób linia łamana nosi nazwę **diagramu (wieloboku liczebności)**. Łącząc odcinkami kolejne górne granice przedziałów klasowych, otrzymujemy **diagram kumulacyjny**, zwany też **kumulacyjnym wielobokiem liczebności**. Tak więc diagram kumulacyjny powstaje przez połączenie punktów, których współrzędnymi są górne granice przedziałów klasowych i odpowiadające im liczebności skumulowane.

W przypadku, kiedy interwały klasowe szeregu rozdzielczego przedziałowego są różne, zachodzi konieczność modyfikacji sposobu budowy histogramu. Aby otrzymać histogram mający pola prostokątów proporcjonalne do liczebności poszczególnych klas, na osi rzędnych należy odkładać **wskaźniki natężenia liczebności**. Natężenie liczebności w danej klasie ustala się następująco:

$$\frac{\text{liczebność w danej klasie} \times \text{interwał przedziału najmniejszego}}{\text{interwał danej klasy}}$$

Obliczone w powyższy sposób natężenie liczebności informuje o częstotliwości występowania jednostek zbiorowości na jednakową rozpiętość (interwał) przedziału klasowego w danym szeregu statystycznym.

W przypadku cechy ciągłej teoretycznie można nieograniczenie zmniejszać rozpiętość przedziałów klasowych, zwiększając tym samym ich liczbę. W rezultacie takiego postępowania środki przedziałów klasowych będą tworzyć gęstą siatkę punktów. Łącząc te punkty otrzymujemy linię ciągłą, która nosi nazwę **krzywej liczebności, krzywej ogiwalnej** lub **ogiwy**.

Czwartym etapem badania statystycznego jest **opis** lub **wnioskowanie statystyczne**. Różnica między opisem a wnioskowaniem statystycznym sprowadza się do tego, że **opis** dotyczy tylko danej zbiorowości generalnej lub próby (niekoniecznie losowej), podczas gdy **wnioskowanie** ma miejsce wówczas, kiedy badanie jest reprezentacyjne (próba losowa), a jego wyniki są uogólniane na całą populację generalną, z której pochodzi próba. Możliwości uogólniania wyników z próby losowej na całą populację generalną daje rachunek prawdopodobieństwa, który stanowi teoretyczną podstawę wnioskowania statystycznego. Metody wnioskowania statystycznego wchodzą w zakres **statystyki matematycznej**. Rezultaty uzyskane na podstawie próby losowej jedynie w przybliżeniu odwzorowują relacje zachodzące w populacji generalnej. Przy przenoszeniu rezultatów z próby na całą populację generalną, z której pochodzi próba, istnieje niebezpieczeństwo popełnienia błędów. Rozmiary tych błędów oceniane są jako prawdopodobieństwa.

Opis statystyczny odnosi się nie do poszczególnych jednostek, ale do całej zbiorowości generalnej. Ma on zatem charakter **sumaryczny**. Przy opisie statystycznym posługujemy się określonymi miarami, zwanymi **parametrami**. Metody opisu statystycznego wchodzą w zakres **statystyki opisowej** (analiza struktury, współzależności i dynamiki zjawisk masowych).

1.5. System informacyjny statystyki publicznej

Zasady gromadzenia danych i prowadzenia badań statystycznych reguluje Ustawa o statystyce publicznej z 29 czerwca 1995 r. (Dz. U., Nr 88, poz. 439). Całokształt działań związanych ze zbieraniem informacji statystycznych, ich gromadzeniem, przechowywaniem, opracowywaniem, ogłaszaniem, udostępnianiem i rozpowszechnianiem wyników badań statystycznych jako oficjalnych danych tworzy **System Informacyjny Statystyki Publicznej** (SISP).

Zadania wynikające z funkcjonowania SISP realizowane są przez **służby statystyki publicznej**, tzn. Główny Urząd Statystyczny (GUS) oraz podległe mu urzędy statystyczne. GUS, utworzony w 1918 r., jest centralnym organem administracji państwowej, którego zadaniem są m.in.: prowadzenie i udostępnianie wyników badań statystycznych, międzynarodowa współpraca w zakresie statystyki, rozwijanie metodologii badań statystycznych oraz ustalanie i aktualizacja podstawowych definicji, kodów, klasyfikacji i nomenklatur (tj. zbioru nazw i terminów używanych w statystyce). Przykładowo, pod nadzorem GUS zostały opracowane, wymienione w ustawie o statystyce państwowej, **standardy klasyfikacyjne**: Europejska Klasyfikacja Działalności (EKD); Systematyczny Wykaz Wyrobów (SWW); Klasyfikacja Środków Trwałych (KŚT); Klasyfikacja Obiektów Budowlanych (KOB) czy Klasyfikacja Zawodów (KZ).

Szczególnie ważnym standardem klasyfikacyjnym jest **EKD**, wprowadzona w Polsce w 1992 r. Przed 1992 r. obowiązywała w Polsce, oparta na stosowanych w krajach RWPG, Klasyfikacja Gospodarki Narodowej (**KGN**). EKD klasyfikuje jednostki gospodarki narodowej na podstawie ich przeważającego przedmiotu działalności, co umożliwia analizę porównawczą rozwoju różnych krajów (np. analizę struktury Produktu Krajowego Brutto – PKB). EKD dzieli podmioty gospodarki narodowej na sekcje, działy, grupy oraz klasy.

GUS gromadzi przede wszystkim dane o wysokim stopniu agregacji. Są one publikowane m.in. w corocznie wydawanych opracowaniach: Małym Roczniku Statystycznym Rzeczpospolitej Polskiej oraz Roczniku Statystycznym RP. GUS wydaje również – z różną częstotliwością – roczniki branżowe (demograficzny, przemysłu, handlu zagranicznego itp.).

W terenie zadania statystyki są realizowane przez podległe GUS-owi urzędy wojewódzkie (zlokalizowane zazwyczaj w miastach wojewódzkich).

Gromadzą one dane dotyczące województw, powiatów, gmin, miast itp. Informacje te stanowią cenne źródło do wszelkiego rodzaju analiz dla samorządów terytorialnych, podmiotów gospodarczych i różnych instytucji funkcjonujących na danym terenie.

Do zadań służb statystyki publicznej należy także prowadzenie, w skali krajowej, dwóch **rejestrów: urzędowego** (zwanego też **administracyjnym**) oraz **podmiotów gospodarki narodowej** (**REGON**). **Rejestr urzędowy** jest wykazem podmiotów informującym o ich działalności. Jest prowadzony przez sądy i organy administracji publicznej na podstawie aktów wykonawczych do ustaw (np. rejestry podatników, działalności gospodarczej, udzielonych zezwoleń i koncesji, gruntów, budynków). **REGON** to powszechny system identyfikacji i kodowania jednostek gospodarki narodowej. W rejestrze tym ujmowane są dane o wszystkich osobach prawnych, jednostkach nieposiadających osobowości prawnej, indywidualnych gospodarstwach rolnych oraz osobach fizycznych prowadzących działalność gospodarczą. Dowodem wpisu do REGON-u jest nadanie określonemu podmiotowi numeru identyfikacyjnego, zawierającego zakodowane informacje o jego cechach.

W pracach statystycznych istotną rolę odgrywa również **rejestr podziału terytorialnego kraju**, który służy do identyfikacji istniejących w Polsce województw, powiatów, gmin, miast i wsi. Nazwy jednostek podziału terytorialnego kraju wraz z ich identyfikatorami są publikowane przez GUS („Wykaz identyfikatorów i nazw jednostek podziału terytorialnego kraju"). W wykazie tym wyszczególniono również system rejonów statystycznych i obwodów spisowych. System ten jest wykorzystywany m.in. przy organizacji masowych badań statystycznych (np. powszechnego spisu ludności). Przydatny jest on także do tworzenia tzw. operatów losowania w badaniach reprezentacyjnych prowadzonych przez GUS.

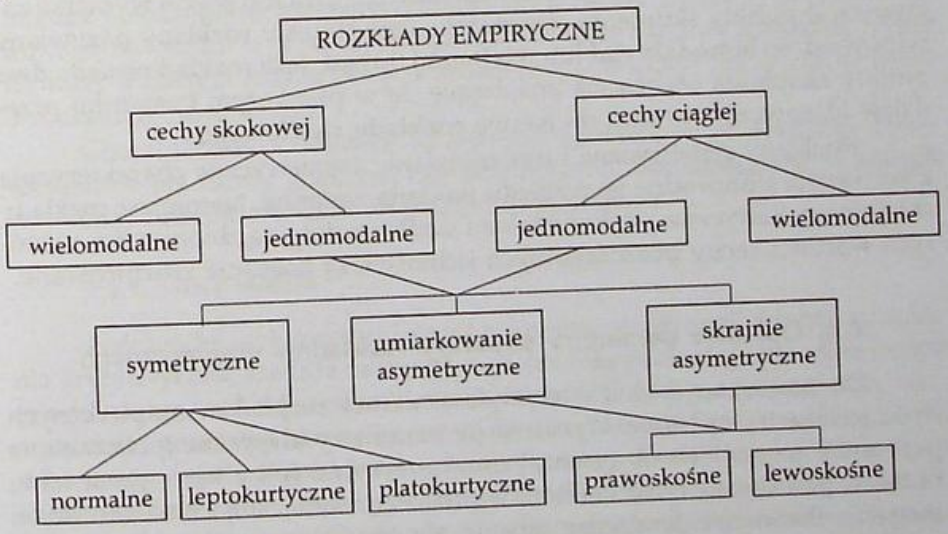
Podkreślić należy, że zebrane dane dotyczące poszczególnych podmiotów, jak również indywidualnych osób są **tajemnicą statystyczną**. Oznacza to, że mogą być one wykorzystywane jedynie do opracowań, zestawień i analiz statystycznych.

Rozdział II

OPISOWA ANALIZA STRUKTURY ZJAWISK MASOWYCH

2.1. Rodzaje rozkładów empirycznych jednowymiarowej zmiennej

O **zmiennej jednowymiarowej** mówimy wówczas, gdy badaniu podlega jedna właściwość (cecha) zbiorowości statystycznej. **Rozkładem empirycznym** zmiennej nazywamy przyporządkowanie kolejnym wartościom zmiennej (x_i) odpowiadających im liczebności (n_i). Rozkłady empiryczne ustalane są na podstawie konkretnych danych statystycznych. Odzwierciedlają one strukturę badanej zbiorowości z punktu widzenia wyróżnionej cechy. Umiejętność odróżniania różnych typów rozkładów jest nieodzownym warunkiem prawidłowej analizy statystycznej. Od rodzaju rozkładu zależy bowiem dobór i zastosowanie odpowiednich charakterystyk służących do opisu zbiorowości. Najczęściej spotykane typy rozkładów empirycznych przedstawia schemat 2.1.



Schemat 2.1. Rodzaje rozkładów empirycznych

Rozkład, którego krzywa liczebności (cecha ciągła) lub diagram (cecha skokowa) ma jedno maksimum nazywa się rozkładem **jednomodalnym**. Wśród rozkładów jednomodalnych można wyróżnić rozkłady **symetryczne**, **umiarkowanie asymetryczne** i **skrajnie asymetryczne**. W symetrycznym jednomodalnym rozkładzie, liczebności odpowiadające wartościom zmiennej rozkładają się symetrycznie wokół liczebności największej.

Wśród symetrycznych, jednomodalnych rozkładów szczególne znaczenie w statystyce ma **rozkład normalny**. Jest to rozkład o ściśle określonym skupieniu wartości cechy wokół średniej arytmetycznej. Z tego też względu nie każdy rozkład symetryczny jest rozkładem normalnym. Na przykład w rozkładzie **leptokurtycznym (wysmukłym)** wartości cechy są bardziej skupione wokół średniej aniżeli w rozkładzie normalnym, natomiast w rozkładzie **platokurtycznym (spłaszczonym)** – mniej skoncentrowane.

Rozkłady empiryczne o charakterze symetrycznym występują rzadko. Częściej spotykamy się z rozkładami zbliżonymi do rozkładu symetrycznego (tj. **rozkładami asymetrycznymi**). Rozkłady asymetryczne mogą być **umiarkowanie asymetryczne** i **skrajnie asymetryczne**, a te z kolei – **prawoskośne** i **lewoskośne**. W rozkładach asymetrycznych większość obserwacji znajduje się w przedziałach położonych bliżej początku szeregu (asymetria lewoskośna) lub w pobliżu końca szeregu (asymetria prawoskośna). W szeregach o asymetrii lewoskośnej dużo jednostek zbiorowości posiada stosunkowo niskie wartości cechy, natomiast niewiele jest obserwacji o wysokich wartościach. W szeregach o asymetrii prawoskośnej sytuacja jest odwrotna: przedział klasowy, zawierający największą liczbę obserwacji, przesunięty jest w prawo.

W niektórych rozkładach można dostrzec dwa lub więcej wyraźnie zarysowane punkty skupienia obserwacji. Tego rodzaju rozkłady nazywamy odpowiednio **bimodalnymi** lub **wielomodalnymi**. Jeśli rozkład posiada dwa punkty skupienia obserwacji znajdujące się w pierwszym i ostatnim przedziale klasowym – to nosi on nazwę **rozkładu siodłowego**.

Rozkłady symetryczne i umiarkowanie asymetryczne charakteryzują zbiorowości jednorodne ze względu na daną zmienną. Natomiast rozkłady skrajnie asymetryczne, wielomodalne i siodłowe dotyczą zbiorowości, w których wartości cechy poszczególnych jednostek są znacznie zróżnicowane.

2.2. Opisowe parametry struktury rozkładów empirycznych

Do sumarycznej charakterystyki struktury rozkładów empirycznych służą **parametry opisowe**. Wyróżnia się parametry **klasyczne** (obliczane na podstawie wszystkich obserwacji) oraz **pozycyjne** (przy ich wyznaczaniu brane są pod uwagę tylko niektóre wartości zmiennej, stojące na określonej pozycji). Parametry klasyczne stosuje się przede wszystkim do analizy rozkładów symetrycznych lub umiarkowanie asymetrycznych. Parametry

pozycyjne są wykorzystywane do badań każdego typu rozkładu, ale zazwyczaj stosowane są w analizie rozkładów silnie asymetrycznych oraz takich, w których występują otwarte przedziały klasowe.

Parametry opisowe rozkładu mogą być **wielkościami absolutnymi** (wyrażonymi w takich jednostkach, jak badana zmienna, np. w kg, godzinach, latach) lub mieć postać **liczb względnych** (ułamkowych lub procentowych). Parametry względne są szczególnie przydatne przy porównywaniu dwóch lub więcej struktur. W szczególności można wyróżnić dwa typy porównań:

1) porównanie dwóch różnych zbiorowości według tej samej cechy badania (np. struktura zgonów według wieku mężczyzn i kobiet w określonym roku);

2) porównanie dotyczące jednej zbiorowości, ale dwóch różnych cech (np. struktura urodzeń żywych według kolejności urodzenia dziecka i wieku matki w Polsce w 2005 r.).

W opisie struktury zbiorowości masowych najczęściej wykorzystywane są następujące parametry:

- **miary przeciętne** (zwane też miarami poziomu wartości zmiennej, położenia lub średnimi). Służą one do określania tej wartości zmiennej opisanej przez rozkład, wokół której skupiają się wszystkie wartości zmiennej;
- **miary rozproszenia** (zmienności, zróżnicowania, dyspersji), służące do badania stopnia zróżnicowania wartości zmiennej;
- **miary asymetrii** (skośności), informujące o kierunku zróżnicowania wartości zmiennej;
- **miary koncentracji i spłaszczenia**. Miary koncentracji służą do badania stopnia nierównomierności rozkładu ogólnej sumy wartości zmiennej między poszczególne jednostki badanej zbiorowości. Miary spłaszczenia informują natomiast o tym, czy skupienie wartości badanej zmiennej wokół średniej w danym rozkładzie jest mniejsze czy większe niż w zbiorowości o rozkładzie normalnym.

Charakterystyki opisowe są bardziej syntetycznymi formami opisu rozkładów niż formy graficzna czy tabelaryczna. Pozwalają one w sposób wymierny określić właściwości badanych rozkładów.

2.2.1. Miary średnie

Najczęściej wykorzystywanymi w analizie struktury średnimi są: **średnia arytmetyczna**, **średnia harmoniczna**, **średnia geometryczna**, **dominanta** (modalna, wartość najczęstsza) oraz **kwantyle**. Średnie: arytmetyczna, harmoniczna i geometryczna zaliczane są do miar klasycznych, pozostałe zaś do miar pozycyjnych. Wśród kwantyli wyróżniamy z kolei **kwartyle** (dzielące zbiorowość na cztery części pod względem liczebności), **kwintyle** (dzielące zbiorowość na pięć części), **decyle** (dzielące zbiorowość na dziesięć części) oraz **percentyle** (dzielące zbiorowość na sto części).

Obydwie grupy średnich (klasyczne i pozycyjne) nie tylko nie wykluczają się, ale nawzajem uzupełniają. Każda z nich opisuje bowiem poziom wartości cechy z innego punktu widzenia. Istnieją jednak sytuacje, kiedy układ informacji liczbowych nie pozwala na obliczenie tej czy innej średniej. Problem ten zostanie wyjaśniony przy omawianiu poszczególnych średnich.

Średnia arytmetyczna

Średnia arytmetyczna jest ilorazem sumy wartości zmiennej i liczebności badanej zbiorowości. Z szeregów wyliczających średnią arytmetyczną obliczamy następująco:

$$\bar{x} = \frac{x_1 + x_2 + \ldots + x_n}{N} = \frac{\sum_{i=1}^{n} x_i}{N}, \qquad (2.1)$$

gdzie: $\bar{x}$ – symbol średniej arytmetycznej, x_i – warianty cechy mierzalnej, N – liczebność badanej zbiorowości.

Średnia określona relacją (2.1) nazywa się średnią arytmetyczną nieważoną (prostą, zwykłą), w przeciwieństwie do **średniej arytmetycznej ważonej**, obliczanej z szeregów rozdzielczych punktowych i przedziałowych.

Jeżeli warianty zmiennej w badanej zbiorowości występują z różną częstotliwością, wówczas oblicza się średnią arytmetyczną ważoną. Wagami są liczebności (częstości) odpowiadające poszczególnym wariantom zmiennej. Wzór na obliczanie średniej arytmetycznej z szeregów rozdzielczych punktowych przyjmuje następującą postać:

$$\bar{x} = \frac{x_1 n_1 + x_2 n_2 + \ldots + x_k n_k}{N} = \frac{\sum_{i=1}^{k} x_i n_i}{N}, \qquad (2.2)$$

gdzie: n_i $(i = 1, 2, \ldots, k)$ – liczebność jednostek odpowiadająca poszczególnym wariantom zmiennej, N – ogólna liczebność badanej zbiorowości $(N = n_1 + n_2 + \ldots + n_k)$.

W szeregach rozdzielczych przedziałowych wartości zmiennej w każdej klasie nie są jednoznacznie określone, ale zawarte są w przedziale od – do. Dolną granicę przedziału klasowego oznaczać będziemy symbolem x_{oi}, górną zaś – x_{1i}. W celu obliczenia średniej arytmetycznej z szeregu rozdzielczego przedziałowego należy uprzednio wyznaczyć środki przedziałów. Środki przedziałów klasowych – oznaczone symbolem $\overset{\circ}{x}_i$ – obliczamy następująco:

$$\overset{\circ}{x}_i = \frac{x_{0i} + x_{1i}}{2}; \ (i = 1, 2, \ldots, k). \qquad (2.3)$$

Wzór na średnią arytmetyczną z szeregu rozdzielczego przedziałowego jest następujący:

$$\bar{x} = \frac{\overset{\circ}{x}_1 n_1 + \overset{\circ}{x}_2 n_2 + \ldots + \overset{\circ}{x}_k n_k}{N} = \frac{\sum_{i=1}^{k} \overset{\circ}{x}_i n_i}{N}. \qquad (2.4)$$

Jeżeli zamiast liczebności absolutnych w obliczeniach zostaną wykorzystane procentowe wskaźniki częstości, to wzór na średnią arytmetyczną przyjmuje postać:

$$\bar{x} = \frac{\sum_{i=1}^{k} \overset{\circ}{x}_i w_i}{N}, \qquad (2.5)$$

gdzie: $w_i = \frac{n_i}{N} \cdot 100$.

Przy obliczaniu średniej arytmetycznej z szeregów rozdzielczych przedziałowych arbitralnie wybiera się wartości reprezentujące poszczególne klasy. Do tego celu wykorzystujemy wzór (2.3). Oznacza to przyjęcie założenia o równomiernym rozkładzie jednostek w poszczególnych przedziałach klasowych. Założenie to nie zawsze jest jednak spełnione. Obliczona średnia arytmetyczna będzie wówczas wielkością przybliżoną. Ogólnie można stwierdzić, że dokładność średniej arytmetycznej obliczonej z szeregów rozdzielczych przedziałowych będzie tym większa, im przedziały klasowe będą węższe.

Sposób obliczania średniej arytmetycznej z szeregu rozdzielczego przedziałowego ilustruje przykład podany w tab. 2.1.

Średnia arytmetyczna obliczona z podanego w tab. 2.1 szeregu rozdzielczego przedziałowego wynosi:

$$\bar{x} = \frac{325}{40} = 8{,}125 \text{ minuty.}$$

Tab. 2.1. Rozkład czasu trwania obsługi w banku X (w minutach)

$x_{0i} - x_{1i}$	n_i	Obliczenia pomocnicze			
		$\overset{\circ}{x}_i$	$\overset{\circ}{x}_i n_i$	$w_i \cdot 100$	$\overset{\circ}{x}_i w_i$
0–5	8	2,5	20,0	20,0	50,00
5–10	20	7,5	150,0	50,0	375,00
10–15	11	12,5	137,5	27,5	343,75
15–20	1	17,5	17,5	2,5	43,75
Ogółem	40	X	325,0	100,0	812,50

Źródło: dane umowne.

Wykorzystując do obliczenia średniej arytmetycznej częstości względne – wzór (2.5) – otrzymujemy:

$$\bar{x} = \frac{812{,}50}{100} = 8{,}125 \text{ minuty.}$$

Uzyskane wyniki są równoważne, gdyż wartość średniej arytmetycznej nie zależy od liczebności absolutnych poszczególnych klas, ale od proporcji między nimi.

Średniej arytmetycznej nie można obliczyć z szeregu rozdzielczego o **otwartych przedziałach klasowych**. Jeśli otwarte przedziały klasowe mają niewielkie liczebności, to przed obliczeniem średniej można je umownie domknąć. Przyjmuje się, że otwarty przedział można domknąć, gdy liczba jednostek w tym przedziale nie przekracza 5% liczebności (czyli 0,05N). Ze względu na to, że główną przyczyną niedomknięcia przedziałów jest wysokie rozproszenie skrajnych wartości cechy, trudno jest trafnie wybrać wartości reprezentujące otwarte klasy.

Często zdarza się, że znamy średnie arytmetyczne dla pewnych grup i na tej podstawie chcemy obliczyć średnią arytmetyczną dla wszystkich grup łącznie. Wykorzystujemy wówczas następujący wzór:

$$\bar{\bar{x}} = \frac{\sum_{i=1}^{k} \bar{x}_i n_i}{N}, \tag{2.6}$$

gdzie: $\bar{x}_i$ jest średnią arytmetyczną i-tej grupy, natomiast n_i – liczebnością i-tej grupy ($N=n_1+n_2+...+n_k$).

Średnia arytmetyczna posiada wiele własności. Najważniejszymi z nich są:

1) jako miara klasyczna jest wypadkową działania wszystkich wartości badanej cechy i spełnia nierówność: $x_{\min} < \bar{x} < x_{\max}$;

2) suma odchyleń poszczególnych wartości zmiennej od średniej arytmetycznej wynosi zero, czyli:

$\sum_{i=1}^{k}(x_i - x) = 0$ w przypadku szeregu wyliczającego,

$\sum_{i=1}^{k}(x_i - \bar{x}) \cdot n_i = 0$ w przypadku szeregu rozdzielczego punktowego,

$\sum_{i=1}^{k}(\overset{\circ}{x}_i - \bar{x}) \cdot n_i = 0$ w przypadku szeregu rozdzielczego przedziałowego;

3) jeśli pomnożymy średnią arytmetyczną przez ogólną liczebność badanej zbiorowości, to otrzymamy sumę wartości wszystkich jednostek, tzn. $N\bar{x} = \sum_{i=1}^{N} x_i$;

4) średnia arytmetyczna sumy (różnicy) zmiennych równa się sumie (różnicy) ich średnich arytmetycznych;

5) jeżeli wszystkie wartości zmiennej powiększymy (pomniejszymy, podzielimy lub pomnożymy) o pewną stałą c, to średnia arytmetyczna będzie równa sumie (różnicy, ilorazowi lub iloczynowi) średniej arytmetycznej i stałej c:

$$\frac{1}{N}\sum_{i=1}^{N}(x_i + c) = \bar{x} + c;$$

6) na poziom średniej arytmetycznej silny wpływ wywierają wartości ekstremalne (skrajne), przy czym wpływ ten jest silniejszy w przypadku wysokich wartości zmiennej;

7) średnia arytmetyczna – jako wypadkowa wszystkich zaobserwowanych wartości cechy – jest wielkością abstrakcyjną. Oznacza to, że w niektórych przypadkach może ona przyjmować wartości w ogóle niewystępujące w zbiorowości;

8) średnia arytmetyczna jest miarą prawidłową tylko w odniesieniu do zbiorowości jednorodnych, o niewielkim zróżnicowaniu wartości zmiennej u poszczególnych jednostek. W miarę wzrostu asymetrii i dyspersji rozkładu, a także w rozkładach bi- i wielomodalnych należy do opisu wykorzystywać przeciętne pozycyjne.

Średnia harmoniczna

Średnia harmoniczna jest odwrotnością średniej arytmetycznej z odwrotności wartości zmiennych. Z szeregów wyliczających średnią harmoniczną obliczamy ze wzoru:

$$H = \frac{N}{\sum_{i=1}^{N}\frac{1}{x_i}}, \tag{2.7}$$

gdzie: H jest symbolem średniej harmonicznej.

Przy obliczaniu średniej harmonicznej z szeregów rozdzielczych (punktowych i przedziałowych) zachodzi konieczność stosowania wag (czyli uwzględniania liczebności). Z szeregów rozdzielczych punktowych średnią harmoniczną obliczamy następująco:

$$H = \frac{N}{\sum_{i=1}^{k}\frac{1}{x_i}n_i}. \tag{2.8}$$

Z szeregów rozdzielczych przedziałowych średnią harmoniczną obliczamy według wzoru (2.8) z tym, że konkretne warianty cechy (x_i) zastępujemy środkami przedziałów klasowych ($\overset{\circ}{x}_i$).

Średnią harmoniczną stosuje się wówczas, gdy wartości zmiennej podane są w jednostkach względnych („łamanych"), np. km/godz., kg/osobę, wagi zaś – w jednostkach liczników tych jednostek względnych. Przykładowo można tu wymienić takie zmienne, jak:

- prędkość pojazdu (zmienna – km/godz., waga – liczba kilometrów),
- gęstość zaludnienia (zmienna – mieszkańcy/km^2, waga – liczba mieszkańców),

❖ spożycie artykułu X na głowę ludności (zmienna – kg/osobę, waga – liczba kilogramów).

Przykładowo, jeśli turysta jechał rowerem przez 2 godziny z prędkością 15 km/h, a przez następne 4 godziny z prędkością 9 km/h, to średnią prędkość jazdy obliczymy za pomocą średniej harmonicznej następująco:

$$H = \frac{30+36}{\frac{1}{15}\cdot 30 + \frac{1}{9}\cdot 36} = 11 \text{ km/h},$$

gdzie: 30 jest przebytą drogą w ciągu 2 godzin, 36 – pokonaną odległością w czasie 4 godzin. Prędkość jest bowiem ilorazem przebytej drogi i czasu. Przebyta przez turystę droga wynosi: $2 \cdot 15 + 4 \cdot 9 = 66$ km. Na pokonanie tej drogi turysta zużył 6 h (2 h + 4 h). Stąd też średnia prędkość wynosi: 66 : 6 = 11 km/h.

Średnia geometryczna

Średnia geometryczna jest pierwiastkiem k-tego stopnia z iloczynu k zmiennych, czyli:

$$G = \sqrt[k]{x_1 \cdot x_2 \cdot \ldots \cdot x_k} = \sqrt[k]{\prod_{i=1}^{k} x_i}, \qquad (2.9)$$

gdzie: π jest znakiem iloczynu określonej liczby wyrazów.

Wzór (2.9) wykorzystywany jest do obliczania średniej geometrycznej z szeregów wyliczających. Średnią geometryczną dla szeregu rozdzielczego punktowego oblicza się ze wzoru:

$$G = \sqrt[N]{x_1^{n_1} \cdot x_2^{n_2} \cdot \ldots \cdot x_k^{n_k}} = \sqrt[N]{\prod_{i=1}^{k} x_i^{n_i}}, \qquad (2.10)$$

gdzie: $N = \sum_{i=1}^{k} n_i$.

Średnią geometryczną dla szeregów rozdzielczych przedziałowych oblicza się na podstawie wzoru (2.10) z tym, że konkretne wartości zmiennej $(x_1, x_2, \ldots, x_k)$ zastępuje się środkami przedziałów klasowych $\overset{\circ}{x}_i$ $(i = 1, 2, \ldots, k)$, gdzie k oznacza liczbę przedziałów klasowych w danym szeregu rozdzielczym.

Średnia geometryczna znajduje zastosowanie przy badaniu średniego tempa zmian zjawisk, których rozwój przedstawiony jest w postaci szeregów dynamicznych. Stąd też do średniej geometrycznej powrócimy w dalszej części podręcznika.

Dominanta (modalna, wartość najczęstsza)

Dominanta jest to najczęściej powtarzająca się wartość zmiennej w szeregu statystycznym. Określa ona najbardziej typową wartość zmiennej

w badanej zbiorowości. Charakterystyczną cechą dominanty jest możliwość jej wyznaczania zarówno z szeregów dotyczących cechy mierzalnej, jak i niemierzalnej. Wartość dominanty można ustalić jedynie z rozkładów jednomodalnych.

W szeregach wyliczających i rozdzielczych punktowych dominanta jest tą wartością cechy, której odpowiada największa liczebność.

W szeregach rozdzielczych przedziałowych bezpośrednio można określić tylko przedział, w którym znajduje się dominanta. Jest to przedział o największej liczebności. Konkretną wartość liczbową, należącą do tego przedziału i będącą dominantą, wyznacza się za pomocą następującego wzoru interpolacyjnego:

$$D = x_D + \frac{n_D - n_{D-1}}{(n_D - n_{D-1}) + (n_D - n_{D+1})} \cdot i_D, \qquad (2.11)$$

gdzie: D – symbol dominanty, x_D – dolna granica klasy, w której znajduje się dominanta, n_D – liczebność przedziału dominanty, n_{D-1} – liczebność przedziału poprzedzającego przedział dominanty, n_{D+1} – liczebność przedziału następującego po przedziale dominanty, i_D – interwal (rozpiętość) przedziału dominanty.

Przy wyznaczaniu dominanty za pomocą wzoru (2.11) wskazany jest bezwzględny wymóg, by rozpiętości przedziału dominanty i obu przedziałów bezpośrednio z nim sąsiadujących były jednakowe. Technikę wyznaczania dominanty z szeregów rozdzielczych przedziałowych zilustrujemy przykładem liczbowym (dane w tab. 2.2).

Tab. 2.2. Liczba podmiotów gospodarczych działających w gminach wiejskich w województwie L

Liczba podmiotów	Liczba gmin
4–8	18
9–13	24
14–18	8
19–23	2

Źródło: dane umowne.

Z danych liczbowych zawartych w tab. 2.2 wynika, że dominanta znajduje się w przedziale 9–13. Wartość dominanty, wyznaczona za pomocą wzoru (2.11) jest zatem równa:

$$D = 9 + \frac{24-18}{(24-18)+(24-8)} \cdot 5 = 10{,}36 \approx 10.$$

Otrzymany wynik oznacza, że najczęściej spotykana liczba podmiotów gospodarczych w badanej zbiorowości gmin oscyluje wokół liczby 10. Wartość ta mieści się w przedziale najliczniejszym.

Jeżeli liczebność przedziałów przed i za przedziałem dominanty są jednakowe, to dominanta jest równa środkowi klasy dominującej. W przypadku, gdy liczebność klasy poprzedzającej jest większa od liczebności klasy następującej, dominanta będzie wartością bliższą dolnej granicy przedziału, w którym jest zawarta i odwrotnie.

Z szeregów rozdzielczych przedziałowych dominantę można również wyznaczyć graficznie. Graficzne wyznaczanie modalnej sprowadza się do wykreślenia histogramu liczebności i połączenia dwoma odcinkami wierzchołków najwyższego prostokąta, po przekątnej, z najbliższymi wierzchołkami sąsiednich prostokątów. Rzut punktu przecięcia tych odcinków na oś odciętych wskazuje wartość dominanty[1].

Kwantyle

Spośród kwantyli najczęściej używanymi miarami są **kwartyle**, wśród których wyróżniamy **kwartyl pierwszy** (dolny), **kwartyl drugi** (zwany też medianą lub wartością środkową) oraz **kwartyl trzeci** (górny)[2]. Każdy z kwartyli dzieli zbiorowość uporządkowaną niemalejąco pod względem liczebności na dwie części, przy czym:

- **kwartyl pierwszy** dzieli zbiorowość na dwie części w ten sposób, że 25% jednostek zbiorowości ma wartości zmiennej mniejsze lub równe kwartylowi pierwszemu, a 75% – równe lub większe od tego kwartyla;
- **mediana** dzieli zbiorowość na dwie części w ten sposób, że 50% jednostek ma wartości mniejsze lub równe medianie oraz 50% – równe lub większe od mediany;
- **kwartyl trzeci** dzieli zbiorowość na dwie części w ten sposób, że 75% jednostek ma wartości zmiennej mniejsze lub równe kwartylowi trzeciemu, a 25% – równe lub większe od kwartyla trzeciego.

Z szeregów wyliczających (składających się zazwyczaj z niewielkiej liczby jednostek) najczęściej wyznacza się medianę. W przypadku, gdy liczba obserwacji jest **nieparzysta**, medianą jest wartość środkowa. Jeśli natomiast liczna jednostek zbiorowości jest **parzysta** – mediana jest średnią arytmetyczną dwóch środkowych wartości zmiennej. Sposób wyznaczania mediany z szeregów wyliczających można zapisać następującymi wzorami:

$$Me = x_{\frac{N+1}{2}}, \text{ gdy } N \text{ jest nieparzyste,} \tag{2.12}$$

$$Me = \frac{1}{2}\left(x_{\frac{N}{2}} + x_{\frac{N}{2}+1} \right), \text{ gdy } N \text{ jest parzyste.} \tag{2.13}$$

[1] Por. M. Sobczyk, *Statystyka. Podstawy teoretyczne, przykłady i zadania*, Wydawnictwo UMCS, Lublin 2000, s. 48.

[2] Pozostałe kwantyle: kwintyle, decyle i percentyle jest sens obliczać tylko wtedy, gdy zbiorowość jest wystarczająco liczna.

Wyznaczanie mediany z szeregu rozdzielczego punktowego sprowadza się do wskazania jednostki środkowej i odczytania wariantu cechy odpowiadającej tej jednostce. Określenie środkowej jednostki ułatwia kumulacja liczebności, polegająca na kolejnym, narastającym sumowaniu liczebności dotyczących poszczególnych wariantów badanej zmiennej. Jeśli kumulacji podlegają częstości względne, a nie liczebności absolutne, to otrzymany zbiór danych określa się mianem **dystrybuanty empirycznej**. Graficznym obrazem kształtowania się liczebności skumulowanych jest histogram liczebności skumulowanej lub diagram liczebności skumulowanej.

Z szeregów rozdzielczych przedziałowych kwartyle wyznaczamy wykorzystując następujące wzory interpolacyjne:

$$Q_1 = x_{Q_1} + \frac{\frac{N}{4} - \sum_{i=1}^{k-1} n_i}{n_{Q_1}} \cdot i_{Q_1}, \tag{2.14}$$

$$Q_2 = Me = x_{Me} + \frac{\frac{N}{2} - \sum_{i=1}^{k-1} n_i}{n_{Me}} \cdot i_{Me}, \tag{2.15}$$

$$Q_3 = x_{Q_3} + \frac{\frac{3}{4}N - \sum_{i=1}^{k-1} n_i}{n_{Q_3}} \cdot i_{Q_3}, \tag{2.16}$$

gdzie: Q_1, $Q_2(Me)$, Q_3 – odpowiednio kwartyl pierwszy, kwartyl drugi (mediana) oraz kwartyl trzeci, x_{Q_1}, x_{Q_2}, x_{Q_3} – dolne granice przedziałów, w których znajdują się odpowiednio kwartyl pierwszy, mediana i kwartyl trzeci, $\sum_{i=1}^{k-1} n_i$ – suma liczebności od klasy pierwszej do poprzedzającej tę, w której znajdują się odpowiednio kwartyl pierwszy, mediana i kwartyl trzeci, n_{Q_1}, n_{Me}, n_{Q_3} – liczebności przedziałów, w których znajdują się odpowiednio kwartyl pierwszy, mediana i kwartyl trzeci, i_{Q_1}, i_{Me}, i_{Q_3} – interwały przedziałów, w których znajdują się odpowiednio kwartyl pierwszy, mediana i kwartyl trzeci.

Technikę obliczania kwartyli z szeregów rozdzielczych przedziałowych pokażemy na przykładzie (tab. 2.3).

Pierwszym krokiem przy wyznaczaniu kwartyli jest określenie klasy, w której znajduje się odpowiedni kwartyl. W tym celu obliczamy kolejno: $\frac{N}{4}$, $\frac{N}{2}$ i $\frac{3N}{4}$, a następnie korzystamy z dystrybuanty empirycznej (bądź też z szeregu skumulowanych liczebności bezwzględnych).

Korzystając z informacji liczbowych zawartych w tab. 2.3 mamy:

$$\frac{N}{4}=\frac{100}{4}=25,\ \frac{N}{2}=\frac{100}{2}=50,\ \frac{3N}{4}=\frac{300}{4}=75.$$

Tak więc, kwartyl pierwszy (Q_1) znajduje się w przedziale 200–300 tys. zł, kwartyl drugi ($Q_2 = Me$) oraz kwartyl trzeci (Q_3) w przedziale 300–400 tys. zł.

Tab. 2.3. Roczna sprzedaż (w tys. zł) uzyskana przez przedstawicieli handlowych

Sprzedaż w tys. zł	Liczba przedstawicieli	Odsetek przedstawicieli	Dystrybuanta empiryczna
0–100	2	4,0	4,0
100–200	4	8,0	12,0
200–300	13	26,0	38,0
300–400	19	38,0	76,0
400–500	10	20,0	96,0
500–600	2	4,0	100,0
Razem	50	100,0	X

Źródło: dane umowne.

Wartość kwartyla pierwszego jest równa (wzór (2.14)):

$$Q_1 = 200 + \frac{25-12}{26}\cdot 100 = 250 \text{ tys. zł.}$$

Oznacza to, że 25% przedstawicieli handlowych uzyskało roczną sprzedaż w wysokości nieprzekraczającej 250 tys. zł, a pozostałych 75% przedstawicieli – nie mniejszą niż 250 tys. zł.

Mediana wynosi (wzór (2.15)):

$$Me = 300 + \frac{50-26}{38}\cdot 100 = 363{,}158 \text{ tys. zł.}$$

Otrzymany wynik pozwala stwierdzić, że 50% przedstawicieli handlowych uzyskało roczną sprzedaż nie większą niż 363,158 tys. zł, a 50% – nie mniejszą niż 363,158 tys. zł.

Wartość kwartyla trzeciego obliczamy następująco (wzór (2.16)):

$$Q_3 = 300 + \frac{75-26}{38}\cdot 100 = 428{,}947 \text{ tys. zł.}$$

Okazuje się więc, że 75% przedstawicieli handlowych osiągnęło roczną sprzedaż mniejszą lub równą kwocie 428,947 tys. zł, a pozostałych 25% przedstawicieli – większą lub równą 428,947 tys. zł.

Kwartyle można również wyznaczać graficznie. W tym celu należy sporządzić wykres skumulowanego wieloboku liczebności (jeśli do obliczeń wykorzystujemy liczebności absolutne) lub dystrybuanty empirycznej (jeśli posługujemy się częstościami względnymi)[3].

Kwartyle (podobnie jak dominantę) można wyznaczać z szeregów o otwartych przedziałach klasowych. Ponadto, w odróżnieniu od średniej arytmetycznej, nie reagują one na wartości ekstremalne (skrajne).

Średnia arytmetyczna, mediana i modalna jako miary tendencji centralnej powiązane są ze sobą odpowiednimi zależnościami, które wyrażają się równościami lub nierównościami (decyduje o tym typ rozkładu empirycznego). Zależności te zostaną omówione przy miarach asymetrii. W tym miejscu podajemy jedynie, że w rozkładach symetrycznych miary te są sobie równe ($\bar{x} = D = Me$). W rozkładach asymetrycznych, często występujących w praktyce badań statystycznych, związek między wymienionymi charakterystykami jest określony wzorem:

$$\bar{x} - D = 3(\bar{x} - Me). \tag{2.17}$$

Wzorem (2.17) można posłużyć się do obliczenia nieznanego parametru, gdy znane są dwa pozostałe.

2.2.2. Miary zmienności

Wartości średnie nie wystarczają do scharakteryzowania struktury zbiorowości. Badana zbiorowość statystyczna może bowiem charakteryzować się różnym stopniem **zmienności (rozproszenia, zróżnicowania, dyspersji)** badanej cechy. Dyspersją nazywamy zróżnicowanie jednostek zbiorowości ze względu na wartość badanej cechy. Zmienność ocenia się za pomocą wielu miar statystycznych, wśród których wyróżnia się miary **klasyczne** i **pozycyjne** oraz **bezwzględne (absolutne)** i **względne (relatywne, stosunkowe)**. Klasyczne miary zmienności obliczane są na podstawie **wszystkich** wartości badanej cechy, pozycyjne – na podstawie **niektórych** (stojących na określonej pozycji) wartości. Do klasycznych miar zmienności zaliczamy **odchylenie standardowe, wariancję** i **współczynnik zmienności** (obliczany przy użyciu odchylenia standardowego i średniej arytmetycznej). W grupie pozycyjnych miar zmienności wyróżnia się: **empiryczny obszar zmienności** (zwany też **rozstępem**), **odchylenie ćwiartkowe** oraz **pozycyjny współczynnik zmienności**.

Bezwzględne miary zmienności (tj. rozstęp, odchylenie ćwiartkowe, wariancja oraz odchylenie standardowe) są wielkościami mianowanymi, posiadającymi miano badanej cechy. **Względną miarą dyspersji** jest współczynnik zmienności, wyrażany w procentach.

[3] Por. M. Sobczyk, *op. cit.*, s. 51.

Rozstęp

Rozstęp jest różnicą między największą i najmniejszą wartością cechy. Oblicza się go ze wzoru:

$$R = x_{max} - x_{min}. \tag{2.18}$$

Jak wynika z relacji (2.18), wartość rozstępu zależy jedynie od dwóch skrajnych wielkości: najmniejszej i największej. Brakuje zatem informacji o zróżnicowaniu pozostałych jednostek zbiorowości pod względem badanej cechy. Dlatego też rozstęp stosowany jest głównie w przypadkach, gdy niezbędna jest wstępna orientacja o obszarze zmienności cechy.

Odchylenie ćwiartkowe

Odchylenie ćwiartkowe oblicza się na podstawie kwartyli. Definiuje się go jako połowę różnicy między kwartylem trzecim i pierwszym, czyli:

$$Q = \frac{Q_3 - Q_1}{2}. \tag{2.19}$$

Różnicę $Q_3 - Q_1$ określa się mianem rozstępu kwartylowego lub rozstępu międzykwartylowego.

Odchylenie ćwiartkowe mierzy poziom zróżnicowania jedynie połowy jednostek (50%), pozostałych po odrzuceniu 25% jednostek o wartościach mniejszych od kwartyla pierwszego ($x_i < Q_1$) i 25% jednostek większych od kwartyla trzeciego ($x_i > Q_3$). Miara ta nie jest więc wrażliwa na skrajne wartości zbioru. Z tego też względu nie należy do zbyt często wykorzystywanych miar zmienności.

Jeżeli do opisu tendencji centralnej użyto mediany, a do opisu zmienności odchylenia ćwiartkowego – to możliwe jest określenie **typowego obszaru zmienności** badanej cechy. Obszar ten określa następująca nierówność:

$$Me - Q < x_{typ} < Me + Q. \tag{2.20}$$

Nietypowe w danej zbiorowości są jednostki o wartości niższej do różnicy $Me - Q$ i wyższej od sumy $Me - Q$.

Odchylenie ćwiartkowe jest szczególnie przydatne w analizie statystycznej szeregów rozdzielczych przedziałowych o klasach otwartych. Odchylenie ćwiartkowe interpretuje się jako przeciętne zróżnicowanie badanych jednostek wokół mediany.

Wariancja i odchylenie standardowe

Najczęściej wykorzystywanymi miarami zróżnicowania są **wariancja i odchylenie standardowe**. Wariancja jest to średnia arytmetyczna kwadratów odchyleń poszczególnych wartości cechy od ich średniej arytmetycznej. Oznaczamy ją symbolem s^2 i obliczamy w następujący sposób[4].

$$s^2 = \frac{1}{N}\sum_{i=1}^{N}(x_i - \bar{x})^2 \text{ dla szeregów wyliczających,} \tag{2.21}$$

$$s^2 = \frac{1}{N}\sum_{i=1}^{k}(x_i - \bar{x})^2 n_i \text{ dla szeregów rozdzielczych punktowych,} \tag{2.22}$$

$$s^2 = \frac{1}{N}\sum_{i=1}^{k}(\mathring{x}_i - \bar{x})^2 n_i \text{ dla szeregów rozdzielczych przedziałowych.} \tag{2.23}$$

Niektórzy statystycy twierdzą, że obliczana z szeregów rozdzielczych przedziałowych (o równych rozpiętościach klas i liczbie przedziałów mniejszych od 12) wariancja jest przeszacowana. Dlatego też zaleca się – w takich przypadkach – stosowanie poprawki Shepparda, która wynosi $\frac{i^2}{12}$. Należy wtedy posługiwać się wzorem[5]:

$$s^2 = \frac{1}{N}\sum_{i=1}^{k}(\mathring{x}_i - \bar{x})^2 n_i - \frac{i^2}{12}. \tag{2.24}$$

Wariancja jako miara zróżnicowania ma dwie ważne właściwości, a mianowicie:

1) wariancja jest różnicą między średnią arytmetyczną kwadratów wartości zmiennej i kwadratem jej średniej arytmetycznej, czyli:

$$s^2 = \bar{x}^2 - (\bar{x})^2, \tag{2.25}$$

2) jeżeli badaną zbiorowość podzielimy na k grup, to wariancja ogólna (całej zbiorowości) jest sumą dwóch składników: wariancji wewnątrzgrupowej i wariancji międzygrupowej:

$$s^2 = \bar{s}_i^2 + s^2(\bar{x}_i) = \frac{\sum_{i=1}^{k} s_i^2 n_i}{N} + \frac{\sum_{i=1}^{k}\left(x_i - \bar{\bar{x}}\right)^2 n_i}{N}, \tag{2.26}$$

gdzie: $N = \sum_{i=1}^{k} n_i$, $i = 1, 2, ..., k$, s_i^2 – wariancja i-tej grupy, $\bar{x}_i$ – średnia arytmetyczna i-tej grupy, $\bar{\bar{x}}$ – średnia ogólna (średnia ze średnich wszystkich grup).

Własność określona wzorem (2.26) nosi nazwę **równości wariancyjnej**.

---

[4] W statystyce matematycznej zamiast symbolu N używa się n. Jeśli wariancja jest obliczana z próby małej ($n \leq 30$), to przy jej obliczaniu sumę kwadratów różnic dzieli się nie przez n, tylko przez $n - 1$. Por. np. A. Zeliaś, B. Pawełek, S. Wanat, *Metody statystyczne. Zadania i sprawdziany*, PWE, Warszawa 2002, s. 46.

[5] Por. A. Zeliaś, *Metody statystyczne*, PWE, Warszawa 2000, s. 60.

Technikę obliczania wariancji z równości wariancyjnej zilustrujemy przykładem. W przedsiębiorstwie składającym się z trzech wydziałów, zebrano następujące informacje dotyczące stażu pracy:

zakład I: $n_1 = 50$ pracowników, $\bar{x}_1 = 12$ lat, $s_1^2 = 3$ (lata)^2;
zakład II: $n_2 = 120$ pracowników, $\bar{x}_2 = 8$ lat, $s_2^2 = 2$ (lata)^2;
zakład III: $n_3 = 70$ pracowników, $\bar{x}_3 = 10$ lat, $s_3^2 = 4$ (lata)^2.

Obliczyć zróżnicowanie stażu pracy w całym przedsiębiorstwie.

Średni staż pracy w całym przedsiębiorstwie obliczamy następująco:

$$\bar{\bar{x}} = \frac{12 \cdot 50 + 8 \cdot 120 + 10 \cdot 70}{50 + 120 + 70} = \frac{2260}{240} = 9{,}4 \text{ lat.}$$

Wewnątrzgrupowe zróżnicowanie stażu pracy jest równe:

$$s_i^2 = \frac{3 \cdot 50 + 2 \cdot 120 + 4 \cdot 70}{50 + 120 + 70} = \frac{670}{240} = 2{,}8 \text{ (lat)}^2.$$

Międzygrupowe zróżnicowanie stażu pracy wynosi:

$$s^2(\bar{x}_i) = \frac{(12 - 9{,}4)^2 \cdot 50 + (8 - 9{,}4)^2 \cdot 120 + (10 - 9{,}4)^2 \cdot 70}{50 + 120 + 70} = \frac{589{,}4}{240} = 2{,}5 \text{ (lat)}^2.$$

Wariancja ogólna jest sumą wariancji wewnątrzgrupowej i międzygrupowej, czyli:

$$s^2 = 2{,}8 + 2{,}5 = 5{,}3 \text{ (lat)}^2.$$

Wynika stąd, że na wariancję ogólną nieco większy wpływ ma zróżnicowanie wewnątrzgrupowe niż międzygrupowe. Ponadto, wariancja ogólna jest zawsze większa aniżeli wariancja poszczególnych grup.

Wariancja jest wielkością nieujemną ($s^2 \geq 0$) i mianowaną. Jej mianem jest kwadrat jednostki fizycznej, w jakiej mierzona jest badana zmienna. Stąd też wariancja jest trudna do merytorycznej interpretacji. W celu uzyskania miary zmienności o mianie zgodnym z mianem badanej cechy, oblicza się dodatni pierwiastek kwadratowy z wariancji. Otrzymana w ten sposób miara jest określana mianem **odchylenia standardowego**:

$$s = \sqrt{s^2}. \tag{2.27}$$

Odchylenie standardowe określa, o ile – średnio biorąc – jednostki zbiorowości różnią się od średniej arytmetycznej badanej zmiennej. Im zbiorowość jest bardziej zróżnicowana, tym wariancja (a więc i odchylenie standardowe) jest większa. Jeśli wartości badanej cechy mierzone są w skali porządkowej (tzn. są rangowane za pomocą liczb naturalnych), to wariancję obliczamy następująco:

$$s^2 = \frac{n^2 - 1}{12}. \tag{2.28}$$

Odchylenie standardowe można wykorzystać do budowy **typowego obszaru zmienności** badanej cechy. Typowy obszar zmienności zawarty jest w przedziale:

$$\bar{x} - s \leq x_{typ} \leq \bar{x} + s. \tag{2.29}$$

Z odchyleniem standardowym wiąże się tzw. **reguła trzech sigm**. Zgodnie z nią, wystąpienie obserwacji o wartości cechy spoza przedziału $(\bar{x} - 3s;\ \bar{x} + 3s)$ jest mało prawdopodobne. W przypadku rozkładów o niewielkiej asymetrii, tylko ok. 0,3% obserwacji wykracza poza przedział:

$$\bar{x} - 3s < x_i < \bar{x} + 3s. \tag{2.30}$$

W rozkładach regularnych (symetrycznych, jednomodalnych) ok. 68% obserwacji odchyla się od średniej arytmetycznej o mniej niż jedno odchylenie standardowe, ok. 95% obserwacji odchyla się o mniej niż 2 odchylenia standardowe i niemal wszystkie – o mniej niż 3 odchylenia standardowe.

Standaryzacja wartości cechy

Odchylenie standardowe może być wykorzystane do **standaryzacji** wartości cechy. Standaryzacja jest przekształceniem pierwotnych wartości cechy x_i w wartości nowej cechy z_i według wzoru:

$$z_i = \frac{x_i - \bar{x}}{s}, \tag{2.31}$$

gdzie: $\bar{x}$ i s są odpowiednio średnią arytmetyczną i odchyleniem standardowym pierwotnych wartości cechy.

Wartość standaryzowana informuje o tym, o ile odchyleń standardowych pierwotna wartość cechy jest większa lub mniejsza od średniej arytmetycznej. Standaryzacji podlegają wyłącznie cechy ilościowe, gdyż tylko wtedy można obliczyć średnią arytmetyczną i odchylenie standardowe. Wartości cechy większej od średniej arytmetycznej odpowiada dodatnia wartość standaryzowana ($z_i > 0$), a wartościom niższym – ujemna wartość standaryzowana ($z_i < 0$).

Średnia arytmetyczna zbioru danych standaryzowanych wynosi zero, a odchylenie standardowe jest równe jedności. Wynika to z następujących przekształceń:

$$\bar{z} = \frac{1}{N}\sum_{i=1}^{N} z_i = \frac{1}{N}\sum_{i=1}^{N} \frac{x_i - \bar{x}}{s} = \frac{1}{Ns}\sum_{i=1}^{N}(x_i - \bar{x}) = 0, \tag{2.32}$$

oraz

$$s_z^2 = \frac{1}{N}\sum_{i=1}^{N}(z_i - \bar{z})^2 = \frac{1}{N}\sum_{i=1}^{N} z_i^2 = \frac{1}{N}\sum_{i=1}^{N}\left(\frac{x_i - \bar{x}}{s}\right)^2 = \frac{1}{Ns^2}\sum_{i=1}^{N}(x_i - \bar{x})^2 = \frac{s^2}{s^2} = 1. \tag{2.33}$$

Dane standaryzowane pochodzące z różnych rozkładów mogą być ze sobą porównywalne. Załóżmy, że uczeń otrzymał 50 punktów z egzaminu z języka angielskiego, podczas gdy średnia arytmetyczna dla całej klasy wynosiła 45 punktów z odchyleniem standardowym 5 punktów. Ten sam uczeń otrzymał 60 punktów z egzaminu z matematyki przy średniej klasowej równej 55 punktów i odchyleniu standardowym 10 punktów. W którym z wymienionych przedmiotów uczeń jest lepszy w porównaniu do reszty klasy?

Udzielenie odpowiedzi na postawione pytanie wymaga dokonania standaryzacji wyników. Standaryzowana ocena z języka angielskiego jest równa 1 $\left(\frac{50-45}{5}=1\right)$, a standaryzowana ocena z matematyki wynosi 0,5 $\left(\frac{60-55}{10}=0{,}5\right)$. Z obliczeń wynika więc, że na tle klasy uczeń jest lepszy z języka angielskiego niż z matematyki.

Współczynnik zmienności

Przedstawione dotychczas miary nie pozwalają na porównywanie zmienności tej samej cechy w różnych zbiorowościach, czy kilku cech wyrażonych w odmiennych mianach. W takich przypadkach wykorzystuje się niemianowaną (najczęściej wyrażaną w procentach) miarę zróżnicowania – **współczynnik zmienności**. Współczynnik zmienności jest ilorazem absolutnej miary zróżnicowania i przeciętnego poziomu wartości cechy. Z uwagi na fakt, że przy analizie rozkładu wartości cechy posługujemy się różnymi miarami dyspersji i przeciętnymi, współczynnik zmienności można obliczyć kilkoma metodami, a mianowicie:

$$V_s = \frac{s}{\bar{x}} \cdot 100, \tag{2.34}$$

$$V_Q = \frac{Q}{Me} \cdot 100, \tag{2.35}$$

$$V_{Q,Q_3} = \frac{Q_3 - Q_1}{Q_3 + Q_1} \cdot 100, \tag{2.36}$$

gdzie: V jest symbolem współczynnika zmienności, a indeks przy nim informuje o rodzaju bezwzględnej miary dyspersji użytej w obliczeniach.

Współczynnik zmienności obliczony za pomocą wzoru (2.34) nosi miano **klasycznego**, relacje (2.35) i (2.36) określają **pozycyjne** współczynniki zmienności.

Jeżeli współczynnik zmienności przyjmuje wysokie wartości liczbowe, to fakt ten świadczy o niejednorodności badanej zbiorowości statystycznej. Umownie przyjmuje się, że jeżeli współczynnik V nie przekracza 10%, to cechy wykazują niewielkie zróżnicowanie. Taką zbiorowość uznaje się za jednorodną, co rzutuje na poprawne wyniki analizy statystycznej.

Poniżej podajemy przykład praktycznego zastosowania współczynnika zmienności.

Analiza wysokości miesięcznych wpływów za świadczenie usług noclegowych przez trzy hotele dostarczyła następujących informacji:

- ❖ średnie wartości usług w hotelach A, B, C były równe odpowiednio (w tys. zł): 60, 30, 50;
- ❖ odchylenia standardowe wartości sprzedanych usług wynosiły odpowiednio (w tys. zł): 11, 9 oraz 12.

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

Podane w tym przykładzie wartości odchyleń standardowych nie mogą stanowić podstawy do wyciągania wniosków o stopniu zmienności ze względu na znaczne różnice w średnim poziomie wpływów w poszczególnych hotelach. Do tego celu należy wykorzystać klasyczny współczynnik zmienności. Podstawiając odpowiednie dane liczbowe do wzoru (2.34) otrzymujemy:

- dla hotelu A: $V_s = \frac{11}{60} \cdot 100 = 18{,}33\%$,
- dla hotelu B: $V_s = \frac{9}{30} \cdot 100 = 30\%$,
- dla hotelu C: $V_s = \frac{12}{50} \cdot 100 = 24\%$.

Tak więc najmniejsza dyspersja miesięcznych wpływów za świadczenie usług noclegowych występuje w hotelu A.

2.2.3. Miary asymetrii

Z punktu widzenia potrzeb analizy struktury zjawisk masowych istotny jest nie tylko przeciętny poziom i wewnętrzne zróżnicowanie zbiorowości, ale także to, czy przeważająca liczba jednostek tworzących badaną zbiorowość ma wartości cechy wyższe czy niższe od przeciętnego poziomu. Problem ten wiąże się z oceną kierunku **asymetrii (skośności) rozkładu**.

Asymetrię rozkładu najłatwiej jest określić przez porównanie takich jego charakterystyk, jak średnia arytmetyczna, mediana, oraz modalna. W **rozkładach symetrycznych** średnie te są sobie równe.

W **rozkładach asymetrycznych** wymienione średnie różnią się między sobą, a różnice między nimi są tym większe, im empiryczny rozkład badanej cechy bardziej odbiega od symetrycznego. Jeśli spełniona jest nierówność: $\bar{x} > Me > D$, to rozkład charakteryzuje się asymetrią **prawostronną (dodatnią)**. Jeśli natomiast zachodzi nierówność: $\bar{x} < Me < D$, to mówimy o asymetrii **lewostronnej (ujemnej)**.

Wskaźnik asymetrii

Do określania **kierunku asymetrii** (tzn. czy asymetria jest prawostronna czy lewostronna) rozkładu wykorzystuje się **wskaźnik** asymetrii (skośności) określony wzorem:

$$W_s = \bar{x} - D. \tag{2.37}$$

Jeśli różnica $\bar{x} - D$ jest dodatnia, mamy do czynienia z asymetrią prawostronną. W przypadku asymetrii lewostronnej wskaźnik skośności jest ujemny. W rozkładzie symetrycznym zachodzi równość: $\bar{x} = D$.

Kierunek asymetrii rozkładu można również określić na podstawie kwartyli. Zachodzą wówczas następujące zależności:

- przy rozkładzie symetrycznym:

$$(Q_3 - Q_2) - (Q_2 - Q_1) = 0, \tag{2.38}$$

- przy rozkładzie o asymetrii prawostronnej:

$$(Q_3 - Q_2) - (Q_2 - Q_1) > 0, \tag{2.39}$$

- przy rozkładzie o asymetrii lewostronnej:

$$(Q_3 - Q_2) - (Q_2 - Q_1) < 0. \tag{2.40}$$

Wskaźnik skośności jest liczbą mianowaną. Określa jedynie kierunek asymetrii, a nie informuje o jej sile, gdyż jest wielkością nieunormowaną.

Moment standaryzowany trzeciego rzędu

Miarą określającą zarówno kierunek, jak i siłę asymetrii jest **współczynnik asymetrii**. Jest on definiowany za pomocą **momentu standaryzowanego trzeciego rzędu** i określony wzorem:

$$A_s = \frac{m_3}{s^3}. \tag{2.41}$$

Licznik wzoru (2.41) wyraża przeciętną wielkość trzecich potęg odchyleń wartości cechy od średniej arytmetycznej, czyli:

$$m_3 = \frac{1}{N}\sum_{i=1}^{N}(x_i - \bar{x})^3 \text{ dla szeregu wyliczającego,} \tag{2.42}$$

$$m_3 = \frac{1}{N}\sum_{i=1}^{k}(x_i - \bar{x})^3 n_i \text{ dla szeregu rozdzielczego punktowego,} \tag{2.43}$$

$$m_3 = \frac{1}{N}\sum_{i=1}^{N}\left(\overset{\circ}{x}_i - \bar{x}\right)^3 n_i \text{ dla szeregu rozdzielczego przedziałowego.} \tag{2.44}$$

Mianownik wzoru (2.41) jest trzecią potęgą odchylenia standardowego badanej cechy.

W przypadku rozkładów o asymetrii prawostronnej $A_s > 0$, a lewostronnej $A_s < 0$. W rozkładach symetrycznych $A_s = 0$. Znak współczynnika A_s określa więc kierunek asymetrii, jego moduł wskazuje na siłę asymetrii. Im większa jest wartość bezwzględna współczynnika, tym silniejsza jest asymetria rozkładu. Jeżeli asymetria nie jest zbyt silna, to wartość standaryzowanego momentu trzeciego rzędu zawiera się w granicach:

$$-1 \le A_s \le +1. \tag{2.45}$$

Jedynie przy ekstremalnie silnej asymetrii, bezwzględna wartość współczynnika asymetrii przekracza 2.

Współczynnik asymetrii

Najpopularniejszą względną miarą asymetrii jest współczynnik asymetrii o następującej postaci:

$$A_s = \frac{\bar{x} - D}{s}. \tag{2.46}$$

Wzór (2.46) określa, ile odchyleń standardowych zawiera różnica między średnią arytmetyczną a dominantą. Własności tego współczynnika asymetrii są takie same jak współczynnika określonego wzorem (2.41). Podkreślić jednak należy, że współczynnik obliczony za pomocą relacji (2.46) jest przybliżoną miarą asymetrii. Za precyzyjniejszą miarę asymetrii uznaje się moment standaryzowany trzeciego rzędu.

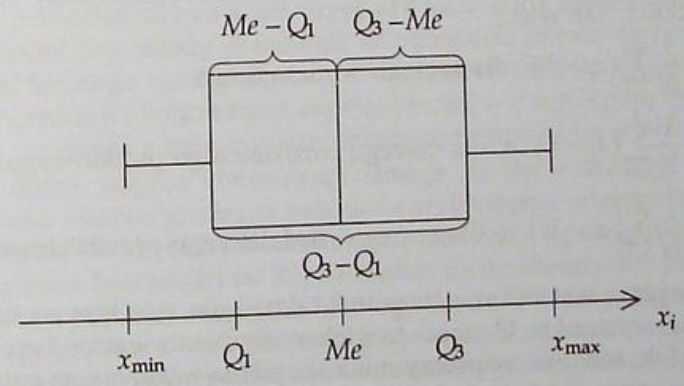
Pozycyjny współczynnik asymetrii

Pozycyjny współczynnik asymetrii wykorzystuje się zazwyczaj wtedy, gdy rozkład empiryczny nie spełnia warunków niezbędnych do obliczenia dominanty czy średniej arytmetycznej. Współczynnik ten jest określony wzorem:

$$A_s = \frac{(Q_3 - Me) - (Me - Q_1)}{(Q_3 - Me) + (Me - Q_1)} = \frac{(Q_3 - Me) - (Me - Q_1)}{Q_3 - Q_1} = \frac{Q_3 + Q_1 - 2Me}{2Q}. \tag{2.47}$$

Z konstrukcji wzoru (2.47) wynika, że pozycyjny współczynnik asymetrii określa kierunek i siłę asymetrii jednostek znajdujących się w drugiej i trzeciej ćwiartce obszaru zmienności, a więc w zawężonej „przestrzeni". Z tego też względu pozycyjny współczynnik asymetrii traktowany jest jako miara uzupełniająca.

W odniesieniu do szeregów wyliczających, użytecznym narzędziem oceny asymetrii rozkładu z wykorzystaniem kwartyli jest tzw. **wykres pudełkowy** (zwany też ramkowym). Na wykresie tym zaznacza się pięć syntetycznych charakterystyk rozkładu: kwartyl pierwszy, medianę, kwartyl drugi, najmniejszy i największy wynik obserwacji (por. rys. 2.1).



Rys. 2.1. Wykres pudełkowy

Wykres pudełkowy składa się z prostokąta, którego szerokość wyznaczają kwarytyle Q_1 i Q_3, a jego wysokość jest dowolna. Wewnątrz prostokąta znajduje się mediana. W liczniku wzoru (2.47) mamy różnicę prawej i lewej części prostokąta, na jakie dzieli go mediana. Przy prawostronnej asymetrii rozkładu różnica ta jest dodatnia, przy lewostronnej – ujemna. Przy rozkładzie symetrycznym jest ona równa zeru. Maksymalną wartość różnicy określa mianownik wzoru (2.47), czyli długość całego pudełka. Tak więc pozycyjny współczynnik asymetrii jest miarą unormowaną, przyjmującą wartości z przedziału [–1; +1].

Jak wynika z rys. 2.1, wyznaczony prostokąt uzupełniają dwa odcinki, zwane **wąsami**. Jeden z nich łączy prostokąt na poziomie Q_1 z minimalną wartością zmiennej, drugi zaś – łączy Q_3 z maksymalną wartością cechy. W ten sposób wykres pozwala na ocenę zróżnicowania wartości cechy.

2.2.4. Miary spłaszczenia i koncentracji

Zbiorowość statystyczną analizuje się także ze względu na stopień skupienia poszczególnych wartości cechy wokół średniej arytmetycznej. Skupienie to jest – w dużym stopniu – uzależnione od poziomu dyspersji. Im większe jest zróżnicowanie, tym mniejsze skupienie i odwrotnie. Miarą skupienia poszczególnych wartości cechy wokół jej średniej arytmetycznej jest **współczynnik skupienia (kurtoza)**.

Kurtoza

Współczynnik skupienia (kurtoza) jest standaryzowanym momentem centralnym czwartego rzędu, czyli:

$$K = \frac{m_4}{s^4}, \tag{2.48}$$

gdzie: m_4 jest momentem centralnym czwartego rzędu, określającym przeciętną wielkość czwartych potęg odchyleń wartości cechy od średniej arytmetycznej. Tak więc:

$$m_4 = \frac{1}{N}\sum_{i=1}^{N}(x_i - \bar{x})^4 \text{ dla szeregu wyliczającego,} \tag{2.49}$$

$$m_4 = \frac{1}{N}\sum_{i=1}^{k}(x_i - \bar{x})^4 n_i \text{ dla szeregu rozdzielczego punktowego,} \tag{2.50}$$

$$m_4 = \frac{1}{N}\sum_{i=1}^{k}\left(\overset{\circ}{x}_i - \bar{x}\right)^4 n_i \text{ dla szeregu rozdzielczego przedziałowego.} \tag{2.51}$$

Im wyższa wartość współczynnika skupienia, tym krzywa liczebności jest bardziej wysmukła. Oznacza to większe skupienie wartości cechy wokół średniej. Małe wartości współczynnika skupienia wskazują na spłaszczenie rozkładu, a więc mniejsze skupienie wartości cechy wokół średniej arytmetycznej. Przyjmuje się, że jeśli zbiorowość ma rozkład normalny (por. rozdział III), to $K = 3$. Jeśli natomiast $K < 3$, to rozkład jest bardziej spłaszczony niż normalny. Taki rozkład nosi nazwę **platokurtycznego**. W przypadku, gdy $K > 3$, rozkład empiryczny badanej cechy jest bardziej wysmukły, a skupienie jest silniejsze od normalnego. Mówimy wówczas o rozkładach **leptokurtycznych**.

Współczynnik skupienia podaje się zazwyczaj w postaci:

$$K' = \frac{m_4}{s^4} - 3 \tag{2.52}$$

i określa mianem **współczynnika ekscesu**. Współczynnik ekscesu jest równy zeru, jeżeli rozkład badanej cechy jest normalny. W rozkładach spłaszczonych (plaktokurtycznych) $K' < 0$, zaś w rozkładach wysmukłych (leptokurtycznych) – $K' > 0$.

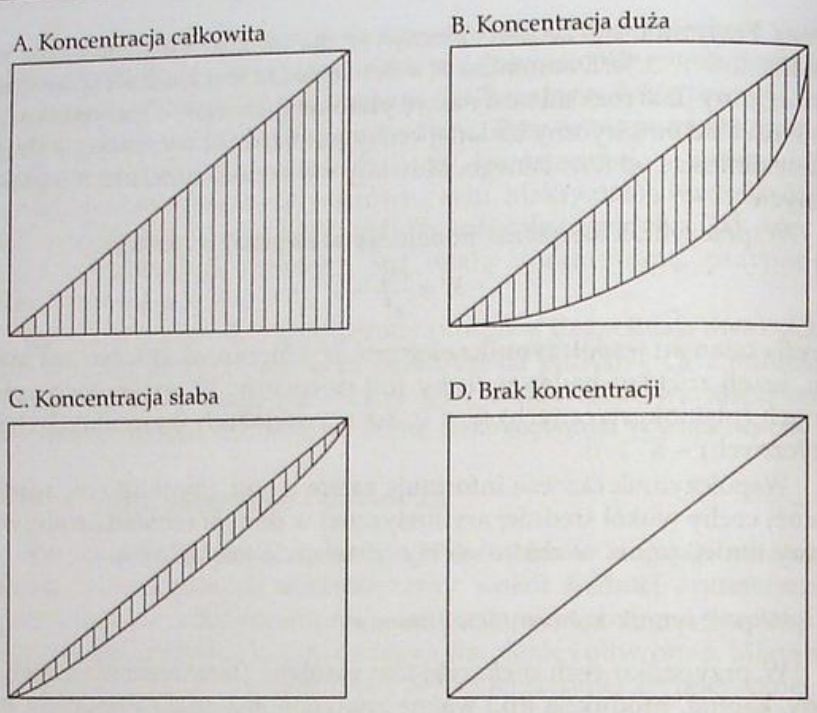
Współczynnik ekscesu informuje zatem o tym, czy skupienie wartości badanej cechy wokół średniej arytmetycznej w danym rozkładzie jest większe czy mniejsze niż w zbiorowości o rozkładzie normalnym.

Współczynnik koncentracji Lorenza

W przypadku cech o charakterze zasobów (powierzchnia ziemi, dochody, kapitał, produkcja itp.) ważne znaczenie ma analiza rozkładu ogólnej sumy wartości badanej cechy (łącznego funduszu cechy) pomiędzy poszczególne jednostki zbiorowości statystycznej. Mówimy wówczas o **koncentracji** badanego zjawiska. Koncentracja jest bezpośrednio związana z asymetrią i dyspersją badanej cechy. Im silniejsza asymetria i większe zróżnicowanie wartości zmiennej -- tym koncentracja jest większa.

Zupełna (całkowita) koncentracja występuje wtedy, gdy łączny fundusz cechy przypada na jedną jednostkę zbiorowości (np. łączny areał powierzchni ziemi w województwie pozostaje w posiadaniu jednego gospodarstwa rolnego). Z **brakiem koncentracji** mamy do czynienia wówczas, gdy na każdą jednostkę zbiorowości przypada taka sama część ogólnej sumy wartości cechy (np. każdy pracownik w przedsiębiorstwie otrzymuje taką samą część łącznego funduszu płac). W badaniach statystycznych zjawiska braku koncentracji i koncentracji zupełnej raczej nie występują. Najczęściej mamy do czynienia z różnym natężeniem koncentracji (por. rys. 2.2).

Do oceny stopnia koncentracji stosuje się dwie metody: **graficzną** i **analityczną**. Metoda graficzna polega na wykreśleniu **wieloboku koncentracji Lorenza**. W tym celu na osi odciętych odmierza się skumulowane częstości względne liczebności (w %), natomiast na osi rzędnych – procentowe skumulowane częstości względne łącznego funduszu cechy. Łącząc punkty o tych współrzędnych otrzymujemy **krzywą koncentracji** (nazywaną też **krzywą Lorenza**). W przypadku nierównomiernego rozdziału łącznego funduszu cechy pomiędzy jednostki zbiorowości wszystkie punkty leżałyby na



Rys. 2.2 Różne przypadki koncentracji
Źródło: A. Zeliaś, *Metody statystyczne*, PWE, Warszawa 2000, s. 73.

przekątnej kwadratu o boku 100. Przekątna tego kwadratu nosi nazwę **linii równomiernego rozdziału**. Powierzchnia zawarta między linią równomiernego rozdziału a krzywą koncentracji Lorenza jest **powierzchnią koncentracji**. Im większy jest stopień koncentracji, tym bardziej krzywa Lorenza odchyla się od linii równomiernego rozdziału, a tym samym większa jest powierzchnia koncentracji. Maksymalna wartość powierzchni koncentracji pozostaje równa połowie kwadratu, tj. 5000, gdyż dwa boki prostokątnego trójkąta równoramiennego mają długość 100, stąd jego pole jest równe 5000.

Stosunek pola zawartego między linią równomiernego rozdziału a krzywą koncentracji do pola połowy kwadratu (pola trójkąta) nosi nazwę **współczynnika koncentracji Lorenza**. Współczynnik ten ma następującą postać:

$$k = \frac{a}{5000}. \qquad (2.53)$$

Współczynnik (2.53) jest miarą niemianowaną, przyjmującą wartości liczbowe z przedziału: $0 \le k \le 1$. Przy braku koncentracji $k = 0$, natomiast przy $k = 1$ występuje koncentracja zupełna (całkowita).

Praktyczny sposób postępowania przy pomiarze siły koncentracji metodą graficzną i analityczną zaprezentujemy na przykładzie (por. tab. 2.4).

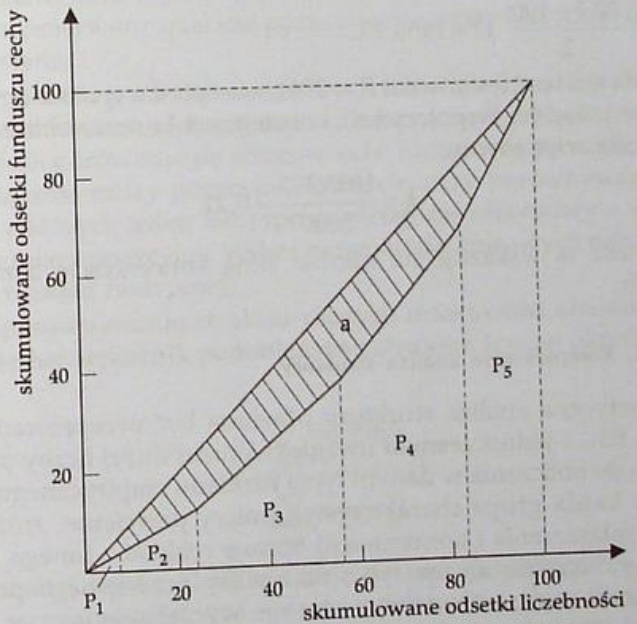
Tab. 2.4. Liczba ludności w gminach w Polsce w 1977 r.

Gminy o liczbie ludności (w tys.)	Liczba gmin	Łączna liczba ludności [a] (w tys.)	Częstości względne		Skumulowane częstości względne	
			liczby gmin	łącznej liczby gmin	liczby gmin	łącznej liczby gmin
powyżej 2	15	23,4	0,7	0,2	0,7	0,2
2–5	490	1972,5	23,7	13,3	24,4	13,5
5–7	663	3951,3	32,0	26,6	56,4	40,1
7–10	551	4551,0	26,6	30,6	83,0	70,7
powyżej 10	351	4364,3	17,0	29,3	100,0	100,0

[a] Łączną liczbę ludności w poszczególnych przedziałach podano na podstawie danych indywidualnych, gdyż umowne domknięcie przedziałów i korzystanie ze środków klas prowadziłoby do dużych błędów.

Źródło: Rocznik Demograficzny 1978, s. 7–8.

Już na podstawie pobieżnej analizy danych zawartych w tab. 2.4 można stwierdzić, że występuje koncentracja liczby ludności w gminach. Odczytując parami odsetki liczby gmin i odsetki łącznej liczby ludności, zauważamy dość znaczne różnice. Świadczy to o występowaniu koncentracji. Potwierdza to również wykres wieloboku koncentracji (rys. 2.3).



Rys. 2.3. Wielobok koncentracji
Źródło: M. Sobczyk, *Statystyka. Podstawy teoretyczne, przykłady i zadania*, UMCS, Lublin 2000, s. 68.

Do obliczenia współczynnika koncentracji za pomocą wzoru (2.53) niezbędna jest znajomość powierzchni a. W tym celu należy uprzednio ustalić powierzchnię znajdującą się poniżej krzywej koncentracji Lorenza. Dokładne ustalenie tej powierzchni wymaga znajomości analitycznej postaci funkcji krzywej koncentracji. W badaniach empirycznych powierzchnię tę znajduje się w sposób przybliżony jako sumę pól trójkąta i trapezów (por. rys. 2.3).

Powierzchnię pod krzywą koncentracji można zatem obliczyć jako:

$$P = P_1 + \sum_{i=1}^{k} P_i, \tag{2.54}$$

gdzie: P_1 – pole trójkąta, P_i – pole i-tego trapezu ($i = 2, 3, ..., k$).

Korzystając ze znanych wzorów na pole trójkąta i trapezu, otrzymujemy:

$$P_1 = \frac{0{,}2 \cdot 0{,}7}{2} = 0{,}070,$$

$$P_2 = \frac{0{,}2 + 13{,}5}{2} \cdot 23{,}7 = 162{,}345,$$

$$P_3 = \frac{13{,}5 + 40{,}1}{2} \cdot 32 = 857{,}6,$$

$$P_4 = \frac{40{,}1 + 70{,}7}{2} \cdot 26{,}6 = 1473{,}64,$$

$$P_5 = \frac{70{,}7 + 100}{2} \cdot 17 = 1450{,}95.$$

Suma tych pól jest równa: $P = 3944{,}605$. Pole a wynosi zatem: $a = 5000 - 3944{,}605 = 1055{,}395$. Współczynnik koncentracji Lorenza obliczony ze wzoru (2.53), jest więc równy:

$$k = \frac{1055{,}395}{5000} = 0{,}21.$$

Wartość ta wskazuje na niezbyt silną koncentrację liczby ludności w gminach.

2.2.5. Kompleksowa analiza struktury

Statystyczna analiza struktury powinna być przeprowadzona **kompleksowo**, tzn. z jednoczesnym uwzględnieniem dużej liczby parametrów, możliwych do obliczenia w danym typie rozkładu empirycznego. Wynika to z faktu, że każda grupa charakterystyk (miary przeciętne, zróżnicowania, asymetrii, spłaszczenia i koncentracji) opisuje rozkład z innego punktu widzenia. Ograniczenie się np. tylko do analizy przeciętnego poziomu czy zróżnicowania wartości badanej cechy nie wyczerpuje jeszcze wszystkich właściwości rozkładu empirycznego. Mogą bowiem występować rozkłady np. o identycznej średniej arytmetycznej i takim samym odchyleniu standardowym, ale różniące się asymetrią.

Jak wiadomo, do opisu struktury zbiorowości wykorzystuje się dwie grupy miar: **klasyczne i pozycyjne**. Miary klasyczne oparte są na wszystkich wartościach badanej cechy. Jest to zarówno ich wada, jak i zaleta. Wada tych miar ujawnia się w przypadku występowania obserwacji **skrajnych** (nietypowych), odbiegających znacznie od pozostałych. Przykładowo, wielkości ekstremalne zniekształcają poziom średniej arytmetycznej, a tym samym i innych miar klasycznych, do obliczania których wykorzystuje się tę średnią (np. odchylenie standardowe, współczynnik asymetrii). Z tego też względu nie należy wówczas stosować miar klasycznych. Tego rodzaju miar nie należy również stosować w rozkładach skrajnie asymetrycznych, mających więcej niż jedno maksimum, jak też w szeregach rozdzielczych o otwartych przedziałach klasowych. Ogólnie można stwierdzić, że miary klasyczne stosowane są przy opisie struktur **rozkładów typowych** (jednomodalnych, o słabej asymetrii lub symetrycznych).

Miary pozycyjne są mniej dokładne, gdyż oparte są na wybranych wartościach cechy, stojących na określonej pozycji. Dlatego też są one wykorzystywane do opisu struktur **rozkładów nietypowych** (skrajnie asymetrycznych, bi- lub wielomodalnych o otwartych przedziałach klasowych). Obliczanie miar klasycznych z rozkładów nietypowych nie jest wskazane.

Tak więc w opisie struktur rozkładów typowych pierwszeństwo mają miary klasyczne. Miary pozycyjne można obliczać dodatkowo. Rozkłady nietypowe są opisywane miarami pozycyjnymi; miary klasyczne nie są tutaj wykorzystywane.

W przypadku porównywania dwóch lub więcej struktur zbiorowości o typowych rozkładach wykorzystujemy miary klasyczne oraz, dodatkowo, pozycyjne. Jeśli porównuje się zbiorowości o nietypowych rozkładach – stosujemy **wyłącznie** miary pozycyjne. Wreszcie, przy porównywaniu dwóch rozkładów, z których jeden jest typowy, drugi zaś nietypowy – wykorzystujemy tylko miary pozycyjne. Wybór parametrów opisowych narzucony jest tutaj przez rozkład nietypowy.

Przy porównywaniu struktur różnych zbiorowości znajduje zastosowanie **względny wskaźnik podobieństwa struktur**. Jest on określony następująco:

$$Z = \frac{\sum_{i=1}^{k} \min(w_i)}{\sum_{i=1}^{k} \max(w_i)}. \tag{2.55}$$

W liczniku wzoru (2.55) znajduje się suma mniejszych wskaźników struktury, w mianowniku zaś – suma większych wskaźników struktury porównywanych rozkładów. Wskaźnik Z przyjmuje wartości z przedziału [0;1], przy czym jeśli $Z = 1$, to porównywane struktury są identyczne. W przypadku, gdy $Z = 0$, porównywane struktury są krańcowo odmienne.

W celu zilustrowania sposobu obliczania względnego wskaźnika podobieństwa struktur posłużymy się poniższym przykładem (tab. 2.5).

Tab. 2.5. Pracownicy zakładów A i B według wieku

Wiek w latach	Zakład A		Zakład B	
	liczebności bezwzględne	wskaźniki struktury (w %)	liczebności bezwzględne	wskaźniki struktury (w %)
18–24	204	14,64	274	14,66
25–34	264	18,95	355	18,99
35–44	356	25,56	472	25,25
45–54	320	22,97	429	22,95
55–59	201	14,43	271	14,50
60–65	48	3,45	68	3,65
Razem	1393	100,00	1869	100,00

Źródło: dane umowne.

Wskaźniki struktury (zwane też częstościami, liczebnościami względnymi, frakcjami lub odsetkami) są ilorazami liczby jednostek o danej wartości cechy do liczebności całej zbiorowości statystycznej (N). Wskaźniki struktury mogą być wyrażane w ułamku, w procentach lub w promilach, a mianowicie:

$$\left.\begin{aligned} &w_i = \frac{n_i}{N}, \quad 0 \le w_i \le 1, \\ &w_i = \frac{n_i}{N}\cdot 100, \quad 0 \le w_i \le 100, \\ &w_i = \frac{n_i}{N}\cdot 1000, \quad 0 \le w_i \le 1000, \\ &i = 1,2,\ldots,k; \quad \sum_{i=1}^{k} w_i = 1 \quad (100 \text{ lub } 1000), \\ &\sum_{i=1}^{k} n_i = N. \end{aligned}\right\} \tag{2.56}$$

Obliczone na podstawie danych liczbowych zawartych w tab. 2.5 wskaźniki struktury wieku pracowników zakładu A są równe:

$$\frac{204}{1393}\cdot 100 = 14{,}64\% \quad \frac{264}{1393}\cdot 100 = 18{,}95\%,$$
$$\frac{356}{1393}\cdot 100 = 25{,}56\% \quad \frac{320}{1393}\cdot 100 = 22{,}97\%,$$
$$\frac{201}{1393}\cdot 100 = 14{,}43\% \quad \frac{48}{1393}\cdot 100 = 3{,}45\%.$$

W analogiczny sposób obliczono wskaźniki struktury dla zbiorowości pracowników zakładu B.

Wykorzystując dane liczbowe zawarte w tab. 2.5, względny wskaźnik podobieństwa struktur obliczamy następująco:

$$Z = \frac{14{,}64 + 18{,}95 + 25{,}25 + 22{,}95 + 14{,}43 + 3{,}45}{14{,}66 + 18{,}99 + 25{,}56 + 22{,}97 + 14{,}50 + 3{,}65} = \frac{99{,}67}{100{,}33} = 0{,}993.$$

Otrzymany wynik oznacza, że między strukturą wieku pracowników zakładów A i B istnieje duże podobieństwo.

ZADANIA

2.1. W pewnym mieście badano zbiorowość sklepów spożywczych ze względu na powierzchnię użytkową w m^2. Otrzymano następujące wyniki:

70	62	65	68	71	75
67	78	72	74	91	80
81	67	74	96	77	54
83	52	64	68	85	99
57	52	56	70	63	61

Utworzyć szereg rozdzielczy przedziałowy o interwale równym 8 m^2. Obliczyć – za pomocą średniej arytmetycznej – przeciętną powierzchnię sklepu spożywczego.

2.2. Poniższy szereg rozdzielczy skumulowany przedstawia powierzchnię (w m^2) 70 magazynów:

Powierzchnia magazynu	Skumulowana liczebność magazynów
poniżej 50	8
poniżej 80	20
poniżej 110	45
poniżej 140	65
poniżej 170	70

Obliczyć średnią arytmetyczną powierzchni magazynu, jeżeli najmniejszy jest magazyn o powierzchni 20 m^2.

2.3. U 18 robotników nie zaobserwowano żadnej usterki w wyprodukowanych przez nich wyrobach, u 28 co najwyżej jedną usterkę, u 33 co najwyżej dwie usterki, u 37 co najwyżej 3 usterki, u 38 co najwyżej cztery usterki, a u 40 robotników zaobserwowano co najwyżej pięć usterek. Jaka jest przeciętna liczba usterek w wyprodukowanych wyrobach w badanej zbiorowości robotników?

2.4. W firmie budowlanej zatrudniającej 80 osób średnia płaca brutto w maju wynosiła 2326,50 zł. Jaka była wysokość funduszu płac w tym miesiącu?

2.5. Oblicz średnią arytmetyczną liczby nieobecności na zajęciach z matematyki w semestrze zimowym w 25-osobowej grupie studenckiej mając następujące informacje:

Liczba studentów	Liczba nieobecności w semestrze zimowym
7	0
10	1
4	2
1	3
1	4
2	5

2.6. Rozkład spółek akcyjnych wg wysokości dywidendy wypłaconej akcjonariuszom przedstawia się następująco:

Udział spółek (w %)	Dywidenda w zł
7,8	1–2
20,0	2–3
33,3	3–4
?	4–5
10,0	5–6
6,7	6–7
5,5	7–8

2.7. Struktura zatrudnienia według wieku w spółce X w pewnym miesiącu przedstawiała się następująco:

Wiek pracowników (w latach)	poniżej 20	21–30	31–40	41–50	51–60	powyżej 60
Udział pracowników (w %)	3	23	27	31	10	6

Oblicz średnią arytmetyczną wieku zatrudnionych w tej spółce. Z ilu osób składa się cała załoga spółki, jeśli wiadomo, że 36 osób przekroczyło 60 rok życia? Za dolną granicę pierwszego przedziału przyjąć 18 lat, natomiast za górną granicę ostatniej klasy 70 lat.

2.8. W czteroosobowej rodzinie średnia miesięczna płaca wynosi 850 zł. Jakie wynagrodzenie otrzymuje córka, jeśli wynagrodzenie pozostałych członków rodziny wynosi: ojca – 900 zł, matki – 750 zł, syna – 720 zł?

2.9. Za wzorowe wykonanie zadania sześciu studentów otrzymało premię w wysokościach równych odpowiednio (w zł): 65, 84, 91, 53, 48 oraz 79. Oblicz średnią arytmetyczną wypłaconych premii. Na tym przykładzie sprawdź następujące własności średniej arytmetycznej:

1) suma odchyleń poszczególnych wartości cechy od jej średniej arytmetycznej wynosi zero,

2) suma kwadratów odchyleń poszczególnych wartości cechy od jej średniej arytmetycznej stanowi minimum.

2.10. Pięciu pracowników sklepu otrzymało premię za ubiegły miesiąc. Premia pierwszego pracownika była niższa od średniej premii w grupie o 23 zł, drugiego niższa od średniej o 52 zł, premie trzeciego i czwartego pracownika były wyższe do średniej grupowej premii odpowiednio o 98 zł oraz 42 zł. Czy premia piątego pracownika była wyższa czy niższa od średniej premii i o ile?

2.11. W pewnej grupie uczniów średnia wieku wynosi 11 lat. Najstarszy z uczniów ma 17 lat, a średnia wieku pozostałych wynosi 10 lat. Ilu uczniów liczy ta grupa?

2.12. W pewnym zakładzie zbadano staż pracy pracowników bezpośrednio produkcyjnych. Okazało się, że 25% spośród nich pracowało krócej niż 6 lat, połowa od 6 do 12 lat, a najdłuższy staż pracy pozostałych pracowników był równy 18 lat. Średni staż pracy pracowników administracyjno-biurowych w tym zakładzie wynosił 12 lat. Jaki jest średni staż pracy ogółu pracowników tego zakładu, jeśli grupa pracowników bezpośrednio produkcyjnych jest 2,5-krotnie liczniejsza od administracyjno-biurowych?

2.13. W dwóch działach firmy pracowało po 2 osoby. W pierwszym dziale miesięczne zarobki brutto każdej z osób wynosiły odpowiednio 1400 zł i 1600 zł, a w drugim dziale – 2000 zł i 4000 zł. Pracownicy obydwu działów domagają się wzrostu średnich zarobków w obydwu komórkach. Jak to zrobić, nie zwalniając nikogo i nie wydając ani grosza na podwyżki?

2.14. Średni miesięczny zarobek 25 pracowników w pewnej spółce wynosił 2000 zł, gdyż 20 osób zarabiało po 1400 zł, 4 osoby po 3000 zł, a 1 osoba 10 000 zł. Dwudziestu najniżej zarabiających pracowników zażądało podwyżki płac do poziomu średniej płacy w spółce. Żądanie ich zostało spełnione. Czy po podwyżce zarabiają oni poniżej początkowej średniej?

2.15. W pewnej grupie studenckiej liczącej 30 osób średnia ocena uzyskana z egzaminu ze statystyki wynosiła 3,8. Druga i trzecia grupa otrzymały z tego egzaminu oceny dające średnie równe odpowiednio: 4,2 oraz 4,6. Jaka była średnia ocena z egzaminu we wszystkich grupach łącznie, jeśli w grupie drugiej było 28 studentów, a w trzeciej 26?

2.16. W czterech bankach zbadano czas obsługi klientów. Średnie arytmetyczne czasu obsługi w trzech bankach były równe odpowiednio: $\bar{x}_1 = 10$ minut, $\bar{x}_2 = 12$ minut, $\bar{x}_3 = 8$ minut. Rozkład czasu obsługi w czwartym banku przedstawiał się następująco:

Czas obsługi w minutach	1–3	3–5	5–7
Liczba klientów	20	20	60

Oblicz średnią arytmetyczną czasu obsługi we wszystkich bankach łącznie, jeśli stosunek liczby klientów w I, II, III i IV banku wynosił odpowiednio: 2:4:3:1.

2.17. Obliczając średnią arytmetyczną stażu pracy pewnej grupy robotników otrzymano następujące rezultaty: $x_4 n_4 = 210$, $x_5 n_5 = 225$, $x_6 n_6 = 110$, $N = 100$, $\bar{x} = 6{,}8$ lat. Wyznaczyć pozostałe iloczyny cząstkowe niezbędne do wyznaczenia średniej arytmetycznej stażu pracy, jeśli wiadomo, że spełniają one proporcję: 1:6:20.

2.18. W firmie pracuje 25 osób. Cztery z nich zarabiają nie więcej niż 1600 zł, osiem zarabia nie więcej niż 2000 zł, piętnaście osób otrzymuje nie więcej niż 2400 zł oraz dwadzieścia jeden osób zarabia nie więcej niż 2800 zł. Pozostałe osoby stanowią kierownictwo firmy, ale żadna z nich nie zarabia więcej niż 4200 zł. Oblicz miesięczny fundusz płac w tej firmie oraz wysokość przeciętnej miesięcznej płacy.

2.19. Oblicz średnią powierzchnię indywidualnych gospodarstw rolnych w województwie L, jeśli rozkład tej cechy przedstawia się następująco:

Powierzchnia gospodarstwa w ha	Skumulowany odsetek gospodarstw
poniżej 2	42,5
poniżej 5	63,2
poniżej 7	76,4
poniżej 10	89,3
poniżej 15	97,2
poniżej 20	99,0

Za gospodarstwo rolne przyjęto obszar użytków rolnych o powierzchni co najmniej 1 ha. Największe obszarowo gospodarstwo rolne w województwie L miało 30 ha.

2.20. W ostatnich 10 meczach koszykarz zdobywał średnio po 22 punkty. Po jedenastym spotkaniu jego średnia liczba punktów uzyskiwanych w jednym meczu wzrosła do 24. Ile punktów zdobył koszykarz w jedenastym meczu?

2.21. Średnia arytmetyczna sześciu liczb jest równa 4,5. O ile zmieni się średnia, jeśli:

a) do sumy sześciu liczb dodamy 8,4,

b) do każdej z sześciu liczb dodamy 8,4?

2.22. Pewien inwestor kupił za 5000 zł dwa rodzaje akcji: A i B. Na akcjach A zarobił 24% kwoty, którą w nie zainwestował, zaś na akcjach B stracił 16% kwoty, którą na nie wydał. W rezultacie inwestor zyskał 800 zł. Jaka była średnia cena akcji, jeśli inwestor nabył 200 akcji A i 50 akcji B?

2.23. W hurtowni są dwa rodzaje cukierków czekoladowych po 15 zł za 1 kg i po 20 zł za 1 kg. Ile kilogramów cukierków każdego gatunku należy wymieszać, aby otrzymać 40 kg mieszanki po 18 zł za 1 kg?

2.24. Za pomocą średniej arytmetycznej oceń przeciętną wielkość gospodarstwa, mając następujący rozkład:

Powierzchnia gospodarstwa (w ha)	Liczba gospodarstw
poniżej 2	2
2–4	4
4–7	6
7–10	5
10 i więcej	3

Dodatkowo wiadomo, że średnia powierzchnia dwóch gospodarstw liczących poniżej 2 ha wynosi 1,5 ha, a gospodarstwa liczące 10 ha i więcej zajmują razem powierzchnię 51 ha.

2.25. W pewnej spółce przeciętna płaca pracowników administracyjno-biurowych wynosi 2100 zł, a pozostałych pracowników (bez zarządu) 1820 zł. Jaka jest przeciętna płaca wszystkich pracowników w tej spółce, jeśli pracownicy administracyjno-biurowi stanowią 8% załogi?

2.26. Motocyklista przez pierwsze dwie godziny podróży jechał z przeciętną prędkością 70 km/h, a przez trzy pozostałe godziny z przeciętną prędkością 60 km/h. Jaka była średnia prędkość motocyklisty podczas tej podróży?

2.27. Gdyby średnią prędkość samochodu zwiększyć o 9 km/h, to przebyłby on drogę 180 km w czasie o 40 minut krótszym. Jaka jest średnia prędkość samochodu?

2.28. Połowę drogi z miasta A do miasta B rowerzysta przebył z prędkością 40 km/h, a drugą połowę z prędkością 30 km/h. Oblicz średnią prędkość rowerzysty.

2.29. Pociąg przejechał 2 km w ciągu 3 minut, a samochód 3 km w ciągu 2 minut. Ile razy prędkość samochodu była większa od prędkości pociągu?

2.30. Oblicz przeciętną cenę targowiskową towaru X zaobserwowaną w czterech punktach miasta, jeśli dysponujemy następującymi informacjami:

Punkty	A	B	C	D
Cena w zł/kg	8	9	10	11
Utarg w zł	360	450	300	220

2.31. Makroregion składa się z pięciu województw: A, B, C, D i E. W województwie A zamieszkuje 700 tys. osób, w województwie B – 520 tys. osób, w województwie C – 490 tys. osób, w województwie D – 680 tys. osób, a w województwie E – 580 tys. osób. Gęstość zaludnienia (tj. liczba ludności przypadająca na 1 km^2) w poszczególnych województwach jest równa odpowiednio: 70, 60, 50, 68 oraz 63 osób/km^2. Obliczyć średnią gęstość zaludnienia w tym makroregionie.

2.32. Za 120 zł zakupiono towar, którego cena wynosiła 3 zł za sztukę, a za 200 zł towar, którego 1 sztuka kosztuje 5 zł. Oblicz średnią cenę zakupionych towarów.

2.33. Czterech spośród dziesięciu robotników potrzebuje na wykonanie jednego wyrobu 0,4 godziny, trzech 15 minut, a pozostałych trzech 1/3 godziny. Jaki jest przeciętny czas potrzebny na wykonanie jednego wyrobu?

2.34. Na targowisku ceny 1 kg gruszek kształtowały się następująco: I gatunek – 1,80 zł, II gatunek – 1,40 zł, III gatunek – 1 zł. Sprzedawca uzyskał ze sprzedaży gruszek w ciągu dnia 56 zł, z tego 18 zł za I gatunek oraz 24 zł za III gatunek. Oblicz średnią cenę sprzedaży 1 kg gruszek.

2.35. Czwartą część drogi rowerzysta jechał z prędkością 20 km/h, połowę drogi z prędkością 24 km/h, a pozostałą część drogi z prędkością 12 km/h. Oblicz przeciętną prędkość rowerzysty na całej trasie.

2.36. W mieście Z zamieszkuje 50 tys. osób, a w mieście W – 70 tys. osób. Gęstość zaludnienia w tych miastach jest równa odpowiednio: 1500 osób/km^2 oraz 1000 osób/km^2. Oblicz łączną gęstość zaludnienia w obu miastach.

2.37. Dla trzech gmin woj. lubelskiego zamieszkałych przez odpowiednio 15 tys., 45 tys. oraz 30 tys. mieszkańców, gęstość zaludnienia kształtuje się następująco: 30 osób/km^2, 67,5 osoby/km^2 oraz 60 osób/km^2. Oblicz średnią gęstość zaludnienia w trzech gminach łącznie oraz ogólną powierzchnię zajmowaną przez te gminy.

2.38. Wydajność pracy grupy robotników (mierzona w szt/h) kształtowała się następująco:

18	15	14	13	17	19	17	20	17	17
12	18	15	16	17	17	17	17	16	14
15	16	16	12	19	20	19	12	20	18

Utworzyć szereg rozdzielczy punktowy i szereg rozdzielczy przedziałowy o interwale równym 2. Obliczyć i zinterpretować średnią arytmetyczną, dominantę i kwartyle. Wykorzystać obydwa rodzaje szeregów.

2.39. Strukturę rodzin według liczby jej członków charakteryzuje poniższy rozkład:

Odsetek rodzin	15	30	20	15	10	5	5
Liczba członków rodziny	2	3	4	5	6	7	8

Za pomocą miar przeciętnych scharakteryzować ten rozkład. Wyniki zinterpretować.

2.40. Ile wynosi dominanta i mediana wagi 10 uczniów klasy I, dla których dane indywidualne (w kg) są następujące: 19, 21, 19, 23, 24, 25, 23, 22, 23 i 20?

2.41. W pewnym mieście dokonano klasyfikacji urzędów pocztowych pod względem liczby zatrudnionych. Okazało się, że 6 urzędów zatrudnia od 2 do 5 osób, 8 urzędów od 5 do 8 osób, a 10 urzędów od 8 do 11 osób. Oblicz i zinterpretuj miary tendencji centralnej.

2.42. Koszty handlowe w 200 firmach kształtowały się następująco:

Liczba firm	10	20	30	110	30
Koszty w tys. zł	1–3	3–5	5–7	7–9	9–11

Wyznaczyć i zinterpretować średnią arytmetyczną, modalną i kwartyle.

2.43. W wyniku pomiaru wzrostu grupy studentów otrzymano nieco asymetryczny rozkład wzrostu, dla którego wartość środkowa była tylko o 2% wyższa od średniej arytmetycznej wzrostu równej 171 cm. Jaki wzrost dominował w badanej grupie?

2.44. Rozkład kosztów produkcji pewnego wyrobu charakteryzuje umiarkowana asymetria. Najczęściej spotykany koszt jednostkowy wynosi 28,2 zł, a koszt połowy wyrobów był wyższy od 44 zł. Jaka jest wartość średniej arytmetycznej kosztu jednostkowego?

2.45. Wyznaczyć miary tendencji centralnej, charakteryzujące ceny artykułu X o następującej strukturze sprzedaży:

Cena w zł	10	15	20
Wartość sprzedaży w zł	150	315	400

2.46. Lekkoatletka wykonała serię skoków w dal. Najczęściej, bo 15 razy, uzyskiwała wyniki w granicach od 7 m do 7,5 m. Osiem skoków lekkoatletki miało długość z przedziału 6,5–7 m. Dominanta skoków w dal była równa 7,35 m. Ile skoków o długości 7,5–8 m wykonała lekkoatletka?

2.47. W pewnej spółdzielni mieszkaniowej dominują mieszkania o powierzchni 48–54 m^2. Najwięcej, bo 30% mieszkań, ma powierzchnię równą 50 m^2, a 20% mieszkań ma powierzchnię w granicach 42–48 m^2. Jaki odsetek mieszkań w tej spółdzielni zajmuje powierzchnię od 54 do 60 m^2?

2.48. Badając wydajność pracy przy produkcji pewnego wyrobu otrzymano następujące wyniki: $\bar{x} = Me = D = 11$ szt./godz. Jaki odsetek pracowników zatrudnionych przy produkcji tego wyrobu uzyskał wydajność od 8 do 10 szt./godz., jeśli największy odsetek (40%) pracowników uzyskał wydajność 10–12 szt./godz., a wydajność 12–14 szt./godz. osiągnęło 20% pracowników?

2.49. Mediana wieku zatrudnionych w pewnym przedsiębiorstwie zawarta jest w przedziale 40–50 lat i wynosi 44 lata. W przedziale tym mieści się 25 pracowników. W zbiorowości zatrudnionych w tym przedsiębiorstwie 40 pracowników liczy mniej niż 40 lat. Ilu pracowników jest zatrudnionych w tym przedsiębiorstwie?

2.50. Mediana zarobków 120-osobowej grupy pracowników znajdowała się w przedziale 1000–1500 zł, do którego należało 20 pracowników i wynosiła 1300 zł. Ilu pracowników w tym przedsiębiorstwie zarabiało mniej niż 1500 zł?

2.51. Miesięczne zużycie gazu (w m^3) przez 12 gospodarstw domowych wynosiło: 96, 135, 82, 116, 84, 101, 90, 68, 106, 118, 122, 130. Za pomocą klasycznego współczynnika zmienności ocenić zróżnicowanie zużycia gazu w gospodarstwach domowych.

2.52. Czas rozwiązywania pewnego zadania (w minutach) przez grupę 200 uczniów charakteryzuje poniższy rozkład:

Liczba uczniów	1	10	48	82	46	12	1
Czas rozwiązywania	0–2	2–4	4–6	6–8	8–10	10–12	12–14

Wyznacz granice liczbowe obszaru zmienności dla typowych jednostek badanej zbiorowości. Jaki odsetek uczniów uzyskał wyniki typowe?

2.53. Zbiorowość 500 małżeństw zbadano pod względem liczby dzieci (x) i czasu trwania małżeństwa (y). Okazało się, że średnia arytmetyczna dla kwadratów wartości cechy y wynosi 149, $\bar{x} = 2$, $\bar{y} = 10$, $s(x) = 0{,}5$. Pod względem której cechy: czasu trwania małżeństwa, czy liczby posiadanych dzieci, małżeństwa są bardziej zróżnicowane?

2.54. W jednym z domów akademickich przeprowadzono badanie dotyczące miesięcznych wydatków na cele kulturalne. Otrzymano następujące wyniki:

Wydatki miesięczne w zł	40–80	80–120	120–160	160–200
Odsetek studentów	10	30	40	20

Za pomocą klasycznych i pozycyjnych miar ocenić zróżnicowanie tej zbiorowości pod względem badanej cechy.

2.55. Lekkoatleta A uzyskał w skoku w dal następujące wyniki na zawodach w całym sezonie (w m): 6,82; 6,96; 7,23; 7,05; 7,80; 7,75. Lekkoatleta B, startujący na tych samych zawodach, uzyskał takie wyniki, że ich średnia arytmetyczna wyniosła 7,5 m, a suma ich kwadratów 450,2592 m^2. Który z lekkoatletów osiągnął regularniejsze wyniki?

2.56. Rozkład klubów sportowych w województwie K według liczby sekcji przedstawia się następująco:

Odsetek klubów	10	15	35	25	10	5
Liczba sekcji w klubie	1	2	3	4	5	6

Badanie klubów sportowych pod względem liczby sekcji w województwie P dało następujące wyniki: $\bar{x} = 2{,}6$ sekcji, $s = 1{,}1$ sekcji. Wyznaczyć typowe obszary zmienności dla badanej cechy. W którym województwie kluby są bardziej zróżnicowane pod względem badanej cechy?

2.57. W pierwszej grupie pracowników, liczącej 20 osób, średnia płaca netto wynosiła 1675 zł z odchyleniem standardowym 80 zł, a w drugiej, 25-osobowej gru-

pie, średnia płaca netto była równa 1800 zł przy typowym obszarze zmienności 1650 $< x_{typ} <$ 1950 zł. Oblicz typowy obszar zmienności płac w obu grupach łącznie.

2.58. W dwóch wydziałach zbadano wydajność pracy, której miernikiem był procent wykonania normy. W wydziale I średnia wydajność była równa 103%, przy wariancji 81 $(\%)^2$. W wydziale II wydajność pracy 10 pracowników kształtowała się następująco (w %): 102; 105; 109; 93; 91; 115; 131; 100; 88; 98. W którym wydziale zróżnicowanie wydajności pracy było mniejsze?

2.59. W pewnym mieście zbadano małe i średnie firmy pod względem zatrudnienia. Otrzymano następujący rozkład zatrudnienia (w osobach):

Zatrudnienie	0–5	5–10	10–15	15–20	20–25	25–30
Liczba firm	8	12	14	3	2	1

Obliczyć współczynnik asymetrii. Wynik zinterpretować.

2.60. Lekkoatleta wykonał serię skoków w dal. Co czwarty skok był krótszy od 7,2 m, a co czwarty dłuższy od 7,8 m. Wiadomo ponadto, że połowa wykonanych przez lekkoatletę skoków była nie większa niż 7,4 m. Obliczyć i zinterpretować współczynnik asymetrii.

2.61. Średnia premia w zakładzie wyniosła w I kwartale 250 zł, przy odchyleniu standardowym równym 50 zł. W IV kwartale każdy z pracowników otrzymał premię wyższą o 100 zł. Ile wyniesie średnia premia w IV kwartale, a ile odchylenie standardowe?

2.62. W dwóch grupach pracowników liczących po 100 osób zbadano miesięczne wydatki na artykuł X. Otrzymano następujące wyniki:

grupa I: $Me = 50$ zł, $V_Q = 24\%$, $Q_1 = 30$ zł;

grupa II: $Q_3 = 48$ zł, $V_Q = 25\%$, $Me = 45$ zł.

Porównać względną i bezwzględną dyspersję wydatków w obu grupach oraz siłę i kierunek asymetrii.

2.63. W dwóch hurtowniach przeprowadzono badanie pracowników pod względem dotychczasowego stażu pracy. Otrzymano następujące wyniki:

hurtownia I: $\bar{x} = 14$ lat, $V_s = 20\%$;

hurtownia II: $\bar{x} = 10$ lat, $V_s = 25\%$.

Oblicz współczynnik zmienności dla całej zbiorowości robotników, jeśli w hurtowni I było zatrudnionych 120 osób, a w hurtowni II – 80 osób.

2.64. Stu pracowników pewnego przedsiębiorstwa (70 mężczyzn i 30 kobiet) zbadano pod względem wieku i otrzymano następujące informacje:

mężczyźni: $\bar{x} = 40$ lat, $D = 35$ lat, $A_s = +0{,}5$;

kobiety: $\bar{x} = 30$ lat, $D = 33$ lata, $A_s = -0{,}5$.

Oblicz współczynnik zmienności i typowy obszar zmienności dla całej zbiorowości 100 pracowników.

2.65. W dwustu sklepach spożywczych przeprowadzono badanie dotyczące kosztów miesięcznego zużycia paliwa silnikowego. Otrzymano następujący rozkład:

Koszty handlowe	1–3	3–5	5–7	7–9	9–11
Liczba sklepów	10	20	30	110	30

Z analogicznego badania kosztów handlowych przeprowadzonego wśród sklepów handlujących artykułami gospodarstwa domowego otrzymano następujące

syntetyczne charakterystyki: $\bar{x} = 5$ tys. zł, $V_s = 20\%$, $A_s = +0{,}3$. Dokonać wszechstronnej analizy porównawczej badanych grup sklepów pod względem wysokości kosztów handlowych.

2.66. W pewnej bibliotece zbadano zbiorowość 120 losowo wybranych czytelników pod względem liczby przeczytanych w ciągu miesiąca książek. Strukturę czytelników według tej cechy ilustruje poniższy rozkład:

Liczba czytelników	44	42	22	7	2	1
Liczba przeczytanych książek	1	2	3	4	5	6

Zbadać siłę i kierunek asymetrii rozkładu.

2.67. Wykorzystując odpowiedni miernik, zbadać asymetrię płac w pewnym przedsiębiorstwie wiedząc, że płaca środkowa wynosi 1080 zł, 25% najniższych płac nie przekracza 800 zł, a 25% najwyższych płac przekracza 1800 zł.

2.68. Na stacji meteorologicznej mierzono – o godz. 12^{00} każdego dnia kwietnia – temperaturę powietrza. Otrzymano następujący rozkład:

Temperatura w °C	0	2	5	8	12	15	17	20	22	24	27
Liczba dni	2	3	3	4	5	6	2	2	1	1	1

Jaki odsetek dni kwietnia miał nietypowe temperatury?

2.69. Analiza wydajności pracy, liczonej w % wykonania normy, dla dwóch wydziałów produkcyjnych dostarczyła następujących informacji:

wydział I: $\bar{x} = D = Me$, $V_s = 20\%$, $s = 22\%$,

wydział II: $\bar{x} = 120\%$, $D = 100\%$, $Me = 110\%$, $s = 24\%$.

Porównać strukturę robotników obu wydziałów ze względu na procent wykonania normy.

2.70. Badanie wieku mężczyzn i kobiet biorących udział w wycieczkach zagranicznych dostarczyło m.in. następujących informacji: a) średni wiek kobiet wynosił 45,23 lat, zaś mężczyzn – 45,37 lat, b) najwięcej było kobiet w wieku 48,48 lat, a mężczyzn w wieku 52,78 lat, c) odchylenie standardowe wieku kobiet było równe 10,52 lat, a wieku mężczyzn – 10,75 lat.

Dokonać analizy porównawczej wieku kobiet i mężczyzn biorących udział w wycieczkach zagranicznych.

2.71. Dokonać analizy porównawczej wieku studentów studiów dziennych i zaocznych, jeśli dysponujemy poniższymi danymi:

studia dzienne: $D = 19$ lat, $\bar{x} = 20$ lat, $V_s = 10\%$,

studia zaoczne: $D = Me = 25$ lat, $s = 2$ lata.

2.72. Różnica wzrostu w 25-osobowej grupie studentów między średnią arytmetyczną i dominantą wynosi 12 cm na korzyść średniej. Wariancja wzrostu jest równa 196 $(\text{cm})^2$. Czy większość studentów w tej grupie ma wzrost niższy czy wyższy od średniej? Jaka jest siła i kierunek asymetrii?

2.73. Czas opóźnienia (w minutach) autobusu nr 24 w ciągu miesiąca kształtował się następująco:

Czas opóźnienia	0–4	4–8	8–12	12–16	16–20
Liczba opóźnień	8	6	5	3	2

Parametry opisowe rozkładu opóźnień autobusu nr 30 były równe: $\bar{x} = 8$ minut, $6 < x_{typ} < 10$, $A_s = -0{,}45$. Dokonać kompleksowej analizy porównawczej czasu opóźnienia dwóch autobusów.

2.74. Czas oczekiwania (w min) na wizytę u lekarza opisuje poniższy rozkład:

Czas oczekiwania	0–4	4–8	8–12	12–16	16–20
Liczba osób	5	10	20	10	5

Dokonać – za pomocą parametrów opisowych – kompleksowej analizy struktury rozkładu czasu oczekiwania na wizytę u lekarza.

2.75. Na trzech zmianach przeprowadzono badanie dotyczące wykonania normy wśród robotników. Na pierwszej zmianie było zatrudnionych 200 robotników. 14 spośród nich wykonywało od 85% do 95% normy, 40 od 95% do 105% normy, 92 od 105% do 115% normy, 14 od 125% do 135% normy, a pozostali – od 115% do 125% normy. Na drugiej zmianie pracowało 100 robotników. Rozkład wykonania normy w tej grupie był następujący:

% wykonania normy	85–95	95–105	105–115	115–125	125–135
Liczba robotników	3	39	26	19	13

Na trzeciej zmianie stwierdzono, że wszyscy robotnicy wykonują normę w granicach od 85% do 135%, przy czym 5% robotników wykonywało do 95% normy, 52% badanych wykonywało do 105% normy, 64% robotników wykonywało do 115% normy, a 79% robotników zatrudnionych na tej zmianie – do 125% normy.

Przeprowadzić wszechstronną analizę porównawczą struktury robotników zatrudnionych na trzech zmianach ze względu na procent wykonania normy.

2.76. W dwóch zakładach produkcyjnych przeprowadzono badanie wydajności pracy. W każdym z zakładów zatrudnionych było po 100 robotników. W zakładzie I przeciętna wydajność na 1 zatrudnionego wynosiła 10 szt./godz., a odchylenie standardowe stanowiło 40% średniej, środkowa wartość wydajności pracy mieściła się w przedziale 8–10 szt./godz., w którym znajdowało się 40% robotników, 30% badanych osiągnęło wydajność niższą niż 8 szt./godz., 25% – niższą niż 7 szt./godz., a 75% – niższą niż 12 szt./godz. Najczęściej spotykana wydajność wynosiła 8 szt./godz. Wyniki badania wydajności w II zakładzie przedstawia poniższy rozkład:

Wydajność w szt./godz.	poniżej 2	2–4	4–6	6–8	8–12	powyżej 12
Odsetek robotników	5	10	15	40	20	10

Dokonać wszechstronnej analizy porównawczej struktury robotników w obu zakładach z punktu widzenia wydajności pracy.

2.77. Średnia arytmetyczna zarobków w grupie 70 mężczyzn była równa 1750 zł, przy współczynniku zmienności 12%. Średnia płaca w grupie 30 kobiet wynosiła 1450 zł, przy typowym obszarze zmienności: $1350 < x_{typ} < 1550$ zł. Obliczyć średnią płacę, współczynnik zmienności i typowy obszar zmienności dla całej zbiorowości (mężczyzn i kobiet łącznie).

2.78. W grupie 100 osób mających dodatkowe – poza głównym – miejsce pracy, dokonano badania wieku. Otrzymano następujące parametry rozkładu wieku: $D = 32{,}3$ lata, $V_s = 15\%$, $A_s = 0{,}02$. Oblicz pozostałe miary tendencji centralnej oraz typowy obszar zmienności wieku.

2.79. Grupę 300 pracowników firmy zbadano pod względem czasu dojazdu do pracy. Otrzymano następujące wyniki:

dojeżdżający własnym środkiem lokomocji: $\bar{x}_1 = 30$ minut, $D_1 = 32$ minuty, $A_s = 0{,}4$;

dojeżdżający komunikacją miejską: $\bar{x}_2 = 45$ minut, $D_2 = 50$ minut, $A_s = -0{,}4$.

Obliczyć współczynnik zmienności dla obydwu grup łącznie, jeśli proporcje liczebnościowe w grupach są określone relacją 3:2.

2.80. Zbadać siłę i kierunek asymetrii, jeśli $Q_1 = 1{,}5$; $Q_3 = 8{,}5$; $V_Q = 0{,}6$.

2.81. Ocenić siłę i kierunek asymetrii, jeśli drugi moment centralny wynosi 3600, a trzeci moment centralny jest równy 90 000.

2.82. W grupie 50 studentów zbadano miesięczne wydatki na prasę. Otrzymano następujące wyniki: $D = 38{,}7$ zł, $V_s = 25{,}6\%$, $A_s = 0{,}04$. Wyznacz typowy obszar zmienności wydatków.

2.83. W grupie 120 wybranych jednostek statystycznych dokonano badania cechy X. Ustalono, że suma wariantów cech jest równa 1258, suma kwadratów wariantów cech 10860, zaś suma sześcianów wartości cech wynosi 175 600. Oblicz możliwe do wyznaczenia parametry struktury.

2.84. W wybranej grupie 10 pracowników zebrano informacje o czasie spóźnień do pracy w okresie tygodnia. Otrzymano następujące informacje (w minutach): 40, 50, 37, 52, 64, 75, 48, 62, 72 i 70. Czy na tej podstawie można stwierdzić, że:

a) średnie spóźnienie jest równe 56 minut,

b) spóźnienie połowy osób nie przekracza 64 minut,

c) rozkład spóźnień jest symetryczny,

d) względna dyspersja rozkładu wynosi 22,4%?

2.85. Wydatki na reklamę w dwóch zakładach w ciągu roku charakteryzują następujące parametry:

zakład I: $\bar{x} = 3000$ zł,

zakład II: $D = 2500$ zł, $Me = 2000$ zł.

W którym z zakładów wydatki na reklamę były – średnio biorąc – wyższe? W którym z zakładów wydatki na reklamę były bardziej zróżnicowane, jeśli wariancje wydatków w obu zakładach były jednakowe?

2.86. Wykorzystując poniższe dane dokonaj za pomocą współczynnika Lorenza oceny koncentracji ludności w miastach:

Grupy miast wg liczby ludności	Liczba miast	Łączna liczba ludności w tys.
poniżej 5000	273	840,3
poniżej 10 000	455	2128,4
poniżej 20 000	632	4678,0
poniżej 40 000	771	8906,1
poniżej 100 000	822	12 333,3
poniżej 200 000	844	15 191,7
powyżej 200 000	20	8711,9

2.87. Zbiorowość 50 pracowników zarządów banków w mieście L zbadano ze względu na wysokość miesięcznej płacy. Otrzymano poniższy rozkład:

Płace w tys. zł	6–8	8–10	10–12	12–14	14–16
Liczba pracowników	3	6	27	13	1

Czy rozkład płac jest leptokurtyczny czy platokurtyczny?

2.88. Struktura obiektów turystycznych w województwie L według liczby miejsc noclegowych przedstawia się następująco:

Liczba miejsc noclegowych	Liczba obiektów	Łączna liczba miejsc noclegowych
do 40	66	2365
41–60	28	1960
61–100	15	1460
101–140	9	910
141–180	1	140
181 i więcej	1	200

Zbadać graficznie i rachunkowo siłę koncentracji miejsc noclegowych w obiektach turystycznych województwa L.

2.89. N podstawie badania 56 magazynów otrzymano poniższe informacje:

Powierzchnia w m^2	20–40	40–60	60–100	100–150	150–200	200–300
Liczba magazynów	14	16	10	8	5	3

Wyznaczyć graficznie i analitycznie siłę koncentracji badanego zjawiska.

2.90. Krzywą koncentracji Lorenza charakteryzują cztery punkty o następujących współrzędnych określonych przez skumulowane wartości łącznego funduszu cechy i liczebności: (0,4;0,1), (0,8;0,4), (0,9;0,5), (1,0;1,0). Co na podstawie powyższych informacji można sądzić o sile koncentracji?

2.91. Rozkład gimnazjów pod względem liczby uczniów przedstawia się następująco:

Liczba uczniów w szkole	0–40	40–80	80–120	120–160	160–200
Liczba szkół	10	60	70	65	100

Zbadać siłę koncentracji.

2.92. Rozkład powierzchni sklepów (w m^2) w mieście L przedstawia się następująco:

Powierzchnia	30–50	50–70	70–90	90–110	110–130	130–150
Liczba sklepów	9	19	23	14	9	7

Jak silna jest koncentracja powierzchni sklepów w mieście L?

2.93. Z wykresu wieloboku koncentracji Lorenza dotyczącego ludności miejskiej zamieszkałej w gminach województwa lubelskiego wynika, iż krzywa koncentracji dzieli pole trójkąta pod linią równomiernego rozdziału w stosunku 1:4 na korzyść pola pod krzywą. Czy można na tej podstawie określić siłę koncentracji badanego zjawiska?

2.94. Częstość zamieszczonych w prasie reklam przez klientów agencji X w ostatnim miesiącu przedstawia się następująco:

Liczba reklam (x_i)	1	2	3	4	5
Częstość (w_i)	0,12	0,14	0,22	0,36	0,16

Wyznaczyć współczynnik kurtozy i ekscesu.

2.95. Czy dla trenera kadry siatkarzy korzystniejszy jest rozkład wzrostu kadrowiczów charakteryzujący się mniejszym, czy większym skupieniem wokół wartości średniej niż normalne?

2.96. W powiecie Ł zbadano powierzchnię plantacji chmielu (w arach), otrzymując następujący rozkład:

Powierzchnia	0,5–3,5	3,5–6,5	6,5–9,5	9,5–12,5
Liczba plantacji	33	59	27	5

Sporządzić wykres wieloboku koncentracji Lorenza i obliczyć współczynnik koncentracji Lorenza.

2.97. W badaniu koncentracji miesięcznych zarobków stwierdzono, że zarobki jednej trzeciej najniżej zarabiających pracowników stanowiły 20% łącznego funduszu płac. Zarobki trzeciej części pracowników o najwyższych pensjach stanowiły natomiast 50% łącznego funduszu płac. Oblicz wartość współczynnika koncentracji Lorenza.

2.98. Rozkład zawartości tłuszczu w serach twardych kształtował się następująco:

Zawartość tłuszczu w %	20–22	22–24	24–26	26–28	28–30
Liczba zakładów	6	8	11	19	6

Oblicz współczynnik ekscesu. Wynik zinterpretuj.

2.99. Skumulowane wskaźniki struktury liczby małych i średnich przedsiębiorstw oraz łącznej liczby pracujących w nim pracowników kształtowały się następująco:

Skumulowane wskaźniki liczby przedsiębiorstw	Skumulowane wskaźniki struktury łącznej liczby pracujących
89,5	24,5
98,5	43,5
99,0	48,7
99,5	57,3
99,8	70,7
99,9	85,7
100,0	100,0

Obliczyć i zinterpretować współczynnik koncentracji Lorenza.

2.100. Rozkład ocen uzyskanych na egzaminie z matematyki przez 1000 studentów dziennych kształtował się następująco:

Oceny	2	3	4	5
Liczba studentów	100	300	400	200

Czy rozkład ocen charakteryzuje się większym czy mniejszym skupieniem wokół wartości średniej niż normalne?

Rozdział III

PROBABILISTYCZNE PODSTAWY WNIOSKOWANIA STATYSTYCZNEGO

Jak już wcześniej stwierdzono, wnioskowanie statystyczne obejmuje procedury właściwe badaniu częściowemu. Badanie częściowe obejmuje tylko część populacji generalnej, a jego wyniki są uogólniane na całą zbiorowość. Podzbiór elementów populacji objęty badaniem częściowym nazywamy **próbą**. Fundamentalną sprawą w badaniu częściowym jest to, aby próba była reprezentatywna, tzn. jej struktura pod względem badanej cechy była zbliżona do struktury populacji. Brak jest jednak metody wyboru, która gwarantowałaby pełną reprezentatywność próby (populacja generalna jest nieznana, a dopiero badanie częściowe ma dostarczyć informacji o niej). Ogólnie stwierdza się, że próba jest reprezentatywna, jeśli spełnione są dwa warunki:

1) elementy populacji generalnej pobierane są do próby w sposób losowy,

2) próba jest dostatecznie liczna.

Próby reprezentatywne z populacji skończonych otrzymuje się w wyniku **losowania**. Losowanie próby statystycznej ze zbiorowości o skończonej liczbie elementów nazywamy **metodą reprezentacyjną**.

3.1. Losowy dobór próby

Wyróżnikiem losowego doboru próby jest to, że stwarza on **jednakową szansę** (takie samo prawdopodobieństwo) dostania się do próby zarówno wszystkim poszczególnym jednostkom populacji generalnej, jak też wyróżnionym w danym badaniu zespołom (grupom) elementów.

Model postępowania prowadzący do losowego wyboru próby nazywamy **schematem losowania**. Wśród wielu schematów losowania wyróżnia się m.in.:

- losowanie niezależne i zależne,
- losowanie indywidualne i zespołowe,

- losowanie jednostopniowe i wielostopniowe,
- losowanie nieograniczone i ograniczone.

Losowanie niezależne (zwane też losowaniem ze zwracaniem) to takie, w którym losujemy jednostkę po jednostce za każdym razem z tej samej populacji i przy jednakowych prawdopodobieństwach wyboru. Ta sama jednostka może być wylosowana wielokrotnie, gdyż po jej pierwszym wylosowaniu zostaje ona ponownie włączona w skład populacji.

W **losowaniu zależnym** (losowaniu ze zwracaniem) jednostka raz wylosowana do próby nie bierze udziału w dalszym losowaniu, gdyż nie jest zwracana do danej populacji generalnej. W miarę losowania kolejnych jednostek do próby prawdopodobieństwo dostania się elementów do próby ulega tu zmianie.

Losowanie jest **indywidualne**, gdy z populacji generalnej pobiera się do próby poszczególne elementy. Losowanie **zespołowe** (grupowe) polega na pobieraniu określonych zespołów, składających się z więcej niż jednego elementu (np. losowanie do próby całych rodzin).

W losowaniu **jednostopniowym** stosowane jest jedno stadium: jednostki do próby są pobierane od razu z całej populacji generalnej. W losowaniu wielostopniowym, np. trzystopniowym, cała populacja generalna dzielona jest najpierw na M dużych grup, zwanych jednostkami losowania pierwszego stopnia. Z kolei każda jednostka losowania stopnia pierwszego składa się z L mniejszych grup, zwanych jednostkami losowania drugiego stopnia. Wreszcie, każda jednostka losowania drugiego stopnia zawiera po K jednostek populacji. Losowanie wielostopniowe polega na tym, że najpierw losuje się m jednostek stopnia pierwszego, z kolei z każdej jednostki stopnia pierwszego pobiera się l jednostek stopnia drugiego, a następnie z każdej jednostki stopnia drugiego – losuje się k jednostek. Ostatecznie próba losowa liczy $m \cdot k \cdot l$ jednostek, przy czym $m < M$, $l < L$, $k < K$.

Losowanie jest **nieograniczone**, jeżeli odbywa się od razu z całej populacji, podczas gdy losowania **ograniczonego** dokonuje się z poszczególnych części populacji, zwanych **warstwami**. Podziału populacji generalnej na wewnętrznie jednorodne warstwy dokonuje się przed przystąpieniem do losowania. Po ustaleniu, ile jednostek losujemy z każdej warstwy, losowania dokonujemy niezależnie z każdej warstwy. Próbę stanowią zatem jednostki wylosowane ze wszystkich warstw.

Do losowania ograniczonego zalicza się również losowanie **systematyczne**. Polega ono na tym, że do próby włączamy co k-ty element populacji, poczynając od losowo wybranego j-tego elementu. Symbolem k oznaczono tu przedział losowania, który jest ilorazem liczebności populacji i liczebności próby.

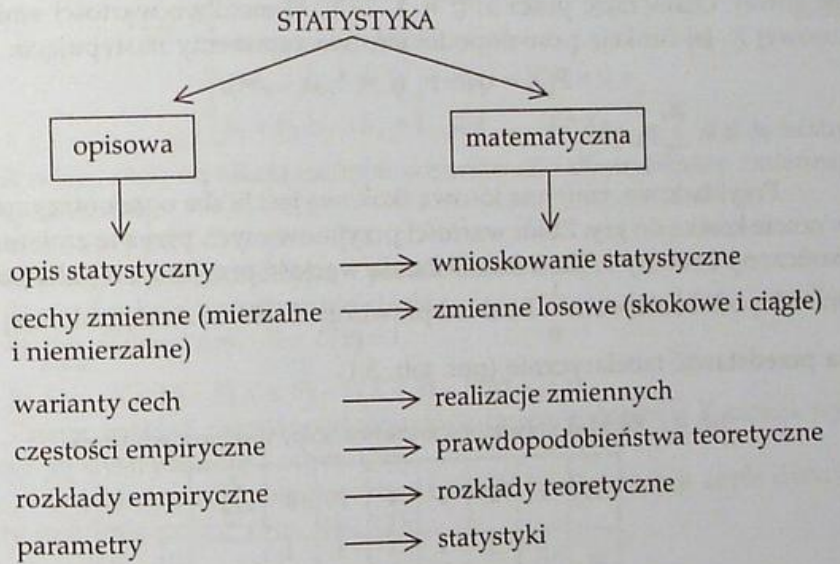
Przedstawione powyżej schematy losowania dotyczą populacji skończonych, których elementy można ponumerować. Mówimy wówczas o tzw. **operacie losowania**. Próby losowe z populacji skończonych otrzymuje się

w wyniku zastosowania określonych **technik losowania** (np. tablic czy generatorów liczb losowych).

Losowanie indywidualne, nieograniczone, niezależne nazywamy **losowaniem prostym**, a otrzymaną próbę określamy mianem **próby losowej prostej**. Używając skrótowo pojęć: próba losowa czy próba, mamy na myśli próbę losową prostą.

3.2. Zmienne losowe i ich rodzaje

Możliwość uogólniania wyników otrzymanych z próby losowej na całą populację generalną, z której ona pochodzi, daje rachunek prawdopodobieństwa. Rachunek ten stanowi teoretyczną podstawę wnioskowania statystycznego. Przejście od opisu statystycznego do wnioskowania statystycznego wymaga rozróżniania używanych pojęć, w zależności od zakresu przeprowadzonej analizy, a mianowicie:



Fundamentalnym pojęciem wnioskowania statystycznego jest termin: **zmienna losowa**. Zmienną losową nazywamy jednoznaczne przyporządkowanie każdemu zdarzeniu elementarnemu wartości liczbowej. Zmienne losowe przyjmujące skończoną lub nieskończoną, ale przeliczalną liczbę wartości nazywamy **skokowymi** lub **dyskretnymi**. W przypadku, gdy możliwe wartości zmiennej losowej tworzą przedział ze zbioru liczb rzeczywistych – nazywamy ją **ciągłą**. Przykładami zmiennej losowej skokowej są: liczba dzieci w rodzinie, liczba ziaren w kłosie itp., a ciągłej: wzrost, waga, wiek itp. W zastosowaniach metod statystycznych zmienne losowe ciągłe (w od-

różnieniu od zmiennych losowych skokowych) nie występują w swojej czystej, matematycznej postaci. Wynika to stąd, że pomiar tego typu zmiennych odbywa się z określoną dokładnością (np. do dwóch miejsc po przecinku). Nie zmienia to faktu, że zbiór możliwych wartości zmiennej losowej ciągłej jest mocym continuum (jest nieprzeliczalny).

Zmienna losowa skokowa

Przyjęło się oznaczać zmienne losowe dużymi literami z końca alfabetu, np. X, Y, Z. Konkretne wartości przybierane przez te zmienne (zwane **realizacjami**) oznacza się odpowiednio małymi literami: x, y, z. Przyporządkowanie wszystkim możliwym wartościom (realizacjom) zmiennej losowej X odpowiadających im – sumujących się do jedności – prawdopodobieństw, określa **rozkład zmiennej losowej skokowej**. Przyporządkowanie to nosi również nazwę **funkcji rozkładu** prawdopodobieństwa zmiennej losowej skokowej. Oznaczając przez x_i ($i = 1, 2, ..., k$) możliwe wartości zmiennej losowej X, jej funkcję prawdopodobieństwa zapiszemy następująco:

$$P(X = x_i) = p_i \ (i = 1, 2, ..., n), \tag{3.1}$$

gdzie: $p_i \geq 0$, $\sum_{i=1}^{n} p_1 = 1$.

Przykładowo, zmienną losową skokową jest liczba oczek otrzymanych w rzucie kostką do gry. Zbiór wartości przyjmowanych przez tę zmienną jest skończony i równy 6. Zmienna ta każdą wartość przyjmuje z jednakowym prawdopodobieństwem $\frac{1}{6}$. Rozkład prawdopodobieństwa tej zmiennej można przedstawić tabelarycznie (por. tab. 3.1)

Tab. 3.1. Rozkład prawdopodobieństwa liczby wyrzuconych oczek

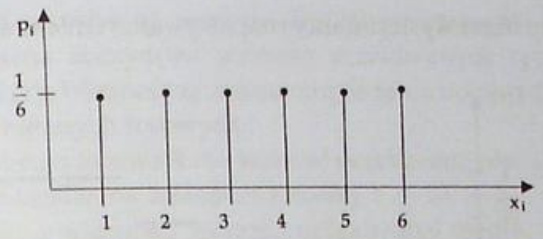
x_i	1	2	3	4	5	6	$\sum_{i=1}^{n}$
p_i	$\frac{1}{6}$	$\frac{1}{6}$	$\frac{1}{6}$	$\frac{1}{6}$	$\frac{1}{6}$	$\frac{1}{6}$	1

Rozkład prawdopodobieństwa zmiennej losowej skokowej można również przedstawić graficznie za pomocą diagramu (rys. 3.1)

Na osi odciętych odkłada się wartości przyjmowane przez zmienną losowa X, a na osi rzędnych – odpowiadające im prawdopodobieństwa.

Do charakterystyki rozkładu zmiennej losowej można również użyć **dystrybuanty**. Dystrybuantą – oznaczaną symbolem $F(x)$ – nazywamy funkcję określającą prawdopodobieństwo tego, że zmienna losowa X przyjmie wartości mniejsze od ustalonego x, czyli:

$$F(x) = P(X < x), \ x \in R. \tag{3.2}$$



Rys. 3.1. Rozkład prawdopodobieństwa liczby wyrzuconych oczek
Źródło: Opracowanie własne.

Zakładając, że zbiór wartości zmiennej losowej X został uporządkowany rosnąco, dystrybuantę można zapisać następująco:

$$F(x) = \begin{cases} 0 & \text{dla} \quad x \leq x_1 \\ p_1 & \text{dla} \quad x_1 < x \leq x_2 \\ p_1 + p_2 & \text{dla} \quad x_2 < x < x_3 \\ \dots & \dots \quad \dots \\ p_1 + p_2 + p_{n-1} & \text{dla} \quad x_{n-1} < x \leq x_n \\ p_1 + p_2 + \dots + p_n = 1 & \text{dla} \quad x > x_n \end{cases}. \tag{3.3}$$

Z relacji (3.3) wynikają następujące własności dystrybuanty zmiennej losowej skokowej:

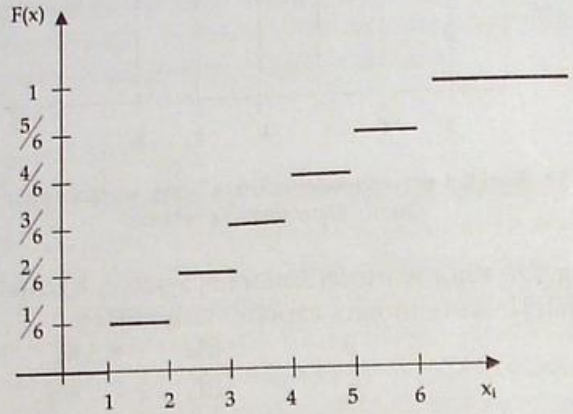
1) $0 \leq F(x) \leq 1$,
2) jest funkcją niemalejącą i przedziałami stałą,
3) jest funkcją lewostronnie ciągłą,
4) $\lim_{x \to -\infty} F(x) = 0$ oraz $\lim_{x \to +\infty} F(x) = 1$,
5) $P(a < X \leq b) = P(X \leq b) - P(X < a) = F(b) - F(a)$.

Znając rozkład prawdopodobieństwa zmiennej losowej X można wyznaczyć jej dystrybuantę i odwrotnie.

W doświadczeniu polegającym na rzucie kostką do gry, zapis dystrybuanty przyjmie postać (por. tab. 3.1):

$$F(x) = \begin{cases} 0 & \text{dla} \quad x \leq 1 \\ \frac{1}{6} & \text{dla} \quad 1 < x \leq 2 \\ \frac{2}{6} & \text{dla} \quad 2 < x \leq 3 \\ \frac{3}{6} & \text{dla} \quad 3 < x \leq 4 \\ \frac{4}{6} & \text{dla} \quad 4 < x \leq 5 \\ \frac{5}{6} & \text{dla} \quad 5 < x \leq 6 \\ 1 & \text{dla} \quad x > 6 \end{cases}.$$

Graficzny obraz dystrybuanty rozpatrywanej zmiennej losowej przedstawia rys. 3.2.



Rys. 3.2. Dystrybuanta liczby oczek w rzucie kostką
Źródło: Opracowanie własne

Znając rozkład prawdopodobieństwa lub dystrybuantę zmiennej losowej mamy pełną informację o niej. Niekiedy interesują nas tylko niektóre własności, charakteryzujące zmienną w sposób syntetyczny. Tego rodzaju charakterystyki noszą nazwę **parametrów rozkładu**. Podstawowymi parametrami rozkładu zmiennej losowej X są: **wartość oczekiwana, wariancja i odchylenie standardowe.**

W przypadku zmiennej losowej skokowej, wartość oczekiwana – oznaczana symbolem $E(X)$ – dana jest wzorem:

$$E(X) = \sum_{i=1}^{n} x_i p_i. \tag{3.4}$$

Dla zmiennej losowej X określonej jako liczba wyrzuconych oczek w rzucie kostką do gry, wartość oczekiwana wynosi (por. tab. 3.1):

$$E(X) = 1 \cdot \frac{1}{6} + 2 \cdot \frac{1}{6} + 3 \cdot \frac{1}{6} + 4 \cdot \frac{1}{6} + 5 \cdot \frac{1}{6} + 6 \cdot \frac{1}{6} + = 3{,}5.$$

Wartość oczekiwana jest średnią wartością zmiennej losowej X. Charakteryzuje się ona poniższymi własnościami:

1. Wartość oczekiwana stałej równa się tej stałej: $E(C) = C$.

2. Wartość oczekiwana sumy dwóch zmiennych losowych jest sumą wartości oczekiwanych tych zmiennych: $E(X+Y) = E(X) + E(Y)$. Własność tę można uogólnić na dowolną liczbę zmiennych.

3. Wartość oczekiwana iloczynu stałej i zmiennej losowej jest równa iloczynowi tej stałej i wartości oczekiwanej zmiennej losowej X: $E(CX) = E(C) \cdot E(X) = CE(X)$.

4. Wartość oczekiwana iloczynu dwóch niezależnych zmiennych losowych jest równa iloczynowi wartości oczekiwanych tych zmiennych: $E(XY) = E(X) \cdot E(Y)$. Własność tę można uogólnić na iloczyn dowolnej liczby niezależnych zmiennych losowych.

Niech zmienna losowa X ma wartość oczekiwaną równą 10. Mamy obliczyć wartość oczekiwaną zmiennej losowej $Y = 2X + 5$.

Korzystając z własności wartości oczekiwanej mamy:

$E(Y) = E(2X + 5) = E(2X) + (E5) = 2E(X) + E(5) = 2 \cdot 10 + 5 = 15.$

Wariancję zmiennej losowej skokowej X obliczamy ze wzoru:

$$D^2(X) = E(X^2) - [E(X)]^2 = \sum_{i=1}^{n} x_i^2 p_i - [E(X)]^2. \tag{3.5}$$

Pierwiastek kwadratowy z wariancji jest jej odchyleniem standardowym:

$$D(X) = \sqrt{D^2(X)}. \tag{3.6}$$

Wariancja i odchylenie standardowe są miarami zróżnicowania w rozkładzie zmiennej losowej.

W przykładzie dotyczącym rozkładu liczby wyrzuconych oczek parametry te wynoszą:

$$D^2(X) = \left(1^2 \cdot \frac{1}{6} + 2^2 \cdot \frac{1}{6} + 3^2 \cdot \frac{1}{6} + 4^2 \cdot \frac{1}{6} + 5^2 \cdot \frac{1}{6} + 6^2 \cdot \frac{1}{6}\right) - 3{,}5^2 =$$

$$= \left(\frac{1}{6} + \frac{4}{6} + \frac{9}{6} + \frac{16}{6} + \frac{25}{6} + \frac{36}{6}\right) - 3{,}5^2 = 15{,}17 - 12{,}25 = 2{,}92,$$

$$D(X) = \sqrt{2{,}92} = 1{,}71.$$

Wartość wariancji nie ma interpretacji. Odchylenie standardowe informuje natomiast o tym, że średnie odchylenie liczby wyrzuconych oczek od wartości oczekiwanej wynosi 1,71.

Wariancja zmiennej losowej ma następujące własności:

1. Wariancja stałej jest równa zeru: $D^2(C) = 0$.

2. Wariancja iloczynu stałej i zmiennej losowej równa się iloczynowi kwadratu tej stałej i wariancji zmiennej losowej: $D^2(CX) = C^2D^2(X)$.

3. Wariancja sumy stałej i zmiennej losowej jest równa wariancji zmiennej losowej: $D^2(C + X) = D^2(X)$.

4. Wariancja sumy (różnicy) niezależnych zmiennych losowych równa się sumie wariancji tych zmiennych: $D^2(X + Y) = D^2(X - Y) = D^2(X) + D^2(Y)$.

Niech zmienne X i Y będą niezależnymi zmiennymi losowymi, przy czym $D^2(X) = 2$, $D^2(Y) = 8$. Obliczyć $D(X - Y)$.

Korzystając z własności wariancji mamy:

$D^2(X - Y) = D^2(X) + D^2(Y) = 2 + 8 = 10$

$D(X - Y) = \sqrt{10}$.

Zmienna losowa ciągła

Dla zmiennej losowej ciąglej niemożliwe jest przypisanie wszystkim jej wartościom prawdopodobieństw sumujących się do jedności. Możliwe jest natomiast przyporządkowanie takich prawdopodobieństw przedziałom liczbowym typu: $P(x< X< x+\Delta x)$, gdzie Δx jest długością przedziału o początku w punkcie x.

Jeśli przy $\Delta x \to 0$ istnieje granica $f(x)$ o postaci:

$$\lim_{\Delta x\to 0}\frac{P(x< X< x+\Delta x)}{\Delta x}=f(x), \tag{3.7}$$

to granicę tę nazywamy **funkcją gęstości prawdopodobieństwa** zmiennej losowej ciąglej X (używa się również terminu: **gęstość prawdopodobieństwa**).

Gęstość prawdopodobieństwa zmiennej losowej ciąglej X można zatem określić jako średnią „ilość prawdopodobieństwa" przypadającą na jednostkę długości przedziału $(x, x+\Delta x)$, gdy długość tego przedziału zmierza do zera.

Określoną na zbiorze liczb rzeczywistych funkcję gęstości prawdopodobieństwa $f(x)$ zmiennej losowej ciąglej X charakteryzują następujące własności:

1. $f(x)\geq 0$,
2. $\int_a^b f(x)dx = P(a< X\leq b) = P(a\leq X< b) = P(a< X< b)$,
3. $\int_a^b f(x)dx = 1$ lub $\int_{-\infty}^{+\infty} f(x)dx = 1$.

Dystrybuantę zmiennej losowej ciąglej X definiuje się podobnie jak dystrybuantę zmiennej skokowej, przy czym sumę zastępuje się całką:

$$F(x)=P(X< x)=\int_{-\infty}^{x} f(u)du, \tag{3.8}$$

gdzie: $f(u)$ jest funkcją gęstości zmiennej losowej X.

Określona wzorem (3.8) dystrybuanta jest funkcją ciągłą i spełnia następujące warunki:

1. $0\leq F(x)\leq 1$,
2. $F(x)$ jest funkcją niemalejącą,
3. $\lim_{x\to-\infty} F(x)=0$ oraz $\lim_{x\to+\infty} F(x)=1$.

Jeżeli gęstość f(x) zmiennej losowej X jest ciągła, wówczas zachodzi relacja:

$$F'(x)=f(x), \tag{3.9}$$

gdzie: $F'(x)$ oznacza pochodną dystrybuanty $F(x)$.

Dla zmiennej losowej ciągłej zachodzi:

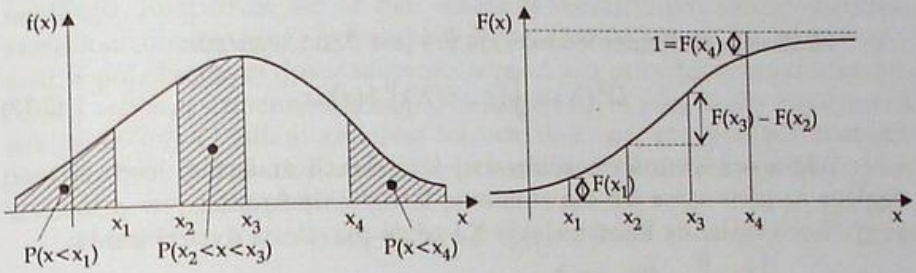
$$P(X-a)=0 \tag{3.10}$$

oraz

$$P(a\leq X\leq b)=P(a< X< b)=\int_a^b f(x)dx=F(b)-F(a), \tag{3.11}$$

gdzie: a i b są stałymi ($a< b$).

Przykładowy wykres funkcji gęstości i dystrybuanty zmiennej losowej ciąglej przedstawia rys. 3.3.



Rys. 3.3. Wykresy funkcji gęstości i dystrybuanty zmiennej losowej ciąglej
Źródło: M. Sobczyk, *Statystyka...*, s. 115.

Załóżmy, że gęstość prawdopodobieństwa zmiennej losowej ciąglej X jest określona wzorem:

$$f(x)=\begin{cases}0 & \text{dla} \quad x\leq 0\\ Cx & \text{dla} \quad 0< x\leq 4.\\ 0 & \text{dla} \quad x>4\end{cases}$$

Naszym zadaniem jest wyznaczenie stałej C, określenie dystrybuanty oraz obliczenie $P(1\leq X\leq 2)$.

Stałą C wyznaczamy z zależności:

$$\int_{-\infty}^{+\infty} f(x)dx=1.$$

Mamy więc:

$$\int_{-\infty}^{+\infty} f(x)dx=\int_{-\infty}^{0} 0dx+\int_0^4 Cxdx+\int_4^{+\infty} 0dx=\int_0^4 Cxdx=C\frac{1}{2}x^2\Big|_0^4=8C=1,$$

stąd $C=\frac{1}{8}$.

Dystrybuantę wyznaczamy ze wzoru (3.8):

$$F(x)=\int_{-\infty}^{x} f(u)du=\int_0^x \frac{1}{8}udu=\frac{1}{16}x^2\Big|_0^x=\frac{1}{16}x^2 \text{ dla } 0< x\leq 4$$

$$F(x)=\begin{cases}0 & \text{dla} \quad x\leq 0\\ \frac{1}{16}x^2 & \text{dla} \quad 0< x\leq 4.\\ 1 & \text{dla} \quad x>4\end{cases}$$

Szukane prawdopodobieństwo obliczamy następująco:

$$P(1 \le X \le 2) = F(2) - F(1) = \frac{4}{16} - \frac{1}{16} = \frac{3}{16}.$$

Wartość oczekiwana zmiennej losowej typu ciągłego jest określona wzorem:

$$E(X) = \int_{-\infty}^{+\infty} xf(x)dx. \tag{3.12}$$

Wariancja zmiennej losowej ciągłej jest dana wzorem:

$$D^2(X) = \int_{-\infty}^{+\infty} [x - E(X)]^2 f(x)dx. \tag{3.13}$$

Własności wartości oczekiwanej i wariancji zmiennej losowej typu ciągłego są takie same jak dla zmiennej losowej dyskretnej.

Niech zmienna losowa ciągła X będzie określona dystrybuantą:

$$F(x) = \begin{cases} 1 - \frac{8}{x^3} & \text{dla} \quad x \ge 2 \\ 0 & \text{dla} \quad x < 2 \end{cases}.$$

Naszym zadaniem jest wyznaczenie wartości oczekiwanej i wariancji tej zmiennej.

Jak wynika ze wzorów (3.12) oraz (3.13), do wyznaczenia $E(X)$ oraz $D^2(X)$ niezbędna jest znajomość postaci funkcji gęstości zmiennej losowej X. Do wyznaczenia $f(x)$ wykorzystujemy wzór (3.9). Mamy więc:

$$f(x) = F'(x) = \begin{cases} \frac{24}{x^2} & \text{dla} \quad x \ge 2 \\ 0 & \text{dla} \quad x < 2 \end{cases}.$$

Stąd też:

$$E(X) = \int_2^{\infty} x\frac{24}{x^4}dx = \int_2^{\infty} \frac{24}{x^3}dx = 24\int_2^{\infty} \frac{1}{x^3}dx = 3,$$

$$D^2(X) = \int_2^{+\infty} \frac{24(x-3)^3}{x^4}dx = 3.$$

3.3. Rozkłady teoretyczne zmiennych losowych

Realizacje zmiennych losowych mogą być skończonymi lub nieskończonymi zbiorami informacji liczbowych, występujących z określonymi prawdopodobieństwami. Informacje te tworzą **rozkłady teoretyczne**. W statystyce opisowej ich odpowiednikiem są rozkłady empiryczne, będące zbiorem informacji o wariantach cech i ich częstościach.

W statystyce matematycznej, najczęściej wykorzystywane są rozkłady:

❖ dla zmiennych losowych skokowych: rozkład **dwumianowy** (zwany też binomialnym lub Bernoulliego) oraz rozkład **Poissona**;

❖ dla zmiennych losowych ciągłych: rozkład **normalny Gaussa-Laplace'a** i jego transformacje (rozkład **chi-kwadrat**, rozkład **t-Studenta**, rozkład **Fishera-Snedecora**).

Rozkłady teoretyczne zmiennej losowej skokowej

Rozkład dwumianowy opiera się na tzw. schemat doświadczeń Bernoulliego. Rozpatruje się w nim n ($n \ge 2$) niezależnych eksperymentów, których rezultatem może być sukces lub porażka. Prawdopodobieństwo sukcesu w pojedynczym doświadczeniu wynosi p, a prawdopodobieństwo porażki $1 - p = q$. Zmienną losową X definiuje się tu jako liczbę uzyskanych sukcesów. Jest to zatem zmienna losowa skokowa, przyjmująca wartości: $k = 0, 1, 2, ..., n$. Prawdopodobieństwo tego, że zmienna losowa X przyjmie wartość k, wyraża się wzorem:

$$P(X = k) = \binom{n}{k} p^k q^{n-k}, \tag{3.14}$$

gdzie: $\binom{n}{k} = \frac{n!}{k!(n-k)!}$ jest liczbą kombinacji k-elementowych z n-elementowego zbioru.

Określona wzorem (3.14) funkcja rozkładu prawdopodobieństwa zmiennej losowej X nazywana jest **rozkładem dwumianowym** (binomialnym, Bernoulliego) z parametrami n oraz p. Rozkład dwumianowy jest więc rozkładem dwuparametrycznym. Wartość oczekiwana oraz wariancja zmiennej losowej X o rozkładzie dwumianowym są odpowiednio równe:

$$E(X) = np \text{ oraz } D^2(X) = npq. \tag{3.15}$$

Załóżmy, że zmienna losowa skokowa X ma rozkład dwumianowy z wartością oczekiwaną 0,5 oraz $n = 5$. Naszym zadaniem jest obliczenie prawdopodobieństw: a) $P(X = 2)$, b) $P(X \ge 2)$.

W rozkładzie dwumianowym $E(X) = np$. Stąd też $p = \frac{E(X)}{n} = \frac{0,5}{5} = 0,1$.

Wykorzystując wzór (3.14) mamy:

a) $P(X = 2) = \binom{5}{2} \cdot 0,1^2 \cdot 0,9^3 = 0,0729$,

b) $P(X \ge 2) = P(X = 2) + P(X = 3) + P(X = 4) + P(X = 5) =$
$= 0,0729 + 0,0081 + 0,00045 + 0,00001 = 0,08146.$

Jeśli prawdopodobieństwo pojawienia się sukcesu w pojedynczym doświadczeniu bardzo małe ($p \le 0,02$), a liczba przeprowadzonych doświadczeń duża ($n \ge 100$), to zmienna losowa X określana jako liczba sukcesów podlega – w przybliżeniu – **rozkładowi Poissona**. Zmienna losowa X ma rozkład Poissona, jeżeli:

$$P(X = k) = \frac{\lambda^k}{k!} e^{-\lambda} \quad (k = 0, 1, 2, ..., n), \tag{3.16}$$

gdzie $\lambda = np$ to jednocześnie wartość oczekiwana i wariancja tego rozkładu, e zaś jest podstawą logarytmów naturalnych ($e = 2{,}718$). Rozkład Poissona zależy zatem od jednego parametru: λ.

Załóżmy, że centrala telefoniczna w małej firmie obsługuje 100 abonentów. Prawdopodobieństwo tego, że abonent zgłosi się do centrali w ciągu godziny, jest równe 0,02. Obliczyć prawdopodobieństwo tego, że w ciągu godziny będą co najmniej cztery zgłoszenia.

Zmienną losową X jest tu liczba abonentów zgłaszających się do centrali w ciągu godziny. Zmienna ta ma rozkład Poissona z $\lambda = 2$ ($\lambda = np = 100 \cdot 0{,}02 = 2$). Szukane prawdopodobieństwo wynosi zatem:

$$P(X \geq 4) = 1 - [P(X=0) + P(X=1) + P(X=2) + P(X=3)] =$$

$$= 1 - \left(\frac{2^0}{0!} e^{-2} + \frac{2^1}{1!} e^{-2} + \frac{2^2}{2!} e^{-2} + \frac{2^3}{3!} e^{-2} \right) =$$

$$1 - (0{,}1353 + 0{,}2707 + 0{,}2707 + 0{,}1804) = 0{,}1429.$$

Rozkłady teoretyczne zmiennej losowej ciągłej

Kluczową rolę w statystyce matematycznej odgrywa rozkład normalny, zwany także rozkładem Gaussa-Laplace'a. Jego znaczenie wynika przede wszystkim stąd, że przy nieograniczonym wzroście liczby niezależnych doświadczeń, wszystkie rozkłady teoretyczne (zarówno zmiennych losowych, jak i ciągłych) są szybko zbieżne do rozkładu normalnego. Ponadto we wnioskowaniu statystycznym opartym na wynikach badań prób losowych popełniane są błędy, których rozkład jest normalny lub granicznie normalny.

Zmienna losowa ciągła X ma **rozkład normalny**, jeśli jej funkcja gęstości prawdopodobieństwa ma postać:

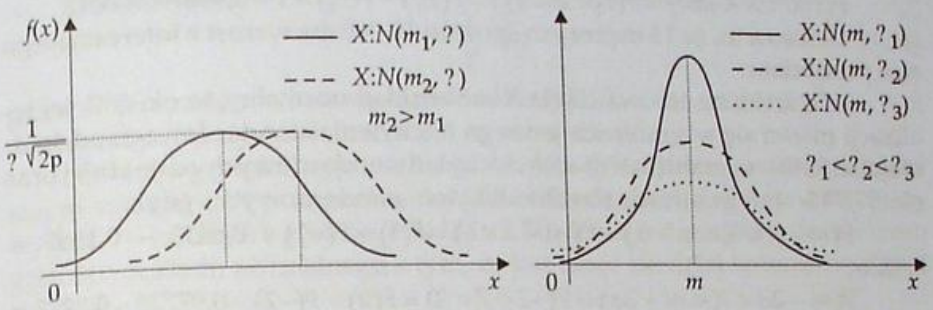
$$f(x) = \frac{1}{\sigma\sqrt{2\pi}} e^{\frac{-(x-m)^2}{2\sigma^2}}, \quad x \in R, \sigma > 0, \tag{3.17}$$

gdzie: $m = E(X)$, $\sigma = D(X)$, $\pi = 3{,}14$, $e = 2{,}718$ (podstawa logarytmów naturalnych).

Rozkład normalny charakteryzują dwa parametry: m i σ, które są odpowiednio wartością oczekiwaną i odchyleniem standardowym. Dla zmiennej losowej X o rozkładzie normalnym z parametrem m i σ stosuje się oznaczenie $N(m,\sigma)$. Tak więc zapis $N(2;0{,}5)$ oznacza, że zmienna losowa X ma rozkład normalny o średniej 2 i odchyleniu standardowym 0,5.

Wykres funkcji gęstości prawdopodobieństwa zmiennej losowej X nosi nazwę **krzywej normalnej** lub krzywej Gaussa-Laplace'a (por. rys. 3.4).

Z rysunku 3.4 wynika, że parametr m rozkładu normalnego przesuwa krzywą wzdłuż osi odciętych, a parametr σ powoduje, że krzywa staje się bardziej spłaszczona lub wysmukła. Im mniejsza jest wartość σ, tym bardziej rozkład jest skupiony wokół wartości oczekiwanej $E(X) = m$.



Rys. 3.4. Rozkłady normalne przy różnych m i σ

Dystrybuanta rozkładu normalnego jest określona wzorem:

$$F(x) = \frac{1}{\sigma\sqrt{2\pi}} \int_{-\infty}^{x} e^{\frac{-(t-m)^2}{2\sigma^2}} dt. \tag{3.18}$$

Wykorzystywanie dystrybuanty (3.18) do obliczania prawdopodobieństw natrafia na trudności rachunkowe. Dla uniknięcia tego mankamentu stosuje się przekształcenie zwane **standaryzacją**. Jeśli zmienna losowa ciągła X ma rozkład normalny $N(m,\sigma)$, to zmienna standaryzowana Z ma postać:

$$Z = \frac{X - m}{\sigma}. \tag{3.19}$$

Zmienna losowa Z ma standaryzowany rozkład normalny określony całkowicie przez dwa parametry: wartość oczekiwaną $E(Z) = 0$ oraz wariancję i odchylenie standardowe $D^2(Z) = D(Z) = 1$, co zapisujemy w skrócie: $N(0,1)$. Wartości funkcji gęstości prawdopodobieństwa i dystrybuanty rozkładu $N(0,1)$ są stablicowane.

Niech wzrost populacji mężczyzn ma rozkład normalny $N(172;6)$. Naszym zadaniem jest obliczenie prawdopodobieństwa tego, że wzrost przypadkowo wybranego mężczyzny będzie zawarty między 190 cm a 200 cm.

Jeśli zmienna losowa X ma rozkład $N(m,\sigma)$, to:

$$P(x_1 < X < x_2) = P(z_1 < Z < z_2) = F(z_2) - F(z_1).$$

Interesuje nas prawdopodobieństwo będące różnicą dystrybuant standaryzowanych wartości cechy. Wartości standaryzowanej zmiennej losowej są równe:

$$z_1 = \frac{190 - 172}{6} = 3 \text{ oraz } z_2 = \frac{200 - 172}{6} = 4{,}67.$$

Z tablic dystrybuanty rozkładu normalnego $N(0,1)$ odczytujemy: $F(z_1) = F(3) = 0{,}9987$ oraz $F(z_2) = F(4{,}67) = 1$. Prawdopodobieństwo tego, że wzrost przypadkowo wybranego mężczyzny będzie zawarty w przedziale 190 cm a 200 cm, wynosi:

$P(100 < X < 200) = P(z_1 < Z < z_2) = F(z_2) - F(z_1) = 1 - 0{,}9987 = 0{,}0013.$

Oznacza to, że 13 mężczyzn spośród 10 000 ma wzrost z interesującego nas przedziału.

Jeśli zmienna losowa ciągła X ma rozkład normalny, to ok. 68% jej realizacji mieści się w granicach jednego odchylenia standardowego od średniej, ok. 95% – w granicach dwóch odchyleń standardowych od średniej oraz ok. 99,73% – w granicach trzech odchyleń standardowych, gdyż:

$P(m - \sigma < X < m + \sigma) = P(-1 < Z < 1) = F(1) - F(-1) = 0{,}8413 - 0{,}1587 = 0{,}6826;$

$P(m - 2\sigma < X < m + 2\sigma) = P(-2 < Z < 2) = F(2) - F(-2) = 0{,}97725 - 0{,}2275 = 0{,}99545;$

$P(m - 3\sigma < X < m + 3\sigma) = P(-3 < Z < 3) = F(3) - F(-3) = 0{,}99865 - 0{,}00135 = 0{,}9973.$

Tak więc niemal wszystkie realizacje zmiennej losowej ciągłej zawierają się w granicach trzech odchyleń standardowych od średniej. W statystyce własność ta nosi nazwę **reguły trzech sigm**. Reguła ta jest często wykorzystywana do eliminacji obserwacji nietypowych. Za obserwacje nietypowe uznaje się te, których wartość różni się od średniej o więcej niż 3 odchylenia standardowe.

Rozkłady związane z rozkładem normalnym

We wnioskowaniu statystycznym ważne znaczenie mają również rozkłady: **chi-kwadrat** (χ^2), **t-Studenta** oraz **Fishera-Snedecora (F)**. Rozkłady te są związane z rozkładem normalnym.

Jeżeli zmienne X_i $(i = 1, 2, ..., n)$ są niezależnymi zmiennymi losowymi o jednakowych rozkładach $N(m, \delta)$, to zmienna losowa chi-kwadrat (χ^2) jest definiowana jako:

$$\chi^2 = \sum_{i=1}^{n} \left(\frac{X_i - m}{\sigma} \right)^2 = \sum_{i=1}^{n} Z_i^2 \qquad (3.20)$$

gdzie Z_i są niezależnymi zmiennym o rozkładzie $N(0,1)$.

Liczba niezależnych składników sumy (3.20) nosi nazwę **liczby stopni swobody**. W tym przypadku liczba stopni swobody v wynosi n $(v = n)$.

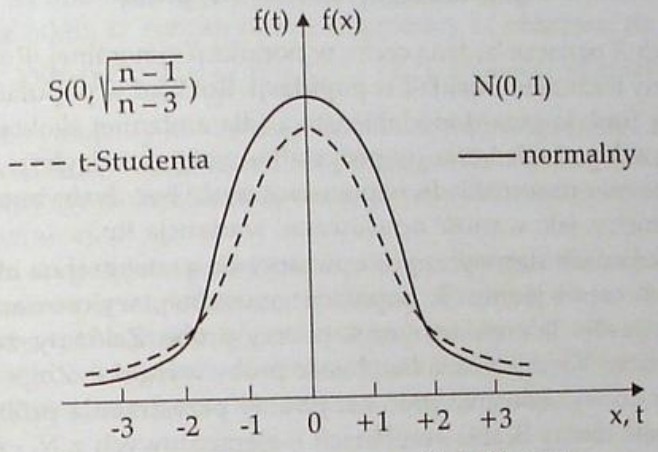
Jak wynika ze wzoru (3.20), zmienna losowa o rozkładzie χ^2 przyjmuje tylko nieujemne wartości. Kształt wykresu funkcji gęstości zależy jedynie od liczby stopni swobody. Rozkład χ^2 jest stablicowany, przy czym w tablicach podawane są zazwyczaj wartości χ^2 dla 30 (i mniej) stopni swobody. W przypadku, gdy liczba stopni swobody przekracza 30, korzysta się z rozkładu normalnego, który jest granicznym rozkładem χ^2. Jeśli bowiem liczba stopni swobody dąży do nieskończoności (praktycznie już wtedy, gdy liczba stopni swobody przekracza 30), to dystrybuanta zmiennej losowej $\sqrt{2\chi^2}$ zmierza do dystrybuanty rozkładu normalnego $N(\sqrt{2n-1};\ 1)$.

Rozkładem t-Studenta[1] nazywamy rozkład zmiennej losowej:

$$t = \frac{X}{Y}\sqrt{n}, \qquad (3.21)$$

gdzie: X jest zmienną losową o rozkładzie $N(0,1)$, Y ma rozkład χ^2 z $v = n$ stopniami swobody (zmienne X i Y są niezależne)[2].

Funkcję gęstości tego rozkładu określa tylko jeden parametr: liczba stopni swobody. W przypadku rozkładu t-Studenta liczba stopni swobody wynosi $n - 1$. Krzywa gęstości rozkładu t-Studenta jest podobna do wykresu gęstości rozkładu normalnego z tym, że jest nieco bardziej splaszczona (por. rys. 3.5).



Rys. 3.5. Rozkład normalny i rozkład t-Studenta

Przy wzroście liczby stopni swobody, rozkład t-Studenta jest zbieżny do rozkładu normalnego $N(0,1)$. Dla celów wnioskowania statystycznego, przybliżenie to można uznać za wystarczające już dla $v > 30$.

Rozkład t-Studenta jest stablicowany. Nie ma zatem potrzeby wykorzystywania dość skomplikowanych postaci analitycznych funkcji gęstości czy dystrybuanty tego rozkładu. Wartości krytyczne są odczytywane z tablic.

Rozkładem Fishera-Snedecora (rozkładem F) nazywamy rozkład zmiennej losowej:

$$F = \frac{U}{k_1} : \frac{V}{k_2} = \frac{U}{V} \cdot \frac{k_2}{k_1}, \qquad (3.22)$$

[1] Rozkład ten wprowadził brytyjski statystyk W. S. Gosset (1867–1937), który swoje prace podpisywał pseudonimem Student.

[2] Zmienną określoną wzorem (3.21) zapisujemy tradycyjnie małą literą, odstępując od zasady odróżniania zmiennej losowej i jej realizacji.

gdzie zmienne losowe U i V są niezależnymi zmiennymi losowymi o rozkładach χ^2 oraz odpowiednio k_1 i k_2 stopniach swobody. Rozkład F zależy od dwóch parametrów: k_1 i k_2, które są stopniami swobody związanymi z licznikiem (k_1) i mianownikiem (k_2) zmiennej losowej F.

Rozkład F jest stablicowany. W tablicach można odczytać wartości krytyczne dla różnej liczby stopni swobody licznika i mianownika oraz danego prawdopodobieństwa α. Dla różnych prawdopodobieństw α konstruowane są oddzielne tablice.

3.4. Rozkłady statystyk z próby

Niech X oznacza badaną cechę w populacji generalnej. Rozkład cechy X określamy terminem: **rozkład w populacji**. Rozkład w populacji może być opisywany funkcją prawdopodobieństwa (dla zmiennej skokowej) lub gęstością prawdopodobieństwa (w przypadku zmiennej ciągłej). Alternatywnym sposobem opisu rozkładu w populacji może być dystrybuanta, czy też takie parametry, jak: wartość oczekiwana, wariancja itp.

W badaniach statystycznych opieramy się zazwyczaj na informacjach dotyczących części jednostek populacji generalnej, wylosowanych w odpowiedni sposób. Ta część jednostek tworzy próbę. Załóżmy, że populacja generalna liczy N jednostek, a liczebność próby wynosi n. Zbiór wszystkich możliwych do wylosowania prób nazywamy **przestrzenią prób losowych**. Zbiór ten jest równy liczbie kombinacji n-elementowych z N, czyli $\binom{N}{n}$.

Niech ciąg o postaci: $x_1, x_2, \ldots, x_n$, oznacza wartości cechy X u wylosowanych do n-elementowej próby jednostek zbiorowości. Jeśli losowanie do próby będzie powtarzane, to za każdym razem otrzymamy inny zbiór wartości: $x_1, x_2, \ldots, x_n$. Oznaczmy przez $\{x_1\}$ zbiór wartości otrzymanych w pierwszym losowaniu do próby, przez $\{x_2\}$ – w drugim losowaniu, ..., przez $\{x_n\}$ – w n-tym losowaniu. Zbiór wartości $\{x_1\}$ można traktować jako realizację zmiennej losowej X_1, zbiór $\{x_2\}$ – jako realizację zmiennej X_2 itd. W każdym losowaniu wybierane są elementy z całej populacji (losowanie ze zwracaniem), każda ze zmiennych losowych $X_1, X_2, \ldots, X_n$ ma taki sam rozkład jak rozkład X w populacji. Z tych też względów zmienne losowe $X_1, X_2, \ldots, X_n$ są niezależne.

Przy wnioskowaniu o parametrach populacji generalnej na podstawie próby losowej posługujemy się funkcjami zmiennych tworzących próbę: $X_1, X_2, \ldots, X_n$. Funkcje te nazywamy **statystykami z próby. Statystyką z próby** nazywamy zmienną losową, będącą funkcją obserwowanych w próbie zmiennych losowych, określoną na przestrzeni prób, czyli:

$$U = f(X_1, X_2, \ldots, X_n). \tag{3.23}$$

Przykładami statystyki z próby są:

$$\overline{X} = \frac{1}{n}\sum_{i=1}^{n} X_i, \tag{3.24}$$

$$S^2 = \frac{1}{n}\sum_{i=1}^{n}\left(X_i - \overline{X}\right)^2, \tag{3.25}$$

czyli średnia arytmetyczna i wariancja.

Omówione w rozdziale drugim parametry: średnia arytmetyczna i wariancja, oznaczane były odpowiednio symbolami: $\bar{x}$ i s^2. Wielkości te były liczbami obliczonymi z n wartości. Statystyki $\overline{X}$ oraz S^2 są natomiast estymatorami parametru m w populacji generalnej oraz wariancji w populacji generalnej. Są one zmiennymi losowymi, posiadającymi określony rozkład. Wynika to z faktu, iż rozpatrywane estymatory są obliczane na podstawie zmiennych losowych: $\overline{X} = \frac{1}{n}(X_1 + X_2 + \ldots + X_n)$. Stosując schemat losowania prostego, otrzymujemy konkretną, n-elementową próbę: $x_1, x_2, \ldots, x_n$, gdzie x_1 jest realizacją X_1, x_2 – realizacją X_2 itd. Średnią arytmetyczną obliczoną z próby oznaczamy symbolem $\bar{x}$ i nazywamy **oceną** średniej arytmetycznej w populacji generalnej.

Każda **statystyka z próby** jako funkcja zmiennych losowych, sama jest zmienną losową o określonym rozkładzie. Rozkład ten nazywamy **rozkładem statystyki z próby**. W praktycznych zastosowaniach wnioskowania statystycznego szczególnie często korzysta się z rozkładów dotyczących średniej arytmetycznej z próby oraz wariancji z próby.

Rozkłady statystyk związane ze średnią w próbie

Niech zmienna losowa X ma w populacji generalnej rozkład $N(m,\sigma)$. Z populacji tej pobieramy n-elementową próbę losową prostą $(X_1, X_2, \ldots, X_n)$. Średnia arytmetyczna z próby jest określona wzorem (3.24). Korzystając z własności wartości oczekiwanej i wariancji mamy:

$$E\left(\overline{X}\right) = E\left(\frac{1}{n}\sum_{i=1}^{n} X_i\right) = \frac{1}{n}\sum_{i=1}^{n} E(X_i) = \frac{1}{n} nm = m$$

oraz

$$D^2\left(\overline{X}\right) = D^2\left(\frac{1}{n}\sum_{i=1}^{n} X_i\right) = \frac{1}{n^2}\sum_{i=1}^{n} D^2(X_i) = \frac{1}{n^2} n\sigma^2 = \frac{\sigma^2}{n},$$

$$D\left(\overline{X}\right) = \frac{\sigma}{\sqrt{n}}.$$

Tak więc wartość oczekiwana średniej z próby jest równa wartości oczekiwanej zmiennej losowej w populacji generalnej. Odchylenie standardowe średniej arytmetycznej z próby, zwane standardowym błędem oceny, jest natomiast wprost proporcjonalne do odchylenia standardowego zmiennej losowej w populacji i odwrotnie proporcjonalne do pierwiastka

kwadratowego z liczebności próby $\left(\overline{X}:N\left(m,\frac{\sigma}{\sqrt{n}}\right)\right)$. $D(\overline{X})$ informuje o wielkości błędu popełnionego przy szacowaniu średniej w populacji za pomocą średniej z n-elementowej próby. Standaryzując zmienną losową $\overline{X}$ otrzymujemy:

$$Z=\frac{\overline{X}-m}{\sigma}\sqrt{n}. \tag{3.26}$$

Jeśli odchylenie standardowe σ w populacji nie jest znane, wówczas wprowadzamy oszacowania S lub $\hat{S}$, tj. odchylenia standardowe z próby:

$$S=\sqrt{\frac{1}{n}\sum_{i=1}^{n}\left(X_i-\overline{X}\right)^2}, \tag{3.27}$$

$$\hat{S}=\sqrt{\frac{1}{n-1}\sum_{i=1}^{n}\left(X_i-\overline{X}\right)^2}. \tag{3.28}$$

Wielkości S oraz $\hat{S}$ są – podobnie jak $\overline{X}$ – zmiennymi losowymi. Jeżeli $X_1, X_2, ..., X_n$ są niezależnymi zmiennymi losowymi o jednakowym rozkładzie normalnym $N(m,\sigma)$, to statystyka

$$t=\frac{\overline{X}-m}{S}\sqrt{n-1}=\frac{\overline{X}-m}{\hat{S}}\sqrt{n} \tag{3.29}$$

ma rozkład t-Studenta z $n-1$ stopniami swobody.

Badania statystyczne nie zawsze dotyczą jednej populacji. W analizach porównawczych zachodzi konieczność porównywania dwóch lub więcej średnich. W przypadku dwóch populacji korzystamy z następującego twierdzenia[3]: jeżeli $X_1, X_2, ..., X_{n1}$ oraz $Y_1, Y_2, ..., Y_{n2}$ są niezależnymi od siebie ciągami zmiennych losowych o jednakowych rozkładach normalnych $N(m,\sigma)$, to zmienna losowa:

$$t=\frac{\overline{X}-\overline{Y}}{\sqrt{n_1S_1^2+n_2S_2^2}}\sqrt{\frac{(n_1+n_2-2)n_1n_2}{n_1+n_2}} \tag{3.30}$$

ma rozkład t-Studenta z n_1+n_2-2 stopniami swobody.

Statystyka (3.30) jest wykorzystywana do wnioskowania o różnicy średnich arytmetycznych z dwóch prób, gdy odchylenia standardowe populacji nie są znane (zakłada się jednak, że są one równe). Jeśli natomiast odchylenia standardowe w populacjach generalnych są znane, wówczas do wnioskowania o różnicy średnich z dwóch prób stosujemy wzór:

[3] A. Balicki, W. Makać, *Metody wnioskowania statystycznego*, Wydawnictwo Uniwersytetu Gdańskiego, Gdańsk 1997, s. 102.

$$Z=\frac{\overline{X}-\overline{Y}}{\sqrt{\frac{\sigma_x^2}{n_1}+\frac{\sigma_y^2}{n_2}}}, \tag{3.31}$$

gdzie σ_1^2 oraz σ_2^2 są odpowiednio odchyleniami standardowymi zmiennych w obu populacjach. Statystyka (3.31) ma rozkład normalny $N(0,1)$.

Rozkłady statystyk związane z wariancją w próbie

Wariancja jest podstawową miarą zmienności. We wnioskowaniu statystycznym rozkłady związane z wariancją pełnią więc istotną rolę.

Jeżeli $X_1, X_2, ..., X_n$ są niezależnymi zmiennymi losowymi o jednakowym rozkładzie normalnym $N(m,\sigma)$, to zmienna losowa:

$$U=\frac{nS^2}{\sigma^2}=\frac{n\hat{S}^2}{n-1}=\sum_{i=1}^{n}\left(X_1-\overline{X}\right) \tag{3.31}$$

ma rozkład chi-kwadrat (χ^2) z $n-1$ stopniami swobody.

Jeżeli natomiast $X_1, X_2, ..., X_{n1}$ oraz $Y_1, Y_2, ..., Y_{n2}$ są niezależnymi zmiennymi losowymi o rozkładach normalnych $N(m_1,\sigma)$ i $N(m_2,\sigma)$, to statystyka:

$$F=\frac{\hat{S}_1^2}{\hat{S}_2^2} \tag{3.32}$$

ma rozkład Fishera-Snedecora z n_1-1 (stopnie swobody licznika) i n_2-1 (stopnie swobody mianownika) stopniami swobody. Wykorzystując wzór (3.32) należy zwrócić uwagę na równość odchyleń standardowych w obu populacjach generalnych. Odchylenia te są wprawdzie nieznane, ale można sprawdzić ich równość wykorzystując odpowiedni test statystyczny. Problem ten będzie przedstawiony w rozdziale dotyczącym weryfikacji hipotez.

ZADANIA

3.1. Dane są możliwe realizacje dyskretnej zmiennej losowej X: $x_1 = -1$; $x_2 = 0$; $x_3 = 1$ oraz wartości oczekiwane tej zmiennej i jej kwadratu: $E(X) = 0{,}1$; $E(X^2) = 0{,}9$. Zapisać w postaci tabelarycznej rozkład prawdopodobieństwa zmiennej losowej X.

3.2. Zmienne losowe X i Y są niezależne. Znaleźć wariancję zmiennej losowej $Z = 3X + 2Y$, jeśli wiadomo, że $D^2(X) = 5$ i $D^2(Y) = 6$.

3.3. Wiadomo, że $E(X) = 10$ oraz $D(X) = 2$. Obliczyć $E(Y)$ oraz $D(Y)$, jeśli wiadomo, że $Y = 2X + 5$.

3.4. Dany jest rozkład prawdopodobieństwa zmiennej losowej skokowej X:

x_i	10	20	30	40	50
p_i	0,1	0,2	0,3	0,4	0,5

Obliczyć wartość oczekiwaną i wariancję tej zmiennej.

3.5. Znaleźć wartość oczekiwaną i wariancję zmiennej losowej X^4, jeśli rozkład zmiennej losowej X jest następujący:

x_i	–1	0	1	2
p_i	0,2	0,2	?	0,1

3.6. Zmienna losowa X ma rozkład:

x_i	0	1	2	3
p_i	0,25	?	0,5	0,1

Znaleźć funkcję rozkładu prawdopodobieństwa zmiennej losowej $Y = 2X - 1$.

3.7. Rozkład prawdopodobieństwa zmiennej losowej skokowej X przedstawia się następująco:

x_i	0	1	2	3
p_i	$\frac{1}{8}$	$\frac{3}{8}$	$\frac{3}{8}$	$\frac{1}{8}$

Wyznacz dystrybuantę zmiennej losowej X.

3.8. Dana jest dystrybuanta zmiennej losowej skokowej X:

$$f(x)=\begin{cases}0 & \text{dla} \quad x \le -100\\ \frac{2}{3} & \text{dla} \quad -100 < x \le 100\\ 1 & \text{dla} \quad x > 100\end{cases}$$

Wyznacz rozkład prawdopodobieństwa tej zmiennej.

3.9. Dana jest dystrybuanta zmiennej losowej skokowej X:

$$F(x)=\begin{cases}0 & \text{dla} \quad x \le -1\\ 0{,}1 & \text{dla} \quad -1 < x \le 0\\ 0{,}6 & \text{dla} \quad 0 < x \le 2\\ 1 & \text{dla} \quad x > 2\end{cases}$$

Obliczyć wartość oczekiwaną i wariancję zmiennej losowej X.

3.10. Dyskretna zmienna losowa ma rozkład prawdopodobieństwa określony następująco: $\{(1;p_1), (2;p_2), (3;p_3)\}$. Wiadomo, że $E(X) = 2{,}3$; $D^2(X) = 0{,}61$. Jakie są wartości p_1, p_2 i p_3?

3.11. Skokowa zmienna losowa X przyjmuje dwie wartości: x_1 i x_2, przy czym $(x_1 < x_2)$. Prawdopodobieństwo tego, że zmienna losowa X przyjmie wartość x_1, wynosi 0,6. Znaleźć rozkład prawdopodobieństwa zmiennej losowej X, jeśli $E(X) = 1{,}4$; $D^2(X) = 0{,}0576$.

3.12. Zmienna losowa X ma następujący rozkład prawdopodobieństwa:

x_i	1	?	5	10	20
p_i	0,5	0,25	0,1	0,1	?

Uzupełnij znaki zapytania liczbami, jeśli $E(X^2) = 37$. Wiadomo, że $x_1 < x_2 < x_3 < x_4 < x_5$. Obliczyć odchylenie standardowe tej zmiennej.

3.13. Zmienna losowa skokowa X przyjmuje następujące wartości: 1, 2, 3, 4, 5 i 6 z jednakowymi prawdopodobieństwami $p = \frac{1}{6}$. Sporządź wykres dystrybuanty tej zmiennej oraz wyznacz odchylenie standardowe.

3.14. Zmienna losowa X przyjmuje wartości 0, 1, 2 i 3, każdą z prawdopodobieństwem 0,25. Oblicz prawdopodobieństwa: $P(X \le 2)$ oraz $P(X > 1)$.

3.15. Zmienna losowa skokowa X ma następujący rozkład: $\{(-1;0), (0;\frac{1}{3}), (1;\frac{1}{3})\}$. Wyznaczyć rozkład prawdopodobieństwa zmiennej losowej $Y = X^2$. Obliczyć $D^2(Y)$ oraz $P(|X| < 0{,}5)$.

3.16. Z bieżącej produkcji pobrano losowo 5 elementów. Wadliwość produkcji wynosi 0,1. Znaleźć rozkład prawdopodobieństwa zmiennej losowej X definiowanej jako liczba sztuk wadliwych. Obliczyć $E(X)$.

3.17. W serii wyprodukowanych detali znajduje się 10% niestandardowych. W losowaniu zwrotnym wybrano 4 detale. Znaleźć rozkład prawdopodobieństwa i dystrybuantę liczby niestandardowych detali wśród czterech wybranych.

3.18. Prawdopodobieństwo tego, że zmienna losowa X przyjmie wartości 1, 2, 3, 4, dane jest funkcją: $P(X = x) = \frac{1}{30}x^2$. Sprawdzić, czy jest to funkcja rozkładu prawdopodobieństwa. Jeśli tak, to znaleźć prawdopodobieństwo tego, że zmienna losowa przyjmie wartość: a) mniejszą od 4, b) większą od 2, c) mniejszą niż 5 i większą niż 2.

3.19. Prawdopodobieństwo tego, że student jest przygotowany do ćwiczeń, wynosi $\frac{2}{3}$. Prowadzący zajęcia wywołuje 4 studentów. Znaleźć $P(X = 3)$, gdzie X jest zmienną losową charakteryzującą liczbę osób nieprzygotowanych do ćwiczeń.

3.20. Niech X będzie zmienną losową skokową o rozkładzie spełniającym warunki: $P(X = 2) = \frac{1}{3}$; $P(X = 3) = 0{,}5$; $P(X = 4) = \frac{1}{6}$. Znaleźć dystrybuantę zmiennej losowej X.

3.21. Niech X oznacza zmienną losową określającą liczbę oczek w rzucie kostką do gry. Znaleźć rozkład prawdopodobieństwa zmiennej losowej $Y = X^2 - 7X + 10$.

3.22. Wyznaczyć dystrybuantę zmiennej losowej o rozkładzie prawdopodobieństwa danym funkcją: $P(X = x) = \frac{x}{3}$ dla $x \in \{1,2,\ldots,5\}$.

3.23. Niech $F(x)$ będzie dystrybuantą zadaną wzorem:

$$F(x)=\begin{cases}0 & \text{dla} \quad x \le 1\\ \frac{1}{3} & \text{dla} \quad 1 < x \le 4\\ \frac{1}{2} & \text{dla} \quad 4 < x \le 6\\ \frac{5}{6} & \text{dla} \quad 6 < x \le 10\\ 1 & \text{dla} \quad x > 10\end{cases}$$

Obliczyć: a) $P(2 \le x < 6)$, b) $P(X = 4)$, c) $P(5 < X < \infty)$.

3.24. Prawdopodobieństwo wylosowania złej puszki z partii wynosi 0,3. Zakupujemy 5 puszek pochodzących z tej partii. Niech X oznacza zmienną losową ok-

reślającą liczbę złych puszek. Znaleźć rozkład prawdopodobieństwa zmiennej losowej X.

3.25. Zmienna losowa X przyjmuje trzy wartości: $x_1 = 3$, $x_2 = 5$ oraz x_3 odpowiednio z prawdopodobieństwami: p; 0,3 i 0,2. Wyznacz x_3 i p, jeśli $E(X) = 5$. Oblicz wariancję zmiennej losowej X.

3.26. Zmienna losowa skokowa X przyjmuje trzy wartości: 0, 1 i 2. Wiadomo, że $E(X) = 1$ oraz $E(X^2) = 1{,}5$. Wyznacz rozkład prawdopodobieństwa zmiennej losowej X.

3.27. Zmienna losowa X przyjmuje z jednakowym prawdopodobieństwem wartości równe trzem kolejnym wyrazom ciągu geometrycznego o pierwszym wyrazie 3 i ilorazie q. Oblicz q, jeśli $E(X) = 7$.

3.28. Wyznacz rozkład prawdopodobieństwa zmiennej losowej X, mając jej dystrybuantę o postaci:

$$F(x)=\begin{cases} 0 & \text{dla} \quad x \le 1 \\ 0{,}2 & \text{dla} \quad 1 < x \le 2 \\ 0{,}6 & \text{dla} \quad 2 < x \le 5 \\ 1 & \text{dla} \quad x > 5 \end{cases}.$$

3.29. Zmienna losowa X ma następujący rozkład prawdopodobieństwa:

x_i	-1	2	5
p_i	$\frac{2}{7}$	$\frac{4}{7}$	$\frac{1}{7}$

Wyznacz dystrybuantę tej zmiennej.

3.30. Zmienna losowa X ma rozkład dwumianowy z wartością oczekiwaną 12 i wariancją 3. Znajdź n i p.

3.31. W partii składającej się z 6 części 4 są wadliwe. Wylosowano 3 części. Zapisz rozkład prawdopodobieństwa liczby wadliwych części wśród wylosowanych oraz wykreśl dystrybuantę tej zmiennej losowej.

3.32. Liczba dni roboczych o normalnym zużyciu wody w pewnej firmie produkcyjnej jest zmienną losową. Wiadomo, że prawdopodobieństwo zaobserwowania normalnego dziennego zużycia wody wynosi 0,75. Oblicz wartość oczekiwaną i wariancję zużycia wody w okresie tygodnia ($n = 5$ dni).

3.33. Oblicz prawdopodobieństwo tego, że wśród 100 elementów znajdują się co najmniej 4 wadliwe, jeśli wadliwość elementów w tej partii wynosi 2%.

3.34. Po mieście jeździ 10 000 samochodów. Prawdopodobieństwo wezwania pogotowia technicznego w ciągu doby przez samochód wynosi 0,0001. Oblicz prawdopodobieństwo tego, że w ciągu losowo wybranej doby pogotowie będzie wzywane: a) najwyżej 1 raz, b) co najmniej 1 raz.

3.35. Zmienna losowa X ma rozkład dwumianowy o parametrach: $E(X) = 4{,}8$ oraz $D^2(X) = 2{,}88$. Oblicz: a) $P(X = 0)$, b) $P(X > 0)$, c) $P(X < 4)$.

3.36. Zmienna losowa X ma rozkład Poissona o wartości oczekiwanej $E(X) = 3$. Oblicz: a) $P(X = 0)$, b) $P(X \ge 1)$, c) $F(4)$, d) $P(0 < X \le 3)$.

3.37. Pomyłka przy prowadzeniu rachunków klientów w pewnym banku nie powinna pojawić się częściej niż raz na tysiąc. Jakie jest prawdopodobieństwo tego, że w losowo wybranych 500 rachunkach wykryje się co najwyżej 1 błąd?

3.38. Niech zmienna losowa ciągła X ma rozkład normalny $N(4;2)$, a zmienna losowa Y – rozkład normalny $N(3;1)$. Oblicz: a) $E(2X + 3Y)$; b) $D^2(2X - 3Y + 5)$; c) $D(2X - 3Y)$.

3.39. Jeśli mańkuci stanowią przeciętnie 1% ludności, to jakie jest prawdopodobieństwo, że wśród 200 wylosowanych osób znajdzie się co najwyżej 2 mańkutów?

3.40. Zmienna losowa skokowa X ma następujący rozkład prawdopodobieństwa: $\left\{\left(-1;\frac{2}{3}\right);\left(0;\frac{1}{6}\right);\left(1;\frac{1}{6}\right)\right\}$. Wyznaczyć rozkład prawdopodobieństwa zmiennej losowej $Y = X^3$. Obliczyć: a) $D^2(Y)$, b) $P(|X| < 1{,}5)$.

3.41. Prawdopodobieństwo występowania braków przy produkcji narzędzi rolniczych wynosi 0,02. Jakie jest prawdopodobieństwo tego, że w partii liczącej 200 narzędzi będzie nie więcej niż dwa narzędzia wybrakowane?

3.42. W rodzinie jest dziesięcioro dzieci. Przyjmując, że prawdopodobieństwo urodzenia się chłopca i dziewczynki jest takie samo, obliczyć, że w rodzinie tej jest: a) pięciu chłopców, b) nie mniej niż trzech i nie więcej niż ośmiu chłopców.

3.43. W czasie transportu psuje się 3% skrzynek owoców. Wylosowano 4 skrzynki. Niech X będzie zmienną losową określającą wylosowane skrzynki z zepsutymi owocami. Znaleźć rozkład prawdopodobieństwa zmiennej losowej X.

3.44. Zmienna losowa X jest określona dystrybuantą o postaci:

$$F(x)=\begin{cases} 0 & \text{dla} \quad x \le 1 \\ 0{,}15 & \text{dla} \quad 1 < x \le 3 \\ 0{,}25 & \text{dla} \quad 3 < x \le 5 \\ 0{,}50 & \text{dla} \quad 5 < x \le 7 \\ 0{,}80 & \text{dla} \quad 7 < x \le 10 \\ 1 & \text{dla} \quad x > 10 \end{cases}.$$

Znaleźć $E(X)$ oraz $D^2(X)$.

3.45. Zakład ubezpieczeniowy ocenia, że corocznie w wypadkach drogowych ginie 1% ubezpieczonych kierowców. Obliczyć prawdopodobieństwo tego, że w ciągu roku zakład będzie musiał wypłacić odszkodowanie więcej niż trzy razy, jeśli ubezpieczyło się w tym zakładzie 100 kierowców.

3.46. W pewnym mieście wylosowano 400 rodzin i zbadano je pod względem liczby osób w rodzinie. Otrzymano następujące informacje:

Liczba rodzin	20	120	160	60	28	12
Liczba osób w rodzinie	1	2	3	4	5	6

Niech X oznacza zmienną losową: liczba osób w rodzinie. Podać rozkład prawdopodobieństwa i dystrybuantę tej zmiennej. Oblicz prawdopodobieństwo tego, że wybrana losowo rodzina liczy nie mniej niż 5 osób.

3.47. Przypuszcza się, że 40% mieszkańców kraju korzysta z kart kredytowych. Jakie jest prawdopodobieństwo tego, że wśród 10 klientów dokonujących zakupów pięciu z nich będzie płaciło kartą kredytową?

3.48. Stwierdzono, że 2 promile książek dla gimnazjów ma uszkodzone oprawy. Jaka jest wartość oczekiwana i odchylenie standardowe liczby książek uszkodzonych wśród 500 wylosowanych?

3.49. Znajdź taką wartość stałej a, aby funkcja $f(x) = ax^3$ dla $0 \le x \le 1$ była funkcją gęstości zmiennej losowej ciągłej X.

3.50. Funkcja gęstości prawdopodobieństwa zmiennej losowej ciągłej X jest określona następująco:

$$f(x)=\begin{cases}0{,}5 & \text{dla} \quad 0 \le x \le 2\\ 0 & \text{dla pozostałych} \quad x\end{cases}.$$

Znajdź dystrybuantę tej zmiennej i oblicz $P(X = 2)$.

3.51. Znaleźć gęstość prawdopodobieństwa zmiennej losowej ciągłej X, jeśli ma ona rozkład normalny o parametrach: $E(X) = 3$; $D^2(X) = 16$.

3.52. Oblicz wariancję i odchylenie standardowe zmiennej losowej ciągłej X o gęstości: $f(x)=\frac{1}{9}x^2$ dla $0 \le x \le 3$ i 0 dla pozostałych wartości x.

3.53. Zmienna losowa ciągła X ma funkcję gęstości o postaci:

$f(x)=\frac{1}{2\sqrt{2\pi}}e^{\frac{(x-3)^2}{8}}$; $x \in R$. Oblicz $P(|X|>2)$.

3.54. Dla jakiej wartości C funkcja:

$$f(x)=\begin{cases}0{,}75x(2-x) & \text{dla} \quad 0 \le x \le C\\ 0 & \text{dla pozostałych} \quad x\end{cases}$$

jest funkcją gęstości? Oblicz $E(X)$ oraz $D(X)$.

3.55. Funkcja gęstości zmiennej losowej ciągłej X jest określona wzorem:

$$f(x)=\begin{cases}0 & \text{dla} \quad x \le 1\\ 0{,}5 & \text{dla} \quad 1 < x \le 3\\ 0 & \text{dla} \quad x > 3\end{cases}.$$

Zapisz dystrybuantę tej zmiennej. Oblicz: a) $P(X < 2)$; b) $P(X > 2{,}5)$; c) $F(1{,}5)$.

3.56. Zmienna losowa ciągła X ma rozkład normalny $N(165;15)$. Wiadomo, że $P(X > x) = 0{,}02018$. Dla jakiej wartości x prawdziwa jest ta równość?

3.57. Wiadomo, że wiek kobiet w chwili urodzenia dziecka ma rozkład $N(26{,}9;5{,}5)$. Oblicz prawdopodobieństwo tego, że dziecko urodzi kobieta w wieku poniżej 30 lat.

3.58. Zaplanowane w firmie podwyżki wynagrodzeń mają rozkład $N(10\%;5\%)$. Ilu spośród 200 zatrudnionych pracowników otrzyma ponad 15% podwyżki?

3.59. Rozkład wyników testu psychologicznego jest rozkładem $N(80;10)$. Ilu spośród 500 studentów uzyskało co najmniej 100 punktów?

3.60. Zmienna losowa ciągła X ma rozkład $N(0,1)$. Oblicz: a) $P(X > 0)$, b) $P(X > 2)$, c) $P(-1 < X < 0)$, d) $P(|X| < 2)$, e) $P(|X| > 1)$, f) $P(-1 < X < 3)$.

3.61. Wskaźnik rentowności w przedsiębiorstwach przemysłowych w województwie L ma rozkład $N(3{,}5\%;1{,}5\%)$. Oblicz prawdopodobieństwo tego, że w losowo wybranym przedsiębiorstwie rentowność nie jest większa niż 4%.

3.62. Sprawdzić, czy funkcja:

$$f(x)=\begin{cases}0 & \text{dla} \quad x < 0 \text{ i } x > \sqrt{6}\\ \frac{1}{3}x & \text{dla} \quad 0 \le x \le \sqrt{6}\end{cases}$$

jest funkcją gęstości zmiennej losowej ciągłej X. Jeśli tak, to obliczyć $P(|X|<1)$ oraz podać analityczną postać dystrybuanty.

3.63. Znaleźć wartość oczekiwaną zmiennej losowej ciągłej X, która jest określona przez dystrybuantę:

$$F(x)=\begin{cases}0 & \text{dla} \quad x \le 0\\ 0{,}25x & \text{dla} \quad 0 < x \le 4\\ 1 & \text{dla} \quad x > 4\end{cases}.$$

3.64. Zmienna losowa ciągła X ma rozkład $N(4;9)$. Obliczyć: a) $P(X > 13)$, b) $P(|X-2|<14)$, c) $P(|X-3|>11)$.

3.65. Staż pracy pracowników ma rozkład $N(11{,}4;5{,}2)$. Jaki odsetek pracowników ma staż od 5 do 8 lat?

3.66. Zmienna losowa ciągła X ma rozkład $N(2;4)$. Wyznaczyć wartość a tak, by $P(|X-2|<4a)=0{,}8$.

3.67. Zmienna losowa ciągła X ma gęstość: $f(x)=\frac{1}{\sqrt{2\pi}}e^{\frac{-(x-1)^2}{2}}$. Obliczyć $P(X < 3)$.

3.68. Zmienna losowa ciągła X podlega rozkładowi zgodnemu z funkcją gęstości:

$$f(x)=\begin{cases}0 & \text{dla} \quad x < 0\\ Cx & \text{dla} \quad 0 \le x \le 4\\ 0 & \text{dla} \quad x > 4\end{cases}.$$

Wyznaczyć stałą C, podać dystrybuantę zmiennej losowej X oraz obliczyć $P(1 \le X \le 2)$.

3.69. Zmienna losowa ciągła X jest określona za pomocą dystrybuanty:

$$F(x)=\begin{cases}1-8x^{-3} & \text{dla} \quad x \ge 2\\ 0 & \text{dla} \quad x < 2\end{cases}.$$

Wyznacz $E(X)$ i $D(X)$.

3.70. Niech X ma rozkład $N(0;1)$. Obliczyć: a) $P(0 < X < 2)$, b) $P(X > 2)$, c) $P(X < -0{,}5)$, d) $P(|X| < 1)$.

3.71. Obliczyć $P(|X| > 2)$ dla zmiennej losowej ciągłej X o gęstości:

$$f(x)=\frac{1}{2\sqrt{2\pi}}e^{-\frac{1}{8}(x+1)^2}.$$

3.72. Wyznaczyć stałą C tak, aby funkcja $f(x)$ była funkcją gęstości prawdopodobieństwa zmiennej losowej ciągłej X:

$$f(x)=\begin{cases}Cx & \text{dla} \quad 0 \le x \le 3\\ 0 & \text{dla pozostałych} \quad x\end{cases}.$$

Podać dystrybuantę rozpatrywanej zmiennej.

3.73. Zmienna losowa ciągła X przyjmuje wartości z przedziału $(0,1)$. Funkcja gęstości prawdopodobieństwa w tym przedziale wynosi: $f(x)=\frac{1}{5}x^3$. Obliczyć $E(X)$.

3.74. Znaleźć $E(X)$ oraz $D^2(X)$ zmiennej losowej ciągłej X, jeśli jej dystrybuanta dana jest wzorem:

$$F(x)=\begin{cases}0 & \text{dla} \quad x\le 0\\ x & \text{dla} \quad 0<x\le 1\\ 1 & \text{dla} \quad x>1\end{cases}$$

3.75. Dla jakiej wartości C funkcja:

$$f(x)=\begin{cases}Cx^2 & \text{dla} \quad 0\le x\le 3\\ 0 & \text{dla pozostałych} \quad x\end{cases}$$

jest gęstością prawdopodobieństwa? Oblicz $P(1\le X\le 4)$.

3.76. Znaleźć gęstość prawdopodobieństwa zmiennej losowej ciągłej X, która ma rozkład normalny o parametrach: $E(X)=3{,}6$; $D^2(X)=16$.

3.77. Zmienna losowa X ma rozkład $N(10;2)$. Oblicz $P(15<X<25)$.

3.78. Zmienna losowa X ma rozkład $N(1,2)$. Oblicz $P(|X|>3)$.

3.79. Rozkład odległości sobotnio-niedzielnych wyjazdów turystów jest normalny z wartością oczekiwaną 20 km i wariancją 16 $(\text{km})^2$. Jaki odsetek turystów wyjeżdża na odległość przekraczającą 30 km?

3.80. Zmienna losowa ciągła X może przybierać wartości z przedziału $(0,4)$, a jej funkcja gęstości w tym przedziale dana jest wzorem: $f(x)=ax$. Wyznaczyć wartość parametru a oraz obliczyć prawdopodobieństwo tego, że $X<2$.

3.81. Obliczyć:

a) $P(0<X<6)$ jeśli X jest zmienną losową o rozkładzie $N(4,2)$,

b) $P(|X|<3)$ jeśli X jest zmienną losową o rozkładzie $N(1,2)$.

3.82. Waga produkowanych wyrobów jest zmienną losową X o rozkładzie $N(100\text{ kg}, 5\text{ kg})$. Ile – przeciętnie – wyrobów spośród tysiąca waży mniej niż 90 kg?

3.83. Zmienna losowa X ma rozkład $N(5,3)$. Obliczyć: a) $P(X>6{,}5)$, b) s jeśli $P(X>s)=0{,}08$.

3.84. Zmienna losowa X ma rozkład normalny $N(0,1)$. Obliczyć: a) $P(X>2)$, b) $P(|X|>2)$.

3.85. Wzrost studentów pewnej uczelni ma rozkład $N(176,10)$. Ilu spośród 2000 studentów ma wzrost niższy od 178 cm?

3.86. Zmienna losowa X ma rozkład $N(0,1)$. Znaleźć wartość z, jeśli: a) $P(X<z)=0{,}95$, b) $P(|X|<z)=0{,}95$.

3.87. Ilość gazu w butli jest zmienną losową o rozkładzie normalnym ze średnią 12 jednostek objętości i odchyleniem standardowym 2 jednostek. Znaleźć taką ilość gazu x, aby 14,92% butli zawierało więcej niż x gazu.

3.88. Zmienna losowa ciągła X ma rozkład $N(97;10)$. Znaleźć takie x, aby $P(102<X<x)=0{,}05$.

3.89. Niech X będzie zmienną losową o rozkładzie normalnym $N(600;100)$. Znajdź takie dwie wartości x_1 i x_2, aby zachodziły równości: $P(X>x_1)=0{,}01$ oraz $P(X<x_2)=0{,}05$.

3.90. Stwierdzono, że wartość kaloryczna porcji napoju dietetycznego ma rozkład normalny ze średnią 5 kalorii i odchyleniem standardowym 0,5 kalorii. Jaki odsetek tego napoju spośród 100 000 porcji ma więcej niż 6 kalorii?

3.91. W pewnej kasie SKOK stwierdzono, że wysokość udzielonych w ciągu ostatniego miesiąca pożyczek ma rozkład normalny ze średnią 280 zł i wariancją 400 $(\text{zł})^2$. Jaki odsetek klientów tej kasy zaciąga pożyczki przekraczające 300 zł, a jaki procent pożyczki nieprzekraczające 260 zł?

3.92. Producent golarek elektrycznych marki X gwarantuje, że ich czas użytkowania ma rozkład $N(5$ lat, 10 miesięcy$)$. Jaki odsetek produkowanych golarek będzie użytkowanych po okresie gwarancji wynoszącym 3 lata?

3.93. Stwierdzono, że iloraz inteligencji (IQ) ma rozkład normalny o wartości oczekiwanej 100 i wariancji 225. Wyznaczyć wartość IQ, której nie przekracza 70% badanej populacji.

3.94. Zgodnie z planem, czas lotu z Warszawy do Frankfurtu ma rozkład $N(90;2)$. Oblicz długość trwania lotu, która nie jest przekraczana w 85% przelotów na tej trasie.

3.95. Waga pewnej grupy osób opisana jest rozkładem normalnym o wartości średniej 75 kg i wariancji 16 $(\text{kg})^2$. Jaki odsetek osób waży więcej niż 83 kg?

3.96. Ciężar jajek dostarczonych do skupu ma rozkład normalny ze średnią 2 dag i wariancją 0,01 $(\text{dag})^2$. Jajko kwalifikuje się do I gatunku, jeżeli jego waga wynosi co najmniej 2,096 dag. Jaki procent jajek dostarczanych do skupu można uznać za jajka I gatunku?

3.97. Średni czas żarzenia się liści tytoniu wynosi 17 sekund. Liście tlące się krócej niż 12 sekund są dyskwalifikowane. Jaki jest procent liści przydatnych do produkcji, jeśli wariancja analizowanej cechy wynosi 6,25 $(\text{sek})^2$?

3.98. Aby zdać egzamin ze statystyki, należy poprawnie rozwiązać co najmniej 70% zadań testowych. Przyjmuje się, że wyniki testu zdających w I terminie mają rozkład $N(76\%; 8\%)$. Jaki odsetek studentów zda egzamin w I terminie?

3.99. Stwierdzono, że czas spóźnień studentów na zajęcia jest zgodny z rozkładem normalnym. Wiadomo, że co piąte spóźnienie było krótsze od 2 minut, a wariancja spóźnień jest równa 0,25. Ustalić średni czas spóźnień.

3.100. Rozkład czasu dojazdu do pracy pewnej grupy pracowników jest normalny z odchyleniem standardowym równym 15 minut. Jaki jest średni czas dojazdu, jeśli 75,8% ogółu pracowników dojeżdża do pracy krócej niż 40 minut?

Rozdział IV

ESTYMACJA PODSTAWOWYCH PARAMETRÓW POPULACJI

4.1. Estymatory i ich pożądane własności

Estymacja – to szacowanie (ocenianie) nieznanych wartości parametrów rozkładu cechy statystycznej w populacji generalnej (**estymacja parametryczna**) lub postaci rozkładu badanych cech (**estymacja nieparametryczna**). Poniżej skoncentrujemy uwagę na estymacji parametrycznej.

W ramach estymacji parametrycznej wyróżnia się **estymację punktową** oraz **estymację przedziałową**. W estymacji punktowej za ocenę wartości parametru przyjmuje się **jedną** konkretną wartość, otrzymaną na podstawie wyników próby losowej. W przypadku estymacji przedziałowej wyznacza się – w odpowiedni sposób – określony **przedział** liczbowy, w którym z zadanym z góry prawdopodobieństwem zawiera się wartość szacowanego parametru. Podstawowym narzędziem estymacji jest **estymator**.

Estymatorem parametru Q rozkładu zmiennej losowej X nazywamy taką statystykę:

$$Z_n = f(X_1, X_2, \ldots, X_n), \tag{4.1}$$

której rozkład prawdopodobieństwa zależy od szacowanego parametru.

Estymator jest zmienną losową. Oznaczamy go więc dużą literą.

W zastosowaniach praktycznych interesuje nas konkretna wartość liczbowa:

$$z_n = f(x_1, x_2, \ldots, x_n), \tag{4.2}$$

jaką przyjmuje estymator Z_n parametru Q dla realizacji próby $(x_1, x_2, \ldots, x_n)$. Tę konkretną wartość z_n, będącą realizacją zmiennej losowej Z_n, nazywamy **oceną parametru** Q. Ocenę tę oznaczamy małą literą.

Do szacowania parametru Q wykorzystuje się wyniki próby losowej. Istnieje zatem możliwość popełnienia błędu. Różnicę między estymatorem a wartością parametru (tzn. $Z_n - Q$) nazywamy błędem szacunku. Za miarę tego błędu przyjmuje się zazwyczaj wyrażenie:

$$\Delta = E(Z_n - Q)^2. \tag{4.3}$$

W przypadku gdy $E(Z_n) = Q$, wyrażenie (4.3) jest wariancją $D^2(Z_n)$ estymatora Z_n. Pierwiastek kwadratowy z wariancji jest natomiast odchyleniem standardowym $D(Z_n)$, które nosi nazwę **standardowego błędu szacunku** parametru Q. Iloraz standardowego błędu szacunku do estymatora Z_n jest nazywany **współczynnikiem zmienności**:

$$V(Z_n) = \frac{D(Z_n)}{Z_n}. \tag{4.4}$$

Współczynnik zmienności najczęściej wyraża się w procentach i określa mianem **względnego** błędu szacunku parametru Q.

Do szacowania parametrów populacji należy wykorzystywać najlepsze estymatory spośród wielu dostępnych statystyk z próby. Przykładowo, do estymacji średniej w populacji m można użyć średniej arytmetycznej z próby, mediany z próby, średniej z pierwszej i ostatniej obserwacji w próbie itp. Powstaje zatem problem wyboru „dobrego" estymatora, tj. takiego, który zapewnia otrzymywanie wyników zbliżonych do rzeczywistości. Przy wyborze należy kierować się kryteriami określającymi pożądane własności estymatora. Są nimi: **nieobciążoność**, **zgodność**, **efektywność** i **dostateczność** (wystarczalność).

Estymator Z_n parametru Q jest **nieobciążony**, jeśli jego wartość oczekiwana jest równa szacowanemu parametrowi, tj.:

$$E(Z_n) = Q. \tag{4.5}$$

Własność nieobciążoności oznacza, że wartości estymatora (oceny parametru Q) uzyskiwane w powtarzanych próbach mają średnią równą wartości szacowanego parametru. Tak więc jeśli estymator jest nieobciążony, to uzyskiwane przy jego zastosowaniu oceny parametru nie są obciążone błędem systematycznym.

Estymator niespełniający relacji (4.5) nazywamy estymatorem **obciążonym**, a różnicę:

$$b_n = E(Z_n) - Q \tag{4.6}$$

nazywamy **obciążeniem estymatora**.

Estymator obciążony, dla którego obciążenie b_n maleje wraz ze wzrostem liczebności próby, nazywamy estymatorem **asymptotycznie nieobciążonym**. Dla estymatora asymptotycznie nieobciążonego zachodzi więc:

$$\lim_{n\to\infty} b_n = 0. \tag{4.7}$$

Estymator Z_n parametru Q nazywamy **zgodnym**, jeśli dla dowolnie małego $\varepsilon > 0$ spełniona jest relacja:

$$\lim_{n\to\infty} P\{|Z_n - Q| < \varepsilon\} = 1. \tag{4.8}$$

Warunek (4.8) oznacza, że wraz ze wzrostem liczebności próby uzyskuje się coraz większe prawdopodobieństwo tego, że estymator będzie

przyjmować wartości nieznacznie różniące się (tzn. o mniej niż ε) od wartości parametru. Jeśli próba jest dostatecznie duża, to ryzyko popełnienia błędu przy stosowaniu estymatora zgodnego jest niewielkie.

Pomiędzy własnościami nieobciążoności i zgodności estymatora zachodzą następujące zależności:

a) jeśli estymator Z_n parametru Q jest nieobciążony (lub asymptotycznie nieobciążony), a jego wariancja zmierza – w miarę wzrostu liczebności próby – do zera, to Z_n jest estymatorem zgodnym,

b) jeśli estymator Z_n parametru Q jest zgodny, to jest asymptotycznie nieobciążony.

Załóżmy, że dysponujemy dwoma nieobciążonymi estymatorami parametru Q. Który z nich wybrać? Intuicyjnie można stwierdzić, że należy wybrać ten z nich, którego wartości – przy tej samej liczebności prób – wykazują silniejszą koncentrację wokół wartości szacowanego parametru. Krótko mówiąc, lepszy jest ten estymator, który ma mniejszą wariancję. Problem ten wiąże się z kolejną własnością dobrego estymatora – **efektywnością**.

Jeśli $\{Z_n^1, Z_n^2, \ldots, Z_n^k\}$ jest zbiorem wszystkich nieobciążonych estymatorów parametru Q, to ten spośród nich jest najefektywniejszy, którego wariancja jest najmniejsza. Tego rodzaju estymator dostarcza ocen, które – przeciętnie biorąc – są najbliższe wartości parametru Q. Tak więc **estymatorem najefektywniejszym** nazywamy estymator o najmniejszej wariancji. Efektywnością estymatora Z_n^i parametru Q nazywamy wyrażenie:

$$e\left(Z_n^i\right) = \frac{D^2\left(Z_n^*\right)}{D^2\left(Z_n^i\right)}, \tag{4.9}$$

gdzie Z_n^* oznacza estymator najefektywniejszy.

Efektywność najefektywniejszego estymatora jest równa jedności, w pozostałych przypadkach $0 < e < 1$. Jeśli zachodzi równość:

$$\lim_{n\to\infty} e\left(Z_n^i\right) = 1, \tag{4.10}$$

to estymator jest **asymptotycznie najefektywniejszy**.

Ostatnią spośród pożądanych własności estymatora jest **dostateczność** (**wystarczalność**). Estymator Z_n parametru Q jest dostateczny, jeżeli zawiera wszystkie informacje, jakie na temat parametru Q występują w próbie. Przykładowo, spośród dwóch nieobciążonych estymatorów wartości oczekiwanej $E(X) = m$, którymi są średnia arytmetyczna z próby; $\overline{X} = \frac{1}{n}\sum_{i=1}^{n} X_i$ oraz $\widetilde{X} = \frac{X_{\min} + X_{\max}}{2}$, $\widetilde{X}$ nie jest dostatecznym estymatorem, gdyż przy jego wyznaczaniu wzięto pod uwagę jedynie dwie wartości z próby ($X_{\min}$ oraz $X_{\max}$).

Do uzyskiwania estymatorów charakteryzujących się pożądanymi własnościami służy **metoda największej wiarygodności (MNW)**[1].

4.2. Estymacja punktowa i przedziałowa

W estymacji punktowej za wartość parametru Q przyjmuje się obliczoną na podstawie konkretnej n-elementowej próby losowej wartość tego parametru ($Q = z_n$). Wynik estymacji punktowej zapisuje się zwykle wraz ze średnim błędem szacunku parametru Q, tzn. $D(Z_n)$:

$$Q = z_n \pm D(Z_n). \tag{4.11}$$

Prawdopodobieństwo tego, że estymator przyjmie wartość równą wartości szacowanego parametru, jest równe zeru. Oznacza to, że punktowe oceny parametrów uzyskiwane z prób losowych z reguły różnią się od wartości tych parametrów w populacji generalnej. Ponadto przy estymacji punktowej nie jesteśmy w stanie ustalić stopnia ufności co do prawdziwości uzyskanych wyników. Z tych też względów najczęściej korzysta się z **estymacji przedziałowej**. Za twórcę tego sposobu szacowania parametrów uznaje się statystyka polskiego pochodzenia Jerzego Splawę-Neymana (1894–1981).

Estymacja przedziałowa polega na tym, że na podstawie próby wyznacza się losowy przedział, który z określonym z góry – bliskim jedności – prawdopodobieństwem pokrywa nieznaną wartość szacowanego parametru. Przedział taki nosi nazwę **przedziału ufności Neymana**, a prawdopodobieństwo, z jakim pokrywa on szacowany parametr, określa się mianem **współczynnika ufności** i oznacza symbolem $1-\alpha$. Wyznaczenie przedziału ufności jest możliwe, gdy znany jest rozkład statystyki używanej do estymacji.

Ogólna postać przedziału ufności jest następująca:

$$P\{g_1(Z_n) < Q < g_2(Z_n)\} = 1-\alpha, \tag{4.12}$$

gdzie $g_1(Z_n)$ oraz $g_2(Z_n)$ są odpowiednio dolną i górną granicą przedziału ufności.

W ujęciu J. Splawy-Neymana, parametr Q jest wielkością **stałą** (nielosową). Zmiennymi losowymi są natomiast zależne od Q granice przedziału ufności. Stąd też interpretacja wyniku estymacji przedziałowej powinna wskazywać na zmienność jego granic, a nie parametru Q. Konstruowana dla dowolnego parametru Q interpretacja przedziału ufności Neymana jest następująca: przy wielokrotnym pobieraniu prób n-elementowych i wyznaczaniu na ich podstawie wartości funkcji $g_1(Z_n)$ i $g_2(Z_n)$ w średnio $(1-\alpha)\cdot 100\%$ przypadków, otrzymamy przedziały pokrywające nieznaną wartość parametru Q, a w $\alpha \cdot 100\%$ przypadków otrzymamy przedziały niepokrywające parametru Q.

[1] Metoda ta jest omówiona m.in. w pracy J. Grenia, *Statystyka matematyczna. Podręcznik programowany*, PWN, Warszawa 1987, s. 258–271.

W praktyce dysponujemy tylko jedną próbą i na jej podstawie wyznaczamy jeden przedział ufności, którego końce są liczbami. W tej sytuacji nie jesteśmy pewni, czy wartość szacowanego parametru Q należy do otrzymanego przedziału. Jednakże „ufamy", że tak jest, jeśli współczynnik ufności $1-\alpha$ pozostaje dostatecznie wysoki. Dlatego też w praktycznych zastosowaniach za współczynniki ufności przyjmuje się prawdopodobieństwa: 0,90; 0,95; 0,99.

Różnicę między górną i dolną granicą przedziału ufności nazywamy długością przedziału ufności. Połowa długości przedziału ufności nosi nazwę **bezwzględnej precyzji oszacowania**. Im krótszy jest przedział ufności, tym precyzyjniejsza jest estymacja przedziałowa. Przyjmowanie zbyt wysokiego – bliskiego jedności – współczynnika ufności powoduje wzrost długości przedziału ufności, czyli pogarsza precyzję oszacowania. Stąd też nie należy, bez potrzeby, przyjmować $1-\alpha$ na bardzo wysokim poziomie.

4.2.1. Estymacja średniej w populacji

Najczęściej stosowanym estymatorem średniej w populacji jest średnia arytmetyczna z próby $\overline{X}$. Estymator ten jest zmienną losową o rozkładzie normalnym ze średnią m i odchyleniem standardowym $\frac{\sigma}{\sqrt{n}}$. Służy on zarówno do punktowego, jak i przedziałowego szacowania średniej w populacji.

Załóżmy, że w mieście L wylosowano 200 ofert sprzedaży mieszkań. Rozkład ceny 1 m^2 powierzchni mieszkania kształtował się w wylosowanej próbie następująco:

Cena 1 m^2 (w tys. zł)	1,0–1,5	1,5–2,0	2,0–2,5	2,5–3,0	3,0–3,5
Liczba ofert	36	44	64	36	20

Na podstawie wyników próby losowej mamy oszacować – punktowo i przedziałowo – średnią cenę 1 m^2 powierzchni mieszkania we wszystkich ofertach sprzedaży mieszkań zgłoszonych w mieście L.

Estymacja punktowa średniej w populacji polega na przyjęciu średniej arytmetycznej z próby ze średnim błędem szacunku $D(\overline{X})$ za wartość parametru m w populacji. Średni błąd szacunku parametru m jest określony następująco:

$$D(\overline{X}) = \frac{\sigma}{\sqrt{n}}. \tag{4.13}$$

Wynika to z faktu, że jeśli cecha X ma w populacji rozkład normalny ze średnią m i odchyleniem standardowym σ i z populacji tej pobieramy n-elementową próbę losową prostą, to średnia arytmetyczna z próby $\overline{X}$ ma rozkład normalny ze średnią $E(\overline{X}) = m$ i odchyleniem standardowym $D(\overline{X}) = \frac{\sigma}{\sqrt{n}}$. Zapisujemy to w następujący sposób: $\overline{X} \approx N\left(n, \frac{\sigma}{\sqrt{n}}\right)$.

Z wyników próby obliczamy średnią arytmetyczną i odchylenie standardowe:

$$\bar{x} = \frac{1}{n}\sum_{i=1}^{k} x_i n_i = \frac{421}{200} = 2{,}105 \text{ tys. zł},$$

$$s = \sqrt{\frac{1}{n}\sum_{i=1}^{k}\left(x_i - \bar{x}\right)^2 n_i} = \sqrt{\frac{74{,}405}{200}} = 0{,}61 \text{ tys. zł}.$$

Ze względu na dużą próbę przyjmujemy, że nieznane odchylenie standardowe w populacji jest równe – w przybliżeniu – odchyleniu standardowemu obliczonemu z próby ($\sigma \approx s$). Stąd też:

$$D(\bar{X}) = \frac{0{,}61}{\sqrt{200}} = 0{,}043 \text{ tys. zł}$$

Punktowa ocena średniej m w populacji jest zatem równa:

$$m \approx \bar{x} \pm \frac{s}{\sqrt{n}} \approx 2{,}105 \pm 0{,}043 \text{ [tys. zł]}.$$

Średnia cena 1 m^2 powierzchni mieszkania we wszystkich zgłoszonych ofertach wynosi 2,105 tys. zł, ze średnim błędem szacunku ±0,043 tys. zł.

W celu określenia przedziału ufności dla średniej standaryzujemy zmienną losową $\bar{X}$, otrzymując:

$$Z = \frac{\bar{X} - m}{\frac{\sigma}{\sqrt{n}}} = \frac{\bar{X} - m}{\sigma}\sqrt{n}. \qquad (4.14)$$

Zmienna Z ma standaryzowany rozkład normalny $N(0,1)$.

Dla danej liczebności próby n i przyjętego współczynnika ufności $1-\alpha$ przedział ufności dla średniej m w populacji ma postać:

$$P(-z_\alpha < Z < z_\alpha) = 1 - \alpha \qquad (4.15)$$

lub

$$P\left(-z_\alpha < \frac{\bar{X} - m}{\sigma}\sqrt{n} < z_\alpha\right) = 1 - \alpha. \qquad (4.16)$$

Rozwiązując nierówność (4.16) względem m, mamy:

$$P\left(\bar{X} - z_\alpha \frac{\sigma}{\sqrt{n}} < m < \bar{X} + z_\alpha \frac{\sigma}{\sqrt{n}}\right) = 1 - \alpha. \qquad (4.17)$$

Symbol z_α jest wartością zmiennej losowej Z odczytaną z tablicy dystrybuanty standaryzowanego rozkładu normalnego według zasady:

$$F(z_\alpha) = 1 - \frac{\alpha}{2}. \qquad (4.18)$$

Dysponując konkretną próbą losową o liczebności $n = 200$ ofert sprzedaży mieszkań, przedział ufności dla średniej w populacji (przy współczynniku ufności 0,95 oraz przyjęciu założenia, że ($\sigma \approx s$), zbudujemy następująco:

$$2{,}105 - 1{,}96\frac{0{,}61}{\sqrt{200}} < m < 2{,}105 + 1{,}96\frac{0{,}61}{\sqrt{200}}$$

$$2{,}02 < m < 2{,}19.$$

Tak więc przedział liczbowy o końcach 2,02 tys. zł oraz 2,19 tys. zł jest jednym z tych możliwych do otrzymania przedziałów, który z prawdopodobieństwem 0,95 pokrywa nieznaną wartość średniej ceny 1 m^2 powierzchni mieszkania w populacji. Nie możemy jednak z całą pewnością twierdzić, że wartość m należy do otrzymanego przedziału ufności Neymana.

W przypadku, gdy cecha X w populacji ma rozkład normalny $N(m,\sigma)$ ze znanym odchyleniem standardowym, niezależnie od liczebności próby przedział ufności dla średniej w populacji budujemy zgodnie z relacją (4.17).

W pewnej spółce postanowiono zbadać średni czas dojazdu pracowników do pracy. W tym celu wylosowano 16 pracowników i otrzymano następujące wyniki (w minutach): 12; 17; 8; 14; 25; 20; 9; 15; 17; 13; 11; 5; 7; 10; 14; 19. Z poprzednio przeprowadzonych badań wynika, że rozkład czasu dojazdu do pracy jest normalny z odchyleniem standardowym $\sigma = 6$ minut. Naszym zadaniem jest wyznaczenie przedziału ufności (przy $1-\alpha = 0{,}96$) dla nieznanego średniego czasu dojazdu w populacji pracowników.

Ze względu na fakt, że populacja generalna ma rozkład $N(m, 6)$, do budowy przedziału ufności wykorzystamy wzór (4.17). Podstawiając dane liczbowe do tego wzoru otrzymujemy:

$$13{,}5 - 2{,}05\frac{6}{\sqrt{16}} < m < 13{,}5 + 2{,}05\frac{6}{\sqrt{16}}$$

$$10{,}425 < m < 16{,}575.$$

Otrzymany przedział ufności o końcach 10,425 min oraz 16,575 min jest jednym z tych wszystkich możliwych do otrzymania przedziałów, które z prawdopodobieństwem 0,96 pokrywają średni czas dojazdu w populacji pracowników.

Jeśli cecha X ma w populacji generalnej rozkład $N(m,\sigma)$, gdzie zarówno średnia m, jak i odchylenie standardowe nie są znane, wówczas – przy małej próbie ($n \le 30$) – do wnioskowania o średniej m populacji wykorzystuje się statystykę o postaci:

$$T = \frac{\bar{X} - m}{S}\sqrt{n-1}, \qquad (4.19)$$

gdzie:

$$S = \sqrt{\frac{1}{n-1}\sum_{i=1}^{n}\left(X_i - \bar{X}\right)^2} \qquad (4.20)$$

oznacza odchylenie standardowe z próby.

Rozkład statystyki T, określonej wzorem (4.19), nie zależy od parametru σ. Rozkład ten nosi nazwę **rozkładu t-Studenta** o $n-1$ stopniach swo-

body. Jedynym parametrem tego rozkładu jest liczba stopni swobody: równa liczbie niezależnych składników tworzących tę statystykę. Z uwagi na równość: $\sum_{i=1}^{n}(X_i - \overline{X}) = 0$, liczba stopni swobody wynosi $n - 1$.

W celu ustalenia przedziału ufności dla średniej w populacji przy zadanym współczynniku ufności $1 - \alpha$ za punkt wyjścia przyjmujemy relację:

$$P(-t_{\alpha,n-1} < T < t_{\alpha,n-1}) = 1 - \alpha. \tag{4.21}$$

Podstawiając do wzoru (4.21) w miejsce T wyrażenie (4.19) otrzymujemy:

$$P\left(-t_{\alpha,n-1} < \frac{\overline{X} - m}{S}\sqrt{n-1} < t_{\alpha,n-1}\right) = 1 - \alpha. \tag{4.22}$$

Rozwiązując nierówność (4.24) względem m mamy:

$$P\left(\overline{X} - t_{\alpha,n-1}\frac{S}{\sqrt{n-1}} < m < \overline{X} + t_{\alpha,n-1}\frac{S}{\sqrt{n-1}}\right) = 1 - \alpha. \tag{4.23}$$

Dla konkretnej próby losowej przedział liczbowy, będący realizacją (4.25), ma końce określone następująco:

$$\left(\overline{x} - t_{\alpha,n-1}\frac{S}{\sqrt{n-1}}; \overline{x} + t_{\alpha,n-1}\frac{S}{\sqrt{n-1}}\right). \tag{4.24}$$

Wartość $t_{\alpha,n-1}$ odczytujemy z tablicy rozkładu t-Studenta przy określonej liczbie stopni swobody i poziomie istotności α. Otrzymane na podstawie rozkładu t-Studenta przedziały ufności mają nie tylko losowe końce, ale również losową długość, która wynosi: $2t_{\alpha,n-1}\frac{S}{\sqrt{n-1}}$ (S jest zmienną losową).

Jeśli liczba stopni swobody rozkładu t-Studenta zmierza do nieskończoności, to rozkład ten zmierza do standaryzowanego rozkładu normalnego. W praktycznych zastosowaniach już przy $n > 30$ wartość $t_{\alpha,n-1}$ zastępuje się przez z_α. Stąd też próbę o liczebności większej od 30 nazywamy **próbą dużą**.

W celu ustalenia nowych norm pracy konieczne było oszacowanie średniego czasu potrzebnego do wykonania pewnego detalu na określonym typie obrabiarki. W tym celu z populacji wszystkich robotników wylosowano próbę prostą liczącą $n = 17$ robotników i u każdego z nich dokonano pomiaru czasu wykonania detalu. Okazało się, że średni czas wykonania detalu wynosił 15 minut, z odchyleniem standardowym równym 2 minuty. Naszym zadaniem jest oszacowanie – przy współczynniku ufności 0,95 – średniego czasu potrzebnego do wykonania detalu w całej populacji robotników, jeśli wiadomo, że rozkład czasu wykonania detalu jest $N(m,\sigma)$.

Z uwagi na to, że odchylenie standardowe σ w populacji nie jest znane, a próba jest mała, przedział ufności dla średniej m zbudujemy wykorzystując wzór (4.23). Z tablic rozkładu t-Studenta dla $\alpha = 0{,}05$ oraz $n - 1 = 16$ stopni swobody, odczytujemy wartość $t_{0,05;16} = 2{,}12$. Podstawiając tę wartość i charakterystyki obliczone na podstawie próby do wzoru (4.23) otrzymujemy:

$$15 - 2{,}12\frac{2}{\sqrt{16}} < m < 15 + 2{,}12\frac{2}{\sqrt{16}}$$

$$13{,}94 < m < 16{,}06.$$

Na poziomie ufności 0,95 możemy zatem stwierdzić, że średni czas wykonania detalu przez wszystkich robotników jest nie mniejszy niż 13,94 minuty i nie większy niż 16,06 minuty.

4.2.2. Estymacja wariancji i odchylenia standardowego populacji

Przy konstrukcji przedziału ufności dla wariancji σ^2 w populacji korzysta się ze statystyki:

$$\chi^2 = \frac{nS^2}{\sigma^2} = \frac{(n-1)\hat{S}^2}{\sigma^2} = \sum_{i=1}^{n}\left(\frac{X_i - \overline{X}}{\sigma}\right)^2 = \frac{\sum_{i=1}^{n}(X_i - \overline{X})^2}{\sigma^2}, \tag{4.25}$$

gdzie $X_i (i = 1, \ldots, n)$ jest ciągiem niezależnych zmiennych losowych mających jednakowy rozkład $N(m,\sigma)$. Statystyka (4.25) ma rozkład chi-kwadrat o $n-1$ stopniach swobody. Symbolami S^2 oraz $\hat{S}^2$ oznaczono estymatory parametru σ^2, przy czym:

$$S^2 = \frac{1}{n}\sum_{i=1}^{n}(X_i - \overline{X})^2, \tag{4.26}$$

$$\hat{S}^2 = \frac{1}{n-1}\sum_{i=1}^{n}(X_i - \overline{X})^2. \tag{4.27}$$

S^2 jest estymatorem zgodnym, ale obciążonym parametru σ^2, natomiast $\hat{S}^2$ estymatorem zgodnym i nieobciążonym parametru σ^2.

Pomiędzy estymatorami S^2 oraz $\hat{S}^2$ zachodzą następujące relacje:

$$\hat{S}^2 = \frac{n}{n-1}S^2 \tag{4.28}$$

oraz

$$S^2 = \frac{n-1}{n}\hat{S}^2. \tag{4.29}$$

Jak łatwo zauważyć, przy małych n iloraz $\frac{n}{n-1}$ istotnie różni się od jedności. Przy $n \to \infty$ iloraz ten szybko zmierza do jedności. Praktycznie już dla $n > 50$ między estymatorami S^2 oraz $\hat{S}^2$ nie ma różnicy.

Przedział ufności dla wariancji σ^2 w populacji, z której pobrano małą próbę losową, ma – przy współczynniku ufności $1 - \alpha$ – następującą postać:

$$P\left(\frac{nS^2}{\chi^2_{\frac{\alpha}{2}}} < \sigma^2 < \frac{nS^2}{\chi^2_{1-\frac{\alpha}{2}}}\right) = 1-\alpha, \tag{4.30}$$

$$P\left(\frac{(n-1)\hat{S}^2}{\chi^2_{\frac{\alpha}{2}}} < \sigma^2 < \frac{(n-1)\hat{S}^2}{\chi^2_{1-\frac{\alpha}{2}}}\right) = 1-\alpha. \tag{4.31}$$

Modele (4.30) oraz (4.31) stosowane są dla małych prób ($n \leq 30$) pochodzących z populacji normalnych.

Wartości $\chi^2_{\frac{\alpha}{2}}$ i $\chi^2_{1-\frac{\alpha}{2}}$ odczytujemy z tablic rozkładu chi-kwadrat przy $n-1$ stopniach swobody w taki sposób, aby przy ustalonym współczynniku ufności $1-\alpha$ spełnione były równości:

$$P\left(\chi^2 > \chi^2_{\frac{\alpha}{2}}\right) = \frac{\alpha}{2} \tag{4.32}$$

oraz

$$P\left(\chi^2 > \chi^2_{1-\frac{\alpha}{2}}\right) = 1-\frac{\alpha}{2}. \tag{4.33}$$

Modele (4.30) oraz (4.31) pozwalają zbudować przedział ufności dla odchylenia standardowego σ w populacji. W tym celu należy spierwiastkować wszystkie człony nierówności podwójnej w relacjach (4.30) lub (4.31).

Załóżmy, że w firmie dokonano kontroli wagi zawartości 10 losowo dobranych puszek rybnych, otrzymując (w gramach): 295; 285; 280; 290; 305; 315; 295; 290; 305; 300. Wiadomo, że waga zawartości puszek jest zmienną losową o rozkładzie normalnym o nieznanych parametrach. Naszym zadaniem jest oszacowanie – przy współczynniku ufności 0,90 – wariancji wagi zawartości puszek.

Wariancja jest tu szacowana na podstawie wyników 10-elementowej próby losowej. Stąd też przedział ufności dla parametru σ^2 budowany jest na podstawie rozkładu chi-kwadrat, określonego przez $n-1$ stopni swobody (wzór 4.31):

$$\frac{9\cdot 110}{16{,}919} < \sigma^2 < \frac{9\cdot 110}{3{,}325}$$

$$58{,}51 < \sigma^2 < 297{,}74.$$

Wariancję z próby $\hat{s}^2$ obliczono następująco:

$$\hat{s}^2 = \frac{1}{n-1}\sum_{i=1}^{n}(x_i-\bar{x})^2 = \frac{1}{9}\cdot 990 = 110.$$

Wartości 16,919 oraz 3,325 zostały odczytane z tablic rozkładu chi-kwadrat przy współczynniku ufności 0,90 oraz 9 stopniach swobody.

Przedział liczbowy o dolnej granicy 58,51 $(g)^2$ i górnej granicy 297,74 $(g)^2$ jest jednym z możliwych do otrzymania na podstawie wyników powtarzalnych 10-elementowych prób. Przedziały te mają tę własność, iż w 90 przypadkach na 100 pokrywają nieznaną wariancję wagi zawartości puszek w całej partii.

Pierwiastkując człony nierówności: 58,51 $(g)^2 < \sigma^2 <$ 297,74 $(g)^2$ otrzymujemy: 7,65 g $< \sigma <$ 17,26 g. Interpretacja przedziału ufności dla odchylenia standardowego σ jest analogiczna jak przedziału ufności dla wariancji σ^2.

W przypadku dysponowania dużą próbą ($n > 30$) można od razu zbudować przedział ufności dla odchylenia standardowego σ. Korzystamy wówczas z faktu, że estymator S parametru σ ma asymptotyczny rozkład normalny $N\left(\sigma, \frac{\sigma}{\sqrt{2n}}\right)$. Standaryzując odchylenie standardowe z próby S mamy:

$$Z = \frac{S-\sigma}{\frac{\sigma}{\sqrt{2n}}} = \frac{S-\sigma}{\sigma}\sqrt{2n}. \tag{4.34}$$

Statystyka Z ma rozkład asymptotycznie normalny $N(0,1)$. Stąd też:

$$P\left(-z_\alpha < \frac{S-\sigma}{\sigma}\sqrt{2n} < z_\alpha\right) = 1-\alpha, \tag{4.35}$$

gdzie z_α jest wartością odczytaną z tablicy dystrybuanty rozkładu normalnego standaryzowanego w ten sposób, aby $F(z_\alpha) = 1-\frac{\alpha}{2}$.

Przekształcając relację (4.35) otrzymujemy:

$$P\left(\frac{S}{1+\frac{z_\alpha}{\sqrt{2n}}} < \sigma < \frac{S}{1-\frac{z_\alpha}{\sqrt{2n}}}\right) = 1-\alpha \tag{4.36}$$

lub

$$P\left(S - z_\alpha\frac{S}{\sqrt{2n}} < \sigma < S + z_\alpha\frac{S}{\sqrt{2n}}\right) = 1-\alpha. \tag{4.37}$$

Sposób konstrukcji przedziału ufności dla odchylenia standardowego w przypadku dużej próby zilustrujemy przykładem.

W banku X wylosowano w ostatnim tygodniu maja br. 150 kont osobistych i zbadano je pod względem liczby przeprowadzonych operacji. Okazało się, że średnia arytmetyczna wynosiła 3 operacje, natomiast odchylenie standardowe 1,2 operacji: Mamy oszacować – punktowo i przedziałowo (przy współczynniku ufności 0,95) – zróżnicowanie liczby operacji w populacji wszystkich kont osobistych.

Estymatorem odchylenia standardowego w populacji jest odchylenie standardowe S z dużej próby. Rozkład odchylenia standardowego z próby S zmierza – przy $n \to \infty$ – do rozkładu normalnego o wartości oczekiwanej σ i odchyleniu standardowym $\frac{\sigma}{\sqrt{2n}}$. Estymacja punktowa odchylenia standardowego w populacji (σ) polega na przyjęciu odchylenia standardowego z próby ze średnim błędem szacunku za wartość σ. Mamy więc:

$$\sigma = s \pm \frac{s}{\sqrt{2n}},$$

$$\sigma = 1{,}2 \pm \frac{1{,}2}{\sqrt{300}} = 1{,}2 \pm 0{,}069.$$

Dokonując przedziałowej estymacji odchylenia standardowego w populacji (jeśli $1-\alpha = 0{,}95$, to $z_\alpha = 1{,}96$) korzystamy ze wzoru (4.37). Otrzymujemy:

$$1{,}2 - 1{,}96\frac{1{,}2}{\sqrt{300}} < \sigma < 1{,}2 + 1{,}96\frac{1{,}2}{\sqrt{300}}$$

$$1{,}064 < \sigma < 1{,}336.$$

Oznacza to, że przedział liczbowy (1,064; 1,336) jest jednym ze wszystkich możliwych przedziałów, które z prawdopodobieństwem 0,95 pokrywają nieznane zróżnicowane liczby operacji w populacji wszystkich kont osobistych w banku X.

4.2.3. Estymacja wskaźnika struktury populacji

W badaniach statystycznych często spotykamy się z cechami **jakościowymi (niemierzalnymi)**. W takich badaniach interesują nas zazwyczaj dwa rodzaje jednostek zbiorowości: **frakcja (wskaźnik struktury)** jednostek posiadających wyróżnioną cechę jakościową oraz frakcję jednostek nie posiadających jej. Frakcję elementów posiadających wyróżnioną cechę w populacji generalnej oznaczamy symbolem p, frakcję jednostek nieposiadających wyróżnionej cechy – symbolem $q = 1 - p$. Powyższe frakcje są wskaźnikami struktury; spełniona jest zatem równość: $p + q = 1$.

Najlepszym estymatorem wskaźnika struktury w populacji jest wskaźnik struktury z próby: $\hat{p} = \frac{m}{n}$, gdzie m jest liczbą jednostek w próbie posiadających wyróżnioną cechę, a n jest liczebnością próby. Estymator ten ma – przy dużej ($n > 100$) próbie – rozkład asymptotycznie normalny $N\left(p, \sqrt{\frac{p(1-p)}{n}}\right)$. Odchylenie standardowe rozkładu estymatora nosi nazwę **średniego błędu szacunku wskaźnika struktury (frakcji)**.

Estymacja punktowa polega na przyjęciu statystyki z próby ze średnim błędem szacunku tej statystyki za wartość parametru p populacji:

$$p = \hat{p} \pm D(\hat{p}) = \hat{p} \pm \sqrt{\frac{\hat{p}(1-\hat{p})}{n}}. \tag{4.38}$$

W celu budowy przedziału ufności dla wskaźnika struktury p w populacji należy dokonać standaryzacji estymatora $\hat{p}$:

$$Z = \frac{\hat{p} - p}{\sqrt{\frac{\hat{p}(1-\hat{p})}{n}}}. \tag{4.39}$$

Statystyka Z ma asymptotyczny rozkład normalny zero-jedynkowy $N(0, 1)$.

Przedział ufności dla wskaźnika struktury w populacji generalnej jest określony wzorem:

$$P\left(-z_\alpha < \frac{\hat{p} - p}{\sqrt{\frac{\hat{p}(1-\hat{p})}{n}}} < z_\alpha\right) = 1 - \alpha. \tag{4.40}$$

Po odpowiednich przekształceniach nierówności podwójnej (4.40) otrzymamy wzór na przedział ufności dla parametru p z dużej próby ($n > 30$):

$$P\left(\hat{p} - z_\alpha\sqrt{\frac{\hat{p}(1-\hat{p})}{n}} < p < \hat{p} + z_\alpha\sqrt{\frac{\hat{p}(1-\hat{p})}{n}}\right) = 1 - \alpha, \tag{4.41}$$

gdzie: $\hat{p} = \frac{m}{n}$ jest wskaźnikiem struktury w próbie (m – liczba elementów wyróżnionych w próbie, n – liczebność próby).

Jak wynika ze wzoru (4.41), dokładność oszacowania parametru p wzrasta, jeśli $n \to \infty$.

W lutym 2004 roku Ośrodek Badania Opinii Publicznej poinformował, że 75% ze 100 pytanych osób z wykształceniem wyższym popiera przystąpienie Polski do Unii Europejskiej. Należy oszacować (punktowo i przedziałowo przy współczynniku ufności 0,95) frakcję obywateli Polski z wykształceniem wyższym popierających wstąpienie naszego kraju w struktury UE.

Populację generalną tworzą tu mieszkańcy Polski z wykształceniem wyższym. W populacji tej zmienną losową X jest opinia na temat przystąpienia Polski do Unii Europejskiej. Przedmiot naszego zainteresowania stanowi frakcja $\hat{p} = \frac{m}{n}$ obywateli popierających to wydarzenie. Statystyka $\hat{p}$ jest estymatorem parametru p w populacji.

Z wyników próby otrzymujemy:

$$\hat{p} = \frac{m}{n} = \frac{75}{100} = 0{,}75 \text{ oraz } D(\hat{p}) = \sqrt{\frac{0{,}75 \cdot 0{,}25}{100}} = 0{,}043.$$

Ocena punktowa parametru p jest zatem równa:

$$\hat{p} = 0{,}75 \pm 0{,}043.$$

Wykorzystując wzór (4.41) mamy:

$$0{,}75 - 1{,}96\sqrt{\frac{0{,}75 \cdot 0{,}25}{100}} < p < 0{,}75 + 1{,}96\sqrt{\frac{0{,}75 \cdot 0{,}25}{100}}$$

$$0{,}75 - 0{,}084 < p < 0{,}75 + 0{,}084$$

$$0{,}666 < p < 0{,}834.$$

Przedział liczbowy o dolnej granicy 66,6% i górnej granicy 83,4% jest jednym z możliwych do otrzymania na podstawie wyników powtarzalnych 100-elementowych prób, które w 95 przypadkach na 100 pokrywają procentową frakcję obywateli naszego kraju z wyższym wykształceniem pozytywnie oceniających przystąpienie Polski do Unii Europejskiej.

4.3. Zagadnienie minimalnej liczebności próby

Przy estymacji parametrów populacji dysponujemy wynikami jednej próby losowej. Otrzymana ocena różni się – w mniejszym lub większym stopniu – od faktycznej wartości parametru. Miary precyzji oszacowania, tzn. wariancja lub odchylenie standardowe estymatora, są zależne od liczebności próby. Im próba jest liczniejsza, tym oszacowanie jest bardziej precyzyjne. Powstaje zatem problem wyznaczenia takiej liczebności próby, która zapewni żądany stopień precyzji oszacowania.

Dokładność estymacji przedziałowej parametru Q mierzy się długością przedziału ufności wyznaczonego na podstawie wyników próby losowej. Połowę długości przedziału ufności określa się mianem **maksymalnego (dopuszczalnego) błędu szacunku** i oznacza symbolem d.

Jak wiadomo, przedział ufności dla średniej m w populacji o rozkładzie $N(m,\sigma)$ ze znanym odchyleniem standardowym (por. wzór (4.17)) ma postać:

$$\overline{X} - z_\alpha \frac{\sigma}{\sqrt{n}} < m < \overline{X} + z_\alpha \frac{\sigma}{\sqrt{n}}. \tag{4.42}$$

a jego długość wynosi: $2z_\alpha \frac{\sigma}{\sqrt{n}}$. Wynika stąd, że dokładność estymacji przedziałowej (jej miarą jest długość przedziału ufności) zależy od rozproszenia cechy X w populacji (σ), od współczynnika ufności $1-\alpha$ (określającego wartość z_α) oraz od liczebności próby (n).

Aby połowa długości przedziału ufności (zwana maksymalnym lub dopuszczalnym błędem szacunku – d) nie przekraczała – przy ustalonym $1-\alpha$ – wartości d, powinna być spełniona relacja:

$$z_\alpha \frac{\sigma}{\sqrt{n}} \le d. \tag{4.43}$$

Przekształcając relację (4.43) otrzymujemy:

$$n \ge \frac{z_\alpha^2 \sigma^2}{d^2}. \tag{4.44}$$

Zatem wielkość:

$$n = \frac{z_\alpha^2 \sigma^2}{d^2} \tag{4.45}$$

jest minimalną liczebnością próby zapewniającą uzyskanie ustalonej z góry dokładności estymacji przedziałowej średniej w populacji. Wynik otrzymany ze wzoru (4.45) zaokrąglamy zawsze w górę.

Załóżmy, że naszym zadaniem jest ustalenie minimalnej liczebności próby niezbędnej do oszacowania średniego wzrostu ogółu noworodków, jeśli cecha ta ma rozkład normalny $N(m, 1{,}5$ cm). Przyjmujemy, że $1-\alpha = 0{,}99$, natomiast maksymalny błąd szacunku nie powinien przekraczać 0,5 cm.

Do wyznaczenia minimalnej liczebności próby wykorzystamy wzór (4.45). Z tablic dystrybuanty rozkładu normalnego $N(0,1)$ odczytujemy wartość $z_\alpha = 2{,}58$. Stąd też:

$$n = \frac{2{,}58^2 \cdot 1{,}5^2}{0{,}5^2} = \frac{14{,}9769}{0{,}25} = 59{,}9 = 60 \text{ noworodków.}$$

Występująca we wzorze (4.45) wariancja σ^2 nie zawsze jest znana. Korzysta się wówczas – w zależności od sytuacji – z wartości wariancji otrzymanej w innych badaniach lub też szacuje się ją na podstawie **próby wstępnej (pilotażowej)**. W tym przypadku najpierw losuje się niewielką próbę o liczebności n_0 i wyznacza z niej statystykę $\hat{s}^2$. W kolejnym kroku określa się liczebność próby właściwej na podstawie wzoru:

$$n = \frac{t_{\alpha,n-1}^2 \hat{s}}{d^2}, \tag{4.46}$$

gdzie: $t_{\alpha,n-1}$ jest wartością odczytaną z tablicy rozkładu t-Studenta dla poziomu istotności α i danej liczby $n_0 - 1$ stopni swobody. Wzór (4.46) należy stosować wyłącznie dla populacji normalnych.

Przedstawiony powyżej sposób postępowania przy ustalaniu minimalnej liczebności próby nosi nazwę **dwustopniowej metody Steina**.

Dyrekcja hipermarketu zamierza usprawnić pracę personelu zatrudnionego w dniach weekendu. W tym celu wylosowano pilotażową próbę $n_0 = 10$ klientów i stwierdzono, że średni czas przebywania tych klientów w hipermarkecie wynosi $\overline{x} = 26{,}5$ minuty, przy $\hat{s}^2 = 36$ (min)^2. Ustalić, na jak licznej próbie należy przeprowadzić badanie właściwe w celu oszacowania średniego czasu przebywania w hipermarkecie w populacji klientów przy współczynniku ufności 0,95 oraz maksymalnym błędzie szacunku $d = 1{,}5$ minuty.

Z tablic rozkładu t-Studenta odczytujemy – dla $\alpha = 0{,}05$ oraz $n_0 - 1 = 9$ stopni swobody – wartość $t_{0{,}05;9} = 2{,}262$. Podstawiając odpowiednie wartości do wzoru (4.46) otrzymujemy:

$$n = \frac{(2{,}262)^2 \cdot 36}{1{,}5^2} = 81{,}866 = 82 \text{ klientów.}$$

Tak więc, do próby właściwej należy jeszcze dolosować $n - n_0 = 82 - 10 = 72$ klientów.

W przypadku estymacji frakcji wyróżnionych elementów (p) maksymalny błąd szacunku przy n-elementowej próbie i ustalonym współczynniku ufności $1 - \alpha$, określa wzór:

$$d = z_\alpha \sqrt{\frac{p(1-p)}{n}}. \tag{4.47}$$

Przekształcając relację (4.47) otrzymujemy:

$$n = z_\alpha^2 \frac{p(1-p)}{d^2}, \tag{4.48}$$

gdzie d oznacza ustaloną z góry wartość maksymalnego (dopuszczalnego) błędu szacunku.

We wzorach (4.47) i (4.48) symbol p nie jest z reguły znany (stanowi on przedmiot estymacji). Do obliczeń przyjmuje się zatem jego ocenę uzyskaną w innych badaniach lub wartość, jakiej oczekujemy na podstawie własnego rozeznania problemu. Jeżeli nie mamy żadnych podstaw do ustalenia nawet przypuszczalnej wielkości p, przyjmuje się, że wartość $p = 0{,}5$. Dla tej wartości iloczyn $p(1-p)$ jest największy i równy 0,25 ($0{,}5 \cdot 0{,}5 = 0{,}25$). Przy takim podejściu wzór na minimalną liczebność próby przyjmie postać:

$$n = \frac{z_\alpha^2}{4d^2}. \tag{4.49}$$

Przypuśćmy, że chcemy oszacować odsetek rodzin pragnących mieć dostęp do internetu. Ile rodzin należy wylosować do próby, aby przy współczynniku ufności 0,90 uzyskać nie dłuższy niż 8% przedział ufności?

Ze względu na nieznajomość rzędu wielkości frakcji w populacji, do obliczeń przyjmujemy $p = 0{,}5$. Ponadto mamy: $z_\alpha = 1{,}64$, a postulowana wartość $d = 4\%$. Podstawiając te wielkości do wzoru (4.49) otrzymamy:

$$n = \frac{1{,}64^2}{4 \cdot (0{,}04)^2} = 420{,}25 = 421 \text{ rodzin.}$$

W celu zagwarantowania postulowanej dokładności należy zbadać minimum 421 rodzin.

ZADANIA

4.1. Czas świecenia żarówek jest zmienną losową o rozkładzie $N(m, 50)$. Z partii wyprodukowanych żarówek pobrano 9-elementową próbę losową. Okazało się, że średni czas świecenia żarówek w wylosowanej próbie jest równy $\bar{x} = 2880$ godzin. Przy współczynniku ufności 0,95, zbudować (z dokładnością do godziny) przedział ufności dla średniego czasu świecenia żarówek w całej populacji.

4.2. W wylosowanej próbie 150 klientów pewnego supermarketu dokonujących zakupów w niedzielę stwierdzono, że średnia wartość zakupów wynosi 68 zł przy odchyleniu standardowym równym 17 zł. Przyjmując współczynnik ufności 0,90, zbudować przedział ufności dla średniej wartości zakupów wszystkich klientów dokonujących zakupów w niedzielę. Określić bezwzględną precyzję szacunku. Zakładamy, że rozkład badanej cechy jest normalny.

4.3. W losowej próbie 17 małych przedsiębiorstw stwierdzono, że przeciętne zatrudnienie wynosiło 8 osób z odchyleniem standardowym 2. Oszacować – przy współczynniku ufności 0,90 – średnie zatrudnienie we wszystkich małych przedsiębiorstwach. Zakładamy, że rozkład badanej cechy jest normalny.

4.4. W 49-elementowej próbie losowej robotników otrzymano $\bar{x} = 120$ jednorodnych operacji wykonywanych w ciągu zmiany roboczej, przy odchyleniu standardowym $s = 10$ operacji. Przyjmując współczynnik ufności 0,95, zbudować przedział ufności dla średniej liczby operacji w ciągu zmiany roboczej w populacji robotników. Zakładamy, że rozkład badanej cechy jest normalny.

4.5. Z populacji studentów studiujących w pewnej uczelni wylosowano 132-elementową próbę w celu oszacowania średniego czasu poświęconego na naukę w czytelni w ciągu tygodnia. Otrzymano następujące wyniki:

Czas nauki w czytelni w godz.	0–2	2–4	4–6	6–8	8–10	10–12
Liczba studentów	10	28	42	30	15	7

Przyjmując współczynnik ufności 0,90; zbudować przedział ufności dla średniego tygodniowego czasu nauki w czytelni w całej populacji studentów. Zakładamy, że rozkład badanej cechy jest normalny.

4.6. Zmienna losowa X ma rozkład $N(m, 40)$. Na podstawie zaobserwowanych w próbie $n = 25$ wartości tej zmiennej zbudowano przedział ufności dla wartości oczekiwanej o długości 32. Jaki poziom współczynnika ufności przyjęto przy budowie przedziału ufności?

4.7. W 8-osobowej losowo wybranej grupie uczniów dokonano pomiaru czasu rozwiązywania pewnego zadania matematycznego. Otrzymano następujące wyniki (w minutach): 25; 16; 12; 10; 12; 21; 25; 20. Oszacować przedział ufności dla średniego czasu niezbędnego do rozwiązania zadania w całej zbiorowości uczniów. Przyjąć współczynnik ufności 0,90. Zakładamy, że rozkład badanej cechy jest normalny.

4.8. Średni dzienny utarg w sklepach spożywczych wynosił 528 zł, z odchyleniem standardowym 75 zł. Przyjmując współczynnik ufności 0,99, oszacować – przedzialowo – średni dzienny utarg we wszystkich sklepach spożywczych. Zakładamy, że rozkład badanej cechy jest normalny.

4.9. W 200-osobowej grupie nauczycieli akademickich wylosowanej niezależnie przeprowadzono badanie stażu pracy. Otrzymano następujące wyniki:

Staż pracy (w latach)	0–4	4–8	8–12	12–16	16–20
Liczba nauczycieli	20	40	80	40	20

Oszacować – przy współczynniku ufności 0,95 – przeciętny staż pracy w całej populacji nauczycieli. Zakładamy, że rozkład badanej cechy jest normalny.

4.10. Wylosowano próbę złożoną z 5 worków cementu. Ich waga (w kg) wynosiła: 50,2; 50,3; 50,5; 50,5; 50,4. Oszacować – przy współczynniku ufności 0,90 – odchylenie standardowe wagi wszystkich worków cementu. Zakładamy, że rozkład badanej cechy jest normalny.

4.11. W pewnym mieście wylosowano 500 mieszkań. Stwierdzono, że w 200 spośród nich był zainstalowany internet. Co na tej podstawie można powiedzieć o frakcji mieszkań wyposażonych w internet w tym mieście? Przyjąć współczynnik ufności 0,99. Zakładamy, że rozkład badanej cechy jest normalny.

4.12. Badanie wydajności pracy w jednej z firm, przeprowadzone na podstawie 100-elementowej próby losowej prostej, dało następujące wyniki:

Wydajność w sztukach	0–4	4–8	8–12	12–16	16–20
Liczba pracowników	5	15	40	30	10

Jakiego zróżnicowania wydajności pracy należy spodziewać się w całej populacji pracowników przy współczynniku ufności 0,96? Zakładamy, że rozkład badanej cechy jest normalny.

4.13. W losowej próbie 728 respondentów oszacowano średni czas poświęcony na oglądanie telewizji. Okazało się, że wyniósł on 5,2 godz./dzień, przy typowym obszarze zmienności $5{,}0 < x_{typ} < 5{,}4$. Oszacować, przy współczynniku ufności 0,90, przeciętny czas poświecony na oglądanie telewizji w populacji generalnej. Zakładamy, że rozkład badanej cechy jest normalny.

4.14. W losowo pobranej próbie 250 studentów 173 posiada komputer w domu. Przy współczynniku ufności 0,95 zbudować przedział ufności dla odsetka studentów posiadających komputer. Przyjmujemy, że rozkład badanej cechy jest normalny.

4.15. W losowej próbie 2120 absolwentów gimnazjum przeciętna liczba punktów uzyskana z testu matematyczno-przyrodniczego wynosiła 32 z 12,32% zmiennością. Przy współczynniku ufności 0,95 oszacować średnią liczbę punktów uzyskanych przez wszystkich absolwentów gimnazjum. Zakładamy, że rozkład badanej cechy jest normalny.

4.16. Pracochłonność 6 losowo wybranych detali (w minutach) kształtowała się następująco: 6,3; 5,9; 6,2; 5,8; 5,7; 6,1. Przyjmując współczynnik ufności 0,90, zbudować przedział ufności określający zróżnicowanie pracochłonności w całej populacji produkowanych detali. Zakładamy, że rozkład badanej cechy jest normalny.

4.17. Z wyników badania opinii publicznej przeprowadzonego w losowej próbie 2325 Polaków 1736 osób było za przystąpieniem Polski do Unii Europejskiej. Przy współczynniku ufności 0,90 zbudować przedział ufności dla frakcji osób popierających przystąpienie Polski do Unii Europejskiej.

4.18. W celu oszacowania średniego miesięcznego przebiegu pewnego typu samochodów ciężarowych (w km) wylosowano niezależnie 456 samochodów, z których 103 miały przebieg większy niż 5000 km. Przy współczynniku ufności 0,95 zbudować przedział ufności dla odsetka samochodów o miesięcznym przebiegu mniejszym niż 5000 km.

4.19. W losowo pobranej próbie 675 maturzystów 63% dostało się na bezpłatne studia dzienne. Oszacować metodą przedziałową frakcję maturzystów, którzy nie dostali się na studia dzienne. Przyjąć współczynnik ufności 0,95.

4.20. W celu oszacowania średniego czasu dojazdu do pracy osób zatrudnionych w pewnym przedsiębiorstwie wylosowano próbę 17 pracowników. Średni czas dojazdu w tej próbie wynosił 40 minut, a współczynnik zmienności był równy 50%. Jaki współczynnik ufności przyjęto przy budowie przedziału ufności dla średniej w populacji generalnej, jeśli długość oszacowanego przedziału wynosiła 21,2 minuty?

4.21. Na podstawie losowej próby 120 jednokilogramowych opakowań cukru otrzymano $\bar{x} = 95$ dag i $s = 10$ dag. Zbuduj przedział ufności dla odchylenia standardowego w rozkładzie wagi wszystkich produkowanych jednokilogramowych opakowań cukru. Przyjąć współczynnik ufności 0,90. Zakładamy, że rozkład badanej cechy jest normalny.

4.22. Struktura wylowionych 200 sztuk karpia pod względem gatunkowym przedstawia się następująco:

Gatunek karpia	I	II	III
Liczba karpi	46	54	100

Przyjmując współczynnik ufności 0,99, oszacować metodą przedziałową odsetek występowania I gatunku karpia.

4.23. W celu zbadania wieku lekarzy zatrudnionych na wsi i w mieście wylosowano 2 próby: 9-elementową próbę lekarzy zatrudnionych na wsi i 11-elementową próbę lekarzy pracujących w mieście. Średni wiek lekarzy wiejskich wyniósł 42 lata, a miejskich – 46 lat. Zakładając, że rozkład wieku ogółu lekarzy jest normalny $N(m;$ 2,4), oszacować przeciętny wiek ogółu lekarzy. Przyjąć współczynnik ufności 0,98.

4.24. W celu ustalenia przeciętnej zawartości witaminy C w owocach dzikiej róży pobrano 15 próbek 100-gramowych miąższu owocowego. Uzyskano następujące wyniki (w miligramach/100 g miąższu): 495; 455; 438; 483; 501; 468; 492; 471; 474; 485; 504; 469; 478; 495 i 481. Przyjmując współczynnik ufności 0,95, zbudować przedział ufności dla wariancji badanej cechy. Zakładamy, że rozkład badanej cechy jest normalny.

4.25. Badano miesięczne wydatki studentów na cele kulturalne. Dla 20 wylosowanych studentów uzyskano średnią 60 zł oraz wariancję 225 $(\text{zł})^2$. Przyjmując współczynnik ufności 0,95, oszacować metodą przedziałową przeciętne wydatki oraz wariancję wydatków na cele kulturalne w całej zbiorowości studentów. Zakładamy, że rozkład badanej cechy jest normalny.

4.26. Na podstawie informacji uzyskanych w 12 losowo wybranych stacjach meteorologicznych wyznaczono (w dniach) średnią długość okresu wegetacyjnego $\bar{x} = 231{,}33$ dnia oraz $s = 31{,}44$ dnia. Zakładając, że rozkład badanej cechy jest normalny, zbuduj przedział ufności dla średniej i wariancji długości okresu wegetacyjnego. Przyjąć współczynnik ufności 0,95.

4.27. W pewnym przedsiębiorstwie zatrudniającym 5000 pracowników 40% spośród nich jest zadłużonych w Spółdzielczej Kasie Oszczędnościowo-Kredytowej.

Spośród pracowników zadłużonych pobrano niezależną próbę 7,5% osób, w której odchylenie standardowe spłacanych miesięcznie rat kredytu wynosiło 80 zł. Zakładając, że rozkład wysokości spłacanych rat jest normalny, oszacuj – z prawdopodobieństwem 0,95 – przedział ufności pokrywający odchylenie standardowe spłacanych kredytów w całej zbiorowości kredytobiorców.

4.28. Z prawdopodobieństwem 0,95 oszacuj, jaka część uczniów szkół średnich pali papierosy, jeśli w próbie liczącej 1000 uczniów wybranych w losowaniu niezależnym, 360 paliło papierosy.

4.29. W celu oszacowania zróżnicowania wagi jaj pochodzących od kur pewnej rasy zważono 15 jaj wylosowanych w sposób niezależny, otrzymując następujące wyniki (w gramach): 62; 70; 57; 58; 59; 67; 65; 69; 55; 57; 60; 54; 72; 66; 74. Zakładając, że rozkład wagi jaj jest normalny, zbudować przedział ufności dla wariancji wagi jaj w populacji generalnej. Przyjąć współczynnik ufności 0,96.

4.30. 70% pracowników spośród wylosowanych w sposób niezależny, 260 osób oceniło warunki pracy w swoim zakładzie pracy jako niezadowalające. Przyjmując współczynnik ufności 0,95, zbudować przedział ufności dla frakcji niezadowolonych z warunków pracy w całej populacji pracowników.

4.31. Wariancja plonów pszenicy na 17 losowo wybranych poletkach doświadczalnych wyniosła 4 $(dt/ha)^2$. Zakładając, że plony pszenicy mają rozkład normalny, zbudować – przy współczynniku ufności 0,90 – przedział ufności dla wariancji i odchylenia standardowego populacji generalnej.

4.32. Jaka powinna być minimalna liczebność próby niezbędna do oszacowania odsetka maturzystów zamierzających podjąć studia, jeśli w próbie liczącej 40 uczniów, 70% spośród nich zamierza studiować? Przyjąć współczynnik ufności 0,90 i maksymalny błąd szacunku równy 5%.

4.33. Rozkład stażu pracy pracowników w pewnej instytucji jest rozkładem normalnym $N(m, 3)$. Zbadano 5 losowo wybranych pracowników tego zakładu i stwierdzono, że ich staż pracy wynosił (w latach): 10; 12; 8; 15; 10. Ilu co najmniej pracowników należy wylosować do próby, aby przy współczynniku ufności 0,95 oszacować średni staż pracy ogółu pracowników, otrzymując przedział ufności o długości nieprzekraczającej 4 lat?

4.34. W spółdzielni mieszkaniowej „Kolejarz" przeprowadzono analizę powierzchni mieszkań, otrzymując:

Powierzchnia w m^2	25–35	35–45	45–55	55–65
Liczba mieszkań	10	40	50	100

Zakładając, że rozkład powierzchni mieszkań jest normalny, należy:

a) wyznaczyć przedział ufności dla odsetka mieszkań o powierzchni powyżej 45 m^2 (przyjąć współczynnik ufności 0,95);

b) ustalić minimalną liczebność próby niezbędną do oszacowania przeciętnej powierzchni mieszkania na poziomie ufności 0,95 i przy maksymalnym błędzie szacunku 1 m^2;

c) zbudować 95% przedział ufności dla odchylenia standardowego powierzchni mieszkań.

4.35. Na podstawie wyników badania ankietowego przeprowadzonego w grupie wylosowanych 500 studentów stwierdzono, że 100 z nich nie ma kłopotów finansowych. Jakiego odsetka studentów mających kłopoty finansowe można oczekiwać w całej zbiorowości studentów? Przyjąć poziom istotności 0,95. Ilu studentów należałoby wylosować do próby, aby przy niezmienionej wiarygodności oszacowania zwiększyć dwukrotnie jego dokładność?

4.36. W wylosowanej próbie 200 studentów studiów zaocznych stwierdzono, iż 10% spośród nich liczy mniej niż 20 lat. Jaką liczną próbę należy wylosować, aby z prawdopodobieństwem 0,95 i przy dopuszczalnym błędzie szacunku 1%, oszacować odsetek studentów, którzy nie przekraczają 20 roku życia wśród ogółu studentów?

4.37. Ilu pracowników należy wylosować niezależnie do próby, aby oszacować odsetek osób, których praca jest zgodna z wyuczonym zawodem? Przyjąć współczynnik ufności 0,95 i maksymalny błąd szacunku nie większy niż 3%.

4.38. W celu zbadania czasu dojazdu do pracy w pewnym dużym przedsiębiorstwie wylosowano próbę pilotażową o liczebności 36 pracowników. Z próby tej obliczono: $\bar{x} = 34{,}6$ minuty oraz $\hat{s}^2 = 272{,}8$ $(\text{minuty})^2$. Ilu pracowników należy wylosować do próby zasadniczej, aby przy szacowaniu średniego czasu dojazdu populacji wszystkich pracowników nie popełnić błędu większego niż 2 minuty? Przyjąć współczynnik ufności 0,95.

4.39. Chcemy oszacować odsetek radnych w pewnym mieście, którzy legitymują się wykształceniem wyższym. Z wcześniej przeprowadzonych badań wynika, że frakcja ta wynosi 15%. Ustalić, jaka powinna być minimalna liczebność próby, przy maksymalnym błędzie szacunku 3%. Przyjąć współczynnik ufności 0,95.

4.40. Wiadomo, że przeciętny stopień zróżnicowania czasu rozwiązywania pewnego zadania matematycznego wynosi $\sigma = 5$ minut. Ilu studentów należy wylosować do próby, aby ocenić przeciętny czas rozwiązywania tego zadania z wiarygodnością 0,90 i maksymalnym błędem szacunku 2 minuty?

4.41. Dyrekcja hipermarketu zamierza ustalić, ile czasu spędzają w nim klienci w soboty. W tym celu wylosowano próbę pilotażową, która dała następujące wyniki:

Czas w minutach	2–6	6–10	10–14	14–22	22–30
Liczba klientów	21	58	43	15	7

Przyjmując współczynnik ufności 0,96 oraz maksymalny błąd szacunku nieprzekraczający 0,5 minuty oszacuj, czy wylosowana próba wstępna jest wystarczająca do określenia średniego czasu spędzonego na zakupach w populacji wszystkich klientów tego hipermarketu.

4.42 Ile osób należy wylosować do próby, aby przy współczynniku ufności 0,96 i maksymalnym błędzie szacunku 5% można było oszacować odsetek klientów sklepów spożywczych używających kart płatniczych przy dokonywaniu zakupów?

4.43. Dzienna liczba klientów obsługiwanych przez jedną kasjerkę w wylosowanych okienkach kasowych placówek pewnego banku kształtowała się następująco: 45; 54; 64; 73; 76; 82; 89; 89; 97; 98. Ile okienek należy wylosować, aby przy współczynniku ufności 0,90 maksymalny błąd szacunku średniej dziennej liczby klientów wynosił 10%?

4.44. Ile osób należy wylosować do próby, aby z maksymalnym błędem szacunku 2% ocenić udział w rynku pewnej firmy, jeżeli wiemy, że w pewnym mieście produkty tej firmy kupuje 28% klientów? Przyjąć współczynnik ufności 0,95.

4.45. Jak liczną próbę faktur należy wylosować do badania, aby przy współczynniku ufności 0,90 maksymalny błąd szacunku frakcji ogółu faktur z błędami nie przekroczył 5%? Według wstępnych ocen, odsetek faktur z błędami nie przekracza 30%.

4.46. Rozkład wielkości stypendiów studenckich jest normalny. Ilu studentów należy wylosować niezależnie do próby, aby przy współczynniku ufności 0,98 zbudować przedział ufności o rozpiętości co najmniej 100 zł dla średniego stypendium pobieranego przez nich? Wiadomo, że odchylenie standardowe stypendium wynosi 63 złote.

4.47. We wstępnej próbie liczącej $n=10$ noworodków odchylenie standardowe wzrostu było równe $\hat{s}=2{,}2$ cm. Jaka duża musi być próba losowa, aby przy współczynniku ufności 0,98 oszacować – z maksymalnym błędem szacunku 1 cm – średni wzrost ogółu noworodków?

4.48. Ile osób należy wylosować niezależnie do próby z maksymalnym błędem szacunku 3% i przy współczynniku ufności 0,90, aby oszacować procent wyborców, którzy poprą kandydaturę polityka XXL w wyborach do Senatu? W poprzednich wyborach kandydaturę tę poparło 30% wyborców.

4.49. Jaka powinna być minimalna liczebność próby losowej gospodarstw domowych dla oszacowania średniego poziomu wydatków na artykuły żywnościowe, jeśli współczynnik ufności wynosi 0,95, a maksymalny błąd szacunku nie powinien przekraczać 20 złotych? Wiadomo, że rozkład wydatków żywnościowych jest normalny z wariancją równą 6889 $(zł)^2$

4.50. Firma CCCC zamierza ustalić frakcję klientów będących potencjalnymi nabywcami jej produktów. Szacuje się, że odsetek klientów chcących nabyć produkty tej firmy wynosi 75%. Ilu – co najmniej – klientów powinno znaleźć się w próbie losowej przy maksymalnym błędzie szacunku 10% i współczynniku ufności 0,95?

Rozdział V

WERYFIKACJA HIPOTEZ STATYSTYCZNYCH

5.1. Zasady testowania hipotez statystycznych

Hipotezą statystyczną nazywamy każdy sąd (przypuszczenie) dotyczący postaci rozkładu cechy w populacji generalnej (**hipotezy nieparametryczne**) lub wartości jego parametrów (**hipotezy parametryczne**). O prawdziwości lub fałszywości tego sądu wnioskujemy na podstawie próby losowej pobranej z populacji generalnej.

Zbiór możliwych sądów (przypuszczeń) jest zazwyczaj ograniczony, gdyż z reguły mamy pewne informacje o populacji generalnej. Przykładowo, jeśli wiemy, że badana zmienna losowa w populacji generalnej podlega rozkładowi normalnemu (jest zmienną losową ciągłą) – to zbiór sądów jest ograniczony do tego rodzaju rozkładów, różniących się jedynie wartościami parametrów m i σ. W takim przypadku, bezpodstawne jest formułowanie hipotezy, że populacja generalna (ściślej: zmienna losowa w populacji generalnej) ma rozkład dwumianowy. Informacje a priori o populacji generalnej wyznaczają tzw. **zbiór hipotez dopuszczalnych**, który oznaczamy symbolem Ω. Zbiór ten jest zbiorem rozkładów, o których wiemy, że mogą charakteryzować populację generalną. Jeśli elementy zbioru Ω różnią się między sobą co najwyżej wartościami parametrów, to formułowane hipotezy nazywamy **parametrycznymi**. W przypadku, gdy elementy zbioru Ω różnią się nie tylko wartościami parametrów, ale również postacią funkcyjną – stawiane hipotezy nazywamy **nieparametrycznymi**. Wnioskowanie statystyczne o słuszności sformułowanej hipotezy nazywamy **sprawdzaniem, testowaniem** lub **weryfikacją** hipotezy.

W procedurze sprawdzania hipotez statystycznych podstawowe znaczenie ma **hipoteza zerowa** (**sprawdzana, weryfikowana**), oznaczana symbolem H_0. Oprócz niej formułuje się **hipotezę alternatywną** (H_1), która jest odpowiednim zaprzeczeniem hipotezy zerowej. Hipotezę alternatywną uznaje się za prawdziwą w przypadku odrzucenia hipotezy zerowej.

Weryfikacji hipotez statystycznych dokonujemy na podstawie wyników próby losowej. Istnieje więc możliwość popełnienia błędu. Wyniki próby mogą być takie, że H_0 uznamy za fałszywą i odrzucimy ją, chociaż w istocie jest ona prawdziwa lub też przyjmiemy H_0, która jest fałszywa. Odrzucenie hipotezy H_0, gdy jest ona prawdziwa, nosi nazwę **błędu I rodzaju** (poziomu istotności). Przyjęcie hipotezy H_0, gdy jest ona fałszywa, określamy mianem **błędu II rodzaju**. Rozmiary tych błędów oceniane są jako prawdopodobieństwa: α (prawdopodobieństwo popełnienia błędu I rodzaju) i β (prawdopodobieństwo popełnienia błędu II rodzaju). Rodzaje błędów popełnianych przy weryfikacji hipotez statystycznych prezentuje tab. 5.1.

Tab. 5.1. Decyzje i ich konsekwencje w teście sprawdzającym hipotezę H_0

Sytuacje	Decyzja	
	przyjęcia H_0	odrzucenia H_0
Hipoteza zerowa prawdziwa	decyzja prawidłowa	błąd I rodzaju (α)
Hipoteza zerowa fałszywa	błąd II rodzaju (β)	decyzja prawidłowa

Źródło: A. Zeliaś, *Metody statystyczne*, PWE, Warszawa 2000, s. 258.

Narzędziem służącym do weryfikacji hipotez statystycznych na podstawie wyników próby losowej są **testy statystyczne**. **Testem statystycznym** nazywamy regułę postępowania służącą do podejmowania – na podstawie wyników próby losowej – decyzji o przyjęciu lub odrzuceniu sprawdzanej hipotezy. Tak więc test statystyczny jest regułą rozstrzygającą, jakie wyniki próby dają podstawę do uznania sprawdzanej hipotezy za prawdziwą, a jakie za fałszywą. W zależności od tego, czy test służy do weryfikacji hipotezy parametrycznej, czy nieparametrycznej – nazywany jest **testem parametrycznym** lub **nieparametrycznym**. Istotnym założeniem występującym przy testach parametrycznych jest założenie normalności populacji, z której pobiera się próbę losową.

Załóżmy, że W oznacza zbiór wszystkich możliwych wyników n-elementowej próby (czyli przestrzeń prób), natomiast $W_n = (x_1, x_2, \ldots, x_n)$ – pewną próbę (punkt w przestrzeni prób). Konstrukcja testu statystycznego polega na określeniu takiego obszaru przestrzeni próby w, że jeśli $W_n \in w$ (tzn. wynik próby znajdzie się w tym obszarze), to sprawdzaną hipotezę H_0 odrzucamy, natomiast gdy $W_n \in W - w$ – to H_0 przyjmujemy. Hipoteza zerowa może – przykładowo – mieć postać: $H_0: Q = Q_0$, co czytamy: „hipoteza H_0, że parametr populacji $Q = Q_0$". Hipotezą alternatywną do H_0 jest tu $H_1: Q \neq Q_0$ (parametr populacji Q różni się do Q_0). W zależności od potrzeb H_1 może również mieć postać: $H_1: Q > Q_0$ lub $H_1: Q < Q_0$.

Obszar w nazywamy **obszarem krytycznym** lub obszarem odrzucenia H_0, natomiast W–w **obszarem przyjęcia** hipotezy zerowej. W tej sytuacji prawdopodobieństwa popełnienia błędów I rodzaju (α) i II rodzaju (β) możemy zapisać jako:

$$P(W_n \in w | H_0) = \alpha \tag{5.1}$$

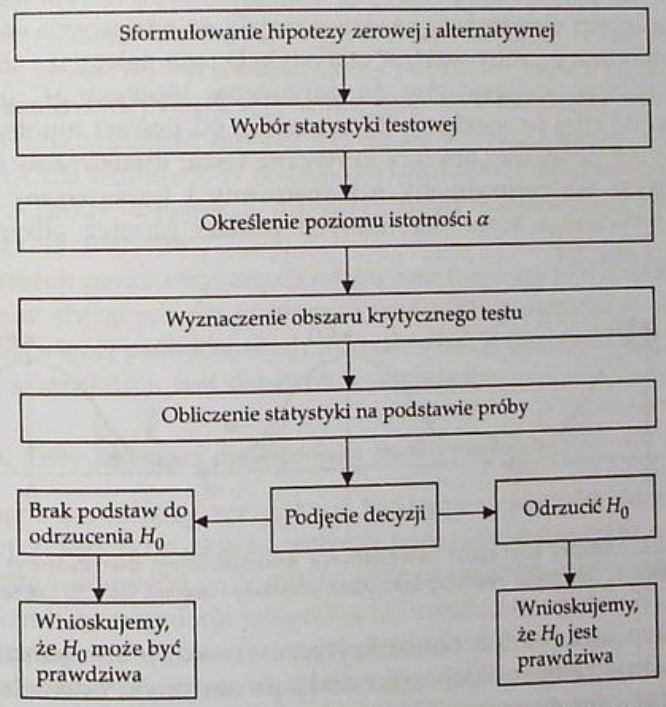
oraz

$$P(W_n \in (W - w) | H_1) = \beta. \tag{5.2}$$

Najlepszy byłby taki test statystyczny, w którym zarówno prawdopodobieństwo α, jak i β byłyby jak najmniejsze. Jednakże przy danej n-elementowej próbie losowej zmniejszenie α powoduje wzrost β i odwrotnie.

W praktycznych zastosowaniach testów statystycznych za bardziej niebezpieczne uznaje się popełnienie błędu I rodzaju (α). Wartości α są przyjmowane z góry, najczęściej jako prawdopodobieństwa: 0,01; 0,02; 0,05; 0,1. Testy, które kontrolują jedynie błąd I rodzaju (zwany poziomem istotności), a nie uwzględniają konsekwencji popełnienia błędu II rodzaju, nazywane są **testami istotności**. Testy te pozwalają na odrzucenie hipotezy H_0 z małym ryzykiem popełnienia błędu I rodzaju (α), a nie pozwalają na jej przyjęcie, gdyż nie wiemy, jak duże może być prawdopodobieństwo popełnienia błędu II rodzaju. Tak więc w testach istotności może być podjęta jedna z dwóch decyzji: H_0 odrzucamy lub stwierdzamy, że brak jest podstaw do jej odrzucenia.

Kolejność postępowania przy weryfikacji hipotez testami istotności przedstawia schemat 5.1.



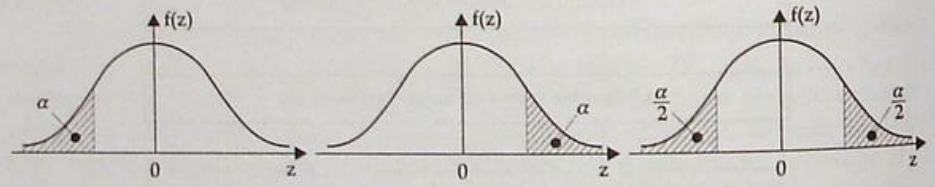
Schemat 5.1. Przebieg weryfikacji hipotez testami istotności
Źródło: A. Balicki, W. Makać, *Metody wnioskowania statystycznego*, Wydawnictwo Uniwersytetu Gdańskiego, Gdańsk 1997, s. 136.

Przy weryfikacji hipotez należy, w pierwszej kolejności, sformułować **hipotezę zerową** (H_0) i **alternatywną** (H_1). W hipotezie zerowej zakłada się, że między analizowanymi parametrami lub rozkładami **nie ma żadnej różnicy**. Przy wnioskowaniu o parametrach hipotezę zerową zapisujemy – w sposób ogólny – jako: $H_0: Q = Q_0$. Hipoteza przeciwstawna do weryfikowanej (hipoteza alternatywna – H_1) może być zapisana w trojaki sposób: $H_1: Q \neq Q_0$, $H_1: Q > Q_0$ lub $H_1: Q < Q_0$.

W kolejnym kroku weryfikacji hipotez budujemy określoną statystykę U_n zwaną **statystyką testową** lub **funkcją testową**. Statystyka ta jest funkcją wyników próby losowej: $U_n = f(x_1, x_2, \ldots, x_n)$. Rozkład statystyki testowej U zależy – przy założeniu, że H_0 jest prawdziwa – od liczebności próby, postaci H_0 oraz rozkładu badanej cechy w populacji generalnej. W szczególności, statystyka U może mieć rozkład normalny, t-Studenta, F-Snedecora czy chi-kwadrat.

Prawdopodobieństwo popełnienia błędu I rodzaju, oznaczane symbolem α i określane mianem poziomu istotności, przyjmowane jest arbitralnie. Żądamy, aby ryzyko popełnienia błędu I rodzaju było jak najmniejsze. Stąd też za α przyjmuje się wartości bliskie zera, najczęściej $\alpha \leq 0{,}1$.

W następnym etapie budowy testu należy wyznaczyć obszar krytyczny i obszar przyjęcia hipotezy H_0. Obszarem krytycznym testu jest taki podzbiór wartości statystyki U_n, który pozwala na odrzucenie sprawdzanej H_0. Jeśli obliczona z próby wartość statystyki U_n nie należy do obszaru krytycznego, to brak jest podstaw do odrzucenia hipotezy H_0 (co nie jest równoznaczne z jej przyjęciem). W zależności od postaci hipotezy alternatywnej , wyróżniamy **trzy obszary krytyczne testu: dwustronny** (zwany też **symetrycznym** lub **centralnym**), **prawostronny** i **lewostronny**. Położenie obszaru krytycznego testu dla różnych postaci hipotez alternatywnych przedstawia rys. 5.1.



5.1. Obszary krytyczne (lewostronny, prawostronny, dwustronny)
Źródło: M. Sobczyk, *Statystyka. Podstawy teoretyczne, przykłady i zadania*, UMCS, Lublin 2000, s. 181.

Jak wynika z rys. 5.1, obszar krytyczny znajduje się zawsze na krańcach rozkładu. Od pozostałej części rozkładu statystyki oddzielają go tzw. wartości krytyczne testu z_α. Wartości krytyczne z_α odczytuje się z tablic rozkładu statystyki, przy zadanym poziomie istotności α, stosownie do sposobu sformułowania hipotezy alternatywnej.

Statystykę testową oblicza się z wyników próby losowej. W testach istotności, statystyki testowe (mające zazwyczaj rozkład normalny, t-Studenta lub graniczny rozkład normalny) oblicza się następująco:

$$\text{statystyka testowa} = \frac{\text{statystyka obliczona z próby} - \text{hipotetyczna wartość parametru}}{\text{odchylenie standardowe rozkładu statystyki}} = \frac{z_n - Q_0}{D(Z_n)} \quad (5.3)$$

Wyznaczoną na podstawie próby losowej wartość statystyki testowej porównuje się z wartością krytyczną testu odczytaną z odpowiednich tablic statystycznych. Jeśli wartość statystyki z próby znalazła się w obszarze krytycznym, to H_0 należy odrzucić jako nieprawdziwą. Prawdziwa jest wówczas H_1, która jest przeciwstawna do H_0. Jeśli natomiast wartość statystyki z próby znalazła się poza obszarem krytycznym, to brak jest podstaw do odrzucenia H_0. Oznacza to, że hipoteza zerowa może być prawdziwa. W szczególnym przypadku, gdy wartość statystyki z próby jest dokładnie równa wartości krytycznej, podejmuje się decyzję o odrzuceniu H_0.

5.2. Parametryczne testy istotności

W tym punkcie skupimy uwagę na weryfikacji hipotez statystycznych dotyczących podstawowych parametrów populacji jednowymiarowych, tj. **wartości oczekiwanej** (średniej), **wskaźnika struktury** oraz **wariancji**. Do weryfikacji odpowiednich hipotez używa się wówczas testów: dla średniej, dla proporcji, dla wariancji. W testach tych porównuje się oceny parametrów uzyskane z próby losowej z hipotetycznymi wielkościami parametrów, pełniącymi rolę wzorców.

W ramach parametrycznych testów istotności wyodrębnia się również grupę testów służących do porównywania ocen parametrów uzyskanych z dwóch lub więcej prób losowych (test dla dwóch średnich, test dla dwóch wskaźników struktury, test dla dwóch wariancji).

5.2.1. Testy istotności dla średniej i dwóch średnich

Postępowanie przy weryfikacji hipotezy o wartości oczekiwanej m w populacji zależy od rozkładu cechy w populacji, znajomości wartości odchylenia standardowego w populacji oraz liczebności próby losowej.

Załóżmy, że populacja generalna ma rozkład normalny $N(m,\sigma)$ o nieznanej średniej m oraz znanym odchyleniu standardowym σ. Z populacji tej wylosowano n-elementową próbę w celu zweryfikowania hipotezy $H_0: m = m_0$, wobec hipotezy alternatywnej $H_1: m \neq m_0$, gdzie m_0 jest hipotetyczną średnią w populacji.

Sprawdzianem hipotezy zerowej jest statystyka testowa Z o postaci:

$$Z=\frac{\overline{X}-m_0}{\frac{\sigma}{\sqrt{n}}}=\frac{\overline{X}-m_0}{\sigma}\sqrt{n}, \tag{5.4}$$

gdzie $\overline{X}$ jest średnią z próby pobranej z populacji. Średnia ta ma rozkład normalny $N\left(m,\frac{\sigma}{\sqrt{n}}\right)$. Jeśli H_0 jest prawdziwa, to statystyka testowa określona wzorem (5.4) ma rozkład $N(0,1)$.

Statystyka testowa Z jest zmienną losową, a jej realizacja – obliczona z wyników próby losowej – jest równa:

$$z=\frac{\bar{x}-m_0}{\sigma}\sqrt{n}. \tag{5.5}$$

Realizację z porównujemy z wartością krytyczną odczytaną z tablic dystrybuanty rozkładu normalnego $N(0,1)$ przy założonym poziomie istotności α. Sposób odczytu wartości krytycznej z_α zależy od postaci hipotezy alternatywnej. Przy dwustronnym obszarze krytycznym wartości z_α odczytujemy w taki sposób, aby spełniona była relacja:

$$F(z_\alpha)=1-\frac{\alpha}{2}. \tag{5.6}$$

Przy $H_1: m\neq m_0$ zachodzi: $P(|z|\geq z_\alpha)=\alpha$. Oznacza to, że hipotezę zerową odrzucamy na korzyść hipotezy alternatywnej, gdyż $z\geq z_\alpha$ lub $z\leq -z_\alpha$. Oznacza to, że różnica między średnią z próby $\bar{x}$ a hipotetyczną wartością średniej w populacji m_0 jest zbyt duża (**statystycznie istotna**). W przeciwnym przypadku, tzn. gdy $|z|<z_\alpha$, stwierdzamy, że wyniki próby nie dają podstaw do odrzucenia H_0.

Hipoteza alternatywna może mieć również postać: $H_1: m<m_0$ (**lewostronny obszar krytyczny**) lub $H_1: m>m_0$ (**prawostronny obszar krytyczny**). Przy lewostronnym obszarze krytycznym odrzucenie H_0 następuje wówczas, gdy spełniona jest nierówność: $z\leq -z_\alpha$. Wartość z_α jest odczytywana z tablic dystrybuanty standaryzowanego rozkładu normalnego według zasady: $F(z_\alpha)=\alpha$. Przy prawostronnym obszarze krytycznym odrzucenie H_0 następuje wtedy, gdy $z\geq z_\alpha$. Wartość krytyczną z_α odczytujemy wówczas z tablicy dystrybuanty rozkładu $N(0,1)$ zgodnie z relacją $F(z_\alpha)=1-\alpha$. Tak więc zbiory wartości krytycznych (Z_k) przy różnych postaciach hipotez alternatywnych są określone następująco:

$Z_k=(-\infty,-z_\alpha]\cup[z_\alpha,\infty)$ przy $H_1: m\neq m_0$,

$Z_k=[z_\alpha,\infty)$ przy $H_1: m>m_0$,

$Z_k=(-\infty,-z_\alpha]$ przy $H_1: m<m_0$.

W przypadku, gdy z populacji o nieznanej średniej m i nieznanym odchyleniu standardowym σ pobrano próbę o liczebności $n>30$, średnia $\overline{X}$ ma asymptotyczny rozkład normalny o parametrach $N\left(m,\frac{S}{\sqrt{n}}\right)$, gdzie S jest estymatorem odchylenia standardowego. Jeśli H_0 jest prawdziwa, to statystyka testowa przyjmuje wówczas postać:

$$Z=\frac{\overline{X}-m_0}{\frac{S}{\sqrt{n}}}=\frac{\overline{X}-m_0}{S}\sqrt{n}. \tag{5.7}$$

Statystyka (5.7) ma asymptotyczny rozkład normalny $N(0,1)$.

Zdarza się również konieczność weryfikacji hipotezy o średniej m na podstawie wyników małej próby wylosowanej z populacji, tzn. $n\leq 30$. Zakłada się przy tym, że rozkład populacji jest normalny $N(m,\sigma)$ o nieznanej średniej m i nieznanym odchyleniu standardowym σ. W tym przypadku statystyka testowa jest definiowana wzorem:

$$T=\frac{\overline{X}-m_0}{S}\sqrt{n-1}=\frac{\overline{X}-m_0}{\hat{S}}\sqrt{n}, \tag{5.8}$$

gdzie: $S=\sqrt{\frac{1}{n}\sum_{i=1}^{n}\left(X_i-\overline{X}\right)^2}$ oraz $\hat{S}=\sqrt{\frac{1}{n-1}\sum_{i=1}^{n}\left(X_i-\overline{X}\right)^2}$.

Przy założeniu prawdziwości H_0, statystyka (5.8) ma rozkład t-Studenta o $n-1$ stopniach swobody. Wartość krytyczną testu t_α odczytujemy wówczas z tablic rozkładu Studenta przy założonym poziomie istotności α oraz $n-1$ stopniach swobody. Tablice te są zazwyczaj zbudowane dla dwustronnego obszaru krytycznego. Chcąc odczytać wartość krytyczną testu dla obszaru jednostronnego (prawostronnego lub lewostronnego), należy podwoić poziomi istotności α.

Decyzja dotycząca sprawdzanej H_0 jest podejmowana – przy uwzględnieniu postaci H_1 – tak, aby spełnione były relacje:

1) dla $H_1: m\neq m_0$ zachodzi relacja $P(|t|\geq t_{\alpha,n-1})=\alpha$. Odrzucenie H_0 następuje zatem wtedy, gdy wartość statystyki obliczona z próby $t\geq t_{\alpha,n-1}$ lub $t\leq -t_{\alpha,n-1}$;

2) dla $H_1: m>m_0$ zachodzi relacja $P(t\geq t_{\alpha,n-1})=\alpha$, stąd też wartość statystyki obliczona z próby $t\geq t_{\alpha,n-1}$ powoduje odrzucenie H_0;

3) dla $H_1: m<m_0$ mamy: $P(t\leq -t_{\alpha,n-1})=\alpha$. Oznacza to, że wartość statystyki obliczona z próby $t\leq -t_{\alpha,n-1}$ powoduje odrzucenie H_0.

Przypuszcza się, że średni stan na kontach osobistych osób pracujących wynosi – na początku miesiąca – 2,5 tys. PLN. Wyniki z próby składającej się z 18 kont dostarczyły następujących informacji: $\bar{x}=1{,}9$ tys. PLN oraz $s=0{,}47$ tys. PLN. Zakłada się, że rozkład środków pieniężnych na kontach osobistych jest normalny o nieznanych parametrach. Na poziomie istotności 0,05 zweryfikować hipotezę, że wylosowane konta osobiste należą do populacji o średnim stanie konta równym 2,5 tys. PLN.

W przypadku tym mamy zweryfikować hipotezę zerową $H_0: m=2{,}5$ wobec hipotezy alternatywnej $H_1: m<2{,}5$. Znak mniejszości użyty w hipotezie alternatywnej wynika z zaobserwowanej relacji między średnią z próby a hipotetyczną średnią w populacji.

Wybór statystyki testowej zależy od liczebności próby oraz od tego, czy parametr σ w populacji generalnej jest znany czy też nie. W naszym przypadku σ jest nieznane, a próba mała. Stąd też wykorzystujemy statystykę o postaci:

$$t = \frac{\bar{x} - m_0}{s}\sqrt{n-1} = \frac{1{,}9 - 2{,}5}{0{,}47}\sqrt{18-1} = -5{,}26.$$

Statystyka ta ma – przy założeniu prawdziwości H_0 – rozkład t-Studenta o $n-1=17$ stopniach swobody. Z tablic rozkładu Studenta odczytujemy przy podwojonym poziomie istotności $2\alpha = 0{,}1$ (jednostronny obszar krytyczny) i 17 stopniach swobody wartość krytyczną $t_{0,t17} = -1{,}74$. Ze względu na to, że $t = -5{,}26 < t_{0,t17} = -1{,}74$, sprawdzaną hipotezę zerową należy odrzucić. Oznacza to, że średni stan kont osobistych na początku miesiąca jest istotnie mniejszy od 2,5 tys. PLN. Różnica ta jest statystycznie istotna.

Podejmowanie decyzji ekonomicznych jest zazwyczaj poprzedzone porównywaniem alternatywnych możliwości. Przykładowo, właściciel sklepu spożywczego porównuje warunki dostaw towarów (ceny, terminy dostaw, sposób płatności itp.) oferowane przez różnych kontrahentów. Zaobserwowane w wyniku tych porównań różnice mogą mieć charakter przypadkowy lub też mogą być statystycznie istotne. W celu rozstrzygnięcia, z którą z tych alternatyw mamy do czynienia, można wykorzystać metody wnioskowania statystycznego, w tym np. testy dla dwóch średnich, dwóch wskaźników struktury czy dwóch wariancji. Zakładamy przy tym, że z badanych populacji generalnych pobrano niezależne próby losowe proste. Oznacza to, że wyniki uzyskane w jednej próbie nie mają wpływu na rezultaty otrzymane w próbie drugiej.

W praktycznych zastosowaniach metod wnioskowania statystycznego dość często zachodzi konieczność **porównania dwóch średnich** m_1 i m_2 w dwóch populacjach (np. porównanie starej i nowej technologii produkcji wyrobów, porównanie populacji zdrowych ze zbiorowością chorych pod względem określonej cechy itp.). Weryfikuje się wówczas hipotezę zerową $H_0: m_1 = m_2$ ($H_0: m_1 - m_2 = 0$) wobec odpowiedniej hipotezy alternatywnej H_1. Sposób postępowania przy weryfikacji H_0 jest tu różny, w zależności od następujących okoliczności:

1) czy znane są wariancje w populacjach generalnych?
2) czy rozkłady badanych zmiennych w populacjach są normalne?
3) czy wariancje w populacjach są jednakowe?
4) jaka jest liczebność próby: duża czy mała?

Załóżmy, że rozważamy strukturę dwóch zbiorowości statystycznych (populacji) pod względem cechy X. Populacje te mają rozkłady normalne $N(m_1, \sigma_1)$ oraz $N(m_2, \sigma_2)$, przy czym odchylenia standardowe σ_1 oraz σ_2 są znane. Wartości średnich m_1 i m_2 nie są znane, a naszym zadaniem jest weryfikacja hipotezy o równości tych średnich: $H_0: m_1 = m_2$ lub $H_0: m_1 - m_2 = 0$. W celu weryfikacji H_0 z każdej populacji wylosowano próby o liczebnościach

n_1 i n_2. Zmienna losowa, będąca różnicą średnich $\bar{X}_1 - \bar{X}_2$, ma również rozkład normalny o parametrach $N\left(m_1 - m_2; \sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}\right)$. Przy założeniu prawdziwości H_0, gdy znane są wariancje w obu populacjach, statystyką testową jest zmienna standaryzowana o rozkładzie $N(0,1)$:

$$Z = \frac{\bar{X}_1 - \bar{X}_2}{\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}}. \tag{5.9}$$

Obszar krytyczny jest tu budowany w taki sam sposób, jak w przypadku weryfikacji H_0 o średniej w populacji.

Jeśli populacje mają rozkłady normalne $N(m_1, \sigma_1)$ oraz $N(m_2, \sigma_2)$ o nieznanych lecz różnych wariancjach $\sigma_1^2 \neq \sigma_2^2$, a próby są duże – statystyka testowa służąca do weryfikacji H_0 o równości średnich w populacjach jest definiowana wzorem:

$$Z = \frac{\bar{X}_1 - \bar{X}_2}{\sqrt{\frac{S_1^2}{n_1} + \frac{S_2^2}{n_2}}}, \tag{5.10}$$

gdzie S_1^2 oraz S_2^2 są wariancjami z prób. Statystyka (5.10) ma – przy założeniu prawdziwości H_0 – rozkład $N(0,1)$.

Przy nieznanych, ale różnych wariancjach w populacjach generalnych i małych próbach ($n_1 < 30$ lub $n_2 < 30$) posługujemy się statystyką o postaci[1]:

$$T = \frac{\bar{X}_1 - \bar{X}_2}{\sqrt{\frac{S_1^2}{n_1 - 1} + \frac{S_2^2}{n_2 - 2}}}. \tag{5.11}$$

Rozkład statystyki T zależy od liczebności prób ($n_1 - 1$ oraz $n_2 - 1$) oraz od ilorazu wariancji w populacjach (z założenia iloraz ten jest nieznany). Odczytując wartość krytyczną z tablic rozkładu Studenta, należy uprzednio ustalić liczbę stopni swobody, która jest równa:

$$\text{liczba stopni swobody} = \frac{\left(\frac{s_1^2}{n_1 - 1} + \frac{s_2^2}{n_2 - 1}\right)^2}{\frac{\left(\frac{s_1^2}{n_1 - 1}\right)^2}{n_1 - 1} + \frac{\left(\frac{s_2^2}{n_2 - 1}\right)^2}{n_2 - 1}}, \tag{5.12}$$

[1] A. Balicki, W. Makać, *Metody wnioskowania statystycznego*, Wydawnictwo Uniwersytetu Gdańskiego, Gdańsk 1997, s. 164.

gdzie s_1^2 oraz s_2^2 są wariancjami obliczonymi z konkretnych prób losowych.

Przy weryfikacji H_0 o równości średnich w dwóch populacjach może wystąpić również sytuacja, w której dwie populacje mają rozkłady normalne o nieznanych, ale równych wariancjach ($\sigma_1^2 = \sigma_2^2$). W takim przypadku, jeśli próby losowe są małe, posługujemy się statystyką testową o rozkładzie t-Studenta z $n_1 + n_2 - 2$ stopniami swobody:

$$T = \frac{\overline{X}_1 - \overline{X}_2}{\sqrt{\frac{n_1 S_1^2 + n_2 S_2^2}{n_1 + n_2 - 2}\left(\frac{1}{n_1} + \frac{1}{n_2}\right)}}. \qquad (5.13)$$

Jeśli próby są duże, to przy tych samych założeniach, w charakterze sprawdzianu $H_0: m = m_0$, występuje statystyka definiowana wzorem:

$$Z = \frac{\overline{X}_1 - \overline{X}_2}{\sqrt{n_1 S_1^2 + n_2 S_2^2}}\sqrt{n_1 n_2}. \qquad (5.14)$$

We wszystkich wymienionych wyżej przypadkach, weryfikacja H_0 przebiega w analogiczny sposób jak w testach parametrycznych dotyczących średniej w populacji. Hipotezę zerową odrzucamy – na przyjętym poziomie istotności – jeśli empiryczna wartość wykorzystywanej przy weryfikacji statystyki testowej należy do zbioru krytycznego. W przeciwnym przypadku stwierdzamy, że nie ma podstaw do odrzucenia H_0.

Załóżmy, że w celu porównania przeciętnego stażu pracy pracowników w dwóch zakładach należących do tej samej branży wylosowano – z każdego z nich – próby o liczebnościach $n_1 = 60$ pracowników oraz $n_2 = 82$ pracowników. Z wyników prób obliczono: $\bar{x}_1 = 6{,}8$ lat, $s_1 = 1{,}7$ lat, $\bar{x}_2 = 8{,}2$ lat, $s_2 = 2{,}5$ lat. Na poziomie istotności 0,01 należy zweryfikować hipotezę o równości średnich staży pracy w populacji pracowników obydwu zakładów. Zakładamy, że wariancje w populacjach są jednakowe.

Hipoteza zerowa zakłada, że średnie w obu populacjach, z których pochodzą próby, są jednakowe, a hipoteza alternatywna stwierdza, że średnie różnią się między sobą. Mamy więc:

$$H_0: m_1 = m_2,$$
$$H_0: m_1 \neq m_2.$$

W przykładzie mamy do czynienia z dużymi próbami losowymi. Wariancje w populacjach – jakkolwiek nieznane – są jednakowe. W takim przypadku w charakterze sprawdzianu H_0 użyjemy statystyki testowej (5.14). Empiryczna wartość tej statystyki jest równa:

$$z = \frac{|6{,}8 - 8{,}2|}{\sqrt{60 \cdot 1{,}7^2 + 82 \cdot 2{,}5^2}}\sqrt{60 \cdot 82} = 3{,}75.$$

Przy poziomie istotności 0,01 i dwustronnym obszarze krytycznym, odczytana z tablic dystrybuanty rozkładu normalnego $N(0,1)$ wartość wynosi: $z_{0,01} = 2{,}58$. Ze względu na to, że $|z| = 3{,}75 > z_\alpha = 2{,}58$, statystyka obliczona z próby znalazła się w obszarze krytycznym. Przyjmujemy więc H_1, głoszącą, że średnie staże pracy pracowników w populacjach badanych zakładów różnią się istotnie.

5.2.2. Testy istotności dla wariancji i dwóch wariancji

Wariancja jest podstawowym parametrem opisującym rozkład cechy w populacji generalnej. Parametr ten charakteryzuje rozkład w sposób bezpośredni (tak jest w przypadku rozkładu normalnego) bądź też jest wykorzystywany przy ocenie takich własności rozkładu, jak asymetria czy kurtoza.

Testy istotności dla wariancji w populacji o rozkładzie normalnym opierają się na estymatorach wariancji z próby (prób) losowej S^2 oraz $\hat{S}^2$ określonych następująco:

$$S^2 = \frac{1}{n}\sum_{i=1}^{n}\left(X_i - \overline{X}\right)^2 \text{ lub } \hat{S}^2 = \frac{1}{n-1}\sum_{i=1}^{n}\left(X_i - \overline{X}\right)^2. \qquad (5.15)$$

Jak łatwo zauważyć:

$$nS^2 = \sum_{i=1}^{n}\left(X_i - \overline{X}\right)^2 \text{ oraz } (n-1)\hat{S}^2 = \sum_{i=1}^{n}\left(X_i - \overline{X}\right)^2. \qquad (5.16)$$

Niech badana cecha w populacji ma rozkład normalny $N(m,\sigma)$ o nieznanych parametrach. Z populacji tej pobrano n-elementową próbę losową prostą, na podstawie której należy zweryfikować hipotezę zerową $H_0: \sigma^2 = \sigma_0^2$, gdzie σ^2 jest wariancją w populacji, natomiast σ_0^2 – hipotetyczną wariancją. Hipoteza alternatywna może mieć różną postać, a mianowicie: $H_1: \sigma^2 \neq \sigma_0^2$; $H_1: \sigma^2 > \sigma_0^2$ lub $H_1: \sigma^2 < \sigma_0^2$. Najczęściej jednak przyjmuje się hipotezę alternatywną z prawostronnym obszarem krytycznym, gdyż zazwyczaj sytuację, w której wariancja zmiennej w populacji jest wysoka, uznaje się za niekorzystną. Przy prawostronnym obszarze krytycznym, wartość krytyczną χ_α^2 odczytujemy z tablic rozkładu chi-kwadrat przy ustalonym poziomie istotności α i $n-1$ stopniach swobody. Jeśli spełniona jest relacja $\chi^2 \geq \chi_\alpha^2$ – hipotezę zerową odrzucamy. W przeciwnym przypadku stwierdzamy, że brak jest podstaw do odrzucenia hipotezy zerowej.

W przypadku $n < 30$ sprawdzianem hipotezy zerowej $H_0: \sigma^2 = \sigma_0^2$ jest statystyka określona wzorem:

$$\chi^2 = \frac{nS^2}{\sigma_0^2} = \frac{(n-1)\hat{S}^2}{\sigma_0^2} = \frac{1}{\sigma_0^2}\sum_{i=1}^{n}\left(X_i - \overline{X}\right)^2. \qquad (5.17)$$

Rozkład chi-kwadrat zmierza – przy wzroście liczby stopni swobody $n-1$ – do granicznego rozkładu normalnego. Zbieżność ta jest wystarczająco duża już dla $n > 30$. Korzysta się wówczas z przekształcenia:

$$Z = \sqrt{2\chi^2} - \sqrt{2(n-1)-1} = \sqrt{2\chi^2} - \sqrt{2n-3}. \qquad (5.18)$$

Zmienna Z ma rozkład asymptotycznie normalny $N(0,1)$. Wartość krytyczną testu odczytujemy zatem z tablic dystrybuanty standaryzowanego rozkładu normalnego.

Przy stosowaniu testu t-Studenta dla małych prób przy weryfikacji hipotezy o równości średnich w dwóch populacjach wymagana jest równość wariancji w porównywanych zbiorowościach generalnych. Do weryfikacji hipotezy zerowej $H_0: \sigma_1^2 = \sigma_0^2$ wobec hipotezy alternatywnej $H_1: \sigma_1^2 > \sigma_0^2$ (taka postać H_1 jest najczęściej wykorzystywana) używa się wariancji $\hat{S}_1^2$ oraz $\hat{S}_2^2$, obliczanych z dwóch niezależnych prób losowych o liczebnościach równych odpowiednio n_1 i n_2. Do weryfikacji H_0 służy statystyka testowa o postaci:

$$F = \frac{\hat{S}_1^2}{\hat{S}_2^2}. \tag{5.19}$$

Statystyka (5.19) ma rozkład F-Snedecora (krótko: rozkład F) z $n_2 - 1$ oraz $n_1 - 1$ stopniami swobody. Wartość krytyczną testu odczytujemy z tablic rozkładu F-Snedecora przy ustalonym poziomie istotności α i liczbie stopni swobody $n_1 - 1$ oraz $n_2 - 1$. W przypadku gdy $F \geq F_\alpha$, hipotezę zerową odrzucamy na korzyść hipotezy alternatywnej.

Załóżmy, że trener lekkoatletów zamierza porównać regularność wyników osiągniętych w skoku w dal przez dwóch sportowców. W tym celu wylosował wyniki 8 skoków pierwszego zawodnika oraz 9 wyników drugiego sportowca. Wyniki losowania były następujące (w m):

Pierwszy zawodnik	7,48	7,88	8,00	7,15	7,26	7,33	7,71	7,18	–
Drugi zawodnik	7,62	7,83	8,03	7,97	7,17	8,08	7,92	7,93	7,58

Należy – na poziomie istotności 0,05 – zweryfikować hipotezę o jednakowej regularności wyników uzyskiwanych przez obydwu zawodników.

Zadaniem naszym jest weryfikacja hipotezy zerowej $H_0: \sigma_1^2 = \sigma_2^2$, wobec hipotezy alternatywnej $H_1: \sigma_1^2 > \sigma_2^2$. Z wyników pobranych prób losowych obliczamy: $\hat{s}_1^2 = 0{,}1072$ oraz $\hat{s}_2^2 = 0{,}0837$. Na podstawie wzoru (5.19) otrzymujemy:

$$F = \frac{0{,}1072}{0{,}0837} = 1{,}28.$$

Dla poziomu istotności $\alpha = 0{,}05$ oraz $n_1 - 1 = 7$ i $n_2 - 1 = 8$ stopni swobody odczytujemy wartość krytyczną $F_\alpha = 3{,}5$. Ponieważ $F = 1{,}28 < F_\alpha = 3{,}5$, stwierdzamy, że brak jest podstaw do odrzucenia H_0, że wariancje wyników osiągniętych w skoku w dal przez obydwu zawodników są jednakowe.

5.2.3. Testy istotności dla frakcji i dwóch frakcji

Test istotności dla frakcji (nazywany również testem dla **proporcji, odsetka, wskaźnika struktury, prawdopodobieństwa**) służy do sprawdzania

hipotezy o udziale w populacji generalnej jednostek posiadających wyróżniony wariant cechy.

W rozpatrywanym teście zakłada się, że populacja generalna ma rozkład binomialny z parametrem p. Parametr ten jest wielkością nieznaną, a określa prawdopodobieństwo, że badana cecha w populacji przyjmuje jedną z dwóch możliwych wartości. Dla takiej populacji chcemy zweryfikować hipotezę zerową, że parametr p ma określoną wartość p_0 ($H_0: p = p_0$). Hipoteza alternatywna może być zapisana w trojaki sposób: $H_1: p \neq p_0$, $H_1: p > p_0$ oraz $H_1: p < p_0$. Sprawdzianem hipotezy zerowej dla próby o liczebności $n \geq 100$ jest następująca statystyka:

$$Z = \frac{\frac{m}{n} - p_0}{\sqrt{\frac{p_0 q_0}{n}}}, \tag{5.20}$$

gdzie m oznacza liczbę elementów wyróżnionych w n-elementowej próbie. Statystyka (5.20) ma – przy założeniu prawdziwości hipotezy zerowej – rozkład asymptotycznie normalny $N(0,1)$. Położenie obszaru krytycznego w rozkładzie tej statystyki zależy od postaci hipotezy alternatywnej, a jego wielkość określa przyjęty poziom istotności α.

Przy badaniu dwu skończonych populacji generalnych ze względu na wyróżnioną cechę, zachodzi konieczność weryfikacji hipotezy o równości wskaźników struktury w obu zbiorowościach. Zakłada się wówczas, że populacje te mają rozkład dwumianowy z parametrami równymi odpowiednio p_1 oraz p_2. Z populacji tych należy pobrać dwie niezależne próby ($n_1 \geq 100$ i $n_2 \geq 100$). Na podstawie wyników tych prób ustala się frakcje:

$$p_1^* = \frac{m_1}{n_1} \text{ oraz } p_2^* = \frac{m_2}{n_2}, \tag{5.21}$$

gdzie m_1 oraz m_2 oznaczają liczbę elementów wyróżnionych odpowiednio w pierwszej i drugiej próbie.

Hipoteza zerowa zakłada tu, że wskaźniki struktury w obu populacjach są identyczne ($H_0: p_1 = p_2$), wobec hipotezy alternatywnej sformułowanej z użyciem jednego z trzech znaków: $\neq$, $>$ lub $<$. Do weryfikacji H_0 wykorzystuje się statystykę testową o postaci:

$$Z = \frac{p_1^* - p_2^*}{\sqrt{\frac{\bar{p}\bar{q}}{n}}}, \tag{5.22}$$

gdzie: $\bar{p} = \frac{m_1 + m_2}{n_1 + n_2}$, $\bar{q} = 1 - \bar{p}$, $n = \frac{n_1 \cdot n_2}{n_1 + n_2}$.

Statystyka (5.22) ma, przy założeniu prawdziwości hipotezy H_0, asymptotyczny (graniczny) rozkład normalny $N(0,1)$. Wartość krytyczną testu odczytujemy więc z tablicy dystrybuanty standaryzowanego rozkładu

normalnego, przy założonym poziomie istotności α (stosownie do postaci hipotezy alternatywnej). Brak podstaw do odrzucenia H_0 oznacza, że obie próby pochodzą z tej samej populacji.

Załóżmy, że mamy zweryfikować przypuszczenie, że palacze papierosów stanowią jednakowy odsetek wśród kobiet i mężczyzn. W celu sprawdzenia tego sądu wylosowano 500 mężczyzn i 600 kobiet. Okazało się, że wśród mężczyzn było 200 palaczy, a wśród kobiet 250. Na poziomie istotności 0,05 zweryfikować hipotezę, że odsetek palących papierosy wśród kobiet i mężczyzn jest jednakowy.

Mamy więc: $H_0: p_1 = p_2$, wobec $H_1: p_1 \neq p_2$. Z wyników prób obliczamy:

$$p_1^* = \frac{m_1}{n_1} = \frac{200}{500} = 0{,}4;\ p_2^* = \frac{m_2}{n_2} = \frac{250}{600} = 0{,}42;\ \bar{p} = \frac{m_1 + m_2}{n_1 + n_2} = \frac{450}{1100} = 0{,}41;\ \bar{q} = 0{,}59;$$

$$n = \frac{500 \cdot 600}{500 + 600} = 272{,}727 \approx 273.$$

Korzystając ze wzoru (5.22) obliczamy wartość z:

$$z = \frac{|0{,}4 - 0{,}42|}{\sqrt{\frac{0{,}41 \cdot 0{,}59}{273}}} = 0{,}671.$$

Z tablic dystrybuanty rozkładu normalnego $N(0,1)$ odczytujemy – dla dwustronnego obszaru krytycznego i poziomu istotności 0,05 – wartość krytyczną $z_\alpha = 1{,}96$, gdyż $F(z_\alpha) = 1 - \frac{0{,}05}{2} = 0{,}975$. Ponieważ zachodzi nierówność: $z = 0{,}671 < z_{0,05} = 1{,}96$, więc z prawdopodobieństwem popełnienia błędu I rodzaju na poziomie 0,05 możemy stwierdzić, że brak jest podstaw do odrzucenia H_0, że palacze papierosów występują jednakowo często wśród kobiet i mężczyzn.

5.3. Nieparametryczne testy istotności

Nieparametryczne testy istotności, w przeciwieństwie do parametrycznych, nie wymagają żadnych założeń co do rozkładu badanej cechy w populacji generalnej. Stąd też są one nazywane **testami niezwiązanymi z rozkładem zmiennej**. W nieparametrycznych testach istotności weryfikowana hipoteza zerowa, odnosząc się do rozkładu badanej cechy w populacji generalnej, nie precyzuje wartości jego parametrów. W szczególności są one stosowane w zastępstwie testów parametrycznych, gdy próby losowe są małe, a postulat normalności rozkładu w populacji nie może być spełniony.

Testy nieparametryczne są przydatne zwłaszcza wtedy, gdy skala pomiaru jest słabsza niż wymagana przy stosowaniu wnioskowania parametrycznego. Testy te wyróżniają się prostotą budowy i mało uciążliwymi rachunkami. Jednakże w porównaniu z testami parametrycznymi, mają zazwyczaj mniejszą moc[2].

W zbiorowości testów nieparametrycznych wyróżnia się trzy podstawowe grupy: **testy zgodności, testy losowości** oraz **testy niezależności**. Testy zgodności i losowości dotyczą zagadnień związanych z analizą struktury zbiorowości. Testy niezależności służą przede wszystkim do wnioskowania o związku między zmiennymi jakościowymi, chociaż można się nimi posługiwać również w odniesieniu do cech ilościowych. Zostaną one zaprezentowane w następnym, szóstym rozdziale, poświęconym metodom analizy współzależności zjawisk.

W odniesieniu do wszystkich wymienionych grup testów nieparametrycznych zakłada się spełnienie następujących warunków[3]:

1) duża liczebność próby n w porównaniu z liczebnością populacji generalnej N,

2) próba powinna być próbą losową prostą (losowanie niezależne),

3) poziom istotności α powinien być równy 0,05 lub 0,01.

5.3.1. Test zgodności chi-kwadrat (χ^2)

Nazwa testu pochodzi od używanej w nim statystyki (mającej asymptotyczny rozkład χ^2) oraz kojarzona jest z nazwiskiem jego twórcy – K. Pearsona. Stąd też często używa się nazwy: **test zgodności χ^2 Pearsona**. Statystyka χ^2 jest tu definiowana następująco:

$$\chi^2 = \sum_{i=1}^{r} \frac{(n_i - np_i)^2}{np_i}, \tag{5.23}$$

gdzie n_i oznacza liczebności empiryczne (zaobserwowane w próbie), a np_i – liczebności teoretyczne (oczekiwane). Jak wynika ze wzoru (5.23), gdyby nie było odchyleń wartości empirycznych od teoretycznych, wartość statystyki (5.23) byłaby równa zeru. Przypadek taki jest raczej niespotykany. Dlatego też bada się, jak duże mogą być dopuszczalne różnice, aby nie przeczyły zgodności obserwacji empirycznych z hipotetycznymi (oczekiwanymi).

Stawiana w teście istotności hipoteza H_0 dotyczy zgodności rozkładu empirycznego z rozkładem hipotetycznym o znanej postaci dystrybuanty $F(x)$. Hipoteza alternatywna H_1 zakłada, że weryfikowany rozkład jest innego typu. Formalny zapis tych hipotez jest następujący (są to hipotezy nieparametryczne):

---

[2] Pod pojęciem mocy testu dla określonej hipotezy rozumie się prawdopodobieństwo odrzucenia hipotezy alternatywnej, gdy jest ona fałszywa. Moc testu jest więc związana z prawdopodobieństwem popełnienia błędu II rodzaju (β). Moc testu, oznaczana symbolem $1-\beta$, jest zatem tym większa, im β jest mniejsze.

[3] A. Zeliaś, *Metody statystyczne*, PWE, Warszawa 2000, s. 279.

$$H_0: F(x) = F_0(x),$$
$$H_1: F(x) \neq F_0(x). \quad (5.24)$$

gdzie $F_0(x)$ jest określoną, hipotetyczną postacią dystrybuanty (np. rozkładu normalnego, Bernoulliego czy Poissona). Do weryfikacji H_0 wykorzystuje się statystykę (5.23). Ma ona – przy założeniu prawdziwości H_0 – asymptotyczny rozkład chi-kwadrat o $r-1$ lub $r-k-1$ stopniach swobody. W tym drugim przypadku liczba k odnosi się do liczby parametrów rozkładu hipotetycznego oszacowanych na podstawie wyników próby losowej.

Obszar krytyczny w tym teście buduje się **zawsze prawostronnie**, tzn. tak, aby spełniona była relacja: $P(\chi^2 \geq \chi^2_\alpha) = \alpha$. W przypadku gdy zachodzi: $\chi^2 \geq \chi^2_\alpha$, to H_0 należy odrzucić (różnica między rozkładem empirycznym a hipotetycznym jest statystycznie istotna). W przeciwnym przypadku nie ma podstaw do odrzucenia H_0. Stwierdzając zatem brak podstaw do odrzucenia hipotezy H_0, nie mamy powodów, by kwestionować zgodność rozkładu badanej cechy w populacji generalnej z przyjętym rozkładem hipotetycznym.

Test zgodności χ^2 Pearsona można stosować zarówno dla zmiennych losowych ciągłych, jak i skokowych. W tym pierwszym przypadku należy dokonać klasyfikacji wyników próby do r rozłącznych klas. W literaturze przedmiotu obserwuje się wyraźny brak zgodności poglądów zarówno co do ogólnej liczebności próby losowej, jak również minimalnej liczebności w poszczególnych klasach. Polska norma PN-83/N-01052.07 zaleca, aby liczebność próby wynosiła co najmniej 100, liczba klas była równa: 10–25, a minimalna liczebność empiryczna w poszczególnych klasach – nie mniejsza od 5[4]. Jeżeli w rozkładzie empirycznym w pewnym wariancie cechy lub klasie występuje liczebność mniejsza od 5, to należy dokonać połączenia sąsiednich klas.

Technikę stosowania testu zgodności χ^2 zilustrujemy przykładami, w których będziemy weryfikować hipotezy o zgodności rozkładu empirycznego z rozkładami hipotetycznymi: normalnym, dwumianowym i Poissona.

Załóżmy, że losowa próba $n = 200$ niezależnych obserwacji wagi noworodków (w kg) dała następujące wyniki:

Waga	1,0–1,4	1,4–1,8	1,8–2,2	2,2–2,6	2,6–3,0
Liczba noworodków	15	45	70	50	20

Na poziomie istotności 0,05 mamy zweryfikować hipotezę, że rozkład wagi noworodków jest rozkładem normalnym.

Sprawdzaną hipotezą zerową jest tu: $H_0: F(x) = F_0(x)$, gdzie $F_0(x)$ jest dystrybuantą rozkładu normalnego, a hipotezą alternatywną: $H_1: F(x) \neq F_0(x)$.

[4] Por. M. Maliński, *Weryfikacja hipotez statystycznych wspomagana komputerowo*, Wydawnictwo Politechniki Śląskiej, Gliwice 2004, s. 214.

Parametry rozkładu normalnego (tzn. wartość oczekiwana i odchylenie standardowe) szacujemy na podstawie wyników próby. Otrzymujemy: $\bar{x} = 2$ kg oraz $s = 0{,}43$ kg. Pozostałe obliczenia niezbędne do wyznaczenia statystyki (5.23) są przeprowadzone w tab. 5.2.

Tab. 5.2. Obliczenia pomocnicze

x_i	n_i	z_i	$F(z_i)$	p_i	np_i	$\frac{(n_i - np_i)^2}{np_i}$
1,4	15	–1,39	0,082	0,082	16,4	0,12
1,8	45	–0,46	0,323	0,241	48,2	0,21
2,2	70	+0,46	0,677	0,354	70,8	0,01
2,6	50	+1,49	0,918	0,241	48,2	0,07
3,0	20	–	–	0,082	16,4	0,79
Razem	200	X	X	1,000	200,0	$\chi^2 = 1{,}2$

Źródło: obliczenia własne.

Wartości x_i występujące w pierwszej kolumnie tab. 5.2 dotyczą górnych granic przedziałów klasowych. Symbolem $F(z_i)$ oznaczono wartości dystrybuanty rozkładu $N(0,1)$ w punkcie $z_i = \frac{x_i - \bar{x}}{s}$. Tak więc, zachodzi relacja:

$$F(z_i) = F\left(\frac{x_i - \bar{x}}{s}\right). \quad (5.25)$$

Z uwagi na to, że suma prawdopodobieństw p_i powinna być równa 1, prawdopodobieństwo dla ostatniego przedziału wyznaczono jako $1 - F(z_i)$.

Wartości dystrybuanty $F(z_i)$ odczytano z tablic rozkładu normalnego $N(0,1)$, a prawdopodobieństwa p_i obliczono następująco: $p_i = F(z_i) - F(z_{i-1})$. Wartość krytyczna testu χ^2_α została odczytana z tablic rozkładu chi-kwadrat dla poziomu istotności α oraz $5 - 2 - 1 = 2$ stopni swobody. Wartość ta jest równa $\chi^2_\alpha = 5{,}991$. Ze względu na to, że: $\chi^2 = 1{,}2 < \chi^2_\alpha = 5{,}991$, brakuje podstaw do odrzucenia hipotezy, że rozkład wagi noworodków jest rozkładem normalnym.

W kolejnym przykładzie zweryfikujemy hipotezę o zgodności rozkładu empirycznego z rozkładem Poissona. Rozkład Poissona dotyczy zmiennej losowej skokowej (dyskretnej).

Na 200 poletkach, z których każde miało powierzchnię 30 m^2, ustalono liczebności roślin ostu. Otrzymano następujące wyniki:

Liczba roślin ostu	0	1	2	3	4	5	6 i więcej
Liczba poletek	22	58	65	35	10	7	3

Na poziomie istotności 0,05 należy zbadać, czy otrzymany rozkład empiryczny jest zgodny z rozkładem Poissona.

Jak wiadomo, rozkład Poissona zależy od jednego parametru λ. Ponieważ do wyznaczania prawdopodobieństw p_i konieczna jest znajomość tego parametru, należy go oszacować. Oceną parametru λ jest średnia arytmetyczna z próby. Wynosi ona $\bar{x} = 1{,}9$. W celu wyznaczenia prawdopodobieństw p_i korzystamy z funkcji określonej wzorem:

$$P(X = k) = \frac{1{,}9^k}{k!} e^{-1{,}9} \text{ dla } k = 0,1,2,3,4,5. \tag{5.26}$$

Ostatni wariant cechy w szeregu empirycznym ma liczebność mniejszą od 5. Połączono go zatem z wariantem przedostatnim. Łączna liczebność ostatniego wariantu jest więc równa 10.

Prawdopodobieństwa p_i można również odczytać z tablic rozkładu Poissona, przy czym ostatnie prawdopodobieństwo jest dopełnieniem do jedności. Kolejne etapy obliczeń związanych z wyznaczaniem statystyki χ^2 przedstawiono w tab. 5.3.

Tab. 5.3. Obliczenia pomocnicze

x_i	n_i	p_i	np_i	$n_i - np_i$	$\frac{(n_i - np_i)^2}{np_i}$
0	22	0,150	30,0	–8,0	2,13
1	58	0,284	56,8	1,2	0,03
2	65	0,270	54,0	11,0	2,24
3	35	0,171	34,2	0,8	0,02
4	10	0,081	16,2	–6,2	2,37
5 i więcej	10	0,044	8,8	1,2	0,16
Razem	200	1,000	200,0	X	$\chi^2 = 6{,}95$

Źródło: obliczenia własne.

Na podstawie wyników próby szacowano jeden parametr ($k = 1$), a r – wobec połączenia klas – jest równe 6. Liczba stopni swobody wynosi zatem: $r - k - 1 = 6 - 1 - 1 = 4$. Z tablic rozkładu χ^2 dla $\alpha = 0{,}05$ i 4 stopni swobody odczytujemy wartość krytyczną $\chi^2_\alpha = 9{,}488$. Ostatecznie mamy więc: $\chi^2 = 6{,}95 < \chi^2_\alpha = 9{,}488$, wobec czego nie ma podstaw do odrzucenia hipotezy zerowej. Z prawdopodobieństwem popełnienia błędu I rodzaju na poziomie 0,05 można zatem twierdzić, że rozkład liczby roślin ostu na poletkach jest rozkładem Poissona.

Załóżmy, że w pewnej firmie zbadano liczbę lekkich wypadków przy pracy. Otrzymano następujące wyniki zaobserwowane w ciągu ostatniego roku:

Kolejne godziny pracy	1	2	3	4	5	6	7	8
Liczba wypadków	18	10	12	14	15	20	20	11

Na poziomie istotności 0,05 mamy zweryfikować hipotezę, że rozkład liczby wypadków jest rozkładem dwumianowym z parametrem $p = \frac{1}{8}$.

Sprawdzaną hipotezą zerową jest tu: $H_0: F(x) = F_0(x)$, wobec hipotezy alternatywnej $H_1: F(x) = F_0(x)$, gdzie $F_0(x)$ jest dystrybuantą rozkładu dwumianowego. Niezbędne obliczenia związane z weryfikacja hipotezy zerowej zawiera tab. 5.4.

Tab. 5.4. Obliczenia pomocnicze

x_i	n_i	np_i	$(n_i - np_i)^2$	$\frac{(n_i - np_i)^2}{np_i}$
1	18	15	9	0,60
2	10	15	25	1,67
3	12	15	9	0,60
4	14	15	1	0,07
5	15	15	0	0,00
6	20	15	25	1,67
7	20	15	25	1,67
8	11	15	16	1,07
Razem	120	120	X	$\chi^2 = 7{,}35$

Źródło: obliczenia własne.

Z uwagi na to, że wartość parametru p w rozkładzie dwumianowym została z góry określona, liczba stopni swobody jest równa: $8 - 1 = 7$. Odczytana z tablic rozkładu χ^2 dla poziomu istotności $\alpha = 0{,}05$ i 7 stopni swobody wartość krytyczna wynosi: $\chi^2_\alpha = 14{,}067$. Ponieważ $\chi^2 = 7{,}35 < \chi^2_\alpha = 14{,}067$, więc nie ma podstaw do odrzucenia H_0, że rozkład liczby wypadków w kolejnych godzinach pracy jest rozkładem dwumianowym.

5.3.2. Test losowości próby

Jak wiadomo, wnioskowaniem statystycznym nazywamy proces uogólniania wyników otrzymanych z próby na całą populację generalną, z której ona pochodzi. Aby uogólnienie to dawało satysfakcjonujące rezultaty, próba musi mieć charakter **losowy**. Do weryfikacji hipotezy o losowości próby najczęściej wykorzystuje się **test liczby serii**.

Seria jest sekwencją elementów jednego rodzaju, przed i po których występują inne elementy lub nie ma żadnego. Wyjaśnienie tej definicji – przy użyciu symboli a i b – przedstawia poniższy schemat:

aaa	bb	a	bbbb	aa	bbbbb	aaaaa
seria	seria	seria	seria	seria	seria	seria

W sekwencji symboli przedstawionej na powyższym schemacie jest $k = 7$ serii.

Przypuśćmy, że z populacji generalnej o określonym rozkładzie pobrano próbę $(x_1, x_2, \ldots, x_n)$ o liczebności n elementów. Zakładając, że pojawianie się kolejnych elementów w próbie ma charakter losowy, ogólna liczba serii w ciągu n-elementowym jest traktowana jako zmienna losowa K. Zmienna ta ma znany i stablicowany rozkład (tablice rozkładu liczby serii). Hipotezy: zerowa i alternatywna, są tutaj formułowane następująco:

H_0: próba jest losowa,

H_1: próba nie jest losowa.

Wartość statystyki z próby wyznaczamy w następujący sposób:

1) kolejne obserwacje w próbie tworzą n-elementowy ciąg podstawowy,

2) obserwacje porządkujemy rosnąco i wyznaczamy z nich medianę (Me),

3) w ciągu podstawowym w miejsce wartości $x_i (i = 1,2,\ldots, n)$ wpisujemy symbole:

a, jeśli $x_i < Me$,

b, jeśli $x_i > Me$,

$x_i = Me$ pomijamy,

4) w otrzymanym ciągu symboli zliczamy utworzoną liczbę serii k, która jest wartością statystyki K otrzymaną z próby.

Obszar krytyczny testu liczby serii jest **dwustronny**, gdyż zarówno wystąpienie zbyt małej, jak i zbyt dużej liczby serii zaprzecza losowemu charakterowi próby.

Z tablic rozkładu liczby serii odczytujemy dwie wartości krytyczne: k_1 oraz k_2, w następujący sposób:

$$k_1 \text{ dla } \frac{\alpha}{2} \text{ oraz } n_1 \text{ i } n_2,$$

$$k_2 \text{ dla } 1 - \frac{\alpha}{2} \text{ oraz } n_1 \text{ i } n_2,$$

gdzie n_1 i n_2 ($n_1 + n_2 = n$) oznaczają odpowiednio liczbę elementów oznaczonych symbolami a i b, natomiast α jest przyjętym z góry poziomem istotności. W celu podjęcia decyzji weryfikacyjnej porównujemy empiryczną liczbę serii z wartościami krytycznymi k_1 i k_2. Jeżeli $k_1 < k < k_2$, to nie ma podstaw do odrzucenia H_0 o losowości próby. Gdy natomiast wystąpi alternatywa: $k_1 \leq k$ lub $k \geq k_2$, odrzucamy H_0 na korzyść H_1. Oznacza to, że próba nie ma charakteru losowego.

Zwrócić należy uwagę na to, że zaprezentowany powyżej sposób postępowania jest właściwy dla $n_1, n_2 \leq 20$. Dla takich liczebności można odczytać z tablicy rozkładu liczby serii wartości krytyczne k_1 i k_2. Jeżeli występuje $n_1, n_2 > 20$, zmienna losowa K zmierza – asymptotycznie – do rozkładu normalnego.

Załóżmy, że w celu oszacowania średniej liczby jednostek internetu zainstalowanych w blokach mieszkalnych przy pewnej ulicy, wylosowano 17 bloków i otrzymano następujące wyniki (liczba zainstalowanych jednostek internetu): 60, 57, 55, 50, 49, 36, 34, 32, 31, 30, 29, 27, 25, 23, 21, 20, 19. Mamy – na poziomie istotności $\alpha = 0{,}05$ – zweryfikować hipotezę, że wybór bloków do próby był losowy.

Hipoteza zerowa ma tu postać H_0: próba jest losowa, natomiast hipoteza alternatywna H_1: próba nie jest losowa. Aby zweryfikować H_0 testem liczby serii, należy najpierw z wyników próby losowej obliczyć medianę. Ponieważ $n = 17$, mediana jest dziewiątym z kolei wynikiem, czyli $Me = 31$. Używając symboli a i b zamiast wartości x_i w ciągu podstawowym otrzymujemy następujący ciąg:

b b b b b b b b a a a a a a a a a

w którym liczba serii $k = 2$. Liczba elementów a wynosi $n_1 = 8$, liczba elementów b jest równa $n_2 = 8$. Z tablic rozkładu liczby serii odczytujemy dla $\alpha = 0{,}05$, tzn. dla $\frac{\alpha}{2} = 0{,}025$ i dla $1 - \frac{\alpha}{2} = 0{,}975$, wartości krytyczne $k_1 = 3$ oraz $k_2 = 12$. Z uwagi na to, że $k = 2 < k_1 = 3$, hipotezę zerową o losowości próby należy odrzucić (otrzymaliśmy zbyt małą liczbę serii).

Test liczby serii, służący do badania losowości próby, można również stosować dla cech jakościowych. Załóżmy, że obserwowano płeć studentów zgłaszających się kolejno na egzamin. Otrzymano następujący ciąg (M – mężczyzna, K – kobieta): K, K, K, M, K, K, K, M, M, K, K, K, K, K, K, K, M, K, K, M. Mamy sprawdzić – na poziomie istotności 0,05 – czy kolejność wchodzenia na egzamin miała charakter losowy.

W otrzymanym ciągu liczba serii $k = 8$, przy czym $n_1 = 15$ oraz $n_2 = 5$. Wartości krytyczne odczytane z tablicy rozkładu liczby serii wynoszą: $k_1 = 2$ i $k_2 = 9$. Tak więc zachodzi następująca relacja: $k_1 = 2 < k = 8 < k_2 = 9$. Oznacza to, że brak jest podstaw do odrzucenia hipotezy zerowej o losowości próby.

ZADANIA

5.1. Zgodnie z normą waga konserw mięsnych powinna być równa 250 g z wariancją 25 $(g)^2$. Zakupiono 100 konserw, których średnia waga wyniosła 245 g. Zakładając, że rozkład wagi konserw jest normalny, zweryfikuj hipotezę, że waga konserw nie różni się od wagi przewidzianej normą. Przyjąć poziom istotności 0,01.

5.2. Uczniowie pewnej lokalnej szkoły średniej, w liczbie 120, dojeżdżali do niej z różnych odległości. Uzyskali oni ze sprawdzianu wiadomości o świecie współczesnym średnią liczbę punktów równą 30 z odchyleniem standardowym 5. Po pewnym czasie uruchomiono stałą komunikację między szkołą a miejscowościami, w których mieszkali uczniowie. Ponownie przeprowadzony sprawdzian wiadomości tych uczniów dał średnią liczbę punktów równą 35, z typowym obszarem zmienności $29 < x_{typ} < 41$. Na poziomie istotności 0,05 zweryfikować hipotezę, że zor-

ganizowanie uczniom dojazdu do szkoły polepszyło ich wyniki w nauce. Wykorzystać lewostronny obszar krytyczny.

5.3. Czas użytkowania baterii R-14 ma rozkład $N(m,100)$. Wylosowano 25 baterii tego typu i stwierdzono, że średni czas ich użytkowania wynosił 3000 godzin. Producent baterii twierdzi, że średni czas ich użytkowania jest równy 3100 godzin. Przyjmując poziom istotności 0,02, zweryfikować hipotezę, że średni czas użytkowania baterii różni się od czasu podawanego przez producenta.

5.4. W wylosowanej niezależnie próbie 81 zakładów zbadano koszty własne produkcji pewnego wyrobu. Z wyników próby otrzymano $\bar{x} = 540$ zł oraz $s = 150$ zł. Przy poziomie istotności 0,05 zweryfikować hipotezę, że średnie koszty własne w populacji wszystkich zakładów są równe 500 zł.

5.5. W instytucji ubezpieczeniowej A wylosowano 100 szkód komunikacyjnych, a w instytucji B – 150 szkód. Średni okres likwidacji szkód w instytucji A wyniósl 40 dni z odchyleniem standardowym 10 dni. W instytucji B średni okres likwidacji szkód był równy 55 dni, przy wariancji wynoszącej 144 $(\text{dni})^2$. Na poziomie istotności 0,01 zweryfikować hipotezę, że:

a) średni czas likwidacji szkody w instytucji A był krótszy niż w instytucji B,

b) zróżnicowanie okresu likwidacji szkody w instytucji B było istotnie większe niż w instytucji A.

5.6. Prezydent pewnego miasta stwierdził, że średnia powierzchnia sklepów w jego mieście wynosi 100 m^2, przy współczynniku zmienności 20%. W celu weryfikacji tego stwierdzenia wylosowano 70 sklepów i otrzymano następujące wyniki (w m^2):

Powierzchnia	50–70	70–90	90–110	110–130
Liczba sklepów	10	27	21	12

Na poziomie istotności 0,04 zweryfikować hipotezę, że prezydent zawyżył zarówno średnią powierzchnię sklepów, jak również dyspersję powierzchni.

5.7. Z dwóch kompanii wojskowych wylosowano po 12 szeregowców. Wariancja wzrostu wylosowanych szeregowców w pierwszej kompanii wynosiła $\hat{s}_1^2 = 5{,}76\ (\text{cm})^2$, a w drugiej $\hat{s}_2^2 = 5{,}41\ (\text{cm})^2$. Zakładając, że rozkład wzrostu szeregowców w obydwu kompaniach jest normalny, zweryfikuj hipotezę, że wariancje wzrostu w obydwu kompaniach są jednakowe. Przyjąć poziom istotności 0,05.

5.8. Twierdzi się, że polowa wypadków drogowych jest efektem nadmiernej prędkości jazdy. Okazało się, że na 330 wypadków, 67 było wynikiem nadmiernej prędkości. Czy – na poziomie istotności 0,05 – można potwierdzić obiegową opinię?

5.9. Przypuszcza się, że dyspersja miesięcznego spożycia mięsa i jego przetworów na 1 osobę w Polsce jest nie większa niż 0,5 kg. W celu sprawdzenia tego przypuszczenia wylosowano próbę o liczebności 16 osób. Z próby tej otrzymano następujące statystyki: $\bar{x} = 5{,}08$ kg/osobę oraz $s = 0{,}67$ kg/osobę. Zakładając, że rozkład spożycia mięsa i jego przetworów jest normalny, zweryfikować hipotezę o wielkości zróżnicowania spożycia w populacji generalnej. Przyjąć poziom istotności 0,05.

5.10. Panuje przekonanie, że kobiety po 50 roku życia chętniej oszczędzają niż mężczyźni. W celu zweryfikowania tego sądu wylosowano 31 kobiet i 31 mężczyzn powyżej 50 roku życia. W próbie tej średnie oszczędności wśród kobiet wyniosły 1800 zł z odchyleniem standardowym 450 zł, wśród mężczyzn $\bar{x} = 1500$ zł i $s = 350$ zł. Na poziomie istotności 0,05 zweryfikować hipotezę:

a) o równości wariancji oszczędności w populacji kobiet i mężczyzn,

b) o wyższym średnim poziomie oszczędności w populacji kobiet w porównaniu ze zbiorowością mężczyzn.

5.11. W jednej z politechnik wylosowano niezależnie próbę 150 studentów, z których jedynie 45 zdało wszystkie egzaminy w pierwszym terminie. Na poziomie istotności 0,05 zweryfikować hipotezę, że mniej niż trzecia część studentów zdaje egzaminy w pierwszym terminie.

5.12. W ubiegłym roku do urzędu wpłynęło 1419 skarg od indywidualnych osób ubezpieczonych. W 1124 przypadkach urząd podjął interwencję. W bieżącym roku ogółem wpłynęło 2430 skarg, wobec których urząd w 1823 przypadkach podjął interwencję. Na poziomie istotności 0,04 zweryfikować hipotezę, że frakcja niepodjętych interwencji w bieżącym roku w porównaniu do poprzedniego istotnie zwiększyla się.

5.13. W dwóch wybranych losowo przedsiębiorstwach przemysłowych zbadano efektywny czas pracy robotników bezpośrednio produkcyjnych. W obu przedsiębiorstwach rozkład efektywnie przepracowanego czasu pracy jest normalny z identycznym odchyleniem standardowym równym 0,8 godziny. Średni efektywny czas pracy w $n_1 = 22$-elementowej próbie robotników w pierwszym przedsiębiorstwie wynosił 6,3 godziny, a w 15-elementowej próbie robotników w drugim przedsiębiorstwie $\bar{x}_2 = 5{,}8$ godziny. Przyjmując poziom istotności 0,10 sprawdzić hipotezę, że średni efektywny czas pracy robotników w obu populacjach robotników jest jednakowy.

5.14. Wysunięto przypuszczenie, że konsumenci, którzy dopiero w punkcie sprzedaży dokonują wyboru marki produktu, stanowią 15% wszystkich nabywców produktu. W celu sprawdzenia tej hipotezy wylosowano próbę $n = 729$ konsumentów i stwierdzono, że 91 spośród nich dokonało wyboru marki dopiero w miejscu zakupu produktu. Na poziomie istotności 0,1 zweryfikować wysunięte przypuszczenie.

5.15. W dwóch losowo wybranych grupach chorych zastosowano różne specyfiki obniżające temperaturę ciała. W pierwszej grupie, 10 chorych, po trzech dniach kuracji zanotowano następujące temperatury (w °C): 37,0; 36,8; 37,9; 37,2; 38,0; 37,0; 37,1; 35,9; 36,5; 37,0. W grupie drugiej, 15 chorych, po trzech dniach kuracji uzyskano średnią temperaturę ciala 37,5°C, z przeciętnym zróżnicowaniem mierzonym odchyleniem standardowym 0,3°C. Czy można twierdzić, że efektywność działania obu specyfików jest, średnio rzecz biorąc, taka sama? Przyjąć poziom istotności 0,02.

5.16. Z populacji rodzin miejskich w pewnym województwie wylosowano 280 rodzin. Okazało się, że 57 spośród nich to rodziny pięcioosobowe lub większe. Wśród wylosowanych 200 rodzin wiejskich w tym samym województwie, 63 rodziny składały się z pięciu lub więcej osób. Przyjmując poziom istotności 0,01, zweryfikować hipotezę, że frakcje dużych rodzin w miastach i na wsi w tym województwie są jednakowe.

5.17. Zbadano kształtowanie się ceny produktu X w różnych sklepach. Wylosowano 12 sklepów i stwierdzono, że średnia cena produktu X wynosi 28 zł, przy typowym obszarze zmienności 26,8 zł $< x_{typ} <$ 29,2 zł. Czy – na poziomie istotności 0,05 – można twierdzić, że średnia cena produktu X jest wyższa od 27,5 zł?

5.18. Na 150 wypadków samochodowych w pewnym województwie 118 spowodowanych było nadużyciem alkoholu przez kierowców. W sąsiednim województwie liczba wypadków spowodowanych nadużyciem alkoholu wyniosla 130 na

185 zgłoszonych. Na poziomie istotności 0,05 zweryfikować hipotezę o jednakowym odsetku wypadków samochodowych spowodowanych nadużyciem alkoholu przez kierujących pojazdami w obydwu województwach.

5.19. Dokonano 8 pomiarów czasu opóźnienia pociągu na przystanku X. Wyniki pomiarów były następujące (w minutach): 4; 6; 3; 1; 3; 3; 9; 1. Zakładamy, że czas opóźnienia ma rozkład normalny. Czy na poziomie istotności 0,1 można przyjąć, że średni czas opóźnienia wynosi 4 minuty?

5.20. Rozkład liczby uczniów w 305 gimnazjach w wybranych miejscowościach województwa lubelskiego przedstawia się następująco:

Liczba uczniów	0–40	40–80	80–120	120–160	160–200
Liczba gimnazjów	10	60	70	65	100

Zakładając, że liczba uczniów w gimnazjach ma rozkład normalny, zweryfikować – na poziomie istotności 0,05 – hipotezę, że średnia liczba uczniów w tego typu szkołach wynosi 125.

5.21. W mieście L wylosowano niezależnie do próby 400 gospodarstw domowych i ustalono, że 280 spośród nich posiada komputer. Czy na poziomie istotności 0,01 można twierdzić, że udział gospodarstw posiadających komputer przekracza 6%?

5.22. Producent wyrobu X twierdzi, że 60% gospodarstw kupuje ten wyrób. Konkurent przeprowadził badanie ankietowe. Okazało się, że na 120 gospodarstw domowych 55 potwierdziło stosowanie wspomnianego produktu. Czy – na poziomie istotności 0,05 – można sądzić, że reklama nie odpowiada stanowi rzeczywistemu?

5.23. Panuje pogląd, że 13% inwestorów giełdowych kupuje akcje spółek. W losowej próbie 250 inwestorów 35 zakupiło akcje spółek. Czy – na poziomie istotności 0,1 – można uznać wyrażany pogląd za słuszny?

5.24. W pewnym wieżowcu w ciągu 150 dni obserwowano liczbę awarii sieci energetycznej. Otrzymano następujące wyniki:

Liczba awarii	0	1	2	3	4	5
Liczba dni	18	40	35	25	15	7

Czy – przy przyjęciu poziomu istotności 0,05 – można twierdzić, że
a) przeciętna liczba awarii jest równa 2,
b) odchylenie standardowe liczby awarii wynosi 1?

5.25. Przypuszcza się, że klienci dokonujący zakupów kierują się kolorem opakowania. W celu sprawdzenia tej sugestii, na oddalonych od siebie półkach sklepowych umieszczono takie same soki pomarańczowe w niebieskich i pomarańczowych kartonach. W alejce z niebieskimi opakowaniami na 200 potencjalnych klientów 85 dokonało zakupu, w alejce z pomarańczowymi opakowaniami soku spośród 300 potencjalnych nabywców 162 osoby dokonały zakupu. Czy – na poziomie istotności 0,05 – można twierdzić, że frakcja zakupów soków w dwóch rodzajach kolorów opakowania jest jednakowa?

5.26. Przez 4 tygodnie stawiano prognozy meteorologiczne dwiema różnymi metodami. Prognozy stawiane pierwszą metodą okazały się trafne dla 21 dni, a metodą drugą – dla 17 dni. Czy na poziomie istotności 0,05 można twierdzić, że pierwsza metoda jest lepsza od drugiej?

5.27. Firma farmaceutyczna testuje nowy lek nasercowy. W grupie leczonej starym specyfikiem, liczącej 500 osób, istotna poprawa stanu zdrowia nastąpiła

wśród 115 pacjentów. Wśród pacjentów poddanych kuracji nowym lekiem, stan zdrowia polepszył się u 27% spośród 600, u których zastosowano terapię. Czy – na poziomie istotności 0,04 – można twierdzić, że nowy lek jest skuteczniejszy od starego?

5.28. Biuro nieruchomości twierdzi, że ceny wynajmu mieszkań w Lublinie i w Rzeszowie różnią się istotnie. W celu sprawdzenia tego sądu pobrano dwie próby ofert wynajmu kawalerek, oferowanych przez pośredników obrotu nieruchomościami w obu miastach. Otrzymano następujące wyniki:

Lublin: $n_1 = 12$; $\bar{x}_1 = 800$ zł, $s_1 = 50$ zł,

Rzeszów: $n_2 = 17$; $\bar{x}_2 = 720$ zł, $s_2 = 80$ zł.

Czy – na poziomie istotności 0,1 – można wysunięte przypuszczenie uznać za słuszne?

5.29. Dzienne utargi firmy handlowej mają rozkład normalny $N(m,\sigma)$. Z próby losowej o liczebności 121 dni wynika, że średnia arytmetyczna utargów wyniosła 950 zł z wariancją 25 $(\text{zł})^2$. Na poziomie istotności 0,01 sprawdzić hipotezę $H_0: m = 1000$ zł. Wykorzystać lewostronny obszar krytyczny.

5.30. Na poziomie istotności 0,1 zweryfikować hipotezę, że dyspersja wagi studentów jest mniejsza niż 5 kg, jeśli z próby losowej 85 studentów otrzymano odchylenie standardowe wagi równe 6 kg.

5.31. Strukturę 1000 losowo wybranych mieszkań na osiedlu Czuby w Lublinie według liczby izb przedstawia poniższy rozkład:

Liczba izb w mieszkaniu	2	3	4	5	6 i więcej
Liczba mieszkań	96	288	404	168	44

Czy słuszne jest stwierdzenie, że udział mieszkań 4-izbowych w ogólnej liczbie mieszkań nie przekracza 40%? Przyjąć poziom istotności 0,05.

5.32. W dwóch spółkach z o.o. wylosowano po 51 pracowników w celu zbadania czasu dojazdu do pracy. Wariancja czasu dojazdu do pracy w pierwszej spółce (położonej poza miastem) wynosiła 110,1 $(\text{min})^2$, a w drugiej spółce (położonej w centrum miasta) – 48 $(\text{min})^2$. Przyjmując poziom istotności 0,05, zweryfikować hipotezę, że wariancje czasu dojazdu do pracy w obu spółkach są jednakowe.

5.33. Sondaż opinii publicznej na temat oczekiwanej frekwencji na wyborach do samorządu lokalnego w gminie X wykazał, że 155 osób spośród 300 wylosowanych zamierza uczestniczyć w głosowaniu. Czy można – na poziomie istotności 0,05 – przyjąć, że połowa ogółu uprawnionych mieszkańców tej gminy zamierza wziąć udział w wyborach do samorządu?

5.34. W dwóch przedsiębiorstwach (A i B) zbadano staż pracy pracowników. Otrzymano następujące wyniki:

Staż w latach	0–5	5–10	10–15	15–20	20–25
Liczba pracowników w zakładzie A	3	4	12	8	1
Liczba pracowników w zakładzie B	5	10	9	2	2

Czy słuszne jest twierdzenie o jednakowym przeciętnym stażu pracy pracowników w obu zakładach? Przyjąć poziom istotności 0,01.

5.35. Zgodnie z normą, kaloryczność obiadów w stołówce szpitalnej powinna być równa 2500 kaloriom. Dla wylosowanych w stołówce szpitalnej 60 obiadów otrzymano średnią arytmetyczną liczby kalorii równą 2430 z odchyleniem standar-

dowym 260 kalorii. Czy można twierdzić, na poziomie istotności 0,05, że średnia kaloryczność obiadów szpitalnych nie osiąga normy?

5.36. Wysunięto przypuszczenie, że średni wiek nabywców płyt kompaktowych z muzyką hip-hop nie przekracza 22 lat. W celu sprawdzenia tego sądu wylosowano 80 nabywców tego rodzaju płyt i otrzymano: $\bar{x} = 23{,}5$ lat oraz $s = 5{,}1$ roku. Czy przypuszczenie to można uznać – przy poziomie istotności 0,05 – za słuszne?

5.37. W celu sprawdzenia, czy wysokość półki, na której eksponuje się ten sam towar, ma wpływ na poziom sprzedaży, notowano wielkość dziennej sprzedaży tego towaru z półki o wysokości 120 cm i 150 cm. Otrzymano następujące wyniki:

Półka	120 cm	150 cm
Liczba dni obserwacji	12	12
Średnia dzienna sprzedaż w sztukach	44,5	35,0
Dzienne odchylenie standardowe sprzedaży w sztukach	5,3	8,8

Przyjmując poziom istotności 0,01, sprawdzić odpowiednią hipotezę.

5.38. Grupę 10 losowo wybranych kobiet poddano miesięcznej diecie odchudzającej. Otrzymano następujący rozkład ich wagi (przed i po diecie):

Przed dietą	86	56	67	82	92	59	90	80	65	92
Po diecie	86	34	66	83	87	60	88	85	65	87

Na poziomie istotności 0,05 sprawdzić, czy zastosowana dieta wpłynęła na spadek wagi.

5.39. W celu porównania regularności wyników uzyskiwanych przez dwóch skoczków w dal odnotowano 5 wyników skoczka A (w cm): 740; 720; 730; 720; 740 oraz 7 wyników skoczka B: 770; 760; 780; 680; 800; 720; 740. Na poziomie istotności 0,05 zweryfikować hipotezę o jednakowej regularności długości skoków obydwu skoczków.

5.40. Sklep spożywczy otrzymał dostawę maku w torebkach, z których każda powinna ważyć 500 g. Częste reklamacje klientów spowodowały konieczność skontrolowania wagi. Wylosowano zatem 17 torebek i uzyskano następujące ich wagi (w g): 500; 485; 480; 500; 480; 485; 465; 475; 480; 480; 491; 489; 503; 492; 475; 465; 500. Zakłada się, że rozkład ciężaru maku w torebkach jest normalny. Na poziomie istotności 0,05 zweryfikować zasadność reklamacji zgłaszanych przez klientów.

5.41. W celu oszacowania średniej punktacji na egzaminie ze statystyki z populacji studentów II r. studiów Akademii Ekonomicznej wylosowano 22 osoby i odnotowano ich wyniki w kolejności losowania: 80; 60; 90; 85; 50; 40; 45; 35; 65; 70; 75; 83; 46; 55; 58; 62; 39; 42; 48; 51; 89; 95. Przyjmując poziom istotności 0,05, zweryfikować hipotezę o losowości próby.

5.42. W pewnym mieście wylosowano niezależnie 500 rodzin i zbadano miesięczne zużycie energii elektrycznej w każdej z nich. Otrzymano następujący rozkład:

Zużycie energii w kWh	35–45	45–55	55–65	65–75	75–85
Liczba rodzin	70	100	140	110	80

Na poziomie istotności 0,01 zweryfikować hipotezę, że rozkład zużycia energii elektrycznej przez rodziny jest normalny.

5.43. Wylosowano 300 robotników i zbadano ich pod względem liczby wytwarzanych braków w partii 100 sztuk wyrobów. Otrzymano następujące wyniki:

Liczba braków	0	1	2	3	4	5
Liczba robotników	99	138	39	15	6	3

Na poziomie istotności 0,05 zweryfikować hipotezę, że rozkład wytwarzanych braków w populacji generalnej jest rozkładem Poissona.

5.44. Kandydatów na kierowców poddano badaniom sprawdzającym refleks. Każdy z kandydatów miał wykonać określone czynności na czterech typach aparatów. W wyniku przebadania 100 losowo wybranych osób otrzymano następujący rozkład liczby wykonanych zadań:

Liczba wykonanych zadań	0	1	2	3	4
Liczba kandydatów	5	12	23	40	20

Na poziomie istotności 0,01 sprawdzić hipotezę, że rozkład ten jest rozkładem Bernoulliego.

5.45. W celu zbadania wadliwości partii pewnych elementów wylosowano do kontroli technicznej 15 tych elementów i dokonano pomiaru ich średnicy. Otrzymano następujące wyniki (w mm): 16; 20; 25; 34; 22; 33; 47; 30; 28; 19; 22; 40; 36; 31; 38. Na poziomie istotności 0,05 zweryfikować hipotezę o losowym wyborze elementów do pomiaru.

5.46. Ankieta zawiera cztery pytania, na które przewidziano dwie odpowiedzi: „tak" albo „nie". Na podstawie odpowiedzi uzyskanych od 60 ankietowanych osób wybranych losowo należy sprawdzić – na poziomie istotności 0,05 – hipotezę, że liczba pozytywnych odpowiedzi udzielonych przez ankietowanych ma rozkład dwumianowy.

5.47. Zbiór jabłek w sadzie pewnego ogrodnika w dziesięciu kolejnych latach był następujący (w dt): 8,2; 6,1; 8,3; 6,5; 6,0; 8,4; 6,4; 8,0; 6,9; 7,9. Na poziomie istotności 0,05 zweryfikować hipotezę, że próba jest losowa.

5.48. Z analizy liczby kradzieży ujawnionych w pewnym supermarkecie otrzymano następujące dane:

Dzienna liczba kradzieży	0	1	2	3	4
Liczba dni	22	30	22	16	10

Czy – na poziomie istotności 0,05 – można sądzić, że rozkład liczby kradzieży jest rozkładem Poissona?

5.49. W kolejce po zakup akcji PKO BP zaobserwowano następującą kolejność kobiet (K) i mężczyzn (M): K, K, K, M, K, K, K, M, M, K, K, K, K, K, K, K, M, K, K, M. Na poziomie istotności 0,05 sprawdzić, czy próba jest losowa.

5.50. Automat paczkuje kawę w opakowania o wadze 250 g. Zważono 200 paczek kawy i otrzymano następujący rozkład:

Waga	248,4–248,8	248,8–249,2	249,2–249,6	249,6–250,0	250,0–250,4
Liczba paczek	15	45	70	50	20

Na poziomie istotności 0,1 zweryfikować hipotezę, że waga paczek kawy ma rozkład $N(250; 0, 4)$.

Rozdział VI

METODY ANALIZY KORELACJI I REGRESJI

Jednostki wchodzące w skład badanej zbiorowości są zazwyczaj charakteryzowane przez wiele cech statystycznych jednocześnie. Cechy te są – z reguły – ściśle ze sobą powiązane i wzajemnie warunkują się. Zbiór właściwości badanych jednostek określamy mianem **statystycznej cechy wielowymiarowej**. Liczba cech wyróżnionych w badaniu każdej jednostki decyduje o wymiarze cechy wielowymiarowej. W związku z tym można wyróżnić cechę **dwuwymiarową** (jednostki charakteryzowane są jednocześnie przez dwie właściwości), trójwymiarową, czterowymiarową itp. Podobnie jak w przypadku cech jednowymiarowych, również cechy wielowymiarowe można podzielić na **mierzalne** (ilościowe) i **niemierzalne** (jakościowe) oraz na **skokowe** i **ciągłe**.

O ile terminu „cecha statystyczna" używa się w przypadku rozkładu empirycznego, to rozpatrując populację generalną – posługujemy się określeniem **zmienna losowa**. Zmienna losowa wielowymiarowa jest funkcją wielowymiarową opisaną na zbiorze zdarzeń elementarnych. Wektor wartości tej funkcji ma tyle składowych, ile wynosi wymiar zmiennej. Jeśli więc zmienna losowa jest *k*-wymiarowa, to każdemu zdarzeniu elementarnemu przyporządkowanych jest *k* wartości tej funkcji.

Rozkład zmiennej losowej wielowymiarowej można – podobnie jak jednowymiarowej – opisać za pomocą funkcji prawdopodobieństwa (zmienna losowa skokowa), funkcji gęstości (zmienna losowa ciągła) oraz dystrybuanty (w przypadku obydwu rodzajów zmiennych). Ponadto dla syntetycznego scharakteryzowania rozkładów dotyczących zmiennych losowych wielowymiarowych korzysta się z określonych parametrów, takich jak: wartość oczekiwana, wariancja czy odchylenie standardowe. W dalszej części podręcznika skoncentrujemy uwagę na dwuwymiarowej zmiennej losowej skokowej.

6.1. Dwuwymiarowa zmienna losowa skokowa

Dwuwymiarową zmienną losową lub **wektorem losowym** (X,Y) nazywamy uporządkowaną parę (X,Y), w której każda ze zmiennych X i Y jest zmienną losową. Dwuwymiarowa zmienna losowa (X,Y) jest zmienną **skokową**, jeśli składowe X i Y mają skończony lub przeliczalny zbiór wartości.

Prawdopodobieństwo zdarzenia polegającego na tym, że zmienna losowa X przyjmie wartość x_i i jednocześnie zmienna losowa Y przyjmie wartość y_i zapisujemy następująco:

$$P(x_i, y_i) = P(X = x_i, Y = y_i) = p_{ij} (i = 1,2,\ldots,k; j = 1,2,\ldots,l). \quad (6.1)$$

Jeżeli zmienne losowe X i Y mogą przybierać skończony ciąg wartości, to:

$$\sum_{i=1}^{k}\sum_{j=1}^{l} p_{ij} = 1, \text{ przy czym } p_{ij} > 0. \quad (6.2)$$

Relację określoną wzorem (6.1) nazywamy **rozkładem** dwuwymiarowej zmiennej losowej skokowej (dyskretnej). Rozkład ten można przedstawić w formie tzw. **tablicy korelacyjnej** (tab. 6.1).

Tab. 6.1. Rozkład prawdopodobieństwa dwuwymiarowej zmiennej losowej dyskretnej

Y / X	y_1	y_2	y_3	...	y_l	$\sum$
x_1	p_{11}	p_{12}	p_{13}	...	p_{1l}	$p_{1\cdot}$
x_2	p_{21}	p_{22}	p_{23}	...	p_{2l}	$p_{2\cdot}$
x_3	p_{31}	p_{32}	p_{33}	...	p_{3l}	$p_{3\cdot}$
...	...	...	...	...	...	...
x_k	p_{k1}	p_{k2}	p_{k3}	...	p_{kl}	$p_{k\cdot}$
$\sum$	$p_{\cdot 1}$	$p_{\cdot 2}$	$p_{\cdot 3}$	...	$p_{\cdot l}$	1

Źródło: opracowanie własne.

Jak wynika z tab. 6.1, sumę określoną wzorem (6.2) otrzymujemy dodając wszystkie prawdopodobieństwa zawarte w tablicy. Suma ostatniej kolumny równa się sumie ostatniego wiersza i równa się jedności, tzn.

$$\sum_{i=1}^{k} p_{i\cdot} = \sum_{i=1}^{k} p_{\cdot j} = \sum_{i=1}^{k}\sum_{j=1}^{l} p_{ij} = 1. \quad (6.3)$$

W rozkładzie dwuwymiarowym zmiennej losowej (X, Y) można określić **rozkłady brzegowe**, które są jednowymiarowymi rozkładami zmiennych X i Y. Rozkład brzegowy zmiennej Y tworzą pierwszy i ostatni wiersz tab. 6.1, natomiast rozkład brzegowy zmiennej X – pierwsza i ostatnia kolumna tab. 6.1.

Dla zmiennej losowej dyskretnej (X, Y) rozkłady brzegowe są określone następująco:

$$P(X = x_i) = p_{i\cdot} \text{ dla } i = 1,2,\ldots,k \quad (6.4)$$

oraz

$$P(Y = y_j) = p_{\cdot j} \text{ dla } j = 1,2,\ldots,l. \quad (6.5)$$

Rozkłady brzegowe określają więc, jaką wartość przyjmie zmienna losowa X niezależnie od tego, jaką wartość przyjmie zmienna Y i odwrotnie. Stąd też rozkłady brzegowe nazywane są **rozkładami bezwarunkowymi**.

Można również określić jednowymiarowe rozkłady każdej ze zmiennych przy założeniu, że druga zmienna przyjmie określoną wartość. Mówimy wówczas o **rozkładach warunkowych**. Rozkłady te są określone następująco:

$$P(X = x_i | Y = y_j) = \frac{P(X = x_i, Y = y_j)}{P(Y = y_j)} = \frac{p_{ij}}{p_{\cdot j}} \quad (6.6)$$

oraz

$$P(Y = y_j | X = x_i) = \frac{P(X = x_i, Y = y_j)}{P(X = x_i)} = \frac{p_{ij}}{p_{i\cdot}}. \quad (6.7)$$

Rozkładów warunkowych zmiennej X jest więc tyle, ile wariantów przyjmuje zmienna Y i odwrotnie.

Zmienne losowe X i Y nazywamy **niezależnymi**, jeśli:

$$P(X = x_i, Y = y_i) = P(X = x_i) \cdot P(Y = y_j) \quad (6.8)$$

lub

$$p_{ij} = p_{i\cdot} p_{\cdot j}. \quad (6.9)$$

Z relacji (6.9) wynika, że zmienne losowe X i Y są niezależne, jeśli wartości łącznej funkcji prawdopodobieństwa pozostają równe iloczynowi odpowiednich wartości brzegowych funkcji prawdopodobieństwa. Relacja (6.9) określa konieczny i dostateczny warunek niezależności dwóch zmiennych losowych dyskretnych.

W przypadku gdy chcemy w sposób syntetyczny scharakteryzować zależność między zmiennymi, posługujemy się dwoma parametrami rozkładu dwuwymiarowego, tj. **kowariancją** lub **współczynnikiem korelacji**.

Kowariancja charakteryzuje sposób „kojarzenia się" wartości zmiennych tworzących wspólny (łączny) rozkład. W przypadku zmiennych losowych skokowych kowariancję, oznaczaną symbolem cov(X,Y), wyznacza się ze wzoru:

$$\begin{aligned} \operatorname{cov}(X,Y) &= \sum_{i=1}^{k}\sum_{j=1}^{l}\left[(x_i - E(X))(y_j - E(Y))\right]p_{ij} = \\ &= \sum_{i=1}^{k}\sum_{j=1}^{l} x_i y_j p_{ij} - E(X)E(Y) = E(XY) - E(X)E(Y), \end{aligned} \quad (6.10)$$

gdzie $E(X)$ oraz $E(Y)$ są odpowiednio wartościami oczekiwanymi zmiennych losowych X i Y, obliczonymi z rozkładów brzegowych.

Dodatnia wartość kowariancji oznacza, że wartości X i Y zmieniają się w tym samym kierunku (tzn. rosną lub maleją). W takim przypadku mówimy, że zmienne X i Y są **skorelowane dodatnio** (pozytywnie). Ujemna wartość kowariancji świadczy o tym, że kierunek zmian wartości X i Y jest odmienny (przy wzroście X maleją wartości Y i odwrotnie). Zmienne losowe X i Y są wówczas **skorelowane ujemnie (negatywnie)**.

W sytuacji, gdy przy wzroście X wartość Y, generalnie biorąc, nie zmienia się, zmienne X i Y są nieskorelowane, a $\text{cov}(X,Y) = 0$. Dla zmiennych niezależnych kowariancja przyjmuje wartość równą 0. Zmienne niezależne są więc również nieskorelowane. Twierdzenie odwrotne nie jest prawdziwe, gdyż zmienne, dla których $\text{cov}(X,Y) = 0$, mogą być zależne.

Kowariancja jest miarą nieunormowaną. Wady tej pozbawiony jest współczynnik korelacji, oznaczany symbolem ρ. Dla zmiennych X *i* Y współczynnik korelacji jest definiowany następująco:

$$\rho = \frac{\text{cov}(X,Y)}{D(X)D(Y)}, \tag{6.11}$$

gdzie $D(X)$ oraz $D(Y)$ są odchyleniami standardowymi obliczonymi z rozkładów brzegowych odpowiednich zmiennych.

Współczynnik korelacji jest wielkością unormowaną, gdyż przybiera wartości liczbowe z przedziału: $-1 \le \rho \le +1$. Dla zmiennych nieskorelowanych $\rho = 0$ (wtedy również $\text{cov}(X,Y) = 0$). Znak współczynnika korelacji wskazuje na kierunek zależności między zmiennymi, wartość bezwzględna – na siłę zależności między zmiennymi X i Y.

Załóżmy, że pewna firma budowlana prowadzi działalność w Lublinie i poza Lublinem. W wyniku badań stwierdzono, że rozkład prawdopodobieństwa liczby inwestycji prowadzonych jednocześnie w Lublinie i poza Lublinem przedstawia się następująco (X – liczba inwestycji w Lublinie, Y – liczba inwestycji poza Lublinem):

X \ Y	0	1	2
0	0,1	0,3	0,2
1	0,2	0,1	0,1

Źródło: dane umowne

Brzegowe funkcje prawdopodobieństwa zmiennych losowych X *i* Y przedstawiają się następująco:

– brzegowa funkcja prawdopodobieństwa zmiennej losowej X:

x_i	0	1
$p_{i\cdot}$	0,6	0,4

– brzegowa funkcja prawdopodobieństwa zmiennej losowej Y:

y_j	0	1	2
$p_{\cdot j}$	0,3	0,4	0,3

Aby sprawdzić, czy zmienne losowe X *i* Y są niezależne, należy wykorzystać warunek (6.9). Mamy więc:

$$p_{11} = P(X=0, Y=0) = 0,1 \neq p_1 p_1 = 0,6 \cdot 0,3 = 0,18$$
$$p_{12} = P(X=0, Y=1) = 0,3 \neq p_1 p_2 = 0,6 \cdot 0,4 = 0,24$$
$$p_{13} = P(X=0, Y=2) = 0,2 \neq p_1 p_3 = 0,6 \cdot 0,3 = 0,18$$
$$p_{21} = P(X=1, Y=0) = 0,2 \neq p_2 p_1 = 0,4 \cdot 0,3 = 0,12$$
$$p_{22} = P(X=1, Y=1) = 0,1 \neq p_2 p_2 = 0,4 \cdot 0,4 = 0,16$$
$$p_{23} = P(X=1, Y=2) = 0,1 \neq p_2 p_3 = 0,4 \cdot 0,3 = 0,12.$$

Zmienne losowe X *i* Y nie są niezależne, jeśli dla którejkolwiek pary (x_i, y_j) wartość funkcji prawdopodobieństwa spełnia warunek $p_{ij} \neq p_{i\cdot} p_{\cdot j}$. W naszym przypadku zmienne X *i* Y nie są niezależne. W kolejnym kroku należy zatem wyznaczyć współczynnik korelacji ρ między liczbą inwestycji w Lublinie i poza Lublinem. Korzystając z rozkładów brzegowych zmiennych X *i* Y, wyznaczamy wartości średnie i wariancje:

$$E(X) = 0 \cdot 0,6 + 1 \cdot 0,4 = 0,4$$
$$E(Y) = 0 \cdot 0,3 + 1 \cdot 0,4 + 2 \cdot 0,3 = 1$$
$$D^2(X) = E(X^2) - [E(X)]^2 = 0^2 \cdot 0,6 + 1^2 \cdot 0,4 - 0,4^2 = 0,24$$
$$D(X) = \sqrt{0,24} = 0,49$$
$$D^2(Y) = E(Y^2) - [E(Y)]^2 = 0^2 \cdot 0,3 + 1^2 \cdot 0,4 + 2^2 \cdot 0,3 - 1^2 = 0,6$$
$$D(Y) = \sqrt{0,60} = 0,775.$$

Wartość oczekiwaną iloczynu $E(XY)$ obliczamy następująco:

$$E(XY) = \sum_{i=1}^{k} \sum_{j=1}^{l} x_i y_j p_{ij}. \tag{6.12}$$

Podstawiając dane liczbowe do wzoru (6.12) otrzymujemy:

$$E(XY) = 0 \cdot 0 \cdot 0,1 + 0 \cdot 1 \cdot 0,3 + 0 \cdot 2 \cdot 0,2 + 1 \cdot 0 \cdot 0,2 + 1 \cdot 1 \cdot 0,1 + 1 \cdot 2 \cdot 0,1 = 0,3$$

Korzystając ze wzoru (6.10) otrzymujemy kowariancję:

$$\text{cov}(X,Y) = 0,3 - 0,4 \cdot 1 = -0,1.$$

Określony wzorem (6.11) współczynnik korelacji ρ wynosi:

$$\rho = \frac{-0,1}{0,49 \cdot 0,775} = \frac{-0,1}{0,37975} = -0,263.$$

Wartość wyznaczonego współczynnika korelacji wskazuje na niezbyt silne ujemne skorelowanie liczby inwestycji prowadzonych jednocześnie w Lublinie i poza Lublinem.

Współczynnik korelacji ρ jest unormowaną miarą zależności liniowej zmiennych losowych X *i* Y, gdyż:

$$|\rho(X,Y)| \le 1 = -1 \le \rho(X,Y) \le 1. \tag{6.13}$$

Jeśli zmienne losowe X *i* Y są niezależne, wówczas $\rho(X,Y) = 0$.

Gdyby zadaniem naszym było wyznaczenie średniej i wariancji łącznej liczby inwestycji prowadzonych jednocześnie w Lublinie i poza nim, to należałoby skorzystać z twierdzeń o wartości oczekiwanej i wariancji sumy dwóch zmiennych losowych X *i* Y, a mianowicie:

$$E(X+Y)=E(X)+E(Y) \tag{6.14}$$

oraz

$$D^2(X\pm Y)=D^2(X)+D^2(Y)\pm 2\operatorname{cov}(X,Y). \tag{6.15}$$

W naszym przypadku mamy:

$$E(X+Y)=0{,}4+1=1{,}4$$
$$D^2(X+Y)=0{,}24+0{,}6+2(-0{,}1)=0{,}64.$$

6.2. Dwuwymiarowa zmienna losowa ciągła

Jak wiadomo, zmienna losowa ciągła przyjmuje wartości nieprzeliczalne. **Funkcja gęstości** $f(x,y)$ dwuwymiarowej zmiennej losowej jest określona następująco:

$$f(x,y)\geq 0 \tag{6.16}$$

$$\int_{-\infty}^{+\infty}\int_{-\infty}^{+\infty} f(x,y)\,dxdy=1. \tag{6.17}$$

Dystrybuanta zmiennej losowej ciągłej (X,Y) przyjmuje postać:

$$F(x_0,y_0)=P(X\leq x_0, Y\leq y_0)=\int_{-\infty}^{x_0}\int_{-\infty}^{y_0} f(x,y)\,dxdy. \tag{6.18}$$

Dwuwymiarową zmienną losową ciągłą (X,Y) można również scharakteryzować określonymi parametrami. Najczęściej parametrami tymi są:

1) wartość oczekiwana:

$$E(X,Y)=[E(X),E(Y)], \tag{6.19}$$

2) wariancja:

$$D^2(X,Y)=[D^2(X),D^2(Y)], \tag{6.20}$$

3) kowariancja

$$\operatorname{cov}(X,Y)=\int_{-\infty}^{+\infty}\int_{-\infty}^{+\infty}(x-E(X))\cdot(y-E(Y))f(x,y)\,dxdy. \tag{6.21}$$

Jak wynika ze wzorów (6.19)–(6.21), do obliczania wyróżnionych parametrów dwuwymiarowej zmiennej losowej ciągłej niezbędna jest znajomość odpowiednich parametrów jednowymiarowej zmiennej losowej ciągłej. Są one wyznaczane następująco:

$$E(X)=\int_{-\infty}^{+\infty} xf(x)\,dx, \tag{6.22}$$

$$E(Y)=\int_{-\infty}^{+\infty} yf(y)\,dy, \tag{6.23}$$

$$D^2(X)=\sigma_x^2=\int_{-\infty}^{+\infty}[x-E(X)]^2 f(x)\,dx=\int_{-\infty}^{+\infty}x^2 f(x)\,dx-[E(X)]^2, \tag{6.24}$$

$$D^2(Y)=\sigma_y^2=\int_{-\infty}^{+\infty}[y-E(Y)]^2 f(y)\,dy=\int_{-\infty}^{+\infty}y^2 f(y)\,dy-[E(Y)]^2. \tag{6.25}$$

W przypadku, gdy zmienne losowe tworzące dwuwymiarową zmienną (X,Y) są niezależne, zachodzą następujące zależności:

$$F(x_0,y_0)=F(x_0)\cdot F(y_0) \tag{6.26}$$

oraz

$$f(x,y)=f(x)\cdot f(y). \tag{6.27}$$

Jeżeli zmienne losowe tworzące zmienną dwuwymiarową są zależne, to ich rozkład można przedstawić w układzie warunkowym. Warunkowa funkcja gęstości zmiennej losowej X jest określona wzorem:

$$f(x|y)=\frac{f(x,y)}{f(y)}, \tag{6.28}$$

warunkowa funkcja zmiennej losowej Y powstaje jako:

$$f(y|x)=\frac{f(x,y)}{f(x)}, \tag{6.29}$$

gdzie $f(x)$ i $f(y)$ są funkcjami gęstości odpowiednich rozkładów brzegowych i spełniają warunki:

$$f(x)=\int_{-\infty}^{+\infty} f(x,y)\,dy, \tag{6.30}$$

$$f(y)=\int_{-\infty}^{+\infty} f(x,y)\,dx. \tag{6.31}$$

Niech zmienna losowa (X,Y) ma rozkład ciągły o gęstości podanej wzorem:

$$f(x,y)=\begin{cases} Cx & \text{dla} \quad 0<x<y<1 \\ 0 & \text{dla} \quad \text{pozostałych } x \text{ i } y \end{cases}$$

Naszym zadaniem jest:

a) wyznaczenie stałej C,
b) obliczenie wartości dystrybuanty $F(0{,}5;2)$,
c) wyznaczenie gęstości brzegowych $f(x)$ oraz $f(y)$,
d) sprawdzenie, czy zmienne losowe X i Y są niezależne,
e) obliczenie współczynnika korelacji $\rho(X,Y)$.

Stałą C wyznaczamy z warunku:

$$\int_{-\infty}^{+\infty}\int_{-\infty}^{+\infty} f(x,y)\,dxdy=1. \tag{6.32}$$

Mamy więc:

$$\int_0^1 dx\int_x^1 Cx\,dy=C\int_0^1 x[y]_x^1\,dx=C\int_0^1 x(1-x)\,dx=C\int_0^1\left(x-x^2\right)dx=C\left[\frac{x^2}{2}-\frac{x^3}{3}\right]_0^1=C\frac{1}{6}$$

$C\frac{1}{6}=1$, stąd $C=6$.

Aby funkcja $f(x,y)$ była funkcją gęstości dwuwymiarowej zmiennej losowej (X,Y), stała C powinna być równa 6.

Dystrybuanta zmiennej losowej (X,Y) typu ciągłego jest określona wzorem:

$$F(x,y)=\int_{-\infty}^{x}\int_{-\infty}^{y} f(u,t)dudt. \tag{6.33}$$

Mamy zatem:

$$F(0{,}5;2)=\int_{-\infty}^{0,5}\int_{-\infty}^{2} f(x,y)dxdy=\int_{0}^{0,5} dx\int_{x}^{1} 6xdy=6\int_{0}^{0,5} x[y]_{x}^{1}dx=6\int_{0}^{0,5} x(1-x)dx=$$

$$=6\left[\frac{x^2}{2}-\frac{x^3}{3}\right]_0^{0,5}=0{,}5.$$

Z kolei, korzystając ze wzorów (6.30) i (6.31) wyznaczamy gęstości brzegowe zmiennych losowych X i Y:

$$f(x)=\int_{x}^{1} 6xdy=6x(1-x) \text{ dla } x\in(0,1),$$

$$f(x)=0 \text{ dla } x\notin(0,1).$$

Tak więc:

$$f(x)=\begin{cases}6x(1-x) & \text{dla } x\in(0,1)\\ 0 & \text{dla } x\notin(0,1)\end{cases}.$$

Gęstość brzegową zmiennej Y wyznaczamy następująco:

$$f(y)=\int_{0}^{y} 6ydx=3y^2 \text{ dla } y\in(0,1)$$

i

$$f(y)=0 \text{ dla } y\notin(0,1).$$

Mamy zatem:

$$f(y)=\begin{cases}3y^2 & \text{dla } y\in(0,1)\\ 0 & \text{dla } y\notin(0,1)\end{cases}.$$

W celu sprawdzenia, czy zmienne losowe X i Y są niezależne, korzystamy z relacji (6.27). Okazuje się, że warunek niezależności nie jest spełniony. Oznacza to, że zmienne losowe X i Y nie są niezależne (czyli są zależne).

Aby wyznaczyć współczynnik korelacji $\rho(X,Y)$ korzystamy ze wzoru:

$$\rho(X,Y)=\frac{\operatorname{cov}(X,Y)}{D(X)D(Y)}. \tag{6.34}$$

Kolejność obliczeń jest następująca:

$$E(X)=6\int_{0}^{1} x^2(1-x)dx=6\left[\frac{x^3}{3}-\frac{x^4}{4}\right]_0^1=0{,}5,$$

$$E(Y)=3\int_{0}^{1} y^3dy=0{,}75,$$

$$D^2(X)=6\int_{0}^{1} x^3(1-x)dx-0{,}25=6\left[\frac{x^4}{4}-\frac{x^5}{5}\right]_0^1-0{,}25=0{,}05$$

$$D(X)=\sqrt{0{,}05}=0{,}224,$$

$$D^2(Y)=3\int_{0}^{1} y^4dy-0{,}5625=0{,}6-0{,}5625=0{,}0375$$

$$D(Y)=0{,}194$$

$$\operatorname{cov}(X,Y)=E(XY)-E(X)E(Y)=\int_{-\infty-\infty}^{+\infty+\infty} xy\cdot f(x,y)dxdy-0{,}5\cdot 0{,}75=$$

$$=6\int_{0}^{1} x^2dx\int_{x}^{1} ydy-0{,}375=3\int_{0}^{1} x^2\left[y^2\right]_{x}^{1}dx-0{,}375=3\left[\frac{x^3}{3}-\frac{x^5}{5}\right]_0^1-0{,}375=$$

$$=0{,}4-0{,}375=0{,}025.$$

Podstawiając dane liczbowe do wzoru (6.34) otrzymujemy:

$$\rho(X,Y)=\frac{0{,}025}{0{,}224\cdot 0{,}194}=\frac{0{,}025}{0{,}043456}=0{,}575.$$

Pomiędzy zmiennymi losowymi X i Y istnieje zatem umiarkowana zależność dodatnia.

6.3. Opis współzależności zjawisk masowych

Zależności między zjawiskami mogą być trojakiego rodzaju: **funkcyjne**, **korelacyjne** oraz **stochastyczne**. **Zależności funkcyjne** występują w naukach fizycznych i przyrodniczych. Przykładem tego może być zależność drogi (s) od czasu (t) i prędkości (v). W podanym przykładzie wartości zmiennych t i v w sposób jednoznaczny określają s. Podobnie wynagrodzenie w akordzie prostym uzależnione jest funkcyjnie od wykonanej pracy, a kwota odsetek prostych – od złożonego kapitału, stopy procentowej oraz czasu.

W naukach ekonomicznych najczęściej występuje **zależność korelacyjna**, w której określonym wartościom jednej zmiennej odpowiadają ściśle określone **średnie** wartości drugiej zmiennej. Takimi relacjami są np. związki pomiędzy wydatkami na żywność a liczbą osób w rodzinie, wiekiem pracownika a jego stażem pracy, poziomem wykształcenia i wynagrodzeniem itp.

Zależność stochastyczna występuje wtedy, gdy wraz ze zmianą jednej zmiennej zmienia się **rozkład prawdopodobieństwa** drugiej zmiennej. Na przykład wydatki na żywność w rodzinach o takiej samej liczbie osób nie są jednakowe. Wynika to z faktu, że rodziny o takiej samej liczbie osób różnią się zazwyczaj innymi cechami determinującymi poziom wydatków na żywność (np. dochodem rodziny, płcią i wiekiem rodziny, wykształceniem członków rodziny, charakterem wykonywanej pracy). Przy ustalonej liczbie osób w rodzinie istnieje zatem określony – warunkowy – rozkład wydatków na żywność. Należy oczekiwać, iż rozkłady te będą, przy odmiennej liczbie osób w rodzinie, różnić się między sobą. Jednakże uzasadnione jest przypuszczenie, że średnie wydatków będą wzrastać w miarę wzrostu liczby osób w rodzinie. Zależność, w której wartość jednej zmiennej wpływa na rozkład prawdopodobieństwa drugiej zmiennej, nazywamy **zależnością stochastyczną**. Gdyby – przy wszystkich możliwych wartościach liczby osób w rodzinie – rozkłady warunkowe wydatków na żywność były jednakowe, to cechy te byłyby **stochastycznie niezależne**.

Można zatem stwierdzić, że zależność korelacyjna jest szczególnym przypadkiem zależności stochastycznej. Jeżeli między badanymi zmiennymi nie ma związku stochastycznego, to nie ma również zależności korelacyjnej. Odwrotne twierdzenie nie jest jednak prawdziwe. Wynika to z faktu, że określonej liczbie identycznych wariantów zmiennej odpowiada zawsze ta sama średnia, ale daną średnią można uzyskać z różnej kombinacji wariantów zmiennej. Na przykład wariantom cechy „waga" 62 kg i 68 kg odpowiada $\bar{x} = 65$ kg, ale średnią $\bar{x} = 65$ można otrzymać z takich wariantów cechy, jak: 60 kg i 70 kg, 63 kg i 67 kg, 61 kg i 69 kg itd.

Zwrócić należy uwagę na to, że badanie związków korelacyjnych ma sens jedynie wtedy, gdy między zmiennymi istnieje więź **przyczynowo-skutkowa**, dająca się logicznie uzasadnić. Znane są w literaturze przykłady badania, nawet istotnej statystycznie zależności między np. liczbą zajętych gniazd bocianich a liczbą urodzeń na danym terenie czy między liczbą zarejestrowanych odbiorników radiowo-telewizyjnych a liczbą chorych umysłowo. Zależności te są jednak nieuzasadnione logicznie. Noszą one nazwę **korelacji iluzorycznej** lub **pozornej**. Tak więc w analizie współzależności zjawisk należy wyróżnić dwa podejścia: **jakościowe** i **ilościowe**. Najpierw na podstawie analizy merytorycznej należy uzasadnić logiczne występowanie związku, a dopiero potem przystąpić do analizy ilościowej współzależności.

Wśród związków przyczynowo-skutkowych można wyróżnić zależności dwustronne i jednostronne. Związki dwustronne cechuje wzajemne oddziaływanie na siebie badanych zjawisk (trudno jest tu określić, co jest skutkiem, a co przyczyną). Przykładem tego rodzaju zależności jest relacja między kwotami wydatkowanymi na reklamę a zyskiem w przedsiębiorstwie. Niewątpliwie im większe są wydatki na reklamę, tym większego zysku należy oczekiwać. Jednakże poziom osiągniętego zysku determinuje wielkość funduszu przeznaczonego na reklamę.

W związkach jednostronnych występuje jednokierunkowe oddziaływanie przyczyny na skutek. Tego typu powiązanie występuje np. między wiekiem samochodu a jego ceną. W tym przypadku cenę samochodu determinuje jego wiek, a nie odwrotnie.

Ilościowa analiza współzależności obejmuje:

1) **analizę korelacji** (pomiar siły i kierunku zależności między badanymi cechami),

2) **analizę regresji** (badanie mechanizmu powiązań między cechami, którego wyrazem są **funkcje regresji**).

6.3.1. Formy prezentacji materiału statystycznego

W analizie współzależności zjawisk dane liczbowe (materiał statystyczny) mogą być ujmowane w postaci:

a) szeregów korelacyjnych (**dane indywidualne**),

b) tablicy korelacyjnej,

c) diagramu korelacyjnego.

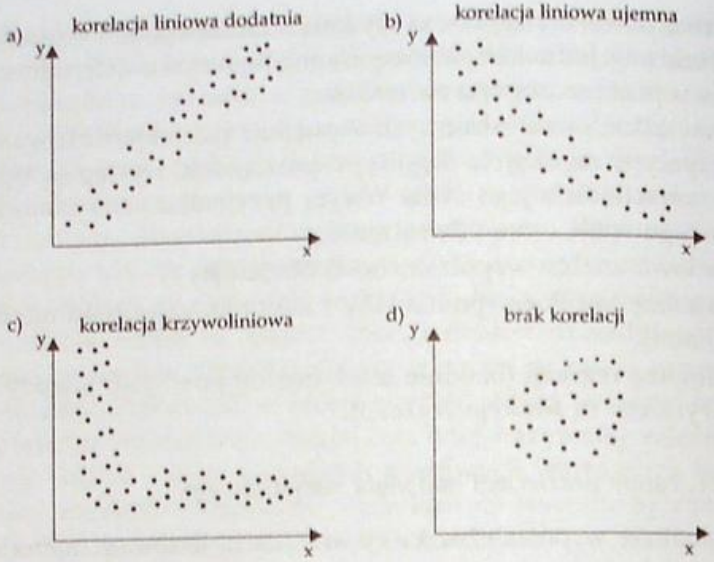
Szereg korelacyjny tworzą dwa wiersze (lub dwie kolumny) zawierające warianty odpowiadających sobie cech statystycznych (x_i oraz y_i):

$$x_i: \quad x_1 \quad x_2 \quad x_3 \quad \dots \quad x_n,$$
$$y_i: \quad y_1 \quad y_2 \quad y_3 \quad \dots \quad y_n.$$

Jeżeli warianty obydwu zmiennych wykazują zmiany **jednokierunkowe** (tzn. wartości obydwu zmiennych na ogół rosną lub maleją), to mamy do czynienia z **korelacją dodatnią**. Gdy wzrostom wartości jednej zmiennej odpowiadają spadki wartości drugiej zmiennej, wówczas mówimy o **korelacji ujemnej**. W przypadku korelacji ujemnej zmiany wartości cech wykazują charakter **różnokierunkowy**.

Prostym sposobem stwierdzania, czy między dwiema zmiennymi występuje korelacja, jest **diagram korelacyjny**. Diagram sporządzany jest w prostokątnym układzie współrzędnych. Na osi odciętych odkładane są wartości jednej zmiennej, a na osi rzędnych – drugiej zmiennej. Zaobserwowane wartości obu cech (x_i, y_i) są współrzędnymi punktów wykresu. Punkty te tworzą mniej lub bardziej wyraźną „smugę", co pozwala na ocenę, czy między zmiennymi istnieje zależność, jaka jest jej siła, kierunek i kształt. Typowe układy punktów na diagramie korelacyjnym przedstawia rys. 6.1.

Jeżeli badana zbiorowość jest liczna, wyniki dwóch cech grupuje się w tablicy korelacyjnej (por. tab. 6.2).



Rys. 6.1. Wykresy korelacyjne: a) korelacja liniowa dodatnia, b) korelacja liniowa ujemna, c) korelacja krzywoliniowa, d) brak korelacji
Źródło: M. Sobczyk, *Statystyka. Podstawy teoretyczne, przykłady i zadania*, UMCS, Lublin 2000, s. 228.

Tab. 6.2. Tablica korelacyjna dwu zmiennych

x_i \ y_j	y_1	y_2	...	y_j	...	y_r	$\sum_j$
x_1	n_{11}	n_{12}	...	n_{1j}	...	n_{1r}	$n_{1\cdot}$
x_2	n_{21}	n_{22}	...	n_{2j}	...	n_{2r}	$n_{2\cdot}$
⋮	⋮	⋮	...	⋮	...	⋮	⋮
x_i	n_{i1}	n_{i2}	...	n_{ij}	...	n_{ir}	$n_{i\cdot}$
⋮	⋮	⋮	...	⋮	...	⋮	⋮
x_k	n_{k1}	n_{k2}	...	n_{kj}	...	n_{kr}	$n_{k\cdot}$
$\sum_i$	$n_{\cdot 1}$	$n_{\cdot 2}$	...	$n_{\cdot j}$	...	$n_{\cdot r}$	n

Źródło: opracowanie własne.

Tablica korelacyjna przedstawia **rozkład dwuwymiarowy**, czyli łączny rozkład zbiorowości według dwóch cech. Na skrzyżowaniu kolumn z wierszami występują liczebności jednostek statystycznych, u których zaobserwowano jednoczesne występowanie określonej wartości x_i oraz y_j. Układ liczebności w tablicy korelacyjnej informuje o sile i kierunku zależności między badanymi cechami. Liczebności skupione wzdłuż przekątnych świadczą o istnieniu związku liniowego, a liczebności rozmieszczone nieregularnie – wskazują na związek krzywoliniowy lub niezależność cech.

W analizie korelacji wykorzystywane są różne mierniki, określone **rodzajem cech** (ilościowe, jakościowe), sposobem prezentacji **informacji statystycznych** (dane indywidualne, tablice korelacyjne) oraz **charakterem związku** (liniowy, krzywoliniowy). Wszystkie mierniki korelacji są liczbami niemianowanymi.

6.3.2. Współczynnik korelacji liniowej Pearsona

W przypadku, gdy obie badane cechy są ilościowe, a związek między nimi jest liniowy, najczęściej stosowaną miarą współzależności pozostaje **współczynnik korelacji liniowej Pearsona**. Współczynnik ten – w przypadku danych indywidualnych – jest określony następująco:

$$r_{xy} = r_{yx} = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{ns(x)s(y)} = \frac{\mathrm{cov}(x,y)}{s(x)s(y)}, \tag{6.35}$$

gdzie $s(x)$ oraz $s(y)$ są odchyleniami standardowymi dotyczącymi odpowiednio cechy x i y. Współczynnik (6.35) jest wyznaczany poprzez **standaryzację kowariancji**. Kowariancja jest średnią arytmetyczną iloczynu odchyleń poszczególnych zmiennych od ich średnich arytmetycznych:

$$\mathrm{cov}(x,y) = \mathrm{cov}(y,x) = \frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y}). \tag{6.36}$$

Kowariancja jest liczbą mianowaną (wyrażona w jednostkach będących iloczynem jednostek cech x i y) i informuje o kierunku współzależności (dodatnia, ujemna). Jeśli $\mathrm{cov}(x,y) = 0$, to między zmiennymi x i y brakuje zależności (zmienne są nieskorelowane). W przypadku gdy $\mathrm{cov}(x,y) > 0$ – między zmiennymi występuje zależność dodatnia, gdy $\mathrm{cov}(x,y) < 0$ – mamy do czynienia z zależnością ujemną. Kowariancja przyjmuje wielkości liczbowe z przedziału:

$$-s(x)s(y) \le \mathrm{cov}(x,y) \le s(x)s(y). \tag{6.37}$$

Standaryzując kowariancję otrzymujemy unormowany miernik siły i kierunku współzależności liniowej dwóch cech mierzalnych x i y (r_{xy}).

Znak współczynnika korelacji liniowej Pearsona informuje o kierunku współzależności, wartość bezwzględna – o sile tej współzależności.

Współczynnik korelacji liniowej Pearsona przyjmuje wartości liczbowe z przedziału domkniętego: $-1 \le r_{xy} \le +1$. W ocenie siły współzależności często korzysta się z następujących przedziałów:

$|0 - 0{,}2|$ – współzależność bardzo słaba,
$|0{,}2 - 0{,}4|$ – współzależność słaba,
$|0{,}4 - 0{,}6|$ – współzależność umiarkowana,
$|0{,}6 - 0{,}8|$ – współzależność silna,
$|0{,}8 - 1{,}0|$ – współzależność bardzo silna.

Wartość współczynnika korelacji liniowej równa 1 lub –1 wskazuje na istnienie zależności funkcyjnej między zmiennymi. Miernik ten określa siłę współzależności, ale nie wskazuje, czy cecha x wpływa na y, czy też odwrotnie. Własność tę określamy mianem **symetryczności** ($r_{xy} = r_{yx}$).

Współczynnik korelacji r_{xy} jest określony wskaźnikiem, a nie pomiarem na skali liniowej o jednakowych jednostkach. Dlatego też nie można mówić, że np. zależność $r_{xy} = 0{,}90$ jest dwukrotnie silniejsza niż przy $r_{xy} = 0{,}45$.

Kwadrat współczynnika r_{xy} nazywamy **współczynnikiem determinacji**. Współczynnik determinacji – wyrażany najczęściej w procentach – informuje o tym, jaka część zmian zmiennej zależnej (skutek) jest wyjaśniona przez zmiany zmiennej niezależnej (przyczyny).

Technikę obliczania współczynnika korelacji liniowej Pearsona ilustruje poniższy przykład. Załóżmy, że w pewnym Urzędzie Stanu Cywilnego przeprowadzono (w jednym dniu) reprezentacyjne badanie zawartych małżeństw według wieku męża i żony. Wyniki badania losowo wybranych par przedstawiają się następująco:

Wiek męża (x_i)	18	19	20	21	23	24	26	27	27	30
Wiek żony (y_i)	19	21	23	21	20	23	26	25	26	34

Naszym zadaniem jest określenie siły i kierunku współzależności między badanymi cechami.

Na podstawie diagramu korelacyjnego można stwierdzić, że zależność między badanymi zmiennymi ma charakter prostoliniowy. Informacje liczbowe o badanych cechach tworzą dane indywidualne. Do oceny siły i kierunku zależności między wiekiem męża i wiekiem żony wykorzystany zostanie współczynnik określony wzorem (6.35). Obliczenia pomocnicze z tym związane przedstawiono w tab. 6.3.

Tab. 6.3. Obliczenia pomocnicze

x_i	y_i	$x_i - \bar{x}$	$y_i - \bar{y}$	$(x_i - \bar{x})(y_i - \bar{y})$	$(x_i - \bar{x})^2$	$(y_i - \bar{y})^2$
18	19	–5,5	–4,8	26,4	30,25	23,04
19	21	–4,5	–2,8	12,6	20,25	7,84
20	23	–3,5	–0,8	2,8	12,25	0,64
21	21	–2,5	–2,8	7,0	6,25	7,84
23	20	–0,5	–3,8	1,9	0,25	14,44
24	23	0,5	–0,8	–0,4	0,25	0,64
26	26	2,5	2,2	5,5	6,25	4,84
27	25	3,5	1,2	4,2	12,25	1,44
27	26	3,5	2,2	7,7	12,25	4,84
30	34	6,5	10,2	66,3	42,25	104,04
235	238	X	X	134,0	142,5	169,60

Źródło: obliczenia własne.

Średni wiek kobiet zawierających w badanym dniu związek małżeński wynosi: $\bar{x} = 23{,}5$ lat, średni wiek mężczyzn jest równy: $\bar{y} = 23{,}8$ lat. Odchylenia standardowe wieku wynoszą natomiast: $s(x) = 3{,}8$ lat oraz $s(y) = 4{,}1$ lat. Wykorzystując wzór (6.35) otrzymujemy:

$$r_{xy} = \frac{134}{10 \cdot 4{,}1 \cdot 3{,}8} = \frac{13{,}4}{4{,}1 \cdot 3{,}8} = 0{,}86.$$

Otrzymany wynik oznacza, że między badanymi zmiennymi istnieje bardzo silna zależność korelacyjna dodatnia. Współczynnik determinacji jest tu równy: $r_{xy}^2 = (0{,}86)^2 = 0{,}7396 = 73{,}96\%$. W niemal 74% zmiany jednej cechy są uwarunkowane zmianami drugiej. W tym przypadku trudno jest stwierdzić, jaki jest kierunek wpływu zmiennych, tzn. czy wiek męża wpływa na wiek żony, czy też odwrotnie.

W przypadku danych przedstawionych w tablicy korelacyjnej, ustalenie siły i kierunku związku między badanymi cechami przebiega w nieco odmienny sposób. Współczynnik korelacji liniowej Pearsona przyjmuje wówczas następującą postać:

$$r_{xy} = r_{yx} = \frac{\sum_{i=1}^{k}\sum_{j=1}^{r}(x_i - \bar{x})(y_i - \bar{y})n_{ij}}{ns(x)s(y)} = \frac{\operatorname{cov}(x,y)}{s(x)s(y)}. \qquad (6.38)$$

Tak więc kowariancja jest tutaj określona wzorem:

$$\operatorname{cov}(x,y) = \operatorname{cov}(y,x) = \frac{1}{n}\sum_{i=1}^{k}\sum_{j=1}^{r}(x_i - \bar{x})(y_j - \bar{y})n_{ij}. \qquad (6.39)$$

Oczywiste jest, że w sytuacji pogrupowania wartości badanych cech w tablicy korelacyjnej w szeregi rozdzielcze przedziałowe, w obliczeniach należy korzystać ze środków przedziałów klasowych.

Załóżmy, że w 100 szkołach przeprowadzono badanie mające na celu określenie zależności między liczbą sal lekcyjnych (x) a liczbą uczniów w szkole (y). Otrzymane wyniki przedstawiono w tablicy korelacyjnej (tab. 6.4).

Tab. 6.4. Liczba sal lekcyjnych i liczba uczniów w 100 szkołach

Liczba uczniów w szkole (x)	Liczba sal lekcyjnych (y)			Razem
	4–8	8–12	12–16	
60–120	10	–	–	10
120–180	–	10	10	20
180–240	–	20	20	40
240–300	–	–	20	20
300–360	–	–	10	10
Razem	10	30	60	100

Źródło: dane umowne.

Do zbadania siły i kierunku zależności między badanymi cechami wykorzystamy współczynnik korelacji liniowej Pearsona z tablic korelacyjnych (wzór 6.38). W pierwszej kolejności z rozkładów brzegowych cech obliczamy średnie arytmetyczne i odchylenia standardowe. Otrzymujemy:

$$\bar{x} = \frac{90 \cdot 10 + 150 \cdot 20 + 210 \cdot 40 + 270 \cdot 20 + 330 \cdot 10}{100} = 210 \text{ uczniów},$$

$$\bar{y} = \frac{6 \cdot 10 + 10 \cdot 30 + 14 \cdot 60}{100} = 12 \text{ sal},$$

$s(x) = 65{,}7$ uczniów oraz $s(y) = 2{,}7$ sal lekcyjnych.

Obliczenia związane z ustaleniem wartości liczbowej licznika wzoru (6.38) zawarte są w tab. 6.5.

Tab. 6.5. Obliczenia pomocnicze

$\hat{x}_i - \bar{x}$ \ $\hat{y}_j - \bar{y}$	–6	–2	2	Razem
–120	10	–	–	10
–60	–	10	10	20
0	–	20	20	40
60	–	–	20	20
120	–	–	10	10
Razem	10	30	60	100

Źródło: obliczenia własne.

Suma ważonych iloczynów par odchyleń badanych zmiennych jest zatem równa:

$$\begin{aligned}
10(-120)(-6) &= 7200\\
10(-60)(-2) &= 1200\\
20 \cdot 0 \cdot (-2) &= 0\\
20 \cdot 0 \cdot 2 &= 0\\
20 \cdot 60 \cdot 2 &= 2400\\
10 \cdot 120 \cdot 2 &= 2400\\
10(-60) \cdot 2 &= -1200\\
&\underline{\quad\quad}\\
&12\,000
\end{aligned}$$

Obliczony z tablicy korelacyjnej współczynnik r_{xy} wynosi:

$$r_{xy} = \frac{12\,000}{100 \cdot 65{,}7 \cdot 2{,}7} = \frac{120}{65{,}7 \cdot 2{,}7} = 0{,}68.$$

Tak więc pomiędzy liczbą sal lekcyjnych a liczbą uczniów w szkole zachodzi silna zależność korelacyjna dodatnia. Kwadrat współczynnika korelacji $r_{xy}^2 = 0{,}4624 = 46{,}24\%$ informuje o tym, że zmienność sal lekcyjnych w szkołach jest w ok. 46% wyjaśniona zmiennością liczby uczniów.

6.3.3. Współczynnik korelacji rang Spearmana

Współczynnik korelacji **rang Spearmana** (zwany też współczynnikiem korelacji **kolejnościowej**) jest wykorzystywany głównie do badania siły i kierunku współzależności między cechami porządkowymi. Nasilenie badanej cechy jest wówczas określane za pomocą **rang. Rangowanie** – to nadawanie wartościom badanych cech kolejnych numerów od 1 do n. Ranga 1 oznacza zazwyczaj największą wartość cechy lub mającą w badaniu pozytywne znaczenie, a ranga n jest przypisywana najmniejszej wartości cechy. Sposób rangowania może być odwrotny; musi być jednak taki sam dla obydwu cech. W przypadku, gdy występują jednakowe wartości cech, przyporządkowujemy im średnią arytmetyczną z kolejnych rang. Mówimy wówczas o występowaniu tzw. **węzłów** lub **rang wiązanych**. Korelacja dodatnia między badanymi cechami występuje wtedy, gdy ma miejsce zgodność uporządkowań (rang), a korelacja ujemna – gdy uporządkowania są przeciwstawne. Współczynnik korelacji rang można również wykorzystywać w przypadku cech mierzalnych, ale tylko dla niewielkiej zbiorowości. Pierwotne wartości zmiennych w takim przypadku należy również poddać operacji rangowania. Współczynnik korelacji rang Spearmana jest obliczany ze wzoru:

$$r_S = 1 - \frac{6\sum_{i=1}^{n} d_i^2}{n\left(n^2 - 1\right)}, \tag{6.40}$$

gdzie d_i określa różnicę między rangami odpowiadających sobie wartości cechy x_i oraz y_i $(i = \ldots 1, 2, \ldots, n)$.

Współczynnik korelacji kolejnościowej przyjmuje wartości z przedziału domkniętego od –1 do +1, przy czym wartość liczbowa określa siłę zbieżności rang, natomiast znak informuje, czy dla obu cech statystycznych występuje zgodność, czy też rozbieżność ich rang.

Dział marketingowy firmy krawieckiej podjął badania, których celem było ustalenie współzależności między popytem na garsonki a ich fasonem. Analizie poddano fason 10 rodzajów garsonek. W odniesieniu do fasonu wyróżniono cztery warianty: super modny, modny, praktyczny oraz niemodny. Każdemu z tych wariantów przypisano skalę porządkową w następujący sposób: fason super modny 4 punkty, modny 3 punkty, praktyczny 2 punkty oraz niemodny 1 punkt. W przypadku popytu rangowanie przebiegało od 10 (dla typu garsonek o największej sprzedaży) aż do 1 punktu (w przypadku garsonek o najmniejszej sprzedaży). Dane wyjściowe przedstawiają się następująco:

Typ garsonki	A	B	C	D	E	F	G	H	I	J
Popyt w tys. szt.	0,5	0,8	0,4	0,2	0,1	0,7	0,4	0,5	0,9	1,0
Fason (w punktach)	3	4	2	1	1	4	2	3	4	4

Pomocnicze obliczenia niezbędne do wyznaczenia współczynnika korelacji rang Spearmana zawiera tab. 6.6.

Tab. 6.6. Obliczenia pomocnicze

Typ garsonki	Rangi dla		d_i	d_i^2
	popytu	fasonu		
A	5,5	5,5	0	0
B	8	8,5	–0,5	0,25
C	3,5	3,5	0	0
D	2	1,5	0,5	0,25
E	1	1,5	–0,5	0,25
F	7	8,5	–1,5	2,25
G	3,5	3,5	0	0
H	3,5	5,5	0	0
I	9	8,5	0,5	0,25
J	10	8,5	1,5	2,25
Razem	X	X	X	5,50

Źródło: obliczenia własne.

Wykorzystując wzór (6.40) otrzymujemy:

$$r_S = 1 - \frac{6 \cdot 5{,}5}{10(10^2 - 1)} = 0{,}97.$$

Współczynnik korelacji rang Spearmana przyjmuje wartości liczbowe z przedziału domkniętego $\langle -1,+1 \rangle$. Badana zależność jest tym silniejsza, im r_S jest bliższy jedności. Otrzymany wynik $r_S = 0{,}97$ oznacza dużą zgodność badanych cech: im bardziej modny jest fason produkowanych garsonek, tym większa jest ich sprzedaż (wyższy popyt).

6.3.4. Wskaźniki (stosunki) korelacyjne Pearsona

Wskaźniki korelacyjne Pearsona wyznacza się z danych ujętych w postaci tablicy korelacyjnej. Służą one do określania siły współzależności między zmiennymi x i y. Stosunki korelacyjne są – z reguły – niesymetryczne. Oznacza to, że oddzielnie należy wyznaczać zależność cechy x od cechy y (wpływ y na x) i oddzielnie zależność cechy y od cechy x (wpływ x na y). Wskaźnik (stosunek) korelacyjny Pearsona, określający pierwszą zależność, jest oznaczany symbolem e_{xy}, natomiast drugą zależność – e_{yx}. Są one miernikami zależności **krzywoliniowej** i mogą być wykorzystywane wtedy, gdy obie cechy są **mierzalne**, jak również wówczas, gdy jedna z nich jest **niemierzalna** (jakościowa). W tym ostatnim przypadku zmienną mierzalną musi być zmienna zależna, co daje możliwość obliczenia wariancji.

Konstrukcja wskaźników korelacyjnych oparta jest na odpowiednich **równościach wariancyjnych**. Równość wariancyjna zmiennej y względem zmiennej x przyjmuje następującą postać:

$$s^2(y) = s^2(\bar{y}_i) + \overline{s_i^2(y)}, \tag{6.41}$$

gdzie $s^2(y)$ jest wariancją ogólną, $s^2(\bar{y}_i)$ – wariancją międzygrupową, natomiast $\overline{s_i^2(y)}$ – wariancją wewnątrzgrupową.

Dzieląc równość (6.41) przez $s^2(y)$ otrzymujemy:

$$1 = \frac{s^2(\bar{y}_i)}{s^2(y)} + \frac{\overline{s_i^2(y)}}{s^2(y)}. \tag{6.42}$$

Wskaźnik korelacyjny y względem x obliczamy następująco:

$$e_{yx} = +\sqrt{\frac{s^2(\bar{y}_i)}{s^2(y)}} = +\sqrt{1 - \frac{\overline{s_i^2(y)}}{s^2(y)}}. \tag{6.43}$$

W analogiczny sposób obliczamy wskaźnik korelacyjnej x względem y, a mianowicie:

$$s^2(x) = s^2(\bar{x}_j) + \overline{s_j^2(x)}, \tag{6.44}$$

$$1 = \frac{s^2(\bar{x}_j)}{s^2(x)} + \frac{\overline{s_j^2(x)}}{s^2(x)}, \tag{6.45}$$

$$e_{xy} = +\sqrt{\frac{s^2(\bar{x}_j)}{s^2(x)}} = +\sqrt{1 - \frac{\overline{s_j^2(x)}}{s^2(x)}}. \tag{6.46}$$

Wielkości $s^2(\bar{y}_i)$ oraz $s^2(\bar{x}_j)$ to wariancje średnich grupowych, mierzące zróżnicowanie cech między grupami, gdyż:

$$s^2(\bar{y}_i) = \frac{1}{n}\sum_{i=1}^{k}(\bar{y}_i - \bar{y})^2 n_i. \tag{6.47}$$

oraz

$$s^2(\bar{x}_j) = \frac{1}{n}\sum_{j=1}^{r}(\bar{x}_j - \bar{x})^2 n_j. \tag{6.48}$$

Wariancje te przyjmują wartość 0 wtedy, gdy wszystkie średnie warunkowe są takie same i równe średniej ogólnej. W takim przypadku zarówno e_{xy}, jak i e_{yx} są równe zeru. Im silniejsza zależność występuje między badanymi cechami, tym większe są różnice średnich warunkowych, a wariancja średnich warunkowych coraz bardziej zbliża się do wariancji ogólnej (e_{xy} i e_{yx} dążą do jedności).

Wskaźniki (stosunki) korelacyjne Pearsona są wielkościami niemianowanymi, zawartymi w przedziale liczbowym 0–1:

$$e_{xy}, e_{yx} \in \langle 0,1 \rangle. \tag{6.49}$$

Do oceny siły zależności stosuje się następujące przedziały:
0–0,2 – zależność bardzo słaba,
0,2–0,4 – zależność słaba,
0,4–0,6 – zależność umiarkowana,
0,6–0,8 – zależność silna,
0,8–1,0 – zależność bardzo silna.

Stosunki korelacyjne nie wskazują kierunku korelacji, a ponadto nie są symetryczne ($e_{yx} \neq e_{xy}$), z wyjątkiem dwóch sytuacji:

1) jeśli zmienne x i y są niezależne korelacyjnie, to $e_{yx} = e_{xy} = 0$;

2) jeśli między zmiennymi x i y zachodzi związek funkcyjny, to $e_{yx} = e_{xy} = 1$.

Pomiędzy wskaźnikami korelacyjnymi a współczynnikiem korelacji liniowej Pearsona zachodzą następujące zależności:

$$e_{xy}^2 \geq r_{xy}^2 \text{ oraz } e_{yx}^2 \geq r_{xy}^2. \tag{6.50}$$

Zależności te są wykorzystywane do oceny, czy badany związek między cechami można uznać za liniowy, czy też nie. Znajdują tu zastosowanie tzw. **mierniki krzywoliniowości** (oznaczane odpowiednio symbolami: m_{xy} oraz m_{yx}), obliczane następująco:

$$m_{xy} = e_{xy}^2 - r_{xy}^2, \tag{6.51}$$

$$m_{yx} = e_{yx}^2 - r_{yx}^2. \tag{6.52}$$

Relacja (6.51) określa stopień krzywoliniowości zależności zmiennej x względem zmiennej y, wzór (6.52) – stopień krzywoliniowości zmiennej y względem x. Mierniki krzywoliniowości przyjmują wartości liczbowe z przedziału:

$$0 \leq m_{xy}, m_{yx} \leq 1. \tag{6.53}$$

Umownie przyjmuje się, że jeśli $m < 0{,}2$, to związek między zmiennymi można uznać za prostoliniowy. W przeciwnym przypadku krzywoliniowość związku należy uznać za zasadną.

W konkretnych sytuacjach badawczych oblicza się tylko jeden stosunek korelacyjny: e_{xy} lub e_{yx}. Wymaga to uprzedniego określenia, która ze zmiennych jest zależna (skutek), a która niezależna (przyczyna). W zapisie wskaźników korelacyjnych: e_{xy} oraz e_{yx} – na pierwszym miejscu występuje zmienna zależna, a na drugim zmienna niezależna.

Badając zmienność wieku osób deklarujących uczestnictwo w wyborach do Sejmu i Senatu RP stwierdzono, że wśród losowo wybranych 1500 osób średni wiek wyniósł 36,4 lat, z przeciętnym zróżnicowaniem 16,6 lat. Wśród 800 wylosowanych kobiet średni wiek wyniósł 42 lata z pięćdziesięcioprocentową zmiennością, wśród wylosowanych mężczyzn średni wiek wyniósł 30 lat, z typowym obszarem zmienności 27 lat $< X_{typ} <$ 33 lata. Zadaniem naszym jest zbadanie zależności między wiekiem a płcią wyborców.

Jedna z wyróżnionych cech jest mierzalna (wiek), a druga niemierzalna (płeć). W takim przypadku miernikiem siły zależności jest wskaźnik korelacyjny Pearsona. Zmienną mierzalną musi być zmienna zależna. Oznaczmy zatem przez x wiek wyborców, a przez y ich płeć. Przy takich oznaczeniach będziemy obliczać stosunek korelacyjny e_{xy}. Mamy więc:

$s(x) = 16{,}6$ lat,

$$\bar{x} = \frac{42 \cdot 800 + 700 \cdot 30}{1500} = \frac{54\,600}{1500} = 36{,}4 \text{ lat},$$

$$s^2(\bar{x}_j) = \frac{(42 - 36{,}4)^2 \cdot 800 + (30 - 36{,}4)^2 \cdot 700}{1500} = \frac{53\,760}{1500} = 35{,}84 \text{ (lat)}^2,$$

$$\overline{s_j^2(x)} = s^2(x) - s^2(\bar{x}_j) = 275{,}56 - 35{,}84 = 239{,}72 \text{ (lat)}^2.$$

Wykorzystując wzór (6.46) otrzymujemy:

$$e_{xy} = +\sqrt{\frac{35{,}84}{275{,}56}} = +\sqrt{1 - \frac{239{,}72}{275{,}56}} = 0{,}36.$$

Uzyskany wynik świadczy o umiarkowanej zależności między badanymi cechami.

6.3.5. Współczynnik korelacji cząstkowej i współczynnik korelacji wielorakiej (wielokrotnej)

Jeżeli na zmienną objaśnianą y oddziałuje więcej niż jedna zmienna objaśniająca $(x_1, x_2, \ldots, x_k)$, a interesuje nas jedynie ścisłość związku korelacyjnego między dwiema zmiennymi przy wyłączeniu wpływu innych, to wykorzystujemy współczynnik korelacji cząstkowej. Współczynniki te oznaczamy symbolem $r_{ij.kl\ldots z}$. Pierwsze dwa indeksy przed kropką (zwane głównymi) oznaczają cechy, których zależność chcemy zbadać. Indeksy po kropce (zwane następczymi) oznaczają cechy eliminowane. Liczba eliminowanych zmiennych wskazuje na rząd współczynnika korelacji cząstkowej. W związku z tym wyróżnia się współczynnik korelacji cząstkowej rzędu pierwszego, drugiego itd.

Do obliczania współczynników korelacji cząstkowej wszystkich rzędów wygodnie jest posłużyć się rachunkiem macierzowym. Współczynnik korelacji cząstkowej wyznaczamy wówczas ze wzoru:

$$r_{ij.kl\ldots z} = \frac{-R_{ij}}{\sqrt{R_{ii} R_{jj}}} = \frac{-S_{ij}}{\sqrt{S_{ii} S_{jj}}}, \tag{6.54}$$

gdzie R_{ij} jest dopełnieniem algebraicznym macierzy R współczynników korelacji par wszystkich włączonych do analizy zmiennych, powstałym przez skreślenie jej i-tego wiersza i j-tej kolumny. R_{ii} oraz R_{jj} są dopełnieniami algebraicznymi macierzy R powstałej przez skreślenie odpowiednio i-tego wiersza i i-tej kolumny oraz j-tego wiersza i j-tej kolumny. Elementami macierzy R są współczynniki korelacji liniowej Pearsona obliczone dla par wszystkich rozpatrywanych zmiennych:

$$R=\begin{bmatrix}1 & r_{12} & r_{13} & \cdots & r_{1z}\\ r_{21} & 1 & r_{23} & \cdots & r_{2z}\\ \cdots & \cdots & \cdots & \cdots & \cdots\\ r_{z1} & r_{z2} & r_{z3} & \cdots & 1\end{bmatrix}$$

Macierz R jest symetryczna, a jej wymiary zależą od liczby wprowadzonych do analizy zmiennych.

Współczynniki korelacji cząstkowej mogą być również obliczone przy wykorzystaniu macierzy wariancji i kowariancji S (por. wzór 6.54). Macierz S jest symetryczna, a jej elementami są wariancje poszczególnych zmiennych (znajdują się one na głównej przekątnej) i kowariancje pomiędzy parami analizowanych zmiennych (znajdują się one poza główną przekątną). Ogólna postać macierzy wariancji i kowariancji jest następująca:

$$S=\begin{bmatrix}s_1^2 & c_{12} & c_{13} & \cdots & c_{1z}\\ c_{21} & s_2^2 & c_{23} & \cdots & c_{2z}\\ \cdots & \cdots & \cdots & \cdots & \cdots\\ c_{z1} & c_{z2} & c_{z3} & \cdots & s_z^2\end{bmatrix}$$

Przykładowo, jeśli zmienne x_1 oraz x_2 oznaczymy arabskimi cyframi 1 i 2, to wzór na kowariancję między nimi ma postać:

$$c_{12}=\frac{1}{n}\sum_{i=1}^{n}(x_{i1}-\bar{x}_1)(x_{i2}-\bar{x}_2), \tag{6.55}$$

gdzie $\bar{x}_1$ i $\bar{x}_2$ są wartościami średnich arytmetycznych odpowiednio cech x_1 i x_2.

Wariancję dla np. cechy x_1 obliczymy ze wzoru:

$$s_1^2=\frac{1}{n}\sum_{i=1}^{n}(x_{i1}-\bar{x}_1)^2. \tag{6.56}$$

Znaki poszczególnych kowariancji informują o kierunkach zależności dla par cech w badanym zbiorze.

Zauważyć należy, że na podstawie macierzy kowariancji S można zbudować macierz korelacji. Budowanie macierzy kowariancji z wykorzystaniem macierzy korelacji jest możliwe tylko wtedy, gdy znane są wartości odchyleń standardowych badanych cech.

Załóżmy, że dana jest macierz kowariancji trzech zmiennych (x_1, x_2, x_3):

$$S=\begin{bmatrix}8 & 5 & 6\\ 5 & 10 & 3\\ 6 & 3 & 12\end{bmatrix}$$

Z zastosowaniem macierzy S należy zbudować macierz korelacji R.

Elementami macierzy R są współczynniki korelacji liniowej między parami wyróżnionego zbioru cech. Współczynniki te obliczymy ze wzoru:

$$r_{ij}=\frac{\operatorname{cov}(x_i,x_j)}{s_i s_j}. \tag{6.57}$$

Mamy zatem:

$$r_{12}=\frac{c_{12}}{s_1 s_2}=\frac{5}{\sqrt{8\cdot 10}}=\frac{5}{8{,}94}\approx 0{,}559,$$

$$r_{13}=\frac{c_{13}}{s_1 s_3}=\frac{6}{\sqrt{8\cdot 12}}=\frac{6}{9{,}798}\approx 0{,}612,$$

$$r_{23}=\frac{c_{23}}{s_2 s_3}=\frac{3}{\sqrt{10\cdot 12}}=\frac{3}{10{,}95}\approx 0{,}274.$$

Macierz korelacji R ma więc następującą postać:

$$R=\begin{bmatrix}1 & 0{,}559 & 0{,}612\\ 0{,}559 & 1 & 0{,}274\\ 0{,}612 & 0{,}274 & 1\end{bmatrix}$$

Zawarte w macierzy R współczynniki korelacji mają znak dodatni. Wskazuje to na taki sam kierunek zmian w wartościach składowych cechy trójwymiarowej.

Niech macierz korelacji ma postać:

$$R=\begin{bmatrix}1 & 0{,}5 & 0{,}8\\ 0{,}5 & 1 & 0{,}4\\ 0{,}8 & 0{,}4 & 1\end{bmatrix}$$

Wariancje poszczególnych zmiennych są natomiast równe: $s_1^2=9$; $s_2^2=16$; $s_3^2=25$. Naszym zadaniem jest zbudowanie macierzy wariancji i kowariancji S.

Korzystając ze wzoru (6.57) mamy:

$$\operatorname{cov}(x_i,x_j)=r_{ij}\cdot s_i\cdot s_j. \tag{6.58}$$

Wykorzystując relację (6.58) mamy:

$c_{12}=r_{12}\cdot s_1\cdot s_2=0{,}5\cdot 3\cdot 4=6,$

$c_{13}=r_{13}\cdot s_1\cdot s_3=0{,}8\cdot 3\cdot 5=12,$

$c_{23}=r_{23}\cdot s_2\cdot s_3=0{,}4\cdot 4\cdot 5=8.$

Ostatecznie macierz wariancji i kowariancji S przyjmuje następującą postać:

$$S=\begin{bmatrix}9 & 6 & 12\\ 6 & 16 & 8\\ 12 & 8 & 25\end{bmatrix}$$

Technikę obliczania współczynników korelacji cząstkowej pokażemy na przykładzie. Współczynniki korelacji między liczbą sklepów (zmienna 1), ich łączną powierzchnią (zmienna 2) oraz liczbą gospodarstw domowych (zmienna 3), obliczone na podstawie informacji liczbowych z 10 wylosowanych województw, kształtowały się następująco: $r_{12}=-0{,}006$; $r_{23}=0{,}018$ oraz

$r_{13} = 0{,}914$. Mamy obliczyć współczynniki korelacji cząstkowej rzędu pierwszego.

W pierwszym kroku budujemy macierz współczynników korelacji R:

$$R = \begin{bmatrix} 1 & -0{,}006 & 0{,}914 \\ -0{,}006 & 1 & 0{,}018 \\ 0{,}914 & 0{,}018 & 1 \end{bmatrix}$$

Wykorzystując wzór (6.54), współczynniki korelacji cząstkowej rzędu pierwszego obliczymy następująco:

$$r_{12\cdot 3} = \frac{-R_{12}}{\sqrt{R_{11}\cdot R_{22}}} = \frac{-\left[-\begin{vmatrix} -0{,}006 & 0{,}018 \\ 0{,}914 & 1 \end{vmatrix}\right]}{\sqrt{\begin{vmatrix} 1 & 0{,}018 \\ 0{,}018 & 1 \end{vmatrix}\cdot\begin{vmatrix} 1 & 0{,}914 \\ 0{,}914 & 1 \end{vmatrix}}} = -0{,}054,$$

$$r_{13\cdot 2} = \frac{-R_{13}}{\sqrt{R_{11}\cdot R_{23}}} = \frac{-\left[+\begin{vmatrix} -0{,}006 & 1 \\ 0{,}914 & 0{,}018 \end{vmatrix}\right]}{\sqrt{\begin{vmatrix} 1 & 0{,}018 \\ 0{,}018 & 1 \end{vmatrix}\cdot\begin{vmatrix} 1 & -0{,}006 \\ -0{,}006 & 1 \end{vmatrix}}} = 0{,}915,$$

$$r_{23\cdot 1} = \frac{-R_{23}}{\sqrt{R_{22}\cdot R_{33}}} = \frac{-\left[-\begin{vmatrix} 1 & -0{,}006 \\ 0{,}914 & 0{,}018 \end{vmatrix}\right]}{\sqrt{\begin{vmatrix} 1 & 0{,}914 \\ 0{,}914 & 1 \end{vmatrix}\cdot\begin{vmatrix} 1 & -0{,}006 \\ -0{,}006 & 1 \end{vmatrix}}} = 0{,}057.$$

Współczynniki korelacji cząstkowej informują zarówno o kierunku, jak i o sile zależności między badanymi zmiennymi (są zawarte w przedziale domkniętym od –1 do +1). Na przykład współczynniki mierzące siłę współzależności pomiędzy liczbą sklepów a liczbą gospodarstw domowych przed i po wyeliminowaniu wpływu łącznej powierzchni sklepów różnią się nieznacznie ($r_{13} = 0{,}914$; $r_{13\cdot 2} = 0{,}915$). Można zatem sądzić, że czynnik wyeliminowany nie odgrywał znaczącej roli w opisie zmienności liczby sklepów.

Jeżeli chcemy zbadać ścisłość związku korelacyjnego pomiędzy wartością jednej cechy (zmienna zależna, objaśniana), a kompleksem innych cech (zmienne niezależne, objaśniające) wówczas właściwą miarą jest **współczynnik korelacji wielorakiej (wielokrotnej)**.Współczynnik ten oznaczamy symbolem R_w lub $R_{123\ldots k}$. Pierwszy indeks (1) oznacza zmienną objaśnianą, pozostałe zaś (2,3...k) – zmienne objaśniające, których łączny wpływ na zmienną objaśnianą chcemy zbadać.

Współczynnik korelacji wielorakiej jest określony wzorem:

$$R_w = R_{1234\ldots k} = \sqrt{1 - \frac{\det R}{\det R_1}}, \qquad (6.59)$$

gdzie R jest macierzą współczynników korelacji par wszystkich rozpatrywanych zmiennych (a więc zmiennej objaśnianej i zmiennych objaśniających), a R_1 jest macierzą współczynników korelacji pomiędzy parami tylko zmiennych objaśniających. Macierz R i R_1 można zapisać następująco:

$$R = \begin{bmatrix} 1 & r_{12} & r_{13} & \cdots & r_{1k} \\ r_{21} & 1 & r_{23} & \cdots & r_{2k} \\ r_{31} & r_{32} & 1 & \cdots & r_{3k} \\ \cdot & \cdot & \cdot & \cdots & \cdot \\ r_{k1} & r_{k2} & r_{k3} & \cdots & 1 \end{bmatrix} = R_1$$

Współczynnik korelacji wielorakiej zawiera się w przedziale:

$$0 \le R_{1234\ldots k} \le 1. \qquad (6.60)$$

Informuje on o natężeniu łącznego wpływu zmiennych objaśniających na zmienną objaśnianą. Nie interpretuje się znaku współczynnika korelacji wielorakiej.

Przy obliczaniu współczynników korelacji cząstkowej dysponowaliśmy następującymi wartościami współczynników korelacji: $r_{12} = -0{,}006$; $r_{23} = 0{,}018$; $r_{13} = 0{,}914$. Zmienna 1 dotyczy liczby sklepów, zmienna 2 – powierzchni sklepów, a zmienna 3 – liczby gospodarstw domowych. Zmienna 1 jest zmienną zależną, a pozostałe – zmiennymi niezależnymi.

Wykorzystywane przy obliczaniu współczynnika korelacji wielorakiej macierze są tutaj równe:

$$R = \begin{bmatrix} 1 & -0{,}006 & 0{,}914 \\ -0{,}006 & 1 & 0{,}018 \\ 0{,}914 & 0{,}018 & 1 \end{bmatrix} = R_1,$$

$\det R = 0{,}1641186$, $\det R_1 = 0{,}999676$.

Współczynnik korelacji wielorakiej (wzór 6.59) jest więc równy:

$$R_{123} = \sqrt{1 - \frac{0{,}1641186}{0{,}999676}} = 0{,}9142.$$

Łączny wpływ powierzchni sklepów oraz liczby gospodarstw domowych na liczbę sklepów jest dość silny. Obliczony współczynnik determinacji ($R^2_{123} \approx 84\%$) wskazuje, że ok. 84% zmienności liczby sklepów mierzonej wariancją wynika ze zmienności powierzchni sklepów i liczby gospodarstw domowych.

6.3.6. Miary współzależności cech jakościowych

W badaniach statystycznych niejednokrotnie zachodzi konieczność ustalenia skojarzeń (asocjacji) między cechami niemierzalnymi bądź między zmienną mierzalną i jakościową. Jeśli każda ze zmiennych niemierzalnych ma dwa warianty (podział dychotomiczny), to zaobserwowane liczebności

można przedstawić w postaci czteropolowej tablicy asocjacji (kontyngencji). Ogólną postać tablicy czteropolowej przedstawia tab. 6.7.

Tab. 6.7. Schemat tablicy czteropolowej

X \ Y	y_1	y_2	Razem
x_1	a	b	a+b
x_2	c	d	c+d
Razem	a+c	b+d	n

W tab. 6.7. symbolami a, b, c, d oznaczono absolutne (lub procentowe) liczebności jednostek mających określone warianty cech. Do pomiaru siły zależności między badanymi cechami, których warianty przedstawiono w formie tablicy czteropolowej, wykorzystuje się tzw. **współczynniki asocjacji** (zwane też współczynnikami zbieżności korelacyjnej). Są nimi:

1) współczynnik zbieżności korelacyjnej Pearsona-Bravaisa:

$$V = \frac{ad - bc}{\sqrt{(a+b)(a+c)(b+d)(c+d)}}, \tag{6.61}$$

2) współczynnik zbieżności korelacyjnej Bykowskiego:

$$W = \frac{(a+d) - (b+c)}{a+b+c+d}, \tag{6.62}$$

3) współczynnik zbieżności korelacyjnej Yula-Kendalla:

$$Q = \frac{ad - bc}{ad + bc}. \tag{6.63}$$

Wartości liczbowe współczynników (6.61) – (6.63) zawierają się w przedziale domkniętym od –1 do +1. Dodatnie wartości świadczą o tym, że I wariant cechy X kojarzy się (współwystępuje) na ogół z I wariantem cechy Y, a II wariant cechy X z II wariantem cechy Y. Ujemne wartości współczynników informują o tym, że I wariant cechy X kojarzy się raczej z II wariantem cechy Y, a II wariant cechy X z I wariantem cechy Y. Im większa jest wartość bezwzględna współczynników, tym silniejszy związek między badanymi cechami.

W 500 rodzinach przeprowadzono obserwacje dotyczące koloru oczu ojca i syna. Otrzymane wyniki przedstawiono w tab. 6.8.

Tab. 6.8. Kolor oczu ojca i syna

Kolor oczu syna	Kolor oczu ojca		Razem
	jasny	ciemny	
Jasny	30	80	110
Ciemny	170	220	390
Razem	200	300	500

Źródło: dane umowne

Czy kolor oczu syna zależny jest od koloru oczu ojca?

Aby odpowiedzieć na powyższe pytanie, obliczymy wartość współczynnika zbieżności korelacyjnej Yula-Kendalla:

$$Q = \frac{30 \cdot 220 - 80 \cdot 170}{30 \cdot 220 + 80 \cdot 170} = 0{,}35.$$

Otrzymany wynik wskazuje na istnienie niezbyt silnej zależności między kolorem oczu ojca i syna.

Wartości pozostałych współczynników zbieżności korelacyjnej są równe:

$$V = \frac{30 \cdot 220 - 170 \cdot 80}{\sqrt{(30+80)(30+170)(80+220)}} = -0{,}138,$$

$$W = \frac{(30+220) - (80+170)}{30+80+170+220} = 0.$$

Różne wartości współczynników zbieżności korelacyjnej są rezultatem odmiennych założeń przyjętych przy konstrukcji poszczególnych wzorów. Dlatego też do oceny ścisłości związku cech w tablicach czteropolowych należy wykorzystywać jeden wzór.

W przypadku, gdy tablica kontyngencji ma więcej niż cztery pola, do wyznaczania współczynników zbieżności korelacyjnej wykorzystuje się statystykę chi-kwadrat (χ^2). Statystyka ta jest określona wzorem:

$$\chi^2 = \sum_{i=1}^{k} \sum_{j=1}^{r} \frac{\left(n_{ij} - \hat{n}_{ij}\right)^2}{\hat{n}_{ij}}, \tag{6.64}$$

gdzie: n_{ij} – to liczebności empiryczne, $\hat{n}_{ij}$ – są liczebnościami teoretycznymi, które powinny wystąpić w i-tej kolumnie tablicy oraz j-tym jej wierszu, gdyby między zmiennymi nie występowała korelacja. Wartości $\hat{n}_{ij}$ obliczamy ze wzoru:

$$\hat{n}_{ij} = \frac{\sum_{i=1}^{k} n_{ij} \cdot \sum_{j=1}^{r} n_{ij}}{n} = \frac{n_{i\cdot} \cdot n_{\cdot j}}{n}. \tag{6.65}$$

Statystyka chi-kwadrat przyjmuje wartości liczbowe z przedziału $\left(0, n\sqrt{(k-1)(r-1)}\right)$, gdzie k jest liczbą kolumn w tablicy, zaś r – liczbą wierszy.

Najczęściej obliczane z wykorzystaniem wielopolowych tablic współczynniki zbieżności korelacyjnej, to:

1) **współczynnik φ Yule'a:**

$$\varphi = \sqrt{\frac{\chi^2}{n}}, \tag{6.66}$$

2) **współczynnik zbieżności Czuprowa:**

$$T_{xy} = +\sqrt{\frac{\chi^2}{n\sqrt{(k-1)(r-1)}}}, \tag{6.67}$$

3) **współczynnik *V* Cramera:**

$$V = \sqrt{\frac{\chi^2}{n\min(k-1, r-1)}},$$

4) **współczynnik *C* Pearsona:**

$$C = \sqrt{\frac{\chi^2}{\chi^2 + n}}. \tag{6.69}$$

Maksymalna wartość współczynnika C Pearsona jest ściśle uzależniona od wymiarów tablicy kontyngencji. Im tablica ma mniejsze wymiary, tym bardziej wartość C oddala się od 1 w kierunku do 0. Dla tablic symetrycznych ($r = k$) maksymalna wartość współczynnika C wynosi:

$$C_{max} = \sqrt{\frac{k-1}{k}} = \sqrt{\frac{r-1}{r}}. \tag{6.70}$$

Dla tablic prostokątnych ($r \neq k$) maksymalna wartość jest równa:

$$C_{max} = \frac{\sqrt{\frac{k-1}{k}} + \sqrt{\frac{r-1}{2}}}{2}. \tag{6.71}$$

Skorygowaną wartość współczynnika C_{kor} obliczamy ze wzoru:

$$C_{kor} = \frac{C}{C_{max}}. \tag{6.72}$$

Interpretacja współczynników φ, T, V oraz C_{kor} jest taka sama, a mianowicie:

1) osiągają one wartość zero w przypadku niezależności badanych cech,

2) są symetryczne,

3) im większa jest siła współzależności między zmiennymi, tym ich wartość liczbowa jest większa,

4) nie wskazują kierunku korelacji badanych zmiennych (ich znak jest zawsze dodatni).

Ponadto każdy z wymienionych współczynników charakteryzują specyficzne własności. I tak, współczynnik φ Yule'a przyjmuje wartości liczbowe z przedziału: $0 \leq \varphi \leq 1$ tylko wtedy, gdy liczba wierszy w tablicy kontyngencji wynosi 2, przy dowolnej liczbie kolumn. Jeśli – przy dowolnej liczbie kolumn – liczba wierszy jest większa od 2, to współczynnik φ Yule'a może przyjmować wartości większe od 1.

Współczynnik zbieżności Czuprowa przyjmuje wartości liczbowe z przedziału: $0 \leq T \leq 1$, ale jedynie wtedy gdy $r = k$ (symetryczna tablica kontyngencji). Dla tablic niesymetrycznych ($r \neq k$) maksymalna wartość T może być mniejsza od jedności.

Współczynnik V Cramera przyjmuje wartości liczbowe z przedziału $0 \leq V \leq 1$, ale maksymalną wartość równą 1 osiąga tylko w przypadku symetrycznych tablic kontyngencji. Jeśli $r = k$, to $V = T$. Gdy natomiast $r \neq k$, to $V > T$.

W celu sprawdzenia, czy między przynależnością partyjną posłów a ich deklarowanymi rocznymi dochodami (w tys. euro) istnieje zależność, przeprowadzono badanie w grupie 300 posłów. Otrzymane wyniki przedstawiono w tab. 6.9.

Tab. 6.9. Przynależność partyjna i deklarowane dochody posłów

Partie	Roczne dochody w tys. euro				Razem
	poniżej 10	10–20	20–30	30 i więcej	
X	52	30	22	14	118
Y	16	37	47	43	143
Z	8	9	12	10	39
Razem	76	76	81	67	300

Źródło: dane umowne.

W pierwszym etapie obliczeń ustalamy wartość liczbową statystyki chi-kwadrat. W tym celu obliczamy za pomocą wzoru (6.65) liczebności teoretyczne. Są one równe:

$$\hat{n}_{11} = \frac{118 \cdot 76}{300} = 29{,}89, \quad \hat{n}_{21} = \frac{143 \cdot 76}{300} = 36{,}23, \quad \hat{n}_{31} = \frac{39 \cdot 76}{300} = 9{,}88,$$

$$\hat{n}_{12} = \frac{118 \cdot 76}{300} = 29{,}89, \quad \hat{n}_{22} = \frac{143 \cdot 76}{300} = 36{,}23, \quad \hat{n}_{32} = \frac{39 \cdot 76}{300} = 9{,}88,$$

$$\hat{n}_{13} = \frac{118 \cdot 81}{300} = 31{,}86, \quad \hat{n}_{23} = \frac{143 \cdot 81}{300} = 38{,}61, \quad \hat{n}_{33} = \frac{39 \cdot 81}{300} = 10{,}53,$$

$$\hat{n}_{14} = \frac{118 \cdot 67}{300} = 26{,}35, \quad \hat{n}_{24} = \frac{143 \cdot 67}{300} = 31{,}94, \quad \hat{n}_{34} = \frac{39 \cdot 67}{300} = 8{,}71.$$

Wartość statystyki χ^2 (wzór 6.64) jest więc równa:

$$\chi^2 = \frac{(52-29{,}89)^2}{29{,}89} + \frac{(30-29{,}89)^2}{29{,}89} + \frac{(22-31{,}86)^2}{31{,}86} + \frac{(14-26{,}35)^2}{26{,}35} +$$

$$+\frac{(16-36{,}23)^2}{36{,}23} + \frac{(37-36{,}23)^2}{36{,}23} + \frac{(47-38{,}61)^2}{38{,}61} + \frac{(43-31{,}94)^2}{31{,}94} + \frac{(8-9{,}88)^2}{9{,}88} +$$

$$+\frac{(9-9{,}88)^2}{9{,}88} + \frac{(12-10{,}53)^2}{10{,}53} + \frac{(10-8{,}71)^2}{8{,}71} = 16{,}3550 + 0{,}0004 + 3{,}0515 + 5{,}7883 +$$

$$+11{,}2960 + 0{,}0164 + 1{,}8232 + 3{,}8298 + 0{,}3577 + 0{,}0784 + 0{,}2052 + 0{,}1911 = 42{,}993.$$

Wsółczynniki zbieżności korelacyjnej wynoszą odpowiednio:

– φ Yule'a:

$$\varphi = \sqrt{\frac{42{,}993}{300}} = 0{,}379,$$

– zbieżności Czuprowa:

$$T_{xy} = +\sqrt{\frac{42{,}993}{300\sqrt{3\cdot 2}}} = 0{,}242,$$

– V Cramera:

$$V = \sqrt{\frac{42{,}993}{300\cdot 2}} = 0{,}268,$$

– C Pearsona:

$$C = \sqrt{\frac{42{,}993}{42{,}993+300}} = 0{,}354.$$

$$C_{max} = \frac{\sqrt{\frac{4-1}{4}} + \sqrt{\frac{3-1}{3}}}{2} = 0{,}8415,$$

$$C_{kor} = \frac{0{,}354}{0{,}8415} = 0{,}521.$$

Obliczone wartości współczynników wskazują na występowanie w stopniu umiarkowanym związku między przynależnością posłów do określonej partii a deklarowanymi rocznymi dochodami. Na tej podstawie można sądzić, że poglądy polityczne posłów są – w pewnym stopniu – związane z ich statusem materialnym lub też przynależność do danej partii może mieć pewien wpływ na wysokość ich dochodów.

6.4. Regresja liniowa

Termin „regresja" (od łac. *regressus*) po raz pierwszy został użyty przez F. Galtona (1822–1911), który badając wzrost mężczyzn „cofał się wstecz", sięgając do wzrostu ojców jako wyjaśnienia wzrostu synów. W dobie współczesnej przez analizę regresji rozumie się metody badania wpływu zmiennych uznanych za objaśniające (niezależne) na zmienną objaśnianą (zależną). Analitycznym narzędziem służącym do opisu mechanizmu powiązań między zmiennymi są **funkcje regresji**. Funkcje regresji wskazują, jak zmieniają się średnie wartości zmiennej zależnej przy zmianie wartości zmiennej niezależnej. Formalny opis zależności między zmiennymi X i Y, którego narzędziem jest funkcja regresji, przyjmuje postać **modelu regresji**. Budowa takiego modelu opiera się na określonych założeniach dotyczących postaci funkcji regresji. W naszych rozważaniach skoncentrujemy uwagę na **liniowych funkcjach regresji**.

W statystyce wyróżnia się **funkcje regresji I i II rodzaju**. Funkcja regresji I rodzaju dotyczy populacji generalnej i można ją wyznaczyć tylko wówczas, gdy dysponujemy wynikami badania pełnego. Funkcja regresji II rodzaju odnosi się do próby i jest **aproksymantą** (przybliżeniem) funkcji regresji I rodzaju. W funkcji regresji I rodzaju wartościom jednej cechy przyporządkowuje się średnie warunkowe wartości drugiej cechy.

W dwuwymiarowym rozkładzie zmiennej (X,Y) badaniu można poddać zarówno regresję Y względem X, jak też regresję X względem Y. W pierwszym przypadku zmienna X jest zmienną objaśniającą (niezależną), a zmienna Y – zmienną zależną (objaśnianą). W takiej sytuacji obserwowaną zmienność Y chcemy wyjaśnić poprzez związki tej zmiennej ze zmienną X. Jeśli odwrócimy role zmiennych, to obserwowaną zmienność X chcemy wyjaśnić wpływem zmiennej Y. Badamy wówczas regresję X względem Y. Zmienna Y jest wówczas zmienną objaśniającą, a X zmienną objaśnianą.

Warunkowa wartość oczekiwana zmiennej Y względem X jest jakąś funkcją zmiennej X, zaś warunkowa wartość oczekiwana X względem Y jest funkcją zmiennej Y. Możemy zatem zapisać:

$$E(Y|X=x) = g_1(x) \text{ lub } y = g_1(x), \tag{6.73}$$

$$E(X|Y=y) = g_2(y) \text{ lub } x = g_2(y). \tag{6.74}$$

Jeżeli zmienna losowa (X,Y) jest skokowa, to obrazami geometrycznymi funkcji (6.73) oraz (6.74) są pewne łamane przedstawione w prostokątnym układzie współrzędnych. Łamane te powstają przez połączenie punktów o współrzędnych: $(x_i; \overline{y}|x_i)$ dla $i = 1,2,\ldots,k$ (w przypadku funkcji regresji Y względem X) oraz $(\overline{x}|y_j)$ dla $j = 1,2,\ldots,r$ (w przypadku funkcji regresji X względem Y).

Jeżeli zmienna (X,Y) jest ciągła, to obrazami geometrycznymi funkcji (6.73) i (6.74) są pewne linie, które nazywamy liniami regresji I rodzaju.

Jak już wcześniej stwierdzono, przybliżenie za pomocą określonej funkcji matematycznej związku występującego pomiędzy cechami statystycznymi nosi nazwę regresji II rodzaju. Rozbieżność między poszczególnymi wartościami cechy uznanej za zależną i jej regresją I rodzaju wyrażana jest za pomocą składnika losowego:

$$Y = \hat{Y} + \varepsilon, \tag{6.75}$$

gdzie: Y – zmienna (cecha) zależna, opisywana przez model regresji, $\hat{Y}$ – regresja I rodzaju, która może przyjąć postać reguły matematycznej (określonej funkcji f) lub też być przez tę regułę przybliżona (w tym ostatnim przypadku regułą matematyczną będzie regresją II rodzaju), ε – zmienna losowa, opisująca odchylenia wartości zmiennej zależnej od jej regresji I rodzaju.

W równaniu (6.75) wielkość $\hat{Y}$ można zastąpić konkretną funkcją matematyczną. Jeśli będzie to funkcja liniowa, to:

$$Y = \alpha_0 + \alpha_1 X + \varepsilon_1. \tag{6.76}$$

W podobny sposób można zapisać zależność cechy X od Y:

$$X = \beta_0 + \beta_1 Y + \varepsilon_2. \tag{6.77}$$

W równaniu (6.76) przyjmuje się, że zmienna niezależna X jest nielosowa i determinuje zmienną Y z dokładnością do składnika losowego ε_1. Podobnie w równaniu (6.77) zakłada się, że zmienna Y jest nielosowa i determinuje zmienną X z dokładnością do składnika losowego ε_2.

Równania regresji służą do przewidywania wartości cech statystycznych uznanych za zależne. Rząd dokładności tego przewidywania jest z jednej strony uwarunkowany tym, jak „blisko" regresji I rodzaju znajduje się regresja II rodzaju, z drugiej zaś zależy od tego, w jakim stopniu składnik losowy ε wyraża oddziaływanie czynników przypadkowych, zakłócających „układanie się" wartości cechy zależnej według równania linii prostej. W związku z tym żąda się, aby składnik ε miał charakter losowy, tzn. by spełniał następujące własności:

1) brak systematycznych oddziaływań ε na zmienna zależną Y $(E(\varepsilon)=0)$;

2) zakres zmienności ε jest niezależny od zmiennej niezależnej $(D^2(\varepsilon)=\sigma^2$, gdzie σ^2 jest liczbą stałą);

3) czynniki kształtujące ε nie są ze sobą powiązane $(\mathrm{cov}(\varepsilon_i,\varepsilon_j)=0$ dla $i\neq j)$.

Jeśli powyższe własności są spełnione, to równania (6.76) i (6.77) określają model regresji liniowej. Jeśli, dodatkowo, składnik losowy ε jest zmienną losową o rozkładzie normalnym z parametrami $(0, D^2(x))$ – to taki model nazywamy **liniowym modelem regresji normalnej**.

W praktyce, postać funkcji matematycznej opisującej związek pomiędzy cechami statystycznymi nie jest znana. W związku z tym z populacji dwuwymiarowej pobiera się próbę losową i sporządza się – w prostokątnym układzie współrzędnych – wykres rozrzutu. Wykres ten stanowi podstawę do wyboru analitycznej postaci funkcji (np. liniowej, wykładniczej, potęgowej itp.). Kolejnym krokiem jest **estymacja parametrów strukturalnych** (występujących przy zmiennych objaśniających) oraz parametrów **struktury stochastycznej** (związanych ze składnikiem losowym).

6.4.1. Estymacja parametrów strukturalnych liniowej funkcji regresji

Do szacowania parametrów strukturalnych linowych funkcji regresji, z jedną zmienną objaśniającą (6.76) i (6.77), wykorzystuje się m.in. **klasyczną metodę najmniejszych kwadratów** (KMNK).

Metoda najmniejszych kwadratów polega na takim oszacowaniu parametrów α_0 i α_1 (bądź β_0 i β_1) funkcji regresji, by dla n wartości (x_i, y_i) danych z próby, funkcje:

$$W=\sum_{i=1}^{n}(y_i-a_0-a_1x_i)^2 \text{ lub } W=\sum_{i=1}^{n}(x_i-b_0-b_1y_i)^2 \tag{6.78}$$

osiągnęły minimum.

Funkcje:

$$\hat{y}_i=a_0+a_1x_i+u_i \text{ oraz } \hat{x}_i=b_0+b_1y_i+z_i \tag{6.79}$$

są aproksymantami odpowiednich liniowych funkcji regresji w próbie losowej. Symbolami $u_i=y_i-\hat{y}_i$ oraz $z_i=x_i-\hat{x}_i$ oznaczono – odpowiednio – składniki resztowe (dla $i=1,2,\ldots,n$).

Wyrażenie (6.78) jest funkcją dwóch zmiennych: a_0 i a_1 lub b_0 i b_1. Warunkiem koniecznym istnienia ekstremum jest zerowanie się pochodnych cząstkowych. Ograniczając rozważania do funkcji regresji y względem x, mamy:

$$\frac{\partial W}{\partial a_0}=-2\sum_{i=1}^{n}(y_i-a_0-a_1x_i), \tag{6.80}$$

$$\frac{\partial W}{\partial a_1}=-2\sum_{i=1}^{n}(y_i-a_0-a_1x_i)x_i. \tag{6.81}$$

Po przyrównaniu pochodnych cząstkowych (6.80) i (6.81) do zera oraz po przeprowadzeniu odpowiednich przekształceń otrzymujemy układ równań normalnych o postaci:

$$\begin{cases} na_0+a_1\sum_{i=1}^{n}x_i=\sum_{i=1}^{n}y_i \\ a_0\sum_{i=1}^{n}x_i+a_1\sum_{i=1}^{n}x_i^2=\sum_{i=1}^{n}x_iy_i \end{cases}. \tag{6.82}$$

Układ (6.82) ma następujące rozwiązanie:

$$a_1=\frac{\sum_{i=1}^{n}(x_i-\bar{x})(y_i-\bar{y})}{\sum_{i=1}^{n}(x_i-\bar{x})^2} \tag{6.83}$$

oraz

$$a_0=\bar{y}-a_1\bar{x}, \tag{6.84}$$

gdzie $\bar{x}$, $\bar{y}$ – to średnie arytmetyczne odpowiednich zmiennych x i y.

Postępując analogicznie w przypadku liniowej funkcji regresji x względem y otrzymujemy następujący układ równań:

$$\begin{aligned} nb_0+b_1\sum_{i=1}^{n}y_i&=\sum_{i=1}^{n}x_i, \\ b_0\sum_{i=1}^{n}y_i+b_1\sum_{i=1}^{n}y_i^2&=\sum_{i=1}^{n}x_iy_i. \end{aligned} \tag{6.85}$$

Rozwiązaniem układu (6.85) jest:

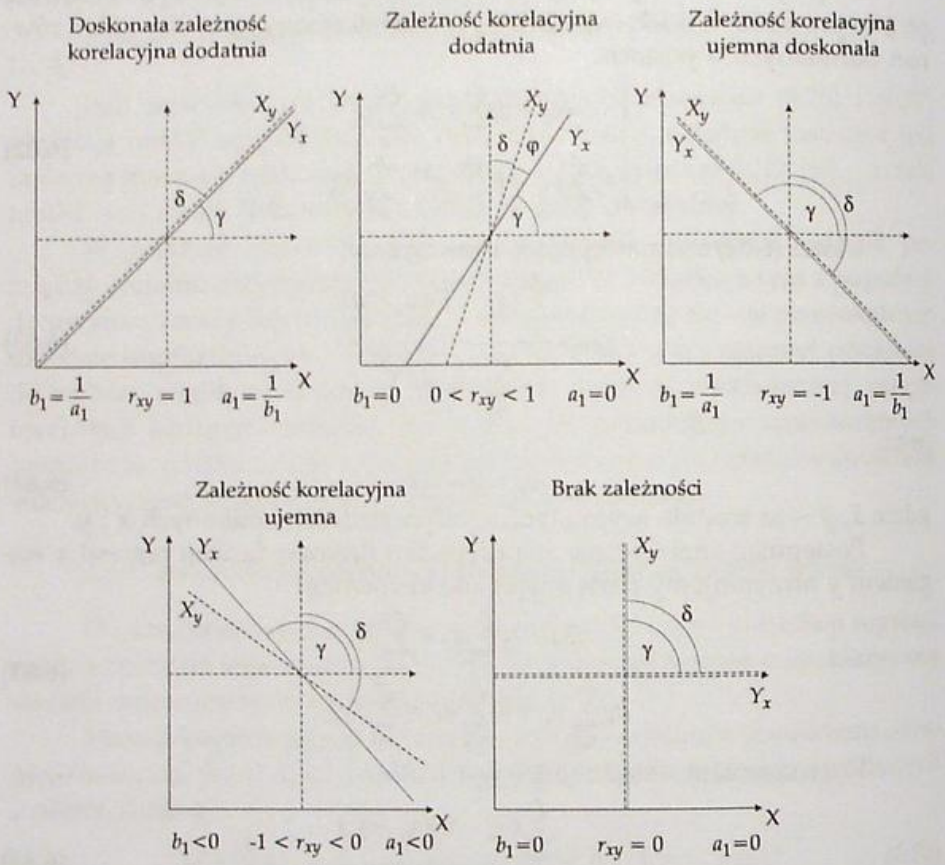
$$b_1=\frac{\sum_{i=1}^{n}(x_i-\bar{x})(y_i-\bar{y})}{\sum_{i=1}^{n}(y_i-\bar{y})^2} \tag{6.86}$$

oraz

$$b_0=\bar{x}-b_1\bar{y}. \tag{6.87}$$

Oceny a_1 i b_1 określane są mianem **współczynników regresji liniowej**. Odpowiadają one na pytanie, jak zmieni się średnia wartość zmiennej zależnej (objaśnianej), jeśli zmienna objaśniająca (niezależna) wzrośnie o jednostkę. Oceny a_1 i b_0 są wyrazami wolnymi odpowiednich funkcji regresji. Są to wartości, jakie przyjmuje zmienna zależna, gdy zmienna niezależna jest równa zeru. W zależności od badanego zjawiska, oceny te mogą mieć interpretację ekonomiczną lub nie. Na przykład w przypadku liniowej funkcji regresji kosztów całkowitych (zmienna zależna) względem wielkości produkcji (zmienna niezależna) wyraz wolny informuje o poziomie kosztów stałych.

Graficznym obrazem funkcji regresji II rodzaju jest rys. 6.2.



Rys. 6.2. Możliwe położenia linii regresji
Źródło: M. Sobczyk, *Statystyka. Podstawy teoretyczne, przykłady i zadania*, UMCS, Lublin 2000, s. 272.

Z rys. 6.2 wynika, że kąt γ jest kątem nachylenia prostej regresji Y względem X, zaś $\operatorname{tg}\gamma = a_1$. Kąt δ jest natomiast katem nachylenia prostej regresji X względem Y, a $\operatorname{tg}\delta = b_1$. Kąty te tworzą linie regresji z dodatnimi kierunkami osi X i Y. Im bliżej siebie położone są obie linie regresji (im mniejszy kąt φ), tym zależność korelacyjna jest silniejsza. W miarę wzrostu siły współzależności korelacyjnej, kąty γ i δ zdążają w sumie do kąta prostego, a iloczyn $\operatorname{tg}\gamma$ i $\operatorname{tg}\delta$ zmierza do jedności. Wynika stąd, że współczynniki regresji są miarami siły współzależności między badanymi zmiennymi. Siłę tę określa średnia geometryczna obliczona ze współczynników linii regresji:

$$r_{xy} = \sqrt{a_1 \cdot b_1}. \tag{6.88}$$

Współczynnik korelacji liniowej określony wzorem (6.88) przyjmuje taki znak, jaki mają współczynniki regresji. Współczynniki regresji mają zawsze takie same znaki (albo obydwa są dodatnie, albo ujemne).

Współczynniki regresji a_1 oraz b_1 można również obliczyć z równoważnych wzorów, a mianowicie:

$$a_1 = r_{xy}\frac{s(y)}{s(x)} \tag{6.89}$$

oraz

$$b_1 = r_{xy}\frac{s(x)}{s(y)}, \tag{6.90}$$

gdzie $s(y)$ i $s(x)$ są odchyleniami standardowymi badanych zmiennych.

Załóżmy, że w modelu regresji liniowej wysokości miesięcznych opłat za gaz (y) względem liczby osób w gospodarstwie domowym (x) dla niezależnej próby losowej 100 gospodarstw otrzymano następujące wyniki (w zł): $\operatorname{cov}(x,y) = 11{,}4$; $\bar{x} = 4$; $s(x) = 0{,}9$; $\bar{y} = 78{,}5$; $s(y) = 17{,}2$. Jakiej wysokości opłat należy oczekiwać w trzyosobowych gospodarstwach domowych?

Odpowiedź na postawione pytanie wymaga znajomości funkcji regresji opisującej mechanizm powiązań zmiennej y względem cechy x, tzn. $\hat{y}_i = a_0 + a_1 x_i$.

Oceny parametrów strukturalnych tej funkcji wyznaczymy ze wzorów (6.89) oraz (6.84). Przedtem jednak należy obliczyć r_{xy}:

$$r_{xy} = \frac{\operatorname{cov}(x,y)}{s(x)s(y)} = \frac{11{,}4}{0{,}9 \cdot 17{,}2} = 0{,}736,$$

$$a_1 = 0{,}736 \cdot \frac{17{,}2}{0{,}9} = 14{,}066,$$

$$a_0 = 78{,}5 - 14{,}066 \cdot 4 = 22{,}236.$$

Równanie regresji liniowej zmiennej y względem x przyjmuje zatem postać: $\hat{y}_i = 22{,}236 + 14{,}066 x_i$. Miesięczne opłaty za gaz w rodzinach trzyosobowych są więc równe: $\hat{y}_{x=3} = 22{,}236 + 14{,}066 \cdot 3 = 64{,}43$ zł.

6.4.2. Weryfikacja oszacowanej funkcji regresji liniowej

Oszacowane na podstawie wyników próby losowej funkcje regresji są jedynie aproksymantami funkcji w populacji generalnej. Dlatego też zachodzi konieczność oceny dopasowania funkcji regresji II rodzaju do danych empirycznych. Podstawą do tej oceny są **reszty**, czyli różnice między empirycznymi i teoretycznymi (wynikającymi z funkcji regresji) wartościami zmiennej zależnej. Dla funkcji regresji Y względem X reszty te są określone następująco:

$$u_i = y_i - \hat{y}_i \ (i = 1,2,\ldots,n), \tag{6.91}$$

dla funkcji regresji X względem Y przyjmują postać:

$$z_i = x_i - \hat{x}_i \ (i = 1,2,\ldots,n), \tag{6.92}$$

gdzie: y_i oraz x_i są wartościami empirycznymi, natomiast $\hat{y}_i$ oraz $\hat{x}_i$ – wartościami teoretycznymi.

Syntetycznym miernikiem odchyleń wartości empirycznych od teoretycznych jest **wariancja resztowa**, wyrażająca się wzorem:

$$s^2(u) = \frac{\sum_{i-1}^{n}(y_i - \hat{y}_i)^2}{n-k} = \frac{\sum_{i-1}^{n} u_i^2}{n-k}, \tag{6.93}$$

gdzie k jest liczbą szacowanych parametrów funkcji regresji.

Wzór (6.93) określa wariancję resztową dla funkcji regresji Y względem X. Dla funkcji regresji X względem Y, wariancję resztową oblicza się następująco:

$$s^2(z) = \frac{\sum_{i-1}^{n}(x_i - \hat{x}_i)^2}{n-k} = \frac{\sum_{i-1}^{n} z_i^2}{n-k}. \tag{6.94}$$

Pierwiastek kwadratowy z wariancji resztowej nosi nazwę **odchylenia standardowego składnika resztowego**. Odchylenie to informuje o przeciętnym odchyleniu empirycznych wartości zmiennej objaśnianej od jej wartości teoretycznych. Dopasowanie danej funkcji regresji do danych empirycznych jest tym lepsze, im odchylenie standardowe składnika resztowego jest mniejsze (bliskie zera).

Odchylenie standardowe składnika resztowego stanowi podstawę do budowy **współczynnika zmienności resztowej**:

$$V_u = \frac{s(u)}{\bar{y}} \cdot 100 \tag{6.95}$$

oraz

$$V_z = \frac{s(z)}{\bar{x}} \cdot 100. \tag{6.96}$$

Współczynniki zmienności resztowej są wielkościami niemianowanymi (wyrażanymi w procentach) i informują o tym, jaką częścią średniej

arytmetycznej zmiennej objaśnianej są odpowiednie odchylenia standardowe składników resztowych. Współczynniki te powinny być możliwie bliskie 0.

Dokładność oszacowania funkcji regresji liniowej można również ocenić za pomocą **współczynnika zbieżności** (nieokreśloności). Współczynniki te są obliczane ze wzorów:

$$\varphi_{yx}^2 = \frac{s^2(u)}{s^2(y)} = \frac{\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}{\sum_{i=1}^{n}(y_i - \bar{y}_i)^2} \tag{6.97}$$

oraz

$$\varphi_{xy}^2 = \frac{s^2(z)}{s^2(x)} = \frac{\sum_{i=1}^{n}(x_i - \hat{x}_i)^2}{\sum_{i=1}^{n}(x_i - \bar{x}_i)^2}. \tag{6.98}$$

Wzór (6.97) dotyczy funkcji regresji Y względem X, a (6.98) odnosi się do funkcji regresji X względem Y.

Współczynnik zbieżności określa, jaką częścią całkowitej zmienności zmiennej zależnej jest zmienność zmiennej zależnej **niewyjaśniona** przez zmienną objaśniającą danej funkcji regresji. Współczynnik zbieżności jest miarą unormowaną zawartą w przedziale $\langle 0,1\rangle$. Im wartość φ^2 jest bliższa 0, tym oszacowana funkcja regresji jest lepiej dopasowana do danych empirycznych.

Ze współczynnikiem zbieżności ściśle łączy się **współczynnik determinacji** (określoności), będący kwadratem współczynnika korelacji wielorakiej obliczanym następująco:

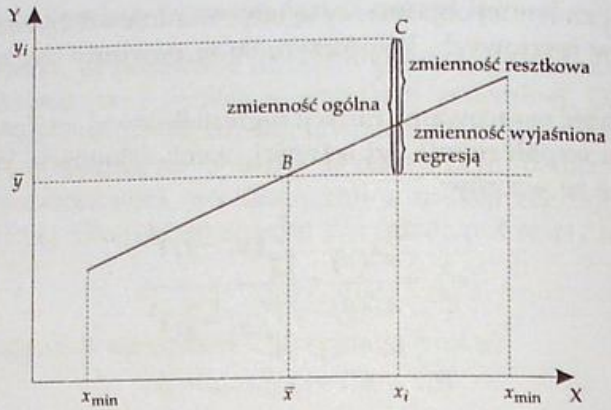
– dla funkcji regresji Y względem X:

$$R_{yx}^2 = 1 - \frac{\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}{\sum_{i=1}^{n}(y_i - \bar{y}_i)^2} = 1 - \varphi_{yx}^2, \tag{6.98}$$

– dla funkcji regresji X względem Y:

$$R_{xy}^2 = 1 - \frac{\sum_{i=1}^{n}(x_i - \hat{x}_i)^2}{\sum_{i=1}^{n}(x_i - \bar{x}_i)^2} = 1 - \varphi_{xy}^2. \tag{6.100}$$

Wartość R^2 zawiera się w przedziale $\langle 0,1\rangle$. Współczynnik R^2 informuje o tym, jaka część całkowitej zmienności zmiennej objaśnianej została wyjaśniona przez zmienną objaśniającą danej funkcji (czyli przez oszacowaną



Rys. 6.3. Geometryczna interpretacja zmienności ogólnej, zmienności wyjaśnionej oraz niewyjaśnionej liniową regresją Y względem X
Źródło: M. Sobczyk, *Statystyka. Podstawy teoretyczne, przykłady i zadania*, UMCS, Lublin 2000, s. 277.

funkcję regresji). Tak więc, im R^2 jest bliższe 1, tym jakość dopasowania funkcji regresji do danych empirycznych jest lepsza.

Jak łatwo zauważyć, całkowity obszar zmienności zmiennej zależnej jest sumą zmienności wyjaśnionej i niewyjaśnionej przez przyjętą funkcje regresji (por. rys. 6.3):

$$\varphi^2 + R^2 = 1. \tag{6.101}$$

Estymacji parametrów funkcji regresji dokonuje się na podstawie wyników próby losowej. Stąd też możliwe jest popełnianie błędów w ocenie poszczególnych parametrów funkcji. Błędy te są określane mianem **standardowych błędów szacunku parametrów** lub krócej: **błędów średnich szacunku**. Błędy te są obliczane na podstawie następujących wzorów:

$$s(a_0) = s(u)\sqrt{\frac{\sum_{i=1}^{n} x_i^2}{n\sum_{i=1}^{n}(x_i - \bar{x})^2}}, \tag{6.102}$$

$$s(a_1) = \frac{s(u)}{\sqrt{\sum_{i=1}^{n}(x_i - \bar{x})^2}}, \tag{6.103}$$

$$s(b_0) = s(z)\sqrt{\frac{\sum_{i=1}^{n} y_i^2}{n\sum_{i=1}^{n}(y_i - \bar{y})^2}}, \tag{6.104}$$

$$s(b_1) = \frac{s(z)}{\sum_{i=1}^{n}(y_i - \bar{y})^2}. \tag{6.105}$$

Błędy średnie szacunku informują o tym, o ile – przeciętnie biorąc – mylimy się (in plus lub in minus), szacując parametry $\alpha_0, \alpha_1, \beta_0, \beta_1$ w populacji generalnej na podstawie wyników otrzymanych z próby losowej. Ilorazy:

$$\frac{s(a_0)}{a_0}; \frac{s(a_1)}{a_1}; \frac{s(b_0)}{b_0}; \frac{s(b_1)}{b_1} \tag{6.106}$$

są względnymi średnimi błędami szacunku odpowiednich parametrów. Zbyt duże błędy (np. przekraczające 50%) przekreślają wartość poznawczą liczbowej oceny parametrów.

Przyczynami powodującymi otrzymywanie dużych błędów szacunku parametrów mogą być:

1) mała liczebność próby losowej wykorzystywanej do estymacji parametrów funkcji regresji,

2) niewłaściwa metoda estymacji parametrów funkcji regresji,

3) przyjęcie nieodpowiedniej zmiennej objaśniającej do funkcji regresji.

Podjęto badania dotyczące zależności między cenami akcji dwóch spółek (w zł). W tym celu zebrano informacje liczbowe o cenach akcji (tab. 6.10).

Tab. 6.10. Ceny akcji spółek A i B

Ceny akcji spółki A	62	62	59	56	57	56	55	55	55	56
Ceny akcji spółki B	83	94	90	86	86	85	82	80	79	80

Źródło: dane umowne.

Na podstawie danych zawartych w tab. 6.10 należy oszacować funkcje regresji oraz dokonać ich weryfikacji.

Do oceny parametrów strukturalnych funkcji regresji Y względem X wykorzystamy układ równań (6.82), do wyznaczenia ocen parametrów funkcji X względem Y – układ równań (6.83). Niezbędne obliczenia zawiera tab. 6.11.

Podstawiając dane liczbowe z tab. 6.11 do układu równań (6.82) otrzymujemy:

$$\begin{cases} 10a_0 + 573a_1 = 845 \\ 573a_0 + 32\,901a_1 = 48\,497 \end{cases}.$$

Rozwiązując powyższy układ mamy: $a_0 = 18{,}61$ zł; $a_1 = 1{,}15$ zł. Równanie regresji Y względem X przyjmuje zatem następującą postać: $\hat{y}_i = 18{,}61 + 1{,}15x_i$. Oznacza to, że jeśli ceny akcji spółki A wzrosną o 1 zł, to ceny spółki B wzrosną – średnio rzecz biorąc – o 1,15 zł.

Podstawiając dane liczbowe z tab. 6.11. do układu równań (6.85) mamy:

$$\begin{cases} 10b_0 + 845b_1 = 573 \\ 845b_0 + 71\,607b_1 = 48\,497 \end{cases}$$

Z układu tego otrzymujemy: $b_0 = 25{,}19$ oraz $b_1 = 0{,}38$. Równanie regresji X względem Y przyjmuje więc postać: $\hat{x}_i = 25{,}19 + 0{,}38y_i$. Jeśli akcje spółki B wzrosną o 1 zł, to, średnio, akcje spółki A wzrosną o 0,38 zł.

Tab. 6.11. Obliczenia pomocnicze do wyznaczenia ocen parametrów funkcji regresji

Cena akcji spółki A (x_i)	Cena akcji spółki B (y_i)	x_i^2	y_i^2	$x_i y_i$
62	83	3844	6889	5146
62	94	3844	8836	5828
59	90	3481	8100	5310
56	86	3136	7396	4816
57	86	3249	7396	4902
56	85	3136	7225	4760
55	82	3025	6724	4510
55	80	3025	6400	4400
55	79	3025	6241	4345
56	80	3136	6400	4480
573	**845**	**32 901**	**71 607**	**48 497**

Źródło: obliczenia własne.

Współczynnik korelacji liniowej między badanymi cechami jest natomiast równy (wzór 6.88):

$$r_{xy} = \sqrt{1{,}15 \cdot 0{,}38} = 0{,}66.$$

Otrzymany wynik wskazuje na umiarkowaną zależność korelacyjną dodatnią pomiędzy cenami akcji spółek A i B.

W celu zbadania, czy otrzymane funkcje regresji są dobrze dopasowane do danych empirycznych, należy wykorzystać następujące mierniki: współczynnik zmienności resztowej, współczynnik determinacji i indeterminacji oraz średnie błędy szacunku parametrów strukturalnych. Niezbędne obliczenia zawiera tab. 6.12.

Teoretyczne wartości zmiennych $\hat{x}_i$ oraz $\hat{y}_i$ otrzymano wstawiając do oszacowanych funkcji regresji odpowiednie wartości zmiennych niezależnych, a mianowicie:

$\hat{x}_1 = 25{,}19 + 0{,}38 \cdot 83 = 56{,}73$ $\quad \hat{y}_1 = 18{,}61 + 1{,}15 \cdot 62 = 89{,}91$
$\hat{x}_2 = 25{,}19 + 0{,}38 \cdot 94 = 60{,}91$ $\quad \hat{y}_2 = 18{,}61 + 1{,}15 \cdot 62 = 89{,}91$
$\hat{x}_3 = 25{,}19 + 0{,}38 \cdot 90 = 59{,}39$ $\quad \hat{y}_3 = 18{,}61 + 1{,}15 \cdot 59 = 86{,}46$
$\hat{x}_4 = 25{,}19 + 0{,}38 \cdot 86 = 57{,}87$ $\quad \hat{y}_4 = 18{,}61 + 1{,}15 \cdot 56 = 83{,}01$
$\hat{x}_5 = 25{,}19 + 0{,}38 \cdot 86 = 57{,}87$ $\quad \hat{y}_5 = 18{,}61 + 1{,}15 \cdot 57 = 84{,}16$
$\hat{x}_6 = 25{,}19 + 0{,}38 \cdot 85 = 57{,}49$ $\quad \hat{y}_6 = 18{,}61 + 1{,}15 \cdot 56 = 83{,}01$

$\hat{x}_7 = 25{,}19 + 0{,}38 \cdot 82 = 56{,}35$ $\quad \hat{y}_7 = 18{,}61 + 1{,}15 \cdot 55 = 81{,}86$
$\hat{x}_8 = 25{,}19 + 0{,}38 \cdot 80 = 55{,}59$ $\quad \hat{y}_8 = 18{,}61 + 1{,}15 \cdot 55 = 81{,}86$
$\hat{x}_9 = 25{,}19 + 0{,}38 \cdot 79 = 55{,}21$ $\quad \hat{y}_9 = 18{,}61 + 1{,}15 \cdot 55 = 81{,}86$
$\hat{x}_{10} = 25{,}19 + 0{,}38 \cdot 80 = 55{,}21$ $\quad \hat{y}_{10} = 18{,}61 + 1{,}15 \cdot 56 = 83{,}01$

Tab. 6.12. Obliczenia pomocnicze do weryfikacji funkcji regresji

x_i	$\hat{x}_i$	z_i	z_i^2	$x_i - \bar{x}$	$(x_i - \bar{x})^2$	y_i	$\hat{y}_i$	u_i	u_i^2	$y_i - \bar{y}$	$(y_i - \bar{y})^2$
62	56,73	5,27	27,7729	4,7	22,09	83	89,91	–6,91	47,7481	–1,5	2,25
62	60,91	1,09	1,1881	4,7	22,09	94	89,91	4,09	16,7281	9,5	90,25
59	59,39	–0,39	0,1521	1,7	2,89	90	86,46	3,54	12,5316	5,5	30,25
56	57,87	–1,87	3,4969	–1,3	1,69	86	83,01	2,99	8,9401	1,5	2,25
57	57,87	–0,87	0,7569	–0,3	0,09	86	84,16	1,84	3,3856	1,5	2,25
56	57,49	–1,49	2,2201	–1,3	1,69	85	83,01	1,99	3,9601	0,5	0,25
55	56,35	–1,35	1,8225	–2,3	5,29	82	81,86	0,14	0,0196	–2,5	6,25
55	55,59	–0,59	0,3481	–2,3	5,29	80	81,86	–1,86	3,4596	–4,5	20,25
55	55,21	–0,21	0,0441	–2,3	5,29	79	81,86	–2,86	8,1796	–5,5	30,25
56	55,59	0,41	0,1681	–1,3	1,69	80	83,01	–3,01	9,0601	–4,5	20,25
573	X	X	**37,9698**	X	**68,10**	**845**	X	X	**114,0125**	X	**204,50**

Źródło: obliczenia własne.

Wyniki obliczeń mierników charakteryzujących stopień dopasowania funkcji: $\hat{x}_i = 25{,}19 + 0{,}38y_i$ oraz $\hat{y}_i = 18{,}61 + 1{,}15x_i$ do danych empirycznych przedstawiono w tab. 6.13.

Tab. 6.13. Mierniki dopasowania funkcji regresji do danych empirycznych

Nazwa miernika	Funkcje regresji	
	$\hat{y}_i = 18{,}61 + 1{,}15x_i$	$\hat{x}_i = 25{,}19 + 0{,}38y_i$
Odchylenie standardowe składnika resztowego	3,78 zł	2,18 zł
Współczynnik zmienności resztowej	4,47%	3,80%
Współczynnik zbieżności	0,5575	0,5576
Współczynnik determinacji	0,4425	0,4426
Średni błąd szacunku wyrazu wolnego	26,271	12,899
Średni błąd szacunku współczynnika regresji	0,4581	0,1524

Źródło: obliczenia własne.

Otrzymane wyniki wskazują na niezbyt dobrą zgodność empirycznych wartości zmiennych objaśnianych z oszacowanymi równaniami regresji liniowej.

Zwrócić należy uwagę na to, że graficznym obrazem obu funkcji regresji są linie proste, przecinające się w punkcie o współrzędnych $(\bar{x}, \bar{y})$.

6.5. Wnioskowanie statystyczne w analizie korelacji i regresji

Jaki wiadomo, badanie współzależności ma sens jedynie wtedy, gdy pomiędzy zmiennymi występują powiązania typu stochastycznego lub co najmniej korelacyjnego. Zmienne losowe są stochastycznie **niezależne**, jeżeli rozkłady warunkowe każdej ze zmiennych są identyczne i takie jak rozkład brzegowy, bez względu na to, jakie wartości przyjmuje druga zmienna.

Do weryfikacji hipotezy o stochastycznej niezależności zmiennych wykorzystuje się **statystykę** χ^2. W tym przypadku hipotezy: zerową i alternatywną, stawia się w następujący sposób:

H_0: cechy X i Y są niezależne,

H_1: cechy X i Y nie są niezależne.

Są to więc hipotezy nieparametryczne.

W celu sprawdzenia wysuniętej H_0, z dwuwymiarowej populacji generalnej (X,Y) pobiera się n-elementową próbę losową, na podstawie której buduje się tablicę korelacyjną o wymiarach $r \times k$, przedstawiającą dwuwymiarowy rozkład obu cech. Liczebności wewnątrz tablicy – oznaczane jako $n_{ij}(i=1,2,\ldots,k; j=1,2,\ldots,r)$ – określają dwuwymiarowy rozkład empiryczny cech X i Y.

Do weryfikacji H_0 o stochastycznej niezależności cech wykorzystuje się statystykę chi-kwadrat określoną następująco:

$$\chi^2 = \sum_{i=1}^{k}\sum_{j=1}^{r}\frac{\left(n_{ij}-\hat{n}_{ij}\right)^2}{\hat{n}_{ij}}, \qquad (6.107)$$

gdzie symbolem $\hat{n}_{ij}$ oznaczono **liczebności teoretyczne**, czyli takie, jakie powinny występować, by cechy uznać za niezależne. Liczebności te otrzymujemy mnożąc sumę wiersza $(n_{i\cdot})$ przez sumę kolumny $(n_{\cdot j})$ i dzieląc ten iloczyn przez liczebność próby (n).

Statystyka (6.107) jest zatem miarą zbieżności pomiędzy rozkładem empirycznym a teoretycznym. Im większa jest obliczona z próby wartość χ^2, tym mniej wiarygodna jest hipoteza o niezależności cech.

Statystyka (6.107) ma – dla dużych prób i przy założeniu prawdziwości H_0 – asymptotyczny rozkład χ^2 z $(k-1)(r-1)$ stopniami swobody. Obszar krytyczny testu jest **prawostronny**, gdyż wysokie wartości χ^2, wywołane dużymi różnicami między liczebnościami empirycznymi i teoretycznymi, powodują odrzucenie H_0, zakładającej niezależność zmiennych. Hipotezę zerową odrzucamy, jeśli $\chi^2 \geq \chi^2_\alpha$. Pozwala to przypuszczać, że między zmiennymi istnieje zależność stochastyczna. W przypadku, gdy $\chi^2 < \chi^2_\alpha$, brakuje podstaw do odrzucenia H_0.

Stosowanie testu niezależności χ^2 wymaga, aby liczebności w polach tablicy nie były zbyt małe. Zaleca się, by w każdej kratce liczebność teoretyczna wynosiła co najmniej 5 $(\hat{n}_{ij} \geq 5)$. Jeśli – po obliczeniu liczebności teoretycznych $\hat{n}_{ij}$ – stwierdzimy, że warunek ten nie jest spełniony, to wiersz

(lub kolumnę), w którym występuje zbyt mała liczebność, należy połączyć z sąsiednim wierszem (kolumną). Dopiero wtedy obliczamy wartość statystyki χ^2 niezbędną do podjęcia decyzji weryfikacyjnej.

Zwrócić należy uwagę na fakt, że tablice rozkładu χ^2 podają zazwyczaj wartości krytyczne dla liczby stopni swobody od 1 do 30. Jeśli liczba stopni swobody przekracza 30, to w celu weryfikacji H_0 o stochastycznej niezależności zmiennych losowych X i Y wykorzystujemy test z o postaci:

$$z = \sqrt{2\chi^2} - \sqrt{2(k-1)(r-1)-1}. \qquad (6.108)$$

Statystyka (6.108) ma rozkład $N(0,1)$. Wartość krytyczną z_α odczytujemy z tablicy dystrybuanty rozkładu normalnego dla dwustronnego obszaru krytycznego.

W 600-osobowej losowo dobranej grupie osób przeprowadzono badanie ankietowe dotyczące zależności między wykształceniem telewidzów a rodzajem programu telewizyjnego, który najchętniej oglądają. Otrzymane wyniki przedstawiono w tab. 6.14.

Tab. 6.14. Zależność między wykształceniem a preferowanym programem telewizyjnym

Wykształcenie (X)	Rodzaj programu (Y)				Razem
	film	teatr	programy rozrywkowe	programy publicystyczne	
Podstawowe	105	10	75	10	200
Średnie	120	60	80	40	300
Wyższe	35	30	15	20	100
Razem	260	100	170	70	600

Źródło: dane umowne.

Naszym zadaniem jest weryfikacja hipotezy o stochastycznej niezależności zmiennych X i Y (na poziomie istotności 0,05).

Liczebności teoretyczne $\hat{n}_{ij}$ w każdej kratce pola tablicy znajdujemy następująco:

$$\hat{n}_{11} = \frac{200\cdot 260}{600} = 86{,}7 \qquad \hat{n}_{21} = \frac{300\cdot 260}{600} = 130{,}0 \qquad \hat{n}_{31} = \frac{100\cdot 260}{600} = 43{,}3$$

$$\hat{n}_{12} = \frac{200\cdot 100}{600} = 33{,}3 \qquad \hat{n}_{22} = \frac{300\cdot 100}{600} = 50{,}0 \qquad \hat{n}_{32} = \frac{100\cdot 100}{600} = 16{,}7$$

$$\hat{n}_{13} = \frac{200\cdot 170}{600} = 57{,}6 \qquad \hat{n}_{23} = \frac{300\cdot 170}{600} = 85{,}0 \qquad \hat{n}_{33} = \frac{100\cdot 170}{600} = 28{,}3$$

$$\hat{n}_{14} = \frac{200\cdot 70}{600} = 23{,}3 \qquad \hat{n}_{24} = \frac{300\cdot 70}{600} = 35{,}0 \qquad \hat{n}_{34} = \frac{100\cdot 70}{600} = 11{,}7$$

Wykorzystując wzór (6.107) mamy:

$$\chi^2 = \frac{(105-86{,}7)^2}{86{,}7} + \frac{(10-33{,}3)^2}{33{,}3} + \frac{(75-56{,}7)^2}{56{,}7} + \frac{(10-23{,}3)^2}{23{,}3} + \frac{(120-130)^2}{130} + \frac{(60-50)^2}{50} + \frac{(80-85)^2}{85} + \frac{(40-35)^2}{35} + \frac{(35-43{,}3)^2}{43{,}3} + \frac{(30-16{,}7)^2}{16{,}7} + \frac{(15-28{,}3)^2}{28{,}3} + \frac{(20-11{,}7)^2}{11{,}7} = 61{,}7623.$$

Z tablic rozkładu χ^2 dla poziomu istotności $\alpha = 0{,}05$ i $(4-1)(3-1) = 6$ stopni swobody odczytujemy wartość $\chi^2_\alpha = 12{,}592$. Z uwagi na to, że $\chi^2 = 61{,}7623 > \chi^2_\alpha = 12{,}592$, hipotezę zerową o niezależności zmiennych odrzucamy. Oznacza to, że między rodzajem oglądanego programy telewizyjnego i wykształceniem telewidzów zachodzi zależność stochastyczna, a więc i korelacyjna. Siłę owej zależności można określić za pomocą odpowiedniego miernika korelacji. W rozpatrywanym przypadku siłę zależności można ustalić za pomocą jednego ze współczynników: Yule'a, Czuprowa, Cramera czy Pearsona. Badana zależność dotyczy bowiem cech niemierzalnych (jakościowych).

6.5.1. Weryfikacja hipotez w analizie współzależności

W podpunkcie tym zajmiemy się oceną istotności mierników współzależności między dwoma i więcej cechami statystycznymi. Mierniki te ustalane są na podstawie wyników próby losowej pobranej z populacji generalnej. Jeśli dwie badane cechy (X,Y) populacji generalnej mają charakter ilościowy, a zmienna losowa (X,Y) ma dwuwymiarowy rozkład normalny – to do oceny ścisłości związku między nimi wykorzystujemy współczynnik korelacji liniowej Pearsona z próby r_{xy} (lub r_{yx}). Wysuwamy wówczas hipotezę, że badane cechy są w populacji generalnej nieskorelowane, tzn. $H_0: \rho = 0$, wobec jednej z hipotez alternatywnych: $H_1: \rho \neq 0$, $H_1: \rho < 0$ lub $H_1: \rho > 0$ (ρ – współczynnik korelacji liniowej Pearsona w populacji generalnej).

Do weryfikacji H_0 stosujemy – w zależności od liczebności próby – test istotności z (dla $n \geq 122$) lub t (dla $n < 122$):

$$z \frac{r_{xy}}{\sqrt{1-r_{xy}^2}}\sqrt{n}, \tag{6.109}$$

$$t \frac{r_{xy}}{\sqrt{1-r_{xy}^2}}\sqrt{n-2}. \tag{6.110}$$

Przy założeniu prawdziwości H_0 statystyka (6.109) ma rozkład normalny $N(0,1)$, natomiast statystyka (6.110) – rozkład t-Studenta o $n-2$ stopniach swobody. Wartości krytyczne statystyki z_α odczytujemy z tablic dystrybuanty rozkładu normalnego, stosownie do postaci hipotezy alternatywnej. Hipotezę zerową o niezależności cech odrzucamy, jeśli statystyka określona wzorem (6.109) znajduje się w obszarze krytycznym, tzn.:

- przy $H_1: \rho \neq 0$, gdy $|z| \geq z_\alpha$;
- przy $H_1: \rho > 0$, gdy $z \geq z_\alpha$;
- przy $H_1: \rho < 0$, gdy $z \leq -z_\alpha$.

Wartości krytyczne statystyki t_α odczytujemy z tablic rozkładu t-Studenta przy założonym poziomie istotności α oraz $n-2$ stopniach swobody. W przypadku jednostronnych obszarów krytycznych, przy odczytywaniu wartości t_α należy podwoić poziom istotności. Hipoteza o braku związku zostanie odrzucona, jeśli wartość statystyki (6.110) znajduje się w obszarze krytycznym, tzn.:

- przy $H_1: \rho \neq 0$, gdy $|t| \geq t_\alpha$;
- przy $H_1: \rho > 0$, gdy $t \geq t_\alpha$;
- przy $H_1: \rho < 0$, gdy $t \leq -t_\alpha$.

Odrzucenie hipotezy zerowej oznacza, że związek między zmiennymi jest statystycznie istotny. Dodać należy, że hipotezy alternatywne o postaci $H_1: \rho > 0$ lub $H_1: \rho < 0$ wysuwamy wtedy, gdy we wnioskowaniu statystycznym uwzględnia się nie tylko siłę, ale również kierunek związku w populacji generalnej.

Załóżmy, że z populacji studentów wylosowano próbę $n = 27$ studentów i dokonano pomiaru ich wzrostu i pojemności klatki piersiowej. Obliczony współczynnik korelacji liniowej Pearsona między tymi cechami wyniósł $r_{xy} = 0{,}3918$. Na poziomie istotności $\alpha = 0{,}01$ mamy zweryfikować hipotezę, że badane cechy są liniowo nieskorelowane.

Formułując hipotezę zerową $H_1: \rho = 0$, przy dwustronnej hipotezie alternatywnej $H_1: \rho \neq 0$, obliczamy wartość statystyki t (mała próba):

$$t = \frac{0{,}3918}{\sqrt{1-0{,}3918^2}}\sqrt{25} = 2{,}129.$$

Dla poziomu istotności $\alpha = 0{,}01$ oraz 25 stopni swobody odczytujemy z tablic rozkładu t-Studenta wartość krytyczną $t_\alpha = 2{,}787$. Ze względu na to, że $t = 2{,}129 < t_\alpha = 2{,}787$, brakuje podstaw do odrzucenia H_0. Taka decyzja weryfikacyjna oznacza, że związek między badanymi zmiennymi nie jest statystycznie istotny.

W przypadku dużej próby korzystamy ze statystyki z określonej wzorem (6.109). Na przykład z populacji pracowników pewnej dużej firmy o zasięgu międzynarodowym wylosowano $n = 324$ pracowników i zbadano ich pod względem tygodniowego czasu poświęconego na podnoszenie poziomu kulturalnego i zawodowego oraz czas wolny. Współczynnik korelacji między tymi zmiennymi wyniósł $r_{xy} = 0{,}94$. Naszym zadaniem jest weryfikacja hipotezy, że związek ten jest – z prawdopodobieństwem popełnienia błędu I rodzaju na poziomie 0,03 – statystycznie istotny.

W pierwszym kroku stawiamy hipotezę zerową i alternatywną: $H_1: \rho = 0$ wobec $H_1: \rho > 0$. Ze względu na dużą próbę, do weryfikacji H_0 wykorzystamy statystykę (6.109):

$$z = \frac{0{,}94}{\sqrt{1-0{,}94^2}}\sqrt{324} = 49{,}76.$$

Z tablic dystrybuanty rozkładu normalnego $N(0,1)$ odczytujemy – dla prawostronnego obszary krytycznego i poziomu istotności $\alpha = 0{,}03$ – wartość krytyczną: $z_\alpha = 1{,}88$. Porównując empiryczną wartość z z jej wartością krytyczną z_α mamy: $z = 49{,}76 > z_\alpha = 1{,}88$. Wartość z znalazła się w obszarze krytycznym, zatem odrzucamy hipotezę o braku korelacji między zmiennymi. Związek między tygodniowym czasem poświęconym na podnoszenie poziomu kulturalnego i zawodowego a czasem wolnym jest statystycznie istotny.

Jak wiadomo, oceniając związek między dwiema zmiennymi ilościowymi (X,Y) lub zmienną ilościową X i jakościową Y, posługujemy się **stosunkiem (wskaźnikiem) korelacyjnym**. W takim przypadku próba musi być duża, a jej wyniki pogrupowane w tablicę korelacyjną o wymiarach $(k \times r)$, gdzie k jest liczbą kolumn w tablicy, a r liczbą wierszy. Obliczony z próby stosunek korelacyjny e_{yx} (lub e_{xy}) jest estymatorem stosunku korelacyjnego η_{yx} (lub η_{xy}) w populacji generalnej. Hipoteza zerowa zakłada, że zmienna losowa Y nie jest korelacyjnie zależna od zmiennej X, a więc stosunek korelacyjny w populacji generalnej jest równy 0. Hipoteza alternatywna zakłada istnienie zależności korelacyjnej zmiennej losowej Y względem X. Mamy zatem: $H_0: \eta_{yx} = 0$ wobec $H_1: \eta_{yx} > 0$ (stosunek korelacyjny przyjmuje tylko dodatnie wartości z przedziału domkniętego od 0 do 1).

Do weryfikacji H_0 stosuje się statystykę F, określoną następująco:

$$F = \frac{e_{yx}^2 (n-k)}{(1-e_{yx}^2)(k-1)}, \qquad (6.111)$$

gdzie: e_{yx} jest stosunkiem korelacyjnym z próby, n – liczebnością próby, k – liczbą rozkładów warunkowych zmiennej niezależnej.

Jeśli prawdziwa jest hipoteza zerowa, to stosunek (6.111) ma rozkład F-Snedecora o $k-1$ oraz $n-k$ stopniach swobody. Obszar krytyczny jest tu **prawostronny** i określa go relacja: $P(F \geq F_\alpha) = \alpha$, gdzie α jest poziomem istotności. Tak więc jeśli wartość statystyki obliczonej z próby jest nie mniejsza od wartości krytycznej odczytanej z tablic rozkładu F-Snedecora, to hipotezę zerową należy odrzucić.

Jeżeli obie zmienne są mierzalne i logicznie uzasadnione jest rozpatrywanie obu kierunków zależności, to postępując w taki sam sposób jak opisano wyżej, można również zweryfikować hipotezę zerową o niezależności korelacyjnej zmiennej losowej X od zmiennej Y.

Załóżmy, że z dwuwymiarowej populacji generalnej o rozkładzie normalnym wylosowano 157-elementową próbę, której wyniki ujęto w postaci

tablicy korelacyjnej o $k = 7$ wierszach (cecha X) oraz $k = 5$ kolumnach (cecha Y). Z tablicy tej wyznaczono stosunki korelacyjne: $e_{yx} = 0{,}387$ oraz $e_{xy} = 0{,}361$. Wnioskując na poziomie istotności $\alpha = 0{,}05$, należy zweryfikować hipotezę o braku zależności między badanymi zmiennymi.

Wiadomo, że stosunki korelacyjne cechuje **asymetryczność**. Dlatego też należy oddzielnie weryfikować hipotezy zerowe o braku korelacji między cechami Y i X oraz X i Y. W pierwszym przypadku mamy: $H_0: \eta_{yx} = 0$, wobec hipotezy alternatywnej $H_1: \eta_{yx} > 0$. Do weryfikacji H_0 wykorzystujemy statystykę określoną wzorem (6.111). Empiryczna wartość tej statystyki wynosi:

$$F = \frac{0{,}387^2}{1-0{,}387^2} \cdot \frac{157-7}{7-1} = 4{,}40.$$

Hipotezę zerową $H_0: \eta_{yx} = 0$, wobec $H_1: \eta_{yx} > 0$ zweryfikujemy testem:

$$F = \frac{e_{xy}^2}{1-e_{xy}^2} \cdot \frac{n-r}{r-1}, \qquad (6.112)$$

gdzie r jest liczbą rozkładów warunkowych zmiennej niezależnej.

Podstawiając dane liczbowe do wzoru (6.112) otrzymujemy:

$$F = \frac{0{,}361^2}{1-0{,}361^2} \cdot \frac{157-5}{5-1} = 5{,}69.$$

Z tablicy rozkładu F-Snedecora odczytujemy – dla $\alpha = 0{,}05$ oraz $k-1 = 6$ i $n-k = 150$ stopnie swobody – wartość krytyczną $F_\alpha = 2{,}16$. W przypadku weryfikacji hipotezy $H_0: \eta_{yx} = 0$ wartość krytyczna dla $\alpha = 0{,}05$ oraz $r-1 = 4$, a także $n-r = 152$ stopnie swobody wynosi $F_\alpha = 2{,}43$. Tak więc w obydwu przypadkach wartości statystyk obliczonych z prób znalazły się w obszarze krytycznym, gdyż $F > F_\alpha$. Fakt ten powoduje konieczność odrzucenia hipotez zerowych, a to pozwala stwierdzić, że w obydwu przypadkach zależność między badanymi cechami jest statystycznie istotna.

Współczynnik korelacji kolejnościowej (rang Spearmana) jest miarą współzależności, której wartości zmiennych X i Y zostały zastąpione rangami. Hipotezy dotyczące istotności związku między zmiennymi X i Y są tu formułowane w analogiczny sposób, jak to miało miejsce przy współczynniku korelacji liniowej Pearsona czy stosunku korelacyjnym. Hipoteza zerowa zakłada, że między zmiennymi w populacji generalnej brak jest związku ($H_0: \rho_S = 0$), hipoteza alternatywna stwierdza natomiast, że związek taki istnieje ($H_1: \rho_S \neq 0$) bądź też jest on dodatni ($H_1: \rho_S > 0$) lub ujemny ($H_1: \rho_S < 0$). Sposób weryfikacji hipotezy zerowej zależy od liczebności próby. Dla dużych prób (praktycznie już dla $n > 10$) wykorzystuje się rozkład współczynnika korelacji ρ_S Spearmana, który jest – w przybliżeniu – normalny $N\left(0, \sqrt{\frac{1}{n-1}}\right)$. W tym przypadku statystykę testową określa wzór:

$$z = r_S \sqrt{n-1}. \qquad (6.113)$$

Wartość krytyczną testu odczytujemy z tablic dystrybuanty rozkładu normalnego przy ustalonym poziomie istotności α. Hipotezę zerową o braku związku (niezależności) odrzucamy, jeśli statystyka testowa znajdzie się w obszarze krytycznym, co będzie miało miejsce wtedy, gdy:

- przy $H_1: \rho_S \neq 0$, gdy $|z| \geq z_\alpha$,
- przy $H_1: \rho_S > 0$, gdy $z \geq z_\alpha$,
- przy $H_1: \rho_S < 0$, gdy $z \leq -z_\alpha$.

Jeżeli próba liczy $n \leq 10$ par obserwacji, to do weryfikacji hipotezy zerowej o niezależności zmiennych wykorzystuje się statystykę:

$$t = \frac{r_S}{\sqrt{\frac{1 - r_S^2}{n-2}}}, \tag{6.114}$$

gdzie r_S jest współczynnikiem korelacji kolejnościowej obliczonym z próby.

Przy założeniu prawdziwości H_0, statystyka (6.114) ma rozkład t-Studenta z $n-2$ stopniami swobody. Wartość krytyczną testu odczytujemy więc z tablic rozkładu t-Studenta przy danym poziomie istotności α i $n-2$ stopniach swobody. Sposób odczytu wartości krytycznej uzależniony jest od postaci hipotezy alternatywnej.

Załóżmy, że na poziomie istotności 0,05 mamy sprawdzić istotność obliczonego z 8-elementowej próby losowej współczynnika korelacji rang $r_S = 0{,}85$ między odległością, jaką studenci studiów zaocznych pokonują do siedziby uczelni, a ich wynikami w nauce.

Sprawdzaną hipotezą zerową jest tu $H_0: \rho_S = 0$, wobec $H_1: \rho_S \neq 0$. Ze względu na liczebność próby ($n = 8$) do weryfikacja hipotezy zerowej zastosujemy test (6.114):

$$t = \frac{0{,}85}{\sqrt{\frac{1 - 0{,}85^2}{8-2}}} = 3{,}95.$$

Dla $\alpha = 0{,}05$ i 6 stopni swobody odczytana wartość krytyczna wynosi: $t_\alpha = 2{,}447$. Ponieważ $|t| = 3{,}95 > t_\alpha = 2{,}447$, więc H_0 z błędem 5% należy odrzucić. Oznacza to istotną statystycznie korelację między badanymi zmiennymi.

W przypadku oceny istotności współczynnika korelacji wielorakiej zakłada się, że z n-wymiarowej populacji generalnej o rozkładzie normalnym wylosowano n-elementową próbę, z której obliczono współczynnik korelacji wielorakiej R_w. Stawiamy hipotezę zerową $H_0: \rho_w = 0$, wobec $H_1: \rho_w > 0$, gdzie ρ_w jest współczynnikiem korelacji wielorakiej w populacji generalnej. W celu weryfikacji H_0 posługujemy się statystyką o postaci:

$$F = \frac{R_w^2}{1 - R_w^2} \cdot \frac{n-m-1}{m}, \tag{6.115}$$

gdzie m jest liczbą zmiennych niezależnych. Statystyka (6.115) ma rozkład F-Snedecora o m oraz $n-m-1$ stopniach swobody.

Przyjmijmy, że z 16-elementowej próby wylosowanej z 3-wymiarowej populacji generalnej o rozkładzie normalnym (Y, X_1, X_2) obliczono współczynnik korelacji wielorakiej $R_w = 0{,}4$. Na poziomie istotności 0,05 należy zweryfikować hipotezę, że współczynnik ten w populacji generalnej jest równy zeru.

W przykładzie tym należy sprawdzić $H_0: \rho_w = 0$, wobec $H_1: \rho_w > 0$. Obliczona za pomocą wzoru (6.115) wartość statystyki testowej wynosi:

$$F = \frac{0{,}16}{1 - 0{,}16} \cdot \frac{16 - 2 - 1}{2} = 1{,}24.$$

Dla 2 oraz 13 stopni swobody i przy $\alpha = 0{,}05$ odczytujemy wartość krytyczną $F_\alpha = 3{,}81$. Obszar krytyczny jest tutaj prawostronny i równy przedziałowi $(3{,}81; +\infty)$. Ze względu na to, że $F = 1{,}24 < F_\alpha = 3{,}81$, nie ma podstaw do odrzucenia hipotezy zerowej.

Wnioskowanie statystyczne o współczynniku regresji wymaga – podobnie jak w przypadku współczynnika korelacji – założenia, że dwuwymiarowy rozkład badanych zmiennych jest normalny.

Załóżmy, że ze zbiorowości generalnej, w której dwuwymiarowa zmienna losowa (X, Y) ma rozkład normalny, pobrano n-elementową próbę losową. Z próby tej wyznaczono funkcję regresji o postaci: $\hat{y}_i = a_0 + a_1 x_i$. Naszym zadaniem jest ocena istotności współczynnika regresji.

Sprawdzaną hipotezą zerową jest tu $H_0: \alpha_1 = 0$, wobec $H_1: \alpha_1 \neq 0$ (lub $H_1: \alpha_1 < 0$ bądź $H_1: \alpha_1 > 0$), gdzie α_1 jest współczynnikiem regresji zmiennej Y względem X w populacji generalnej. Odrzucenie H_0 na korzyść H_1 oznacza istotność współczynnika regresji. Do weryfikacji H_0 wykorzystujemy następującą statystykę testową:

$$t = \frac{a_1}{s(a_1)} = \frac{s(u)}{\sqrt{\sum_{i=1}^{n} x_i^2 - n\bar{x}^2}}, \tag{6.116}$$

gdzie a_1 jest oceną punktową parametru α_1, a $s(a_1)$ średnim błędem szacunku parametru α_1. Symbolem $s(u)$ oznaczono odchylenie standardowe składnika resztkowego funkcji regresji y względem x.

Wyraz wolny w funkcji regresji nie ma samodzielnej interpretacji. Dlatego też – na ogół – weryfikuje się tylko hipotezę dotyczącą istotności współczynnika regresji.

Z dwuwymiarowej populacji generalnej (X, Y) o rozkładzie normalnym wylosowano 15-elementową próbę losową i na podstawie jej wyników oszacowano liniową funkcję regresji:

$$\begin{array}{rcrcl} \hat{y}_i & = & 63{,}8 & + & 2{,}28 x_i \\ & & (10) & & (0{,}44) \end{array},$$

gdzie w nawiasach pod ocenami parametrów podano ich średnie błędy szacunku. Chcemy zbadać – na poziomie istotności 0,05 – istotność współczynnika regresji.

Hipotezy zerowa i alternatywna są tutaj następujące: $H_0: \alpha_1 = 0$, wobec $H_1: \alpha_1 \neq 0$. Wartość statystyki testowej (6.116) jest równa:

$$t = \frac{2,28}{0,44} = 5,18.$$

Z tablic rozkładu t-Studenta odczytujemy przy poziomie istotności 0,05, dwustronnym obszarze krytycznym oraz 13 stopniach swobody wartość krytyczną $t_\alpha = 2,16$. Ponieważ $|t| = 5,18 > t_\alpha = 2,16$, wiec H_0 należy odrzucić z prawdopodobieństwem popełnienia błędu I rodzaju na poziomie istotności 0,05. Oznacza to, że oddziaływanie zmiennej niezależnej X na zmienną zależną Y jest statystycznie istotne.

Jeśli jest to merytorycznie uzasadnione, w formułowanej hipotezie alternatywnej można uwzględnić kierunek oddziaływania zmiennej objaśniającej na zmienną objaśnianą. Obszar krytyczny testu będzie wtedy odpowiednio prawostronny lub lewostronny.

Ważnym zagadnieniem analizy regresji jest wybór analitycznej postaci funkcji regresji. Poniżej przedstawimy dwa sposoby weryfikacji hipotezy zakładającej, że regresja w populacji generalnej jest liniowa. W pierwszym sposobie – stosowanym dla danych niepogrupowanych – wykorzystuje się **test serii**, w drugim zaś – **test F** dla danych ujętych w postaci tablicy korelacyjnej.

Jednym z podstawowych założeń liniowości funkcji regresji jest **losowość reszt**, czyli przypadkowy charakter odchyleń empirycznych wartości y_i zaobserwowanych w próbie od wartości teoretycznych $\hat{y}_i$ otrzymanych z oszacowanej funkcji regresji. Do sprawdzenia tego założenia posługujemy się testem serii. W teście tym weryfikowana hipoteza zerowa stanowi przypuszczenie, że regresja badanych zmiennych w populacji generalnej jest liniowa, a hipoteza alternatywna to zaprzeczenie H_0. Zapisujemy to w następujący sposób:

$$\begin{aligned} H_0&: E(Y|X = x) = \alpha_0 + \alpha_1 x \\ H_1&: E(Y|X = x) = \alpha_0 + \alpha_1 x \end{aligned} \qquad (6.117)$$

Po oszacowaniu klasyczną metodą najmniejszych kwadratów – na podstawie n-elementowej próby losowej – parametrów liniowego modelu regresji otrzymujemy aproksymantę o postaci: $\hat{y}_i = a_0 + a_1 x_i$. Dalsza procedura testowania H_0 przebiega w następujących etapach:

1) dla każdej zaobserwowanej w próbie wartości y_i oblicza się – przy wykorzystaniu oszacowanej liniowej funkcji regresji – wartości $\hat{y}_i$;

2) oblicza się różnice $(y_i - \hat{y}_i)$ i przyporządkowuje się im symbole a i b w następujący sposób: jeśli $(y_i - \hat{y}_i) > 0$, to używamy symbolu a, jeśli natomiast $(y_i - \hat{y}_i) < 0$ – wykorzystujemy symbol b. Reszty równe 0 są pomijane;

3) w uporządkowanym według rosnących wartości zmiennej niezależnej x_i ciągu ustalamy liczbę serii k oraz liczbę symboli n_a i n_b $(n = n_a + n_b)$;

4) z tablic rozkładu liczby serii odczytujemy – przy przyjętym poziomie istotności α oraz ustalonych liczebnościach n_a i n_b – dwie wartości krytyczne: k_1 (dla $\frac{\alpha}{2}$, n_a i n_b) oraz k_2 (dla $1 - \frac{\alpha}{2}$, n_a i n_b);

5) jeśli $k < k_1$ lub $k > k_2$, to H_0 o losowości reszt należy odrzucić. Brak podstaw do odrzucenia H_0 ma miejsce wtedy, gdy $k_1 \leq k \leq k_2$.

Wysunięto przypuszczenie, że czas trwania zwolnienia lekarskiego z powodu choroby jest m.in. uzależniony od długości stażu pracy w warunkach nieodpowiedniego mikroklimatu. W celu sprawdzenia słuszności tej hipotezy przeprowadzono badanie losowej próby 8 chorych. Otrzymano następujące wyniki:

Staż pracy w latach (x_i)	1	8	3	10	7	15	5	2
Czas choroby w dniach (y_i)	3	5	4	8	5	9	8	4

Na podstawie tych informacji mamy sprawdzić, czy wysunięte przypuszczenie jest słuszne przy poziomie istotności 0,05. Dodatkowo mamy zweryfikować, że funkcja regresji Y względem X jest liniowa.

Aby sprawdzić, czy wysunięte przypuszczenie jest słuszne, należy obliczyć współczynnik korelacji liniowej Pearsona. W tym celu obliczamy: $\bar{x} = 6,375$; $\bar{y} = 5,75$; $s(x) = 4,36$; $s(y) = 2,11$; $\sum_{i=1}^{8} (x_i - \bar{x})(y_i - \bar{y}) = 59,75$. Stąd współczynnik korelacji liniowej jest równy:

$$r_{yx} = \frac{59,75}{8 \cdot 4,36 \cdot 2,11} = 0,81.$$

Chcąc zbadać istotność współczynnika korelacji liniowej stawiamy hipotezy: $H_0: \rho_{yx} = 0$, wobec $H_1: \rho_{yx} \neq 0$. Ze względu na małą próbę, wysuniętą hipotezę zerową sprawdzamy testem t:

$$t = \frac{r_{yx}}{\sqrt{1 - r_{yx}^2}} \cdot \sqrt{n-2} = \frac{0,81}{0,59} \cdot \sqrt{6} = 3,362.$$

Wartość krytyczna, odczytana z tablic rozkładu t-Studenta dla $\alpha = 0,05$ oraz 6 stopni swobody (dwustronny obszar krytyczny), wynosi $t_\alpha = 2,447$. Ponieważ $|t| = 3,362 > t_\alpha = 2,447$, więc H_0 należy odrzucić. Oznacza to, że wysunięte przypuszczenie jest słuszne, gdyż współczynnik korelacji liniowej jest statystycznie istotny.

Funkcje regresji Y względem X szacujemy ze wzoru:

$$\hat{y}_i = \bar{y} + r_{xy} \frac{s(y)}{s(x)} (x_i - \bar{x}). \qquad (6.118)$$

Podstawiając odpowiednie dane liczbowe do wzoru (6.118) otrzymujemy:

$$\hat{y}_i = 5{,}75 + 0{,}81\frac{2{,}11}{4{,}36}(x_i - 6{,}375) = 3{,}25 + 0{,}392x_i.$$

Porządkując wartości zmiennej zależnej Y według niemalejących wartości liczbowych zmiennej niezależnej X, mamy:

x_i	1	2	3	5	7	8	10	15
y_i	3	4	4	8	5	5	8	9

Do wyznaczenia reszt niezbędne są wartości teoretyczne zmiennej zależnej Y. Znajdujemy je z funkcji regresji w następujący sposób:

$\hat{y}_{x=1} = 3{,}25 + 0{,}392 \cdot 1 = 3{,}642,$
$\hat{y}_{x=2} = 3{,}25 + 0{,}392 \cdot 2 = 4{,}034,$
$\hat{y}_{x=3} = 3{,}25 + 0{,}392 \cdot 3 = 4{,}426,$
$\hat{y}_{x=5} = 3{,}25 + 0{,}392 \cdot 5 = 5{,}210,$
$\hat{y}_{x=7} = 3{,}25 + 0{,}392 \cdot 7 = 5{,}994,$
$\hat{y}_{x=8} = 3{,}25 + 0{,}392 \cdot 8 = 6{,}386,$
$\hat{y}_{x=10} = 3{,}25 + 0{,}392 \cdot 10 = 7{,}170,$
$\hat{y}_{x=15} = 3{,}25 + 0{,}392 \cdot 15 = 9{,}130.$

Reszty znajdujemy jako różnice $y_i - \hat{y}_i$: –0,642; –0,034; –0,426; 2,79; –0,994; –0,386; 0,83; –0,13. Na podstawie reszt określamy następujący ciąg symboli a i b: *bbbabbab*. Liczba serii $k = 5$, przy czym $n_a = 2$ oraz $n_b = 6$. Przyjmując poziom istotności 0,10, z tablic rozkładu liczby serii, odczytujemy dwie wartości krytyczne: $k_1 = 2$ oraz $k_2 = 5$. Ze względu na zachodzącą nierówność: $k_1 = 2 < k = 5 \leq k_2 = 5$, brak jest podstaw do odrzucenia H_0. Reszty mają zatem charakter losowy, a funkcja regresji zmiennej Y względem X jest liniowa.

Jeżeli otrzymane z dużej próby wyniki obserwacji dwóch zmiennych ilościowych X i Y pogrupujemy w tablicę korelacyjną, to możemy obliczyć zarówno współczynnik korelacji liniowej Pearsona, jak również wskaźniki korelacyjne e_{xy} i e_{yx}.

W takim przypadku hipotezę zerową o liniowości regresji weryfikujemy za pomocą statystyki F określonej wzorem:

$$F = \frac{e_{yx}^2 - r^2}{1 - e_{yx}^2} \cdot \frac{n-k}{k-2}. \qquad (6.119)$$

Hipotezę zerową można tutaj zapisać w następujący sposób: $H_0: \eta_{yx}^2 = \rho^2$, wobec $H_1: \eta_{yx}^2 \neq \rho^2$. Wynika to z faktu, że w sytuacji występowania związku liniowego zachodzi: $e_{yx} = e_{xy} = r$. Uogólniając to na całą populację możemy zapisać: $\eta_{yx} = \eta_{xy} = \rho$.

Jeżeli hipoteza zerowa jest prawdziwa, to statystyka F określona wzorem (6.119) ma rozkład F-Snedecora o $k - 2$ oraz $n - k$ stopniach swobody.

Załóżmy, że z populacji gospodarstw domowych pobrano próbę losową $n = 320$ gospodarstw i zbadano je pod względem wielkości dziennego spożycia ziemniaków (X) oraz wielkości dziennego spożycia przetworów zbożowych (Y). Z wyników próby utworzono tablicę korelacyjną o wymiarach 7×9 (7 klas dla cechy X i 9 dla cechy Y). Na podstawie utworzonej tablicy obliczono: $e_{yx} = 0{,}252$; $e_{xy} = 0{,}279$; $r_{xy} = r_{yx} = 0{,}23$. Na poziomie istotności 0,05 należy zweryfikować hipotezy o prostoliniowości: a) regresji Y względem X, b) regresji X względem Y.

W pierwszym przypadku wyznaczamy – za pomocą wzoru (6.119) – wartość statystyki F:

$$F = \frac{0{,}252^2 - 0{,}23^2}{1 - 0{,}252^2} \cdot \frac{320 - 7}{7 - 2} = 0{,}7089.$$

Ze względu na to, że wartość statystyki F jest mniejsza od jedności, za F należy przyjąć jej odwrotność, czyli: $F = \frac{1}{0{,}7089} = 1{,}4106$. Trzeba również pamiętać o zmianie kolejności liczby stopni swobody przy odczytywaniu wartość krytycznej F_α. W tym przypadku wartość krytyczna wynosi: $F_\alpha = 2{,}24$. Obszar krytyczny jest więc następujący: $(2{,}24; +\infty)$. Ze względu na to, że $F = 1{,}4106 < F_\alpha = 2{,}24$, brakuje podstaw do odrzucenia H_0 o prostoliniowości regresji Y względem X (na poziomie istotności $\alpha = 0{,}05$).

W przypadku weryfikacji hipotezy zerowej o prostoliniowości regresji X względem Y, korzystamy również ze wzoru (6.119), ale nieco zmodyfikowanego. W tym przypadku mamy:

$$F = \frac{0{,}279^2 - 0{,}23^2}{1 - 0{,}279^2} \cdot \frac{320 - 9}{9 - 2} = 1{,}202.$$

Tym razem wartość krytyczna dla $\alpha = 0{,}05$ oraz 7 i 311 stopni swobody wynosi: $F_\alpha = 3{,}25$. Ponieważ $F = 1{,}202 < F_\alpha = 3{,}25$, nie ma podstaw do odrzucenia – na poziomie istotności 0,05 – hipotezy o prostoliniowości regresji X względem Y.

6.5.2. Estymacja przedziałowa w analizie współzależności

W tym podpunkcie zajmiemy się estymacją przedziałową współczynnika korelacji liniowej Pearsona oraz parametrów strukturalnych liniowej funkcji regresji.

Jak wiadomo, współczynnik korelacji liniowej z próby ($r_{xy} = r_{yx}$) jest estymatorem współczynnika korelacji w populacji generalnej (ρ), a konkretna jego wartość liczbowa otrzymana z wyników próby losowej stanowi ocenę punktową. Dla dużych prób ($n \geq 122$) współczynnik z próby ma rozkład asymptotycznie normalny $N\left(\rho, \frac{1-\rho^2}{\sqrt{n}}\right)$. Przyjmując współczynnik ufności $1 - \alpha$ oraz odczytując z tablic rozkładu normalnego $N(0,1)$ taką wartość z_α, aby zachodziło: $P(-z_\alpha < Z < z_\alpha) = 1 - \alpha$, możemy zapisać następującą – przybliżoną – równość:

$$P\left(-z_\alpha < \frac{r-\rho}{1-r^2}\sqrt{n} < z_\alpha\right) = 1 - \alpha. \qquad (6.120)$$

Przekształcając relację (6.120) otrzymujemy:

$$P\left(r - z_\alpha \frac{1-r^2}{\sqrt{n}} < \rho < r + z_\alpha \frac{1-r^2}{\sqrt{n}}\right) = 1-\alpha. \qquad (6.121)$$

Jest to wzór na przedział ufności z dużej próby dla współczynnika korelacji ρ o rozkładzie normalnym.

W pewnym liceum przeprowadzono wśród 150 uczniów badania dotyczące wyników osiąganych w nauce i warunków domowych uczniów. Dla obu zmiennych przygotowano odrębne skale pomiarowe ujmujące wartości zmiennych w punktach. Z uzyskanych danych liczbowych obliczono $r_{xy} = 0{,}77$. Przy współczynniku ufności 0,90 należy zbudować przedział ufności dla współczynnika korelacji liniowej Pearsona w populacji generalnej.

Ponieważ liczebność próby jest duża ($n = 150$), więc do wyznaczenia przedziału ufności dla współczynnika ρ w populacji generalnej wykorzystamy wzór (6.121). Z tablic dystrybuanty rozkładu normalnego $N(0,1)$ odczytujemy wartość z_α tak, aby zachodziło: $F(z_\alpha) = 1 - \frac{\alpha}{2}$, tzn. $F(z_\alpha) = 1 - \frac{0{,}10}{2} = 0{,}95$.

Stąd $z_\alpha = 1{,}64$. Podstawiając odpowiednie dane liczbowe do wzoru (6.121) otrzymujemy:

$$0{,}77 - 1{,}64\frac{1-0{,}5929}{12{,}25} < \rho < 0{,}77 + 1{,}64\frac{1-0{,}5929}{12{,}25}$$

$$0{,}72 < \rho < 0{,}82.$$

Otrzymany przedział (0,72;0,82) jest jednym z tych przedziałów, które z prawdopodobieństwem 0,90 pokrywają nieznaną wartość współczynnika korelacji między badanymi zmiennymi w populacji generalnej uczniów.

Jeśli mamy zbudować przedział ufności dla współczynnika korelacji ρ w populacji generalnej na podstawie współczynnika r obliczonego z małej próby ($n < 122$), korzystamy ze wzoru:

$$P\left(z - u_\alpha \frac{1}{\sqrt{n-3}} < \rho < z + u_\alpha \frac{1}{\sqrt{n-3}}\right) = 1-\alpha, \qquad (6.122)$$

gdzie:

$$z = \frac{1}{2}\ln\frac{1+r}{1-r}, \qquad (6.123)$$

natomiast u_α jest wartością odczytaną z tablic dystrybuanty rozkładu normalnego $N(0,1)$ w taki sposób, aby zachodziło: $F(u_\alpha) = 1 - \frac{\alpha}{2}$.

Warto zauważyć, że określone wzorem (6.123) wartości z można odczytać z tablic podających przekształcone wartości r na z.

Badanie, które objęło 90 niezależnie wylosowanych uczniów szkół średnich w pewnym województwie, dostarczyło m.in. informacji o wysokości miesięcznego „kieszonkowego" oraz o wysokości miesięcznych dochodów rodziców. Obliczony współczynnik korelacji liniowej dla tych cech wyniósł: $r_{xy} = 0{,}57$. Należy – przy współczynniku ufności 0,95 – zbudować przedział ufności dla współczynnika korelacji ρ między rozpatrywanymi zmiennymi w populacji generalnej.

Ze względu na małą próbę, przy konstrukcji przedziału ufności dla ρ w populacji generalnej wykorzystamy wzór (6.122). Najpierw transformujemy, zgodnie ze wzorem (6.123), wartość współczynnika korelacji liniowej r na z:

$$z = \frac{1}{2}\ln\frac{1+0{,}57}{1-0{,}57} = 0{,}65.$$

Wartość u_α odczytana z tablic dystrybuanty rozkładu normalnego $N(0,1)$ wynosi 1,96 (dla $\alpha = 0{,}05$). Podstawiając odpowiednie dane liczbowe do wzoru (6.122) otrzymujemy:

$$0{,}65 - 1{,}96\frac{1}{\sqrt{90-3}} < \rho < 0{,}65 + 1{,}96\frac{1}{\sqrt{90-3}},$$

$$0{,}44 < \rho < 0{,}86.$$

Korzystając z tablicy podającej przekształcenie współczynnika korelacji r na z, mamy:

dla $z = 0{,}44$, $r = 0{,}4136$,

dla $z = 0{,}86$, $r = 0{,}6963$.

Ostatecznie zatem poszukiwaną realizacją przedziału ufności, przy współczynniku $1-\alpha = 0{,}95$, dla ρ jest:

$$0{,}4136 < \rho < 0{,}6963.$$

Przedział (0,4136; 0,6963) jest jednym z tych, które z prawdopodobieństwem 0,95, pokrywają nieznaną wartość współczynnika korelacji liniowej między wysokością miesięcznego „kieszonkowego" uczniów a poziomem dochodów ich rodziców w populacji generalnej uczniów.

Estymację przedziałową parametrów funkcji regresji I rodzaju Y względem X dla małej próby przeprowadza się na podstawie następujących wzorów:

$$P(a_0 - t_\alpha s(a_0) < \alpha_0 < a_0 + t_\alpha s(a_0)) = 1-\alpha \qquad (6.124)$$

oraz

$$P(a_1 - t_\alpha s(a_1) < \alpha_1 < a_1 + t_\alpha s(a_1)) = 1-\alpha, \qquad (6.125)$$

gdzie a_0 i a_1 są odpowiednio otrzymanymi z próby ocenami wyrazu wolnego i współczynnika regresji liniowej, t_α – to wartość odczytana z tablic rozkładu t-Studenta dla poziomu istotności α oraz $n-2$ stopni swobody, natomiast $s(a_0)$ i $s(a_1)$ są średnimi błędami szacunku odpowiednich parametrów.

W analogiczny sposób buduje się przedziały ufności dla parametrów funkcji regresji X względem Y, tzn. β_0 i β_1.

Załóżmy, że dla 25-osobowego zespołu wylosowanych pracowników wyznaczono funkcję regresji określającą zależność między przeciętną liczbą popełnianych błędów przy rozwiązywaniu testu (y) a kolejnymi godzinami pracy (x). Otrzymano następującą funkcję regresji:

$$\hat{y}_i = \underset{(0,1)}{2} + \underset{(0,002)}{0,5x_i},$$

gdzie w nawiasach pod ocenami parametrów podane są ich średnie błędy szacunku.

Przy współczynniku ufności 0,95 mamy zbudować przedział ufności dla parametrów α_0 i α_1 w populacji generalnej. Przedział ufności dla wyrazu wolnego wyznaczymy ze wzoru (6.124):

$$2 - 2{,}069 \cdot 0{,}1 < \alpha_0 < 2 + 2{,}069 \cdot 0{,}1$$
$$1{,}7931 < \alpha_0 < 2{,}22069.$$

Przedział (1,7931; 2,22069) z ufnością 0,95 pokrywa wyraz wolny liniowej funkcji regresji liczby popełnionych błędów przy rozwiązywaniu testu względem kolejnych godzin pracy.

W analogiczny sposób buduje się przedział ufności dla współczynnika regresji, który w naszym przypadku ma postać:

$$0{,}5 - 2{,}069 \cdot 0{,}002 < \alpha_1 < 0{,}5 + 2{,}069 \cdot 0{,}002$$
$$0{,}4959 < \alpha_1 < 0{,}5041.$$

Tak więc, przedział liczbowy (0,4959; 0,5041) jest jednym z tych przedziałów, które z prawdopodobieństwem 0,95 pokrywają nieznaną wartość współczynnika regresji liczby popełnionych błędów względem kolejnych godzin pracy.

ZADANIA

6.1. Dwuwymiarowa zmienna losowa (X,Y) ma następujący rozkład:

X \ Y	1	2	3
0	0,1	0,2	0,3
1	0,1	0,1	0,2

Obliczyć wartość oczekiwaną i wariancję zmiennej losowej X i zmiennej losowej Y.

6.2. Dany jest rozkład prawdopodobieństwa dwuwymiarowej zmiennej losowej (X,Y):

X \ Y	3	4	5
2	0,1	0	0
3	0,3	0,2	0
4	0,1	0,2	0
5	0	0	0,1

Wyznaczyć cov(X,Y) oraz współczynnik korelacji liniowej między zmiennymi losowymi X i Y.

6.3. Łączny rozkład dwuwymiarowej zmiennej losowej (X,Y) charakteryzują prawdopodobieństwa: $p_{11} = 0{,}15$; $p_{12} = 0{,}35$; $p_{21} = 0{,}15$; $p_{22} = 0{,}35$. Czy zmienne losowe X i Y są niezależne?

6.4. Łączny rozkład prawdopodobieństwa dwuwymiarowej zmiennej losowej (X,Y) przedstawia się następująco:

X \ Y	2	3	4	5	6
1	0,01	0,01	0,02	0,04	0,03
2	0,06	0,09	0,10	0,11	0,10
3	0,09	0,11	0,11	0,07	0,05

Wyznaczyć rozkłady brzegowe zmiennych X i Y. Znaleźć rozkład warunkowy zmiennej Y pod warunkiem, że $X = 2$.

6.5. Zmienne losowe $X=\{1,2\}$ oraz $Y=\{1,3,5\}$ są niezależne. Zapisać rozkład prawdopodobieństwa dwuwymiarowej zmiennej losowej (X,Y), jeśli $p_{i\cdot} = [0{,}3;0{,}7]$, $p_{\cdot j} = [0{,}2;0{,}3;0.5]$. Znaleźć rozkłady warunkowe P(Y/X=1) oraz P(Y/X=2).

6.6. Rozkład prawdopodobieństwa zmiennej losowej dwuwymiarowej (X,Y) przedstawia się następująco:

X \ Y	5	6	7
1	0	0	0,1
2	0,1	0,2	0,1
3	0,3	0,1	0,1

Znaleźć rozkłady brzegowe zmiennych losowych X i Y. Czy zmienne losowe X i Y są zależne, czy niezależne?

6.7. Znaleźć rozkład łączny dwuwymiarowej zmiennej losowej (X,Y), gdzie X ma rozkład dwumianowy z parametrami $n = 5$ i $p = 1/3$, natomiast Y ma rozkład dwumianowy z parametrami $n = 4$ i $p = 0{,}25$. Obliczyć: $P(X = 2, Y = 3)$; $P(1 < X \le 4, 1 \le Y \le 3)$; $P(X < 3, Y = 3)$. Założyć, że zmienne X i Y są niezależne.

6.8. Rozkład prawdopodobieństwa dwuwymiarowej zmiennej losowej (X,Y) przedstawia się następująco:

X \ Y	1	2	3
1	0,24	0,18	0,18
2	0,16	0,12	0,12

Porównując warunkowe rozkłady prawdopodobieństwa zmiennych losowych X i Y z ich rozkładami brzegowymi zbadać, czy zmienne losowe X i Y są zależne, czy niezależne.

6.9. Dwuwymiarowa zmienna losowa (X,Y) ma następujący rozkład:

X \ Y	−1	0
0	0,4	0,4
1	0,1	0,1

Zbadać, czy zmienne losowe X i Y są niezależne.

6.10. Dany jest dwuwymiarowy rozkład zmiennej losowej skokowej (X, Y):

X \ Y	1	2	3
2	0,2	0,1	0,1
4	0,1	0,2	0,3

Czy zmienne losowe X i Y są dodatnio skorelowanymi zmiennymi losowymi?

6.11. Dana jest funkcja:

$$f(x,y)=\begin{cases} Cxy & \text{dla} \quad 1\le x\le 2;\ 1\le y\le 2 \\ 0 & \text{dla pozostałych} \quad (x,y) \end{cases}.$$

Wyznaczyć stałą C, aby podana funkcja była gęstością prawdopodobieństwa dwuwymiarowej zmiennej losowej (X, Y).

6.12. Funkcja gęstości dwuwymiarowej zmiennej losowej (X, Y) ma postać:

$$f(x,y)=\begin{cases} 1{,}5x^2y & \text{dla} \quad 0\le x\le 1;\ 0\le y\le 2 \\ 0 & \text{dla pozostałych} \quad (x,y) \end{cases}.$$

Wyznaczyć funkcje gęstości rozkładów brzegowych zmiennych losowych X i Y.

6.13. Łączny rozkład dwuwymiarowej zmiennej losowej (X, Y) przedstawia się następująco:

X \ Y	1	2	3	4
1	0,2	0,1	0	0
2	0,1	0,2	0,1	0,1
3	0	0	0,1	0,1

Wyznaczyć rozkłady brzegowe dla zmiennych losowych X i Y oraz obliczyć wartości oczekiwane i wariancje tych rozkładów. Obliczyć również kowariancję i współczynnik korelacji pomiędzy zmiennymi losowymi X i Y.

6.14. Funkcja gęstości dwuwymiarowej zmiennej losowej (X, Y) ma postać:

$$f(x,y)=\begin{cases} 1{,}5x^2y & \text{dla} \quad 0\le x\le 1;\ 0\le y\le 2 \\ 0 & \text{dla pozostałych} \quad x,y \end{cases}.$$

Sprawdzić, czy zmienne losowe X i Y są niezależne.

6.15. W badaniach nad zależnością między wielkością produkcji (w tys. sztuk) a jednostkowymi kosztami produkcji (w PLN) w pięciu firmach uzyskano następujące wyniki:

Wielkość produkcji	23	30	41	45	50
Jednostkowe koszty	50	38	34	30	28

Obliczyć i zinterpretować współczynnik determinacji liniowej.

6.16. W mieście wojewódzkim L ogłoszono przetarg nieograniczony na sprzedaż działek przeznaczonych pod budownictwo jednorodzinne. Informacje o powierzchni działek (w m^2) oraz o ich cenach wywoławczych (w tys. zł) przedstawiają się następująco:

Numer działki	1	2	3	4	5	6	7	8	9	10
Powierzchnia działki	919	929	945	986	1002	1006	1007	1015	1155	1363
Cena wywoławcza	49,70	53,30	51,00	57,30	57,35	58,30	57,65	54,80	65,30	77,30

Obliczyć i zinterpretować współczynnik determinacji liniowej.

6.17. Średnia dobowa temperatura w °C oraz liczba sprzedanych butelek napojów chłodzących kształtowały się następująco:

Temperatura	18	24	20	35	18	29
Sprzedaż	62	78	66	160	52	107

Obliczyć siłę i kierunek zależności między badanymi cechami.

6.18. Liczba emitowanych tygodniowo reklam wyrobu X i wysokość obrotów (w tys. zł) kształtowały się następująco:

Liczba reklam	3	5	7	7	8
Obroty	115	140	155	160	180

Jaki procent zmian zmiennej zależnej nie może być wyjaśniony zmianami zmiennej niezależnej?

6.19. Poziom zatrudnienia (w tys. osób) oraz wielkość produkcji (w tys. ton) w ośmiu przedsiębiorstwach przemysłowych kształtowały się następująco:

Zatrudnienie	0,9	1,0	1,2	1,2	1,4	1,4	1,5	1,6
Produkcja	2,0	2,3	2,6	2,5	3,0	3,1	3,2	3,4

Obliczyć kowariancję i współczynnik korelacji liniowej Pearsona między badanymi zmiennymi. Wyniki zinterpretować.

6.20. W pewnym przedsiębiorstwie zbadano zależność między stażem pracy i wydajnością pracy sześciu pracowników. Otrzymano następujące wyniki:

Staż pracy (w latach)	3	3	5	8	10	13
Dzienna wydajność pracy (w sztukach)	10	7	12	17	16	22

Obliczyć i zinterpretować współczynnik determinacji liniowej.

6.21. W pewnym zakładzie zbadano pracowników ze względu na czas dojazdu do pracy w minutach (x) oraz liczbę spóźnień (y) w okresie jednego miesiąca. Otrzymane wyniki przedstawiono w tablicy korelacyjnej:

y_i \ x_j	5–15	15–25	25–35
1	3	2	
2	2	2	
3	1	3	2
4		1	4

Obliczyć współczynnik korelacji liniowej między zmiennymi. Wyniki zinterpretować.

6.22. Niech zmienna y oznacza liczbę dzieci w rodzinie, a zmienna x – czas trwania małżeństwa. Mając następujące dane: $y_1 = 0$; $y_2 = 1$; $y_3 = 2$; $y_4 = 3$; $y_5 = 4$; $x_{01} = 0$; $x_{02} = 5$; $x_{03} = 10$; $x_{04} = 15$ (rozpiętość przedziałów klasowych zmiennej x wynosi 5 lat), $n_{11} = 2$; $n_{21} = 2$; $n_{23} = 8$; $n_{22} = 13$; $n_{33} = 14$; $n_{32} = 13$; $n_{43} = 8$; $n_{34} = 5$; $n_{44} = 15$; $n_{54} = 20$, zbadać siłę i kierunek zależności pomiędzy badanymi zmiennymi.

6.23. Dyrekcja stalowni przedsiębiorstwa postanowiła sprawdzić, czy płace są odpowiednim bodźcem do podnoszenia jakości produkcji. W tym celu zbadano 10 losowo wybranych robotników stalowni i odnotowano wysokość ich płac oraz liczbę braków wyprodukowanych przez każdego z nich w okresie miesiąca. Otrzymano następujące informacje:

Płace w tys. zł	1,2	1,3	1,5	2,1	2,3	2,5	2,6	3,0	3,5	4,0
Liczba braków w sztukach	50	40	35	40	45	30	25	30	10	10

W jakim stopniu różnice w liczbie braków są zdeterminowane zróżnicowaniem płac?

6.24. Spośród 100 gospodarstw ogrodniczych o powierzchni 2,3–3,5 ha 10 osiąga roczny dochód 15–16,5 tys. zł, a 10 innych w tej samej grupie obszarowej 16,5–17,5 tys. zł. Następnych 10 gospodarstw o powierzchni 3,5–4,5 ha osiąga dochód wynoszący również 16,5–17,5 tys. zł; 40 gospodarstw w tej samej grupie obszarowej osiągnęło dochód 17,5–18,5 tys. zł, a dalszych 10 gospodarstw dochód równy 18,8–19,5 tys. zł. Pozostałe gospodarstwa miały obszar 4,5–5,5 ha, przy czym 10 miało dochód 17,5–18,5 tys. zł, a innych 10 w tejże grupie – dochód 19,5–20,5 tys. zł. Zbadaj, jaki jest kierunek i siła zależności pomiędzy badanymi zmiennymi.

6.25. Dane o liczbie błędów popełnianych przy przepisywaniu strony tekstu w zależności od stażu pracy przedstawia poniższa tablica:

Liczba błędów	Staż pracy w latach			
	1–3	3–5	5–7	7–9
0–2	6	3		
2–4	6	5		
4–6		5	8	5
6–8		3	8	5

Zbadać siłę i kierunek zależności między badanymi zmiennymi.

6.26. W grupie 33 przedsiębiorców zbadano koszty produkcji (w tys. PLN) i zyski (w mln PLN). Otrzymano następujące wyniki:

Koszty produkcji	Zyski			
	2	4	5	6
50	2	3	5	
45	2	5		
44		10	4	
40		1	1	

Obliczyć współczynnik korelacji liniowej Pearsona dla badanych zmiennych.

6.27. Badanie budżetów rodzinnych 100 rodzin dostarczyło m.in. informacji dotyczących zależności między liczbą osób w rodzinie a rocznymi wydatkami na żywienie w zakładach gastronomicznych:

Wydatki na żywienie w zł	Liczba osób w rodzinie		
	1	2	3
300	10		
200	10	30	10
100		20	20

Określić siłę i kierunek zależności między liczbą osób w rodzinie a wydatkami na żywienie w zakładach gastronomicznych (zakładamy prostoliniowy charakter związku).

6.28. Zarząd spółki X chciałby ustalić, czy na podstawie liczby reklam zamieszczanych w prasie można prognozować wielkość sprzedaży produktu Y w następnym tygodniu. W tym celu zebrano dane dotyczące liczby reklam oraz poziomu sprzedaży (w tys. PLN) w 6 losowo wybranych tygodniach. Otrzymano następujące wyniki:

Liczba reklam	4	8	10	3	2	6
Wielkość sprzedaży	5	7	7	4	5	6

Wykorzystując współczynnik korelacji kolejnościowej, określić siłę i kierunek korelacji między zmiennymi.

6.29. W wyborach prezydenckich przeprowadzono sondaż preferencji wyborców na losowej próbie 50 tys. dorosłych Polaków. W trakcie tego sondażu różne cechy wyborców kojarzone były z częstotliwością głosowania na poszczególnych kandydatów. Otrzymano następujące wyniki:

Nr kandydata	1	2	3	4	5	6	7	8
Odsetek głosów mieszkańców miast (x)	31,4	33,4	13,3	7,0	0,5	4,6	3,5	4,3
Odsetek głosów mieszkańców wsi (y)	33,4	33,4	6,3	7,2	9,5	2,5	2,7	1,4

Stosując współczynnik korelacji rang Spearmana ocenić, czy występuje korelacja (zgodność uporządkowań kandydatów) wyników głosowania mieszkańców miast i wsi.

6.30. Za pomocą współczynnika korelacji rang Spearmana zbadać siłę i kierunek zależności między liczbą ludności (w mln osób) a powierzchnią (w km^2) w 10 województwach:

Nr województwa	1	2	3	4	5	6	7	8	9	10
Ludność	3,95	2,41	1,44	1,34	1,23	1,14	1,13	1,13	1,12	1,03
Powierzchnia	6650	3788	7394	8151	3254	9211	6287	1523	10349	8435

6.31. W pewnym województwie przeprowadzono badanie dotyczące stosunku obywateli do niektórych instytucji publicznych, które oznaczono literami od A do F. Do próby wybrano jedno małżeństwo, które poproszono o uporządkowanie instytucji według ich indywidualnych preferencji oraz nadano im rangi. Otrzymano następujące wyniki:

Instytucja	A	B	C	D	E	F
Ranga męża	6	3	2	1	4	5
Ranga żony	6	5	3	1	2	4

Za pomocą współczynnika korelacji kolejnościowej dokonać oceny zgodności poglądów małżonków pod względem ich preferencji do instytucji.

6.32. Na losowej próbie dorosłych Polaków przeprowadzono badanie o czynnikach szkodzących politykom w ich karierze. Czynniki nieszkodzące (Nie) zostały uporządkowane od najczęściej do najmniej wymienianych. Czynniki szkodzące (Tak) podano w procentach wskazań:

Czynniki	Ranga (Nie)	Ranga (Tak)
Rozwód	1	21
Życie bez ślubu	2	38
Zdrada	3	26
Niesłowność	4	64
Współpraca z SB	5	77
Kłamstwo	6	91
Przestępstwo	7	93
Alkoholizm	8	96

Za pomocą współczynnika korelacji rang sprawdzić, czy między czynnikami, które szkodzą bądź nie szkodzą w karierze polityków istnieje zależność (największemu procentowi wskazań Tak należy przyporządkować rangę 1).

6.33. Za pomocą współczynnika korelacji rang ustalić siłę zależności pomiędzy opiniami klientów i pracownika kontroli jakości o towarze Z. Ujęte w punktach oceny przedstawiają się następująco:

Punkty uzyskane od pracownika kontroli	53	47	41	38	36	35	35	33	27	25	25
Punkty uzyskane od klientów	52	47	38	43	29	34	31	29	24	27	25

6.34. Siedmiu uczniów rozwiązywało dwa testy. Ich wyniki podane w punktach przedstawiają się następująco:

Test I	20	19	18	18	17	16	15
Test II	19	20	20	18	17	15	15

Wykorzystując współczynnik korelacji rang Spearmana ocenić, jak silna zależność występuje między wynikami obu testów.

6.35. Ustalić natężenie współzależności między opiniami o nauczycielach dyrektora szkoły i wizytatora. Opinie te zostały wydane na podstawie kontroli całokształtu pracy zawodowej i kwalifikacji nauczycieli. Wyniki kontroli ujęte w punktach przedstawiają się następująco:

Nauczyciele	A	B	C	D	E	F	G	H	I	J	K
Punkty od dyrektora	41	27	35	33	25	47	38	53	43	35	36
Punkty od wizytatora	38	24	34	29	27	47	43	52	39	31	29

6.36. Ranking firm w opinii klientów krajowych i zagranicznych kształtował się następująco:

Numer firmy	1	2	3	4	5	6
Rangi uzyskane od odbiorców krajowych	4	1	3	2	5	6
Rangi uzyskane od odbiorców zagranicznych	5	1	6	4	2	3

Czy opinie odbiorców krajowych i zagranicznych dotyczące badanych firm są zbieżne?

6.37. Wśród 12 uczniów badano związek pomiędzy pozycją ucznia w klasie w zakresie jego popularności i pozycją wyznaczoną przez nauczyciela matematyki z tytułu aktywności podczas lekcji. Otrzymano następujące wyniki w punktach:

Numer ucznia	1	2	3	4	5	6	7	8	9	10	11	12
Popularność w klasie	16	9	28	16	10	7	25	8	17	12	18	20
Opinia nauczyciela	12	8	16	10	9	8	16	10	15	13	14	14

Zbadaj siłę i kierunek zależności między badanymi zmienymi przy wykorzystaniu współczynnika korelacji rang Spearmana.

6.38. Dziesięć rodzin podało tygodniowe wydatki w złotych na pieczywo i na produkty mleczne. Odpowiednie kwoty wydatków przedstawiają się następująco:

Pieczywo	23	24	29	27	33	29	19	22	21	23
Produkty mleczne	25	28	30	30	35	41	22	25	26	26

Za pomocą współczynnika korelacji rang Spearmana ocenić siłę i kierunek zależności pomiędzy wyróżnionymi wydatkami.

6.39. Restauracja „Bristol", urządzająca przyjęcia weselne, podała poniższe informacje dotyczące kosztów (w tys. zł) oraz czasu trwania (w godzinach) przyjęć weselnych w określonym czasie:

Czas trwania przyjęcia	Koszt przyjęcia					
	do 4	4–8	8–12	12–16	16–20	20–24
0–2	2					
2–4	3					
4–6		6	10	11		3
6–8		8	12	16	4	
8–10				7	2	3

Obliczyć: a) współczynnik determinacji liniowej, b) stosunek korelacyjny kosztów przyjęcia względem czasu jego trwania, c) ocenić stopień krzywoliniowości związku.

6.40. Trzystu pracowników pewnej firmy pracuje na nowych maszynach. Osiągają oni średnią wydajność pracy 172 sztuk/h, przy odchyleniu standardowym równym 7 sztuk/h. Pozostałych 200 pracowników pracuje na starych maszynach, osiągając przeciętnie 170 sztuk/h, przy odchyleniu standardowym 6 sztuk/h. Jak silna jest zależność korelacyjna wydajności pracy od typu maszyn (stare – nowe)?

6.41. Na podstawie poniższych danych ustal – za pomocą stosunku korelacyjnego – siłę współzależności pomiędzy badanymi zmiennymi, jeśli: $n = 100$, $\sum_{i=1}^{r} s_i^2(x) n_i = 50$, $\sum_{i=1}^{r} \bar{x}_i^2 n_i = 575$, $\sum_{i=1}^{r} x_i n_i = 200$.

6.42. Średnie liczby punktów uzyskane na egzaminie ze statystyki przez studentki i studentów wynosiły po 26. Odchylenie standardowe liczby punktów dla całej zbiorowości (studentek i studentów łącznie) stanowiło 10% średniej ogólnej. Wiedząc, że stosunek liczby studentek do liczby studentów wynosi 2:3, ustal siłę zależności między płcią studentów a uzyskanymi przez nich wynikami.

6.43. W zbiorowości 15 pracowników podjęto próbę określenia zależności między liczbą spóźnień a zmianą, na której pracowali. Otrzymano następujące wyniki:

Zmiana I	2	3	0	3	1
Zmiana II	0	2	3	1	2
Zmiana III	2	3	3	3	2

W jakim stopniu liczba spóźnień zależy od wielozmianowego systemu pracy?

6.44. Z badania wydajności pracy pewnej grupy robotników wynika, że średnia wydajność pracy kobiet wynosi 22 sztuki/h, a mężczyzn 26 sztuk/h. Kobiety stanowią 40% ogółu zatrudnionych. Zróżnicowanie wydajności pracy – mierzone odchyleniem standardowym – dla całej zbiorowości robotników stanowiło 10% średniej wydajności pracy. Ustalić siłę zależności między wydajnością pracy a płcią pracowników.

6.45. Badając zmienność wieku osób deklarujących uczestnictwo w wyborach do Sejmu i Senatu RP stwierdzono, że wśród losowo wybranych 1500 osób średni wiek wyniósł 36,4 lat z przeciętnym zróżnicowaniem 16,6 lat. Wśród 800 zbadanych kobiet średni wiek wyniósł 42 lata z 50% zmiennością, wśród 700 zbadanych mężczyzn średni wiek wyniósł 30 lat z 10% zmiennością. Zbadać siłę zależności między wiekiem wyborców a płcią.

6.46. Na podstawie danych zwartych w poniższej tablicy, oceń siłę związku między liczbą zarejestrowanych bezrobotnych a ich płcią:

Płeć	Czas pozostawania bez pracy w miesiącach				
	0–1	1–6	6–12	12–24	24–36
Kobiety	50	111	116	114	90
Mężczyźni	35	105	97	68	32

6.47. Związek korelacyjny dwóch zmiennych określają następujące mierniki: $s^2(y) = 81$; $s^2(\bar{y}_i) = 64$; $r_{xy} = 0{,}9$; $e^2_{xy} = 0{,}92$. Czy korelacja między zmiennymi ma charakter prostoliniowy, czy krzywoliniowy?

6.48. Za pomocą stosunku korelacyjnego ocenić siłę związku między wiekiem a czasem pozostawania bez pracy:

Wiek w latach	Czas pozostawania bez pracy w miesiącach				
	0–1	1–6	6–12	12–24	24–36
18–24	52	65	82	58	20
25–34	13	62	55	47	35
35–44	14	60	51	47	40
45–54	5	26	23	25	21
Powyżej 55	1	3	3	5	6

6.49. Na podstawie danych uzyskanych w pewnej letniskowej miejscowości, obliczono współczynniki korelacji liniowej między trzema zmiennymi: tygodniową sprzedażą napojów chłodzących (x), tygodniową liczbą turystów przebywających w tej miejscowości (y) oraz panującą w niej przeciętną tygodniową temperaturą dnia (z). Otrzymano następujące wyniki: $r_{xy} = -0{,}30$; $r_{xz} = -0{,}70$; $r_{yz} = 0{,}80$. Jaka jest siła związku korelacyjnego między zmiennymi x i y, jeśli zostanie wyeliminowany wpływ temperatury?

6.50. W dziesięciu szkołach podstawowych przeprowadzono badanie dotyczące średnich osiągnięć szkolnych uczniów (x), średnich liczebności klas (y) oraz średniego stażu pracy nauczycieli zatrudnionych w tych szkołach (z). Otrzymano następujące wyniki:

x_i	3,7	4,0	3,3	3,1	4,3	3,6	4,0	3,2	3,5	3,4
y_i	25	19	30	32	18	27	20	34	24	28
z_i	10	19	11	6	21	15	25	8	12	10

Obliczyć: a) współczynniki korelacji liniowej między parami zmiennych, b) współczynniki korelacji cząstkowej (osiągnięć szkolnych od liczebności klas z wyłączeniem wpływu stażu pracy nauczycieli oraz osiągnięć szkolnych od stażu pracy nauczycieli z wyłączeniem wpływu liczebności klas), c) współczynnik korelacji wielorakiej.

6.51. Zbiorowość pracowników pewnego przedsiębiorstwa badano pod względem wydajności pracy (x_1), stażu pracy (x_2) oraz wynagrodzenia (x_3). Macierz współczynników korelacji między tymi zmiennymi jest równa:

$$\begin{bmatrix} 1{,}000 & 0{,}936 & 0{,}900 \\ & 1{,}000 & 0{,}965 \\ & & 1{,}000 \end{bmatrix}$$

Obliczyć współczynniki korelacji cząstkowej: $r_{x_1x_2 \cdot x_3}$; $r_{x_1x_3 \cdot x_2}$; $r_{x_2x_3 \cdot x_1}$ oraz współczynnik korelacji wielorakiej $R_{x_1 \cdot x_2x_3}$.

6.52. Badanie zależności między trzema zmiennymi: stażem pracy (x_1), wydajnością pracy (x_2) i poziomem wykształcenia (x_3), dostarczyło m.in. następujących informacji: $r_{x_1x_2} = 0{,}4$; $r_{x_1x_3} = 0{,}2$; $r_{x_2x_3} = 0{,}5$. Doświadczony kierownik jednego z wydziałów badanego przedsiębiorstwa twierdzi, że różnice w stażu i wykształceniu pracowników tylko w ok. 33% decydują o zróżnicowaniu wydajności pracy. Czy ma on rację?

6.53. W celu opisu zależności korelacyjnej pomiędzy szybkością zapamiętywania (x) a wiekiem (y) oraz ilorazem inteligencji (z) dokonano odpowiednich pomiarów, otrzymując: $\bar{x} = 7$ wyrazów/min, $\bar{y} = 10$ lat, $\bar{z} = 105$, $s(x) = 2$ wyrazy/min, $s(y) = 1{,}5$ roku, $s(z) = 6{,}5$, $r_{xy} = 0{,}326$, $r_{xz} = 0{,}852$, $r_{yz} = 0{,}231$. Na podstawie powyższych danych obliczyć: a) współczynniki korelacji cząstkowej, b) współczynnik korelacji wielorakiej.

6.54. Współczynniki korelacji liniowej Pearsona między produkcją dóbr trwałego użytkowania (x), podażą tych dóbr (y) oraz ich importem (z) są równe: $r_{xy} = 0{,}3$, $r_{xz} = -0{,}5$, $r_{yz} = 0{,}5$. Oblicz – posługując się współczynnikami korelacji cząstkowej – wszystkie uzasadnione logicznie związki między badanymi zmiennymi. Czy prawdą jest, że ok. 4% zmian podaży dóbr trwałego użytkowania zależy od innych czynników niż produkcja i import tych dóbr?

6.55. W czasie badania sprawności studentów otrzymano m.in. następujące informacje dotyczące wagi w kg (y) oraz prędkości w sekundach w biegu ma 100 m (z):

Waga	50	54	54	56	58	68	60	62	68	70	60	60
Prędkość	12,0	13,1	13,2	12,5	12,5	13,9	14,0	12,8	13,0	12,5	13,5	13,0

Ponadto obliczono współczynniki korelacji: $r_{xy} = 0{,}7$, $r_{xz} = 0{,}7$. Oblicz współczynnik korelacji cząstkowej między wagą i prędkością przy eliminacji wzrostu (x). Czy prawdą jest, że 40% zmian w prędkości jest uzależnionych od innych zmiennych niż wzrost i waga studentów?

6.56. W 250-osobowej grupie uczniów klas licealnych zbadano poziom wiadomości z języka polskiego (x) i matematyki (y). Współczynnik korelacji liniowej Pear-

sona między tymi zmiennymi wyniósł: $r_{xy} = 0,74$. Przypuszcza się, że na poziom tego współczynnika wywierają wpływ warunki domowe uczniów (z). Zbadać zależność między zasobem wiadomości z języka polskiego i matematyki, jeśli wyeliminujemy wpływ warunków domowych na uzyskiwane z tych przedmiotów wyniki. Wiadomo, że współczynniki korelacji między wynikami nauczania oraz warunkami domowymi są równe: $r_{xz} = 0,24$, $r_{yz} = 0,21$.

6.57. Wylosowano 1000 klientów PZU zamieszkujących w miastach liczących do 20 tys. mieszkańców oraz powyżej 20 tys. mieszkańców. Wylosowane osoby pytano o ich sytuację materialną. Otrzymano następujące wyniki:

Sytuacja materialna	Liczba klientów z miast o liczbie mieszkańców	
	do 20 tys.	powyżej 20 tys.
Zła	175	505
Średnia	35	245
Dobra	20	20

Na poziomie istotności 0,02 zweryfikować hipotezę o niezależności badanych zmiennych. Obliczyć – jeśli jest to uzasadnione – współczynnik korelacji Czuprowa.

6.58. Wylosowano 400 klientów pewnego banku i zbadano zależność między poziomem miesięcznych dochodów (w zł/osobę) w rodzinie i korzystaniem z kredytów gotówkowych w ciągu ostatniego roku. Otrzymano następujące wyniki:

Miesięczny dochód	Korzystał z kredytu	Nie korzystał z kredytu
Poniżej 1000 zł	180	70
Powyżej 1000 zł	30	120

Na poziomie istotności 0,01 zweryfikować hipotezę o braku zależności między badanymi zmiennymi. W przypadku wystąpienia zależności obliczyć współczynnik T Czuprowa.

6.59. Zaobserwowane w 250 rodzinach liczebności dotyczące zawodów ojca i syna przedstawia poniższa tablica:

Zawód ojca	Zawód syna		
	nauczyciel	lekarz	ekonomista
Nauczyciel	30	15	45
Lekarz	20	40	10
Ekonomista	10	10	70

Określić siłę związku między zawodem ojca i syna.

6.60. Wylosowano próbę 400 nabywców wprowadzanego na rynek nowego produktu i zadano im pytanie, skąd dowiedzieli się o istnieniu produktu. Drugie pytanie dotyczyło wykształcenia nabywców. Otrzymano następujące wyniki:

Wykształcenie	Źródło informacji		
	reklama w TV	reklama w czasopismach	inne
Wyższe	5	70	15
Średnie	70	140	30
Podstawowe i zawodowe	55	10	5

Na poziomie istotności 0,05 zweryfikować hipotezę o niezależności badanych zmiennych.

6.61. Analizując wadliwość wyrobów otrzymano następujące wyniki:

Przyczyny wady	Zakłady	
	A	B
Niska jakość	21	72
Gorszy surowiec	46	26

Zweryfikować, testem chi-kwadrat, hipotezę o niezależności między miejscem powstania wyrobu (zakład A lub B) a przyczyną uznania wyrobu za wadliwy. Przyjąć poziom istotności 0,01. W przypadku odrzucenia H_0 obliczyć siłę zależności między badanymi cechami, wykorzystując do tego celu mierniki właściwe dla cech niemierzalnych.

6.62. Na poziom koncentracji niewątpliwy wpływ wywiera – oprócz innych czynników – stopień zmęczenia. Stopień ten jest zależny od pory dnia. Na podstawie poniższych danych zweryfikować, na poziomie istotności 0,01, hipotezę o niezależności poziomu uwagi uczniów na lekcjach od pory dnia:

Pora dnia	Poziom uwagi	
	niski	wysoki
Rano	17	39
Południe	26	13
Popołudnie	19	12

6.63. Wśród 300 wylosowanych członków „Fitness Club" przeprowadzono ankietę dotyczącą stosowania diety oraz zmiany wagi ciała. Otrzymano następujące wyniki:

Waga	Dieta	
	nie stosują	stosują
Brak utraty	80	20
Utrata	100	100

Na poziomie istotności 0,05 zweryfikować hipotezę o niezależności badanych zmiennych.

6.64. W pewnym rejonie przeprowadzono badanie dotyczące życia zawodowego nauczycieli szkół średnich. Uzyskano następujące informacje:

Grupa wieku	Stan cywilny			
	wolni	pozostający w związku małżeńskim	owdowiali	rozwiedzeni
Adaptacyjna	115	85		
Produkcyjna	35	163	1	1
Emerytalna	24	152	21	3

Stosując współczynnik T Czuprowa zbadać, czy istnieje związek między stanem cywilnym a grupą wieku nauczycieli szkół średnich.

6.65. Związek odległości miejsca pracy od miejsca zamieszkania (x) i czasu dojazdów do pracy (y) pracowników pewnej firmy przedstawia poniższa tablica korelacyjna:

Odległość w km	Czas jednostronnych dojazdów (w minutach)				
	1–10	10–20	20–30	30–40	40–50
1–5	7				
5–10	5	14			
10–15		2	2		
15–20			12		
20–25				5	5

Wyznaczyć równania regresji prostoliniowych. Wyniki zinterpretować.

6.66. W wyniku badania zależności między zużyciem surowców (w kg) i kosztami produkcji (w USD) uzyskano następujące wyniki:

Koszty produkcji	Zużycie surowca		
	15–24	25–34	35–44
6–8	10		
8–10		10	
10–12		20	10
12–14		20	20
14–16			10

Na podstawie powyższych informacji należy wyznaczyć funkcje regresji oraz ustalić siłę i kierunek zależności między badanymi zmiennymi. Jaki będzie – teoretyczny – koszt produkcji przy zużyciu 50 kg surowca?

6.67. Empiryczny rozkład czasu remontu obrabiarek w pewnym zakładzie remontowym przedstawia się następująco:

Czas remontu w dniach	10–20	20–30	30–40	40–50	50–60
Liczba obrabiarek	10	30	50	40	20

Określić siłę i kierunek zależności między czasem remontu a wiekiem obrabiarek, jeśli wiadomo, że średni wiek remontowanych obrabiarek wynosi 16 lat, a jego względna dyspersja 30%. Wiadomo ponadto, że wydłużenie czasu eksploatacji o 1 rok powoduje przedłużenie czasu ich remontu przeciętnie o 2 dni. Jaki jest czas remontu 10-letnich obrabiarek?

6.68. W rezultacie badania cen działek budowlanych (y) w zł za 1 m^2 i odległości działek od centrum miasta w km (x) otrzymano następujące wyniki:

Ceny działek	1000	900	500	500	270	300	100
Odległość	0	1	2	3	4	5	6

Wyznacz równanie regresji opisujące ceny działek względem odległości od centrum. Zbadaj dokładność dopasowania tej funkcji do danych empirycznych.

6.69. W pewnym osiedlu mieszkaniowym przeprowadzono badanie zależności między liczbą osób w rodzinie (x) i wielkością mieszkania określonego liczbą izb (y). Otrzymano następujące informacje: $r_{xy} = 0,6$, $\bar{x} = 3,5$ osoby na izbę, $s(x) = 1,4$ osoby na izbę. Ponadto stwierdzono, że 5% rodzin zajmuje mieszkanie jednoizbowe, 20% – dwuizbowe, 40% – trzyizbowe, 30% – czteroizbowe oraz 5% – pięcioizbowe. Wyznacz parametry strukturalne liniowych funkcji regresji. Oblicz przeciętną liczbę izb zajmowanych przez rodziny czteroosobowe.

6.70. Na podstawie poniższych danych oszacowano liniową funkcję regresji o postaci: $\hat{y}_i = 0,9x_i + 2,2$. Zmienne x i y przyjęły następujące wartości:

Zmienna x	4,3	5,3	6,4	7,0	8,6	11,0
Zmienna y	2	4	5	5	7	10

Zbadać dokładność dopasowania funkcji regresji do danych empirycznych.

6.71. Wyznacz równania regresji liniowej na podstawie następujących danych: $r_{xy} = 0,8$, $b_1 = 40$, $\bar{x} = 2800$, $\bar{y} = 12$.

6.72. Badanie zależności między stażem pracy (y) i wiekiem pracowników (x) dostarczyło następujących informacji: $\bar{x} = 33$ lata, $a_1 = 0,84$ lat, $\hat{y}_i = 0,43x_i - 5,5$. Obliczyć współczynnik korelacji liniowej między badanymi zmiennymi.

6.73. Na podstawie poniższych danych ustal siłę i kierunek zależności pomiędzy stażem robotników bezpośrednio produkcyjnych a ich wydajnością pracy oraz wyznacz równania liniowych funkcji regresji: a) średni staż pracy wynosi 8 lat, b) przyrostowi stażu pracy o jeden rok towarzyszy wzrost wydajności pracy o 2 jednostki produktu na godzinę, c) wydajność pracy robotników różni się od wydajności średniej przeciętnie o ±5 jednostek produktu na godzinę, d) współczynnik zmienności stażu pracy wynosi 25%, e) średnia wydajność pracy wynosi 25 jednostek produktu na godzinę.

6.74. Mając następujące dane: $\bar{x} = 2,5$, $\bar{y} = 625$, $V(x) = 20\%$, $s(y) = 25$, $a_1 = 40$, wyznacz równania liniowych funkcji regresji. Jaka jest siła i kierunek zależności między badanymi zmiennymi?

6.75. Wyznacz równanie regresji liniowej opisujące zależność między cechami x i y, jeśli: a) $\hat{x}_i = 30 + 0,5y_i$, b) $6 < y_{typ} < 14$, c) $V(x) = 20\%$. Ustal również siłę i kierunek zależności między badanymi zmiennymi.

6.76. W modelu regresji liniowej zadłużenia z tytułu zaciągniętych kredytów w tys. zł (y) 80 podmiotów gospodarczych względem wartości produkcji w mln zł (x) otrzymano następujące miary: $\text{cov}(x,y) = -2,88$, $\bar{x} = 5,5$, $s(x) = 0,6$, $s(y) = 8$, $\bar{y} = 183$. Jaki jest poziom zadłużenia przy produkcji równej 9 mln zł? Zbadaj dokładność dopasowania funkcji do danych empirycznych.

6.77. Zależność między wiekiem pracowników a rozmiarami ich absencji chorobowej charakteryzują następujące mierniki: 1) odchylenie standardowe wieku wynosi 12 lat, 2) kowariancja badanych cech jest równa 68,32, 3) wariancja absencji chorobowej jest równa 6,25. Czy takie wyniki są możliwe?

6.78. Współczynnik indeterminacji liniowej funkcji regresji dochodów względem przeciętnych miesięcznych wydatków na cele kulturalne wynosi 0,6864. Średni miesięczny dochód w badanej grupie wynosił 1300 zł, przy odchyleniu standardowym równym 260 zł. Przeciętne miesięczne wydatki na cele kulturalne były równe 165 zł, przy współczynniku zmienności 35%. Jakich wydatków na cele kulturalne należy oczekiwać w gospodarstwach domowych o miesięcznych dochodach 1500 zł? Jaki jest – przeciętnie rzecz biorąc – poziom miesięcznych dochodów w gospodarstwach, w których na cele kulturalne wydaje się miesięcznie 150 zł?

6.79. W celu zbadania zależności między wiekiem kobiet a liczbą posiadanych dzieci wylosowano, w sposób niezależny, 100 kobiet. Otrzymane wyniki przedstawia poniższa tablica korelacyjna:

Liczba dzieci (x)	Wiek kobiet w latach (y)			
	15–25	25–35	35–45	45–55
1	2	1		
2	10	12	15	
3	8	19	10	5
4			5	4
5				3
6				6

Wyznacz funkcje regresji oraz oblicz współczynnik korelacji liniowej Pearsona.

6.80. W 30 losowo wybranych firmach zbadano koszt produkcji detalu w zł (x) względem liczby produkowanych detali w tys. sztuk (y), otrzymując m.in.: $\bar{x} = 10$, $s(x) = 3$, $\bar{y} = 12$, $V(y) = 30\%$, $a_1 = -720$. W jakim stopniu zmienność kosztów produkcji jest wyjaśniona zmianami wielkości produkcji?

6.81. Na podstawie następujących danych: $s(x) = 12$, $s(y) = 16$, $a_1 = 0{,}95$, oblicz współczynnik determinacji liniowej.

6.82. Badając zależność między stażem pracy w latach (y) a wydajnością pracy w sztukach (x) uzyskano następujące informacje: $\hat{y}_i = 0{,}3x_i + 13$, $\bar{x} = 18$, $r_{xy}^2 = 0{,}81$. Wyznacz równanie regresji wydajności pracy względem stażu pracy.

6.83. W badaniach nad korelacją między wydatkami rodzin na artykuł A (x) a ich miesięcznymi dochodami (y) otrzymano następujące równanie regresji: $\hat{y}_i = 14 - 2{,}5x_i$ oraz współczynnik korelacji $r_{xy} = 0{,}64$. Czy takie wyniki są możliwe? Odpowiedź uzasadnić.

6.84. Przy współczynniku ufności 0,89 oszacować – metodą przedziałową – współczynnik regresji zużycia surowca stosowanego w produkcji pewnego wyrobu (y) względem wielkości produkcji (x), jeśli na podstawie wyników 7-elementowej próby losowej otrzymano następujące wyniki: $s(u) = 2$, $\hat{y}_i = 7{,}4 + 2{,}1x_i$, $\sum_{i=1}^{n}(x_i - \bar{x})^2 = 28$.

6.85. Na podstawie 15-elementowej próby losowej obliczono $r_S = 0{,}6$ między wielkością produkcji a stopniem automatyzacji procesu wytwarzania w przedsiębiorstwie Z. Na poziomie istotności 0,05 zweryfikować hipotezę o istotności otrzymanego współczynnika korelacji rang Spearmana.

6.86. Ze zbiorowości uczniów szkół średnich wylosowano niezależnie 200 uczniów i zbadano ich pod względem wysokości miesięcznych dochodów rodziców i wysokości tygodniowego „kieszonkowego". Obliczony współczynnik korelacji liniowej Pearsona między tymi cechami wyniósł: $r_{xy} = 0{,}62$. Przy współczynniku ufności 0,95 zbudować przedział ufności dla współczynnika korelacji w populacji generalnej.

6.87. Badając zależność między wagą samochodu ciężarowego wraz z ładunkiem a zużyciem paliwa, otrzymano – dla losowej próby 20 przewozów – współczynnik korelacji liniowej $r_{xy} = 0{,}42$. Na poziomie istotności 0,1 zweryfikować hipotezę o braku korelacji między zużyciem paliwa a ciężarem samochodów wraz z ładunkiem w całej populacji przewozów.

6.88. Dla stu wylosowanych przedsiębiorstw handlowych wyznaczono liniową funkcję regresji sprzedaży (w tys. zł) względem kosztów handlowych (w tys. zł): $\hat{x}_j = 13{,}5 + 14{,}1y_j$. Standardowy błąd oceny współczynnika regresji wyniósł 0,17. Przyjmując współczynnik ufności 0,95, zbudować przedział ufności dla współczynnika regresji sprzedaży względem kosztów handlowych dla ogółu przedsiębiorstw.

6.89. W pewnym przedsiębiorstwie zbadano zależność między zarobkami pracowników (y) a wydajnością pracy (x). Dla losowo wybranej 400-osobowej próby oszacowano dwa równania regresji: $\hat{x}_i = 2 + 5y_i$ oraz $\hat{y}_i = 0{,}05x_i + 3{,}4$. Przyjmując współczynnik ufności 0,95, zbudować przedział ufności dla współczynnika korelacji liniowej między wyróżnionymi cechami w populacji generalnej.

6.90. Na podstawie 10 wylosowanych rodzin zbadano zależność między dochodami miesięcznymi w tys. zł a liczbą pracujących członków rodziny. Otrzymano następującą funkcję regresji: $\hat{y}_i = 0{,}68 + 0{,}4628x_i$. Wiadomo ponadto, że odchylenie standardowe reszt tej funkcji regresji wynosi 0,56, a suma kwadratów odchyleń zmiennej niezależnej od jej średniej arytmetycznej jest równa 18,56. Na poziomie istotności 0,05 zweryfikować hipotezę, że współczynnik regresji jest statystycznie istotny.

6.91. Wylosowano grupę 82 bezrobotnych zarejestrowanych w pewnym urzędzie pracy w celu oceny istotności związku między ich wykształceniem (x) a liczbą dzieci na utrzymaniu (y). Z uzyskanych informacji utworzono tablicę korelacyjną o wymiarach 5×5. Z tablicy tej obliczono stosunek korelacyjny: $e_{yx} = 0{,}222$. Na poziomie istotności 0,05, zweryfikować hipotezę o braku korelacji między badanymi zmiennymi.

6.92. W 187 wylosowanych niezależnie rodzinach zbadano zależność pomiędzy wydatkami ogółem (zmienna objaśniana), wydatkami na żywność (pierwsza zmienna objaśniająca) oraz liczbą osób w rodzinie (druga zmienna objaśniająca). Obliczony współczynnik korelacji pomiędzy tymi zmiennymi wyniósł: $R_{1\cdot23}$. Na poziomie istotności 0,05 zweryfikuj hipotezę o istotności tego współczynnika.

6.93. Wyniki badania zależności między średnim czasem obsługi klienta w minutach (x) a powierzchnią sprzedaży w m^2 (y) w 12 losowo wybranych sklepach aproksymowano do postaci liniowej funkcji regresji: $\hat{x}_j = -0{,}5y_j + 3$. Poziomy składnika resztowego modelu regresji kształtują się następująco: –0,2; 0; 0,01; 0,1; –0,05; 0,04; 0; 0,15; –0,03; –0,02; 0,08; –0,08. Zbadać losowość składnika resztowego modelu (przyjąć poziom istotności 0,1). Oszacować przedziałowo, o ile przeciętnie zmniejsza się czas obsługi klienta we wszystkich sklepach przy wzroście powierzchni sprzedaży o 1 m^2 (średni błąd szacunku współczynnika regresji x względem y wynosi 0,1).

6.94. Ze zbiorowości generalnej pobrano próbę losową $n = 150$ studentów. Na próbie tej przeprowadzono badanie dotyczące związku między ocenami na świadectwie maturalnym a wynikami egzaminów wstępnych do szkół wyższych. Z uzyskanych informacji zbudowano tablicę korelacyjną o wymiarach 7×9 (7 wariantów dla cechy x i 9 wariantów dla cechy y). Na podstawie tej tablicy obliczono m.in.: $r_{xy} = 0{,}763$ oraz $e_{yx} = 0{,}838$. Na poziomie istotności 0,05, zweryfikować hipotezę o liniowości regresji.

6.95. Oszacowane na podstawie wyników 7-elementowej próby losowej równanie regresji zużycia surowca w produkcji pewnego detalu względem wielkości produkcji przyjęło postać: $\hat{y}_i = 7{,}4 + 2{,}1x_i$. Wiadomo ponadto, że odchylenie standardowe składnika resztowego tej funkcji wynosi 2, a suma kwadratów odchyleń empirycznych wartości zmiennej niezależnej od jej średniego poziomu jest równa 28. Przyjmując współczynnik ufności 0,98, oszacować – metodą przedziałową – współczynnik regresji zużycia surowca względem wielkości produkcji.

6.96. Na podstawie 9 wylosowanych państw UE ustalono zależność między liczbą oddziałów banków w tys. (x) a liczbą pracowników w bankach w tys. (y): $\hat{y}_i = 47{,}976 + 11{,}923x_i$, $\varphi^2 = 0{,}23$. Na poziomie istotności 0,05 zweryfikować hipotezę o istotnie dodatniej zależności między badanymi cechami. Obliczyć teoretyczną liczbę pracowników w bankach, w których bylo 48,1 tys. oddziałów.

6.97. Z dwuwymiarowej populacji normalnej (X,Y) wylosowano 15-elementową próbę losową i na jej podstawie oszacowano liniową funkcję regresji o postaci:

$$\begin{array}{ccccc} \hat{y}_i & = & 63{,}8 & + & 2{,}28x_i \\ & & (10) & & (0{,}44) \end{array}.$$

Na poziomie istotności 0,05 zweryfikować hipotezę o istotności współczynnika regresji tej funkcji.

6.98. W modelu regresji liniowej opisującym zależność wysokości miesięcznych opłat za energię elektryczną w zł (y) od liczby osób w gospodarstwie domowym (x) dla próby losowej 100 rodzin otrzymano: $\operatorname{cov}(x,y) = 11{,}4$, $\bar{x} = 4$, $s(x) = 0{,}9$, $\bar{y} = 78{,}5$, $s(y) = 17{,}2$. Sprawdzić istotność współczynnika korelacji liniowej ($\alpha = 0{,}01$).

6.99. Badając zależność między dwoma cechami, oszacowano następującą funkcję regresji: $\hat{y}_i = 2{,}14x_i + 7{,}4$. Funkcję tę oszacowano na podstawie poniższych informacji:

x_i	1	2	3	4	5	6	7
y_i	8	13	14	17	18	20	23

Na poziomie istotności $\alpha = 0{,}05$ sprawdzić losowość odchyleń wartości empirycznych od teoretycznych.

6.100. Czy na poziomie istotności 0,001 można odrzucić hipotezę o braku związku między sposobem operacji usunięcia woreczka żółciowego a zadowoleniem pacjenta z operacji, jeśli dla tablicy niezależności o wymiarach 2×3 dla 200 wybranych losowo pacjentów obliczona wartość statystyki chi-kwadrat wyniosła 23,01?

Rozdział VII

ANALIZA SZEREGÓW CZASOWYCH

7.1. Pojęcie, rodzaje i składowe szeregu czasowego

Szeregi czasowe (dynamiczne, chronologiczne) służą do badania dynamiki zjawisk masowych. Szeregiem czasowym nazywamy uporządkowany chronologicznie ciąg wartości badanego zjawiska obserwowanego w kolejnych okresach lub momentach czasu. W szeregach czasowych zmienną niezależną jest czas (t), zmienną zależną (y_t) – wartości charakteryzujące określone zjawisko. Można to zapisać jako:

$$y_t = f(t). \tag{7.1}$$

W zależności od celu badania i właściwości zjawisk masowych, zmienna niezależna w szeregach czasowych może dotyczyć ściśle określonego momentu (np. według stanu w dniu czy nawet godzinie) lub też pewnych – dłuższych bądź krótszych – okresów (np. lat, kwartałów, półroczy). W związku z tym wyróżnia się szeregi czasowe **momentów** i szeregi czasowe **okresów**. Szeregi czasowe momentów dotyczą **zasobów** (np. liczba ludności na terenie województwa, stan zatrudnienia w firmie), a szeregi czasowe okresów – **strumieni** (np. liczba urodzeń w roku, liczba mieszkań oddanych do użytku w ostatnim kwartale). Tak więc szeregi czasowe momentów zawierają informację o poziomie określonego zjawiska w chwili dokonywania obserwacji. Informacji prezentowanych w tego rodzaju szeregach nie można agregować. Szeregi czasowe okresów zawierają dane dotyczące kształtowania się poziomu danego zjawiska w całym okresie przyjętym za jednostkę czasu. Istnieje tutaj możliwość wydłużania okresów (np. przechodzenia z danych miesięcznych na kwartalne czy też roczne). Dokonuje się tego za pomocą agregacji (sumowania) danych.

Przeciętny poziom zjawisk prezentowanych w postaci szeregów czasowych momentów oblicza się za pomocą **średniej chronologicznej**:

$$y_{ch} = \frac{\frac{y_1 + y_2}{2} + \frac{y_2 + y_3}{2} + \ldots + \frac{y_{n-1} + y_n}{2}}{n-1} = \frac{0{,}5y_1 + y_2 + y_3 + \ldots + 0{,}5y_n}{n-1}, \tag{7.2}$$

gdzie $y_1, y_2, \ldots, y_n$ oznaczają poziomy badanego zjawiska w kolejnych momentach.

Przeciętny poziom zjawisk przedstawionych za pomocą szeregu czasowego okresów wyznacza się za pomocą **średniej arytmetycznej**:

$$\bar{y} = \frac{1}{n}\sum_{t=1}^{n} y_t. \tag{7.3}$$

Rozwój zjawisk przedstawiony za pomocą szeregów czasowych może być właściwie oceniony wówczas, gdy poszczególne wyrazy szeregów czasowych są wielkościami **jednorodnymi** i **porównywalnymi**. Nie można np. porównywać szeregów czasowych okresów z szeregami czasowymi momentów czy szeregów wyrażonych w różnych jednostkach miary (np. w cenach bieżących i w cenach stałych).

Poziom zjawiska w określonym czasie (okresie lub momencie) jest skutkiem oddziaływania zespołu przyczyn, a mianowicie:

1) przyczyn głównych, które na badane zjawisko działają stale z niezmiennym nasileniem, wytyczając ogólny kierunek zmian zjawiska w czasie zwany **trendem (tendencją rozwojową)**,

2) przyczyn okresowych typu:

– **koniunkturalnego** (cyklicznego) o zmiennym kierunku i nasileniu. Przejawiają się one w dłuższych (mierzonych w latach) okresach i związane są ze zmianami w otoczeniu danego zjawiska (np. w gospodarce światowej). Tego rodzaju wahania noszą nazwę cyklu koniunkturalnego.

– **sezonowego**, które działają regularnie w krótkich rocznych cyklach wahań. Wahania sezonowe odzwierciedlają wpływ zachowań ludzi wynikających z warunków atmosferycznych i kalendarza (pór dnia i roku, świąt itp.) na kształtowanie się zjawisk gospodarczych;

3) przyczyn **przypadkowych (losowych)**. Wywołują one nieregularne odchylenia wielkości zjawiska od poziomu, jakiego oczekujemy na podstawie działania innych czynników (przyczyn głównych, wahań cyklicznych i sezonowych). Ich wpływ na poziom zjawiska jest nieprzewidywalny i to zarówno co do kierunku, jak i siły oddziaływania.

Zwrócić należy uwagę na to, że w konkretnym szeregu czasowym niekoniecznie muszą występować wszystkie składowe. W niektórych szeregach może nie być trendu (występuje wówczas stały, przeciętny poziom zmiennej) i wahań cyklicznych. W przypadku danych rocznych nie można mówić o wahaniach sezonowych. Jak się wydaje, jedynym składnikiem, który zawsze występuje w szeregach czasowych, są wahania przypadkowe (losowe).

Wyodrębnianie poszczególnych składowych szeregu czasowego nosi nazwę **dekompozycji szeregu czasowego**. Sposób dekompozycji zależy od natury danego szeregu czasowego, tj. od formy „nakładania się" poszczególnych składowych. Składowe te mogą łączyć się ze sobą poprzez **dodawanie** lub **mnożenie**. W związku z tym wyróżnia się dwa modele wahań w czasie:

addytywny i **multiplikatywny**. Ideowy schemat modelu **addytywnego** jest następujący:

$$Y = T + S + P, \tag{7.4}$$

gdzie: Y oznacza poziom badanego zjawiska, T – trend (tendencja rozwojowa), S – wahania sezonowe, P – wahania przypadkowe. Podkreślić należy, że w schemacie (7.4) symbole T, S, P nie traktuje się jako zmiennych, ale jako ogólne oznaczenia pojęć. W schemacie tym pominięto wahania koniunkturalne (cykliczne) jako niezwykle trudne do wyodrębnienia.

Multiplikatywny model wahań w czasie przyjmuje postać:

$$Y = T \cdot S \cdot P, \tag{7.5}$$

gdzie oznaczenia są takie same jak we wzorze (7.4).

W addytywnym modelu wahań w czasie zakłada się, że funkcja trendu jest liniowa (lub istnieje możliwość sprowadzenia jej do takiej postaci). Ponadto przyjmuje się, że między poszczególnymi składowymi szeregu w tym modelu nie występują interakcje, tzn. składowe te są niezależne. Poszczególne składniki szeregu czasowego w modelu addytywnym są wyrażane jako wielkości absolutne (posiadające miano).

W modelu multiplikatywnym trend jest wyrażany w takich samych jednostkach jak badane zjawisko, pozostałe składowe są wielkościami względnymi (wskaźnikami). W ten sposób w modelu tym wahania sezonowe i przypadkowe są proporcjonalne do wielkości trendu.

W badaniach empirycznych dysponujemy informacjami pochodzącymi z pewnego wycinka czasowego od $t = 1$ do $t = n$. Chcąc dokonać dekompozycji szeregu czasowego, modele wahań w czasie zapisujemy następująco:

$$y_t = f(t) + g_{it} + z_{t(a)}, \tag{7.6}$$

$$y_t = f(t) \cdot o_{it} \cdot z_{t(m)}, \tag{7.7}$$

gdzie: y_t – poziom badanego zjawiska zaobserwowany w okresie lub momencie t, $f(t)$ – funkcja trendu (tendencji rozwojowej), g_{it} – absolutne poziomy wahań sezonowych, o_{it} – wskaźniki sezonowości, $z_{t(a)}$ i $z_{t(m)}$ – addytywne i multiplikatywne reszty modeli (reprezentują one wahania przypadkowe).

7.2. Metody wyodrębniania trendu

Trendem (tendencją rozwojową) nazywamy powolne, regularne i systematyczne zmiany określonego zjawiska, obserwowane w dostatecznie długim przedziale czasowym i będące rezultatem działania przyczyn głównych.

Do wyodrębniania trendu z szeregów czasowych najczęściej wykorzystuje się **mechaniczną** metodę opartą o **średnie ruchome (kroczące)** oraz **analityczną** metodę najmniejszych kwadratów.

Mechaniczna metoda wyrównywania szeregów czasowych (wyodrębniania trendu) polega na zastępowaniu danych empirycznych średnimi poziomami z okresu badanego i kilku okresów sąsiednich. Te średnie poziomy nazywamy **średnimi ruchomymi**.

Do wygładzania szeregu czasowego za pomocą średnich ruchomych stosuje się dwa rodzaje średnich – **zwykłe** i **scentrowane**. Średnie ruchome zwykłe stosuje się do danych rocznych lub do innych szeregów chronologicznych, w których nie występują wahania sezonowe. Tak więc średnie ruchome zwykłe są wykorzystywane do wyrównywania szeregów zawierających trend i wahania przypadkowe.

Jeśli kolejne wartości szeregu empirycznego oznaczymy przez $y_1, y_2, y_3, \ldots, y_{n-2}, y_{n-1}, y_n$, to średnie ruchome zwykłe trzyokresowe ($k = 3$) obliczamy w sposób następujący:

$$\bar{y}_2 = \frac{y_1 + y_2 + y_3}{3},$$
$$\bar{y}_3 = \frac{y_2 + y_3 + y_4}{3}, \qquad (7.8)$$
$$\ldots\ldots\ldots\ldots$$
$$\bar{y}_{n-1} = \frac{y_{n-2} + y_{n-1} + y_n}{3}.$$

Ciąg średnich ruchomych $\bar{y}_2, \bar{y}_3, \ldots, \bar{y}_{n-1}$ jest nowym, wygładzonym (i krótszym o dwa wyrazy: jeden na początku i jeden na końcu) szeregiem reprezentującym tendencję rozwojową zjawiska.

Średnie ruchome zwykłe są obliczane z nieparzystej liczby wyrazów szeregu. Tylko wtedy możliwe jest przyporządkowanie obliczonej średniej konkretnemu okresowi (całkowitej wartości t). Dzięki temu można porównywać pierwotne (empiryczne) wartości szeregu czasowego z wartościami szeregu wygładzonego (wyrównanego).

Rozważmy przypadek szeregu czasowego dotyczącego sprzedaży artykułów gospodarstwa domowego w hurtowni X w latach 1998–2005 (tab. 7.1)

Tab. 7.1. Sprzedaż artykułów gospodarstwa domowego w hurtowni X w latach 1998–2005 (w mln zł)

Lata	t	Sprzedaż (y_t)	Średnie ruchome 3-okresowe
1998	1	18	–
1999	2	21	$\bar{y}_2 = (18 + 21 + 22):3 = 20{,}3$
2000	3	22	$\bar{y}_3 = (21 + 22 + 24):3 = 22{,}3$
2001	4	24	$\bar{y}_4 = (22 + 24 + 26):3 = 24{,}0$
2002	5	26	$\bar{y}_5 = (24 + 26 + 25):3 = 25{,}0$
2003	6	25	$\bar{y}_6 = (26 + 25 + 28):3 = 26{,}3$
2004	7	28	$\bar{y}_7 = (25 + 28 + 27):3 = 26{,}7$
2005	8	27	–

Źródło: dane umowne; obliczenia własne.

Średnie ruchome scentrowane są obliczane za pomocą średniej chronologicznej i wykorzystywane do eliminacji wahań sezonowych. Średnie ruchome scentrowane obliczane są z liczby wyrazów równej długości cyklu wahań sezonowych. Aby zatem wyeliminować wahania sezonowe z szeregu danych miesięcznych, należy użyć 12-okresowych średnich ruchomych scentrowanych, w przypadku danych kwartalnych – średnich czterookresowych, a w przypadku danych półrocznych – średnich dwuokresowych.

Technikę obliczania średnich ruchomych scentrowanych dla danych kwartalnych zaprezentowano w tab. 7.2.

Tab. 7.2. Transakcje kupna-sprzedaży nieruchomości w mieście L w latach 2003–2005

Lata	kwartały	t	Liczba transakcji (y_t)	Czterookresowe średnie ruchome scentrowane
2003	I	1	29	–
	II	2	30	–
	III	3	26	(0,5·29+30+26+30+0,5·32):4=29,125
	IV	4	30	(0,5·30+26+30·32+0,5·34):4=30,000
2004	I	5	32	(0,5·26+30+32+35+0,5·28):4=30,750
	II	6	34	(0,5·30+32+34+28+0,5·37):4=31,875
	III	7	28	(0,5·32+34+28+37+0,5·36):4=33,250
	IV	8	37	(0,5·34+28+37+36·0,5·33):4=33,625
2005	I	9	36	(0,5·28+37·36+33+0,5·30):4=33,750
	II	10	33	(0,5·37+36+33+30+0,5·42):4=34,625
	III	11	30	–
	IV	12	42	–

Źródło: dane umowne. Obliczenia własne.

Jak wynika z tab. 7.2, szereg wygładzony jest krótszy – w stosunku do wyjściowego – o cztery wyrazy. Szereg wygładzony jest jednak wolny od wahań sezonowych i przypadkowych i prezentuje wyłącznie tendencję rozwojową analizowanego zjawiska. Szereg wygładzony charakteryzuje się mniejszym – w porównaniu do szeregu wyjściowego – obszarem zmienności (tj. różnicą między najmniejszą i największą wartością).

Wykorzystanie metody mechanicznej jest przydatne zwłaszcza wtedy, gdy nie ma możliwości doboru dostatecznie dokładnej analitycznej postaci funkcji trendu.

Analityczna metoda wyodrębniania tendencji rozwojowej polega na dopasowaniu określonej funkcji matematycznej do całego szeregu czasowego. Istotnym zagadnieniem w tej metodzie jest wybór odpowiedniej klasy funkcji trendu oraz prawidłowe oszacowanie jej parametrów. Zaleca się, aby wybrany typ funkcji trendu był możliwie prosty, a parametry funkcji miały merytoryczne znaczenie w opisie rozwoju badanego zjawiska. Sposób pos-

tępowania jest tu zbieżny z regułami wyznaczania postaci funkcji regresji metodą najmniejszych kwadratów, przy założeniu, że czas jest nielosową zmienną niezależną, determinującą przebieg zjawiska.

Do najczęściej stosowanych funkcji trendu należy funkcja liniowa o postaci:

$$Y_t = \alpha_0 + \alpha_1 t + \xi_t, \tag{7.9}$$

gdzie: Y_t – zmienna zależna mierząca poziom badanego zjawiska w okresie (momencie) t, t – zmienna czasowa, α_0, α_1 – nieznane parametry strukturalne funkcji trendu, ξ_t – składnik losowy dla okresu (momentu) t.

Z zapisu modelu (7.9) wynika, że uwzględnia on dwa składniki: trend (reprezentowany przez liniową funkcję trendu) oraz wahania przypadkowe, których odzwierciedleniem są zmienne losowe ξ_t.

Zakładając, że rozkład ξ_t spełnia warunki: $E(\xi_t) = 0$, $D^2(\xi_t) = \sigma^2$, $\operatorname{cov}(\xi_s, \xi_t) = 0$ dla $s \neq t$, uzyskujemy podstawę do szacowania parametrów strukturalnych funkcji trendu za pomocą klasycznej metody najmniejszych kwadratów.

Funkcja trendu (7.9) jest **funkcją trendu I rodzaju** (właściwą dla populacji generalnej) o numeracji okresów (momentów) t od $-\infty$ do $+\infty$. Empiryczny szereg czasowy traktujemy jako próbę o n okresach obserwacji rozpatrywanego zjawiska ($t = 1, 2, \ldots, n$ lub $t = 0, 1, \ldots, n-1$). Otrzymana z wyników próby losowej funkcja trendu jest **aproksymantą** (przybliżeniem) funkcji trendu I rodzaju. Przyjmuje ona postać:

$$\hat{y}_t = a_0 + a_1 t + z_t, \tag{7.10}$$

gdzie: $\hat{y}_t$ – teoretyczne wartości trendu w okresie t wynikające z danej funkcji, a_0 i a_1 – oszacowania parametrów α_0 i α_1, z_t – składnik resztowy.

Funkcja trendu (7.10) nosi nazwę **funkcji trendu II rodzaju**. Wykorzystując metodę najmniejszych kwadratów do szacowania parametrów strukturalnych liniowej funkcji trendu otrzymujemy następujący układ równań normalnych:

$$\begin{cases} \sum_{t=1}^{n} y_t = n a_0 + a_1 \sum_{t=1}^{n} t \\ \sum_{t=1}^{n} y_t t = a_0 \sum_{t=1}^{n} t + a_1 \sum_{t=1}^{n} t^2 \end{cases}. \tag{7.11}$$

Rozwiązując układ (7.11) otrzymujemy:

$$a_0 = \bar{y} - a_1 \bar{t} \tag{7.12}$$

oraz

$$a_1 = \frac{\sum_{t=1}^{n} (t - \bar{t})(y_t - \bar{y})}{\sum_{t=1}^{n} (t - \bar{t})^2}, \tag{7.13}$$

gdzie: $\bar{t}$ i $\bar{y}$ – średnie arytmetyczne zmiennych t i y_t.

Procedura wyznaczania ocen a_0 i a_1 znacznie uprości się, jeśli jednostki czasu tak ponumerujemy, aby ich suma wynosiła zero. W takim przypadku układ równań (7.11) przyjmuje postać uproszczoną:

$$\begin{cases} \sum_{t=1}^{n} y_t = n a_0 \\ \sum_{t=1}^{n} y_t t' = a_1 \sum_{t=1}^{n} (t')^2 \end{cases}. \tag{7.14}$$

Z układu równań (7.14) wynika, że:

$$a_0 = \frac{\sum_{t=1}^{n} y_t}{n} = \bar{y} \tag{7.15}$$

oraz

$$a_1 = \frac{\sum_{t=1}^{n} y_t t'}{\sum_{t=1}^{n} (t')^2}, \tag{7.16}$$

gdzie $t' = t - \bar{t}$.

W celu dokonania oceny oszacowania modelu trendu (7.10) stosujemy takie same mierniki jak w przypadku regresji liniowej (por. rozdział VI). Punktem wyjścia tej oceny są reszty określone wzorem:

$$z_t = y_t - \hat{y}_t \ (t = 1, 2, \ldots, n). \tag{7.17}$$

Miernikami „dobroci" dopasowania funkcji trendu do danych empirycznych są:

1) **odchylenie standardowe składnika resztowego**:

$$s(z_t) = \sqrt{\frac{1}{n-k} \sum_{t=1}^{n} (y_t - \hat{y}_t)^2} = \sqrt{\frac{1}{n-k} \sum_{t=1}^{n} z_t^2}. \tag{7.18}$$

Wartość odchylenia standardowego składnika resztowego informuje o tym, o ile średnio odchylają się poszczególne wartości zmiennej y_t od wartości teoretycznych otrzymanych z oszacowanej funkcji trendu.

2) **współczynnik zmienności resztowej**:

$$V(z_t) = \frac{s(z_t)}{\bar{y}} \cdot 100. \tag{7.19}$$

Wartość tego współczynnika informuje o tym, jaki procent średniego poziomu zmiennej objaśnianej (zależnej) y_t stanowią – przeciętnie rzecz biorąc – odchylenia przypadkowe w danym równaniu trendu. Ze statystycznego punktu widzenia sytuacja jest tym lepsza, im współczynnik zmienności resztowej (a więc również odchylenie standardowe składnika resztowego) jest bliższy zeru.

3) **współczynnik zbieżności** (indeterminacji, nieokreśloności):

$$\varphi^2 = \frac{\sum_{t=1}^{n}(y_t - \hat{y}_t)^2}{\sum_{t=1}^{n}(y_t - \bar{y})^2} = \frac{(n-k)s^2(z_t)}{\sum_{t=1}^{n}(y_t - \bar{y})^2}. \quad (7.20)$$

Współczynnik zbieżności mierzy tę część całkowitej zaobserwowanej zmienności zmiennej zależnej y_t, która wynika z działania czynników przypadkowych (nie została wyjaśniona przez oszacowaną funkcję trendu).

4) **współczynnik determinacji** R^2:

$$R^2 = 1 - \varphi^2. \quad (7.21)$$

Współczynnik ten informuje więc o tym, jaka część całkowitej zaobserwowanej zmienności zmiennej zależnej y_t została wyjaśniona przez oszacowaną funkcję trendu.

Wartości współczynników: R^2 i φ^2 zawierają się w przedziale domkniętym od 0 do 1, przy czym $\varphi^2 = 0$, gdy $R^2 = 1$ i $\varphi^2 = 1$, jeśli $R^2 = 0$. Ogólnie można zatem stwierdzić, że funkcja trendu tym lepiej jest dopasowana do danych empirycznych, im bardziej φ^2 zbliża się do zera (a R^2 do 1), a tym gorzej – im bardziej φ^2 zbliża się do 1 (a R^2 do 0).

5) **średnie błędy szacunku parametrów** α_0 i α_1 liniowej funkcji trendu:

$$D(a_0) = \sqrt{\frac{s^2(z_t)\sum_{t=1}^{n} t^2}{n\left(\sum_{t=1}^{n} t^2 - n\bar{t}^2\right)}}, \quad (7.22)$$

$$D(a_1) = \frac{s(z_t)}{\sqrt{\sum_{t=1}^{n} t^2 - n\bar{t}^2}}, \quad (7.23)$$

gdzie $s^2(z_t)$ oraz $s(z_t)$ są odpowiednio wariancją i odchyleniem standardowym składnik resztowego.

Błędy średnie szacunku informują o tym, jaki – średnio rzecz biorąc – popełnia się błąd szacując wartość parametru α_0 bądź α_1 w kolejnych próbach, z których każda składa się z n elementów.

Estymacji liniowej funkcji trendu oraz sprawdzenia dobroci jej dopasowania do danych empirycznych dokonamy na podstawie danych zawartych w tab. 7.3.

Tab. 7.3. Kredyty na cele konsumpcyjne udzielone przez bank X w poszczególnych miesiącach 2005 r. (w tys. zł)

Miesiące	I	II	III	IV	V	VI	VII	VIII	IX	X	XI	XII
Wysokość kredytu	41,2	42,1	43,3	43,7	44,1	44,6	45,6	46,3	46,9	47,7	49,0	50,1

Źródło: dane umowne.

Do szacowania parametrów strukturalnych liniowej funkcji trendu wykorzystujemy układ równań normalnych (7.11). Niezbędne obliczenia zawarte są w tab. 7.4.

Tab. 7.4. Obliczenia pomocnicze do szacowania parametrów liniowej funkcji trendu

t	y_t	$y_t t$	t^2	$\hat{y}_t$	$y_t - \hat{y}_t$	$(y_t - \hat{y}_t)^2$	t'	$y_t t'$	$(t')^2$	$y_t - \bar{y}$	$(y_t - \bar{y})^2$
1	41,2	41,2	1	41,26	–0,06	0,0036	–5,5	–226,60	30,25	–4,18	17,4724
2	42,1	84,2	4	42,01	0,09	0,0081	–4,5	–189,45	20,25	–3,28	10,7584
3	43,3	129,9	9	42,76	0,54	0,2916	–3,5	–151,55	12,25	–2,08	4,3264
4	43,7	174,6	16	43,51	0,19	0,0361	–2,5	–109,25	6,25	–1,68	2,8224
5	44,1	220,5	25	44,26	–0,16	0,0256	–1,5	–66,15	2,25	–1,28	1,6384
6	44,6	267,6	36	45,01	–0,41	0,1681	–0,5	–22,30	0,25	–0,78	0,6084
7	45,6	319,2	49	45,76	–0,16	0,0256	0,5	22,80	0,25	0,22	0,0484
8	46,3	370,4	64	46,51	–0,21	0,0441	1,5	69,45	2,25	0,92	0,8464
9	46,9	422,1	81	47,26	–0,36	0,1296	2,5	117,25	6,25	1,52	2,3104
10	47,7	477,0	100	48,01	–0,31	0,0961	3,5	166,95	12,25	2,32	5,3824
11	49,0	539,0	121	48,76	0,24	0,0576	4,5	220,50	20,25	3,62	13,1044
12	50,1	601,2	144	49,51	0,59	0,3481	5,5	275,55	30,25	4,72	22,2784
78	**544,6**	**3646,9**	**650**	X	X	**1,2342**	**0**	**107,20**	**143,00**	X	**81,5968**

Źródło: obliczenia własne.

Podstawiając odpowiednie wartości z tab. 7.4 do układu równań (7.11) otrzymujemy:

$$\begin{cases} 544{,}6 = 12a_0 + 78a_1 \\ 3646{,}9 = 78a_0 + 650a_1 \end{cases}.$$

Rozwiązaniem tego układu są wartości: $a_0 = 40{,}51$ oraz $a_1 = 0{,}75$.

Funkcja trendu kredytów udzielanych na cele konsumpcyjne jest zatem następująca:

$$\hat{y}_t = 40{,}51 + 0{,}75t. \quad (7.24)$$

Z funkcji (7.24) wynika, że wysokość udzielanych kredytów wzrastała z miesiąca na miesiąc średnio o 0,75 tys. zł. Wyraz wolny funkcji trendu ($a_0 = 40{,}51$ tys. zł) informuje o teoretycznym poziomie kredytów w okresie $t = 0$, tj. w grudniu 2004 r.

Funkcję trendu można również wyznaczyć numerując jednostki czasu w taki sposób, aby ich suma wynosiła zero ($\sum t = 0$). W tym przypadku oceny parametrów strukturalnych trendu liniowego otrzymamy z układu równań (7.14). Podstawiając odpowiednie dane liczbowe z tablicy (7.4) otrzymujemy:

$$\begin{cases} 544{,}6 = 12a_0 \\ 107{,}20 = 143a_1 \end{cases}.$$

Stąd:

$$a_0 = \frac{544,6}{12} = 45,38 = \bar{y},$$

$$a_1 = \frac{107,20}{143} = 0,75.$$

Równanie tendencji rozwojowej przy spełnieniu warunku $\sum t = 0$ przyjmuje zatem postać:

$$\hat{y}_t = 45,38 + 0,75t. \tag{7.25}$$

Tym razem ocena $a_0 = 45,38$ jest interpretowana jako przeciętny miesięczny poziom udzielanych kredytów w analizowanym okresie. Ocena $a_1 = 0,75$ określa natomiast średni miesięczny przyrost udzielanych kredytów.

Z równania (7.25) można przejść na równanie określone wzorem (7.24), wykorzystując $\bar{t} = 6,5$:

$$\hat{y}_t = 45,38 + 0,75(t - 6,5) = 40,51 + 0,75t.$$

W toku dalszych obliczeń sprawdzamy, czy oszacowana liniowa funkcja trendu jest dobrze dopasowana do danych empirycznych.

Odchylenie standardowe składnika resztowego (wzór 7.18) wynosi (por. tab. 7.4):

$$s(z_t) = \sqrt{\frac{1,2342}{12-2}} = 0,35 \text{ tys. zł.}$$

Oznacza to, że empiryczne wartości udzielonych kredytów w poszczególnych miesiącach 2005 r. odchylają się od wartości teoretycznych wyznaczonych na podstawie równania trendu o postaci (7.24) średnio o ±0,36 tys. zł. Niewielka wartość odchylenia standardowego składnika resztowego świadczy o dobrym dopasowaniu funkcji trendu do danych empirycznych. Potwierdza to współczynnik zmienności resztowej obliczony za pomocą wzoru (7.19):

$$V(z_t) = \frac{0,35}{45,38} \cdot 100 = 0,8\%.$$

Wykorzystując obliczenia zawarte w tab. 7.4, obliczamy współczynnik zbieżności (7.20):

$$\varphi^2 = \frac{1,2342}{81,5968} = 0,015.$$

Równanie trendu (7.24) tylko w 1,5% nie wyjaśnia zmienności wysokości udzielonych kredytów na cele konsumpcyjne w poszczególnych miesiącach 2005 r.

Współczynnik determinacji R^2 (wzór 7.21) wynosi natomiast:

$$R^2 = 1 - 0,015 = 0,985.$$

Oznacza to, że zmienność poziomu udzielonych kredytów aż w 98,5% jest wyjaśniona przez funkcję trendu liniowego o postaci (7.24).

Średnie błędy szacunku (wzory (7.22) i (7.23)) są równe:

$$D(a_0) = \sqrt{\frac{0,35^2 \cdot 650}{12\left(650 - 12 \cdot 6,5^2\right)}} = 0,215,$$

$$D(a_1) = \frac{0,35}{\sqrt{650 - 12 \cdot 6,5^2}} = 0,029.$$

Średnie błędy szacunku parametrów strukturalnych liniowej funkcji trendu są zatem niewielkie.

Na podstawie obliczonych mierników można więc stwierdzić, że oszacowana funkcja trendu dobrze opisuje kształtowanie się poziomu udzielonych kredytów w poszczególnych miesiącach 2005 r. Wskazują na to niskie wartości średnich błędów szacunku parametrów strukturalnych, niewielkie wartości odchylenia standardowego składnika resztowego, współczynnika zmienności resztowej i współczynnika zbieżności oraz wysoka – zbliżona do jedności – wielkość współczynnika determinacji R^2.

Oszacowaną funkcję trendu zapisujemy w następujący sposób:

$$\begin{array}{ccccccc} \hat{y}_t & = & 40,51 & + & 0,75t & + & z_t \\ & & (0,215) & & (0,029) & & (0,35) \end{array} \tag{7.26}$$

gdzie w nawiasach podane są średnie błędy szacunku parametrów strukturalnych oraz odchylenie standardowe składnika resztowego.

7.3. Pomiar wahań sezonowych

Rozwój wielu zjawisk masowych charakteryzuje nierównomierność i zmienność w czasie. Mówimy wówczas, że zjawiska te podlegają wahaniom okresowym. Szczególnym przypadkiem wahań okresowych są **wahania sezonowe**. Pod pojęciem wahań sezonowych rozumie się, powtarzające się z roku na rok w tych samych jednostkach kalendarzowych, dość regularne zmiany ilościowe w rozmiarach przebiegu zjawisk masowych. Charakterystycznymi cechami wahań sezonowych są:

1) roczny cykl wahań, w ramach którego wyróżnia się zazwyczaj podokresy miesięczne ($d = 12$), kwartalne ($d = 4$) i półroczne ($d = 2$);

2) systematyczne powtarzanie się w każdym roku;

3) wskazywanie określonej regularności.

Wahaniami sezonowymi charakteryzuje się produkcja roślinna i zwierzęca, podaż surowców dla przemysłu rolno-spożywczego, ruch turystyczny, obroty w handlu itp. Wahania sezonowe wywierają – na ogół – niekorzystny wpływ na działalność zarówno całej gospodarki narodowej, jak i poszczególnych podmiotów. Powodują one odchylenia od rytmicznego przebiegu procesów gospodarczych, pociągając za sobą z jednej strony nadmierne koszty, z drugiej zaś – niepełne wykorzystanie mocy produkcyjnych w jednych okresach oraz przeciążenie w innych.

Regularność ilościowych zmian występujących w ramach cyklu wahań umożliwia wykrywanie prawidłowości w kształtowaniu się wahań sezonowych poprzez wykorzystanie metod statystycznych. Do tego celu wykorzystuje się **wskaźniki sezonowości** lub **absolutne poziomy wahań sezonowych**. Wskaźniki sezonowości są wyrażane w procentach, absolutne poziomy wahań sezonowych – w jednostkach mianowanych (bezwzględnych). W pierwszym przypadku mówimy o **wahaniach sezonowych multiplikatywnych**, w drugim zaś – o **addytywnych wahaniach sezonowych**. Wahania sezonowe multiplikatywne występują wówczas, gdy w poszczególnych podokresach cyklu (miesiącach, kwartalach, półroczach) badane zjawisko odchyla się od przeciętnego poziomu lub trendu o pewną stałą wielkość względną (np. poziom sprzedaży napojów chłodzących w lipcu jest większy od poziomu wynikającego z trendu o 5%). Wahania sezonowe addytywne oznaczają stałe – co do wartości bezwzględnej – odchylenia poziomu badanego zjawiska w poszczególnych podokresach cyklu sezonowości od przeciętnego poziomu lub trendu (np. spożycie piwa na jednego mieszkańca w III kwartale każdego roku jest wyższe od przeciętnego spożycia o 10 litrów).

Jeżeli w szeregu czasowym nie występuje trend, to do wyodrębniania wahań sezonowych wykorzystuje się metodę **średnich jednoimiennych okresów**. Wskaźniki sezonowości oblicza się wówczas według wzoru:

$$S_i = \frac{\bar{y}_i \cdot d}{\sum_{i=1}^{d} \bar{y}_i} \cdot 100, \tag{7.27}$$

gdzie: S_i – wskaźnik sezonowości dla i-tego podokresu cyklu sezonowości, $\bar{y}_i$ – średnia jednoimiennych okresów obliczona ze wszystkich badanych cykli rocznych, d – liczba podokresów cyklu sezonowości ($d = 12$ dla sezonowości miesięcznej, $d = 4$ dla sezonowości kwartalnej, $d = 2$ dla sezonowości półrocznej).

Obliczone za pomocą wzoru (7.27) wskaźniki sezonowości nazywamy **oczyszczonymi**, jeśli ich suma wynosi 1200 (w przypadku sezonowości miesięcznej), 400 (dla sezonowości kwartalnej) oraz 200 (w odniesieniu do sezonowości półrocznej). Wskaźniki sezonowości niespełniające powyższego warunku nazywamy **surowymi** (**nieoczyszczonymi**). W takim przypadku zachodzi konieczność wprowadzenia **współczynnika korygującego**, określonego wzorem:

$$k = d : \sum_{i=1}^{d} S_i. \tag{7.28}$$

Mnożąc współczynnik korygujący k przez kolejne nieoczyszczone wskaźniki sezonowości S_i, otrzymujemy skorygowane wskaźniki sezonowości ${}_kS_i$:

$${}_kS_i = k \cdot S_i. \tag{7.29}$$

Suma skorygowanych wskaźników sezonowości powinna wynosić odpowiednio 1200, 400 lub 200.

Absolutne poziomy wahań sezonowych dla poszczególnych podokresów cyklu sezonowości są ustalane następująco:

$$g_i = S_i \cdot \bar{y} - \bar{y} = \bar{y}(S_i - 1), \tag{7.30}$$

gdzie: S_i – oczyszczone wskaźniki sezonowości dla i-tego podokresu cyklu wahań, g_i – absolutne poziomy wahań sezonowych wyrażane w jednostkach absolutnych (w takich samych jak badane zjawisko), $\bar{y}$ – średni poziom badanego zjawiska w całym rozpatrywanym okresie.

Suma absolutnych poziomów wahań sezonowych jest równa zeru:

$$\sum_{i=1}^{d} g_i = 0. \tag{7.31}$$

Mówimy wówczas o **oczyszczonych absolutnych poziomach wahań sezonowych**. Jeśli relacja (7.31) nie jest spełniona – mamy do czynienia z **nieoczyszczonymi** absolutnymi poziomami wahań sezonowych. W celu sprowadzenia tego rodzaju absolutnych poziomów wahań do oczyszczonych wprowadzamy poprawkę:

$$p = \frac{\sum_{i=1}^{d} g_i'}{d}. \tag{7.32}$$

Odejmując od surowych (nieoczyszczonych) absolutnych poziomów wahań sezonowych, otrzymujemy oczyszczone mierniki:

$$g_i = g_i' - p, \tag{7.33}$$

gdzie g_i' – to nieoczyszczone absolutne poziomy wahań sezonowych.

Absolutne poziomy wahań sezonowych (o sumie zero) odpowiadają na pytanie, ile – w wyrażeniu bezwzględnym – wzrósł lub zmalał, przeciętnie rzecz biorąc, poziom badanego zjawiska w i-tym podokresie cyklu sezonowości tylko na skutek działania sezonowości.

Sposób obliczania wskaźników sezonowości i absolutnych poziomów wahań sezonowych przedstawimy na przykładzie dotyczącym wynajętych miejsc noclegowych przez właściciela pensjonatu w kolejnych kwartałach lat 2001–2005 (tab. 7.5).

Tab. 7.5. Liczba wynajętych miejsc noclegowych w latach 2001–2005

Kwartały	Lata					$\sum_{i=1}^{5} y_i$	$\bar{y}_i$	S_i (w %)
	2001	2002	2003	2004	2005			
I	1860	1853	1855	1861	1858	9287	1857,4	88,33
II	2200	2210	2254	2245	2213	11 122	2224,4	105,79
III	2415	2418	2423	2419	2420	12 095	2419,0	115,04
IV	1908	1912	1910	1913	1907	9550	1910,0	90,84
Razem	X	X	X	X	X	X	8410,8	400,00

Źródło: dane umowne. Obliczenia własne.

Kwartalne wskaźniki sezonowości zostały obliczone na podstawie wzoru (7.27). Przykładowo, wskaźnik sezonowości dla I kwartału wynosi:

$$S_i = \frac{1857{,}4 \cdot 4}{8410{,}8} \cdot 100 = 88{,}33\%.$$

Wyniki obliczeń wskaźników sezonowości dla kolejnych kwartałów zamieszczono w ostatniej kolumnie tab. 7.5.

Ze względu na to, że suma wskaźników sezonowości jest równa 400, nie ma potrzeby wprowadzania współczynnika korygującego. Przedstawione w tab. 7.5 wskaźniki sezonowości są zatem oczyszczonymi wskaźnikami sezonowości.

Z obliczonych wskaźników sezonowości wynika, że w badanym okresie (2001–2005) tylko na skutek działania sezonowości liczba wynajętych miejsc noclegowych była w pierwszym kwartale niższa od przeciętnej kwartalnej (równej 100%) o 11,67%, w II kwartale wyższa o 5,79%, w III kwartale wyższa o 15,4%, a w IV kwartale – niższa o 9,16%.

Absolutne poziomy wahań sezonowych – obliczone na podstawie wzoru (7.30) – są natomiast równe:

$$\bar{y} = \frac{1857{,}4 + 2224{,}4 + 2419{,}0 + 1910{,}0}{4} = 2102{,}7,$$

$$g_I = 2102{,}7(0{,}8833 - 1) = -245{,}38509 \approx -245 \text{ miejsc},$$

$$g_{II} = 2102{,}7(1{,}0579 - 1) = 121{,}74633 \approx 122 \text{ miejsca},$$

$$g_{III} = 2102{,}7(1{,}1504 - 1) = 316{,}24608 \approx 316 \text{ miejsc},$$

$$g_{IV} = 2102{,}7(0{,}9084 - 1) = -192{,}60732 \approx -193 \text{ miejsca}$$

$$\sum_{i=1}^{4} g_i = 0.$$

Otrzymane wyniki informują o tym, że w I kwartale liczba miejsc noclegowych była niższa od przeciętnej kwartalnej ($\bar{y} = 2102{,}7$) o 245, w II kwartale większa o 122 miejsca, w III kwartale wyższa o 316 miejsc, a w IV kwartale – niższa o 193 miejsca.

Jeśli w szeregu czasowym występuje tendencja rozwojowa (co zazwyczaj ujawnia się po naniesieniu danych na wykres), to pierwszą czynnością jest wyodrębnienie funkcji trendu. Wyznaczony trend stanowi punkt odniesienia przy obliczaniu mierników sezonowości. Kolejne działania zmierzające do wyznaczenia wskaźników sezonowości są następujące:

1) dla każdego okresu t obliczamy z funkcji trendu teoretyczne wartości badanego zjawiska ($\hat{y}_t$);

2) dla jednoimiennych okresów obliczamy wskaźniki sezonowości jako średnią z ilorazów zaobserwowanych w kolejnych cyklach rocznych (c):

$$S_i = \frac{1}{c} \sum_{c} \frac{y_{it}}{\hat{y}_{it}}, \qquad (7.34)$$

3) gdy suma wskaźników sezonowości nie spełnia warunku $\sum_{i=1}^{d} S_i = d$, obliczamy współczynnik korygujący:

$$k = \frac{d}{\sum_{i=1}^{d} S_i}, \qquad (7.35)$$

4) mnożąc nieoczyszczone (surowe) wskaźniki sezonowości przez współczynnik korygujący, otrzymujemy oczyszczone wskaźniki sezonowości:

$${}_k S_i = k \cdot S_i. \qquad (7.36)$$

Informacje liczbowe o kwartalnej sprzedaży wyrobu X w tys. zł w latach 2002–2005 przedstawiają się następująco (tab. 7.6):

Tab. 7.6. Sprzedaż wyrobu X w poszczególnych kwartałach w latach 2002–2005 (w tys. zł)

Lata	kwartały	Sprzedaż (y_t)	t	$y_t t$	t^2	$\hat{y}_t$	$\frac{y_t}{\hat{y}_t}$
2002	I	116,3	1	116,3	1	169,575	0,686
	II	158,8	2	317,6	4	192,565	0,825
	III	278,3	3	834,9	9	215,555	1,291
	IV	238,0	4	952,0	16	238,545	0,998
2003	I	221,9	5	1109,5	25	261,535	0,848
	II	293,2	6	1759,2	36	284,525	1,030
	III	486,1	7	3402,7	49	307,515	1,581
	IV	346,9	8	2775,2	64	330,505	1,050
2004	I	235,7	9	2121,3	81	353,495	0,667
	II	312,4	10	3124,0	100	376,485	0,830
	III	536,8	11	5904,8	121	399,475	1,344
	IV	360,3	12	4323,6	144	422,465	0,853
2005	I	331,2	13	4305,6	169	445,455	0,744
	II	437,4	14	6123,6	196	468,445	0,934
	III	740,4	15	11106,0	225	491,435	1,507
	IV	378,3	16	6052,8	256	514,425	0,735
Σ		5472,0	136	54 329,1	1496	X	X

Źródło: dane umowne. Obliczenia własne.

Podstawiając odpowiednie dane liczbowe zawarte w tab. 7.6 do układu równań normalnych (7.11) otrzymujemy:

$$\begin{cases} 5472 = 16a_0 + 136a_1 \\ 54329{,}1 = 136a_0 + 1496a_1 \end{cases}.$$

Z układu tego wyznaczamy: $a_0 = 146{,}585$ oraz $a_1 = 22{,}99$. Funkcja trendu liniowego przyjęła zatem postać:

$$\hat{y}_t = 146{,}585 + 22{,}99t \quad (\sum t \neq 0).$$

W IV kwartale 2001 r. sprzedaż wyrobu X wynosiła – teoretycznie – 146,585 tys. zł. Wielkość sprzedaży wyrobu wzrastała z kwartału na kwartał, średnio o 22,99 tys. zł.

W celu obliczenia wskaźników sezonowości należy – korzystając z oszacowanej funkcji trendu – wyliczyć teoretyczne wartości $\hat{y}_t$. Są one zawarte w tab. 7.6. Wskaźniki sezonowości dla jednoimiennych kwartałów obliczono zgodnie ze wzorem (7.34):

$$S_I = \frac{0{,}686 + 0{,}848 + 0{,}667 + 0{,}744}{4} = \frac{2{,}945}{4} = 0{,}73625,$$

$$S_{II} = \frac{0{,}825 + 1{,}030 + 0{,}830 + 0{,}934}{4} = \frac{3{,}619}{4} = 0{,}90475,$$

$$S_{III} = \frac{1{,}291 + 1{,}581 + 1{,}344 + 1{,}507}{4} = \frac{5{,}723}{4} = 1{,}43075,$$

$$S_{IV} = \frac{0{,}998 + 1{,}050 + 0{,}853 + 0{,}735}{4} = \frac{3{,}636}{4} = 0{,}9090,$$

$$\sum_{i=1}^{4} S_i = 3{,}98075.$$

Suma wskaźników sezonowości kwartalnej nie jest zgodna z wartością postulowaną $\left(\sum_{i=1}^{4} S_i = 4\right)$. Zachodzi zatem potrzeba wprowadzenia współczynnika korygującego (7.35):

$$k = \frac{4}{3{,}98075} = 1{,}0048357.$$

Mnożąc nieoczyszczone (surowe) wskaźniki sezonowości przez współczynnik korygujący k otrzymujemy (po zaokrągleniu): $S_I = 0{,}739$; $S_{II} = 0{,}909$; $S_{III} = 1{,}438$; $S_{IV} = 0{,}914$. Suma tych wskaźników sezonowości wynosi 4. Wskutek działania czynnika sezonowego sprzedaż w każdym I kwartale była mniejsza od sprzedaży wyznaczonej trendem liniowym przeciętnie o 26,1% (73,9% – 100%), w każdym drugim kwartale niższa o 9,1%, w każdym trzecim kwartale wyższa o 43,8% (143,8% – 100%), a w każdym IV kwartale – niższa o 8,6%.

W nieco odmienny sposób wyodrębnia się absolutne poziomy wahań sezonowych, które dotyczą addytywnego modelu wahań w czasie. Absolutne poziomy wahań sezonowych są wielkościami mianowanymi, obliczanymi zgodnie ze wzorem:

$$g_i = \frac{1}{c}\sum_{c}(y_{it} - \hat{y}_{it}), \qquad (7.37)$$

gdzie: y_{it} – empiryczne poziomy badanego zjawiska w jednostce czasu t ($t = 1, 2, ..., n$) tylko w sezonie i ($i = 1, 2, ..., d$), $\hat{y}_{it}$ – teoretyczne poziomy badanego zjawiska (wyznaczone z funkcji trendu), c – liczba lat z sezonem i ($i = 1, 2, ..., d$).

Sposób postępowania przy wyodrębnianiu absolutnych poziomów wahań sezonowych zaprezentujemy na przykładzie danych zawartych w tab. 7.7.

Tab. 7.7. Wypłaty odsetek od lokat terminowych według kwartałów w latach 2001–2005 (w tys. zł)

Lata kwartały		Odsetki (y_t)	t	$\hat{y}_t$	$y_t - \hat{y}_t$
2001	I	833	1	883,2	–50,2
	II	837	2	909,7	–72,7
	III	839	3	936,2	–97,2
	IV	910	4	962,7	–52,7
2002	I	891	5	989,2	–98,2
	II	986	6	1015,7	–29,7
	III	1053	7	1042,2	10,8
	IV	1256	8	1068,7	187,3
2003	I	1147	9	1095,2	51,8
	II	1135	10	1121,7	13,3
	III	1247	11	1148,2	98,8
	IV	1405	12	1174,7	230,3
2004	I	1319	13	1201,2	117,8
	II	1309	14	1227,7	81,3
	III	1270	15	1254,2	15,8
	IV	1293	16	1280,7	12,3
2005	I	1254	17	1307,2	–53,2
	II	1149	18	1333,7	–184,7
	III	1241	19	1360,2	–119,2
	IV	1320	20	1386,7	–66,7
$\sum$		22 694	210	X	X

Źródło: dane umowne. Obliczenia własne.

Liniowa funkcja trendu wypłat odsetek od lokat terminowych w poszczególnych kwartałach lat 2001–2005 przyjęła postać:

$$\hat{y}_t = 856{,}7 + 26{,}5t \quad (\sum t \neq 0).$$

Teoretyczne – obliczone z funkcji trendu – poziomy wypłat odsetek zawiera tab. 7.7. Wartości te będą – w dalszych obliczeniach – stanowić podstawę do wyodrębnienia absolutnych poziomów wahań sezonowych.

W kolejnym kroku obliczamy różnice: $y_t - \hat{y}_t$ (por. kolumnę 4 w tab. 7.7). Wykorzystując wzór (7.37) obliczamy absolutne poziomy wahań sezonowych:

$$g_I = \frac{-50{,}2 - 98{,}2 + 51{,}8 + 117{,}8 - 53{,}2}{5} = \frac{-32}{5} = -6{,}4 \text{ tys. zł},$$

$$g_{II} = \frac{-72{,}7 + 29{,}7 + 13{,}3 + 81{,}3 - 184{,}7}{5} = \frac{-192{,}5}{5} = -38{,}5 \text{ tys. zł},$$

$$g_{III} = \frac{-97{,}2 + 10{,}8 + 98{,}8 + 15{,}8 - 119{,}2}{5} = \frac{-91}{5} = -18{,}2 \text{ tys. zł},$$

$$g_{IV} = \frac{-52{,}7 + 187{,}3 + 230{,}3 + 12{,}3 - 66{,}7}{5} = \frac{310{,}5}{5} = 62{,}1 \text{ tys. zł},$$

$$\sum_{i=1}^{4} g_i = -1{,}0 \text{ tys. zł}.$$

Suma absolutnych poziomów wahań sezonowych odbiega od oczekiwanego zera. Zachodzi więc konieczność wprowadzenia współczynnika korygującego (7.38):

$$p = \frac{-1}{4} = -0{,}25.$$

Oczyszczone addytywne mierniki sezonowości są zatem równe:

$${}_k g_I = -6{,}4 - (-0{,}25) = -6{,}15 \text{ tys. zł},$$

$${}_k g_{II} = -38{,}5 - (-0{,}25) = -38{,}25 \text{ tys. zł},$$

$${}_k g_{III} = -18{,}2 - (-0{,}25) = -17{,}95 \text{ tys. zł},$$

$${}_k g_{IV} = 62{,}1 - (-0{,}25) = 62{,}35 \text{ tys. zł},$$

$$\sum_{i=1}^{4} g_i = 0.$$

Tak więc z powodu sezonowości wypłaty odsetek w każdym I kwartale są niższe od wypłat wyznaczonych trendem liniowym przeciętnie o 6,15 tys. zł, w każdym II kwartale – niższe o 38,25 tys. zł, w każdym III kwartale – niższe o 17,95 tys. zł, a w każdym IV kwartale – wyższe o 62,35 tys. zł.

7.4. Eliminacja wahań przypadkowych (losowych)

Wahania przypadkowe są spowodowane działaniem czynników losowych. Występują one nieregularnie i z różnym natężeniem. Z reguły nieznane są przyczyny ich występowania. W modelu wahań w czasie wahania przypadkowe są reprezentowane przez reszty $z(t)$, których syntetycznym miernikiem jest odchylenie standardowe składnika resztowego $s(z_t)$.

W addytywnym modelu wahań w czasie poziomy indywidualnych wahań przypadkowych określa się następująco:

$$z_t = y_t - \hat{y}_t - g_{it}, \tag{7.38}$$

gdzie: y_t – empiryczne wartości badanego zjawiska w poszczególnych podokresach, $\hat{y}_t$ – wartości teoretyczne (wynikające z oszacowanej funkcji trendu), g_{it} – absolutne poziomy wahań sezonowych ($i = 1, 2, ..., d$).

W multiplikatywnym modelu wahań w czasie wielkość indywidualnych wahań przypadkowych ustala się ze wzoru:

$$z_t = \frac{y_t}{\hat{y}_t \cdot S_i}, \tag{7.39}$$

gdzie S_i $(i = 1, 2, \ldots, d)$ – to wielkość sezonowości w poszczególnych podokresach cyklu sezonowości.

Jeżeli tendencję rozwojową wyodrębniono za pomocą metody mechanicznej, to we wzorach (7.38) i (7.39) symbol $\hat{y}_t$ zostaje zastąpiony przez średnie ruchome $\bar{y}_t$.

Schemat obliczeń składników resztowych dla addytywnego modelu wahań w czasie, opisującego wypłaty odsetek od lokat terminowych w poszczególnych kwartałach lat 2001–2005 (por. tab. 7.7), przedstawia tab. 7.8.

Tab. 7.8. Schemat obliczeń składników resztowych

Lata kwartały		t	y_t	$\hat{y}_t$	g_{it}	$z_t = y_t - \hat{y}_t - g_{it}$	z_t^2
2001	I	1	833	883,2	–6,15	–44,05	1940,4025
	II	2	837	909,7	–38,25	–34,45	1186,8025
	III	3	839	936,2	–17,95	–79,25	6280,5625
	IV	4	910	962,7	62,35	–115,05	13 236,5025
2002	I	5	891	989,2	–6,15	–92,05	8473,2025
	II	6	986	1015,7	–38,25	8,55	73,1025
	III	7	1053	1042,2	–17,95	28,75	826,5625
	IV	8	1256	1068,7	62,35	124,95	15 612,5025
2003	I	9	1147	1095,2	–6,15	57,95	3358,2025
	II	10	1135	1121,7	–38,25	51,55	2657,4025
	III	11	1247	1148,2	–17,95	116,75	13 630,5625
	IV	12	1405	1174,7	62,35	167,95	28 207,2025
2004	I	13	1319	1201,2	–6,15	123,95	15 363,6025
	II	14	1309	1227,7	–38,25	119,55	14 292,2025
	III	15	1270	1254,2	–17,95	33,75	1139,0625
	IV	16	1293	1280,7	62,35	–50,05	2505,0025
2005	I	17	1254	1307,2	–6,15	–47,05	2213,7025
	II	18	1149	1333,7	–38,25	–146,45	21 447,6025
	III	19	1241	1360,2	–17,95	–101,25	10 251,5625
	IV	20	1320	1386,7	62,35	–129,25	16 653,9025
$\sum$		**210**	**22 694**	X	X	X	**179 349,65**

Źródło: obliczenia własne.

Składniki resztowe obliczone dla poszczególnych kwartałów lat 2001–2005 informują o sile wpływu wahań przypadkowych (losowych) na poziom badanego zjawiska. Odczytując przykładowo pierwszy wiersz ($t = 1$) tab. 7.8 możemy stwierdzić, że na empiryczny poziom wypłat odsetek w I kwartale 2001 r. ($y_1 = 833$ tys. zł) składają się: kwota odsetek, będąca wynikiem oddziaływania przyczyn głównych ($\hat{y}_1 = 883{,}2$ tys. zł), kwota odsetek, będąca rezultatem działania wahań sezonowych ($g_1 = -6{,}15$ tys. zł) oraz kwota odsetek, będąca wypadkową działania wahań przypadkowych ($z_1 = -44{,}05$ tys. zł). Składniki resztowe informują zatem o wpływie na badane zjawisko tych czynników, które nie zostały wyjaśnione tendencją rozwojową (funkcją trendu) i wahaniami sezonowymi.

Syntetycznym miernikiem wahań przypadkowych jest odchylenie standardowe składnika resztowego, które oblicza się według wzoru:

$$s(z_t) = \sqrt{\frac{\sum_t (y_t - \hat{y}_t - g_{it})^2}{n-k-d+1}} = \sqrt{\frac{\sum_t z_t^2}{n-k-d+1}} \qquad (7.40)$$

oraz

$$s(z_t) = \sqrt{\frac{\sum_t \frac{y_t}{\hat{y}_t \cdot S_i}}{n-k-d+1}} = \sqrt{\frac{\sum_t z_t^2}{n-k-d+1}}, \qquad (7.41)$$

gdzie: n – długość szeregu czasowego, d – liczba podokresów cyklu sezonowości, k – liczba szacowanych parametrów funkcji trendu.

W najczęściej spotykanych przypadkach szacowania dwóch parametrów (liniowa funkcja trendu) zamiast $n-k-d+1$ wprowadzamy: $n-d-1$.

Podstawiając odpowiednie dane liczbowe z tab. 7.8 do wzoru (7.40) otrzymujemy:

$$s(z_t) = \sqrt{\frac{179349{,}65}{20-4-1}} = 109{,}35 \text{ tys. zł.}$$

Otrzymany wynik oznacza, że przeciętna siła wahań przypadkowych w całym szeregu czasowym wynosi ±109,35 tys. zł na kwartał. Im mniejsza wielkość odchylenia standardowego składnika resztowego, tym dany model wahań w czasie lepiej opisuje badaną rzeczywistość.

Stosunek odchylenia standardowego składnika resztowego do średniego poziomu badanego zjawiska ($\bar{y} = 22694{:}20 = 1134{,}7$) określa się mianem **współczynnika zmienności resztowej**, czyli:

$$V(z_t) = \frac{s(z_t)}{\bar{y}} \cdot 100. \qquad (7.42)$$

W naszym przykładzie współczynnik zmienności resztowej jest równy:

$$V(z_t) = \frac{109{,}35}{1134{,}7} \cdot 100 = 9{,}6\%.$$

Oznacza to, że odchylenia przypadkowe stanowią – średnio rzecz biorąc – 9,6% przeciętnego poziomu zaobserwowanej zmienności wypłat odsetek. Ze statystycznego punktu widzenia sytuacja jest tym lepsza, im $V(z_t)$ jest bliższe zeru.

7.5. Wnioskowanie statystyczne w analizie szeregów czasowych

Empiryczny szereg czasowy o numeracji okresów (momentów) $t = 1, 2, \ldots, n$ (lub $t = 0, 1, \ldots, n-1$) może być traktowany jako próba losowa wylosowana z populacji generalnej o numeracji od $t = -\infty$ do $t = +\infty$. W takim przypadku uprawnione jest wykorzystanie metod wnioskowania statystycznego (estymacji i weryfikacji hipotez) w analizie szeregów czasowych.

Sposób konstrukcji przedziałów ufności dla parametrów liniowej funkcji trendu zależy m.in. od liczebności próby. Najczęściej mamy do czynienia z próbami małymi ($n \leq 120$). Dlatego też konstrukcja przedziałów ufności dla parametrów α_0 i α_1 oparta jest na zmiennej standaryzowanej t-Studenta:

$$P\{a_0 - t_{\alpha,s} D(a_0) < \alpha_0 < a_0 + t_{\alpha,s} D(a_0)\} = 1 - \alpha \qquad (7.43)$$

oraz

$$P\{a_1 - t_{\alpha,s} D(a_1) < \alpha_1 < a_1 + t_{\alpha,s} D(a_1)\} = 1 - \alpha, \qquad (7.44)$$

gdzie $D(a_0)$ oraz $D(a_1)$ są średnimi błędami szacunku określonymi wzorami (7.22) i (7.23). Wartość $t_{\alpha,s}$ odczytuje się z tablic rozkładu t-Studenta dla przyjętego poziomu istotności α oraz $n-k$ stopni swobody.

W przypadku dużej próby ($n > 120$) zmienną t-Studenta zastępuje się – przy konstrukcji przedziałów ufności dla parametrów strukturalnych funkcji trendu – zmienną z posiadającą rozkład normalny $N(0,1)$.

W przypadku liniowej funkcji trendu merytoryczną interpretację ma współczynnik kierunkowy trendu (α_1). W wielu przypadkach dokonuje się sprawdzenia, czy otrzymana na podstawie wyników próby losowej ocena tego parametru jest statystycznie istotna. Stawia się zatem hipotezę zerową $H_0: \alpha_1 = 0$ oraz hipotezę alternatywną w jednej z postaci: $H_1: \alpha_1 \neq 0$, $H_1: \alpha_1 > 0$, $H_1: \alpha_1 < 0$. Wybór postaci hipotezy alternatywnej związany jest z obszarem krytycznym testu. Jeśli ocena parametru a_1 jest dodatnia, to można przypuszczać, że wartość nieznanego parametru α_1 jest również dodatnia ($H_1: \alpha_1 > 0$), a jeśli a_1 jest ujemne, to można się spodziewać, że $\alpha_1 < 0$ ($H_1: \alpha_1 < 0$). Stawiając hipotezę alternatywną w postaci: $H_1: \alpha_1 \neq 0$, nie precyzuje się znaku parametru α_1. W tym przypadku obszar odrzucenia hipotezy zerowej jest dwustronny.

Do weryfikacji hipotezy zerowej o istotności parametru α_1 wykorzystuje się statystykę o postaci:

$$t = \frac{a_i}{D(a_i)}. \qquad (7.45)$$

Statystyka (7.45) ma – przy założeniu prawdziwości H_0 – rozkład t-Studenta.

Oszacowana na podstawie wyników próby losowej funkcja trendu jest dobrą aproksymantą funkcji trendu w populacji generalnej, jeśli spełnione są następujące warunki:

1) suma kwadratów odchyleń między wartościami empirycznymi i teoretycznymi stanowi minimum,

2) odchylenia między wartościami empirycznymi i teoretycznymi są losowe,

3) nie występuje autokorelacja składników losowych.

Pierwszy warunek jest zawsze spełniony, gdy do szacowania parametrów funkcji trendu wykorzystano metodę najmniejszych kwadratów. Warunek drugi może być sprawdzony w wyniku weryfikacji hipotezy zerowej o losowości odchyleń wartości empirycznych szeregu czasowego od linii trendu (wartości teoretycznych). Hipoteza zerowa ma wówczas postać H_0: odchylenia są losowe, wobec hipotezy alternatywnej H_1: odchylenia nie są losowe. Hipotezę zerową można zweryfikować testem serii, prezentowanym przy badaniu różnic pomiędzy empirycznym i teoretycznym rozkładem regresji dwuwymiarowej (por. punkt 6.10 w rozdziale VI). Podkreślić jednak należy, że tablice rozkładu liczby serii podają zazwyczaj wartości dla $n_1 \le 20$ i $n_2 \ge 20$. Dla większych liczebności, korzystamy z tablic dystrybuanty rozkładu normalnego $N(0,1)$, gdyż przy weryfikacji hipotezy zerowej posługujemy się statystyką:

$$z = \frac{k - \bar{k}}{s_k}, \tag{7.46}$$

gdzie:

$$\bar{k} = \frac{2n_1n_2}{n_1 + n_2} \tag{7.47}$$

oraz

$$s_k = \sqrt{\frac{2n_1n_2(2n_1n_2 - n_1 - n_2)}{(n_1 + n_2)^2(n_1 + n_2 - 1)}}. \tag{7.48}$$

Odrzucenie hipotezy zerowej oznacza, że funkcja trendu nie ma charakteru liniowego.

Przy wyodrębnianiu tendencji rozwojowej z szeregów czasowych należy liczyć się z możliwością wystąpienia autokorelacji składników losowych. Autokorelacja oznacza liniową zależność między odchyleniami losowymi pochodzącymi z różnych jednostek czasu. Miarą siły i kierunku autokorelacji składników losowych jest współczynnik autokorelacji rzędu τ:

$$\rho_\tau = \rho(\xi_t, \xi_{t-\tau}). \tag{7.49}$$

Oszacowaniem współczynnika (7.49) jest współczynnik autokorelacji reszt z_t i $z_{t-\tau}$:

$$r_\tau = \frac{\sum_{t=\tau+1}^{n}(z_t - \bar{z}_t)(z_{t-\tau} - \bar{z}_{t-\tau})}{\sqrt{\sum_{t=\tau+1}^{n}(z_t - \bar{z}_t)^2 \sum_{t=1}^{n-\tau}(z_{t-\tau} - \bar{z}_{t-\tau})^2}}. \tag{7.50}$$

Jeśli $\tau = 1$, to zależność dotyczy następujących po sobie składników losowych ξ_t oraz ξ_{t-1}. Mówimy wówczas o **autokorelacji rzędu pierwszego**. Estymatorem współczynnika autokorelacji rzędu pierwszego jest współczynnik autokorelacji z próby wyznaczany jako:

$$r_1 = \frac{\sum_{t=2}^{n} z_t z_{t-1}}{\sqrt{\sum_{t=1}^{n} z_t^2 \sum_{t=2}^{n} z_{t-1}^2}}. \tag{7.51}$$

Do weryfikacji hipotezy zerowej o braku autokorelacji składnika losowego rzędu pierwszego ($H_0: \rho_1 = 0$) najczęściej wykorzystuje się test Durbina-Watsona. Statystyka Durbina-Watsona jest zdefiniowana następująco:

$$DW = \frac{\sum_{t=2}^{n}(z_t - z_{t-1})^2}{\sum_{t=1}^{n} z_t^2} \approx 2(1 - r_1), \tag{7.52}$$

gdzie z_t oznacza reszty funkcji trendu ($z_t = y_t - \hat{y}_t$).

Jak wiadomo, współczynnik korelacji jest zawarty w przedziale: $-1 \le r_1 \le 1$. Oznacza to, że: $0 \le DW \le 4$. Dla współczynnika autokorelacji z przedziału (–1,0) mamy do czynienia z ujemną autokorelacją, której odpowiada wartość DW z przedziału (2,4). Dla współczynnika autokorelacji z dodatniej części przedziału wartości DW zawierają się w przedziale (0,2). Dla dodatniej autokorelacji hipoteza alternatywna przyjmuje postać: $H_1: \rho_1 > 0$, a dla ujemnej: $H_1: \rho_1 < 0$.

Wartości krytyczne testu Durbina-Watsona są odczytywane z tablic rozkładu Durbina-Watsona przy liczebności próby n, liczbie zmiennych objaśniających k (bez wyrazu wolnego) oraz określonego poziomu istotności α. Z tablic tych odczytuje się dwie wartości: d_l i d_u ($d_l < d_u$). Wartość d_l nazywana jest dolną wartością krytyczną, natomiast d_u – górną wartością krytyczną. W przypadku testowania autokorelacji dodatniej, podejmowane mogą być następujące decyzje:

1) jeśli $DW < d_l$ – to H_0 odrzucamy. Oznacza to występowanie dodatniej autokorelacji składnika losowego,

2) jeśli $DW > d_u$ – to stwierdzamy, że brak jest podstaw do odrzucenia H_0 (brak istotnej dodatniej autokorelacji),

3) jeśli $d_l \le DW \le d_u$ – to nie można podjąć decyzji (jest to tzw. **obszar niekonkluzywności** testu).

W przypadku testowania autokorelacji ujemnej ($H_1: \rho_1 < 0$) postępujemy następująco:

1) jeśli $4 - DW < d_l$ ($DW > 4 - d_l$) – to H_0 odrzucamy (występuje ujemna autokorelacja składnika losowego),

2) jeśli $4 - DW > d_u$ ($DW < 4 - d_u$) – to brak jest podstaw do odrzucenia H_0 (wnioskujemy, że nie występuje istotna ujemna autokorelacja),

3) jeśli $d_l \le 4 - DW \le d_u$ ($4 - d_u \le DW \le 4 - d_l$) – to nie możemy podjąć żadnej decyzji (obszar niekonkluzywności testu).

W zasadzie hipotezę alternatywną powinno się formułować na podstawie znajomości oszacowania współczynnika autokorelacji r_1. Dopuszcza się jednak formułowanie postaci H_1 na podstawie wartości statystyki *DW*. Jeśli *DW* jest większe od dwóch, to testujemy autokorelację ujemną, gdy *DW* < 2 – to weryfikujemy autokorelację dodatnią ($H_1: \rho_1 > 0$). Generalnie zaś przyjmuje się, że jeśli wartość statystyki *DW* znacznie różni się od dwóch, to występuje autokorelacja składnika losowego.

Sposób postępowania przy weryfikacji hipotezy o występowaniu autokorelacji składnika losowego zilustrujemy przykładem liczbowym. Załóżmy, że dla pewnego liniowego modelu tendencji rozwojowej otrzymano następujący ciąg reszt odpowiadający kolejnym latom badanego okresu: 5; 0; –1; 2; –3; 8; 1; –2; –4; –1; 0; 1; 2; 2; –1. Na poziomie istotności 0,05 mamy sprawdzić, czy występuje autokorelacja składnika losowego rzędu pierwszego.

W pierwszym kroku należy – korzystając ze wzoru (7.51) – wyznaczyć ocenę estymatora współczynnika autokorelacji rzędu pierwszego. Niezbędne obliczenia zawiera tab. 7.9.

Tab. 7.9. Obliczenia pomocnicze do wyznaczenia współczynnika autokorelacji

t	z_t	z_{t-1}	$z_t \cdot z_{t-1}$	z_t^2	z_{t-1}^2	$z_t - z_{t-1}$	$(z_t - z_{t-1})^2$
1	5	–	–	25	–	–	–
2	0	5	0	0	25	–5	25
3	–1	0	0	1	0	–1	1
4	2	–1	–2	4	1	3	9
5	–3	2	–6	9	4	–1	1
6	8	–3	–24	64	9	11	121
7	1	8	8	1	64	–7	49
8	–2	1	–2	2	1	–3	9
9	–4	–2	8	16	2	–2	4
10	–1	–4	4	1	16	3	9
11	0	–1	0	0	1	1	1
12	–1	0	0	1	0	–1	1
13	–2	–1	2	4	1	–1	1
14	–2	–2	4	4	4	0	0
15	–1	–2	2	1	4	1	1
$\sum$	0	X	–6	133	134	X	232

Źródło: obliczenia własne.

Zgodnie z relacją (7.51) mamy:

$$r_1 = \frac{-6}{\sqrt{133 \cdot 134}} = \frac{-6}{133{,}5} = -0{,}045.$$

Wartość $r_1 = -0{,}045$ wskazuje na możliwość występowania autokorelacji ujemnej. Stawiamy zatem hipotezy: $H_0: \rho_1 = 0$ wobec hipotezy alternatywnej $H_1: \rho_1 < 0$. Obliczona, zgodnie ze wzorem (7.52), wartość statystyki Durbina-Watsona wynosi:

$$DW = \frac{232}{133} = 1{,}744.$$

Z tablic rozkładu Durbina-Watsona dla $k = 1$ (liczba szacowanych parametrów funkcji trendu bez wyrazu wolnego), $n = 15$ obserwacji oraz poziomu istotności $\alpha = 0{,}05$, odczytujemy wartości krytyczne: $d_l = 1{,}007$ i $d_u = 1{,}361$. Porównując obliczoną wartość statystyki $DW = 1{,}744$ z odczytanymi wartościami krytycznymi stwierdzamy, że:

$$4 - DW = 4 - 1{,}744 < d_u = 1{,}361$$

lub

$$DW = 1{,}744 < 4 - d_u = 4 - 1{,}361 = 2{,}639.$$

Oznacza to, że brak jest podstaw do odrzucenia hipotezy zerowej. Wnioskujemy zatem, iż nie występuje istotna ujemna autokorelacja.

Test Durbina-Watsona jest najpopularniejszym, ale nie jedynym testem służącym do weryfikacji hipotezy o istnieniu autokorelacji. Stosuje się go do weryfikacji hipotezy o autokorelacji pierwszego rzędu. W przypadku konieczności weryfikacji hipotez o autokorelacji wyższych rzędów, korzysta się np. z testu Z. Pawłowskiego, którego opis można znaleźć w pracy: *Prognozy ekonometryczne*, PWN, Warszawa 1973.

ZADANIA

7.1. Przychody ze sprzedaży oraz poziom zapasów materiałów w przedsiębiorstwie X w latach od $t = 1$ do $t = 5$ przedstawiają się następująco:

Lata (t)	1	2	3	4	5
Przychody (w tys. zł)	378	394	428	496	582
Zapasy w tys. zł (stan na 31 XII)	28,3	28,5	27,6	27,0	23,9

Obliczyć przeciętny roczny przychód ze sprzedaży oraz średni roczny poziom zapasów w badanym okresie.

7.2. Liczba zleceń wykonywanych przez firmę Y zajmującą się naprawą sprzętu gospodarstwa domowego w poszczególnych kwartałach lat $t = 1, 2, 3, 4$, przedstawia się następująco:

Kwartały	Lata (t)			
	1	2	3	4
I	8	9	10	11
II	12	16	22	28
III	11	14	18	20
IV	10	11	12	14

Stosując metodę mechaniczną wyodrębnić z podanego szeregu tendencję rozwojową. Przyjąć $k = 3$ oraz $k = 4$. Jakie wnioski wynikają z porównania wyodrębnionych trendów?

7.3. Wielkość importu owoców cytrusowych w latach od $t = 1$ do $t = 10$ kształtowała się następująco (w tys. ton): 19651,2; 20086,9; 20809,9; 21491,8; 21819,5; 22342,7; 22919,2; 24052,2; 25169,2; 23788,8. Z podanego szeregu wyodrębnić, metodą mechaniczną, tendencję rozwojową importu przy wykorzystaniu trzyletniej średniej ruchomej.

7.4. W latach $t = 1$ do $t = 6$ sprzedaż wyrobów gotowych w firmie X (w tys. zł) kształtowała się następująco: 1280; 1330; 1290; 1340; 1420; 1440. Wyodrębnić metodą analityczną (przy warunku $\sum t \neq 0$) tendencję rozwojową analizowanego zjawiska. Wyznaczyć i zinterpretować mierniki dopasowania trendu do danych empirycznych. Jakiego poziomu sprzedaży w badanej firmie należy oczekiwać w roku $T = n + 2$?

7.5. Kształtowanie się kosztów działalności operacyjnej w przedsiębiorstwie X przedstawiają następujące dane:

Lata	1998	1999	2000	2001	2002	2003	2004	2005
Koszty (w tys. zł)	15	12	18	13	19	16	26	17

Wyodrębnić metodą analityczną ($t = 1, 2, ...$) tendencję rozwojową kosztów.

7.6. W poszczególnych kwartałach lat od $t = 1$ do $t = 4$ produkcja (w sztukach) wyrobu X kształtowała się następująco:

Kwartały	Lata			
	1	2	3	4
I	8	9	10	11
II	12	16	22	28
III	11	14	18	20
IV	10	11	12	14

Na podstawie powyższych danych wyodrębnić trend metodą analityczną przy warunkach $\sum t \neq 0$ oraz $\sum t = 0$. Wyniki zinterpretować.

7.7. Dwie osoby oszacowały funkcje trendu dla poniższego szeregu:

Lata	2001		2002		2003		2004		2005	
Półrocza	I	II	I	II	I	II	I	II	I	II
y_t	66	58	60	58	58	54	48	52	54	50

Jedna z nich otrzymała funkcję trendu o postaci: $\hat{y}_t = 55{,}8 - 1{,}49t$, druga zaś: $\hat{y}_t = 64 - 1{,}49t$. Czy w szacunkach jednej z osób tkwi błąd? Jeśli tak, to która osoba otrzymała błędne wyniki? Jeżeli nie, to uzasadnij prawidłowość szacunku obydwu funkcji i przedstaw to na wykresie.

7.8. Na podstawie poniższych informacji oszacuj parametry strukturalne i struktury stochastycznej liniowej funkcji trendu:

t	1	2	3	4	5
y_t	2	5	5	8	10

Czy liniowy trend dobrze opisuje badane zjawisko?

7.9. Reszty liniowej funkcji trendu wynoszą: –1, 2, –1, –5, –5, 10, a wariancja zmiennej y wynosi 500. Oblicz i zinterpretuj R^2 i φ^2 oraz średnie błędy szacunku parametrów strukturalnych.

7.10. Liczba dni nieprzepracowanych z powodu choroby, za które wypłacono zasiłki chorobowe w poszczególnych półroczach lat od $t = 1$ do $t = 8$, kształtowała się następująco:

Lata	1		2		3		4		5		6		7		8	
Półrocza	I	II	I	II	I	II	I	II	I	II	I	II	I	II	I	II
Dni	59	36	40	37	37	36	39	39	42	43	47	46	43	45	57	49

Wyodrębnić wahania sezonowe, obliczając wskaźniki sezonowości oraz absolutne poziomy wahań sezonowych. Wykorzystać metodę średnich jednoimiennych okresów.

7.11. Na podstawie poniższych danych wyodrębnić analitycznie trend liniowy i oszacować stopień jego dopasowania do danych empirycznych.

t	1	2	3	4	5	6
y_t	41,175	45,827	52,033	60,752	67,981	74,632

7.12. Wyznaczyć i zinterpretować parametry strukturalne trendu liniowego ($\sum t \neq 0$). Zakładając, że trend nie ulegnie zmianie, oszacować poziom badanego zjawiska w okresie $T = n + 2$.

t	1	2	3	4	5	6	7
y_t	191,4	189,6	150,2	134,2	136,8	133,0	94,4

7.13. Liczba pracujących w sektorze prywatnym w tys. osób w województwie L w latach $t = 1, 2, ..., 8$ kształtowała się następująco:

t	1	2	3	4	5	6	7	8
y_t	7,9	8,4	8,4	8,7	9,0	9,4	10,1	10,6

Wyodrębnić metodą mechaniczną liniową funkcję trendu oraz zbadać dokładność jej dopasowania do danych empirycznych. Wyniki zinterpretować.

7.14. W poszczególnych kwartałach lat od $t = 1$ do $t = 5$ liczba mieszkań oddanych do użytku (w tys.) kształtowała się następująco:

Kwartały	Lata				
	1	2	3	4	5
I	23,5	23,3	23,7	47,3	18,7
II	35,8	28,1	29,9	19,6	19,6
III	29,4	31,0	28,5	22,4	19,1
IV	61,5	51,9	54,4	43,7	30,6

Wyodrębnić wskaźniki sezonowości przy wykorzystaniu metody odchyleń od trendu.

7.15. Liczbę samochodów na 1000 osób (w sztukach) zarejestrowanych w latach od $t = 1$ do $t = 9$ przedstawia poniższy szereg dynamiczny:

t	1	2	3	4	5	6	7	8	9
y_t	98	106	112	120	127	138	159	169	176

Wyznaczyć liniową funkcję trendu i ocenić dobroć jej dopasowania do danych empirycznych.

7.16. Dla danych kwartalnych z lat od $t = 1$ do $t = 5$ wyznaczono funkcję trendu o postaci $\hat{y}_t = 3 + 2{,}2t$ $(\sum t \neq 0)$. Absolutne poziomy wahań sezonowych w poszczególnych kwartałach są równe: $g_{II} = -1{,}7$; $g_{III} = -0{,}8$; $g_{IV} = 0{,}3$. Obliczyć przewidywaną wielkość zjawiska w I kwartale roku $T = n + 1$.

7.17. Poziom zjawiska (y_t) w latach od $t = 1$ do $t = 5$ kształtował się następująco (w tys. zł): 57,2 58,9; 59,6; 60,2; 61,5. Oszacować liniowe funkcje trendu przyjmując kolejno warunki: a) $\sum t = 0$, b) $\sum t \neq 0$. Wskazać na różnice w interpretacji otrzymanych wyników. Oszacować i zinterpretować odchylenie standardowe reszt.

7.18. Dynamikę kwartalnych obrotów (w tys. zł) biura turystycznego opisuje funkcja trendu o postaci: $\hat{y}_t = 2{,}5 + 0{,}2t + 0{,}01t^2$. Teoretyczna wielkość obrotów w III kwartale 2004 r. wynosiła 7,75 tys. zł. Wskaźnik sezonowości w III kwartale osiągnął poziom 115%, a wariancja składnika resztowego była równa 0,04. Oszacuj przewidywaną wielkość obrotów tego biura w III kwartale 2006 r.

7.19. Przeprowadzono badanie dynamiki obrotów w sklepach spożywczych w ostatnich pięciu latach. Otrzymano trend roczny o postaci: $\hat{y}_t = 100 + 50t$ $(\sum t = 0)$ oraz kwartalne absolutne poziomy wahań sezonowych: 6; –1; –21; 16. Jakiego poziomu obrotów należy oczekiwać w poszczególnych kwartałach roku T = n + 1?

7.20. Liniowa funkcja trendu przewozów towarów (w tys. ton) w poszczególnych kwartałach ostatnich trzech lat przyjęła postać: $\hat{y}_t = 100 + 50t$ $(\sum t = 0)$. Oszacuj wielkość przewozów w IV kwartale piątego roku, jeśli $S_{IV} = 120\%$.

7.21. Trend skupu mleka (w mln litrów) w poszczególnych półroczach ostatnich czterech lat opisuje równanie trendu o postaci: $\hat{y}_t = 0{,}1t + 2$ $(t = 0, 1, 2, \ldots)$. Absolutny poziom wahań sezonowych dla II półrocza wynosi: –0,3 mln litrów, wariancja składnika resztowego 0,0004. Jakiej wielkości skupu mleka należy oczekiwać w I półroczu roku piątego?

7.22. Funkcja trendu liniowego dla półroczy ostatnich 5 lat przedstawia równanie: $\hat{y}_t = 1{,}2t + 3{,}6$ $(t = 0, 1, \ldots)$. Oszacuj poziom zjawiska w II półroczu siódmego roku, jeżeli wiadomo, że $S_I = 110\%$, a wariancja składnika resztowego jest równa 0,25.

7.23. Spożycie artykułu A na jednego mieszkańca w ostatnich sześciu latach opisuje trend roczny o postaci: $\hat{y}_t = 5t + 52$ $(\sum t = 0)$. Wskaźniki sezonowości spożycia są równe: $S_I = 90\%$, $S_{II} = 110\%$, $S_{III} = 140\%$. Obliczyć przewidywane spożycie w IV kwartale siódmego roku.

7.24. W latach od $t = 1$ do $t = 4$ sprzedaż artykułu X (w tys. kg) na terenie pewnej aglomeracji miejskiej przedstawia się następująco:

Kwartały	Lata			
	1	2	3	4
I	274,4	261,9	180,4	310,3
II	262,4	233,5	244,4	284,3
III	224,9	190,6	218,0	273,5
IV	284,1	274,1	335,6	372,1

Wyznaczyć kwartalne wskaźniki sezonowości oraz absolutne poziomy wahań sezonowych. Wyniki zinterpretować.

7.25. Tendencję rozwojową sprzedaży masła (w tys. kostek po 0,25 kg) w pewnym hipermarkecie w poszczególnych kwartałach lat od $t = 1$ do $t = 6$ przedstawia funkcja trendu o postaci: $\hat{y}_t = 45{,}2 + 0{,}45t$. Wiedząc, że suma jednostek czasu wynosi 300, a wariancja składnika resztowego 6,25, oszacować przewidywaną wielkość sprzedaży masła w II kwartale toku $T = n + 1$.

7.26. W pewnym przedsiębiorstwie wyznaczono kwartalną funkcję trendu dla lat od $t = 1$ do $t = 5$ dotyczącą produkcji podstawowej w cenach porównywalnych (w tys. zł): $\hat{y}_t = 8t + 80$ $(\sum t = 0)$. Ponadto ustalono, że wskaźniki sezonowości były równe: $S_I = 70\%$, $S_{II} = 120\%$, $S_{III} = 80\%$. Oszacować, jaka powinna być produkcja w poszczególnych kwartałach roku $T = n + 1$.

7.27. Suma kwadratów reszt dla 72-elementowego szeregu czasowego miesięcznych rozmiarów produkcji napojów alkoholowych wynosi 590 000. Ile wynosi odchylenie standardowe reszt, jeśli uprzednio zastosowano liniową funkcję do wyodrębniania tendencji rozwojowej badanego zjawiska?

7.28. W rezultacie dekompozycji empirycznego półrocznego szeregu czasowego (obejmującego lata od $t = 1$ do $t = 4$) otrzymano m.in. następujące addytywne i multiplikatywne wartości reszt: –1; 5; 4; –2; –1; 3; 2; 1 oraz 4; 10; 12; 8; 7; 16; 12; 3. Ile wynoszą odchylenia standardowe reszt dla obu rozkładów?

7.29. Zbadano wielkość sprzedaży (w tys. zł) pewnego towaru w hurtowni artykułów spożywczych. Empiryczny szereg czasowy (z półroczną sezonowością) został opisany następującymi informacjami z modelu wahań w czasie: $\hat{y}_t = -t + 50$; z_t: 2; –1; 3; –2; 2; –4; $S_I = 0{,}85$; $\sum_{t=1}^{n} (y_t - \hat{y}_t)^2 = 200$. Jakiej sprzedaży należy oczekiwać w drugim półroczu pierwszego roku prognozy?

7.30. Oszacowano wskaźniki sezonowości oraz absolutne poziomy wahań sezonowych czasu pracy, przy czym zostały utracone dwie informacje. Jakie są poziomy brakujących informacji i jaka jest ich interpretacja?

Mierniki sezonowości	Kwartały			
	I	II	III	IV
Wskaźniki sezonowości	1,03	?	0,956	1,098
Absolutne poziomy wahań sezonowych	+31,4	–84,8	?	+99,6

7.31. Produkcja piwa w pewnym browarze (w tys. litrów) w poszczególnych kwartałach lat od $t = 1$ do $t = 6$ przedstawia się następująco: 26; 34; 35; 33; 32; 36; 40; 35; 38; 45; 54; 47; 43; 55; 59; 56; 53; 60; 65; 59; 60; 72; 66; 64. Na podstawie tych danych oszacowano liniową funkcję trendu o postaci: $\hat{y}_t = 26{,}5 + 1{,}8t$. Na poziomie istotności

0,05 zweryfikować hipotezę o losowym charakterze odchyleń wartości empirycznych od teoretycznych.

7.32. W latach od $t = 1$ do $t = 15$ przewozy pasażerów w pewnym oddziale PKS kształtowały się następująco (w tys. osób): 14,4; 14,6; 14,9; 15,2; 15,5; 15,7; 15,9; 16,4; 16,6; 16,8; 17,1; 17,4; 17,6; 18,1; 18,5. Na podstawie powyższych informacji oszacowano liniową funkcję trendu o postaci: $\hat{y}_t = 14{,}03 + 0{,}28t$ ($t = 1, 2, ...$). Na poziomie istotności 0,05 zweryfikować hipotezę o braku autokorelacji składnika resztowego.

7.33. Liczbę widzów w kinach lubelskich (w tys. osób) aproksymowano do postaci liniowej funkcji trendu: $\hat{y}_t = 25{,}6 - 0{,}5t$ ($\sum t \neq 0$). Odchylenie standardowe składnika resztowego było równe 0,6 tys. osób. Zbuduj przedziały ufności dla parametrów strukturalnych funkcji trendu. Przyjąć współczynnik ufności 0,95.

7.34. Oszacowano liniową funkcję trendu, a następnie policzono reszty otrzymując: 3; 5; –1; –4; –3; 0; 1; 2; –3; –4; 2; 2; 3; –1; –2. Przyjmując poziom istotności 0,05, zweryfikować hipotezę o braku autokorelacji składnika losowego.

7.35. Liczba zwolnień z pracy pracowników pewnego zakładu w kolejnych półroczach ostatnich pięciu lat wyraża się równaniem:

$$\hat{y}_t = \underset{(0,3)}{3t} + \underset{(2,8)}{22} \quad \left(\sum t = 0\right).$$

Czy zaobserwowany liniowy trend półroczny można uznać za statystycznie istotny?

7.36. Na podstawie rocznych danych oszacowano tendencję rozwojową produkcji (w tys. sztuk) otrzymując: $\hat{y}_t = 80 + 0{,}85t$ ($t = 1, 2, ..., 7$) oraz $s_e = 2{,}1$. Czy zaobserwowana wzrostowa tendencja produkcji ma istotny charakter? Przyjąć poziom istotności 0,05.

7.37. Na podstawie 12-elementowej próby losowej oszacowano liniowy model trendu, a następnie obliczono reszty. Liczba reszt o ujemnym znaku wynosi 4, a empiryczna liczba serii jest równa 10. Czy otrzymany ciąg reszt ma – na poziomie istotności 0,05 – charakter losowy?

7.38. Na podstawie danych kwartalnych oszacowano tendencję rozwojową liczby remontów mieszkań w latach od $t = 1$ do $t = 5$, otrzymując:

a) przeciętny kwartalny przyrost liczby remontów oszacowany na poziomie 0,114;

b) teoretyczny poziom liczby remontów w I kwartale roku $t = 1$ (oszacowany na podstawie liniowej funkcji trendu przy numeracji kwartałów od 0 do 19) wynoszący 9,771.

Zbudować – przy współczynniku ufności 0,95 – przedziały ufności dla parametrów liniowej funkcji trendu. Wiadomo, że wariancja składnika resztowego wynosi 0,0064.

7.39. Dynamikę nakładów inwestycyjnych (w tys. zł) w poszczególnych kwartałach lat od $t = 1$ do $t = 8$ określa funkcja trendu o postaci: $\hat{y}_t = 0{,}4728t + 7{,}3284$. Suma kwadratów reszt wynosi 53,823. Czy – na poziomie istotności 0,01 – można twierdzić, że współczynnik kierunkowy funkcji trendu jest statystycznie istotny?

7.40. W latach $t = 1, 2, ..., 15$ empiryczne poziomy sprzedaży wyrobu X (w tys. zł) kształtowały się następująco: 10,9; 11,6; 12,3; 12,0; 11,2; 12,6; 13,4; 13,6; 11,7; 12,9; 13,8; 14,3; 14,2; 13,9; 14,4. Wyznaczona na podstawie tych informacji liniowa funkcja

trendu przyjęła postać: $\hat{y}_t = 11{,}26 + 0{,}23t$. Za pomocą testu Durbina-Watsona zweryfikować – na poziomie istotności 0,05 – hipotezę o braku składnika losowego.

7.41. Informacje o trendzie i sezonowości sprzedaży towaru X w poszczególnych półroczach lat $t = 1, 2, ..., 5$ przedstawiają się następująco: +2; –3; +1; 0; +2; –1; +2; –2; 0; –1. Absolutny poziom wahań sezonowych dla pierwszego półrocza wynosi: –1. Zbadać, na poziomie istotności 0,05, losowość odchyleń reszt addytywnego modelu wahań w czasie.

7.42. Trend skupu owoców (w tys. ton) w poszczególnych półroczach lat od $t = 1$ do $t = 5$ przedstawia następujące równanie: $\hat{y}_t = 0{,}1t + 2$ ($t = 0, 1, ...$). Wiadomo ponadto, że wielkość skupu owoców w każdym drugim półroczu była wyższa – średnio rzecz biorąc – od przeciętnej półrocznej o 0,3 tys. kg, a suma kwadratów reszt równa 0,0024. Wiedząc, że rzeczywisty poziom skupu owoców w roku $T = 6$ wynosił 5 tys. kg, sprawdzić, czy statystyczna prognoza skupu byłaby potwierdzeniem rzeczywistości.

7.43. Oszacowany na podstawie próby losowej model tendencji rozwojowej wielkości przewozów kolejowych w latach od $t = 1$ do $t = 15$ przyjął postać: $\hat{y}_t = -3{,}88t + 249{,}9$. Wiadomo ponadto, że $S(a_1) = 0{,}784$, a $S(a_0) = 7{,}127$. Odchylenie standardowe reszt było równe 13,12. Przyjmując współczynnik ufności 0,95 zbudować przedziały ufności dla parametrów α_0 i α_1. Wyniki zinterpretować.

7.44. Długość dróg w km na 10 tysięcy ludności w Polsce w latach 1998–2002 kształtowała się następująco: 63,2; 64,3; 64,6; 64,3; 65,5. Oszacować parametry strukturalne liniowej funkcji trendu przy założeniu: a) $t = 0, 1, ..., n - 1$; b) $\sum t = 0$. Wskazać na różnice w interpretacji otrzymanych wyników. Oszacować współczynnik zmienności resztowej.

7.45. W pewnym sklepie osiedlowym zanotowano następujące wartości sprzedanych kartonów mleka 2% (0,5 litra) w poszczególnych kwartałach lat od $t = 1$ do $t = 5$ (w tys. zł): 5,7; 7,05; 8,3; 9,35 oraz 10,5. Zakładając, że $\sum t \neq 0$, oszacować liniowy model trendu oraz zbadać dokładność jego dopasowania do danych empirycznych.

7.46. Dysponujemy następującymi informacjami:

y_t	298	320	350	358	370	400	408	450	464	500	510	532	560
$\hat{y}_t$	295	317	338	360	381	403	425	446	468	489	511	533	554

Na poziomie istotności 0,05, zweryfikować hipotezę o braku autokorelacji składnika losowego.

7.47. Dysponujemy następującymi informacjami: $\hat{y}_t = 1000 + 120t$ ($t = 0, 1, 2, ..., 9$). Jednostki czasu dotyczą poszczególnych półroczy ostatnich pięciu kolejnych lat. Wiadomo ponadto, że wartości reszt są równe: –0,2; –0,3; 0,4; 0,1; 0,2; –0,5; 0,3; 0,2; –0,1; –0,1. Na poziomie istotności 0,05 zweryfikować hipotezę o losowości reszt.

7.48. Wiedząc, że: 1) liczba sprzedanych komputerów w latach 1986–2005 malała średnio z roku na rok o 0,23 sztuk; 2) teoretyczny poziom sprzedaży komputerów w 1985 r. wynosił 10,09 sztuk; 3) wartości empiryczne liczby sprzedanych komputerów różniły się od wartości teoretycznych średnio o ±0,3016 sztuk, i przyjmując współczynnik ufności 0,95, oszacować przedziały ufności dla parametrów strukturalnych liniowej funkcji trendu.

7.49. Oszacowano następujący model trendu dla lat od $t = 1$ do $t = 5$:

$$\hat{y}_t = \underset{(0,283)}{6,928t} + \underset{(0,483)}{57,067} + \underset{(1,184)}{z_t} \quad (\sum t = 0).$$

Na poziomie istotności 0,05 zweryfikować hipotezę o liniowości trendu.

7.50. Czy liniowa funkcja trendu oszacowana na podstawie 18-elementowego szeregu czasowego jest funkcją II rodzaju, jeśli empiryczna liczba serii $k = 9$, a liczba składników o dodatnim znaku $n_a = 8$. Nie zanotowano sytuacji $y_t = \hat{y}_t$. Przyjąć poziom istotności 0,05.

Rozdział VIII

INDEKSY STATYSTYCZNE

Wyodrębnianie z szeregów czasowych trendu wahań sezonowych czy przypadkowych jest możliwe tylko w przypadku dysponowania dużą liczbą obserwacji badanego zjawiska (długie szeregi chronologiczne). Najkrótszy szereg czasowy tworzy dwuelementowy zbiór wartości. Jedna z tych wartości dotyczy okresu badanego ($t = 1$), druga zaś – okresu przyjętego za podstawę porównań ($t = 0$). Do analizy tego rodzaju szeregów czasowych wykorzystuje się **metody indeksowe**, wśród których wyróżnia się przyrosty absolutne i względne, indeksy jednopodstawowe i łańcuchowe, indeksy dla wielkości absolutnych i stosunkowych (ilorazowych).

8.1. Proste metody badania zmian szeregu dynamicznego

8.1.1. Przyrosty absolutne

Dwie liczby możemy porównać ze sobą przez odejmowanie lub dzielenie. Odejmowanie dwóch wielkości liczbowych daje w wyniku dodatni lub ujemny **przyrost absolutny** (zwany też przyrostem **bezwzględnym**). Przyrosty absolutne mogą być obliczane w stosunku do jednego okresu (momentu) lub też okresu (momentu) stale zmieniającego się. W pierwszym przypadku mówimy o przyrostach absolutnych **o podstawie stałej (jednopodstawowych)**, w drugim zaś – o przyrostach absolutnych **o podstawie zmiennej (łańcuchowych)**.

Jeżeli poszczególne wyrazy szeregu dynamicznego oznaczymy przez: $y_1, y_2, \ldots, y_n$, a za podstawę porównań przyjmiemy wielkość y_1, to ciąg przyrostów absolutnych o podstawie stałej przyjmuje postać:

$$y_2 - y_1, y_3 - y_1, \ldots, y_{n-1} - y_1, y_n - y_1. \tag{8.1}$$

Ciąg przyrostów absolutnych łańcuchowych przedstawia się następująco:

$$y_2 - y_1, y_3 - y_2, \ldots, y_{n-1} - y_{n-2}, y_n - y_{n-1}. \tag{8.2}$$

Przyrosty absolutne informują o tym, o ile jednostek wzrósł (+) lub zmalał (–) poziom badanego zjawiska w okresie (momencie) badanym w po-

równaniu z okresem (momentem) przyjętym za podstawę. Przyrosty absolutne (zarówno jednopodstawowe, jak i łańcuchowe) są wielkościami **mianowanymi**, wyrażonymi w tych samych jednostkach miary co badane zjawisko. Są one zatem wykorzystywane do porównań poziomu zjawisk wyrażonych w takich samych jednostkach fizycznych.

Istotnym zagadnieniem przy obliczaniu przyrostów absolutnych jednopodstawowych jest **wybór podstawy porównań**. Wybrany do porównań okres (moment) powinien być na tyle charakterystyczny, aby pozwalał poznać istotę zachodzących zmian. Dlatego też nie należy za podstawę porównań przyjmować okresu zupełnie wyjątkowego pod względem poziomu badanego zjawiska, gdyż wszystkie porównania byłyby wówczas zniekształcone. Badając np. rozwój produkcji roślinnej nie można za podstawę porównań przyjmować plonów ziemiopłodów z tego roku, w którym zanotowano wysoki nieurodzaj. Przyrosty absolutne dawałyby bowiem niesłychanie korzystny obraz rozwoju produkcji roślinnej.

8.1.2. Przyrosty względne

Przyrostem względnym nazywamy iloraz przyrostów absolutnych zjawiska do jego poziomu w okresie (momencie) przyjętym za podstawę porównań. Przyrosty względne, podobnie jak absolutne, mogą być **jednopodstawowe** lub **łańcuchowe**. Ciąg wartości przyrostów względnych o podstawie stałej (y_1) jest następujący:

$$\frac{y_2 - y_1}{y_1}, \frac{y_3 - y_1}{y_1}, \ldots, \frac{y_{n-1} - y_1}{y_1}, \frac{y_n - y_1}{y_1}. \tag{8.3}$$

Ciąg wartości przyrostów względnych łańcuchowych można natomiast zapisać w postaci:

$$\frac{y_2 - y_1}{y_1}, \frac{y_3 - y_2}{y_2}, \ldots, \frac{y_{n-1} - y_{n-2}}{y_{n-2}}, \frac{y_n - y_{n-1}}{y_{n-1}}. \tag{8.4}$$

Przyrosty względne wyrażane są w **ułamkach** lub w **procentach**. Informują one o tym, o ile wyższy lub niższy jest poziom badanego zjawiska w danym okresie w stosunku do okresu bezpośrednio poprzedzającego (przyrosty względne łańcuchowe) lub w stosunku do okresu przyjętego za podstawę porównań (przyrosty względne jednopodstawowe). Przyrosty względne mogą być wielkościami dodatnimi, ujemnymi lub równymi zero. Przyrosty względne łańcuchowe określane są mianem **wskaźników tempa zmian**.

8.1.3. Wskaźniki dynamiki (indeksy)

Wskaźnikiem dynamiki (indeksem) nazywamy stosunek wielkości badanego zjawiska w danym okresie (momencie) do wielkości tego samego zjawiska w innym okresie (momencie), przyjętym za podstawę porównań.

Okres, z którego bierzemy wielkość porównywaną, nazywamy okresem **badanym (sprawozdawczym)**, a okres, do którego odnosimy wartość porównywaną, określamy mianem okresu **bazowego (podstawowego)**. Jeżeli poziom zjawiska w okresie (momencie) badanym oznaczymy symbolem y_1, a w okresie podstawowym przez y_0, to wzór na indeks przyjmie następującą postać:

$$i = \frac{y_1}{y_0}. \tag{8.5}$$

Indeks jest wielkością niemianowaną i może być wyrażony w ułamku lub w procentach. Jeżeli przyjmie on wartość z przedziału: $0 \leq i \leq 1$, świadczy to o spadku poziomu zjawiska w okresie badanym w stosunku do okresu podstawowego. Większa od jedynki (lub od 100%) wartość indeksu informuje o wzroście poziomu zjawiska w okresie badanym w porównaniu z okresem podstawowym. Wartość indeksu równa 1 (lub 100%) oznacza, że wielkość zjawiska w porównywanych okresach była taka sama.

Wskaźniki dynamiki (indeksy) można podzielić na **jednopodstawowe i łańcuchowe**. Ciąg indeksów o podstawie stałej można zapisać – przy podstawie porównań y_1 – następująco:

$$\frac{y_1}{y_1}, \frac{y_2}{y_1}, \ldots, \frac{y_{n-1}}{y_1}, \frac{y_n}{y_1}. \tag{8.6}$$

Ciąg indeksów łańcuchowych można zapisać jako:

$$\frac{y_2}{y_1}, \frac{y_3}{y_2}, \ldots, \frac{y_{n-1}}{y_{n-2}}, \frac{y_n}{y_{n-1}}. \tag{8.7}$$

Technikę obliczania przyrostów absolutnych względnych oraz indeksów przedstawia tablica 8.1.

Tab. 8.1. Obliczanie przyrostów absolutnych, względnych oraz indeksów

Lata (t)	y_t	Przyrosty absolutne		Przyrosty względne (w %)		Indeksy (w %)	
		jednopodstawowe ($t = 1$) = 100	łańcuchowe	jednopodstawowe ($t = 1$) = 100	łańcuchowe	jednopodstawowe ($t = 1$) = 100	łańcuchowe
1	34,6	0,0		0,0		100,0	
2	36,3	1,7	1,7	4,9	4,9	104,9	104,9
3	37,4	2,8	1,1	8,1	3,0	108,1	103,0
4	39,7	5,1	2,3	14,7	6,1	114,7	106,1
5	39,7	5,1	0,0	14,7	0,0	114,7	100,0
6	41,3	6,7	1,6	19,4	4,0	119,4	104,0
7	38,0	3,4	–3,3	9,8	–8,0	109,8	92,0
8	43,2	8,6	5,2	24,9	13,7	124,9	113,7

Źródło: dane umowne. Obliczenia własne.

Jak wynika z tab. 8.1, między przyrostami względnymi a indeksami istnieje ścisły związek. Indeksy jednopodstawowe można otrzymać z przyrostów względnych o podstawie stałej przez dodanie 100 (lub 1, jeśli posługujemy się ułamkami, a nie wielkościami procentowymi).

W analogiczny sposób można dokonać zamiany indeksów łańcuchowych na przyrosty względne łańcuchowe. Można również dokonać operacji odwrotnej, tzn. zamienić indeksy na przyrosty. Wielkość przyrostu względnego otrzymujemy jako pomniejszoną o 100 (lub 1) wartość indeksu.

W praktyce badań statystycznych częściej wykorzystuje się indeksy niż przyrosty. Na indeksach wygodniej jest bowiem dokonywać określonych przekształceń algebraicznych. Przekształcenia te dotyczą zazwyczaj zamiany indeksów jednopodstawowych na łańcuchowe i odwrotnie oraz zmiany podstawy w szeregach indeksów jednopodstawowych.

Zamiany indeksów **jednopodstawowych na łańcuchowe** dokonujemy przez **dzielenie** indeksów jednopodstawowych przez siebie zgodnie ze wzorem:

$$\frac{y_i}{y_1} : \frac{y_{i-1}}{y_1} = \frac{y_i}{y_{i-1}}. \tag{8.8}$$

Tab. 8.2. Operacje na szeregach indeksowych

t	Absolutny poziom zjawiska (y_t)	Indeksy jednopodstawowe o podstawie $t = 3$	Zamiana indeksów jednopodstawowych na łańcuchowe	Zamiana indeksów łańcuchowych na jednopodstawowe	Zmiana podstawy w indeksach jednopodstawowych z $t = 3$ na $t = 1$
1	y_1	$y_1 : y_3$	–	$1 : \left(\frac{y_3}{y_2} \cdot \frac{y_2}{y_1}\right) = \frac{y_1}{y_3}$	$\frac{y_1}{y_3} : \frac{y_1}{y_3} = 1$
2	y_2	$y_2 : y_3$	$\frac{y_2}{y_3} : \frac{y_1}{y_3} = \frac{y_2}{y_1}$	$1 : \frac{y_3}{y_2} = \frac{y_2}{y_3}$	$\frac{y_2}{y_3} : \frac{y_1}{y_3} = \frac{y_2}{y_1}$
3	y_3	$y_3 : y_3$	$\frac{y_3}{y_3} : \frac{y_2}{y_3} = \frac{y_3}{y_2}$	$y_3 : y_3 = 1$	$\frac{y_3}{y_3} : \frac{y_1}{y_3} = \frac{y_3}{y_1}$
4	y_4	$y_4 : y_3$	$\frac{y_4}{y_3} : \frac{y_3}{y_3} = \frac{y_4}{y_3}$	$y_4 : y_3 = \frac{y_4}{y_3}$	$\frac{y_4}{y_3} : \frac{y_1}{y_3} = \frac{y_4}{y_1}$
5	y_5	$y_5 : y_3$	$\frac{y_5}{y_3} : \frac{y_4}{y_3} = \frac{y_5}{y_4}$	$\frac{y_4}{y_3} \cdot \frac{y_5}{y_4} = \frac{y_5}{y_3}$	$\frac{y_5}{y_3} : \frac{y_1}{y_3} = \frac{y_5}{y_1}$

Źródło: dane umowne. Obliczenia własne.

Zamiany indeksów **łańcuchowych na jednopodstawowe** dokonujemy według następujących zasad:

1) indeks jednopodstawowy w okresie następującym bezpośrednio po okresie przyjętym za podstawę jest taki sam jak indeks łańcuchowy;

2) indeks jednopodstawowy w okresie przyjętym za podstawę wynosi 100%;

3) dalsze indeksy jednopodstawowe po okresie przyjętym za podstawę otrzymujemy mnożąc w sposób narastający kolejne indeksy łańcuchowe, licząc od wskaźnika łańcuchowego znajdującego się tuż po okresie podstawowym;

4) indeksy jednopodstawowe przed okresem podstawowym są odwrotnością iloczynów kolejnych indeksów łańcuchowych, licząc od okresu przyjętego za podstawę.

Zmiany podstawy w indeksach jednopodstawowych dokonujemy przez dzielenie poszczególnych indeksów przy danej podstawie przez indeks jednopodstawowy tego okresu, który przyjmujemy za nową podstawę.

Schemat ilustrujący operacje na szeregach czasowych przedstawiono w tab. 8.2.

8.1.4. Obliczanie średniego tempa zmian zjawisk w czasie

Zarówno indeksy jednopodstawowe, jak i łańcuchowe pozwalają na ocenę zmian badanego zjawiska między wyróżnionymi okresami (momentami). Często jednak zachodzi konieczność oceny zmian danego zjawiska w całym okresie objętym obserwacją. Do tego celu wykorzystuje się **średnią geometryczną**.

Średnia geometryczna ($\overline{y}_g$) jest pierwiastkiem n-tego stopnia z iloczynu n zmiennych. Z n wielkości absolutnych można utworzyć $n - 1$ indeksów łańcuchowych. Dysponując zatem ciągiem indeksów łańcuchowych, średnią geometryczną obliczymy następująco:

$$\overline{y}_g = \sqrt[n-1]{\frac{y_2}{y_1} \cdot \frac{y_3}{y_2} \cdots \frac{y_{n-1}}{y_{n-2}} \cdot \frac{y_n}{y_{n-1}}} = \sqrt[n-1]{\prod_{i=2}^{n} \frac{y_i}{y_{i-1}}}. \tag{8.9}$$

Jeśli opieramy się na informacjach podanych w postaci wielkości absolutnych, średnią geometryczną możemy obliczyć ze wzoru:

$$\overline{y}_g = \sqrt[n-1]{\frac{y_n}{y_1}}. \tag{8.10}$$

Z relacji (8.10) wynika, że średnia geometryczna jest pierwiastkiem stopnia $n - 1$ z ilorazu absolutnych poziomów badanego zjawiska z ostatniego i pierwszego okresu (momentu).

W przypadku, gdy podstawą do wyznaczenia średniej geometrycznej jest ciąg indeksów jednopodstawowych, wykorzystujemy wzór:

$$\overline{y}_g = \sqrt[n-1]{\text{ostatni indeks jednopodstawowy}} \Big/ \text{pierwszy indeks jednopodstawowy}. \tag{8.11}$$

Średnie tempo zmian w okresie, którego dotyczy szereg czasowy, w wyrażeniu procentowym, definiuje się jako:

$$\text{średnie tempo zmian} = (\overline{y}_g - 1) \cdot 100. \tag{8.12}$$

Załóżmy, że dysponujemy informacjami dotyczącymi indeksów łańcuchowych w okresie $t = 1, 2, ..., 8$: 1,005; 0,928; 0,958; 0,978; 0,833; 0,970; 0,989. Zauważmy, że dla $n = 8$ utworzono $n - 1 = 7$ indeksów łańcuchowych. Wykorzystując wzór (8.9) obliczamy średnią geometryczną:

$$\bar{y}_g = \sqrt[7]{1{,}005 \cdot 0{,}928 \cdot 0{,}958 \cdot 0{,}978 \cdot 0{,}833 \cdot 0{,}970 \cdot 0{,}989} = 0{,}965.$$

Średnie tempo zmian w całym rozpatrywanym okresie od $t = 1$ do $t = 8$ jest zatem równe:

$$(0{,}965 - 1) \cdot 100 = -3{,}5\%.$$

Oznacza to, że w latach od $t = 1$ do $t = 8$ poziom badanego zjawiska zmniejszał się z roku na rok średnio o 3,5%.

8.2. Indeksy indywidualne i agregatowe

Wskaźniki dynamiki (indeksy) można podzielić na **indywidualne (proste)** i **agregatowe (złożone)**. Indeksy indywidualne dotyczą pojedynczych zmiennych reprezentujących jednorodne, nierozkładalne zjawiska (np. spożycie ziemniaków w gospodarstwach domowych, liczba sprzedanych biletów w kinie). Indeksy agregatowe (zespołowe) obliczane są dla niejednorodnego zbioru elementów (np. wielkość produkcji różnych asortymentów wyrobów).

8.2.1. Indeksy indywidualne

Indeksem indywidualnym nazywamy iloraz poziomów tego samego pojedynczego zjawiska z dwóch różnych okresów (momentów). W statystyce społeczno-ekonomicznej rozpatruje się zwykle trzy rodzaje indywidualnych wskaźników dynamiki: **indeksy cen, ilości i wartości**.

Indywidualny indeks cen wyraża relację poziomu cen określonego dobra w okresie badanym i w okresie podstawowym, czyli:

$$i_p = \frac{p_1}{p_0}, \tag{8.13}$$

gdzie: i_p – indywidualny indeks cen, p_1 – cena jednostki wyrobu w okresie badanym (sprawozdawczym), p_0 – cena jednostki wyrobu w okresie podstawowym (bazowym).

Indywidualny indeks ilości (masy fizycznej) oblicza się jako stosunek ilości określonego wyrobu wytworzonego w okresie badanym i w okresie podstawowym, tzn.:

$$i_q = \frac{q_1}{q_0}, \tag{8.14}$$

gdzie: i_q – indywidualny indeks ilości, q_1 – ilość wyrobu wyprodukowanego w okresie badanym, q_0 – ilość wyrobu wyprodukowanego w okresie podstawowym. Iloczyn ilości wytworzonego wyrobu w okresie badanym i ceny tego wyrobu z tego samego okresu daje w wyniku wartość ($w_1 = q_1 p_1$). Podobnie oblicza się wartość w okresie podstawowym: $w_0 = q_0 p_0$. Iloraz wartości wyrobu wytworzonego w okresie badanym i w okresie podstawowym nazywamy **indywidualnym indeksem wartości**:

$$i_w = \frac{q_1 p_1}{q_0 p_0} = \frac{w_1}{w_0}, \tag{8.15}$$

gdzie: i_w – indywidualny indeks wartości, w_1 – wartość w okresie badanym, w_0 – wartość w okresie podstawowym.

Indywidualne indeksy: cen, ilości i wartości, informują o zmianie (wzroście lub spadku) tych wielkości w okresie badanym w porównaniu do okresu przyjętego za bazę. Między tymi indeksami zachodzi następująca relacja:

$$i_w = i_q \cdot i_p. \tag{8.16}$$

Relacja (8.16) nosi nazwę **równości indeksowej dla indeksów indywidualnych**. Równość ta umożliwia – przy znajomości dowolnej pary spośród wyróżnionych indeksów – obliczenie trzeciego indeksu.

8.2.2. Indeksy zespołowe dla wielkości absolutnych

W praktyce badań statystycznych niejednokrotnie zachodzi potrzeba obliczania indeksów dotyczących całego zespołu (agregatu) jednostek. Do tego celu stosuje się **indeksy zespołowe (agregatowe)**. W zależności od przedmiotu badań wyróżniamy **indeksy zespołowe dla wielkości absolutnych** i **indeksy zespołowe dla wielkości stosunkowych**.

W grupie zespołowych indeksów dla wielkości absolutnych wyróżnia się **agregatowy indeks wartości, cen i ilości**. Agregatowy indeks wartości jest ilorazem sum wartości badanych dóbr w okresie badanym i w okresie podstawowym, czyli:

$$I_w = \frac{\sum w_1}{\sum w_0} = \frac{\sum q_1 p_1}{\sum q_0 p_0}, \tag{8.17}$$

gdzie: I_w – agregatowy indeks wartości badanego zespołu wyrobów, $\sum q_1 p_1 = \sum w_1$ – suma wartości badanego agregatu w okresie badanym, $\sum q_0 p_0 = \sum w_0$ – suma wartości w okresie podstawowym.

W zapisie indeksu (8.17) opuszczono – dla uproszczenia – granice sumowania; rozciągają się one na wszystkie elementy agregatu.

Agregatowy indeks wartości wyraża zmiany, jakie nastąpiły w okresie badanym w porównaniu z okresem podstawowym zarówno w ilościach określonego agregatu, jak również w ich cenach. W celu obliczenia siły i kierunku zmian wyłącznie ilości lub tylko cen wyrobów wchodzących w skład analizowanego agregatu, oblicza się **agregatowe indeksy ilości** (fizycznych rozmiarów, masy fizycznej) i **agregatowe indeksy cen**.

Konstrukcja agregatowych indeksów ilości i cen oparta jest na **metodzie eliminacji**, zwanej też **standaryzacją wskaźników dynamiki**. Standaryzacja polega tu na ustaleniu stałego poziomu jednego z dwóch czynników: cen lub ilości. Do uzyskania agregatowego indeksu ilości unieruchamiane są (ustalane na stałym poziomie) ceny. Do uzyskania agregatowego indeksu cen, unieruchamia się ilości. Wybór okresu standaryzacji zależy od celu badania i posiadanych informacji statystycznych. Najczęściej wykorzystuje się dwie formuły standaryzacyjne: **Laspeyresa** i **Paaschego**.

Standaryzacja według formuły Laspeyresa polega na unieruchomieniu ilości (przy obliczaniu agregatowego indeksu cen) lub cen (przy obliczaniu agregatowego indeksu ilości) na poziomie okresu podstawowego (bazowego). Standaryzacja według formuły Paaschego polega na unieruchomieniu ilości w indeksie cen lub cen w indeksie ilości na poziomie okresu badanego (sprawozdawczego).

W zależności od przyjętej formuły standaryzacyjnej można zatem wyróżnić dwa rodzaje agregatowych indeksów ilości i cen: Laspeyresa i Paaschego. **Agregatowy indeks ilości według formuły Laspeyresa** przyjmuje następującą postać:

$$I_q^L = \frac{\sum q_1 p_0}{\sum q_0 p_0}. \tag{8.18}$$

Agregatowy indeks ilości według formuły Paaschego oblicza się ze wzoru:

$$I_q^P = \frac{\sum q_1 p_1}{\sum q_0 p_1}. \tag{8.19}$$

Agregatowe indeksy ilości informują o tym, o ile – przeciętnie biorąc – wzrosła lub zmalała ilość określonego zbioru artykułów w okresie badanym w porównaniu z okresem podstawowym (przy odpowiednim założeniu przyjętym w formule standaryzacyjnej).

Agregatowy indeks cen jest ilorazem sumy wartości określonego zbioru artykułów w okresie badanym i sumy wartości tych samych artykułów w okresie podstawowym przy stałym „koszyku" ilości. Przy obliczaniu agregatowych indeksów cen rolę wag spełniają ilości. Standaryzując ilości na poziomie okresu podstawowego, otrzymujemy **agregatowy indeks cen typu Laspeyresa**:

$$I_p^L = \frac{\sum p_1 q_0}{\sum p_0 q_0}. \tag{8.20}$$

Jeśli przyjmiemy za niezmienne ilości na poziomie okresu badanego, to otrzymamy **agregatowy indeks cen według formuły Paaschego**:

$$I_p^P = \frac{\sum p_1 q_1}{\sum p_0 q_1}. \tag{8.21}$$

Agregatowe indeksy cen odpowiadają na pytanie, jak zmieniły się – przeciętnie biorąc – ceny danego zbioru artykułów w okresie badanym w porównaniu z okresem podstawowym, przy unieruchomieniu ilości w obu okresach, zgodnie z przyjętą formułą standaryzacyjną.

Agregatowe indeksy cen i ilości obliczone według formuł Laspeyresa i Paaschego dla tego samego agregatu zwykle różnią się między sobą. Powstaje zatem pytanie, która z tych formuł jest poprawniejsza? Zaleca się, aby – przy pełnym dostępie do informacji – obliczać indeksy agregatowe według obydwu formuł standaryzacyjnych. Formuły te wyznaczają bowiem granice zmian w dynamice badanego agregatu.

W przypadku niezbyt odległych okresów porównawczych (tzn. okresów podstawowego i badanego) obliczane są również **agregatowe indeksy według formuły Fishera**. Agregatowy indeks typu Fishera jest średnią geometryczną indeksów standaryzowanych (cen i ilości) według formuł Laspeyresa i Paaschego. Zgodnie z podaną definicją, agregatowe indeksy ilości i cen według formuły Fishera przyjmują postać:

$$I_q^F = \sqrt{I_q^L \cdot I_q^P} \tag{8.22}$$

oraz

$$I_p^F = \sqrt{I_p^L \cdot I_p^P}. \tag{8.23}$$

Jak łatwo zauważyć, pomiędzy agregatowymi indeksami wartości, ilości i cen zachodzą następujące związki:

$$I_w = I_p^L \cdot I_q^P = I_p^P \cdot I_q^L = I_p^F \cdot I_q^F. \tag{8.24}$$

Związek określony relacją (8.24) nosi nazwę **równości indeksowej dla indeksów agregatowych (zespołowych)**. Jeśli dysponujemy informacjami o poziomach dwóch spośród trzech omawianych indeksów agregatowych, to możemy obliczyć wielkość indeksu trzeciego.

W przedsiębiorstwie produkującym trzy rodzaje wyrobów zebrano informacje dotyczące produkcji (w tys. sztuk) oraz cen jednostkowych (w zł) w dwóch latach: $t = 0$ oraz $t = 1$ (por. tab. 8.3).

Tab. 8.3. Produkcja i ceny jednostkowe

Wyroby	Produkcja w tys. sztuk		Ceny jednostkowe w zł	
	$t = 0$	$t = 1$	$t = 0$	$t = 1$
A	1,2	1,5	10,0	9,0
B	1,5	1,4	6,5	7,0
C	0,8	1,0	5,0	5,8

Źródło: dane umowne.

Naszym zadaniem jest odpowiedź na pytanie: jak zmieniła się wartość produkcji w porównywanych latach? Jaki wpływ na zmianę wartości miała dynamika cen, a jaki dynamika ilości produkowanych wyrobów?

Odpowiedzi na postawione pytania wymagają przeprowadzenia określonych obliczeń. Są one zawarte w tab. 8.4.

Tab. 8.4. Obliczenia pomocnicze

Wyroby	q_0p_0	q_1p_1	q_0p_1	q_1p_0
A	12,00	13,50	10,80	15,00
B	9,75	9,80	10,50	9,10
C	4,00	5,80	4,64	5,00
Razem	**25,75**	**29,10**	**25,94**	**29,10**

Źródło: obliczenia własne.

Podstawiając odpowiednie dane liczbowe z tab. 8.4 do wzoru (8.17), otrzymujemy:

$$I_w = \frac{29,10}{25,75} \cdot 100 = 113\%.$$

Wynik ten oznacza, że łączna produkcja wyrobów w roku $t = 1$ wzrosła w porównaniu z rokiem $t = 0$ o 13%. Złożyły się na to zarówno zmiany w ilościach produkowanych wyrobów, jak też w ich cenach.

W celu oszacowania wpływu dynamiki ilości i dynamiki cen na 13% wzrost wartości produkcji, należy wyznaczyć agregatowe indeksy ilości i cen. Agregatowe indeksy ilości, obliczone według formuł standaryzacyjnych Laspeyresa (wzór (8.18)) i Paaschego (wzór (8.19)) są równe:

$$I_q^L = \frac{29,10}{25,75} \cdot 100 = 113,010\%,$$

$$I_q^P = \frac{29,10}{25,94} \cdot 100 = 112,182\%.$$

Fizyczne rozmiary produkcji wszystkich wyrobów łącznie wzrosły o 13,01% (przy założeniu stałych cen z roku $t = 0$) oraz o 12,182% (przy założeniu stałych cen z roku $t = 1$). Różnica między tymi indeksami jest niewielka i można ją uznać za nieistotną statystycznie. Zastosowanie formuły Fishera (wzór (8.22)) daje wynik:

$$I_q^F = \sqrt{1,1301 \cdot 1,12182} = 1,1259 = 112,59\%.$$

Stosując wzory (8.20) i (8.21) otrzymujemy oszacowania dynamiki cen przy założeniach właściwych formułom Laspeyresa i Paaschego:

$$I_p^L = \frac{25,94}{25,75} \cdot 100 = 100,74\%,$$

$$I_p^P = \frac{29,10}{29,10} \cdot 100 = 100\%.$$

Tak więc przy założeniu stałych ilości na poziomie roku $t = 0$, ceny wszystkich wyrobów wzrosły o 0,74% w roku $t = 1$ w porównaniu z $t = 0$. Gdyby przyjąć stałe ilości produkcji z roku $t = 1$, to ceny wszystkich

wyrobów w roku $t = 1$ nie zmieniłyby się w porównaniu z rokiem $t = 0$. Indeks cen obliczony według formuły Fishera jest natomiast równy:

$$I_p^F = \sqrt{1,0074 \cdot 1,00} = 100,37\%.$$

Dekompozycję agregatowego indeksu wartości można ostatecznie zapisać w postaci równości indeksowej (wzór (8.24)):

$$1,13 = 1,0074 \cdot 1,12182 = 1 \cdot 1,13.$$

W zaprezentowanej powyżej procedurze obliczania agregatowych indeksów wartości, cen i ilości konieczna jest znajomość cen i ilości zarówno z okresu podstawowego, jak i badanego. W przypadku braku tego rodzaju informacji korzysta się z równoważnych wzorów o postaci **średniej arytmetycznej** lub **średniej harmonicznej**.

Agregatowe indeksy wartości przyjmują postać:

$${}_A I_w = \frac{\sum w_0 \cdot i_w}{\sum w_0} \qquad (8.25)$$

oraz

$${}_H I_w = \frac{\sum w_1}{\sum \frac{w_1}{i_w}}. \qquad (8.26)$$

Agregatowe indeksy ilości są natomiast określone wzorami:

$${}_A I_q = \frac{\sum w_0 \cdot i_q}{\sum w_0} \qquad (8.27)$$

oraz

$${}_H I_q = \frac{\sum w_1}{\sum \frac{w_1}{i_q}}. \qquad (8.28)$$

Formuły obliczania agregatowych indeksów cen przedstawiają się następująco:

$${}_A I_p = \frac{\sum w_0 \cdot i_p}{\sum w_0} \qquad (8.29)$$

oraz

$${}_H I_p = \frac{\sum w_1}{\sum \frac{w_1}{i_p}}. \qquad (8.30)$$

Wzory (8.25), (8.27) i (8.29) dotyczą postaci średnioarytmetycznej, wzory (8.26), (8.28) oraz (8.30) – postaci średnioharmonicznej.

8.2.3. Indeksy zespołowe dla wielkości stosunkowych

Wielkości stosunkowe (ilorazowe) są wskaźnikami natężenia wyrażającymi stosunek dwóch zjawisk logicznie ze sobą powiązanych. Przykładem wielkości stosunkowych są takie kategorie ekonomiczne, jak:

- wydajność pracy (iloraz produkcji i czasu pracy),
- koszt jednostkowy (iloraz kosztów całkowitych i wielkości produkcji),
- średnia płaca (iloraz funduszu płac i liczby zatrudnionych).

Każdą wielkość stosunkową można rozpatrywać jako ogólną (zespołową) i cząstkową (jednostkową). Wielkości stosunkowe cząstkowe zapisujemy w postaci ułamka:

$$x = \frac{a}{b} \tag{8.31}$$

stąd

$$a = x \cdot b \text{ oraz } b = \frac{a}{x} = \frac{1}{x} \cdot a.$$

Wielkości stosunkowe ogólne (dotyczące całej zbiorowości) zapisujemy w postaci ułamka:

$$X = \frac{A}{B} = \frac{\sum a}{\sum b} = \frac{\sum xb}{\sum b} = \frac{\sum a}{\sum \frac{a}{x}}. \tag{8.32}$$

Narzędziem analizy dynamiki wielkości stosunkowych są **indeksy wielkości stosunkowych**. Oznaczając przez X_1 ogólną wielkość stosunkową w okresie badanym, a przez X_0 wielkość stosunkową w okresie podstawowym, możemy zapisać

$$I_x = \frac{X_1}{X_0} = \frac{\sum a_1}{\sum b_1} : \frac{\sum a_0}{\sum b_0} = \frac{\sum x_1 b_1}{\sum b_1} : \frac{\sum x_0 b_0}{\sum b_0} = \frac{\sum a_1}{\sum \frac{a_1}{x_1}} : \frac{\sum a_0}{\sum \frac{a_0}{x_0}}. \tag{8.33}$$

Indeks określony wzorem (8.33) nosi nazwę **agregatowego** (zespołowego) **indeksu wszechstronnego** albo **indeksu o zmiennej strukturze**. Jak wynika z relacji (8.33), indeks ten można obliczać za pomocą trzech różnych, ale równoważnych sposobów. Zastosowanie każdego z nich uzależnione jest od informacji, którymi dysponujemy.

Wartość indeksu wszechstronnego jest wypadkową działania dwóch czynników, a mianowicie:

1) dynamiki cząstkowych wielkości stosunkowych,
2) zmian w strukturze czynników a lub b.

Wpływ dynamiki cząstkowych wielkości stosunkowych na poziom indeksu wszechstronnego określają indeksy zespołowe wielkości stosunkowych o **stałej strukturze**. Przyjmując za niezmienne wskaźniki struktury czynników a i b z okresu podstawowego (formuła Laspeyresa), otrzymujemy wzory na indeksy o stałej strukturze:

$$I_{x/a_0} = \frac{\sum a_0}{\sum \frac{a_0}{x_1}} : \frac{\sum a_0}{\sum \frac{a_0}{x_0}}, \tag{8.34}$$

$$I_{x/b_0} = \frac{\sum x_1 b_0}{\sum b_0} : \frac{\sum x_0 b_0}{\sum b_0}. \tag{8.35}$$

Stabilizując współczynniki struktury czynników a i b na poziomie okresu badanego (formuła Paaschego), otrzymujemy:

$$I_{x/a_1} = \frac{\sum a_1}{\sum \frac{a_1}{x_1}} : \frac{\sum a_1}{\sum \frac{a_1}{x_0}}, \tag{8.36}$$

$$I_{x/b_1} = \frac{\sum x_1 b_1}{\sum b_1} : \frac{\sum x_0 b_1}{\sum b_1}. \tag{8.37}$$

Wpływ zmian w strukturze czynników a i b na poziom indeksu wszechstronnego określają **indeksy zespołowe zmian strukturalnych**. Indeks zmian strukturalnych typu **Laspeyresa** otrzymujemy przyjmując za stałe cząstkowe wielkości stosunkowe z okresu podstawowego. Agregatowy indeks wpływu zmian w strukturze czynnika b typu Laspeyresa ma następującą postać:

$${}_bI_{x/x_0} = \frac{\sum x_0 b_1}{\sum b_1} : \frac{\sum x_0 b_0}{\sum b_0}. \tag{8.38}$$

Zespołowy indeks wpływu zmian w strukturze czynnika a obliczony według formuły **Laspeyresa** określa wzór:

$${}_aI_{x/x_0} = \frac{\sum a_1}{\sum \frac{a_1}{x_0}} : \frac{\sum a_0}{\sum \frac{a_0}{x_0}}. \tag{8.39}$$

Jeżeli za niezmienne przyjmiemy cząstkowe wielkości stosunkowe z okresu badanego, to otrzymamy indeks wpływu zmian strukturalnych typu **Paaschego**. Agregatowy indeks wpływu zmian w strukturze czynnika b, obliczony według formuły Paaschego, jest równy:

$${}_bI_{x/x_1} = \frac{\sum x_1 b_1}{\sum b_1} : \frac{\sum x_1 b_0}{\sum b_0}. \tag{8.40}$$

Agregatowy indeks wpływu zmian w strukturze czynnika a, według formuły Paaschego, określony jest wzorem:

$${}_aI_{x/x_1} = \frac{\sum a_1}{\sum \frac{a_1}{x_1}} : \frac{\sum a_0}{\sum \frac{a_0}{x_1}}. \tag{8.41}$$

W konkretnym badaniu statystycznym dotyczącym analizy dynamiki wielkości stosunkowych o doborze odpowiedniej postaci indeksów o stałej i zmiennej strukturze decydują względy merytoryczne. Oblicza się mianowicie te indeksy, które wywierają bezpośredni wpływ na dynamikę badanej wielkości stosunkowej. Na przykład bezpośredni wpływ na dynamikę urodzeń mają zmiany w strukturze kobiet w wieku rozrodczym (czynnik b). Dlatego przy badaniu dynamiki płodności[1] nie ma potrzeby obliczania indeksów o stałej i zmiennej strukturze czynnika a (liczba urodzeń żywych).

Pomiędzy indeksem wszechstronnym a indeksami o stałej i o zmiennej strukturze czynnika b, przy zachowaniu przemienności formuł standaryzacyjnych, zachodzą następujące relacje:

$$I_x = I_{x/b_1} \cdot_b I_{x/x_0}, \tag{8.42}$$

$$I_x = I_{x/b_0} \cdot_b I_{x/x_1}. \tag{8.43}$$

Analogiczny związek ma miejsce między indeksem wszechstronnym a indeksem o stałej strukturze czynnika a i indeksem wpływu zmian w strukturze czynnika a, tzn.:

$$I_x = I_{x/a_1} \cdot_a I_{x/x_0}, \tag{8.44}$$

$$I_x = I_{x/a_0} \cdot_a I_{x/x_1}. \tag{8.45}$$

Korzystając z relacji od (8.42) do (8.45) – zwanych **równościami indeksowymi dla wielkości stosunkowych** – można metodą pośrednią obliczyć jeden z trzech indeksów, jeżeli tylko znane są dwa pozostałe.

Zasady konstrukcji indeksów dla wielkości stosunkowych oraz sposób interpretacji otrzymanych wyników zilustrujemy przykładem modelowym.

W tab. 8.5 podano informacje dotyczące produkcji (w tys. sztuk) i efektywnego czasu pracy (w tys. robotnikogodzin) w czterech kooperujących ze sobą zakładach.

Tab. 8.5. Produkcja i efektywny czas pracy zakładów w latach $t = 0$ i $t = 1$

Zakłady	Produkcja w tys. sztuk		Efektywny czas pracy (w tys. rbg)	
	$t = 0$	$t = 1$	$t = 0$	$t = 1$
A	2842,0	1967,6	104,1	84,1
B	2952,5	2539,2	88,2	86,5
C	2950,0	3513,7	98,3	117,7
D	2426,2	1857,3	88,4	92,2

Źródło: dane umowne.

Naszym zadaniem jest przeprowadzenie wszechstronnej analizy dynamiki wydajności pracy we wszystkich zakładach łącznie.

[1] Płodność jest ilorazem urodzeń żywych do liczby kobiet w wieku rozrodczym.

Informacje liczbowe zawarte w tab. 8.5 pozwalają na obliczenie wszechstronnego indeksu wydajności pracy według wzoru:

$$I_x = \frac{\sum x_1 b_1}{\sum b_1} : \frac{\sum x_0 b_0}{\sum b_0}. \tag{8.46}$$

Obliczenia pomocnicze zawiera tab. 8.6.

Tab. 8.6. Obliczenia pomocnicze

Zakłady	$x_0 = \frac{a_0}{b_0}$	$x_1 = \frac{a_1}{b_1}$	b_0	b_1	$x_0 b_0$	$x_1 b_1$
A	27,3	23,4	104,1	84,1	2841,93	1967,94
B	33,5	29,4	88,2	86,5	2954,70	2543,10
C	30,0	29,9	98,3	117,7	2949,00	3519,13
D	27,4	20,1	88,4	92,2	2422,16	1853,22
Razem	X	X	**379,0**	**380,5**	**11 167,79**	**9883,49**

Źródło: obliczenia własne.

Wykorzystując dane z tab. 8.6, otrzymujemy następującą wartość wszechstronnego indeksu wydajności pracy:

$$I_x = \frac{9883,49}{380,5} : \frac{11167,79}{379,0} = \frac{25,975}{29,466} = 0,8815 = 88,15\%.$$

Otrzymany wynik oznacza, że poziom wydajności pracy we wszystkich czterech zakładach łącznie w roku $t = 1$ zmalał w porównaniu z rokiem $t = 0$ o 11,85%. Na spadek ten miały wpływ zmiany zarówno w cząstkowych poziomach wydajności pracy (w poszczególnych zakładach), jak i w strukturze efektywnego czasu pracy. Przyjmując za stałą (standardową) strukturę czasu pracy na poziomie roku $t = 0$, można obliczyć indeks o stałej strukturze czynnika b według formuły Laspeyresa. Standaryzując strukturę czasu pracy na poziomie roku $t = 1$, otrzymujemy indeks o stałej strukturze czynnika b według formuły Paaschego. Obliczenia z tym związane zawarte są w tab. 8.7.

Tab. 8.7. Obliczenia pomocnicze

Zakłady	$x_1 b_0$	$x_0 b_0$	$x_1 b_1$	$x_0 b_1$	b_0	b_1
A	2435,94	2841,93	1967,94	2295,93	104,1	84,1
B	2593,08	2954,70	2543,10	2897,75	88,2	86,5
C	2939,17	2949,00	3519,23	3531,00	98,3	117,7
D	1776,84	2422,16	1853,22	2526,28	88,4	92,2
Razem	**9745,03**	**11 167,79**	**9883,49**	**11 250,96**	**379,0**	**380,5**

Źródło: obliczenia własne.

Indeks wpływu zmian w cząstkowych poziomach wydajności pracy – przy stałej strukturze czasu pracy z roku $t = 0$ – wynosi:

$$I_{x/b_0} = \frac{\sum x_1 b_0}{\sum b_0} : \frac{\sum x_0 b_0}{\sum b_0} = \frac{9745{,}03}{379} : \frac{11167{,}79}{379} = 0{,}8726 = 87{,}26\%.$$

Tak więc przy wyeliminowaniu różnych struktur czasu pracy przez przyjęcie standardowej struktury z roku $t = 0$, ogólny poziom wydajności pracy w roku $t = 1$ jest niższy o 12,74% w porównaniu z rokiem $t = 0$.

Jeżeli przyjmiemy za stałą strukturę czasu pracy z roku $t = 1$, to możemy obliczyć indeks o stałej strukturze czynnika b według formuły Paaschego:

$$I_{x/b_1} = \frac{\sum x_1 b_1}{\sum b_1} : \frac{\sum x_0 b_1}{\sum b_1} = \frac{9883{,}49}{380{,}5} : \frac{11250{,}96}{380{,}5} = 0{,}8785 = 87{,}85\%.$$

Oznacza to, że przy założeniu stałej struktury czasu pracy z roku $t = 1$, ogólny poziom wydajności pracy w roku $t = 1$ zmalał o 12,15% w porównaniu z rokiem $t = 0$.

Wpływ zmian w strukturze czasu pracy na dynamikę wydajności pracy można zbadać za pomocą indeksu wpływu zmian strukturalnych. Przyjmując do obliczeń stałe cząstkowe poziomy wydajności pracy z roku $t = 0$ otrzymujemy:

$${}_bI_{x/x_0} = \frac{\sum x_0 b_1}{\sum b_1} : \frac{\sum x_0 b_0}{\sum b_0} = \frac{11250{,}96}{380{,}5} : \frac{11167{,}79}{379} = 1{,}00348 = 100{,}348\%.$$

Uzyskany wynik wskazuje, że gdyby cząstkowe poziomy wydajności pracy w roku $t = 1$ były takie same jak w roku $t = 0$, to na skutek zmian w strukturze czasu pracy nastąpiłby wzrost ogólnego poziomu wydajności pracy o 0,348%.

Jeżeli za niezmienne przyjmiemy cząstkowe poziomy wydajności pracy z roku $t = 1$, to otrzymamy agregatowy indeks wpływu zmian w strukturze czynnika b według formuły Paaschego:

$${}_bI_{x/x_0} = \frac{\sum x_0 b_1}{\sum b_1} : \frac{\sum x_0 b_0}{\sum b_0} = \frac{9883{,}49}{380{,}5} : \frac{9745{,}03}{379} = 1{,}01021 = 101{,}021\%.$$

Oznacza to, że gdyby w obu porównywanych latach cząstkowe poziomy wydajności pracy nie zmieniły się i były na poziomie roku $t = 1$, to ogólny poziom wydajności pracy w toku $t = 1$ wzrósłby o 1,021% w porównaniu z rokiem $t = 0$. Wzrost ten byłby spowodowany tylko i wyłącznie zmianami w strukturze efektywnego czasu pracy.

Prawidłowość przeprowadzonych obliczeń można sprawdzić, posługując się równościami indeksowymi dla indeksów stosunkowych, określonych wzorami (8.42) – (8.43):

$$I_x = I_{x/b_1} \cdot {}_bI_{x/x_0} = 0{,}8785 \cdot 1{,}00348 = 0{,}8815,$$
$$I_x = I_{x/b_0} \cdot {}_bI_{x/x_1} = 0{,}8726 \cdot 1{,}01021 = 0{,}8815.$$

Spełnienie powyższych równości świadczy o poprawności przeprowadzonych obliczeń.

ZADANIA

8.1. Indeksy jednospodstawowe (o podstawie z roku $t = 1$) liczby zatrudnionych w jednym dziale gospodarki narodowej w ostatnich pięciu latach kształtowały się następująco:

Lata	1	2	3	4	5
Indeksy	100,0	116,4	160,1	250,6	294,0

Wiedząc, że liczba zatrudnionych w roku $t = 1$ wynosiła 3200 tys. osób, oblicz absolutne poziomy zatrudnienia w poszczególnych latach badanego okresu oraz wyznacz średnie roczne tempo wzrostu zatrudnienia w okresie od $t = 1$ do $t = 5$.

8.2. Dynamika zatrudnienia mierzona indeksami łańcuchowymi w ostatnich 8 latach kształtowała się następująco:

Lata	1	2	3	4	5	6	7	8
Indeksy	101,7	103,1	104,4	103,9	103,5	102,3	101,0	101,4

Dokonaj zamiany indeksów łańcuchowych na jednopodstawowe o podstawie z roku $t = 1$.

8.3. Dane są następujące indeksy jednopodstawowe wartości produkcji pewnego zakładu w ostatnich dziewięciu latach: 1,0; 1,2; 1,2; 1,4; 1,3; 1,5; 1,65; 1,7; 1,8. Wyznaczyć indeksy łańcuchowe oraz średnie tempo zmian.

8.4. Uzupełnij poniższą tablicę:

Lata (t)	1	2	3	4
y_t	300		396	
Indeksy jednopodstawowe ($t = 1$)		1,2		
Indeksy łańcuchowe				1,15

8.5. Porównać średnie tempo zmian wartości obrotów w zakładach X i Y w ostatnich pięciu latach, jeśli:

a) indeksy jednopodstawowe o podstawie $t = 1$ dla zakładu X są równe: 1,0; 1,6; 2,1; 2,0; 2,2;

b) indeksy łańcuchowe dla zakładu Y: 1,2; 1,2; 1,0; 1,8.

8.6. Ceny akcji pewnej firmy w ciągu jednego tygodnia zmieniły się w stosunku do ceny z poniedziałku następująco: 1; 0,98; 1,02; 0,99; 1,05. Jak zmieniały się ceny z dnia na dzień? Wyznaczyć indeksy zmian cen akcji w porównaniu z ceną ze środy.

8.7. Wyznaczyć średnie tempo zmian, mając dane:

t	1	2	3	4
$\frac{y_t - y_{t_0}}{y_{t_0}}$	1,8	1,9	0,2	0,3

8.8. Jednopodstawowe indeksy produkcji w latach od $t = 1$ do $t = 8$ kształtowały się następująco:

Lata	1	2	3	4	5	6	7	8
Indeksy	90	92	96	100	105	112	11	6121

Korzystając z powyższych danych, utworzyć indeksy jednopodstawowe o podstawie z roku $t = 1$.

8.9. Korzystając z poniższych danych, oblicz przyrosty względne jednopodstawowe o podstawie z roku $t = 3$:

Lata	1	2	3	4	5	6
Indeksy łańcuchowe	–	110	105	120	115	110

8.10. Korzystając z poniższych danych, wyznacz przyrosty względne łańcuchowe:

Lata	1	2	3	4	5	6	7
Indeksy jednopodstawowe	90	92	96	100	104	110	115

8.11. Produkcję wyrobu X w ostatnich 7 latach charakteryzują poniższe indeksy łańcuchowe:

Lata	1	2	3	4	5	6	7
Indeksy	94,28	93,79	56,54	40,96	56,49	98,85	93,02

Oblicz średnie roczne tempo zmian produkcji wyrobu w analizowanym okresie. O ile zmieni się produkcja wyrobu X w roku $t = 7$ w porównaniu z produkcją z roku $t = 3$?

8.12. Zużycie gazu sieciowego przez wiejskie gospodarstwa domowe w pewnym województwie charakteryzuje następujący szereg:

Lata	1	2	3	4	5	6	7
Zużycie w m^3	466	586	710	774	893	1103	1407

Jakiego zużycia gazu należy oczekiwać w roku $T = 9$?

8.13. Przyrosty względne łańcuchowe dotyczące sprzedaży detalicznej artykułu A w latach $t = 1, 2, ..., 7$ przedstawiają się następująco: 0,2; –0,11; 0,0; 0,15; –0,14; 0,23. Utworzyć ciąg indeksów jednopodstawowych o podstawie z roku $t = 3$.

8.14. W ostatnich dziesięciu latach średnie roczne tempo wzrostu liczby zarejestrowanych samochodów osobowych w województwie L było równe 8,2%. Jakiej liczby rejestracji należy oczekiwać w roku $T = n + 2$, jeśli w ostatnim roku zarejestrowano ich 8 tysięcy?

8.15. Indeksy łańcuchowe liczby samochodów ciężarowych posiadanych przez firmę transportową przedstawiają się następująco:

Lata	1	2	3	4	5
Indeksy	93	102,1	106,0	105,4	111,0

Wyznacz średnie roczne tempo zmian ilości samochodów ciężarowych w badanym okresie. Ustalić ilość samochodów ciężarowych w roku $T + 1$, jeśli w roku $T + 2$ firma posiadała ich 200.

8.16. Średnioroczne tempo produkcji telewizorów kolorowych w latach od $t = 1$ do $t = 3$ wynosiło 4%. Jakiej wielkości produkcji telewizorów należy oczekiwać w roku czwartym, jeśli wiadomo, że w roku $t = 1$ wyprodukowano 10 tys. sztuk?

8.17. W latach od $t = 1$ do $t = 5$ przewozy pasażerów kolejką linową kształtowały się następująco (w tys. osób): 13,4; 14,0; 16,8; 18,5; 20,0. Zbadać dynamikę przewozów pasażerów za pomocą indeksów łańcuchowych i jednopodstawowych o podstawie z roku $t = 3$. Jakie było średnioroczne tempo wzrostu przewozów w badanym okresie?

8.18. Dynamika dostaw aparatów fotograficznych do sprzedaży detalicznej w mieście X w latach od $t = 1$ do $t = 5$ mierzona indeksami jednopodstawowymi o podstawie z roku $t = 0$ kształtowała się następująco: 104; 117; 135; 156; 174. Obliczyć indeksy łańcuchowe dostaw oraz wyznaczyć średnioroczne tempo wzrostu sprzedaży aparatów w badanym okresie.

8.19. Zatrudnienie w sektorze prywatnym w pewnym województwie w latach $t = 1, 2, 3, 4, 5$, przedstawiało się następująco: 10568; 10980; 11368; 11632; 11744. Obliczyć przyrosty absolutne jednopodstawowe (o podstawie z roku $t = 1$) i łańcuchowe. Dokonać zamiany przyrostów na indeksy jednopodstawowe i łańcuchowe. Jakie było średnioroczne tempo wzrostu zatrudnienia w badanym okresie?

8.20. Tempo przyrostu powierzchni zasiewów pszenicy (w tys. ha) w latach od $t = 1$ do $t = 7$ przedstawiało się następująco: +0,08; 0,00; 0,05; –0,03; 0,07; –0,09; 0,14. Obliczyć średnioroczne tempo wzrostu powierzchni zasiewów pszenicy w badanym okresie. Oszacować, jakiej powierzchni zasiewów należy oczekiwać w roku $T = 9$ jeśli wiadomo, że w roku $t = 7$ powierzchnia ta wynosiła 2420 tys. ha.

8.21. W ostatnim kwartale 2001 roku sprzedano 100 samochodów pewnej marki. W kolejnych kwartałach 2002 i 2003 r. w stosunku do IV kwartału 2001 r. sprzedaż wzrastała odpowiednio o: 5%, 10%, 12%, 18%, 20%, 25%, 40% i 60%. W I kwartale 2004 r. sprzedano 200 samochodów. Ile wynosi średni kwartalny wzrost sprzedaży samochodów tej marki?

8.22. Indeksy łańcuchowe charakteryzujące dynamikę liczby miejsc noclegowych w latach od $t = 1$ do $t = 7$ były następujące: 102,2; 98,3; 104,2; 100,7; 95,9; 94,5. Dokonać zamiany tych indeksów na: a) indeksy jednopodstawowe (o podstawie $t = 1$), b) indeksy jednopodstawowe (o podstawie $t = 4$). Obliczyć średnioroczne tempo zmian liczby miejsc noclegowych w badanym okresie.

8.23. W latach od $t = 1$ do $t = 16$ łańcuchowe różnice stosunkowe (tempa przyrostu) przewozów pasażerów drogą morską kształtowały się następująco: 0,05; 0,02; –0,06; –0,09; 0,00; –0,11; 0,05; –0,01; –0,12; –0,09; –0,06; –0,16; –0,26; –0,10; –0,25; –0,13. Wyznaczyć indeksy o podstawie stałej z roku $t = 5$. Obliczyć średnioroczne tempo zmian przewozów pasażerów w badanym okresie. Oszacować liczbę osób przewiezionych w roku $t = 16$, jeśli w roku $t = 15$ przewieziono ich 218 tys.

8.24. Dynamikę rozmiarów sprzedaży soków owocowych firmy „Hortex" w latach od $t = 1$ do $t = 5$ charakteryzuje następujący ciąg indeksów łańcuchowych: 120%, 180%, 230%, 190%. Jakie było średnioroczne tempo wzrostu sprzedaży soków owocowych w badanym okresie?

8.25. W przedsiębiorstwie X produkowane są 2 wyroby. Kierownictwo tego przedsiębiorstwa zaplanowało zwiększenie produkcji w roku $t = 1$ w stosunku do roku $t = 0$ o 5%. Do analizy wykonania planu przygotowano następujące informacje:

Wyroby	Produkcja w tonach		Cena w tys. zł/t	
	$t = 0$	$t = 1$	$t = 0$	$t = 1$
A	200	250	5	8
B	300	280	7	6

Czy przedsiębiorstwo wykonało założenia planu w zakresie ilości produkcji? Jaki był wpływ cen na zmiany w wartości produkcji?

8.26. W pewnej firmie zmiany w wartości produkcji w roku $t = 1$ w porównaniu z rokiem $t = 0$ przedstawiały się następująco:

Wyroby	Wartość produkcji w tys. zł w roku $t = 0$ w cenach bieżących	Produkcja w tys. sztuk	
		$t = 0$	$t = 1$
A	120	15	18
B	100	10	11
C	80	10	9

Wiedząc, że w roku $t = 1$ wartość produkcji wyniosła 360 tys. zł, ustalić łączne zmiany wartości, ilości i cen produkcji w roku $t = 1$ w porównaniu do roku $t = 0$.

8.27. Wartość sprzedanych telewizorów zmalała o 2,8% w stosunku do roku poprzedniego, natomiast cena telewizorów wzrosła o 8,3%. Jak zmieniła się ilość sprzedanych telewizorów?

8.28. Obroty materiałami budowlanymi w pewnym sklepie w roku t kształtowały się następująco:

Materiały	A	B	C
Obroty w mln zł	0,4	0,8	0,2

Wiadomo ponadto, że ceny materiału A w roku t w porównaniu z rokiem poprzednim zmalały o 5%, materiału B – wzrosły o 20%, materiału C – pozostały bez zmian. Łączne obroty materiałami w roku $t = 1$ wynosiły 1 mln zł. Jaki wpływ na dynamikę wartości sprzedaży tych materiałów miały ceny, a jaki zmiany ilości zakupów?

8.29. Na podstawie poniższych danych ocenić działalność przedsiębiorstwa X w okresie badanym w stosunku do roku podstawowego?

Wyroby	A	B	C
Wartość produkcji w tys. zł w okresie badanym	4,0	2,1	1,0
Indywidualne indeksy cen	0,8	1,4	1,0

Wiadomo ponadto, że łączna wartość produkcji w okresie podstawowym w cenach bieżących wynosiła 9 tys. zł.

8.30. Wartość produkcji w okresie badanym w cenach bieżących była wyższa o 5% od wartości produkcji w roku ubiegłym. O ile procent wzrosła produkcja ilościowo, a jak zmieniły się ceny, jeżeli:

a) wartość produkcji w roku ubiegłym w cenach z roku ubiegłego wyrobu A wynosiła 100 tys. zł, wyrobu B – 50 tys. zł, wyrobu C – 250 tys. zł,

b) indeksy łańcuchowe cen wynosiły odpowiednio: 105%, 120%, 90%.

8.31. Korzystając z poniższych danych, oblicz agregatowy indeks cen:

Wyroby	A	B	C	D
Udział wartości produkcji w okresie bazowym	0,3	0,4	0,2	0,1
Indywidualne indeksy cen	3,2	2,5	1,0	0,4

8.32. Wydatki z dochodów osobistych ludności na dobra materialne (liczone wg cen stałych z roku bazowego) w roku bazowym wynosiły 1050,7 mln zł, natomiast

w roku badanym – 1061,9 mln zł. O ile procent wzrosło spożycie dóbr materialnych w analizowanym okresie?

8.33. Na podstawie poniższych danych dokonaj analizy zmian dynamiki produkcji w pewnym przedsiębiorstwie:

Detale	I	II	III
Wartość produkcji w cenach bieżących w roku badanym (w tys. zł)	4,0	2,1	4,5
Indywidualne indeksy ilości	1,0	0,7	1,5
Indywidualne indeksy cen	0,8	1,4	1,0

8.34. Na podstawie poniższych danych dokonaj porównania działalności gospodarczej przedsiębiorstwa X w okresie sprawozdawczym w stosunku do okresu bazowego:

Wyroby	A	B	C
Ilość produkcji	200	50	100
Ceny jednostkowe w okresie bazowym	6	30	20
Ceny jednostkowe w okresie badanym	8	20	15

8.35. Obrót artykułem A zmniejszył się z 200 tys. zł w okresie bazowym do 190 tys. zł w okresie badanym, artykułem B – wzrósł z 400 tys. zł do 450 tys. zł. Jak zmieniła się ilość sprzedaży obydwu artykułów łącznie w okresie badanym w porównaniu z okresem bazowym, jeśli cena towaru A zmniejszyła się o 5%, a towaru B o 10%?

8.36. Na podstawie poniższych danych oblicz agregatowe indeksy wartości, cen i ilości:

Wyroby	A	B	C
Wielkość produkcji w sztukach w okresie bazowym	590	459	18
Wielkość produkcji w sztukach w okresie badanym	1393	487	23
Wartość produkcji w tys. zł w okresie bazowym	31,5	4,7	0,2
Wartość produkcji w tys. zł w okresie badanym	48,8	5,4	2,2

8.37. Obroty w pewnym przedsiębiorstwie w roku bazowym wynosiły 1 mln zł. W okresie badanym wartość obrotów dla wyrobu A wynosiła 800 mln zł, wyrobu B – 800 mln zł, a wyrobu C – 200 mln zł. Cena wyrobu A w okresie badanym w porównaniu okresem bazowym spadła o 5%, wyrobu B – wzrosła o 10%, a wyrobu C – nie zmieniła się. Jaki wpływ na dynamikę wartości wywarły ceny, a jaki ilości?

8.38. Wielkość produkcji trzech wyrobów (w tys. sztuk) oraz ceny jednostkowe (w zł) w okresie bazowym i badanym przedstawiały się następująco:

Wyroby	A	B	C
Produkcja w okresie bazowym	100	150	450
Produkcja w okresie badanym	140	200	300
Ceny w okresie bazowym	5	4	9
Ceny w okresie badanym	7	6	8

Oblicz i zinterpretuj agregatowe indeksy cen, ilości i wartości.

8.39. W okresie badanym obroty w sklepie warzywniczym wzrosły o 40% w porównaniu z okresem bazowym. Przy stałych cenach z okresu bazowego obroty wzrosłyby o 5%. W jakim stopniu na wzrost obrotów w okresie badanym wpłynęła dynamik cen sprzedawanych artykułów?

8.40. W okresie badanym wartość produkcji przedsiębiorstwa wynosiła 125 tys. zł. Ta sama wartość produkcji liczona w cenach z okresu bazowego wynosiła 120 tys. zł. W okresie bazowym wartość produkcji wynosiła 95 tys. zł. Jaki był wzrost wartości produkcji w okresie badanym i co spowodowało ten wzrost?

8.41. Na podstawie poniższych danych oblicz agregatowe indeksy wartości, cen i ilości:

Wyroby	A	B	C
Odsetek wartości z okresu badanego	50	20	30
Indeksy cen (w %)	108	90	110
Indeksy ilości (w %)	105	95	100

8.42. W badanym roku średnia wydajność pracy w przedsiębiorstwie wzrosła – w porównaniu z okresem bazowym – o 25%. Wiadomo, że przy założeniu stałej struktury zatrudnienia na poziomie okresu bazowego średnia wydajność pracy spadałaby o 6%. W jaki stopniu dynamika przeciętnej wydajności pracy w tym przedsiębiorstwie była uzależniona od zmian w strukturze zatrudnienia?

8.43. Plony czterech zbóż spadły o 10% w roku sprawozdawczym w porównaniu do roku podstawowego. Gdyby struktura powierzchni zasiewów nie uległa zmianie, to plony spadłyby przeciętnie o 20%. Jaki wpływ na dynamikę przeciętnych plonów miały zmiany w strukturze zasiewów?

8.44. Wydajność pracy zwiększyła się w okresie badanym w porównaniu do okresu podstawowego o 56%. Indeks wydajności pracy o stałej strukturze czasu pracy według formuły Laspeyresa wynosi 1,2. Oblicz indeks wpływu zmian w strukturze czasu pracy na dynamikę wydajności pracy.

8.45. W okresie badanym średnia płaca w pewnym przedsiębiorstwie wzrosła w porównaniu z okresem bazowym o 20%. Przy założeniu stałej struktury zatrudnienia na poziomie okresu podstawowego przeciętna płaca spadłaby przeciętnie o 5%. Ustalić wpływ zmian w strukturze zatrudnienia na dynamikę przeciętnej płacy w tym przedsiębiorstwie.

8.46. Czas zużyty na produkcję wyrobów A i B w okresie badanym zmniejszył się w porównaniu z okresem podstawowym o 3%, produkcja – w tym samym okresie – wzrosła o 5%. W okresie badanym w porównaniu z okresem podstawowym nastąpiły zmiany w strukturze produkcji tych wyrobów, które spowodowały spadek pracochłonności ogólnej o 2%. Jaka byłaby dynamika pracochłonności ogólnej, gdyby nie było zmian w strukturze produkcji?

8.47. Przeciętne koszty produkcji wyrobów X i Y w okresie badanym wzrosły w porównaniu z okresem podstawowym o 5%. Poziom średnich kosztów produkcji wyrobu X wzrósł z 300 zł w okresie bazowym do 360 zł w okresie badanym, a wyrobu Y – z 250 zł do 270 zł. W okresie badanym udział produkcji wyrobu X stanowił 60% ogólnej produkcji obu wyrobów łącznie. Jaki był wpływ zmian w strukturze produkcji na dynamikę przeciętnych kosztów obu wyrobów łącznie?

8.48. Liczba stacji benzynowych na 1 tys. km dróg (gęstość stacji benzynowych) wzrosła w roku badanym w stosunku do roku bazowego o 16%. Gdyby terytorialna (w znaczeniu podziału na stacje miejskie i pozamiejskie) struktura stacji benzynowych w obu latach była taka sama jak w roku podstawowym, to średnia ich gęstość wzrosłaby o 120%. Jaki wpływ na średnią dynamikę gęstości stacji miały zmiany w terytorialnej strukturze stacji benzynowych?

8.49. Jak zmieniła się przeciętna płaca w całym przedsiębiorstwie, jeżeli:

a) przeciętna płaca w całym przedsiębiorstwie wzrosła w okresie badanym o 9,4% w stosunku do okresu podstawowego, przy założeniu stałej struktury zatrudnienia z okresu badanego,

b) cząstkowe przeciętne płace byłyby stałe na poziomie okresu podstawowego, to przeciętna płaca w całym przedsiębiorstwie spadałby o 11,7% w okresie badanym w stosunku do okresu podstawowego.

8.50. Ocenić wpływ różnic w strukturze wieku kobiet w okresie rozrodczym na płodność (iloraz urodzeń żywych i liczby kobiet w wieku rozrodczym) kobiet w dwóch regionach, A i B, jeśli wartość indeksu o stałej strukturze wieku kobiet typu Paaschego wynosiła 88%, a rzeczywisty ogólny współczynnik płodności w regionie A był o 3,2% niższy w porównaniu z regionem B.

ODPOWIEDZI DO ZADAŃ

Rozdział II

2.1.

Powierzchnia sklepu w m^2	Liczba sklepów
52–59	5
60–67	7
68–75	9
76–83	5
84–91	2
92–99	2

$\bar{x} = 70{,}97\ m^2$.

2.2. $\bar{x} = 95{,}86\ m^2$.

2.3. $\bar{x} = 1{,}15$ usterki.

2.4. 186,12 tys. zł.

2.5. $\bar{x} = 1{,}4$ nieobecności.

2.6. $\bar{x} = 3{,}93$ zł.

2.7. $\bar{x} = 39{,}6$ lat; $N = 600$ pracowników.

2.8. 1030 zł.

2.9. $\bar{x} = 70$ zł.

2.10. Niższa od średniej o 65 zł.

2.11. $N = 7$ uczniów.

2.12. $\bar{x} = 9{,}86$ lat.

2.13. Przesunąć pracownika zarabiającego 2000 zł z działu II do I.

2.14. Nie, gdyż 2480 zł > 2000 zł.

2.15. $\bar{\bar{x}} = 4{,}18$.

2.16. $\bar{\bar{x}} = 9{,}68$ min.

2.17. $x_1 n_1 = 5$, $x_2 n_2 = 30$, $x_3 n_3 = 100$.

2.18. $\bar{x} = 2312$ zł, $\sum^{*} x_i n_i = 57\,800$ zł.

2.19. $\bar{x} = 4{,}8$ ha.

2.20. 44 punkty.

2.21. a) wzrośnie o 1,4, b) wzrośnie o 8,4.

2.22. $\bar{\bar{x}} = 3400$ zł.

2.23. 16 kg i 24 kg.
2.24. $\bar{x} = 7{,}075$ ha.
2.25. $\bar{\bar{x}} = 1842{,}40$ zł.
2.26. $H = 64$ km/h.
2.27. 45 km/h.
2.28. $H = 34{,}3$ km/h.
2.29. 2,25 razy.
2.30. $H = 9{,}17$ zł/kg.
2.31. $H = 62{,}3$ osób/km^2.
2.32. $H = 4$ zł/sztukę.
2.33. $H = 0{,}32$ godz./wyrób.
2.34. $H = 1{,}27$ zł/kg.
2.35. $H = 18{,}46$ km/h.
2.36. $H = 1161$ osób/km^2.
2.37. $H = 54$ osoby/km^2 oraz 1666,67 km^2.
2.38. Szereg punktowy: $\bar{x} = 16{,}43$ szt./godz., $D = Me = 17$ szt./godz., $Q_1 = 15$ szt./godz., $Q_3 = 18$ szt./godz.
Szereg przedziałowy: $\bar{x} = 16{,}2$ szt./godz., $Me = 16{,}36$ szt./godz., $D = 16{,}8$ szt./godz.
2.39. $\bar{x} = 4{,}1$ osób, $Me = 4$ osoby, $D = Q_1 = 3$ osoby, $Q_3 = 5$ osób.
2.40. $D = 23$ kg, $Me = 22{,}5$ kg.
2.41. $\bar{x} = 7$ osób, $D = 8{,}5$ osoby, $Me = 7{,}25$ osób.
2.42. $\bar{x} = 7{,}3$ tys. zł, $D = 8$ tys. zł, $Q_1 = 6{,}3$ tys. zł, $Me = 7{,}7$ tys. zł, $Q_3 = 8{,}6$ tys. zł.
2.43. $D = 181{,}26$ cm.
2.44. $\bar{x} = 46{,}9$ zł.
2.45. $\bar{x} = 15{,}45$ zł, $D = Me = 15$ zł.
2.46. 12 skoków.
2.47. 10%.
2.48. 20%.
2.49. 100 pracowników.
2.50. 68 pracowników.
2.51. $V_S = 19{,}1\%$.
2.52. 5 min $< x_{typ} <$ 9 min; 64,5%.
2.53. $V(x) = 25\%$, $V(y) = 70\%$, pod względem czasu trwania małżeństwa.
2.54. $V = 28{,}125\%$.
2.55. $V_A = 19{,}6\%$, $V_B = 57{,}8\%$. Regularniejsze wyniki osiągał zawodnik A.
2.56. $V_K = 38{,}8\%$, $V_P = 42{,}3\%$.
2.57. 1605,87 zł $< x_{typ} <$ 1833,01 zł.
2.58. Mniejsze zróżnicowanie wydajności pracy było na wydziale I ($V_I = 8{,}74\%$, $V_{II} = 11{,}66\%$).
2.59. $A_S = -0{,}088$.
2.60. $A_S = 0{,}333$.
2.61. $\bar{x} = 350$ zł, $s = 50$ zł.
2.62. Grupa I: $Q = 12$ zł, $A_S = -0{,}67$. Grupa II: $Q = 11{,}25$ zł, $A_S = -0{,}74$.
2.63. $V_S = 26{,}77\%$.
2.64. $V_S = 27{,}3\%$, 26,9 lat $< x_{typ} <$ 47,1 lat.
2.65. Sklepy spożywcze: $\bar{x} = 7{,}3$ tys. zł, $Me = 7{,}7$ tys. zł, $D = 8$ tys. zł, $s = 2$ tys. zł, $V_S = 27{,}4\%$, $A_S = -0{,}35$.

2.66. $A_S = 1{,}11$.
2.67. $A_S = 0{,}44$.
2.68. 33%.
2.69. Wydział I: $\bar{x} = D = Me = 110\%$, Wydział II: $V_S = 20\%$.
2.70. Kobiety: $V_S = 23{,}3\%$, $A_S = -0{,}31$. Mężczyźni: $V_S = 23{,}7\%$, $A_S = -0{,}69$.
2.71. Studia dzienne: $Me = 19{,}75$ lat, $s = 2$ lata, $A_S = 0{,}5$.
Studia zaoczne: $\bar{x} = 25$ lat, $V_S = 8\%$, $A_S = 0$.
2.72. $A_S = 0{,}86$.
2.73. Dla autobusu nr 24 mamy: $\bar{x} = 7{,}5$ min., $s = 5{,}14$ min., $D = 3{,}2$ min., $A_S = 0{,}837$, $V_S = 68{,}53\%$.
2.74. $\bar{x} = 10$ minut, $D = 10$ minut, $Me = 10$ minut, $Q_1 = 9$ minut, $Q_3 = 13$ minut, $s = 4{,}38$ minut, 5,62 minuty $< x_{typ} <$ 14,38 minuty, $V_S = 43{,}8\%$, $A_S = 0$.
2.75.

Zmiany	I	II	III
$\bar{x}$	110%	110%	110%
D	110%	102,3%	–
Me	110%	108,3%	104,2%
Q_1	104%	100,6%	99,3%
Q_3	116%	118,7%	122,3%
s	9,3%	11,9%	13%
A_S	0	0,65%	0,55%
V_S	0,4%	10,8%	11,8%

2.76. Zakład I: $\bar{x} = 10$ szt./godz., $Me = 9$ szt./godz., $D = 8$ szt./godz., $Q_1 = 7$ szt./godz., $Q_3 = 12$ szt./godz., $Q = 2{,}5$ szt./godz., $V_Q = 28\%$, $A_S = 0{,}2$, $s = 4$ szt./godz.
Zakład II: $Me = 7$ szt./godz., $Q_1 = 5{,}3$ szt./godz., $Q_3 = 9$ szt./godz., $Q = 1{,}85$ szt./godz., $V_Q = 26\%$, $A_S = 0{,}08$.
2.77. $\bar{\bar{x}} = 1660$ zł, $V_S = 13{,}84\%$, $1430{,}28 < x_{typ} < 1889{,}72$.
2.78. $\bar{x} = 32{,}4$ lat, $s = 4{,}86$ lat, 27,54 lat $< x_{typ} <$ 37,26 lat, $Me = 32{,}36$ lat.
2.79. $V_S = 31{,}86\%$.
2.80. $A_S = -0{,}237$.
2.81. $A_S = 0{,}417$.
2.82. 29,1 zł $< x_{typ} <$ 49,1 zł.
2.83. $\bar{x} = 10{,}48$, $V_S = 90{,}74\%$, $0{,}97 < x_{typ} < 19{,}99$, $s = 9{,}51$, $A_S = 1{,}701$.
2.84. a) nie, b) nie, c) nie, d) tak.
2.85. W zakładzie II: $\bar{x} < 2000$ zł, $V_S^I < V_S^{II}$.
2.86. $k = 0{,}69$.
2.87. Rozkład leptokurtyczny.
2.88. $k = 0{,}25$.
2.89. $k = 0{,}375$.
2.90. $k = 0{,}52$.
2.91. $k = 0{,}22$.
2.92. $k = 0{,}19$.
2.93. Tak, $k = 0{,}2$.
2.94. Współczynnik kurtozy wynosi 3,38. Współczynnik ekscesu jest równy 0,38.
2.95. Większe skupienie.
2.96. $K = 0{,}251$.

2.97. $K = 0{,}02$.

2.98. Współczynnik ekscesu: –0,127. Rozkład platokurtyczny.

2.99. $K = 702$.

2.100. Współczynnik ekscesu: –0,742. Rozkład platokurtyczny.

Rozdział III

3.1.

x_i	–1	0	1
p_i	0,4	0,1	0,5

3.2. $D^2(X) = 69$.

3.3. $E(Y) = 25$; $D(Y) = 4$.

3.4. $E(X) = 31$; $D^2(X) = 129$.

3.5. $E(X) = 3{,}9$; $D^2(X) = 87{,}89$.

3.6.

y_i	–1	1	3	5
p_i	0,25	0,15	0,5	0,1

3.7.

$$F(x)=\begin{cases}0 & \text{dla} \quad x\le 0\\ \frac{1}{8} & \text{dla} \quad 0\le x\le 1\\ \frac{4}{8} & \text{dla} \quad 1<x\le 2\\ \frac{7}{8} & \text{dla} \quad 2<x\le 3\\ 1 & \text{dla} \quad x>3\end{cases}$$

3.8.

x_i	–100	100
p_i	$\frac{2}{3}$	$\frac{1}{3}$

3.9. $E(X) = 0{,}7$; $D^2(X) = 1{,}21$.

3.10. $p_1 = 0{,}2$; $p_2 = 0{,}3$; $p_3 = 0{,}5$.

3.11.

x_i	1	2
p_i	0,6	0,4

3.12. $x_2 = 4$; $p_2 = 0{,}05$; $D(X) = 4{,}58$.

3.13. $D(X) = \sqrt{2{,}9166} = 1{,}7$.

3.14. $P(X \le 2) = 0{,}75$; $P(X > 1) = 0{,}5$.

3.15.

y_i	0	1
p_i	1/3	2/3

$D^2(Y) = \frac{2}{9}$; $P(|X| < 0{,}5) = \frac{1}{3}$.

3.16.

x_i	0	1	2	3	4	5
p_i	0,59049	0,32805	0,0729	0,0081	0,00045	0,00001

$E(X) = 0{,}5$.

3.17.

x_i	0	1	2	3	4
p_i	0,6561	0,2916	0,0486	0,0036	0,0001

3.18.

x_i	1	2	3	4
p_i	$\frac{1}{30}$	$\frac{4}{30}$	$\frac{9}{30}$	$\frac{16}{30}$

a) $P(X < 4) = 0{,}467$; b) $P(X > 2) = 0{,}833$; c) $P(2 < X < 5) = 0{,}833$.

3.19. $P(X=3)=\binom{4}{3}\left(\frac{1}{3}\right)^3\cdot\frac{2}{3}=\frac{8}{81}=0{,}0988$.

3.20.

$$F(x)=\begin{cases}0 & \text{dla} \quad x\le 2\\ \frac{1}{3} & \text{dla} \quad 2<x\le 3\\ \frac{5}{6} & \text{dla} \quad 3<x\le 4\\ 1 & \text{dla} \quad x>4\end{cases}$$

3.21.

y_i	–2	0	4
p_i	$\frac{1}{3}$	$\frac{1}{3}$	$\frac{1}{3}$

3.22.

$$F(x)=\begin{cases}0 & \text{dla} \quad x\le 1\\ \frac{1}{15} & \text{dla} \quad 1<x\le 2\\ \frac{3}{15} & \text{dla} \quad 2<x\le 3\\ \frac{6}{15} & \text{dla} \quad 3<x\le 4\\ \frac{10}{15} & \text{dla} \quad 4<x\le 5\\ 1 & \text{dla} \quad x>5\end{cases}$$

3.23. a) $\frac{1}{6}$; b) $\frac{1}{3}$; c) $\frac{1}{2}$.

3.24.

x_i	0	1	2	3	4	5
p_i	0,16807	0,36015	0,3087	0,1323	0,02835	0,00243

3.25. $x_3 = 10$; $p = 0{,}5$; $D^2(X) = 7$.

3.26.

x_i	0	1	2
p_i	0,25	0,5	0,25

3.27. $q = -3$ lub $q = 2$.

3.28.

x_i	1	2	5
p_i	0,2	0,4	0,4

3.29.

$$F(x) = \begin{cases} 0 & \text{dla} \quad x \le -1 \\ \frac{2}{7} & \text{dla} \quad -1 < x \le 2 \\ \frac{6}{7} & \text{dla} \quad 2 < x \le 5 \\ 1 & \text{dla} \quad x > 5 \end{cases}$$

3.30. $n = 16$; $p = 0,75$.

3.31.

x_i	0	1	2	3
p_i	0	0,2	0,6	0,2

3.32. $E(X) = 3,75$; $D^2(X) = 0,9375$.

3.33. $\lambda = 2$; $P(X \ge 4) = 0,143$.

3.34. a) $P(X \le 1) = 0,736$; b) $P(X \ge 1) = 0,632$.

3.35. a) $P(X = 0) = 0,0022$;

b) $P(X > 0) = 1 - P(X = 0) = 0,9978$;

c) $P(X < 4) = P(X = 0) + P(X = 1) + P(X = 2) + P(X = 3) = 0,0022 + 0,0174 + 0,0639 + 0,1419 = 0,2254$.

3.36. a) $P(X = 0) = 0,0498$;

b) $P(X \ge 1) = 1 - P(X = 0) = 0,9502$;

c) $F(4) = P(X < 4) = P(X = 0) + P(X = 1) + P(X = 2) + P(X = 3) = 0,0498 + 0,1494 + 0,2242 + 0,2241 = 0,6474$;

d) $P(0 < X \le 3) = 0,6474$.

3.37. $P(X \le 1) = P(X = 0) + P(X = 1) = 0,6066 + 0,3033 = 0,9099$.

3.38. a) 17; b) 25; c) 5.

3.39. $P(X \le 2) = 0,6768$.

3.40. Rozkład prawdopodobieństwa $Y = X^3$:

y_i	-1	0	1
p_i	$\frac{2}{3}$	$\frac{1}{6}$	$\frac{1}{6}$

a) $D^2(Y) = 0,584$; b) $P(|X| < 1,5) = 1$

3.41. $P(X \le 2) = 0,238$.

3.42. a) $P(X = 5) = 0,246$;

b) $P(3 \le X \le 8) = 0,934$.

3.43.

x_i	0	1	2	3	4
p_i	0,885	0,110	0,005	0,000	0,000

3.44. $E(X) = 5,8$; $D^2(X) = 8,36$.

3.45. $P(X > 3) = 1 - P(X \le 3) = 1 - 0,981 = 0,019$.

3.46.

x_i	1	2	3	4	5	6
p_i	0,05	0,30	0,40	0,15	0,007	0,03

$P(X \ge 5) = 0,1$.

3.47. $P(X = 5) = 0,2$.

3.48. $E(X) = 1$; $D(X) = 1$.

3.49. $a = 4$.

3.50. $P(X = 2) = 0$.

3.51. $f(x) = \frac{1}{4\sqrt{2\pi}} e^{\cdot \frac{-(x-3)^2}{32}}$.

3.52. $D(X) = 0,58$.

3.53. $P(|X| > 2) = 0,3174$.

3.54. $C = 2$; $E(X) = 1$; $D(X) = 0,45$.

3.55. $F(x) = \begin{cases} 0 & \text{dla} \quad x \le 1 \\ 0,5\,(x-1) & \text{dla} \quad 1 < x \le 3 \\ 1 & \text{dla} \quad x > 3 \end{cases}$

a) $P(X < 2) = 0,5$; b) $P(X > 2,5) = 0,25$; c) $F(1,5) = 0,25$.

3.56. $x = 2,05$.

3.57. 0,725.

3.58. 32 pracowników.

3.59. 12 studentów.

3.60. a) 0,5; b) 0,023; c) 0,683; d) 0,954; e) 0,317; f) 0,84.

3.61. $P(X \le 4) = 0,996$.

3.62.

$$F(x) = \begin{cases} 0 & \text{dla} \quad x < 0 \\ \frac{1}{6}x^2 & \text{dla} \quad 0 \le x \le \sqrt{6}; \\ 1 & \text{dla} \quad x > \sqrt{6} \end{cases}$$

$P(|X| < 1) = \frac{1}{6}$.

3.63. $E(X) = 2$.

3.64. a) 0,1587; b) 0,872; c) 0,2254.

3.65. 14,53%.

3.66. $a = 1,28$.

3.67. 0,97725.

3.68.

$$F(x) = \begin{cases} 0 & \text{dla} \quad x \le 0 \\ \frac{1}{16}x^2 & \text{dla} \quad 0 < x \le 4 \\ 1 & \text{dla} \quad x > 4 \end{cases}$$

$C = \frac{1}{8}$; $P(1 \le X \le 2) = F(2) - F(1) = \frac{4}{16} - \frac{1}{16} = 0,1875$.

3.69. $E(X) = 3$; $D(X) = \sqrt{3} = 1,732$.

3.70. a) 0,47725; b) 0,02275; c) 0,3085; d) 0,6826.

3.71. $P(|X| > 2) = 0,3753$.

3.72.
$$F(x)=\begin{cases}0 & \text{dla} \quad x<0\\ \frac{x^2}{9} & \text{dla} \quad 0\le x\le 3;\\ 1 & \text{dla} \quad x>3\end{cases} \qquad C=\frac{2}{9}.$$

3.73. $E(X) = 0{,}04$.

3.74. $E(X) = D^2(X) = 0{,}5$.

3.75. $C=\frac{1}{9}$; $P(1\le X\le 4)=\frac{26}{27}=0{,}963$.

3.76. $f(x)=\frac{1}{4\sqrt{\pi}}e^{\frac{-(x-3)^2}{32}}$.

3.77. $P(15 < X < 25) = 0{,}0062$.

3.78. $P(|X|>3) = 0{,}18145$.

3.79. 0,627%.

3.80. $a=\frac{1}{8}$; $P(X < 2) = F(2) = 0{,}25$.

3.81. a) 0,8186; b) 0,8186.

3.82. 23 wyroby.

3.83. a) 0,001; b) $s = 1{,}41$.

3.84. a) 0,023; b) 0,0027.

3.85. 1159 studentów.

3.86. a) 1,64; b) 1,96.

3.87. $x = 14{,}08$ jednostek gazu.

3.88. $x = 103{,}5$.

3.89. $x_1 = 833$; $x_2 = 436$.

3.90. 2275 porcji.

3.91. Po 15,87% klientów.

3.92. 99,72% golarek.

3.93. $x = 107{,}8$.

3.94. 92,08 minut.

3.95. 2,275%.

3.96. 16,85%.

3.97. 97,725%.

3.98. 77,34%.

3.99. 2,425 minuty.

3.100. 29,5 minuty.

Rozdział IV

4.1. $2847 < m < 2913$.

4.2. $66 < m < 70$.

4.3. $7 < m < 9$.

4.4. $117 < m < 123$.

4.5. $4{,}6 < m < 6{,}4$.

4.6. 0,9545.

4.7. $13{,}6 < m < 21{,}6$.

4.8. $503{,}81 < m < 552{,}19$.

4.9. $9{,}39 < m < 10{,}61$.

4.10. $0{,}086 < \sigma < 0{,}314$.

4.11. $34{,}35\% < p < 45{,}65\%$.

4.12. $3{,}39 < \sigma < 4{,}55$.

4.13. $5{,}188 < m < 5{,}212$.

4.14. $63{,}48\% < p < 74{,}92\%$.

4.15. $31{,}832 < m < 32{,}168$.

4.16. $0{,}16 < \sigma < 0{,}50$.

4.17. $73{,}5\% < p < 76{,}5\%$.

4.18. $76{,}2\% < p < 78{,}6\%$.

4.19. $33{,}3\% < p < 40{,}7\%$.

4.20. $1-\alpha = 0{,}95$.

4.21. $8{,}94 < \sigma < 11{,}06$.

4.22. $15\% < p < 31\%$.

4.23. $43{,}2 < m < 45{,}2$.

4.24. $176{,}55 < \sigma^2 < 607{,}77$.

4.25. $52{,}8 < m < 67{,}2$ oraz $149{,}3 < \sigma^2 < 444{,}8$.

4.26. $210{,}47 < m < 252{,}19$ oraz $602{,}73 < \sigma^2 < 2592{,}06$.

4.27. $71{,}9 < \sigma < 90{,}2$.

4.28. $33{,}02\% < p < 38{,}98\%$.

4.29. $21{,}7 < \sigma^2 < 109{,}4$.

4.30. $64{,}5\% < p < 75{,}5\%$.

4.31. $2{,}59 < \sigma^2 < 8{,}54$ oraz $1{,}61 < \sigma < 2{,}92$.

4.32. $n = 226$.

4.33. $n = 9$.

4.34. a) $69\% < p < 81\%$; b) $n = 331$; c) $8{,}44 < \sigma < 10{,}28$.

4.35. $n = 97$.

4.36. $n = 1329$.

4.37. $n = 1068$.

4.38. $n = 262$ (dolosować 226 pracowników).

4.39. $n = 545$.

4.40. $n = 17$.

4.41. $n = 455$ (dolosować 311 klientów).

4.42. $n = 421$.

4.43. $n = 10$.

4.44. $n = 1937$.

4.45. $n = 226$.

4.46. $n = 9$.

4.47. $n = 39$.

4.48. $n = 628$.

4.49. $n = 67$.

4.50. $n = 73$.

Rozdział V

5.1. $z = -10 < -z_\alpha = -2{,}33$. H_0 odrzucić.

5.2. $z = -7{,}01 < -z_\alpha = -1{,}64$. H_0 odrzucić.

5.3. $|z| = 5 > |z_\alpha| = 2{,}33$. H_0 odrzucić.

5.4. $z = 2{,}25 > z_\alpha = 1{,}96$. H_0 odrzucić.

5.5. a) $z = -10{,}33 < -z_\alpha = -2{,}33$. H_0 odrzucić; b) $z = 2{,}02 < -z_\alpha = 2{,}33$. Brak podstaw do odrzucenia H_0.
5.6. a) $z = -4{,}183 < -z_\alpha = -1{,}75$. H_0 odrzucić; b) $z = -0{,}6 < -z_\alpha = -1{,}75$. Brak podstaw do odrzucenia H_0.
5.7. $F = 1{,}063 < F_\alpha = 2{,}82$. Brak podstaw do odrzucenia H_0.
5.8. $|z| = 11{,}88 > |z_\alpha| = 1{,}96$. H_0 odrzucić.
5.9. $\chi^2 = 28{,}736 > \chi^2_\alpha = 24{,}996$. H_0 odrzucić.
5.10. a) $F = 1{,}653 < F_\alpha = 1{,}84$. Brak podstaw do odrzucenia H_0. b) $t = 2{,}882 > t_\alpha = 1{,}67$. H_0 odrzucić.
5.11. $z = -2{,}58 > -z_\alpha = -1{,}65$. H_0 odrzucić.
5.12. $z = -3 < -z_\alpha = -1{,}75$. H_0 odrzucić.
5.13. $z = 1{,}667 > z_\alpha = 1{,}65$. H_0 odrzucić.
5.14. $|z| = 1{,}938 > |z_\alpha| = 1{,}64$. H_0 odrzucić.
5.15. $|t| = 2{,}473 < |t_\alpha| = 2{,}5$. Brak podstaw do odrzucenia H_0.
5.16. $|z| = 3 > |z_\alpha| = 2{,}58$. H_0 odrzucić.
5.17. $t = 1{,}38 < t_\alpha = 1{,}8$. Brak podstaw do odrzucenia H_0.
5.18. $|z| = 1{,}75 < z_\alpha = 1{,}96$. Brak podstaw do odrzucenia H_0.
5.19. $|t| = 0{,}27 < t_\alpha = 1{,}89$. Brak podstaw do odrzucenia H_0.
5.20. $|z| = 0{,}264 < |z_\alpha| = 1{,}96$. Brak podstaw do odrzucenia H_0.
5.21. $z = 4{,}35 > z_\alpha = 2{,}33$. H_0 odrzucić.
5.22. $|z| = 3{,}17 > |z_\alpha| = 1{,}96$. H_0 odrzucić.
5.23. $z = 0{,}45 < z_\alpha = 1{,}64$. Brak podstaw do odrzucenia H_0.
5.24. a) $|z| = 0, < z_\alpha = 1{,}96$. Brak podstaw do odrzucenia H_0. b) $z = 5{,}964 > z_\alpha = 1{,}64$. H_0 odrzucić.
5.25. $|z| = 2{,}52 > |z_\alpha| = 1{,}96$. H_0 odrzucić.
5.26. $|z| = 1{,}144 < |z_\alpha| = 1{,}96$. Brak podstaw do odrzucenia H_0.
5.27. $z = -1{,}538 > z_\alpha = -1{,}75$. H_0 odrzucić.
5.28. $|t| = 2{,}961 > |t_\alpha| = 1{,}703$. H_0 odrzucić.
5.29. $t = -100 < -t_\alpha = -2{,}358$. H_0 odrzucić.
5.30. $z_\alpha = 1{,}28$.
5.31. $|z| = 0{,}27 < |z_\alpha| = 1{,}96$. Brak podstaw do odrzucenia H_0.
5.32. $F = 2{,}294 > F_\alpha = 1{,}6$. H_0 odrzucić.
5.33. $|z| = 0{,}57 < |z_\alpha| = 1{,}96$. Brak podstaw do odrzucenia H_0.
5.34. $|t| = 1{,}76 < |t_\alpha| = 2{,}67$. Brak podstaw do odrzucenia H_0.
5.35. $z = 2{,}631 > z_\alpha = 1{,}64$. H_0 odrzucić.
5.36. $|t| = 3{,}065 > |t_\alpha| = 2{,}819$. H_0 odrzucić.
5.37. $z = -2{,}08 < -z_\alpha = -1{,}65$. H_0 odrzucić.
5.38. $|t| = 0{,}393 < |t_\alpha| = 1{,}734$. Brak podstaw do odrzucenia H_0.
5.39. $F = 16{,}3 > F_\alpha = 6{,}16$. H_0 odrzucić.
5.40. $t = -5{,}128 < -t_\alpha = -1{,}746$. H_0 odrzucić.
5.41. $k_1 \le k$. H_0 odrzucamy (zbyt mała liczba serii).
5.42. $\chi = 10{,}5$. H_0 odrzucić.
5.43. $\chi^2 = 15{,}336$. H_0 odrzucić.
5.44. $\chi^2 = 25{,}43$. H_0 odrzucić.
5.45. $k = 10$. Brak podstaw do odrzucenia H_0.
5.46. $\chi^2 = 43{,}978$. H_0 odrzucić.
5.47. $k_1 = 2 < k = 9 \le k_2 = 9$. Brak podstaw do odrzucenia H_0.
5.48. $\chi^2 = 1{,}788$. Brak podstaw do odrzucenia H_0.

5.49. $k_1 = 4 < k = 8 \le k_2 = 11$. Brak podstaw do odrzucenia H_0.
5.50. $\chi^2 = 374{,}93$. H_0 odrzucamy.

Rozdział VI

6.1. $E(X) = 0{,}4$; $D^2(X) = 0{,}24$; $E(Y) = 2{,}3$; $D^2(Y) = 0{,}61$.
6.2. $\operatorname{cov}(X,Y) = 0{,}36$; $\rho_{xy} = 0{,}678$.
6.3. Tak.
6.4. a) rozkład brzegowy zmiennej X:

x_i	2	3	4	5	6
$p_{i\cdot}$	0,16	0,21	0,23	0,22	0,18

b) rozkład brzegowy zmiennej Y:

y_i	1	2	3
$p_{\cdot j}$	0,11	0,46	0,43

c) rozkład warunkowy zmiennej Y:

y_j	1	2	3
$P(Y\|X=2)$	0,0625	0,375	0,5625

6.5. $p_{11} = 0{,}06$; $p_{12} = 0{,}09$; $p_{13} = 0{,}15$; $p_{21} = 0{,}14$; $p_{22} = 0{,}21$; $p_{23} = 0{,}35$; $P(Y|X=1) = [0{,}2;\ 0{,}3;\ 0{,}5]$; $P(Y|X=2) = [0{,}2;\ 0{,}3;\ 0{,}5]$.
6.6. Są zależne, gdyż: $P(X=5,\ Y=1) = 0 \neq P(X=5)\cdot P(Y=1) = 0{,}4\cdot 0{,}1 = 0{,}04$.
6.7. $P(X=2,\ Y<3) = 0{,}3125$; $P(X<3,\ Y=3) = 0{,}0370$; $P(1<X\le 4;\ 1\le Y\le 3) = 0{,}3385$.
6.8. Zmienne losowe X i Y są niezależne.
6.9. Zmienne losowe X i Y są niezależne.
6.10. Zmienne losowe X i Y są dodatnio skorelowane, gdyż $\operatorname{cov}(X,Y) = 0{,}6 > 0$.
6.11. $C = \frac{4}{9} = 0{,}44(4)$.
6.12. $f_1(x) = \begin{cases} 3x^2 & \text{dla } 0 \le x \le 1 \\ 0 & \text{dla pozostałych } x \end{cases}$, $f_2(x) = \begin{cases} 0{,}5 & \text{dla } 0 \le y \le 2 \\ 0 & \text{dla pozostałych } y \end{cases}$.
6.13. Rozkład brzegowy zmiennej losowej X:

x_i	1	2	3
$p_{i\cdot}$	0,3	0,5	0,2

Rozkład brzegowy zmiennej losowej Y:

y_i	1	2	3	4
$y_{\cdot j}$	0,3	0,3	0,2	0,2

$E(X) = 1{,}9$, $E(Y) = 2{,}3$; $\operatorname{cov}(X,Y) = 0{,}53$; $D^2(X) = 0{,}49$; $D^2(Y) = 1{,}21$; $\rho_{XY} = 0{,}688$.
6.14. Zmienne losowe X i Y są niezależne.
6.15. $r^2_{xy} = 0{,}916$.
6.16. $r^2 = 92{,}31\%$.
6.17. $r_{xy} = 0{,}98$.
6.18. 3,96%.
6.19. $r_{xy} = 0{,}986$.
6.20. 92,16%.
6.21. $r_{xy} = 0{,}69$.
6.22. $r_{xy} = 0{,}78$.
6.23. $r^2_{xy} = 0{,}7225$.

6.24. $r_{xy} = 0{,}91$.
6.25. $r_{xy} = 0{,}712$.
6.26. $r_{xy} = -0{,}18$.
6.27. $r_{xy} = -0{,}6$.
6.28. $r_S = 0{,}89$.
6.29. $r_S = 0{,}542$.
6.30. $r_S = -0{,}312$.
6.31. $r_S = 0{,}71$.
6.32. $r_S = -0{,}98$.
6.33. $r_S = 0{,}92$.
6.34. $r_S = 0{,}84$.
6.35. $r_S = 0{,}92$.
6.36. $r_S = 0{,}09$.
6.37. $r_S = 0{,}91$.
6.38. $r_S = 0{,}91$.
6.39. a) $r_{xy}^2 = 0{,}22$; b) $e_{yx} = 0{,}584$; c) $m_{yx} = 0{,}121$.
6.40. $e_{yx} = 0{,}15$.
6.41. $e_{yx} = 0{,}88$.
6.42. $e_{xy} = 0{,}803$.
6.43. $e_{xy} = 0{,}417$.
6.44. $e_{yx} = 0{,}803$.
6.45. $e_{yx} = 0{,}93$.
6.46. $e_{yx} = 0{,}139$.
6.47. $m_{yx} = -0{,}11$; $m_{xy} = 0{,}02$; $e_{yx}^2 = 0{,}79$.
6.48. $e_{yx} = 0{,}185$.
6.49. $r_{xy \cdot z} = 0{,}61$
6.50. a) $r_{xy} = -0{,}953$; $r_{xz} = 0{,}891$; $r_{yz} = -0{,}867$; b) $r_{xy \cdot z} = -0{,}801$; $r_{xz \cdot y} = 0{,}436$; c) $R_{x \cdot yz} = 0{,}963$.
6.51. $r_{x_1x_2 \cdot x_3} = 0{,}59$; $r_{x_1x_3 \cdot x_2} = 0{,}037$; $r_{x_2x_3 \cdot x_1} = 0{,}815$; $R_{x_1 \cdot x_2x_3} = 0{,}936$.
6.52. Tak, bo $R^2_{x_2 \cdot x_1x_3} = 0{,}34$.
6.53. a) $r_{xy \cdot z} = 0{,}253$; $r_{xz \cdot y} = 0{,}918$; $r_{yz \cdot x} = -0{,}095$; b) $R_{x \cdot yz} = 0{,}743$.
6.54. $r_{xy \cdot z} = 0{,}74$; $r_{xz \cdot y} = -0{,}95$; $r_{yz \cdot x} = 0{,}95$; $R_{y \cdot xz} = 0{,}98$.
6.55. $r_{yz \cdot x} = -0{,}36$; $R^2_{z \cdot xy} = 0{,}4$.
6.56. $r_{xy \cdot z} = 0{,}73$.
6.57. H_0 należy odrzucić. Oznacza to, że związek jest statystycznie istotny. $T = 0{,}16$.
6.58. $T = 0{,}5$.
6.59. $C = 0{,}49$; $C_{max} = 0{,}6$.
6.60. $\chi^2 = 99{,}57 > \chi^2_\alpha = 9{,}488$; H_0 odrzucić.
6.61. $\chi^2 = 28{,}72 > \chi^2_\alpha = 6{,}63$; H_0 odrzucić.
6.62. $\chi^2 = 14{,}448 > \chi^2_\alpha = 9{,}21$; H_0 odrzucić.
6.63. $\chi^2 = 25 > \chi^2_\alpha = 3{,}841$; H_0 odrzucić.
6.64. $T = 0{,}145$.
6.65. $\hat{x}_i = 1{,}36 + 0{,}97y_i$; $\hat{y}_i = 1{,}43 + 0{,}81x_i$.
6.66. $\hat{y}_i = 3{,}39 + 0{,}249x_i$; $\hat{x}_i = 8{,}55 + 2{,}11y_i$; $r_{xy} = 0{,}724$; $\hat{y}_{x=50} = 15{,}84$ USD.
6.67. $r_{xy} = 0{,}87$; $\hat{y}_i = 5 + 2x_i$; $\hat{y}_{x=10} = 25$ dni.
6.68. $\hat{y}_i = 952{,}5 - 147{,}5x_i$; $\varphi^2 = 0{,}08$.
6.69. $\hat{y}_i = 1{,}7 + 0{,}4x_i$; $\hat{x}_i = 0{,}741 + 0{,}89y_i$; 3,3 izby.
6.70. $\varphi^2 = 0{,}0196$; $V_u = 5{,}3\%$.

6.71. $\hat{y}_i = -32{,}8 + 0{,}01x_i$; $\hat{x}_i = 2320 + 40y_i$.
6.72. $r_{xy} = 0{,}6$.
6.73. $r_{xy} = 0{,}8$; $\hat{y}_i = 9 + 2x_i$; $\hat{x}_i = 0{,}32y_i$.
6.74. $r_{xy} = 0{,}8$; $\hat{y}_i = 40x_i + 525$; $\hat{x}_i = 0{,}016y_i - 7{,}5$.
6.75. $\hat{y}_i = 0{,}16x_i + 4{,}29$; $r_{xy} = 0{,}286$.
6.76. 155 tys. zł, $\varphi^2 = 64\%$.
6.77. Nie.
6.78. 189,88 zł oraz 1262,20 zł.
6.79. $\hat{x}_i = -0{,}73 + 0{,}076y_i$; $\hat{y}_i = 24{,}64 + 5{,}24x_i$; $r_{xy} = 0{,}63$.
6.80. $r_{xy}^2 = 36\%$.
6.81. $r_{xy}^2 = 51\%$.
6.82. $\hat{x}_i = 2{,}7y_i + 1{,}4$.
6.83. Wyniki są błędne, gdyż $a_1 = -2{,}5 < 0$, natomiast $r_{xy} = 0{,}64 > 0$.
6.84. (0,828; 3,372).
6.85. H_0 odrzucić.
6.86. $0{,}534 < \rho < 0{,}706$.
6.87. $t = 1{,}963 > t_\alpha = 1{,}734$; H_0 odrzucić.
6.88. $13{,}77 < \alpha_1 < 14{,}43$.
6.89. $0{,}4165 < \rho < 0{,}5735$.
6.90. $|t| = 3{,}56 > t_\alpha = 2{,}306$; H_0 odrzucić.
6.91. $F = 0{,}973 < F_\alpha = 2{,}50$; brak podstaw do odrzucenia H_0.
6.92. $F = 93{,}52 > F_\alpha = 3{,}05$; H_0 odrzucić.
6.93. $k = 7$. Brak podstaw do odrzucenia H_0. $-0{,}723 < \beta_1 < -0{,}277$.
6.94. $F = 11{,}68 > F_\alpha = 2{,}29$. H_0 odrzucić.
6.95. $0{,}827 < \alpha_1 < 3{,}373$.
6.96. $r = 0{,}8775$; $t = 4{,}841 > t_\alpha = 1{,}895$; H_0 odrzucić.
6.97. H_0 odrzucamy.
6.98. H_0 odrzucamy.
6.99. $k = 5$. Brak podstaw do odrzucenia H_0.
6.100. $\chi^2 = 23{,}01 > \chi^2_\alpha = 13{,}815$. H_0 odrzucić.

Rozdział VII

7.1. $\bar{y} = 455{,}6$ tys. zł oraz $\bar{y}_{ch} = 27{,}3$ tys. zł.
7.2. Dla $k = 3$: 10,33; 11,00; ...; 19,67; 20,67. Dla $k = 4$: 10,375; 11,000; ...; 17,500; 18,000.
7.3. 20182,7; 20796,2; ...; 24336,7; 24046,9.
7.4. $\hat{y}_t = 1238 + 32t$; $s_e = 31{,}9$ tys. zł; $\varphi^2 = 0{,}185$; $R^2 = 0{,}815$; $y^*_{T=n+2} = 1494$ tys. zł.
7.5. $\hat{y}_t = 13{,}5 + t$.
7.6. $\hat{y}_t = 0{,}66t + 8{,}6$ ($\sum t \neq 0$) oraz $\hat{y}_t = 0{,}65t + 14{,}125$ ($\sum t = 0$).
7.7. Obydwa szacunki są prawidłowe.
7.8. $\hat{y}_t = -5{,}4 + 3{,}8t$.
7.9. $\varphi^2 = 0{,}078$; $R^2 = 0{,}922$; $s(a_0) = 5{,}81$; $s(a_1) = 1{,}49$.
7.10. $S_I = 104{,}7\%$; $S_{II} = 95{,}3\%$; $g_I = -2{,}06$ dni; $g_{II} = 2{,}06$ dni.
7.11. $\hat{y}_t = 6{,}928t + 57{,}067$ ($\sum t = 0$).
7.12. $\hat{y}_t = 206{,}77 - 14{,}92t$; 72,49.
7.13. $\hat{y}_t = 7{,}42 + 0{,}365t$ ($\sum t \neq 0$); $s_e = 0{,}216$ tys. osób; $V_e = 2{,}4\%$.
7.14. $S_I = 84{,}2\%$; $S_{II} = 81{,}4\%$; $S_{III} = 81{,}4\%$; $S_{IV} = 153\%$.

7.15. $\hat{y}_t = \underset{(3,527)}{51,91} + \underset{(0,419)}{8,35t} + \underset{(6,327)}{z}$; $\varphi^2 = 0,029$.

7.16. 42,6.

7.17. a) $\hat{y}_t = 59,48 + 0,99t$ $\sum t = 0$; b) $\hat{y}_t = 56,51 + 0,99t$ $(\sum t \neq 0)$; $s(z_t) = 0,34$.

7.18. 13,32 ± 0,2 tys. zł.

7.19. I kwartał: 68,5; II kwartał: 61,5; III kwartał: 41,5; IV kwartał: 78,5.

7.20. 103,5 tys. ton.

7.21. 2,7 ± 0,02.

7.22. 9,72 ± 0,5.

7.23. 8,6 kg.

7.24. $S_I = 104,2\%$; $S_{II} = 94,7\%$; $S_{III} = 83,9\%$; $S_{IV} = 117,2\%$; $g_I = 11,35$ tys. kg; $g_{II} = -14,32$ tys. kg; $g_{III} = -43,78$ tys. kg; $g_{IV} = 46,19$ tys. kg.

7.25. 58,7 ± 2,5.

7.26. 140 tys. zł, 188 tys. zł, 204 tys. zł, 172 tys. zł (przy wykorzystaniu absolutnych poziomów wahań sezonowych).

7.27. 91,81.

7.28. Dla modelu addytywnego: 3,188. Dla modelu multiplikatywnego: 11,416.

7.29. 48,3 ± 7,07.

7.30. 0,919 oraz –46,2.

7.31. $n = 24$; $k = 12$; dla $\alpha = 0,05$ oraz $n_1 = 11$ i $n_2 = 13$ mamy $k_\alpha = 8$. Brak podstaw do odrzucenia H_0.

7.32. $d = 1,0419$; $d_1 = 0,811$; $d_2 = 1,070$. Nie można podjąć żadnej decyzji.

7.33. $24,655 < \alpha_0 < 26,545$ oraz $-0,652 < \alpha_1 < -0,348$.

7.34. $d = 1,4$. Nie można podjąć żadnej decyzji.

7.35. H_0 odrzucić.

7.36. H_0 odrzucić.

7.37. Brak podstaw do odrzucenia H_0.

7.38. $0,108 < \alpha_1 < 0,120$ oraz $9,697 < \alpha_0 < 9,845$.

7.39. $t = 18,18$. H_0 odrzucić.

7.40. $d = 1,9074$; $d_1 = 1,08$; $d_2 = 1,36$. Brak podstaw do odrzucenia H_0.

7.41. $k = 10$; $n_z = 5$; $n_b = 5$; $k_1 = 2$; $k_2 = 9$. H_0 odrzucić.

7.42. Prognoza: 6,5 ± 0,02 tys. kg.

7.43. $-5,57 < \alpha_1 < -2,19$ oraz $234,51 < \alpha_0 < 256,29$.

7.44. $\hat{y}_t = 0,46t + 63,46$ dla t = 0, 1, ... oraz $\hat{y}_t = 0,46t + 64,38$ $(\sum t = 0)$;
$V(z_t) = \frac{0,444}{64,38} \cdot 100 = 0,69\%$.

7.45. $\hat{y}_t = 1,19t + 4,61$; $\varphi^2 = 0,23\%$; $R^2 = 99,77\%$.

7.46. $d = 1,986$. Brak podstaw do odrzucenia H_0.

7.47. Brak podstaw do odrzucenia H_0.

7.48. $-0,26 < \alpha_1 < -0,20$ oraz $9,69 < \alpha_0 < 10,49$.

7.49. $k = 3$. Brak podstaw do odrzucenia H_0.

7.50. Tak.

Rozdział VIII

8.1. Średnie roczne tempo wzrostu: 40,9%.

8.2. 100,0; 103,1; 107,6; 111,8; 115,7; 118,4; 119,6; 121,6.

8.3. Średnie tempo zmian: 11,2%.

8.4. y_t: 360 i 455. Indeksy jednopodstawowe: 1,32 i 1,52. Indeksy łańcuchowe: 1,2 i 1,1.

8.5. Zakład A: 22%. Zakład B: 27%.

8.6. 0,980; 1,041; 0,971; 1,061 oraz 0,980; 0,961; 1,000; 0,971; 1,029.

8.7. Średnie tempo spadku: 22,6%.

8.8. 100,0; 102,2; 106,7; 111,1; 116,7; 124,4; 128,9; 134,4.

8.9. –13,5; –4,8; 0; 20; 32; 45,2.

8.10. 2,2; 4,3; 4,2; 4,0; 5,8; 4,5.

8.11. Średnie roczne tempo spadku: 27,39%. Produkcja spadła o 78,73%.

8.12. 2033 tys. m^3.

8.13. 100,4; 112,4; 100,0; 100,0; 115,0; 98,9; 121,6.

8.14. 9,336 tys. osób.

8.15. Średnie roczne tempo: 3,3%; 228 samochodów.

8.16. 10 816 sztuk.

8.17. Średnioroczne tempo wzrostu: 10,53%.

8.18. Średnioroczne tempo wzrostu: 13,6%.

8.19. Średnioroczne tempo wzrostu: 2,7%.

8.20. Średnioroczne tempo wzrostu: 2,9%.

8.21. 1,0605%.

8.22. $\overline{y}_g = \sqrt[6]{0,9553267}$.

8.23. Średnioroczne tempo spadku: 0,6%.

8.24. 75,28%.

8.25. $I_w = 119\%$; $I_q^L = 103,6\%$; $I_p^P = 115\%$.

8.26. $I_w = 120\%$; $I_q^P = 108,7\%$; $I_p^P = 110,4\%$.

8.27. Spadła o 10,25%.

8.28. $I_w = 140\%$; $I_p^P = 96,2\%$; $I_q^L = 129\%$.

8.29. $I_w = 112,2\%$; $I_p^P = 96,2\%$; $I_q^L = 116,6\%$.

8.30. $I_w = 105\%$; $I_p^L = 97,5\%$; $I_q^P = 107,7\%$.

8.31. $I_p^L = 220\%$.

8.32. $I_q^L = 98,95\%$.

8.33. $I_q^P = 106\%$; $I_p^P = 96,4\%$.

8.34. $I_w = 117,1\%$; $I_p^P = 87,2\%$; $I_q^L = 134,3\%$.

8.35. $I_q^L = 97,53\%$; $I_q^P = 116,36\%$.

8.36. $I_w = 154,9\%$; $I_q^L = 216,43\%$; $I_q^P = 198,1\%$; $I_p^P = 71,57\%$; $I_p^L = 78,19\%$.

8.37. $I_w = 140\%$; $I_p^P = 103,8\%$; $I_q^L = 134,9\%$.

8.38. $I_w = 88,93\%$; $I_q^L = 81,55\%$; $I_q^P = 88,08\%$; $I_p^P = 109,05\%$; $I_p^L = 100,97\%$.

8.39. $I_p^P = 133,33\%$.

8.40. $I_w = 131,58\%$; $I_p^P = 104,16\%$; $I_q^L = 126,32\%$.

8.41. $I_w = 105,5\%$; $I_q^L = 101,1\%$; $I_q^P = 101,3\%$; $I_p^P = 104,4\%$; $I_p^L = 104,1\%$.

8.42. $I_x = 125\%$; $I_{x/b_0} = 94\%$; ${}_bI_{x/x_1} = 132,98\%$.

8.43. Spowodowały wzrost o 12,5%.

8.44. ${}_bI_{x/x_1} = 130\%$.

8.45. ${}_bI_{x/x_1} = 126,32\%$.

8.46. $I_x = 92,4\%$; $I_{x/b_1} = 94,3\%$; ${}_bI_{x/x_0} = 98\%$.

8.47. $I_x = 105\%$; $I_{x/b_1} = 94,3\%$; ${}_bI_{x/x_0} = 90,8\%$.

8.48. Wzrost o 18,2%.

8.49. $I_x = 96,6\%$.

8.50. $I_x = 96,8\%$; 110%.