5 PRACE DYDAKTYCZNE

Seria: Nauki społeczne

1996 Wałbrzyska Wyższa Szkoła Zarządzania i Przedsiębiorczości WWSZiP

Zygmunt Bobowski

Wybrane metody statystyki opisowej i wnioskowania statystycznego

Wałbrzych 2004

WAŁBRZYSKA WYŻSZA SZKOŁA
ZARZĄDZANIA I PRZEDSIĘBIORCZOŚCI

PRACE DYDAKTYCZNE NR 5
SERIA: NAUKI SPOŁECZNE

ZYGMUNT BOBOWSKI

WYBRANE METODY STATYSTYKI OPISOWEJ I WNIOSKOWANIA STATYSTYCZNEGO

Wydanie 1

Wydawnictwo WWSZiP
2004

KIEROWNIK NAUKOWY SERII
prof. dr hab. Zbigniew Kurcz

REDAKTOR WYDAWNICTWA
Michał Lesman

PROJEKT OKŁADKI
Stanisław Gola

REDAKTOR TECHNICZNY
Władysław Ramotowski

Printed in Poland

ISBN 83-915717-9-3

Wydawnictwo WWSZiP, Wałbrzych 2004
Wydanie I
Druk: Drukarnia D&D Spółka z o.o. 44-100 Gliwice, ul. Moniuszki 6

SPIS TREŚCI

Przedmowa

Niniejszy podręcznik jest przeznaczony przede wszystkim dla studentów studiów dziennych i zaocznych kierunków humanistycznych, chociaż z uwagi na uniwersalny charakter przekazanej w niej podstawowej wiedzy teoretycznej może być z powodzeniem wykorzystywany również przez studentów innych kierunków.

Wychodząc z założenia, że studentom tego typu kierunków opanowanie przedmiotów ilościowych, do jakich należy statystyka, może sprawiać pewne trudności Autor – kierując się sentencją sformułowaną przez Einsteina: „wszystko należy robić tak prosto, jak to jest możliwe, ale nie prościej", podjął próbę przekazania wiedzy z tego zakresu w możliwie przystępnej formie unikając nadmiernego jej „zmatematyzowania". Każdy z rozdziałów zawiera więc jedynie minimum teorii niezbędnej dla wyjaśnienia omawianych zagadnień, a dla ułatwienia ich zrozumienia zamieszczono przykładowe zadania wraz z rozwiązaniami.

Treść podręcznika zawiera się w pięciu rozdziałach. W rozdziale pierwszym omówiono podstawowe pojęcia statystyczne oraz przedstawiono procedurę programowania badania statystycznego, w szczególności badania ankietowego. W rozdziale drugim dokonano prezentacji metod opisu statystycznego. Omówiono i zilustrowano przykładami zasady wykorzystania metody tabelarycznej i graficznej. Zasadniczą część tego rozdziału poświęcono opisowi parametrycznemu wskazując na rolę i możliwości wykorzystania poszczególnych grup parametrów i momentów statystycznych w tym opisie. Rozdział trzeci zawiera podstawowe metody analizy współzależności zmiennych i mierniki w tym zakresie wykorzystywane. W rozdziale czwartym omówiono metody analizy dynamiki zjawisk ze szczególnym uwzględnieniem metod indeksowych, analizy tendencji rozwojowej oraz analizy sezonowości. Rozdział piąty poświęcono metodom wnioskowania statystycznego. Dokonano w nim prezentacji procedur szacowania podstawowych parametrów statystycznych dla populacji generalnej oraz testowania hipotez statystycznych, zarówno parametrycznych jak i nieparametrycznych.

W końcowej części publikacji zamieszczono tablice wybranych rozkładów teoretycznych wykorzystywane w procedurach wnioskowania statystycznego oraz zestawienie podstawowych wzorów statystycznych.

Autor zdaje sobie sprawę z mankamentów tej publikacji i będzie wdzięczny za uwagi, które pozwolą na jej udoskonalenie.

Zygmunt Bobowski

Rozdział I

Przedmiot i podstawowe pojęcia statystyki

W literaturze wyróżnia się kilka znaczeń terminu *„statystyka”*, chociaż jego pierwotne znaczenie wywodzi się od wyrazu *„status”*, czyli *państwo*, co pozwalałoby określić ją mianem *„państwoznawstwa”*, tj. nauką zajmującą się opisem procesów zachodzących w państwie. Wśród spotykanych współcześnie znaczeń tego terminu należy wyróżnić:

a) *statystykę jako zbiór danych*; w tym znaczeniu może być ona utożsamiana z rocznikiem statystycznym (np. Rocznik Statystyczny Woj. Dolnośląskiego),
b) *statystykę jako proces gromadzenia informacji* (czynnościowe znaczenie tego terminu); potocznie o osobach zatrudnionych w firmach, instytucjach zajmujących się przygotowywaniem wszelkiego rodzaju sprawozdań, w tym obligatoryjnych przekazywanych urzędom statystycznym można powiedzieć, iż zajmują się „statystyką” tego podmiotu,
c) *statystykę jako parametr statystyczny*, czyli liczbę syntetycznie opisującą badaną zbiorowość (np. średnia arytmetyczna jako bardzo często używana podstawowa „statystyka” badanej zbiorowości),
d) *statystykę jako naukę zajmującą się badaniem procesów, zjawisk masowych i opisującą je najczęściej za pomocą liczb w celu ustalenia prawidłowości charakteryzujących ich kształtowanie się.* Takie znaczenie statystyce przypisał m. in. O. Lange, według którego;
„...statystyka jest dyscypliną traktującą o metodach ilościowych badania zjawisk masowych”.

Procedura badania zjawisk masowych, w tym gromadzenia danych i ustalania prawidłowości, określana jest mianem *badania statystycznego*. Obejmuje ono najczęściej następujące trzy etapy:
I) *programowanie badania statystycznego*, które ma charakter etapu przygotowującego właściwe badanie. Etap ten obejmuje następujące czynności określające:

- problem badawczy, czyli sformułowanie celu badania,
- zakres rzeczowy badania, czyli sprecyzowanie jednostki i zbiorowości statystycznej w proponowanym badaniu,
- zakres czasowy badania, czyli przyjęcie momentu czasowego bądź przedziału czasowego dla badania,
- zakres przestrzenny badania, czyli sprecyzowanie obszaru, z którego wywodzi się badana zbiorowość,
- cechy statystyczne, ze względu na które będzie dokonana charakterystyka zbiorowości bądź zjawiska,
- metodę badania zbiorowości bądź zjawiska, czyli wybór jednej z dwóch metod, tj. badania całkowitego lub częściowego,
- sposób gromadzenia informacji gwarantujący uzyskanie danych o najwyższym stopniu wiarygodności,
- treść formularza badawczego; czynność szczególnie istotna w przypadku badań ankietowych.

II) *gromadzenie danych*, tj. etap o zasadniczym znaczeniu, bowiem „jakość" zebranych informacji decyduje o „jakości" i prawidłowości formułowanych w dalszej części badania wniosków. Proces gromadzenia informacji winien być zakończony kontrolą kompletności zgromadzonego materiału jak również jego kontrolą merytoryczną,

III) *opis statystyczny*, w ramach którego odbywa się opracowywanie i analiza zgromadzonego materiału statystycznego przy wykorzystaniu różnorodnych metod opisu statystycznego w celu ustalenia określonych prawidłowości. Ze względu na jego postać wyróżnia się opis: tabelaryczny, graficzny i parametryczny, zaś z uwagi na jego cel: opis struktury zbiorowości, opis współzależności wybranych cech, opis dynamiki badanego zjawiska.

Wymienione wyżej trzy etapy badania są charakterystyczne dla każdego badania statystycznego zarówno całkowitego jak i częściowego i tworzą one tzw. statystykę opisową. W przypadku badań częściowych wyróżnić należy dodatkowy etap czwarty określany mianem *wnioskowania statystycznego*. Jego zadaniem jest wnioskowanie, przy wykorzystaniu odpowiednich metod statystycznych, o prawidłowościach występujących w zbiorowości całkowitej na podstawie wyników badań stwierdzonych w badanej próbie statystycznej. Zbiór stosowanych w tym przypadku metod tworzy tzw. statystykę matematyczną.

Pierwszy z wymienionych etapów badania statystycznego ma charakter przygotowawczy. Prawidłowe przygotowanie badania wymaga znajomości podstawowych pojęć statystycznych. Należą do nich: jednostka, zbiorowość i cecha statystyczna.

Jednostką statystyczną jest każda pojedyncza jednostka (np. osoba, przedmiot, zjawisko) podlegająca badaniu, zaś *zbiorowością* zbiór tych jednostek wyodrębnionych według określonego kryterium ich podobieństwa[1] (np. pracownicy firmy „K", studenci uczelni „W", mieszkańcy miasta „P"). Ze

[1] To kryterium podobieństwa jest często określane mianem cech stałych, wśród których wyróżnia się cechy: rzeczowe, czasowe i przestrzenne. Ze względu na te cechy jednostki badanej zbiorowości są identyczne.

względu na przyjęte kryterium podobieństwa poddana badaniu zbiorowość jest jednorodna.

Zgodnie z podaną wyżej definicją wymaga się by zbiorowość taka miała charakter „masowy", gdyż tylko wówczas ujawniają się prawidłowości, których nie można zaobserwować w pojedynczym przypadku. Liczebność badanej zbiorowości przyjmuje się zwykle oznaczać jako „*N*". Warto tu przytoczyć kilka najczęściej spotykanych klasyfikacji zbiorowości statystycznych. I tak:

I) ze względu na ich liczebność wyodrębnia się zbiorowości:

a) *skończenie liczne*, składające się ze skończonej, przeliczalnej liczby jednostek statystycznych (np. zbiorowość studentów II roku),
b) *nieskończenie liczne*, tworzone przez nieskończoną liczbę jednostek statystycznych (np. zbiór istot żywych na kuli ziemskiej); w przypadku tego typu zbiorowości stosowana jest metoda badań częściowych ograniczających się jedynie do badania pewnego jej „wycinka",

II) ze względu na przyjęty zakres czasowy badania dokonuje się podziału zbiorowości na:

a) *statyczne*, gdy wszystkie jednostki statystyczne są badane według stanu na ten sam moment czasowy (np. studenci według miejsca zamieszkania na dzień 30.05.2003 r.),
b) *dynamiczne*, gdy zbiorowość jest charakteryzowana ze względu na określoną cechę w pewnym przedziale czasowym (np. struktura mieszkańców Wałbrzycha w latach 1990 – 2003 według wieku),

III) ze względu liczbę cech niezbędnych dla ich opisu wyróżnia się zbiorowości[2]:

a) *proste* (jednorodne), które są opisywane niewielką liczbą cech (1 – 2), np. zbiór ławek w sali dydaktycznej,
b) *złożone* (niejednorodne), dla opisu których wykorzystuje się liczny zbiór cech, np. studenci określonej uczelni.

Wyodrębniona zbiorowość jest poddawana badaniu ze względu na określone właściwości określane mianem cech statystycznych. Według B. Szulca *cecha statystyczna* „...jest to właściwość obiektu odróżniająca go od innych obiektów". Podobną definicję formułują M. Krzysztofiak i A. Luszniewicz, według których „...*cechami statystycznymi* nazywamy właściwości, których odmiany lub wartości wyróżniają jednostki wchodzące w skład zbiorowości statystycznej".

Cechy statystyczne przyjmuje się oznaczać dużymi literami: *X, Y, Z*, zaś ich realizacje: x_i, y_i, z_i i należy je interpretować jako wartości cechy *X, Y, Z* zaobserwowane w i-tej jednostce.

Z punktu widzenia dalszych rozważań istotne znaczenie mają spotykane w literaturze przedmiotu klasyfikacje cech według różnych kryteriów.

I tak z punktu widzenia sposobu zapisu wartości cechy wyróżnia się:

a) *cechy opisowe* (określane również jako werbalne, niemierzalne, jakościowe), których realizacje są wyrażane na skali nominalnej lub porząd-

[2] W literaturze podkreśla się często, iż podział ten jest mało jednoznaczny, zależy bowiem w dużej mierze od poziomu szczegółowości i celu badań

kowej (np. płeć, narodowość, poziom wykształcenia); cecha opisowa odwzorowuje więc zbiór obiektów w zbiór określeń słownych,

b) *cechy liczbowe* (zwane również ilościowymi, mierzalnymi), których wartości są mierzone na skalach interwałowej bądź ilorazowej (np. waga, wiek, temperatura); ta cecha z kolei odwzorowuje zbiór obiektów w zbiór liczb.

Uszczegółowienie tego podziału stanowi spotykany coraz częściej w literaturze podział cech według rodzaju skali pomiarowej. Kryterium to pozwala wyodrębnić cechy, których wartości są wyrażane na skali:

a) nominalnej,
b) porządkowej,
c) interwałowej,
d) ilorazowej.

W przypadku *skali nominalnej* wartości cech są zwykle wyrażane słownie, niekiedy wartościom cechy przyporządkowuje się arbitralnie liczby, które pełnią jedynie rolę nazwy kategorii (na liczbach tych nie można przeprowadzać operacji matematycznych). Wykorzystywane w tym przypadku symbole pozwalają na stwierdzenie tożsamości lub rozróżnienie obiektów ze względu na badaną cechę. W naukach społecznych z tego typu cechami mamy dość często do czynienia, np. płeć, status społeczny, wykonywany zawód, wyznawana religia. Ten rodzaj skali umożliwia klasyfikację (grupowanie) badanego zbioru obiektów na grupy (kategorie) obiektów do siebie podobnych, np. podział społeczeństwa na wyznawców różnych religii, podział ludności na pracujących i niepracujących. Należy podkreślić, iż może budzić wątpliwości zaliczenie skali nominalnej do skal pomiarowych. Oznacza to bowiem, że cechy wyrażane na tej skali są cechami mierzalnymi.

Skala porządkowa (rangowa) umożliwia bardziej precyzyjny, niż w przypadku skali nominalnej, pomiar wartości cechy. Wartości te mogą być również wyrażane słownie lub liczbowo, jednak istnieje możliwość ich uporządkowania według „natężenia" cechy. Przykładami cech, których wartości są wyrażane na tej skali są: poziom wykształcenia, wzrost (wyrażony jako: niski, średni, wysoki), stopnie w służbach mundurowych. Wartości cechy wyrażonej na tej skali umożliwiają uporządkowanie badanego zbioru obiektów od najniższego do najwyższego poziomu badanej cechy. Należy jednak pamiętać, że nie istnieje możliwość określania wielkości różnic między tymi wartościami. Nie ma więc możliwości wykonywania operacji matematycznych na wartościach takiej cechy. Cecha ta pozwala na stwierdzanie tożsamości bądź różności obiektów, a ponadto pozwala na ich uporządkowanie jednak bez określania odległości.

Wartości cechy wyrażanej na *skali interwałowej (przedziałowej)* są liczbami ze zbioru liczb rzeczywistych (zarówno dodatnich jak i ujemnych). Zbiór wartości tej cechy nie posiada „naturalnej" wartości zerowej; wartość zerową przyjmuje się zwykle umownie. Do cech mierzonych na tej skali można zaliczyć: temperaturę powietrza (wody), wynik finansowy firmy. Dla wartości cech wyrażonych na tej skali możliwe jest wykonywanie operacji odejmowania (badanie odległości). Na wartościach tej cechy nie można wykonywać operacji mnożenia i dzielenia.

Skala ilorazowa może być traktowana jako szczególny przypadek skali interwałowej, bowiem w przypadku skali ilorazowej wartości cechy należą do zbioru liczb rzeczywistych nieujemnych. Skala ta posiada naturalny punkt zerowy. Do cech wyrażanych na tej skali można zaliczyć: wzrost, wagę bądź wiek człowieka, dochód przypadający na 1 osobę w rodzinie. Dla wartości cech wyrażonych na skali ilorazowej dopuszczalne jest stosowanie wszystkich operacji matematycznych, tj.: stwierdzanie tożsamości obiektów, ich porządkowanie, ustalanie odległości, określanie równości ilorazów.

Zbliżona do powyższego podziału jest klasyfikacja cech według możliwości ich pomiaru fizycznego. Według tego kryterium wyróżnia się:

a) *cechy mierzalne*, czyli takie których wartości są wynikiem pomiaru fizycznego i wyrażone są w określonych jednostkach miary, np. wzrost w cm, waga w kg,
b) *cechy pośrednio mierzalne*, tzn. cechy których wartości wyrażone są liczbowo, ale liczby te są jedynie wynikiem oceny (przyporządkowane są odpowiednim wyrażeniom słownym), np. ocena z egzaminu, nota sędziowska w skokach narciarskich,
c) *cechy niemierzalne*, czyli cechy których wartości są wyrażane słownie, np. płeć, pochodzenie społeczne, marka samochodu.

Ze względu na liczebność zbioru wartości cechy wyróżnia się:

a) *cechy stałe*, dla których zbiór wartości jest jednoelementowy; innymi słowy każdemu obiektowi badanej zbiorowości przypisywana jest ta sama wartość cechy. Niekiedy wśród nich wyróżnia się dodatkowo cechy stałe rzeczowe, czasowe i przestrzenne umożliwiające precyzyjne zdefiniowanie jednostki i zbiorowości statystycznej dla określonego badania. Cechy te odpowiadają zakresowi rzeczowemu, czasowemu i przestrzennemu badania.
b) *cechy zmienne*, dla których zbiór wartości jest co najmniej dwuelementowy. Cechy te stanowią przedmiot badań statystycznych.

Wśród cech zmiennych dokonuje się zwykle dalszego ich podziału z punktu widzenia podanego wyżej kryterium i wyróżnia się dodatkowo:

1) dla cech opisowych:
   a) *cechy dwuwariantowe* (zero-jedynkowe) np. płeć,
   b) *cechy wielowariantowe*, np. poziom wykształcenia, zawód,
2) dla cech liczbowych:
   a) *cechy skokowe* (dyskretne); posiadają one przeliczalny zbiór wartości zawierający się w zbiorze liczb naturalnych (liczby całkowite nieujemne), np. liczba osób w rodzinie,
   b) *cechy ciągłe*, których zbiór wartości jest nieprzeliczalny i należy do zbioru liczb rzeczywistych; wartości takiej cechy są nieskończenie podzielne i mogą być wyrażane z dowolnie dużą dokładnością, np. wzrost, wiek, waga.

Z punktu widzenia znaczenia poszczególnych cech dla określonego badania wyróżnia się:

a) *cechy istotne* ze względu na sformułowany cel badawczy,
b) *cechy nieistotne*, które w badaniu mogą być pominięte.

Biorąc pod uwagę wpływ cechy na poziom rozwoju obiektów, gdy celem badania jest hierarchiczne porządkowanie obiektów, wyróżnia się:

a) *cechy preferencyjne*,
b) *cechy neutralne*.

Wśród *cech preferencyjnych* wyodrębnia się dodatkowo cechy:

a) stymulanty tzn. takie cechy których wzrost wartości powoduje wzrost poziomu rozwoju obiektu (cechy te pobudzają, stymulują rozwój obiektu); najkorzystniejszą wartością takiej cechy jest jej wartość maksymalna, zaś najmniej korzystną – wartość minimalna,

b) destymulanty, czyli cechy, których wzrost wartości jest niekorzystny dla oceny poziomu rozwoju obiektu; najkorzystniejszą jej wartością jest wartość minimalna, zaś najmniej korzystną – wartość maksymalna,

c) nominanty, czyli cechy dla których wzrost wartości do wielkości nominalnej jest korzystny z punktu widzenia badanego obiektu, natomiast ich dalszy wzrost staje się niekorzystny. Wartością optymalną w tym przypadku jest wartość nominalna.

Cechy neutralne, z uwagi na ich charakter, mogą być pominięte w zestawie cech stanowiących podstawę hierarchicznego porządkowania obiektów.

Kolejną istotną czynnością w fazie wstępnej badania statystycznego jest wybór jednej z dwóch metod badania statystycznego, tj. *badania całkowitego* (wyczerpującego) lub *badania częściowego. Badanie całkowite* obejmuje całą zbiorowość statystyczną, a więc daje ono bardzo precyzyjny i wiarygodny jej opis. Ta metoda badania winna być szczególnie preferowana.

W praktyce często mamy jednak do czynienia z sytuacjami, w których – z różnych względów – nie ma możliwości zbadania wszystkich elementów tworzących całkowitą zbiorowość statystyczną (inaczej zwaną populacją generalną) i jedyną możliwością jej poznania jest zbadanie tylko jej części (tzw. próby statystycznej). W związku z powyższą sytuacją zachodzi konieczność wnioskowania o całej zbiorowości na podstawie wyników uzyskanych dla próby. Można wymienić kilka powodów, dla których prowadzone są *badania częściowe*:

- badanie może mieć charakter niszczący, np. badania jakościowe niektórych wyrobów; badanie całej populacji byłoby równoznaczne z jej zniszczeniem,
- badana populacja może być bardzo liczna lub nieskończenie liczna,
- badanie całkowite może być zbyt kosztowne lub zbyt czasochłonne,
- badanie częściowe gwarantuje wyższą aktualność wyników,
- badanie częściowe - w niektórych przypadkach - jest jedynym możliwym, np. badanie krwi pacjentów,
- badanie częściowe gwarantuje wyższą rzetelność, ponieważ może być zrealizowane mniejszym zespołem przeszkolonych fachowców; istnieje możliwość kontroli ich pracy.

Niezależnie od wybranej metody badania statystycznego istotnym problemem jest dobór techniki gromadzenia informacji. Musi ona gwarantować wysoki stopień wiarygodności gromadzonych informacji. Podstawowym źródłem informacji jest wszelkiego rodzaju sprawozdawczość statystyczna prezentowana na różnym poziomie agregacji danych (np. kraju, województwa,

powiatu, gminy, przedsiębiorstw, instytucji). Udostępniane danych w takim przypadku odbywa się głównie w formie roczników statystycznych bądź różnego rodzaju sprawozdań.

W praktyce badań socjologicznych szczególne znaczenie należy przypisać *badaniom ankietowym* jako formie gromadzenia informacji. *Ankieta* stanowi zbiór odpowiednio sprecyzowanych pytań służących realizacji określonego celu badania. Zasadnicze problemy związane z tego typu badaniami dotyczą w szczególności:

a) liczby i treści pytań zawartych w ankiecie,
b) sposobu przeprowadzenia badania ankietowego,
c) wyboru grupy respondentów tj. osób ankietowanych.

Ankieta składa się zwykle z dwóch zasadniczych części, tj. *metryczki i części merytorycznej. Metryczka* stanowi istotny element ankiety, ponieważ pozwala klasyfikować zbiorowość respondentów według różnych kryteriów i prowadzić analizę uzyskanych wyników badania ankietowego w ramach wyodrębnionych podzbiorowości respondentów. Najczęściej w metryczce uwzględnia się takie cechy respondentów jak: płeć, stan cywilny, wiek, poziom wykształcenia, miejsce zamieszkania, wykonywany zawód. Należy jednak wyraźnie zaznaczyć, że sposób gromadzenia tych danych winien zapewniać anonimowość respondenta.

Druga część ankiety zawiera zestaw pytań dotyczących określonego zagadnienia. Właściwe skonstruowanie tej części ankiety stanowi podstawowy warunek powodzenia przeprowadzanego badania. Nie może być ona zbyt obszerna, ponieważ może zniechęcać respondentów do poddania się badaniu ankietowemu (zwłaszcza w przypadku, gdy badanie ma być prowadzone w formie „sondażu ulicznego"). Projektując *merytoryczną część ankiety* należy przestrzegać następujących zasad:

1) pytania winny być jasno, jednoznacznie i prosto sformułowane,
2) pytania winny być umieszczone w logicznej kolejności,
3) winny umożliwiać udzielenie precyzyjnej odpowiedzi, w miarę możliwości liczbowej,
4) pytania natury osobistej winny być bardzo taktowne.

Umieszczone w ankiecie pytania mogą mieć charakter zamknięty lub otwarty. W pierwszym przypadku, do pytania dołączony jest rozłączny i wyczerpywalny zestaw odpowiedzi. Odpowiedzi te mogą być ujęte w formie:

a) alternatywy (dwóch wykluczających się możliwości),
b) zamkniętego zestawu wielu możliwych odpowiedzi,
c) zestawu półotwartego, gdy oprócz podanego zestawu odpowiedzi respondent ma możliwość dopisania innych możliwości,
d) odpowiedzi – skali, gdy respondent dokonuje oceny natężenia zachowań, poglądów, opinii (skala taka może mieć charakter liczbowy bądź porządkowy).

W drugim przypadku respondent ma pełną swobodę formułowania odpowiedzi.

Wybór zamkniętej formy pytań znacznie ułatwia respondentom udzielanie na nie odpowiedzi, jak również istotnie ułatwia opracowanie zgromadzonego materiału badawczego, tj. kodowanie odpowiedzi, analizę wyników badania. Otwarta forma pytań daje respondentom większą swobodę w for-

mułowaniu odpowiedzi, jednak znacznie utrudnia ona opracowanie i analizę zebranych informacji. Tej formy pytań nie zaleca się umieszczać w ankietach, które nie będą realizowane w postaci kontaktu bezpośredniego ankietera z respondentem.

Opracowany formularz ankiety winien być przetestowany w badaniu pilotażowym na niewielkiej próbie 30 – 50 osób. Badanie takie pozwala zweryfikować poprawność postawionych pytań jak również zamieszczonych wariantów odpowiedzi.

Właściwe badanie ankietowe może być realizowane w różnych formach kontaktów z respondentami. K. Mazurek-Łopacińska wyróżnia następujące metody przeprowadzenia badania ankietowego:

a) wywiad bezpośredni,
b) ankieta telefoniczna,
c) ankieta pocztowa,
d) ankieta prasowa,
e) ankiety rozdawane.

Zaletą pierwszej z metod jest gwarancja prawidłowego zrozumienia przez respondenta treści poszczególnych pytań (zwłaszcza, gdy ich treść może budzić wątpliwości) i właściwego sformułowania odpowiedzi. W takim przypadku z powodzeniem można wykorzystywać ankiety, w których zastosowano otwartą formę pytań. Ten sposób gromadzenia ankiet pozwala „na bieżąco" kontrolować ilość i jakość zgromadzonego materiału badawczego. W przypadku pozostałych metod, a zwłaszcza pocztowej, prasowej i ankiet rozdawanych należy liczyć się z możliwością wystąpienia niewielkiej zwrotności ankiet, jak również z przypadkami błędnie wypełnionych kwestionariuszy.

Badania ankietowe mają najczęściej charakter badań częściowych i w związku z tym pojawia się problem doboru respondentów. Wytypowana grupa respondentów winna być reprezentatywna dla całej populacji, o czym decydują dwa zasadnicze czynniki:

a) metoda doboru jednostek próby,
b) liczebność tej próby.

W literaturze wymienia się następujące sposoby pobierania próby:

a) *dobór arbitralny* – jest to metoda subiektywna; jednostki do próby są typowane przez ekspertów bądź ankieterów. W ramach tej metody można dodatkowo wyodrębnić:

- *dobór kwotowy*, którego celem jest uzyskanie próby o strukturze identycznej jak cała populacja potencjalnych respondentów. Po ustaleniu liczebności próby dokonuje się doboru określonych „kwot" respondentów gwarantujących uzyskanie odpowiedniej jej struktury,
- *losowanie „według wygody"* – stosowane głównie w badaniach marketingowych; przeprowadzane jest na grupach konsumentów określonych towarów,
- *losowanie metodą „przechwytywania respondentów po drodze"* – badanie odbywa się na grupie respondentów wybieranych losowo spośród przechodniów, klientów sklepu itp.,

b) *dobór losowy*, gdy mają zastosowanie określone schematy pobierania prób losowych, jak np. losowanie indywidualne (ograniczone i nieograni-

czone), losowanie warstwowe, losowanie systematyczne, losowanie wielostopniowe. Schematy te zostały dokładniej scharakteryzowane w rozdziale V.

O reprezentatywności próby decyduje również jej liczebność. Im większa jest liczba zgromadzonych ankiet tym bardziej wiarygodne dla całej populacji są wyniki badań. Pojawia się zatem problem ustalenia minimalnej liczebności próby. Analiza zgromadzonego drogą ankietową materiału opiera się zwykle na odpowiednio skonstruowanych tablicach uwzględniających strukturę respondentów ze względu na wyodrębnione kryteria (następuje tym samym podział tej zbiorowości na podgrupy). Im większa jest liczba wyodrębnianych podgrup, tym liczniejsza winna być grupa respondentów objętych badaniem. O liczebności populacji respondentów decyduje również zasięg badań (lokalny, regionalny bądź krajowy). Im jest on większy tym większa winna być losowana próba respondentów. W teorii badań ankietowych podaje się różne zalecenia dotyczące wielkości populacji objętych takimi badaniami. Jedną z propozycji uzależniającą wielkość populacji respondentów od zasięgu badań i liczby wyodrębnionych podgrup zawiera poniższe zestawienie.

Liczba wyodrębnionych podgrup	*Liczba jednostek w badaniu o zasięgu*	
	krajowym	*regionalnym*
do 9	1000 – 1500	200 – 500
10 – 30	1500 – 2500	500 – 1000
powyżej 30	powyżej 2500	powyżej 1000

Inna propozycja nakazuje, by populacja była na tyle liczna, aby liczebności najważniejszych dla analizy komórek tablicy wynosiły 100, a mniej ważnych od 20 do 50.

Właściwie zaprojektowane badanie statystyczne stanowi podstawowy warunek uzyskania wiarygodnych jego wyników. Wymienione wyżej metody opisu i wnioskowania statystycznego są przedmiotem rozważań w kolejnych rozdziałach niniejszego opracowania.

Rozdział II

Metody opisu zbiorowości statystycznej

2.1. OPIS TABELARYCZNY

Zgromadzony w ramach drugiego etapu badania statystycznego zbiór danych ma postać *szeregu szczegółowego, nieuporządkowanego*; tym samym jest mało przejrzysty, chaotyczny. Podejmowane w tej fazie czynności w pierwszej kolejności zmierzają do jego uporządkowania. W przypadku cech opisowych dokonuje się najczęściej jego pogrupowania według wariantów tej cechy, zaś w przypadku cech liczbowych polega ono zwykle na uporządkowaniu realizacji cechy od najniższej do najwyższej. Prowadzi to do uzyskania *szeregu uporządkowanego*. W przypadku licznych zbiorowości taka postać materiału statystycznego jest dość obszerna, co może znacznie utrudnić jego dalszą analizę. Problem ten można rozwiązać tworząc *szeregi rozdzielcze*, w których badana zbiorowość dzielona jest na określoną liczbę klas na podstawie wartości badanej cechy i każdej z klas zostaje przyporządkowana odpowiednia liczba jednostek statystycznych. *Szeregi rozdzielcze* mogą mieć postać szeregu *punktowego* bądź *przedziałowego*.

W *szeregu rozdzielczym punktowym* każda klasa zawiera tylko jedną wartość cechy (opisowej bądź liczbowej) i w związku z tym szereg taki posiada tyle klas, ile wariantów cechy występuje w zgromadzonym materiale statystycznym. W *szeregu z przedziałami klasowymi* dokonuje się grupowania wartości cechy w przedziały klasowe określając ich dolną i górną granicę. Szeregi z przedziałami klasowymi są najczęściej tworzone dla cech liczbowych o charakterze ciągłym, chociaż należy zaznaczyć, iż przedziały takie można również utworzyć dla cech liczbowych skokowych, a nawet opisowych.

Budowa szeregu z przedziałami klasowymi wymaga rozstrzygnięcia następujących kwestii:

a) ile klas winno być utworzonych?
b) jaka winna być ich wielkość (rozpiętość)?

c) jak winny być ustalone (zamknięte) granice przedziałów?

Przy podejmowaniu decyzji o liczbie klas należy mieć na uwadze, iż nie ma w teorii statystyki jednoznacznych wskazań w tym zakresie. Formułowane są jedynie pewne zalecenia, by liczba klas nie była ani zbyt duża, ani zbyt mała i była uzależniona głównie od liczebności zbiorowości. Zgodnie z tą sugestią wielu autorów proponuje wyznaczać orientacyjną (przybliżoną) liczbę klas według wzoru:

$$k \approx \sqrt{N} \qquad 2.1.$$

gdzie:

k – przybliżona liczba klas,
N – liczebność badanej zbiorowości.

Uzyskaną według powyższej formuły liczbę klas należy traktować jako orientacyjną, a rzeczywista liczba klas może być o jedną niższa bądź wyższa od wielkości ***k***. Takie postępowanie odnosi się do cech liczbowych ciągłych, w przypadku cech liczbowych skokowych bądź opisowych o liczbie klas decyduje głównie liczba wariantów cechy.

Przy określaniu wielkości przedziałów klasowych należy dążyć by klasy te posiadały jednakową rozpiętość, gdyż znacznie ułatwia to dalszą analizę ujętego w szeregu materiału statystycznego jak również wyznaczanie niektórych parametrów statystycznych. Należy również zwrócić uwagę, by w szeregu nie występowały klasy „puste", tzn. klasy, dla których nie można przyporządkować żadnej jednostki statystycznej. Uwzględniając powyższe zastrzeżenia przybliżoną wielkość przedziałów klasowych *(h)* można ustalić według formuły:

$$h \approx \frac{R(x)}{k} \qquad 2.2.$$

gdzie:

R(x) – rozpiętość wartości cechy obliczona według wzoru $R(x) = x_{\max} - x_{\min}$,

k – przyjęta liczba klas.

Taka reguła postępowania będzie miała miejsce w przypadku cech ciągłych. W przypadku cech opisowych bądź liczbowych skokowych równa rozpiętość klas będzie oznaczała jednakową liczbę wariantów cechy w każdej klasie.

Konieczność określenia zasad zamykania poszczególnych klas odnosi się do cech ciągłych. Przyjmuje się, iż każda z klas musi być jednostronnie zamknięta (z góry bądź z dołu), co powoduje, że są one rozłączne i umożliwiają jednoznaczne przyporządkowanie poszczególnych jednostek statystycznych konkretnym przedziałom.

Skonstruowany szereg rozdzielczy (zarówno punktowy jak i przedziałowy) może mieć postać szeregu liczebności (prostej), liczebności skumulowanej, częstości (prostej) bądź częstości skumulowanej. W szeregu liczebności

bądź częstości każdej klasie zostaje przyporządkowana określona liczba jednostek statystycznych lub ich udział (najczęściej wyrażony w procentach) w całej zbiorowości. W szeregach skumulowanych każdej klasie przyporządkowana jest odpowiadająca jej liczebność bądź częstość skumulowana.

Procedurę tworzenia szeregów z przedziałami klasowymi zilustrujemy na poniższych przykładach.

Przykład 2.1

Zbiorowość pracowników pewnej firmy scharakteryzowano ze względu na poziom wykształcenia i uzyskano następujące dane: *wyższe, podstawowe, średnie, średnie, zasadnicze zawodowe, podstawowe, niepełne podstawowe, średnie, wyższe, policealne, wyższe, średnie, zasadnicze zawodowe, podstawowe, podstawowe, średnie, średnie, wyższe, średnie, podstawowe, wyższe, wyższe, średnie, średnie, podstawowe, podstawowe, wyższe, niepełne podstawowe, podstawowe, policealne*

Dokonać grupowania badanej zbiorowości według poziomu wykształcenia. Uzyskane wyniki skomentować.

Rozwiązanie.

Zgromadzony materiał statystyczny ma postać szeregu szczegółowego nieuporządkowanego. Grupowania dokonamy ujmując podane informacje w postaci szeregu rozdzielczego punktowego. W szeregu tym każdemu poziomowi wykształcenia przyporządkujemy odpowiednią liczbę pracowników. Efekt tego grupowania ujęto w kolumnie 2. poniższej tablicy roboczej. W kolumnie 3. utworzono szereg liczebności skumulowanej. W kolumnach 4. i 5. umieszczono szeregi częstości i częstości skumulowanej. Częstość wyznaczono jako udział procentowy pracowników o określonym poziomie wykształcenia w całej populacji pracowników.

Poziom wykształcenia (x_i)	*Liczba pracowników* (n_i)	*Liczba pracowników skumulowana* ($cum\ n_i$)	*Częstość w %* (f_i)	*Częstość skumulowana w %* ($cum\ f_i$)
1	*2*	*3*	*4*	*5*
niepełne podstawowe	2	2	6,67	6,67
podstawowe	8	10	26,66	33,33
zasadnicze zawodowe	2	12	6,67	40,00
średnie	9	21	30,00	70,00
policealne	2	23	6,67	76,67
wyższe	7	30	23,33	100,00
Razem	30	X	100,00	X

Można również dokonać innego grupowania tworząc klasy (przedziały) jak w kolejnej tabeli:

Poziom wykształcenia	*Liczba pracowników*	*Liczba pracowników skumulowana*	*Częstość w %*	*Częstość skumulowana w %*
Niepełne podstawowe i podstawowe	10	10	33,33	33,33
Zasadnicze zawodowe i średnie	11	21	36,67	70,00
Policealne i wyższe	9	30	30,00	100,00
Razem	30	X	100,00	X

W badanej zbiorowości w podobnym stopniu dominują pracownicy o wykształceniu: podstawowym (26,66% badanej zbiorowości), średnim (30% badanej grupy) i wyższym (23,33% badanej populacji).

Przykład 2.2

Dla wylosowanej grupy 25 gospodarstw domowych zebrano informacje dotyczące wysokości dochodu (w zł) przypadającego na 1 osobę. Uzyskano następujące dane: *201,50; 556,10; 220,00; 820,40; 482,50; 339,60; 682,40; 398,40; 238,70; 180,50; 627,40; 421,60; 354,50; 289,70; 645,00; 654,30; 728,60; 453,90; 678,20; 721,00; 804,50; 487,60; 564,70; 732,40; 810,00.* Powyższe informacje ująć w postaci szeregu rozdzielczego z przedziałami klasowymi. Uzyskane wyniki skomentować.

Rozwiązanie.

Zauważmy, iż badana cecha ma charakter cechy liczbowej ciągłej i w takim przypadku nie tworzy się zazwyczaj szeregów rozdzielczych punktowych. Procedurę budowy szeregu przedziałowego przeprowadzimy zgodnie z podanymi wyżej wskazówkami. W pierwszej kolejności dokonujemy ustalenia liczby klas według *wzoru 2.1*. Otrzymujemy $k = 5$, co oznacza, iż liczba klas winna wynosić około pięciu (można rozważać wariantowo liczbę klas od 3 – 5). Przyjmijmy liczbę klas równą 5. W dalszej kolejności celem wyznaczenia rozpiętości klas (zakładamy, że wszystkie klasy będą posiadały identyczną rozpiętość) ustalamy rozpiętość wartości badanej cechy; wartość minimalna ($x_{\min}$) wynosi 180,50 zł, zaś wartość maksymalna ($x_{\max}$) – 820,40 zł, co daje nam rozpiętość wynoszącą 639,90 zł (*820,40 – 180,50*). Zgodnie z *wzorem 2.2*. uzyskujemy dla każdej z 5 klas rozpiętość 128 zł (dokładnie 127,98 zł, ale ze względów praktycznych dokonuje się zwykle jej zaokrąglenia „w górę"). Przyjmijmy również, iż każda z klas będzie zamknięta lewostronnie. W ten sposób otrzymamy przedziały klasowe ujęte w poniższej tabeli roboczej, a po zakwalifikowaniu wartości cechy do określonych przedziałów ujęte w kolejnych kolumnach rozkłady: liczebności i częstości oraz liczebności i częstości skumulowanej.

Dochód na 1 osobę w zł (x_i)	*Liczba gospodarstw* (n_i)	*Skumulowana liczba gospodarstw* ($cum\ n_i$)	*Częstość w %* (f_i)	*Częstość skumulowana w %* ($cum\ f_i$)
<180,50 – 308,50)	5	5	20,00	20,00
<308,50 – 436,50)	4	9	16,00	36,00
<436,50 – 564,50)	4	13	16,00	52,00
<564,50 – 692,50)	6	19	24,00	76,00
<692,50 – 820,50)	6	25	24,00	100,00
Razem	25	X	100,00	X

Na podstawie powyższej tabeli można sformułować następujące wnioski:

a) występuje brak wyraźnej dominacji gospodarstw o określonej wielkości dochodów,
b) blisko połowę badanych gospodarstw charakteryzują dochody na osobę przekraczające 564 zł, w przypadku pozostałych dochód na osobę jest niższy od tej wielkości,
c) najmniej liczne są gospodarstwa z przedziałów dochodowych 308,50 – 436,50 zł i 436,50 – 564,50 zł.

Zaprezentowane wyżej szeregi statystyczne można zaliczyć do tzw. *szeregów strukturalnych,* umożliwiają one bowiem określenie struktury badanej zbiorowości ze względu na przyjętą cechę statystyczną. Zazwyczaj prezentują one tę strukturę w określonym momencie czasowym, co oznacza statyczne ujęcie badanej zbiorowości. Odmianę szeregów strukturalnych stanowią *szeregi przestrzenne*, w których struktura zbiorowości określana jest w oparciu o jej terytorialne rozmieszczenie. Ten typ szeregu ilustruje tablica 2.1.

Tablica 2.1. Banki ogółem w krajach Unii Europejskiej w 1994 r.

Kraj	*Liczba banków*
Belgia	131
Dania	174
Francja	1468
Grecja	49
Hiszpania	217
Irlandia	48
Niemcy	1212
Wielka Brytania	486
Włochy	258

Źródło: W. Jaworski: Banki polskie – warunki przetrwania. Warszawa 1997

W literaturze wyodrębnia się dodatkowo szeregi: *dynamiczne (chronologiczne, czasowe*) obrazujące kształtowanie się badanego zjawiska w czasie[3]. Przykład takiego szeregu zawiera tablica 2.2.

Tablica 2.2. Przewozy pasażerów koleją w Polsce w latach 1990 – 1996

Lata	*Liczba pasażerów (w mln osób)*
1990	787,5
11991	650,2
1992	548,1
1993	540,1
1994	493,7
1995	465,1
1996	433,5

Źródło: Rocznik Statystyczny GUS 1997. Warszawa: GUS. 1997

Podane wyżej przykłady dotyczą prezentacji tabelarycznej badanej zbiorowości opisywanej *ze względu na jedną cechę statystyczną*. Badania statystyczne mogą obejmować również przypadki jednoczesnego opisu zbiorowości *ze względu na dwie bądź większą liczbę cech*. W takim przypadku opis tabelaryczny będzie polegał na ujęciu zebranego materiału statystycznego w postaci *tablicy złożonej* (w odróżnieniu od niej szereg statystyczny bywa również określany mianem tablicy prostej). Sytuacja ta często ma miejsce, gdy dokonuje się opracowania materiału zgromadzonego w trakcie badań ankietowych. W tabelach prezentujących wyniki takich badań dokonuje się zwykle „skorelowania" uzyskanych odpowiedzi na poszczególne pytania z określonymi cechami respondentów (np. ich płcią, wiekiem, miejscem zamieszkania).

Właściwie skonstruowana tablica winna składać się z *tytułu, makiety tablicy oraz źródła danych*. *Tytuł tablicy* winien precyzyjnie określać badaną zbiorowość pod względem rzeczowym, czasowym i przestrzennym oraz zawierać ujęte w tablicy cechy statystyczne; innymi słowy winien odpowiadać na następujące pytania: co stanowi opisywaną zbiorowość statystyczną?; jakiego momentu bądź przedziału czasowego dotyczy badanie?; z jakiego obszaru pochodzi badana zbiorowość?; jakie cechy statystyczne zbiorowości ujęto w tablicy?

Makieta tablicy (zwana również tablicą właściwą) składa się z wierszy i kolumn oraz ich tytułów (tytuły wierszy określa się „boczkiem" tablicy, zaś tytuły kolumn „główką" tablicy). Wnętrze tablicy, czyli „pola" znajdujące się na skrzyżowaniach poszczególnych wierszy i kolumn są wypełniane zgromadzonym materiałem statystycznym. Należy tu zaznaczyć, iż każde pole tablicy musi być bezwzględnie wypełnione. Jeśli z różnych względów nie ma możliwości wypełnienia pola tablicy danymi liczbowymi wówczas wykorzystywane są odpowiednie znaki umowne. Należą do nich następujące znaki:

[3] Ten typ szeregów będzie przedmiotem szczegółowej analizy w rozdziale IV

„ – „ *(kreska)*	oznacza, że zjawisko nie występuje,
„ . „ *(kropka)*	oznacza zupełny brak informacji lub brak informacji wiarygodnych,
„0" (zero)	oznacza, że zjawisko występuje w niewielkich ilościach (mniej niż 50% przyjętej jednostki miary),
„×" (ukośny krzyżyk)	oznacza, że wypełnienie danego pola ze względu na układ tablicy jest niemożliwe bądź niecelowe,
„Δ" (pusty trójkąt)	oznacza, że nazwy zostały skrócone w stosunku do obowiązującej klasyfikacji,
„▲" (pełny trójkąt)	oznacza, że dane nie mogą być opublikowane ze względu na konieczność zachowania tajemnicy statystycznej,
„w tym"	oznacza, że nie podaje się wszystkich składników sumy.

Bezpośrednio pod tablicą umieszcza się *źródło danych* zawartych w tablicy. Źródło takie mogą stanowić: rocznik statystyczny, wyniki spisu, badania własne, sprawozdawczość firmy bądź instytucji, itp. Ilustracją prawidłowo skonstruowanej tablicy opisującej zbiorowość studentów ze względu na trzy cechy jest poniższa tablica 2.3.

Tablica 2.3. Studenci Wyższej Szkoły Pedagogicznej w Warszawie według płci, wieku i miejsca zamieszkania (stan na 30.09.2003 r.)

Wyszczególnienie		*Wiek w latach*			*Razem*
		19 – 21	*21 – 23*	*23 – 25*	
Warszawa	Kobiety	211	185	321	717
	Mężczyźni	122	252	87	461
Woj. Mazowieckie	Kobiety	187	124	97	408
	Mężczyźni	142	95	48	285
Inne województwa	Kobiety	45	32	25	102
	Mężczyźni	67	15	-	82
Razem		774	703	578	2055

Źródło: Badania własne

2.2. GRAFICZNY OPIS ZBIOROWOŚCI STATYSTYCZNEJ

Opis graficzny polega na prezentacji zgromadzonego materiału statystycznego ujętego w formie szeregów statystycznych w postaci różnego rodzaju *wykresów statystycznych*. Ta forma opisu statystycznego nie funkcjonuje zazwyczaj samodzielnie. Wykresy stanowią zwykle ilustrację materiału statystycznego ujętego w postaci szeregów bądź tablic i ich zadaniem jest poglądowa prezentacja treści tablic umożliwiająca wstępną, pobieżną

ocenę prawidłowości występujących w badanej zbiorowości. Do najczęściej wykorzystywanych form graficznych zalicza się wykresy: powierzchniowe, liniowe, bryłowe, obrazkowe, mapowe i kombinowane. Ich dobór uzależniony jest od typu szeregu, który ma być zaprezentowany graficznie oraz od celu, jakiemu ma on służyć.

Wykresy powierzchniowe służą głównie do prezentacji szeregów strukturalnych. Wykorzystywane są tu takie figury geometryczne jak: prostokąty, kwadraty, koła. Na podstawie danych ujętych w szeregu statystycznym następuje podział powierzchni tych figur na części proporcjonalnie do struktury badanego zjawiska. W praktyce najczęściej wykorzystywany jest wykres słupkowy zwany również histogramem, który jest umieszczany zazwyczaj w układzie współrzędnych. Na osi odciętych nanoszone są wartości cechy (zarówno liczbowej jak i opisowej), a na osi rzędnych liczebności bądź częstości odpowiadające poszczególnym wariantom cechy. W zależności od typu ilustrowanego szeregu może to być histogram liczebności bądź częstości prostej lub skumulowanej. Wykresy *2.1 i 2.2* stanowią ilustrację tego typu prezentacji graficznych.

Wykresy liniowe (diagramy) są najczęściej wykorzystywane do prezentacji szeregów czasowych. Wykres tego typu umieszczany jest w układzie współrzędnych, gdzie na osi odciętych odkładane są poszczególne jednostki czasu, zaś na osi rzędnych odmierzane są wielkości badanego zjawiska. Ilustruje to wykres *2.3*.

Wykresy punktowe posiadają zastosowanie podobne do wykresów liniowych, a więc mogą służyć ilustracji szeregów czasowych (por. wykres *2.4*)

Wykresy bryłowe pełnią podobną rolę jak wykresy powierzchniowe. W tym przypadku poziom zjawiska ewentualnie jego struktura prezentowana jest przy wykorzystaniu brył (np. sześcianu, prostopadłościanu, kuli). Ten typ prezentacji graficznej ilustruje wykres *2.5*.

Wykresy kombinowane stanowią kompilację wymienionych typów wykresów, np. wykresu powierzchniowego i liniowego, obrazkowego i mapowego itp. (por. wykres *2.6*)

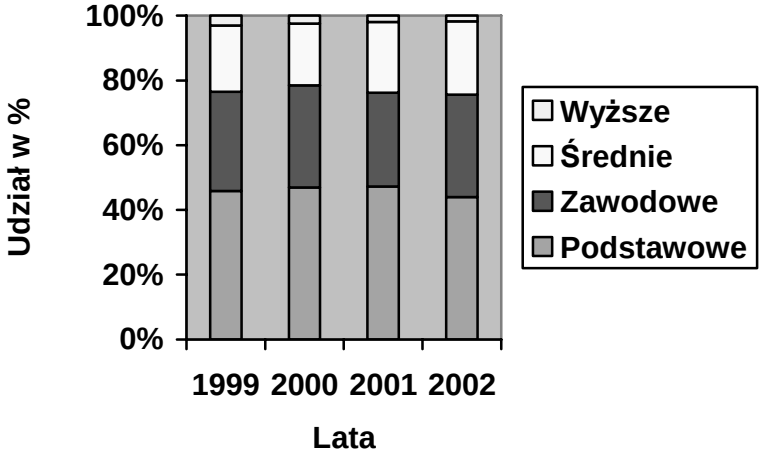
Wykresy mapowe (kartogramy) służą do prezentacji przestrzennego rozmieszczenia badanego zjawiska. Wielkość badanego zjawiska (jego natężenie) na poszczególnych obszarach może być ukazywana poprzez rozmieszczenie odpowiednich symboli o zróżnicowanej wielkości, stosowanie zabarwień o określonych kolorach lub określonego typu zakreskowania. Wykresy tego typu wymagają zwykle zamieszczenia legendy objaśniającej zastosowaną symbolikę. Ilustruje to wykres *2.7*.

Wykresy obrazkowe prezentują wielkość badanego zjawiska za pomocą odpowiednich symboli graficznych (sylwetek obrazków prezentujących badane zjawisko). Tę formę prezentacji ilustruje wykres *2.8*.

Należy dodać, iż w przypadku wykresu, podobnie jak tablicy, należy umieścić jego tytuł oraz źródło danych (tym źródłem najczęściej jest odpowiedni szereg lub tablica statystyczna, na podstawie których został on sporządzony).

Wykres 2.1.

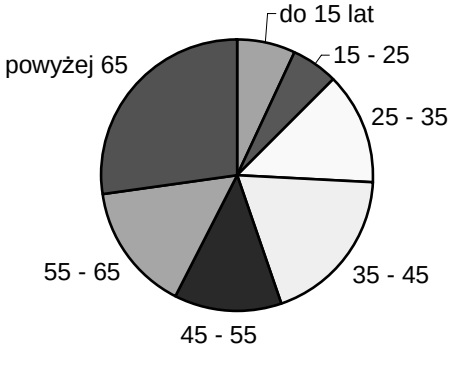
Ludność miasta „Z" wg poziomu wykształcenia w latach 1999-2002



Źródło: Opracowanie własne

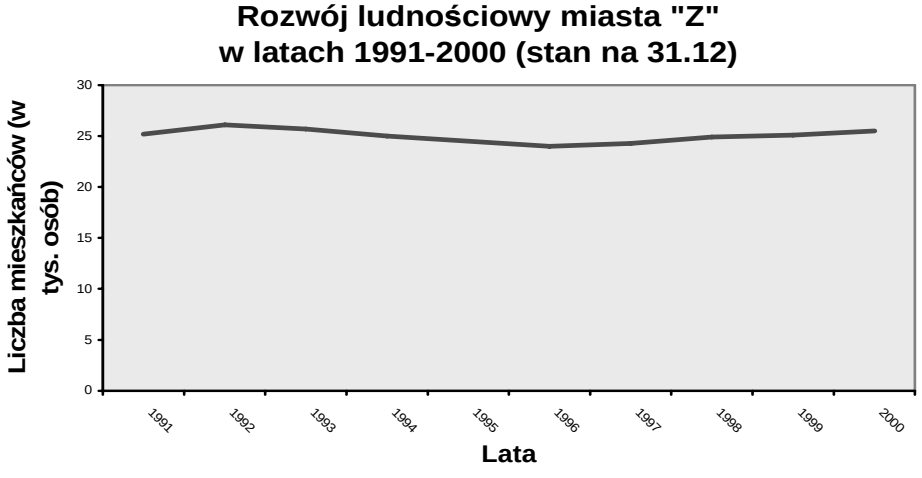
Wykres 2.2.

Struktura wiekowa mieszkanców miasta "Z" w 2001 r. (stan na 31.12.)



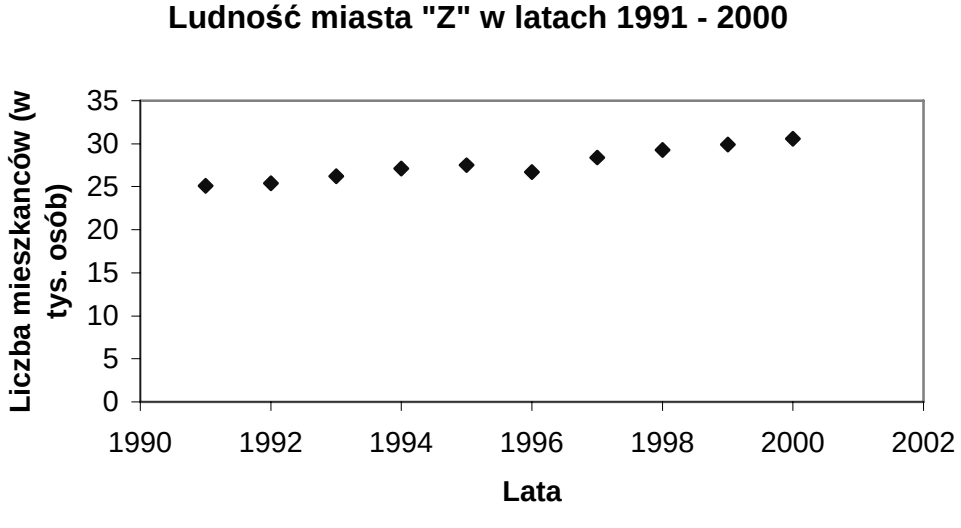
Źródło: Opracowanie własne

Wykres 2.3.



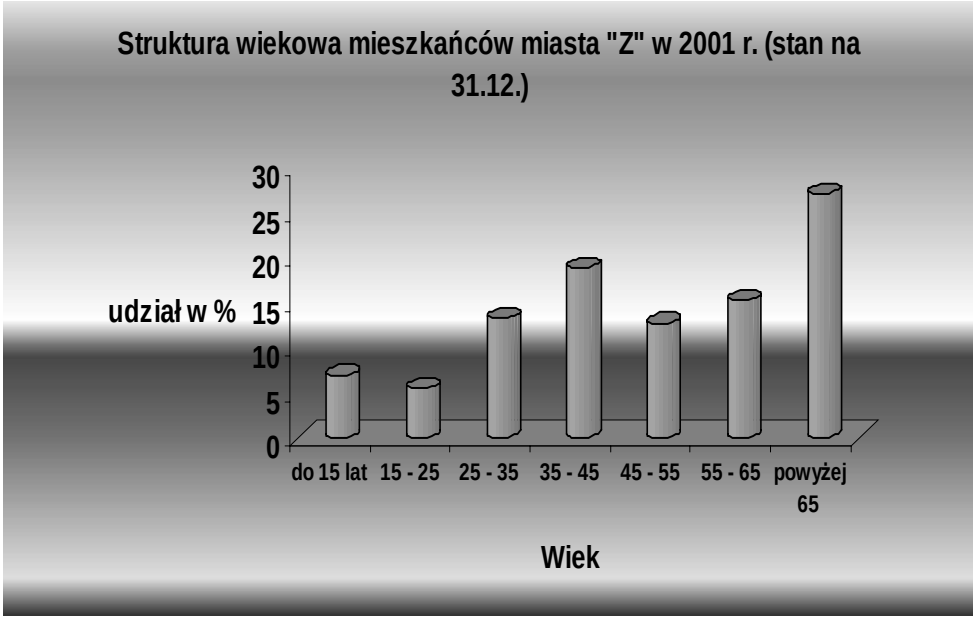
Źródło: Opracowanie własne

Wykres 2.4.



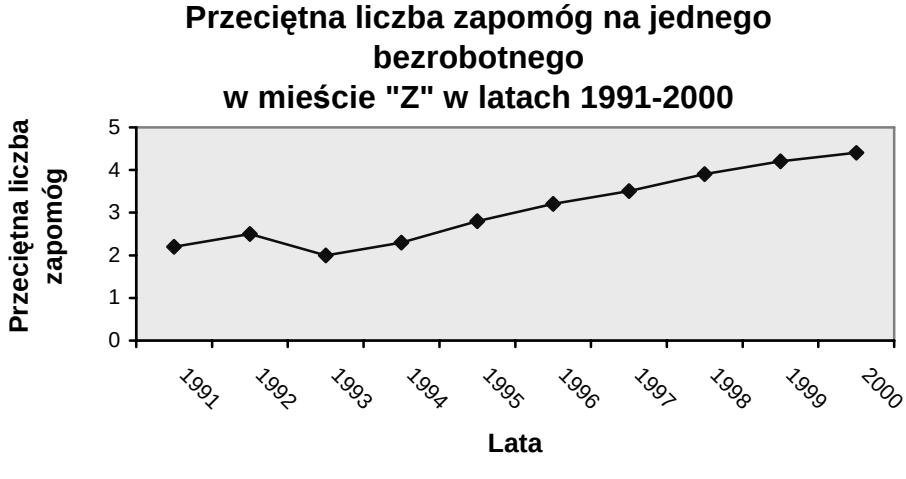
Źródło: Opracowanie własne

Wykres 2.5



Źródło: Opracowanie własne

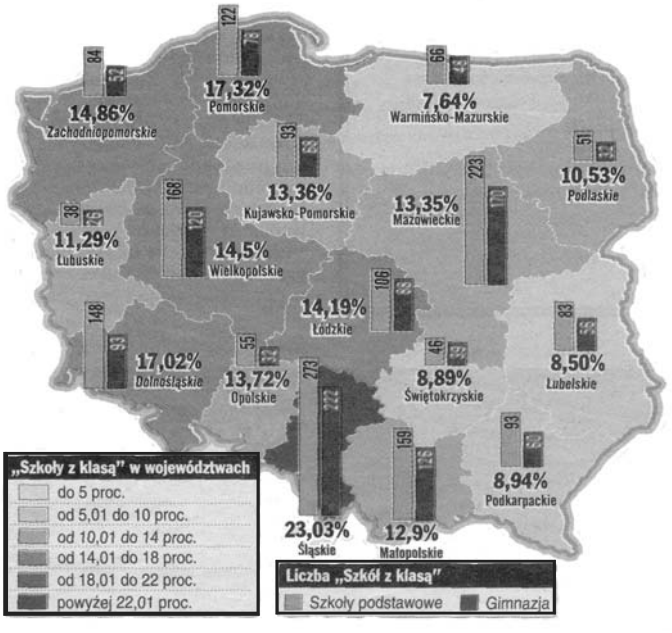
Wykres 2.6



Źródło: Opracowanie własne

Wykres 2.7

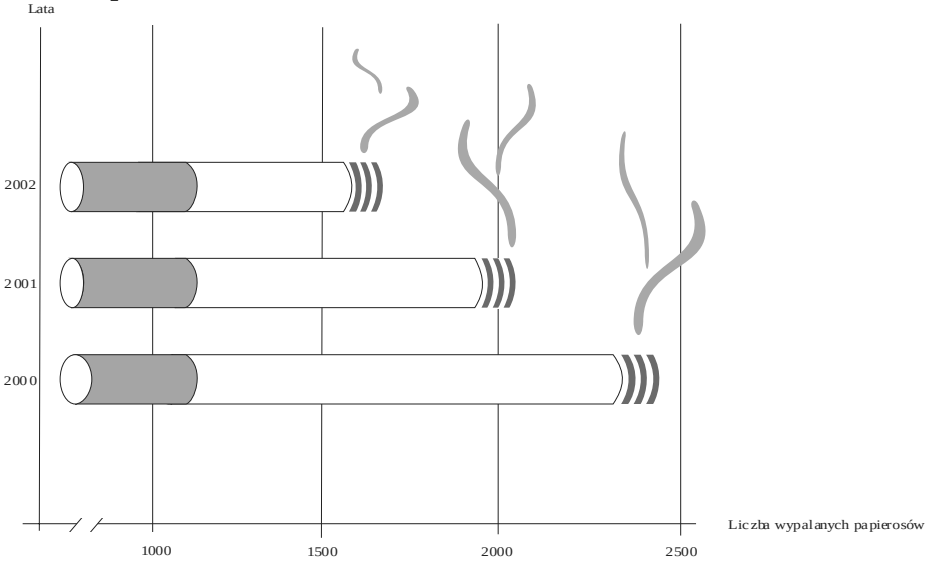
Rozmieszczenie „szkół z klasą" w Polsce we wrześniu 2003 r.



Źródło: „Gazeta Wyborcza" z 14 października 2003 r.

Rys. 2.8

Przeciętna roczna liczba wypalanych papierosów przez mieszkańców miasta „Z" w latach 2000-2002



Źródło: opracowanie własne

2.3. PARAMETRYCZNY OPIS ZBIOROWOŚCI STATYSTYCZNEJ

2.3.1. Istota i klasyfikacja parametrów i momentów statystycznych

Opis parametryczny stanowi jeden z najczęściej wykorzystywanych sposobów opisu rozkładu cechy statystycznej głównie z uwagi na jego syntetyczną i skróconą postać. Ta forma opisu wykorzystuje *parametry statystyczne*, tj. charakterystyki liczbowe opisujące rozkład wartości badanej cechy w szeregu statystycznym. Przy ich pomocy dokonujemy opisu określonych własności analizowanego zbioru wartości cechy. Opis tych własności może obejmować:

a) określenie średniego poziomu zbioru wartości cechy, czyli wybór pojedynczej wartości cechy reprezentującej cały zbiór analizowanych wartości; ma tu zastosowanie grupa parametrów zwanych miarami średnimi i położenia,
b) określenie poziomu zróżnicowania (zmienności) analizowanego zbioru wartości cechy, do czego wykorzystywane są parametry zmienności (rozproszenia),
c) określenie poziomu odchylenia analizowanego rozkładu wartości cechy od rozkładu symetrycznego; badanie tej własności dokonywane jest przy pomocy miar skośności,
d) określenie poziomu skupienia bądź nierównomierności rozłożenia wartości cechy mierzone przy pomocy miar koncentracji,
e) określenie rozmieszczenia realizacji wartości cechy w analizowanym zbiorze poprzez wyznaczenie odpowiednich momentów statystycznych.

Jak wynika z powyższego zestawienia do badania własności ujętych w punktach ***a*** - ***d*** wykorzystywane są odpowiednie grupy parametrów statystycznych (ich szczegółowa charakterystyka zostanie dokonana w dalszej części tego podrozdziału), zaś własność ***e*** badana jest przy pomocy *momentów statystycznych.*

Momenty statystyczne są często wykorzystywanymi charakterystykami rozkładów cechy statystycznej. Wśród nich wyróżnia się dwie podstawowe grupy:

a) *momenty zwykłe*, które są średnimi odchyleń wartości cechy od punktu zerowego podniesionych do potęgi *k*; ich ogólną postać można wyrazić wzorem:

$$M_k(x) = \frac{\sum_{i=1}^{l} (x_i - 0)^k \cdot n_i}{N} = \frac{\sum_{i=1}^{l} x_i^{\,k} \cdot n_i}{N} \qquad 2.3$$

gdzie: $M_k(x)$ – oznacza moment zwykły rzędu *k*,

i = 1, 2, ..., l – warianty (klasy) cechy X

b) *momenty centralne*, które są średnimi odchyleń wartości cechy od ich średniej arytmetycznej podniesionych do potęgi *k*; ich ogólną postać wyraża wzór:

$$m_k(x) = \frac{\sum_{i=1}^{l} (x_i - \bar{x})^k \cdot n_i}{N} \qquad 2.4.$$

gdzie: $m_k(x)$ - oznacza moment centralny rzędu *k*.

Celowa wydaje się prezentacja wybranych momentów statystycznych. I tak wśród momentów zwykłych:

a) moment zwykły rzędu pierwszego jest miarą średnią i nosi nazwę *średniej arytmetycznej*: $\bar{x}$

$$M_1(x) = \frac{\sum_{i=1}^{l} x_i \cdot n_i}{N} = \bar{x},$$

b) moment zwykły rzędu drugiego jest średnią kwadratów wartości cechy:

$$M_2(x) = \frac{\sum_{i=1}^{l} x_i^{\,2} \cdot n_i}{N}$$

c) moment zwykły rzędu trzeciego jest średnią sześcianów wartości cechy:

$$M_3(x) = \frac{\sum_{i=1}^{l} x_i^{\,3} \cdot n_i}{N}$$

Wśród momentów centralnych:

a) moment centralny rzędu pierwszego jest zawsze równy zero na mocy jednej z własności średniej arytmetycznej:

$$m_1(x) = \frac{\sum_{i=1}^{l} (x_i - \bar{x}) \cdot n_i}{N} = 0$$

b) moment centralny rzędu drugiego jest miarą rozproszenia wartości cechy i nosi nazwę *wariancj:*

$$m_2(x) = \frac{\sum_{i=1}^{l} (x_i - \bar{x})^2 \cdot n_i}{N} = s^2(x)$$

c) moment centralny rzędu trzeciego jest wykorzystywany do pomiaru skośności w rozkładzie wartości cechy:

$$m_3(x) = \frac{\sum_{i=1}^{l} (x_i - \bar{x})^3 \cdot n_i}{N}$$

d) moment centralny rzędu czwartego (w postaci standaryzowanej) wykorzystywany jest do pomiaru spłaszczenia rozkładu wartości cechy:

$$m_4(x) = \frac{\sum_{i=1}^{l} (x_i - \bar{x})^4 \cdot n_i}{N}$$

Należy dodać, iż momenty statystyczne mogą być ustalane zarówno dla szeregów szczegółowych jak i rozdzielczych liczebności i częstości. Jak wynika z tej krótkiej prezentacji momenty statystyczne znajdują szerokie zastosowanie w badaniach statystycznych, a znaczna ich część pełni rolę parametrów statystycznych.

2.3.2. Miary średnie i położenia

Ta grupa miar wykorzystywana jest do opisu przeciętnego poziomu analizowanego zbioru wartości cechy. Miary te wskazują wartość najlepiej charakteryzującą wszystkie realizacje występujące w szeregu statystycznym i na ogół możliwe jest nadanie im odpowiedniej treści. Analiza porównawcza zbiorowości statystycznych za pomocą tej grupy miar ma sens, gdy są one w miarę jednorodne. W literaturze spotykany jest najczęściej następujący podział tej grupy miar:

1) *miary średnie klasyczne*, czyli takie, które są ustalane na podstawie pełnego zbioru wartości cechy (innymi słowy: wszystkie jednostki statystyczne badanej zbiorowości mają wpływ na wartość takiej miary). Najczęściej wykorzystywane miary średnie klasyczne to: średnia arytmetyczna, średnia geometryczna, średnia harmoniczna, średnie potęgowe,
2) *miary średnie pozycyjne*, czyli takie o wartości których decydują jedynie jednostki zajmujące określoną pozycję w szeregu statystycznym (innymi słowy: nie wszystkie wartości cechy decydują o poziomie tej miary). Należy tu podkreślić, iż nie wszystkie z nich można określić mianem miar średnich w ścisłym tego słowa znaczeniu. Część z nich ma raczej charakter *miar położenia* i służy do opisu rozmieszczenia wariantów wartości cechy. Podobny charakter posiadają wykorzystywane w opisie parametrycznym momenty statystyczne. Zastosowanie niektórych z nich zostanie zaprezentowane w dalszej części opracowania. Najczęściej stosowane pozycyjne miary średnie to wartość dominująca oraz wartość środkowa. Spośród miar położenia należy wymienić przede wszystkim kwartyle pierwszy, drugi i trzeci oraz decyle.

Średnia arytmetyczna

Średnia arytmetyczna jest ilorazem sumy wszystkich wartości cechy i liczebności tego zbioru. W zależności od postaci materiału statystycznego nieco odmienne są sposoby wyznaczania średniej arytmetycznej. I tak:
a) w przypadku *szeregu szczegółowego* jest ona wyznaczana według następującego wzoru:

$$\bar{x} = \frac{\sum_{i=1}^{N} x_i}{N} \qquad 2.5.$$

gdzie: x_i - oznacza wartości cechy przypisane i-tym jednostkom statystycznym,
N - liczebność analizowanego zbioru wartości cechy.

Tak wyznaczaną średnią określa się niekiedy średnią arytmetyczną prostą (nieważoną, chociaż można również przyjąć, iż mamy tu do czynienia z „wagami" jednostkowymi). Procedurę obliczania średniej arytmetycznej dla tego typu szeregu ilustruje poniższy przykład.

Przykład 2.3.
Zebrano informacje dotyczące wyników egzaminu ze statystyki uzyskanych przez studentów grupy 3 drugiego roku studiów WSP w Wałbrzychu. Otrzymano następujący szereg*: 3,0; 4,5; 4,0; 5,0; 2,0; 3,5; 4,0; 3,0; 4,0; 5,0; 4,5; 3,5; 3,5; 4,0; 3,0; 2,0; 3,5; 5,0; 4,0; 3,5; 3,0; 5,0; 4,0; 3,0; 3,5; 3,5; 4,0; 3,5; 3,5*. Obliczyć średnią ocen uzyskanych z egzaminu ze statystyki.

Rozwiązanie
Średnią arytmetyczną podanych ocen ustalimy według *wzoru 2.5*. Suma ujętych w szeregu 29 ocen wynosi 103,5, podstawiając ją do wzoru otrzymujemy:

$$\bar{x} = \frac{103{,}5}{29} \approx 3{,}57$$

Uzyskana ocena średnia jest nieznacznie wyższa od „dostatecznej plus".
b) w przypadku *szeregu rozdzielczego punktowego* jej ustalanie odbywa się według wzoru:

$$\bar{x} = \frac{\sum_{i=1}^{k} x_i \cdot n_i}{N} \quad \text{lub} \quad \bar{x} = \frac{\sum_{i=1}^{k} x_i \cdot f_i}{\sum_{i=1}^{k} f_i} \qquad 2.6.$$

gdzie: x_i – oznacza występujące w szeregu warianty wartości cechy ($i = 1,2,\ldots, k$)
n_i, f_i – oznaczają przypisane tym wartościom „wagi"; w szczególnym przypadku mogą to być liczebności (n_i) lub częstości (f_i).

Średnia tak wyznaczana określana jest mianem średniej arytmetycznej ważonej, a procedurę jej obliczania ilustruje przykład 2.4.

Przykład 2.4.

Zebrano informacje dotyczące wielkości gospodarstw domowych zamieszkujących miejscowość „K". Otrzymano następujący szereg:

Liczba osób w gospodarstwie	1	2	3	4	5	6	7
Liczba gospodarstw	3	10	25	20	12	5	5

Ustalić średnią wielkość gospodarstwa domowego za pomocą średniej arytmetycznej.

Rozwiązanie

Prezentowany szereg ma charakter szeregu rozdzielczego punktowego (z cechą statystyczną liczbową, skokową). Wyznaczana średnia będzie miała charakter średniej ważonej, gdzie „wagami" są liczebności gospodarstw o określonej wielkości. Zgodnie z *wzorem 2.6* wyznaczamy iloczyny $x_i \cdot n_i$ (są one umieszczone w kolumnie 3 poniższej tabeli)

Liczba osób w gospodarstwie domowym (x_i)	*Liczba gospodarstw* (n_i)	$x_i \cdot n_i$
1	*3*	*3*
2	10	20
3	25	75
4	20	80
5	12	60
6	5	30
7	5	35
Razem	80	303

Sumę iloczynów podstawiamy do wzoru i otrzymujemy:

$$\bar{x} = \frac{303}{80} \approx 3{,}8 \text{ osoby}$$

Średnia liczebność gospodarstwa domowego wynosi 3,8 osoby.

c) w przypadku *szeregu rozdzielczego z przedziałami klasowymi* do jej wyznaczenia wykorzystywany jest poniższy wzór:

$$\bar{x} = \frac{\sum_{i=1}^{k} \dot{x}_i \cdot n_i}{N} \quad \text{lub} \quad \bar{x} = \frac{\sum_{i=1}^{k} \dot{x}_i \cdot f_i}{\sum_{i=1}^{k} f_i} \qquad 2.7.$$

gdzie: $\dot{x}_i$ - oznacza środek i-tego przedziału klasowego ($i = 1,2,\ldots,k$),

n_i, f_i - oznaczają „wagi" przypisane poszczególnym przedziałom klasowym; w szczególnym przypadku mogą to być liczebności (n_i) lub częstości (f_i).

Podanej formuły nie można stosować, gdy w szeregu występuje chociaż jeden otwarty skrajny przedział klasowy (nie ma wówczas możliwości wyznaczenia jego środka). Powszechnie przyjmuje się jednak możliwość wyznaczenia średniej w takiej sytuacji, gdy liczebność takich przedziałów jest niewielka (do 5 % badanej zbiorowości). Wówczas dla przedziału otwartego przyjmuje się rozpiętość taką, jak dla pozostałych klas. Wyznaczanie średniej dla szeregu z przedziałami klasowymi ilustruje przykład 2.5.

Przykład 2.5.

Wśród studentów II roku Wydziału Socjologii zebrano informacje dotyczące ich wzrostu. Uzyskane dane ujęto w poniższym szeregu z przedziałami klasowymi:

Wzrost w cm (x_i)	*Liczba studentów* (n_i)
150 – 158	12
158 – 166	26
166 – 174	45
174 – 182	32
182 – 190	15
Razem	130

Ustalić średni wzrost studentów II roku Wydziału Socjologii.

Rozwiązanie:

Uzyskane informacje ujęto w postaci szeregu z przedziałami klasowymi, co powoduje konieczność ustalenia środków tych przedziałów. W poniższej tablicy roboczej ujęto je w kolumnie 3 i zostały one wyznaczone jako średnia arytmetyczna z górnej i dolnej granicy każdego przedziału.

Wzrost w cm (x_i)	*Liczba studentów* (n_i)	$\dot{x}_i$	$\dot{x}_i \cdot n_i$
1	*2*	*3*	*4*
150 – 158	12	154	1848
158 – 166	26	162	4212
166 – 174	45	170	7650
174 – 182	32	178	5696
182 – 190	15	186	2790
Razem	130	X	22196

W kolumnie 4. ustalono iloczyny $\dot{x}_i \cdot n_i$, a po wstawieniu ich sumy do *wzoru 2.7* otrzymujemy:

$$\bar{x} = \frac{27892}{130} \approx 170{,}74 \text{ cm}$$

Średni wzrost studentów wynosi około 170,74 cm.

Należy podkreślić, że o ile w przypadku dwóch pierwszych formuł ustalona wartość średnia bardzo wiarygodnie odzwierciedla przeciętny poziom wartości cechy w badanej zbiorowości, to w trzecim przypadku ma ona raczej charakter szacunkowy, odbiegający niekiedy dość znacznie od faktycznego poziomu średniego i odchylenie to może wzrastać wraz ze zwiększaniem rozpiętości przedziałów klasowych.

Powszechnie średnia arytmetyczna uważana jest za parametr średni o najkorzystniejszych własnościach. Za najważniejsze spośród nich uważa się następujące:

- jako parametr klasyczny ustalana jest na podstawie wszystkich wartości cechy, a więc posiada wysoką wartość poznawczą (w odróżnieniu np. od parametrów pozycyjnych),
- suma ważona odchyleń poszczególnych wartości cechy od ich średniej arytmetycznej wynosi zawsze zero, co wynika z faktu, że średnia ta pełni rolę „środka ciężkości" analizowanego zbioru wartości cechy. Własność tę można zapisać relacją:

$$\sum_{i=1}^{N}(x_i-\bar{x})=\sum_{i=1}^{k}(x_i-\bar{x})\cdot n_i=\sum_{i=1}^{k}(\dot{x}_i-\bar{x})\cdot n_i=0$$

- ważona suma kwadratów odchyleń poszczególnych wartości cechy od ich średniej arytmetycznej jest najmniejsza z możliwych, co można zapisać następującą zależnością:

$$\sum_{i=1}^{N}(x_i-\bar{x})^2=\sum_{i=1}^{k}(x_i-\bar{x})^2\cdot n_i=\sum_{i=1}^{k}(\dot{x}_i-\bar{x})^2\cdot n_i=\min$$

Własność ta jest wykorzystywana w znanej z ekonometrii metodzie najmniejszych kwadratów.

- jeśli w szeregu rozdzielczym wszystkie wagi (w_i) - w szczególnym przypadku będą to liczebności (n_i) bądź częstości (f_i) - pomnożymy (bądź podzielimy) przez ten sam czynnik q, to średnia arytmetyczna wartości cechy z nowym systemem wag (v_i), gdzie:

$$v_i=w_i\cdot q \qquad \text{lub} \qquad v_i=\frac{w_i}{q},$$

będzie identyczna jak średnia liczona według pierwotnych wag (w_i). Można to ująć w następującej relacji:

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

Wynika to z faktu, że wartość średniej arytmetycznej nie zależy od absolutnych wielkości wag, lecz od proporcji występujących między nimi,

- jeśli wszystkie wartości cechy X podzielimy (bądź pomnożymy) przez tę samą wielkość q to średnia arytmetyczna tak zmienionych wartości cechy będzie q-krotnie mniejsza (lub q-krotnie większa) od średniej pierwotnych wartości cechy. Własność tę można zapisać następującymi relacjami:

$$\frac{\sum_{i=1}^{k}\left(\frac{x_i}{q}\right)\cdot n_i}{N} = \frac{\frac{\sum_{i=1}^{k} x_i \cdot n_i}{N}}{q} = \frac{\overline{x}}{q} \quad \text{lub}$$

$$\frac{\sum (x_i \cdot q)\cdot n_i}{N} = \left[\frac{\sum_{i=1}^{k} x_i \cdot n_i}{N}\right]\cdot q = \overline{x}\cdot q$$

gdzie: $\overline{x}$ - średnia arytmetyczna pierwotnych wartości cechy.

- jeśli do wszystkich wartości cechy X dodamy (bądź od wszystkich wartości odejmiemy) tę samą wielkość q to średnia arytmetyczna tak zmienionych wartości cechy będzie o wielkość q większa (lub o wielkość q mniejsza) od średniej liczonej dla pierwotnych wartości cechy. Własność tę ujmują poniższe relacje:

$$\frac{\sum_{i=1}^{k}(x_i + q)\cdot n_i}{N} = \left[\frac{\sum_{i=1}^{k} x_i \cdot n_i}{N}\right] + q = \overline{x} + q \quad \text{lub}$$

$$\frac{\sum_{i=1}^{k}(x_i - q)\cdot n_i}{N} = \left[\frac{\sum_{i=1}^{k} x_i \cdot n_i}{N}\right] - q = \overline{x} - q$$

gdzie: $\overline{x}$ - średnia arytmetyczna pierwotnych wartości cechy.

- jeśli badaną zbiorowość podzielimy na kilka podzbiorowości to średnia arytmetyczna dla całej zbiorowości będzie średnią arytmetyczną ze średnich tych podzbiorowości,

$$\overline{x} = \frac{\sum_{j=1}^{l} \overline{x}_j \cdot n_j}{\sum_{j=1}^{l} n_j} = \frac{\sum_{j=1}^{l} \overline{x}_j \cdot f_j}{\sum_{j=1}^{l} f_j}$$

gdzie: $\bar{x}_j$ - średnie arytmetyczne wyodrębnionych podzbiorowości ($j = 1,2,...,l$),

n_j, f_j - liczebności bądź częstości odpowiadające tym podzbiorowościom.

- jest wykorzystywana do wyznaczania wielu innych parametrów statystycznych.

Średnia arytmetyczna z uwagi na sposób jej wyznaczania może być wykorzystywana dla oznaczania przeciętnego poziomu wartości cech jedynie typu liczbowego. Dla cech opisowych wyrażanych na skali nominalnej właściwą miarą średnią jest dominanta, zaś dla cech mierzonych na skali porządkowej – mediana bądź dominanta. Miary te zostaną omówione w dalszej części opracowania.

Średnia geometryczna

Należy do klasycznych parametrów średnich i jest stosowana do wyznaczania średniego poziomu wartości cechy podanych w postaci szeregu czasowego momentów. Jej typowe wykorzystanie to badanie średniego tempa zmian. Wyznaczana jest ona według wzoru:

$$x_g = \sqrt[N]{x_1 \cdot x_2 \cdot ... \cdot x_N} \qquad 2.8.$$

gdzie: $x_1, x_2, ..., x_N$ - oznaczają występujące w szeregu czasowym wartości cechy.

Jako podstawowe uznaje się następujące własności średniej geometrycznej:

- obliczana jest jedynie dla dodatnich wartości cechy; jeśli choć jedna wartość cechy wynosiłaby zero, miara również przyjmie wwartość zero,
- jest ona mniej wrażliwa od średniej arytmetycznej na zróżnicowanie wartości cechy.

Procedurę wyznaczania średniej geometrycznej ilustruje poniższy przykład.

Przykład 2.6.

Informacje dotyczące liczby mieszkańców miejscowości „K" w latach 1991 – 1997 zawiera poniższa tablica 2.2.

Tablica 2.2. Mieszkańcy miejscowości „K" w latach 1991 – 1997 (stan na 31.12.)

Lata	*Liczba mieszkańców (w tys. osób)*
1991	15,9
1992	16,3
1993	15,7
1994	16,0
1995	17,1
1996	17,5
1997	18,0

Źródło: Dane umowne

Na podstawie podanych informacji obliczyć przeciętną liczbę mieszkańców tej miejscowości w badanym okresie.

Rozwiązanie:
Informacje podano w postaci szeregu czasowego momentów (liczba mieszkańców na koniec każdego roku), a więc dla obliczenia poziomu średniego wykorzystana zostanie średnia geometryczna. Jej obliczenia dokonamy według *wzoru 2.8*. i po podstawieniu danych otrzymujemy:

$$x_g = \sqrt[7]{15{,}9 \cdot 16{,}3 \cdot 15{,}7 \cdot 16{,}0 \cdot 17{,}1 \cdot 17{,}5 \cdot 18{,}0} = 16{,}62$$

Uzyskany wynik informuje, że przeciętna liczba mieszkańców miejscowości „K" w latach 1991 – 1997 wynosiła 16,62 tys. osób.

Dominanta

Dominanta, oznaczana jako *D(x)*, zwana również *wartością modalną* bądź *typową*, jest wartością cechy występującą najliczniej (najczęściej) w badanej zbiorowości. Zaleca się wykorzystywanie tego parametru głównie w przypadku badania zbiorowości ze względu na cechy opisowe, gdyż ograniczone są wówczas możliwości wykorzystywania innych parametrów średnich.

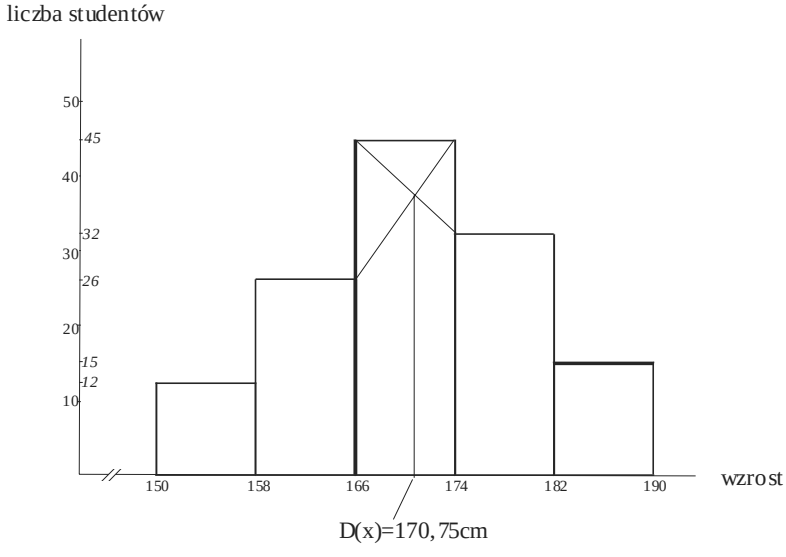
Procedura wyznaczania tego parametru w przypadku *szeregu szczegółowego i rozdzielczego punktowego* polega na wskazaniu takiej wartości cechy, która w zgromadzonym materiale statystycznym powtarza się najczęściej (o ile taka wartość występuje). W *przykładzie 2.3* będzie to ocena 3,5, ponieważ występuje ona najliczniej (aż 9-krotnie). W szeregu rozdzielczym punktowym (zob. *przykład 2.4*), najliczniej występującą wielkością gospodarstwa domowego jest gospodarstwo 3-osobowe; takich gospodarstw jest 25, a więc dominanta wynosi 3.

W przypadku *szeregu z przedziałami klasowymi* procedura wyznaczania dominanty jest bardziej skomplikowana. Określa się w takim przypadku jej wartość przybliżoną. Można tego dokonać *metodą graficzną bądź analityczną*.

Metoda graficzna opiera się na wykorzystaniu histogramu, na którym w najwyższym słupku prowadzone są dwa odcinki biegnące po przekątnej do wysokości sąsiadujących słupków, a punkt ich przecięcia „zrzutowany" na oś odciętych wyznacza przybliżoną wartość dominanty. Procedurę tę opartą na danych z *przykładu 2.5* ilustruje rys. 2.9.

Rys. 2.9

Wyznaczanie dominanty metodą graficzną



Źródło: opracowanie własne

Podejście analityczne umożliwia oszacowanie wartości dominanty metodą interpolacyjną przy wykorzystaniu wzoru:

$$D(x) = x_0 + \frac{n_0 - n_{0-1}}{(n_0 - n_{0-1}) + (n_0 - n_{0+1})} \cdot h_0 \qquad 2.9.$$

gdzie: x_0 – dolna granica przedziału dominującego,

h_0 – rozpiętość przedziału dominującego,

n_0 – liczebność (częstość) przedziału dominującego,

n_{0-1} – liczebność (częstość) przedziału poprzedzającego przedział dominujący,

n_{0+1} – liczebność (częstość) przedziału następnego po przedziale dominującym.

Przy wykorzystaniu podanego wzoru oszacujmy dominujący wzrost dla danych z *przykładu 2.5*. Przedziałem dominanty będzie przedział 166 – 174 cm, w którym występuje najwyższa liczebność ($n_3 = 45$). Wstawiając odpowiednie dane do *wzoru 2.9*. otrzymujemy:

$$D(x) = 166 + \frac{45-26}{(45-26)+(45-32)} \cdot 8 = 166 + 4{,}75 = 170{,}75 \text{ cm}$$

Otrzymany wynik wskazuje, iż dominujący wzrost w badanej zbiorowości wynosi około 170,75 cm. Jest to oczywiście wielkość przybliżona i w badanej zbiorowości może w ogóle nie występować. W takim przypadku dominantę należy traktować jako wartość cechy, wokół której skupiona jest największa liczba jednostek badanej zbiorowości.

Wyznaczanie dominanty w szeregu z przedziałami klasowymi jest możliwe, gdy przedział dominanty i przedziały bezpośrednio z nim sąsiadujące posiadają jednakową rozpiętość. Jeśli przedziałem dominanty jest jeden ze skrajnych przedziałów w szeregu (pierwszy lub ostatni), wówczas przyjąć należy liczebność równą zero dla przedziału poprzedzającego - w pierwszym przypadku lub dla przedziału następnego - w drugim przypadku. Jeśli w szeregu występują przedziały klasowe o zróżnicowanej rozpiętości wówczas dominantę, a ściślej ujmując jej przedział określamy przez wyznaczenie przedziału o najwyższym natężeniu liczebności lub częstości.

Podsumowując stwierdzić również należy, że jakkolwiek dominanta kojarzy się najczęściej z występowaniem w szeregu jednej wartości dominującej, to w praktyce można się spotkać również z przypadkami występowania rozkładów bimodalnych (występują dwie dominanty) bądź nawet trimodalnych (trzy dominanty). Takie przypadki wymagają jednak dodatkowej analizy zebranego materiału statystycznego w celu stwierdzenia, czy ta „wielomodalność" ma charakter systematyczny (regularny).

Mediana

Mediana, oznaczana w dalszym ciągu jako Me(x), zwana również *wartością środkową* należy do grupy miar pozycyjnych określanych mianem kwartyli (jest kwartylem drugim). Jest ona wartością cechy dzielącą badaną zbiorowość na dwie równoliczne części: połowę jednostek o wartościach cechy mniejszych lub równych medianie i drugą połowę o wartościach cechy większych lub równych medianie. Sposób wyznaczania mediany uzależniony jest od typu szeregu, w którym ujęto zgromadzony materiał statystyczny.

Procedura wyznaczania mediany w *szeregu szczegółowym i rozdzielczym punktowym* jest podobna. W szeregu szczegółowym procedura ta wymaga w pierwszej kolejności uporządkowania (w ciągu rosnącym bądź malejącym) takiego szeregu (szereg rozdzielczy punktowy najczęściej jest już uporządkowany). Dalsze postępowanie uzależnione jest od liczebności badanej zbiorowości:

a) gdy liczebność ta jest nieparzysta, medianę stanowi wartość cechy występująca u jednostki środkowej. Jeśli liczebność zbiorowości oznaczymy jako N, to mediana będzie wartością cechy występującą u jednostki o numerze $\frac{N+1}{2}$, co można wyrazić następująco:

$$Me(x) = x_{\frac{N+1}{2}},$$

b) w przypadku parzystej liczby jednostek mediana będzie średnią arytmetyczną wartości cechy występujących u dwóch jednostek środkowych o numerach: $\frac{N}{2}$ i $\frac{N}{2}+1$, co można wyrazić wzorem:

$$Me(x) = \frac{x_{\frac{N}{2}} + x_{\frac{N}{2}+1}}{2} \qquad 2.10.$$

W przypadku wyznaczania mediany w szeregu rozdzielczym dla bardzo licznej zbiorowości wskazane jest skumulowanie takiego szeregu, gdyż znacznie ułatwia to ustalenie tego parametru.

Procedurę wyznaczania mediany dla wymienionych typów szeregów ilustrują przykłady 2.6 i 2.7.

Przykład 2.6

Na podstawie informacji podanych w *przykładzie 2.3* wyznaczyć medianę uzyskanych ocen z egzaminu ze statystyki.

Rozwiązanie:

Badany zbiór liczy 29 ocen (liczebność nieparzysta), wobec tego medianę stanowi ocena o numerze (29 + 1)/2 = 15. Aby ją wyznaczyć dokonujemy uporządkowania danych i w rezultacie otrzymujemy szereg o postaci: *2,0; 2,0; 3,0; 3,0; 3,0; 3,0; 3,0; 3,5; 3,5; 3,5; 3,5; 3,5; 3,5; 3,5; **3,5**; 3,5; 4,0; 4,0; 4,0; 4,0; 4,0; 4,0; 4,0; 4,5; 4,5; 5,0; 5,0; 5,0; 5,0*. Ocena zajmująca środkowe miejsce w szeregu (tj. ocena o numerze 15) wynosi 3,5. Ocena ta stanowi medianę.

Przykład 2.7.

Na podstawie informacji podanych w *przykładzie 2.4* określić medianę wielkości gospodarstw domowych.

Rozwiązanie

Informacje zostały ujęte w postaci szeregu rozdzielczego punktowego. Dla ułatwienia wskazania mediany dokonujemy kumulowania liczebności (kolumna 3. poniższej tabeli):

Liczba osób w gospodarstwie domowym (x_i)	*Liczba gospodarstw* (n_i)	*Liczba gospodarstw skumulowana* ($cum\ n_i$)
1	*2*	*3*
1	3	3
2	10	13
3	25	38
4	20	58
5	12	70
6	5	75
7	5	80
Razem	80	X

Badana zbiorowość obejmuje parzystą liczbę gospodarstw, medianę stanowi wobec tego średnia arytmetyczna liczby osób w rodzinie dla gospodarstw o numerach: 80/2 = 40 i (80/2) +1 = 41. W szeregu skumulowanym zauważamy, iż są to gospodarstwa o jednakowej liczbie osób wynoszącej 4, wobec czego mediana liczby osób w gospodarstwie wynosi 4.

W przypadku *szeregu rozdzielczego z przedziałami klasowymi* opisującego zwykle zbiorowości ze względu na cechę liczbową ciągłą wyznaczanie mediany – podobnie jak dominanty – ma charakter szacunkowy. M. in. z uwagi na ten fakt nie wyróżnia się tu odmiennego postępowania dla parzystej i nieparzystej liczebności badanej zbiorowości. Wartość mediany może być oszacowana dwiema metodami: *graficzną bądź analityczną*.

Metoda graficzna wymaga wykreślenia w układzie współrzędnych histogramu i diagramu kumulacyjnego. Na osi rzędnych wyznaczamy punkt odpowiadający wielkości $\frac{N}{2}$ (mediana dzieli zbiorowość na dwie równe części). Z wysokości tego punktu prowadzimy prostą równoległą do osi odciętych, aż do punktu jej przecięcia z linią diagramu. Ten punkt przecięcia „zrzutowany" na oś odciętych wyznacza przybliżoną wartość mediany.

Metoda analityczna zakłada interpolację wartości mediany według poniższego wzoru:

$$Me(x) = x_0 + \frac{\frac{N}{2} - cum_{n_{0-1}}}{n_0} \cdot h_0 \qquad 2.11.$$

gdzie: x_0 – dolna granica przedziału mediany,

n_0 – liczebność (częstość) przedziału mediany,

h_0 – rozpiętość przedziału mediany,

$cum_{n_{0-1}}$ – liczebność (częstość) skumulowana do przedziału poprzedzającego przedział mediany.

W przypadku obu metod zakłada się równomierne rozłożenie jednostek w całym przedziale mediany, co uprawnia do korzystania z podanych wyżej reguł postępowania.
Procedurę szacowania mediany metodami analityczną i graficzną ilustruje *przykład 2.8.*

Przykład 2.8.
Na podstawie informacji podanych w *przykładzie 2.5* ustalić medianę wzrostu w badanej grupie studentów.

Rozwiązanie
Dla potrzeb metody analitycznej i graficznej dokonujemy kumulowania liczebności jak w kolumnie 3. poniższej tabeli:

Wzrost w cm (x_i)	*Liczba studentów* (n_i)	*Liczba studentów skumulowana* ($cum\ n_i$)
1	*2*	*3*
150 – 158	12	12
158 – 166	26	38
166 – 174	45	83
174 – 182	32	115
182 – 190	15	130
Razem	130	X

Metoda analityczna – zgodnie z *wzorem 2.11* - wymaga wyznaczenia w pierwszej kolejności przedziału mediany; jest to przedział, dla którego $cumn_i$ jest nie mniejsze niż $\frac{N}{2}$ (w naszym przykładzie jest to przedział 166 – 174 cm). Po podstawieniu odpowiednich wielkości do podanego wzoru otrzymujemy:

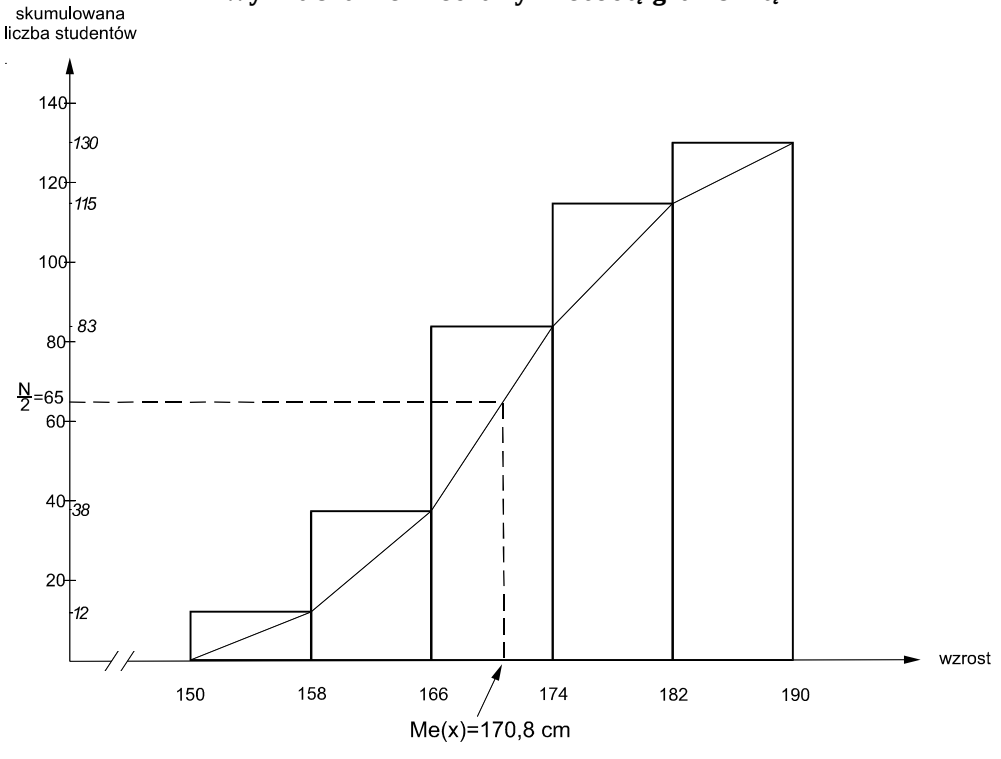
$$Me(x) = 166 + \frac{\frac{130}{2} - 38}{45} \cdot 8 = 170{,}8 \text{ cm}$$

Uzyskana wielkość oznacza, iż wzrost osoby „środkowej" wynosi około 170,8 cm.

W przypadku metody graficznej wykreślamy histogram i diagram liczebności skumulowanej, a następnie zgodnie z podaną procedurą wyznaczamy wartość mediany (zob. rys 2.10).

Rys. 2.10

Wyznaczanie mediany metodą graficzną



Źródło: opracowanie własne

Oszacowana tą metodą wartość mediany jest zbliżona do uzyskanej metodą analityczną.

Podsumowując należy wskazać, że wartość mediany nie zależy od skrajnych wartości cechy, co oznacza, iż może być ona wyznaczana dla szeregów z otwartymi przedziałami skrajnymi. Należy również podkreślić najważniejszą własność tego parametru: suma bezwzględnych odchyleń wszystkich wartości cechy od mediany jest minimalna, co można ująć następującą relacją:

$$\sum_{i=1}^{N}\left|x_i - Me(x)\right| < \sum_{i=1}^{N}\left|x_i - Z\right|,$$

gdzie: Z – dowolna wartość liczbowa.

Inne parametry położenia

Dzieląc badaną zbiorowość na dowolną liczbę części uzyskujemy możliwość wyznaczenia parametrów określanych ogólnym mianem *kwantyli*. Spośród nich najczęściej wykorzystywane są *kwartyle* dzielące zbiorowość

na cztery równoliczne części (inne to: *decyle* dzielące zbiorowość na dziesiętne części oraz centyle dzielące zbiorowość na setne części).

Kwartyle dzielą uporządkowaną zbiorowość na jednakowo liczne „ćwiartki": kwartyl pierwszy oddziela pierwsze 25% badanej zbiorowości od pozostałych 75 %, kwartyl drugi oddziela pierwsze 50% zbiorowości od pozostałej połowy; jest to więc poznana już mediana, kwartyl trzeci oddziela pierwsze 75% badanej zbiorowości od pozostałych 25%. Procedura wyznaczania kwartyli: pierwszego i trzeciego jest zbliżona do metodyki ustalania mediany; różnica sprowadza się jedynie do innego położenia tych parametrów. Dodatkowo przyjmuje się, iż parametry te winny być wyznaczane dla licznych zbiorowości; unika się wówczas sytuacji, gdy wyznaczona wartość parametru nie może być przyporządkowana konkretnej jednostce.

W przypadku wyznaczania kwartyli dla *szeregu szczegółowego* (oczywiście w postaci uporządkowanej) bądź *rozdzielczego punktowego*: kwartyl pierwszy będzie odpowiadał wartości cechy występującej u jednostki o numerze $\frac{N}{4}$, zaś kwartyl trzeci – o numerze $\frac{3N}{4}$. W *szeregu rozdzielczym z przedziałami klasowymi* można – podobnie jak w przypadku mediany – zastosować *metodę analityczną lub graficzną*.

W przypadku *metody analitycznej* kwartyl pierwszy będzie wyznaczany zgodnie z wzorem:

$$Q_1(x) = x_0 + \frac{\frac{N}{4} - cum_{n_{0-1}}}{n_0} \cdot h_0 \qquad 2.12.$$

zaś kwartyl trzeci według wzoru:

$$Q_3(x) = x_0 + \frac{\frac{3N}{4} - cum_{n_{0-1}}}{n_0} \cdot h_0 \qquad 2.13.$$

gdzie: x_0 – dolna granica przedziału zawierającego kwartyl pierwszy bądź trzeci,

n_0 – liczebność (częstość) przedziału zawierającego kwartyl pierwszy bądź trzeci,

h_0 – rozpiętość przedziału zawierającego kwartyl pierwszy bądź trzeci,

$cum_{n_{0-1}}$ – liczebność (częstość) skumulowana do przedziału poprzedzającego przedział zawierający kwartyl pierwszy bądź trzeci.

Metoda graficzna wymaga wykreślenia histogramu i diagramu kumulacyjnego, na którym – podobnie jak w przypadku mediany – wskazujemy przybliżoną wartość wymienionych kwartyli. Procedurę wyznaczania kwar-

tyli pierwszego i trzeciego dla szeregu z przedziałami klasowymi ilustruje przykład 2.9.

Przykład 2.9.
Na podstawie informacji podanych w *przykładzie 2.5*. oszacować wartości kwartyli pierwszego i trzeciego dla wzrostu badanej grupy studentów.

Rozwiązanie
Dla potrzeb metody analitycznej i graficznej dokonujemy kumulowania liczebności jak w poniższej tabeli:

Wzrost w cm (x_i)	*Liczba studentów* (n_i)	*Liczba studentów skumulowana* (*cum* n_i)
1	*2*	*3*
150 – 158	12	12
158 – 166	26	38
166 – 174	45	83
174 – 182	32	115
182 – 190	15	130
Razem	130	X

Metoda analityczna.

Kwartyl pierwszy znajduje się w przedziale 158 – 166, ponieważ *cum* n_i dla tego przedziału jest nie mniejsze niż $\frac{N}{4}$, tj. 32,5, zaś kwartyl trzeci w przedziale 174 – 182, ponieważ *cum* n_i dla tego przedziału jest nie mniejsze od $\frac{3N}{4}$, tj. 97,5. Po podstawieniu odpowiednich danych do *wzorów 2.12 i 2.13* otrzymujemy:

$$Q_1(x) = 158 + \frac{32{,}5 - 12}{26} \cdot 8 = 164{,}3 \text{ cm},$$

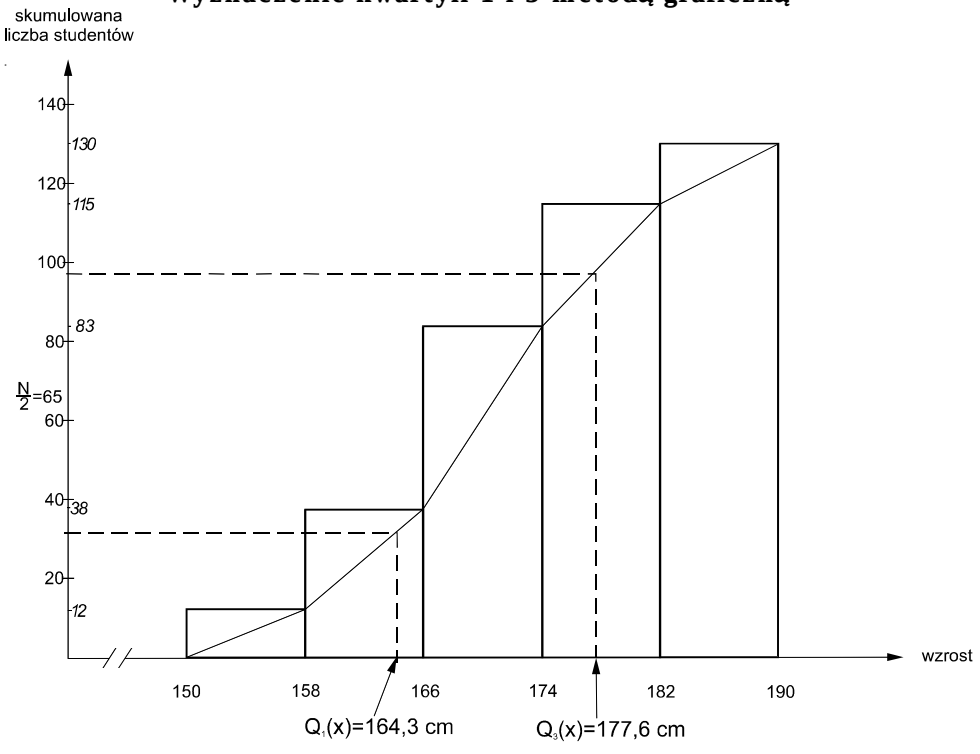
$$Q_3(x) = 174 + \frac{97{,}5 - 83}{32} \cdot 8 = 177{,}6 \text{ cm}.$$

Metoda graficzna

Po wykreśleniu histogramu i diagramu powyższego szeregu kumulacyjnego dokonujemy oszacowania wartości odpowiednich kwartyli jak na rys 2.11.

Rys. 2.11

Wyznaczenie kwartyli 1 i 3 metodą graficzną



Źródło: opracowanie własne

2.3.3. Miary zmienności

Miary zmienności, określane również mianem miar rozproszenia bądź dyspersji, służą do pomiaru poziomu zróżnicowania wartości cechy w badanej zbiorowości. Pomiar zmienności może być wyrażony w jednostkach naturalnych lub stosunkowych. W związku z tym wyróżnia się dwie grupy tych miar:

a) *absolutne miary zmienności*, które mają charakter miar mianowanych,
b) *względne miary zmienności*, ustalane w relacji do przeciętnej wartości badanej cechy.

Biorąc pod uwagę sposób ustalania wartości tych miar można dodatkowo wyodrębnić:

a) *pozycyjne miary zmienności*, które są ustalane w oparciu o wartości cechy zajmujące określone miejsce (pozycję) w badanym szeregu,
b) *klasyczne miary zmienności*, które są ustalane przy wykorzystaniu pełnego zbioru wartości cechy.

Do *absolutnych, pozycyjnych miar zmienności* zalicza się *obszar zmienności* (rozstęp) i *odchylenie ćwiartkowe*; zaś do *absolutnych, klasycznych*

miar zmienności: *wariancję, odchylenie standardowe, odchylenie przeciętne*. Grupę *względnych miar zmienności* reprezentuje *współczynnik zmienności*, który jest ilorazem absolutnych miar zmienności i wybranych miar średnich badanej cechy.

Obszar zmienności

Obszar zmienności zwany również rozstępem wartości cechy oznaczany jako $R(x)$ stanowi najprostszą miarę zmienności i jest określany jako różnica między najwyższą (x_{max}) i najniższą (x_{min})[4] wartością badanej cechy, co można wyrazić wzorem:

$$R(x) = x_{max} - x_{min} \qquad 2.14.$$

Z uwagi na fakt, że miara ta uwzględnia jedynie dwie skrajne wartości cechy, może ona mało wiarygodnie odzwierciedlać poziom zróżnicowania badanej cechy (zwłaszcza, gdy wartości skrajne istotnie odbiegają od typowych wartości cechy). Miara ta może być stosowana do wstępnego badania zmienności. Praktyczne jej wykorzystanie ma miejsce w przypadku konstrukcji szeregu rozdzielczego z przedziałami klasowymi, gdyż zachodzi wówczas konieczność określenia liczby i wielkości klas w ten sposób, by objęły one całą rozpiętość wartości cechy.

Odchylenie ćwiartkowe

Odchylenie ćwiartkowe oznaczane jako $O_c(x)$ stanowi połowę różnicy między kwartylami trzecim - $Q_3(x)$ i pierwszym – $Q_1(x)$, co można wyrazić wzorem:

$$O_c(x) = \frac{Q_3(x) - Q_1(x)}{2} \qquad 2.15.$$

W literaturze podaje się również sposób inny sposób wyznaczania odchylenia ćwiartkowego. Zgodnie z nim parametr ten definiuje się jako średnią arytmetyczną odchyleń kwartyli pierwszego i trzeciego od mediany. Można to wyrazić wzorem:

$$O_c(x) = \frac{[Q_3(x) - Me(x)] + [Me(x) - Q_1(x)]}{2} \qquad 2.16.$$

Jak łatwo zauważyć po niewielkim przekształceniu ta postać wzoru może być zredukowana do *postaci 2.15*. W porównaniu z poprzednią miarą odchylenie ćwiartkowe pozbawione jest wpływu jednostek „nietypowych" dla

[4] w przypadku szeregu rozdzielczego z przedziałami klasowymi wielkościami tymi są odpowiednio: górna granica ostatniego i dolna granica pierwszego przedziału.

badanej zbiorowości, dla których wartości cechy wyraźnie odbiegają od pozostałych, na wartość tej miary. Wartości „nietypowe" znajdą się bowiem albo poniżej kwartyla pierwszego, albo powyżej kwartyla trzeciego. Mankamentem tej miary jest problem z nadaniem jej treści merytorycznej, dlatego miara ta jest rzadko stosowana. Zalecana jest wówczas, gdy w szeregu występują skrajne klasy otwarte i nie ma możliwości zastosowania klasycznych miar zmienności.

Wyznaczając omówione wyżej miary zmienności dla danych z *przykładu 2.3* otrzymujemy:

$$R(x) = x_{\max} - x_{\min} = 7 - 1 = 6 \text{ osób}$$

i

$$O_c(x) = \frac{Q_3(x) - Q_1(x)}{2} = \frac{5-3}{2} = 1 \text{ osoba}$$

Wariancja

Wariancja (*od łacińskiego słowa variare- zmieniać się, różnić się)* należy do klasycznych, absolutnych miar zmienności i można ją definiować jako średnią kwadratów odchyleń poszczególnych wartości cechy od ich średniej arytmetycznej. W zależności od postaci materiału statystycznego wariancja, oznaczana w dalszej części jako $s^2(x)$ będzie wyznaczana według poniższych wzorów:

a) *dla szeregu szczegółowego*:

$$s^2(x) = \frac{\sum_{i=1}^{N} (x_i - \bar{x})^2}{N} \qquad 2.17.$$

b) *dla szeregu rozdzielczego punktowego*:

$$s^2(x) = \frac{\sum_{i=1}^{k} (x_i - \bar{x})^2 \cdot n_i}{N} \qquad 2.18.$$

c) *dla szeregu rozdzielczego z przedziałami klasowymi*:

$$s^2(x) = \frac{\sum_{i=1}^{k} (\dot{x}_i - \bar{x})^2 \cdot n_i}{N} \qquad 2.19.$$

Jakkolwiek *wzory 2.18 i 2.19* odnoszą się do szeregów liczebności, to jednak - podobnie jak w przypadku średniej arytmetycznej - liczebności te (n_i) mogą być zastąpione częstościami (f_i), natomiast liczebność ogólna (N) sumą częstości równą 1 lub 100 (w zależności od sposobu jej wyrażenia).

Istotnym mankamentem wariancji jest to, że jej miano nie jest naturalnym dla badanej cechy, co wynika z potęgowania odchyleń wartości cechy od średniej arytmetycznej. Utrudnia to nadanie jej wartościom treści merytorycznej. W praktyce wykorzystuje się pierwiastek kwadratowy z wariancji określany mianem *odchylenia standardowego*. Niemniej wariancja posiada wiele pozytywnych właściwości, do których można zaliczyć następujące:

a) jako klasyczna miara zmienności liczona jest w oparciu o wszystkie wartości cechy, a warunkiem jej wyznaczenia jest znajomość średniej arytmetycznej w stosunku do której jest obliczana,

b) przyjmuje tylko wartości nieujemne; wartość zerową osiąga w przypadku cechy stałej (wówczas wszystkie wartości cechy są identyczne),

c) jeśli w szeregu rozdzielczym wszystkie wagi - w_i (w szczególnym przypadku n_i lub f_i) pomnożymy bądź podzielimy przez tę samą wielkość q, to wariancja liczona przy tak zmienionym systemie wag będzie identyczna jak wariancja pierwotna (zauważmy, że wariancja „zachowuje się" w tej sytuacji identycznie jak średnia arytmetyczna); własność tę można wyrazić w sposób następujący:

$$\frac{\sum_{i=1}^{k}(x_i-\bar{x})^2\cdot(w_i\cdot q)}{\sum_{i=1}^{k}(w_i\cdot q)}=\frac{\sum_{i=1}^{k}(x_i-\bar{x})^2\cdot w_i}{\sum_{i=1}^{k}w_i}$$

gdzie: w_i – wagi przypisane poszczególnym wartościom cechy (w szczególnym przypadku będą to liczebności n_i lub częstości f_i,

q – wielkość stała,

d) jeśli wszystkie wartości cechy pomnożymy bądź podzielimy przez tę samą wielkość q, to wariancja tak zmienionych wartości cechy będzie q^2 razy większa w przypadku mnożenia lub q^2 razy mniejsza w przypadku dzielenia od wariancji pierwotnych wartości cechy; własność tę wyrażają poniższe równości:

$$s^2(x\cdot q)=s^2(x)\cdot q^2 \qquad\qquad s^2\left(\frac{x}{q}\right)=\frac{s^2(x)}{q^2}$$

e) jeśli do wszystkich wartości cechy dodamy lub od wszystkich wartości cechy odejmiemy tę samą wielkość q, to wariancja tak zmienionych wartości cechy będzie identyczna jak wariancja pierwotnych wartości cechy; wyraża to poniższy zapis:

$$s^2(x+q)=s^2(x) \qquad\qquad s^2(x-q)=s^2(x)$$

f) jeśli zbiorowość podzielimy na dowolną liczbę podzbiorowości cząstkowych (grup), to wariancja ogólna badanej cechy będzie sumą wariancji średnich grupowych i średniej wariancji wewnątrzgrupowych, co można wyrazić wzorem:

$$s^2(x) = \bar{s}_j{}^2(x) + s^2(\bar{x}_j),$$

gdzie: $s^2(x)$ – wariancja ogólna cechy X,

$\bar{s}^2{}_j(x)$ – średnia wariancji wewnątrzgrupowych, *(j=1,2,.., k; k = liczba wyodrębnionych grup)*

$s^2(\bar{x}_j)$ – wariancja średnich grupowych.

Na podstawie podanej formuły można stwierdzić, że na ogólne zróżnicowanie wartości cechy wpływa łącznie wewnątrzgrupowe zróżnicowanie tej cechy, jak również zróżnicowanie międzygrupowe. Podana relacja jest określana mianem *równości wariancyjnej*

g) wariancja stanowi różnicę między średnią arytmetyczną kwadratów wartości cechy a kwadratem średniej arytmetycznej wartości tej cechy:

$$s^2(x) = \overline{(x^2)} - (\bar{x})^2$$

Jak już wspomniano w badaniach zmienności wykorzystuje się pierwiastek z wariancji zwany odchyleniem standardowym. Z uwagi na tę prostą zależność procedury obliczania obu parametrów zostaną zilustrowane łącznie.

Odchylenie standardowe

Stanowi ono najczęściej wykorzystywaną miarę zmienności. Procedurę obliczania odchylenia standardowego oznaczanego zwykle jako $s(x)$ można wyrazić ogólnym wzorem:

$$s(x) = \sqrt{s^2(x)} \qquad 2.20.$$

W zależności od postaci materiału statystycznego odpowiednie wzory umożliwiające w takich przypadkach wyznaczenie odchylenia standardowego przyjmą postać:

a) *dla szeregu szczegółowego*

$$s(x) = \sqrt{\frac{\sum_{i=1}^{N}(x_i - \bar{x})^2}{N}} \qquad 2.21.$$

b) *dla szeregu rozdzielczego punktowego*

$$s(x) = \sqrt{\frac{\sum_{i=1}^{k} (x_i - \bar{x})^2 \cdot n_i}{N}} \qquad 2.22.$$

c) *dla szeregu rozdzielczego z przedziałami klasowymi*

$$s(x) = \sqrt{\frac{\sum_{i=1}^{k} (\dot{x}_i - \bar{x})^2 \cdot n_i}{N}} \qquad 2.23.$$

Występujące we *wzorach 2.21 i 2.23* liczebności mogą być zastąpione częstościami. Ustalona wartość odchylenia standardowego informuje, o ile przeciętnie różnią się wartości cechy występujące u poszczególnych jednostek statystycznych od ich wielkości średniej.
Podstawowe własności odchylenia standardowego wynikają z odpowiednich własności wariancji; niektóre stanowią ich niewielką modyfikację.

Ustalenie wartości średniej arytmetycznej i odchylenia standardowego pozwala na określenie typowego obszaru zmienności dla badanej cechy. Obszar ten zawiera się w przedziale: $\bar{x} \pm s(x)$. W obszarze tym zawiera się około 2/3 wszystkich obserwacji. Jednostki, którym odpowiadają te obserwacje uważa się za typowe dla badanej zbiorowości.

Procedurę wyznaczania wariancji i odchylenia standardowego ilustrują poniższe przykłady.

Przykład 2.9.

Zebrano informacje dotyczące wieku wylosowanej grupy przestępców odbywających wyroki w Zakładzie Karnym. Uzyskano następujące dane (w latach): *22; 45, 32; 48; 35; 28; 21; 29; 33; 26; 24; 30; 41; 44; 22*. Określić zmienność wieku przestępców za pomocą odchylenia standardowego.

Rozwiązanie

Odchylenie standardowe dla informacji ujętych w szeregu szczegółowym ustalone zostanie według *formuły 2.21*. Wymaga ona wcześniejszego ustalenia wieku średniego. Wynosi on 32 lata. W dalszej kolejności ustalane są kwadraty odchyleń wieku poszczególnych przestępców od wieku średniego. Podstawiając sumę ustalonych kwadratów odchyleń do wzoru otrzymujemy:

$$s(x) = \sqrt{\frac{1110}{15}} = 8{,}6 \text{ lat.}$$

Otrzymany wynik oznacza, że wiek każdego z przestępców różni się średnio o 8,6 lat od wieku średniego.

Przykład 2.10.

Poniższa tablica zawiera informacje dotyczące liczby błędów popełnianych przez osoby zdające test teoretyczny w ramach egzaminu na prawo jazdy w pewnym Ośrodku Egzaminacyjnym.

Liczba popełnionych błędów (x_i)	*Liczba zdających* (n_i)
0	5
1	15
2	20
3	15
4	5
5	3
6	2
Razem	65

Zbadać zmienność liczby popełnionych błędów przy pomocy odchylenia standardowego.

Rozwiązanie

Dla wyznaczenia odchylenia standardowego konieczne jest ustalenie średniej arytmetycznej liczby popełnionych błędów. Obliczenia pomocnicze zawiera poniższa tablica robocza.

Liczba popełnionych błędów (x_i)	*Liczba zdających* (n_i)	$x_i \cdot n_i$	$(x_i - \bar{x})^2$	$(x_i - \bar{x})^2 \cdot n_i$
1	*2*	*3*	*4*	*5*
0	5	0	5,11	25,55
1	15	15	1,59	23,85
2	20	40	0,07	1,40
3	15	45	0,55	8,25
4	5	20	3,03	15,15
5	3	15	7,51	22,53
6	2	12	13,99	27,98
Razem	65	147	X	124,71

Na podstawie obliczeń zawartych w kolumnie 3 otrzymujemy:

$$\bar{x} = \frac{147}{65} = 2{,}26$$

W kolejnych kolumnach wykonano obliczenia pomocnicze dla ustalenia odchylenia standardowego. Po podstawieniu uzyskanych wielkości do *wzoru 2.22* otrzymujemy:

$$s(x) = \sqrt{\frac{124,71}{65}} = 1,4 \text{ błędu}$$

Uzyskana wartość miary oznacza, iż przeciętnie liczba błędów popełnionych przez każdego ze zdających różniła się o około 1,4 od średniej liczby tych błędów.

Przykład 2.11.

Dla 120 uczniów jednej ze szkół podstawowych przeprowadzono badanie warunków środowiska domowego. Uzyskane wyniki badania ujęto w poniższej tablicy:

Wyniki pomiaru w punktach (x_i)	*Liczba uczniów* (n_i)
11 – 25	10
26 – 40	35
41 – 55	55
56 - 70	20
Razem	120

Źródło: Badanie własne

Na podstawie podanych informacji zbadać zmienność uzyskanych wyników pomiaru za pomocą odchylenia standardowego.

Rozwiązanie

Celem ustalenia wartości tej miary należy w pierwszej kolejności wyznaczyć średnią arytmetyczną wyników pomiaru. Obliczenia pomocnicze wykonano w kolumnie 4 poniższej tablicy roboczej.

Wyniki pomiaru (w punktach) (x_i)	*Liczba uczniów* (n_i)	*Środki przedziałów* ($\dot{x}_i$)	$x_i \cdot n_i$	$(\dot{x}_i - \bar{x})^2$	$(\dot{x}_i - \bar{x})^2 \cdot n_i$
1	*2*	*3*	*4*	*5*	*6*
11 – 25	10	18	180	655,4	6554
26 – 40	35	33	1155	112,4	3934
41 – 55	55	48	2640	19,4	1067
56 - 70	20	63	1260	376,4	7528
Razem	120	X	5235	X	19083

Średnia ta wynosi: $\frac{5235}{120} = 43,6$ punktu. Kolejne obliczenia pomocnicze dla wyznaczenia odchylenia standardowego - zgodnie z *wzorem 2.23* - wy-

konano w kolumnach 5. i 6. tablicy. Po podstawieniu uzyskanych wyników obliczeń do wzoru otrzymujemy:

$$s(x) = \sqrt{\frac{19083}{120}} = 12{,}6 \text{ punktu}$$

Interpretacja: Przeciętnie wyniki pomiaru środowiska domowego każdego z uczniów różnią się o około 12,6 punktu od pomiaru średniego.

Odchylenie przeciętne

Odchylenie przeciętne, oznaczane dalej jako $d(x)$, definiowane jest jako średnia bezwzględnych odchyleń poszczególnych wartości cechy od ich średniej arytmetycznej. W zależności od typu szeregu, w którym ujęty został materiał statystyczny będzie ono wyznaczane według poniższych wzorów:

a) *dla szeregu szczegółowego*:

$$d(x) = \frac{\sum_{i=1}^{N} |x_i - \bar{x}|}{N} \qquad 2.24.$$

b) *dla szeregu rozdzielczego punktowego*:

$$d(x) = \frac{\sum_{i=1}^{k} |x_i - \bar{x}| \cdot n_i}{N} \qquad 2.25.$$

c) *dla szeregu z przedziałami klasowymi*:

$$d(x) = \frac{\sum_{i=1}^{k} |\dot{x}_i - \bar{x}| \cdot n_i}{N} \qquad 2.26.$$

Również w tym przypadku występujące we *wzorach 2.2 i 2.26* liczebności mogą być zastąpione częstościami.

Interpretacja odchylenia przeciętnego jest identyczna jak odchylenia standardowego, co jednak nie oznacza, że wartości obu miar są identyczne. Odchylenie standardowe przyjmuje wartości nie mniejsze od odchylenia przeciętnego, co można ująć następującą relacją:

$$s(x) \geq d(x)$$

Omawiana miara zmienności odgrywa niewielką rolę w teorii statystyki, dlatego też jest zdecydowanie rzadziej wykorzystywana do pomiaru rozproszenia wartości cechy.

Współczynnik zmienności

Parametr ten należy do grupy względnych miar zmienności i jak wcześniej sygnalizowano jest on ustalany jako iloraz absolutnej miary rozproszenia i wybranej miary średniej. W zależności od wykorzystanych absolutnych miar zmienności i poziomu średniego występują różne postacie współczynnika zmienności. Najczęściej wykorzystywana jest postać wyrażająca się wzorem:

$$W_z(x) = \frac{s(x)}{\bar{x}} \qquad 2.27.$$

Dla celów praktycznych wyraża się go często w procentach mnożąc otrzymany wynik przez 100. Mierzy on, jaką część wartości średniej badanej cechy stanowi jej odchylenie standardowe, a więc ta postać oparta jest na parametrach klasycznych. Im wyższa wartość tej miary tym większa zmienność badanej cechy. Jest oczywiste, że przyjmuje on wartości nieujemne.

Podana postać współczynnika zmienności może być ustalona jedynie wówczas, gdy możliwe jest ustalenie parametrów tworzących tę miarę. Gdy jest to niemożliwe, mogą być wykorzystane współczynniki zmienności oparte na miarach pozycyjnych. Należy do nich m. in. kwartylowy współczynnik zmienności wyrażony poniższymi wzorami:

$$W_z(x) = \frac{O_c(x)}{Me(x)} \qquad 2.28.$$

lub

$$W_z(x) = \frac{Q_3(x) - Q_1(x)}{Q_3(x) + Q_1(x)} \qquad 2.29.$$

Współczynnik zmienności jest miarą niemianowaną i ten brak miana umożliwia porównywanie stopnia zróżnicowania różnych cech statystycznych (wyrażanych w różnych jednostkach miary). Takich porównań nie można w zasadzie dokonywać przy wykorzystaniu absolutnych miar rozproszenia.

Wykorzystując wyniki badania zmienności uzyskane w *przykładach 2.9 - 2.11* określimy względną zmienność badanych cech. Dla wszystkich trzech przypadków możliwe jest wykorzystanie w tym celu współczynnika zmienności o postaci *2.27*. Dla *przykładu 2.9* przyjmie on wartość:

$$W_z(x) = \frac{8,6}{32} = 0,27$$

Wynik ten oznacza, że odchylenie standardowe wieku przestępców stanowi 27% ich wieku średniego.

W *przykładzie 2.10* miernik ten przyjmie postać:

$$W_z(x) = \frac{1,4}{2,26} = 0,62$$

Uzyskana wartość współczynnika zmienności oznacza, że odchylenie standardowe liczby popełnionych błędów stanowi 62% średniej ich liczby.

Dla *przykładu 2.11* współczynnik zmienności osiągnie wielkość:

$$W_z(x) = \frac{12,6}{43,6} = 0,29$$

Otrzymana wielkość oznacza, że odchylenie standardowe wyników pomiaru środowiska domowego stanowi 29% ich pomiaru średniego.

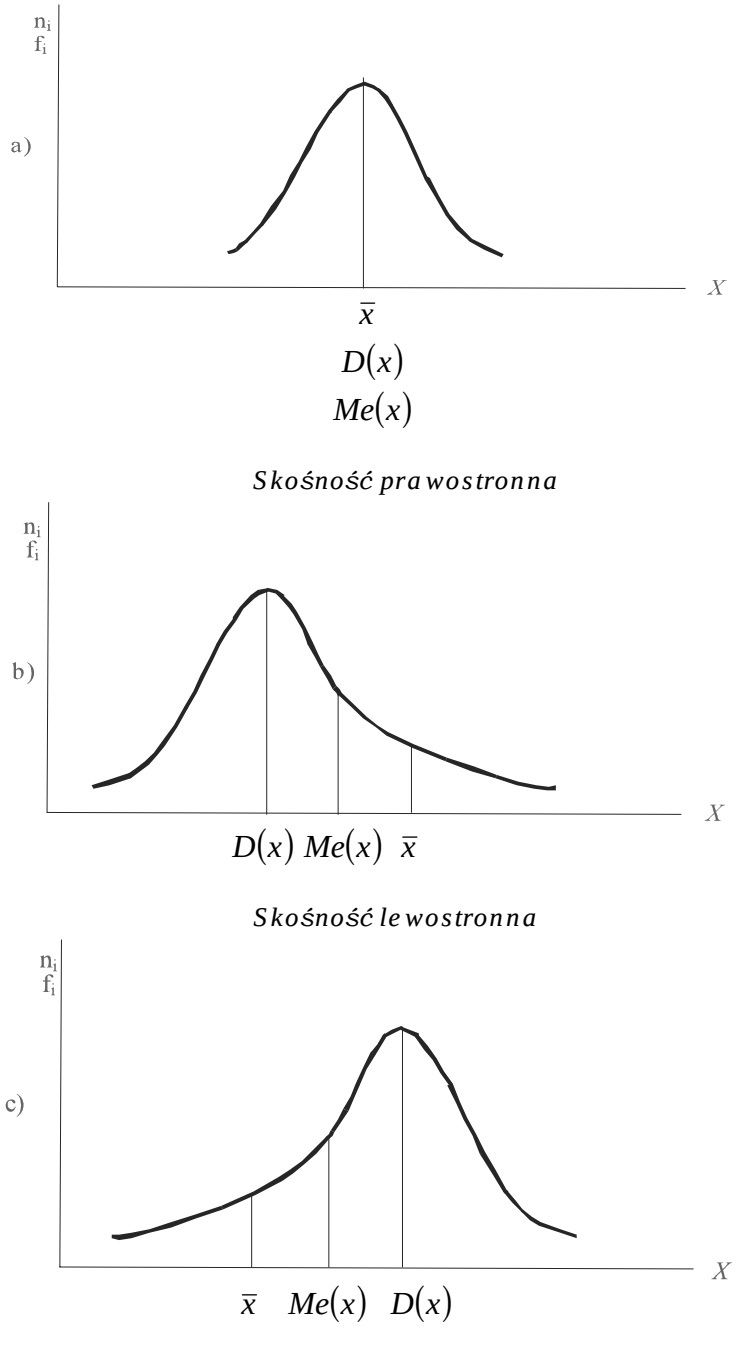
Jak wcześniej podkreślono, współczynnik zmienności jako miara niemianowana umożliwia porównanie rozproszenia różnych cech. W świetle tego można stwierdzić, że w badanych przypadkach zdecydowanie najwyższe zróżnicowanie wartości cechy występuje w *przykładzie 2* („liczba błędów popełnianych w trakcie zdawania testu teoretycznego"), natomiast w pozostałych przykładach poziom zmienności jest zbliżony.

2.3.4. Miary skośności

Zjawisko *skośności* (w literaturze używa się zamiennie pojęcia *asymetrii*) określane jest jako brak symetrii w rozłożeniu wartości cechy względem ich średniej arytmetycznej, co jest równoznaczne z niesymetrycznym rozłożeniem jednostek statystycznych. Jeśli w zbiorowości statystycznej opisywanej ze względu na określoną cechę przeważają liczebnie jednostki o wartościach cechy niższych od średniej arytmetycznej, to sytuację tę określa się mianem skośności prawostronnej, zaś w przypadku dominacji jednostek o wartościach cechy wyższych od średniej arytmetycznej - mówimy o skośności lewostronnej. Przypadki te oraz kształt rozkładu symetrycznego zilustrowano na rys. 2.12, na którym zaznaczono również orientacyjne położenie podstawowych parametrów średnich

Rys. 2.12

Graficzna ilustracja skośności



Źródło: Opracowanie własne

Badanie skośności oznacza określenie jej rodzaju i poziomu odchylenia rozkładu badanej cechy od rozkładu symetrycznego. Do pomiaru tego zjawiska wykorzystywane są miary określane mianem *miar skośności* bądź *asymetrii*, wśród których można wyróżnić *miary pozycyjne oraz klasyczne*.

Do *pozycyjnych* zalicza się mierniki oparte na badaniu relacji między miarami przeciętnymi, w tym pozycyjnymi. Zalicza się do nich następujące miary:

a) różnicę między średnią arytmetyczną i dominantą bądź medianą,
b) kwartylowy współczynnik skośności,
c) standaryzowany współczynnik skośności.

Jako *klasyczną miarę skośności* wykorzystuje się różne postacie *momentu centralnego trzeciego rzędu*.

Różnica między wartością średnią i dominantą

Dla stwierdzenia faktu występowania zjawiska asymetrii bądź jej braku i ewentualnego określenia jej kierunku można wykorzystać relację zachodzącą między średnią arytmetyczną i dominantą. I tak:

- jeśli $\bar{x} = D(x)$, czyli $\bar{x} - D(x) = 0$ zjawisko skośności nie występuje; rozkład jest symetryczny, tzn. występuje jednakowa liczba jednostek statystycznych poniżej i powyżej średniej arytmetycznej (por. rys. 2.4),
- jeśli $\bar{x} > D(x)$, czyli $\bar{x} - D(x) > 0$ rozkład jest asymetryczny; występuje skośność prawostronna, tzn. w zbiorowości dominują jednostki o wartościach cechy niższych od średniej (por. rys. 2.4),
- jeśli $\bar{x} < D(x)$, czyli $\bar{x} - D(x) < 0$ rozkład jest asymetryczny; występuje skośność lewostronna, tzn. w zbiorowości dominują jednostki o wartościach cechy wyższych od średniej (por. rys. 2.4).

W oparciu o powyższe relacje można zaproponować najprostszą miarę skośności wyrażającą się wzorem:

$$M_s(x) = \bar{x} - D(x) \qquad 2.30.$$

Określona w powyższy sposób miara skośności jest wielkością mianowaną i nienormowaną, co nie pozwala na ustalenie natężenia skośności. W literaturze podaje się również zmodyfikowaną postać podanej wyżej miary (traktowaną jako jej równoważną) o postaci:

$$M_s(x) = 3 \cdot [\bar{x} - Me(x)] \qquad 2.31.$$

Według Pearsona formuła ta pozwala uzyskiwać wyniki zbliżone do osiąganych według *wzoru 2.30*, ale jedynie w przypadku rozkładów umiarkowanie skośnych.

Współczynniki skośności

Badania porównawcze skośności różnych cech wymagają stosowania względnych (niemianowanych) mierników. Należą do nich współczynniki skośności. Najczęściej wykorzystywaną formułę można wyrazić następująco:

$$W_{s1}(x) = \frac{\bar{x} - D(x)}{s(x)} \qquad 2.32.$$

lub

$$W_{s1}(x) = \frac{3 \cdot [\bar{x} - Me(x)]}{s(x)} \qquad 2.33.$$

W przypadku braku możliwości wyznaczenia średniej arytmetycznej i w konsekwencji odchylenia standardowego może być wykorzystana podana niżej postać współczynnika skośności oparta ona kwartylach:

$$W_{s2}(x) = \frac{[Q_3(x) - Me(x)] - [Me(x) - Q_1(x)]}{[Q_3(x) - Me(x)] + [Me(x) - Q_1(x)]} = \frac{[Q_3(x) - Me(x)] - [Me(x) - Q_1(x)]}{2 \cdot O_c(x)} \qquad 2.34.$$

Podane współczynniki skośności są wielkościami względnymi i unormowanymi. W większości przypadków przyjmują wartości z przedziału od -1 do +1; przy dużym natężeniu skośności współczynnik skośności o *postaci 2.32* może przyjmować wartości wykraczające poza podany przedział.

Trzeci moment centralny

Kolejna miara skośności oparta jest na wykorzystaniu momentu statystycznego rzędu trzeciego. W literaturze można spotkać różne propozycje wykorzystania tego miernika do pomiaru skośności. I tak dla stwierdzenia jedynie faktu występowania skośności może być wykorzystany trzeci moment centralny o postaci[5]:

$$m_3(x) = \frac{\sum_{i=1}^{k} (x_i - \bar{x})^3 \cdot n_i}{N} \qquad 2.35.$$

Określanie na jego podstawie faktu występowania skośności i jego kierunku odbywa się w sposób następujący:

- jeśli $m_3(x)$ = 0 skośność nie występuje, rozkład jest symetryczny,

[5] Zjawisko skośności badane jest głównie w oparciu o dane ujęte w szeregach rozdzielczych, dlatego też podane w dalszej części formuły odnoszą się do tego typu szeregów. Występujące w podanych wzorach liczebności mogą być zastąpione częstościami

- jeśli $m_3(x) > 0$ rozkład jest asymetryczny, występuje skośność prawostronna,
- jeśli $m_3(x) < 0$ rozkład jest asymetryczny, występuje skośność lewostronna.

W celu dodatkowego określenia natężenia skośności trzeci moment centralny poddajemy standaryzacji, co prezentuje poniższa formuła

$$m_3(t) = \frac{m_3(x)}{s^3(x)} = \frac{\dfrac{\sum_{i=1}^{k}(x_i - \bar{x})^3 \cdot n_i}{N}}{\left(\sqrt{\dfrac{\sum_{i=1}^{k}(x_i - \bar{x})^2 \cdot n_i}{N}}\right)^3} \qquad 2.36.$$

Występujący w mianowniku podanej formuły sześcian odchylenia standardowego umożliwia uzyskanie miary niemianowanej, która może być wykorzystywana do analizy porównawczej skośności rozkładów cech o różnych jednostkach miary. Mankamentem tej miary jest brak ściśle określonych granic przedziału, z którego może ona przyjmować wartości liczbowe. Uniemożliwia to ocenę natężenia skośności. M. Krzysztofiak proponuje w tym celu następującą modyfikację powyższej postaci standaryzowanej:

$$m_3^*(t) = \frac{m_3(x)}{|m_3(x)| + s^3(x)} \qquad 2.37.$$

Zaproponowana modyfikacja pozwala na uzyskanie wartości tej miary z przedziału <-1 ; 1>, co umożliwia określenie kierunku i natężenia skośności, a jej wartości graniczne oznaczają występowanie skrajnej skośności ujemnej bądź dodatniej.

Przeprowadzając badanie porównawcze skośności należy pamiętać, by posługiwać się w jego trakcie tymi samymi miarami, a dokonując wyboru parametru skośności należy preferować te, które są oparte na klasycznej procedurze ustalania ich wartości. Należy również dodać, że wyższy poziom skośności jest zawsze związany z większą zmiennością wartości badanej cechy. Dla zilustrowania procedury badania skośności prześledźmy przykład 2.12.

Przykład 2.12.

Na podstawie informacji ujętych w *przykładzie 2.11* zbadać skośność w rozkładzie wyników pomiaru warunków środowiska domowego dla wyodrębnionej zbiorowości uczniów.

Rozwiązanie

Dla badania skośności wykorzystamy zaprezentowane wyżej miary. Wymagają one przeprowadzenia wielu obliczeń pomocniczych, które zostały ujęte w poniższej tablicy roboczej (pominięte zostaną obliczenia wykonane w *przykładzie 2.11*):

Wyniki pomiaru (w punktach) (x_i)	*Liczba uczniów* (n_i)	*Środki przedziałów* $(\dot{x}_i)$	*cum* n_i	$(\dot{x}_i - \bar{x})^3$	$(\dot{x}_i - \bar{x})^3 \cdot n_i$
1	*2*	*3*	*4*	*5*	*6*
11 – 25	10	18	180	-16777,2	-167772,0
26 – 40	35	33	1155	-1201,6	-42056,6
41 – 55	55	48	2640	80,8	4444,0
56 - 70	20	63	1260	7320,8	146416,0
Razem	120	X	5235	X	-59068,6

Przypomnijmy, że uzyskana wartość średniej arytmetycznej wyniosła 43,6 punktu, zaś odchylenie standardowe – 12,6 punktu.

Dla potrzeb określenia pierwszej z miar skośności - $M_s(x)$ należy dodatkowo wyznaczyć wartości dominanty i mediany. Dominanta wyniesie:

$$D(x) = 41 + \frac{(55-35)}{(55-35)+(55-20)} \cdot 14 = 46{,}1 \text{ punktu}$$

zaś mediana:

$$Me(x) = 41 + \frac{\frac{120}{2} - 45}{55} \cdot 14 = 44{,}8 \text{ punktu}$$

Miara skośności ustalona według *wzoru 2.30* wyniesie:

$$M_s(x) = 43{,}6 - 46{,}1 = -2{,}5 \text{ punktu}$$

zaś według *wzoru 2.31:*

$$M_s(x) = 3(43{,}6 - 44{,}8) = -3{,}6 \text{ punktu}$$

Na podstawie obu miar można stwierdzić występowanie w badanym szeregu skośności lewostronnej (ujemnej), co oznacza, że w badanej zbiorowości dominują uczniowie, dla których wyniki pomiarów warunków środowiska domowego są wyższe od średnich.

Dla określenia natężenia skośności (jak również jej kierunku) wykorzystamy zarówno współczynniki skośności jak i trzeci moment centralny standaryzowany.

Standaryzowane współczynniki skośności o *postaciach 2.32 i 2.33* przyjmują odpowiednio wartości:

$$W_{s1}(x)=\frac{43{,}6-46{,}1}{12{,}6}=-0{,}20 \quad W_{s1}(x)=\frac{3\cdot(43{,}6-44{,}8)}{12{,}6}=-0{,}29$$

Obie miary wskazują na niewielką skośność lewostronną.

Kwartylowy współczynnik skośności zostanie ustalony zgodnie z *wzorem 2.34*. Wymaga on dodatkowego ustalenia kwartyli pierwszego i trzeciego. Kwartyl pierwszy wynosi:

$$Q_1(x)=26+\frac{\frac{120}{4}-10}{35}\cdot 14=34 \quad \text{punkty}$$

zaś kwartyl trzeci:

$$Q_3(x)=41+\frac{\frac{3\cdot 120}{4}-45}{55}\cdot 14=52{,}5 \text{ punktu}$$

Wobec tego współczynnik skośności przyjmie wartość:

$$W_{s2}(x)=\frac{(52{,}5-44{,}8)-(44{,}8-34{,}0)}{(52{,}5-44{,}8)+(44{,}8-34{,}0)}=\frac{-3{,}1}{18{,}5}=-0{,}17$$

Uzyskany wynik potwierdza występowanie niewielkiej skośności lewostronnej.

Spośród wymienionych różnych postaci trzeciego momentu centralnego przyjmijmy tę ostatnią *(wzór 2.37)*. Dla jej wyznaczenia wykonano obliczenia pomocnicze zawarte w powyższej tabeli roboczej w kolumnach 5. i 6. Podstawiając odpowiednie dane do wzoru uzyskujemy:

$$m_3^{*}(t)=\frac{\frac{-59068{,}6}{120}}{\left|\frac{-59068{,}6}{120}\right|+(12{,}6)^3}=\frac{-492{,}2}{492{,}2+2000{,}4}=-0{,}20$$

Kolejna miara również wskazuje niewielką skośność lewostronną.

2.3.5. Miary koncentracji

Zjawisko koncentracji wartości cechy jest w statystyce dwojako interpretowane. Po pierwsze, może oznaczać skupienie wartości cechy wokół średniej arytmetycznej; po drugie – może być utożsamiane z nierównomiernym rozkładem globalnego funduszu wartości cechy wśród jednostek statystycznych badanej zbiorowości.

Koncentracja jako nierównomierność rozłożenia globalnego funduszu wartości cechy

Globalny fundusz wartości cechy stanowi sumę wartości występujących u wszystkich jednostek statystycznych badanej zbiorowości. W przypadku szeregu rozdzielczego punktowego wyznaczany jest on jako suma iloczynów poszczególnych wariantów wartości cechy i odpowiadających im liczebności ($\sum_{i=1}^{k} x_i \cdot n_i$), zaś w przypadku szeregu rozdzielczego z przedziałami klasowymi jako suma iloczynów środków poszczególnych przedziałów klasowych i odpowiadających im liczebności ($\sum_{i=1}^{k} \dot{x}_i \cdot n_i$). Koncentracja w tym znaczeniu jest powiązana z poziomem zróżnicowania wartości cechy oraz natężeniem skośności; *im wyższa zmienność i skośność tym wyższy poziom nierównomierności rozłożenia wartości cechy*. Skrajne przypadki tak rozumianej koncentracji należy interpretować następująco: koncentracja zupełna ma miejsce wówczas, gdy całym funduszem wartości cechy dysponuje jedna jednostka, zaś całkowity brak koncentracji występuje, gdy każdej jednostce przypada taka sama część tego funduszu.

Badanie nierównomierności odbywa się przez porównanie rozkładu tego funduszu i rozłożenia jednostek statystycznych badanej zbiorowości. Jeśli proporcje rozkładu globalnego funduszu są odmienne niż proporcje rozkładu jednostek statystycznych, to występuje koncentracja w tym znaczeniu. Badanie natężenia tak rozumianej koncentracji może się odbywać graficznie – przy wykorzystaniu krzywej Lorenza lub analitycznie przy wykorzystaniu współczynnika koncentracji o postaci

$$k = \frac{a}{5000} \qquad 2.38.$$

gdzie: a – jest powierzchnią zawartą między krzywą Lorenza a linią równomiernego podziału globalnego funduszu wartości cechy.

Współczynnik ten jest wielkością unormowaną w przedziale <0 ; 1>. Im wyższa jest jego wartość tym większa nierównomierność rozłożenia globalnego funduszu wartości cechy, a wielkości skrajne oznaczają:

0 - całkowity brak koncentracji,

1 - koncentrację zupełną (praktycznie nie ma ona miejsca).

Wyznaczając analitycznie współczynnik koncentracji można wykorzystać informacje ujęte na wykresie krzywej (w praktyce nie jest to krzywa, lecz linia łamana) Lorenza. Sposób postępowania w przypadku graficznego i analitycznego badania koncentracji ilustruje poniższy przykład.

Przykład 2.13.

Poniższa tablica zawiera informacje dotyczące struktury przedsiębiorstw woj. „D" według wielkości zatrudnienia:

Zatrudnienie (x_i)	*Liczba przedsiębiorstw* (n_i)	*Środki przedziałów* $(\dot{x}_i)$	$\dot{x}_i \cdot n_i$	$f_i\left(\sum_{i=1}^{k} \dot{x}_i \cdot n_i\right)$	$f_i(n_i)$	$cumf_i\left(\sum_{i=1}^{k} \dot{x}_i \cdot n_i\right)$	$cumf_i(n_i)$
1	*2*	*3*	*4*	*5*	*6*	*7*	*8*
1-10	450	5	2250	14,00	75,00	14,00	75,00
11-50	100	30	3000	18,66	16,67	32,66	91,67
51-200	35	125	4375	27,22	5,83	59,88	97,50
201-500	12	350	4200	26,12	2,00	86,00	99,50
501-1000	3	750	2250	14,00	0,50	100,00	100,00
Razem	600	X	16075	100,00	10,00	X	X

Źródło: Dane umowne

Na podstawie powyższych danych zbadać poziom nierównomierności rozłożenia zatrudnienia w przedsiębiorstwach woj. „D".

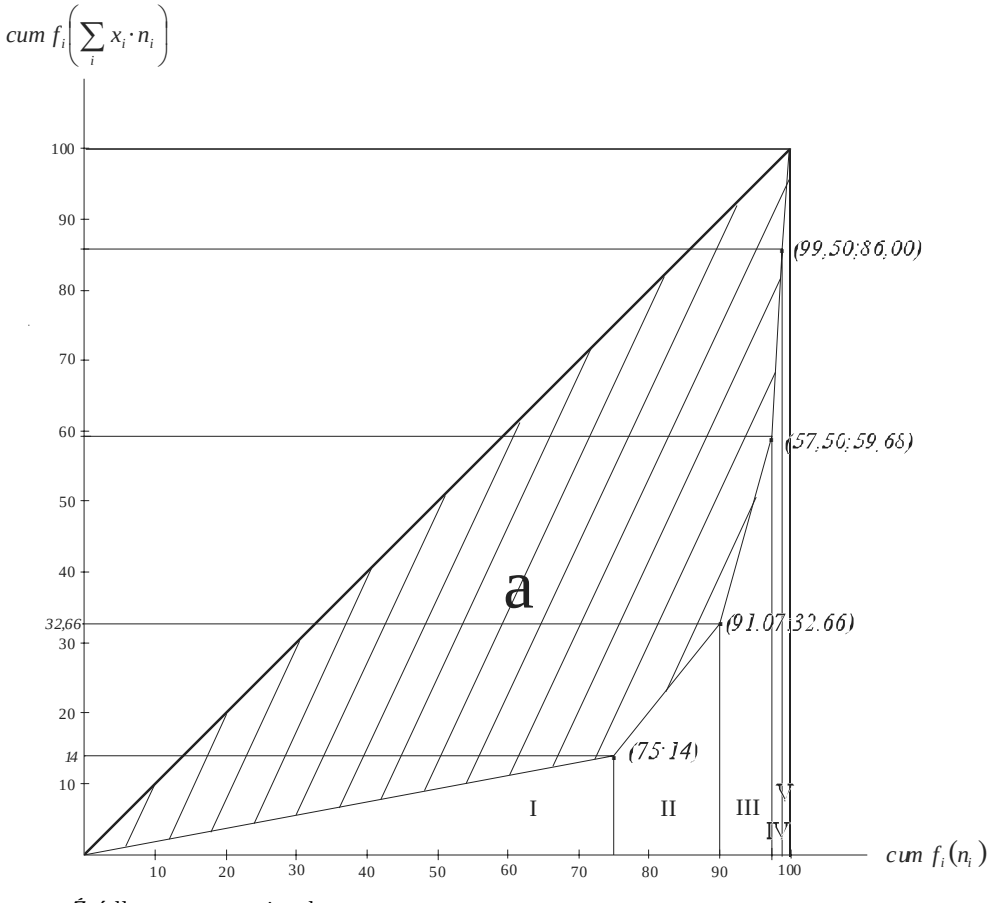
Rozwiązanie.

Globalny fundusz wartości cechy w podanym przykładzie stanowi łączna wielkość zatrudnienia w badanych 600 przedsiębiorstwach. Stanowi ona sumę iloczynów zawartych w kolumnie 4. i wynosi 16075 (taka jest przybliżona łączna wielkość zatrudnienia w badanych przedsiębiorstwach). W kolumnie 5. zawarto częstość występowania globalnego funduszu wartości cechy $f_i\left(\sum_{i=1}^{k} \dot{x}_i \cdot n_i\right)$, której kolejne wartości oznaczają udział zatrudnienia poszczególnych kategorii przedsiębiorstw (według wielkości zatrudnienia) w zatrudnieniu łącznym. W kolejnej kolumnie 5. umieszczono częstość występowania przedsiębiorstw według liczby zatrudnionych w łącznej ich liczbie - $f_i(n_i)$. W kolumnach 7 i 8 dokonano kumulacji częstości zawartych w kolumnach 5 i 6. Analiza informacji zawartych w tych kolumnach pozwala na wstępną ocenę koncentracji wartości cechy; w przypadku, gdyby częstości skumulowane w poszczególnych parach z obu kolumn były identyczne, oznaczałoby to występowanie równomiernego rozłożenia globalnego funduszu wartości cechy. W analizowanym przykładzie częstości te w poszczególnych parach są zróżnicowane; pozwala to na stwierdzenie faktu występowania koncentracji (nierównomierności rozłożenia). Jej poziom zostanie wyznaczony graficznie (por. rys. 2.13) oraz analitycznie.

Wykorzystując informacje zawarte w kolumnach 7. i 8. powyższej tabeli roboczej wykreślamy w układzie współrzędnych krzywą Lorenza, przy czym na osi odciętych nanosimy skumulowane częstości zawarte w kolumnie 8, zaś na osi rzędnych częstości z kolumny 6. Jej przebieg wyznaczają punkty, których współrzędnymi są poszczególne pary częstości skumulowanych (*np. 75,00 i 14,00; 91,67 i 32,66; itd*). W ten sposób pomiędzy krzywą Lorenza, a przekątną naniesionego w układzie współrzędnych kwadratu o boku 100 (określaną mianem linii równomiernego podziału globalnego funduszu wartości cechy) powstało pole „a", które jest graficznym obrazem koncentracji (nierównomierności). Oceniając wzrokowo udział tego pola w powierzchni trójkąta zawartego pod wykreśloną przekątną dokonujemy szacunku poziomu koncentracji.

Rys. 2.13

Krzywa Lorenza



Źródło: opracowanie własne

Pomiar natężenia koncentracji metodą analityczną odbywa się najczęściej przy wykorzystaniu powyższego wykresu, na podstawie którego możliwe jest precyzyjne określenie powierzchni powstałego pola. Ponieważ jego kształt nie stanowi regularnej figury geometrycznej, wobec tego zaleca się ustalenie wielkości pola znajdującego się pod krzywą Lorenza, tj. dopełnienia pola „a". Dopełnienie pola „a" składa się w naszym przykładzie z 5 figur geometrycznych (ponumerowanych I - V) powstałych w wyniku „zrzutowania" poszczególnych punktów wyznaczających przebieg krzywej Lorenza na oś odciętych (uzyskaliśmy 1 trójkąt i 4 trapezy). Ich pola wyznaczamy posługując się współrzędnymi poszczególnych punktów. I tak:

pole trójkąta (figura I) = 75 · 14/2 = 525
pole trapezu (figura II) = (32,66 + 14) /2 · 16,67 = 388,91
pole trapezu (figura III) = (59,88 + 32,66)/2 · 5,83 = 269,75
pole trapezu (figura IV) = (86 + 59,88)/2 · 2,00 = 145,88
pole trapezu (figura V) = (100 + 86)/2 · 0,50 = 46,50

Dopełnienie pola „a" jest sumą powierzchni powyższych figur i wynosi 1376,04.

Odejmując tę powierzchnię od 5000 (taka jest powierzchnia trójkąta znajdującego się pod linią równomiernego podziału) otrzymujemy pole „a", czyli 5000 - 1376,04 = 3623,96

Tę wielkość podstawiamy do *wzoru 2.38* i otrzymujemy:

$$k = \frac{3623{,}96}{5000} = 0{,}725$$

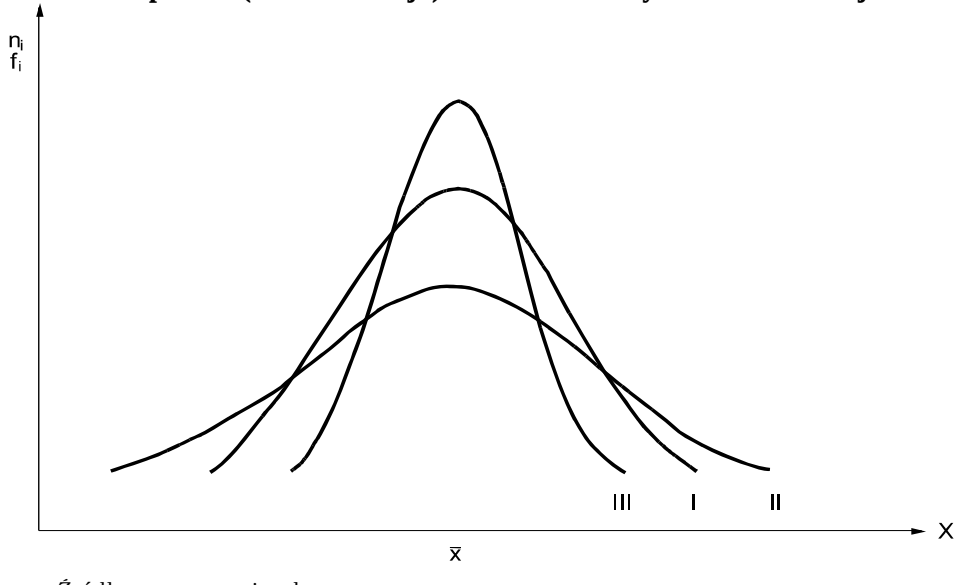
Uzyskana wielkość wskazuje na występowanie dużej nierównomierności rozłożenia (tj. koncentracji) łącznej liczby zatrudnionych w badanym zbiorze przedsiębiorstw.

Koncentracja jako skupienie wartości cechy wokół średniej

Dokonując prezentacji graficznej rozkładu cech statystycznych za pomocą diagramu można zauważyć, iż wykreślona krzywa (w praktyce najczęściej jest to linia łamana) jednomodalna może mieć kształt mniej lub bardziej spłaszczony i odpowiednio mniej lub bardziej wydłużony. Jest to związane z poziomem skupienia wartości cechy wokół ich średniej. Powyższą sytuację ilustruje rys. 2.14.

Rys. 2.14

Skupienie (koncentracja) wartości cechy wokół średniej



Źródło: opracowanie własne

Zakładając, że we wszystkich przypadkach występuje identyczna wartość średnia zaprezentowane krzywe I - III prezentują różny poziom skupie-

nia wartości cechy wokół tej średniej. W przypadku krzywej I można mówić o występowaniu normalnego skupienia (normalnej koncentracji) wokół średniej. Dwie pozostałe krzywe prezentują odpowiednio: II - poziom skupienia mniejszy od normalnego (koncentrację mniejszą od normalnej); krzywa ta ma kształt spłaszczony, platykurtyczny, III - poziom skupienia większy od normalnego (koncentrację większą od normalnej); krzywa ma kształt wysmukły, leptokurtyczny.

Pomiaru tak rozumianej koncentracji dokonujemy przy pomocy *miary kurtozy* (można również spotkać określenie miary pułapu) ustalanej według wzoru:

$$m_4(t) = \frac{m_4(x)}{s^4(x)} = \frac{\dfrac{\sum_{i=1}^{k}(x_i - \bar{x})^4 \cdot n_i}{N}}{\left\{\sqrt{\dfrac{\sum_{i=1}^{k}(x_i - \bar{x})^2 \cdot n_i}{N}}\right\}^4} \qquad 2.39.$$

Jak łatwo zauważyć w liczniku podanej miary występuje moment centralny rzędu czwartego, zaś wykorzystanie w mianowniku odchylenia standardowego podniesionego do potęgi czwartej ma na celu uzyskanie miary niemianowanej. Im wyższa wartość tej miary tym większy poziom skupienia wartości cechy wokół średniej arytmetycznej (wiąże się to ze spadkiem rozproszenia wartości cechy, jednak wartość mianownika maleje szybciej niż wartość czwartego momentu centralnego). W przypadku normalnego skupienia wartości cechy wartość tej miary wynosi 3, w przypadku wartości większych od 3 występuje koncentracja większa od normalnej, zaś przy wartościach mniejszych od 3 - koncentracja mniejsza od normalnej.

W literaturze spotyka się również modyfikację tak sformułowanej miary koncentracji. Nosi ona nazwę *miary ekscesu*. Mierzy ona poziom odchylenia badanego przypadku koncentracji od koncentracji normalnej i wyraża się ona wzorem:

$$e(t) = m_4(t) - 3 \qquad 2.40.$$

Gdy jej wartość wynosi zero - występuje koncentracja normalna; wartości dodatnie wskazują na występowanie koncentracji większej od normalnej, zaś mniejsze od zera - koncentrację mniejszą od normalnej.

W literaturze zaleca się badanie tak rozumianej koncentracji jedynie w przypadku rozkładów symetrycznych lub nieznacznie skośnych. W przypadku dużej skośności należy badać poziom nierównomierności rozłożenia globalnego funduszu wartości cechy

Procedurę wyznaczania miary kurtozy ilustruje poniższy przykład.

Przykład 2.14.

Na podstawie danych z *przykładu 2.11* zbadać poziom skupienia (koncentracji) wyników pomiaru warunków środowiska domowego wokół ich wartości średniej.

Rozwiązanie:

Przypomnijmy, że ustalona wartość średnia wynosi 43,6 punktu, a odchylenie standardowe 12,6 punktu. Obliczenia pomocnicze dla wyznaczenia czwartego momentu centralnego wykonano w poniższej tablicy roboczej:

Wyniki pomiaru (w punktach) (x_i)	*Liczba uczniów* (n_i)	*Środki przedziałów* $(\dot{x}_i)$	$(\dot{x}_i - \bar{x})^4$	$(\dot{x}_i - \bar{x})^4 \cdot n_i$
1	*2*	*3*	*4*	*5*
11 – 25	10	18	429496,7	4294967,0
26 – 40	35	33	12624,8	441868,0
41 – 55	55	48	374,8	20614,0
56 - 70	20	63	141646,8	2832936,0
Razem	120	X	X	7590385,0

Podstawiając sumę iloczynów zawartych w kolumnie 5. oraz odchylenie standardowe do *wzoru 2.39* otrzymujemy:

$$m_4(t) = \frac{\frac{7590395}{120}}{(12{,}6)^4} = \frac{63253{,}2}{25204{,}7} = 2{,}51,$$

a ustalona na tej podstawie miara ekscesu:

$$e(t) = 2{,}51 - 3 = -0{,}49$$

Zarówno czwarty moment centralny standaryzowany jak i miara ekscesu wskazują na koncentrację nieco niższą od normalnej (rozkład ma kształt nieznacznie spłaszczony). Inaczej ujmując: skupienie wyników pomiaru wokół ich średniej jest nieznacznie mniejsze od normalnego.

Rozdział III

Analiza współzależności zmiennych

3.1. WPROWADZENIE

Zagadnienie to stanowi segment wielowymiarowej analizy statystycznej, która może być rozpatrywana w dwóch ujęciach. *Po pierwsze*, dotyczyć może opisu zbiorowości statystycznej, w którym każda jednostka statystyczna jest charakteryzowana przez dowolnie liczny (liczący co najmniej 2) zbiór cech statystycznych. Oznacza to, iż mamy w tym przypadku do czynienia z wielowymiarowym opisem rozkładu określonej zbiorowości. *Po drugie* - rozpatrywać można również inną sytuację, gdy tą samą cechą opisujemy kilka zbiorowości statystycznych.

Analizując pierwszy przypadek (w praktyce najczęściej rozpatrywany) w zbiorze cech przyjętych do opisu badanej zbiorowości – na podstawie analizy merytorycznej - wyodrębniamy takie, które występują we wzajemnym powiązaniu i celem prowadzonego w takim przypadku badania współzależności może być udzielenie odpowiedzi na pytanie; *„Czy pomiędzy wybranymi cechami opisującymi badaną zbiorowość występuje zależność?”* W drugiej sytuacji interesuje nas odpowiedź na pytanie: *„Czy pomiędzy wartościami wybranej cechy opisującymi różne zbiorowości występuje zależność?”*. Ilustracją pierwszego problemu może być próba ustalenia: *Czy istnieje zależność pomiędzy czasem poświęcanym na naukę określonego przedmiotu przez badaną grupę studentów a uzyskiwanymi przez nich wynikami na egzaminie?* W drugim przypadku może to być próba udzielenia odpowiedzi na pytanie: *Czy istnieje zależność między poziomem wykształcenia rodziców a ich dzieci na podstawie zebranego materiału statystycznego o wykształceniu z jednej strony populacji rodziców a z drugiej - ich dzieci?* Należy podkreślić, że badanie współzależności dotyczy głównie pierwszej z omawianych sytuacji.

Prowadzenie analizy współzależności wymaga wyodrębnienia w zbiorze cech opisujących badaną zbiorowość takich cech, co do których zachodzi „podejrzenie”, że występuje pomiędzy nimi związek przyczynowo-skutkowy. Z powyższego wynika, że właściwe badanie współzależności musi być po-

przedzone analizą merytoryczną (określaną również mianem analizy jakościowej) badanego związku w celu uniknięcia badania tzw. *współzależności pozornej*. Współzależność taka ma miejsce wówczas, gdy występuje zależność między wartościami badanych cech, ale nie ma między nimi powiązań przyczynowo-skutkowych, np. *„zależność" między liczbą przebywających na określonym obszarze bocianów a liczbą urodzeń dzieci*. Stwierdzenie wspomnianych powiązań przyczynowych pomiędzy cechami upoważnia do prowadzenia właściwej analizy współzależności. W analizie takiej wyodrębnia się zwykle dwa rodzaje cech: niezależną (jedną bądź kilka) - utożsamianą z przyczyną oraz zależną – stanowiącą skutek oddziaływania wspomnianej przyczyny.

Z teoretycznego punktu widzenia można mówić o dwóch rodzajach zależności pomiędzy cechami: *funkcyjnej i statystycznej*. W pierwszym przypadku mamy do czynienia z jednoznacznym przyporządkowaniem wartościom cechy niezależnej odpowiednich wartości cechy zależnej (każdej wartości zmiennej niezależnej odpowiada tylko jedna wartość zmiennej zależnej). Ten typ zależności nie odnosi się w zasadzie do relacji zachodzących w przypadku zjawisk społeczno-gospodarczych. Wynika to między innymi z następujących przyczyn:

- zjawiska tego typu podlegają zwykle oddziaływaniu bardzo wielu czynników,
- w większości przypadków trudno jednoznacznie zidentyfikować wszystkie czynniki,
- nie wszystkie z ustalonych czynników mają charakter mierzalny by można je było uwzględnić w analizie współzależności,
- wpływ wielu czynników, nawet tych mierzalnych, trudno jednoznacznie określić liczbowo z uwagi na często występujące złożone powiązania z innymi czynnikami,
- uwzględnienie w badaniach zbyt dużej liczby czynników znacznie komplikuje procedury obliczeniowe, a niekiedy wręcz je uniemożliwia.

W takich warunkach badanie współzależności odbywa się zwykle na zasadzie eksperymentowania, bowiem w naukach społeczno-ekonomicznych niemożliwe jest wyizolowanie badanych zjawisk od oddziaływania przyczyn nieistotnych.

W związku z powyższym w przypadku zjawisk społeczno-gospodarczych można mówić jedynie o występowaniu *zależności typu statystycznego*. Jest to zależność niejednoznaczna, tzn. każdej wartości zmiennej niezależnej odpowiadają różne wartości zmiennej zależnej.

Analizując oba rodzaje współzależności należy zauważyć, że o ile w przypadku związku funkcyjnego nie ma uzasadnienia pytanie: czy określona zmienna występuje w silniejszej bądź słabszej zależności funkcyjnej od innej zmiennej, o tyle w przypadku związku statystycznego postawienie podobnego pytania jest sensowne. Można bowiem mówić o mniejszym bądź większym natężeniu zależności typu statystycznego.

W literaturze wyodrębnia się różne podejścia do badania współzależności o zróżnicowanym stopniu precyzji wyników jej badania. Należą do nich:

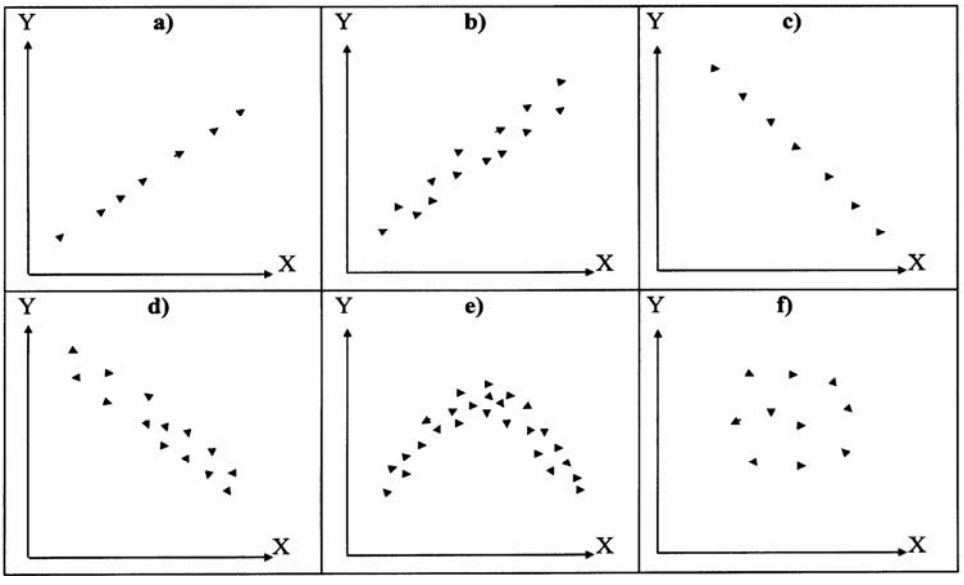
a) metoda graficzna,
b) metoda tabelaryczna,

c) *metody formalne* oparte na wykorzystaniu parametrycznych i nieparametrycznych miar zależności.

Punktem wyjścia dla *metody graficznej* są szeregi szczegółowe informacji o wartościach dwóch wybranych cech Y i X opisujących badaną zbiorowość. Sporządzony na ich podstawie diagram korelacyjny stanowi wykres punktowy umieszczony w układzie współrzędnych prostokątnych, na którym zaznacza się punkty o współrzędnych x_i, y_i (współrzędne te należy traktować jako wartości cech X i Y zaobserwowane w i-tej jednostce). Poniższy rys. 3.1 ilustruje najczęściej spotykane kształty diagramu korelacyjnego.

Rys. 3.1.

Typowe kształty diagramów korelacyjnych



Źródło: opracowanie własne na podstawie: Makać W., Urbanek-Krzysztofiak D.: Metody opisu statystycznego. Wyd. Uniwersytetu Gdańskiego. Gdański 1995, s. 124

Zaprezentowane sytuacje oznaczają:

a) występuje *zależność funkcyjna, dodatnia* (ma miejsce jednoznaczne porządkowanie wzajemne wartości cech X i Y; każdej wartości cechy X odpowiada tylko jedna wartość cechy Y przy czym rosnącym wartościom cechy X odpowiadają rosnące wartości cechy Y lub malejącym wartościom cechy X odpowiadają malejące wartości cechy Y,

b) zależność *prostoliniową o kierunku dodatnim*; w tym przypadku poszczególnym wartościom cechy X może odpowiadać dowolna liczba wartości cechy Y; dodatkowo - wraz ze wzrostem wartości cechy X wzrastają (średnio) wartości cechy Y (lub odwrotnie),

c) ten kształt diagramu ilustruje zależność *funkcyjną o kierunku ujemnym;* w tym przypadku również występuje jednoznaczne, wzajemne przyporządkowanie wartości cech X i Y, przy czym rosnącym wartościom cechy

X odpowiadają malejące wartości cechy Y lub malejącym wartościom cechy X odpowiadają rosnące wartości cechy Y,

d) ten przypadek ilustruje zależność *prostoliniową o kierunku ujemnym*; poszczególnym wartościom cechy X może odpowiadać dowolna liczba wartości cechy Y, przy czym rosnącym wartościom cechy X odpowiadają malejące (średnio) wartości cechy Y (lub odwrotnie),

e) ilustruje jeden z przypadków *związku krzywoliniowego*; w tym przypadku występuje niejednoznaczne przyporządkowanie wartości obu cech, a dodatkowo nie ma miejsca jednolita tendencja zmian wartości tych cech,

f) ten przypadek jest ilustracją *braku zależności* między cechami.

Na podstawie powyższego można stwierdzić, że metoda graficzna oprócz informacji o charakterze związku (liniowy lub krzywoliniowy) i kierunku związku (dodatni lub ujemny) dostarcza również wskazówek umożliwiających wstępną ocenę siły związku miedzy cechami. Może ona stanowić punkt wyjścia dla stosowania formalnych metod opisu współzależności.

Metoda tabelaryczna badania współzależności – wykorzystująca informacje ujęte zarówno w szeregach jak i tablicach statystycznych - pozwala na uzyskanie wyników badania o stopniu precyzji zbliżonym do metody graficznej. Diagram korelacyjny to przecież nic innego, jak zaprezentowany w postaci graficznej materiał statystyczny ujęty w szeregach bądź tablicy statystycznej. W przypadku informacji o wartościach cech ujętych w postaci szeregów statystycznych ocena charakteru związku jak i natężenia oraz kierunku zależności odbywa się na podstawie określenia charakteru wzajemnych powiązań wartości dwóch badanych cech. Analiza taka pozwala na wstępne wyodrębnienie jednej z powyższych sytuacji oznaczonych jako (***a***) –(***f***).

Dla licznych zbiorowości (przyjmuje się zwykle, że ich liczebność przekracza 30) materiał statystyczny opisujący je ujmuje się w formie tablicy statystycznej zwanej w tym przypadku tablicą korelacyjną[6]. Tablica taka prezentuje jednoczesny rozkład badanej zbiorowości ze względu na dwie cechy, stąd też spotykane w literaturze określenie, iż przedstawia ona dwuwymiarowy rozkład zbiorowości. W tablicy takiej dla cech typu liczbowego ich wartości ujmowane są najczęściej w postaci przedziałów klasowych[7], zaś dla cech typu opisowego przyjmuje się występujące w zbiorowości ich warianty. Sformułowane przedziały klasowe bądź warianty cech ujmowane są w „główce" i „boczku" tablicy. Zgromadzony w postaci szeregów szczegółowych materiał statystyczny może być „przeniesiony" do przygotowanej makiety tablicy korelacyjnej tzw. metodą kreskową", tzn. po odczytaniu wartości cech dla określonej jednostki statystycznej jest ona umieszczana za pomocą kreski w odpowiednim polu tablicy korelacyjnej. Wypełnioną „kreskami" tablicę korelacyjną traktujemy jako roboczą, a po ich zliczeniu otrzymujemy wynikową tablicę liczebności, która może być przekształcona w tablicę częstości. W tym celu należy wszystkie liczebności podzielić przez

[6] Pierwotną postacią takiego materiału są jednak szeregi statystyczne

[7] Zasady konstrukcji takich przedziałów są identyczne jak w przypadku szeregów z przedziałami klasowymi

liczebność ogólną (N). Ilustracją takiej tablicy liczebności i jej przekształcenia w tablicę częstości jest poniższa tablica 3.1.

Tabela 3.1. Pracownicy firmy „Z" ze względu na wiek i staż pracy (w latach)

Staż pracy / *Wiek*	*0 - 5*	*5 - 10*	*10 – 15*	n_i / f_i
20 - 30	15 0,30	10 0,20	- -	*25* *0,50*
30 - 40	5 0,10	10 0,20	10 0,20	*25* *0,50*
n_j f_j	*20* *0,40*	*20* *0,40*	*10* *0,20*	*50* *1,00*

Źródło: Badania własne

Na jej podstawie dokonana zostanie charakterystyka podstawowych typów rozkładów występujących w tablicy korelacyjnej, tj. *rozkładu łącznego, brzegowych i warunkowych.*

a) *rozkład łączny* opisuje badaną zbiorowość jednocześnie ze względu na obie cechy (w tablicy znajduje się w jej zacieniowanej części). Liczebności w tym rozkładzie oznaczane są symbolem n_{ij} (zaś częstości f_{ij}) i należy je odczytywać jako liczebności (częstości) i-tego wiersza i j-tej kolumny. W naszym przykładzie *rozkład łączny liczebności* odczytujemy następująco: w badanej zbiorowości pracowników 15 osób posiada wiek 20 – 30 lat i staż pracy 0 – 5 lat, 10 osób w wieku 20 – 30 lat i stażu 5 – 10 lat, 5 osób posiadających wiek 30 – 40 lat oraz staż 0 – 5 lat, 10 osób w wieku 30 – 40 lat i stażu pracy 5 – 10 lat jak również 10 osób posiadających wiek 30 – 40 lat i staż 10 – 15 lat; *rozkład łączny częstości* odczytywany jest zwykle w ujęciu procentowym w sposób następujący: 30 % pracowników tej firmy posiada wiek 20 – 30 lat i staż pracy 0 – 5 lat, 20 % jest w wieku 20 – 30 lat i stażu 5 – 10 lat, 10 % posiada wiek 30 – 40 lat oraz staż 0 – 5 lat, 20 % jest w wieku 30 – 40 lat i stażu pracy 5 – 10 lat, również 20 % posiada wiek 30 – 40 lat i staż 10 – 15 lat,

b) w tablicy występują dwa *rozkłady brzegowe*: *wierszowy i kolumnowy*. Każdy z nich opisuje badaną zbiorowość ze względu na jedną cechę. *Rozkład brzegowy wierszowy* tworzą sumy liczebności (częstości) liczone po poszczególnych wierszach. Są one oznaczane symbolem n_i (częstości - f_i) i należy je odczytywać jako liczebności (częstości) i-tego wiersza. W naszym przykładzie liczebności (częstości) te są umieszczone w ostatniej kolumnie i ilustrują one rozkład zbiorowości pracowników ze względu na wiek. Na ich podstawie można stwierdzić, że w badanej zbiorowości 25 pracowników, tj. 50 % zbiorowości pracowników jest w wieku 20 – 30 lat i taka sama liczba posiada wiek 30 – 40 lat. *Rozkład brzegowy kolumnowy* tworzą sumy liczebności (częstości) liczone w poszczególnych kolumnach. Oznaczane są

one symbolem n_j (częstości – f_j). Należy je odczytywać jako liczebności (częstości) j-tej kolumny. W podanym przykładzie liczebności (częstości) te opisują rozkład zbiorowości pracowników ze względu na ich staż pracy. Na podstawie tego rozkładu można stwierdzić, że 20 pracowników, tj. 40 % zbiorowości posiada staż pracy z przedziału 0 – 5 lat, również 20 (40 %) – staż od 5 – 10 lat oraz 10 (20 %)– staż pracy 10 – 15 lat,

c) *rozkłady warunkowe* opisują badaną zbiorowość ze względu na jedną z cech przy założeniu określonego warunku nałożonego na cechę drugą. I tak *rozkłady warunkowe cechy Y* charakteryzują rozkład zbiorowości ze względu na warianty tej cechy pod warunkiem, że cecha X przyjmuje określoną realizację; cecha Y posiada tyle rozkładów warunkowych, ile wariantów przyjmuje cecha X. Z kolei rozkłady warunkowe cechy X charakteryzują rozkład zbiorowości ze względu na realizacje tej cechy pod warunkiem, że cecha Y przyjmuje określony wariant; liczba rozkładów warunkowych cechy X jest równa liczbie wariantów cechy Y.

W przypadku omawianej tablicy występują dwa rozkłady warunkowe cechy Y (tzn. stażu pracy) i odczytujemy je następująco: przy założeniu, że pracownicy są w wieku 20 – 30 lat 15 z nich ma staż pracy do 5 lat i 10 ma staż 5 – 10 lat, natomiast wśród pracowników w wieku 30 – 40 lat 10 posiada staż do 5 lat, również 10 - staż od 5 – 10 lat oraz 5 – staż od 10 do 15 lat. Dla cechy X (tzn. wieku) występują trzy rozkłady warunkowe odczytywane następująco: przy założeniu, że staż pracy pracowników zawiera się w przedziale 0 – 5 lat 15 z nich jest w wieku 20 – 30 lat i 5 w wieku 30 – 40 lat; przy założeniu, że ich staż wynosi 5 – 10 lat po 10 pracowników jest w wieku 20 – 30 lat oraz 30 – 40 lat; wszyscy pracownicy o stażu 10 – 15 lat posiadają wiek z przedziału 30 – 40 lat. Należy dodać, że w przypadku tablicy korelacyjnej częstości procedura ustalania rozkładów warunkowych jest bardziej skomplikowana i zostanie tu pominięta.

Wstępna ocena zależności na podstawie materiału statystycznego ujętego w tablicy korelacyjnej opierać się może na ocenie rozkładu liczebności (częstości) w tablicy jak również na ocenie podobieństwa rozkładów warunkowych. Koncentracja liczebności (częstości) wzdłuż przekątnych tablicy korelacyjnej wskazuje na występowanie znacznego natężenia zależności; jeśli jest to przekątna biegnąca z lewego górnego narożnika tablicy do prawego dolnego – to sytuacja taka oznacza występowanie zależności o kierunku dodatnim; w przeciwnym przypadku będzie to zależność o kierunku ujemnym.

Dokonując oceny zależności na podstawie rozkładów warunkowych należy kierować się zasadą: im wyższy stopień podobieństwa rozkładów warunkowych określonej cechy (przy zmieniających się warunkach nałożonych na cechę przeciwną) tym mniejsze jest natężenie zależności. Jeśli są one identyczne, zależność nie występuje. Dodać należy, iż na badaniu podobieństwa rozkładów warunkowych opierają się niektóre z miar zależności.

Omówione wyżej sposoby badania zależności pozwalają jedynie na wstępną jej ocenę. Precyzyjniejszych wyników badania dostarczają *metody formalne* wykorzystujące miary zależności cech. Przedmiotem dalszych rozważań będą wyodrębniane w literaturze dwie metody badania współzależ-

ności statystycznej, tj. *metoda nieparametrycznego (stochastycznego) i metoda parametrycznego (korelacyjnego) badania współzależności. Pierwsza z metod* opiera się na badaniu podobieństwa rozkładów warunkowych (analiza dotyczy jedynie rozkładów cech a nie ich wartości) cechy zależnej. Natężenie zależności w tym przypadku określamy na podstawie stopnia podobieństwa warunkowych rozkładów tej cechy. *W drugim przypadku* przedmiotem analizy jest badanie podobieństwa parametrów warunkowych (średnich warunkowych) cechy zależnej. Wyższe podobieństwo średnich warunkowych cechy zależnej oznaczać będzie mniejsze natężenie zależności. W ramach obu metod wykorzystywanych jest wiele miar współzależności cech o ściśle określonych właściwościach i wynikających stąd warunkach ich stosowania. Ich prezentacja zostanie dokonana na tle hipotetycznej „idealnej" miary zależności.

3.2. WŁASNOŚCI „IDEALNEJ" MIARY ZALEŻNOŚCI

W stosunku do miar wykorzystywanych w procedurach badania współzależności formułowane są określone postulaty dotyczące ich własności. Na ich podstawie można sformułować model miary idealnej, która winna posiadać następujące własności:

a) winna być niemianowana, gdyż umożliwia to prowadzenie analizy porównawczej zależności różnych cech,
b) winna być unormowana, tzn. winna przyjmować wartości ze skończonego przedziału liczbowego; umożliwia to ocenę natężenia zależności pomiędzy badanymi cechami. Miary spełniające ten postulat przyjmują najczęściej wartości z przedziału liczbowego <0 ; 1> Dla oceny natężenia zależności można przyjąć następujące kryteria:

Wartość miary zależności	*Natężenie zależności*
0	niezależność (brak zależności)
(0 – 0,33>	zależność słaba
(0,33 – 0,66>	zależność wyraźna
(0,66 – 1,00)	zależność silna
1,00	zależność funkcyjna

c) oprócz natężenia winna wskazywać również kierunek zależności; jej wartość winna informować, czy w określonym przypadku mamy do czynienia z zależnością o kierunku dodatnim bądź ujemnym. W przypadku zależności dodatniej rosnącym (malejącym) wartościom cechy niezależnej odpowiadają rosnące (malejące) wartości cechy zależnej. Zależność ujemna oznacza, iż rosnącym (malejącym) wartościom cechy niezależnej towarzyszą malejące (rosnące) wartości cechy zależnej. Miary wskazujące kierunek zależności przyjmują wartości zarówno dodatnie jak i ujemne; w przypadku miar unormowanych przyjmują one wartości z przedziału liczbowego <-1; 1>. Badanie kierunku zależności odnosi się do relacji zachodzących między cechami, których wartości są wyrażone przynajmniej na skali porządkowej,

d) winna być symetryczna; wówczas wartość miary jest identyczna bez względu na kierunek badania zależności, co oznacza, iż wartość miary zależności Y od X jest identyczna jak miara zależności X od Y. Własność ta jest spełniona w przypadku związków prostoliniowych lub w przypadku badaniach związków zachodzących między cechami opisowymi,
e) istnieje możliwość jej stosowania do badania zależności w związkach prosto- i krzywoliniowych. Spełnienie tej własności wyklucza konieczność badania „charakteru" związku przed zastosowaniem określonej miary do badania zależności. W przypadku miar, które mogą być stosowane do badania zależności w związkach prostoliniowych, właściwe badanie zależności musi być poprzedzone badaniem potwierdzającym występowanie związku prostoliniowego między badanymi cechami. Negatywny wynik takiego badania zmusza nas do wyboru innej miary zależności. Brak możliwości zbadania charakteru związku (np. gdy informacje zawarte są w postaci tablicy korelacyjnej) wymaga przynamniej przyjęcia założenia o występowaniu związku prostoliniowego. Wykorzystanie miary typowej do badania zależności w związkach krzywoliniowych w przypadku związku prostoliniowego nie stanowi błędu; należy tu dodać oczywistą uwagę, iż problem badania „charakteru" związku nie odnosi się do zależności występujących między cechami opisowymi,
f) istnieje możliwość jej stosowania do badania zależności w dowolnym układzie rodzajowym cech; badanie zależności może dotyczyć trzech następujących sytuacji: badamy zależność między dwiema cechami liczbowymi, np. między stażem pracy i zarobkami pracowników; badanie zależności między dwiema cechami opisowymi, np. między poziomem wykształcenia pracowników a miejscem zajmowanym w strukturze organizacyjnej firmy; wreszcie zależności między cechą liczbową a opisową, np. między poziomem wykształcenia pracowników a ich zarobkami. Idealną miarą zależności byłaby taka, którą można zastosować w każdej z wymienionych sytuacji,
g) winna spełniać własność jednolitej preferencji wartości, co oznacza, iż wzrostowi wartości miary towarzyszy wzrost natężenia zależności między cechami,
h) winna być prosta rachunkowo.

Należy zauważyć, że mało prawdopodobne jest jednoczesne „pogodzenie" wszystkich ww. własności (zwłaszcza ostatniej z wcześniej wymienionymi), dlatego poszczególne miary w różnym stopniu spełniają kryteria miary idealnej.

3.3. METODA NIEPARAMETRYCZNEGO BADANIA WSPÓŁZALEŻNOŚCI

Ta metoda badania zależności polega na badaniu prawidłowości występujących w zakresie współwystępowania wariantów cechy zależnej przyporządkowanych poszczególnym wariantom cechy niezależnej i stwierdzaniu, na ile rozkład wariantów cechy zależnej jest zdeterminowany zmieniającymi się odmianami cechy niezależnej. Praktycznie oznacza to badanie podobieństwa rozkładów warunkowych częstości cechy zależnej. Natężenie zależno-

ści między badanymi cechami może być oceniane na podstawie stopnia podobieństwa tych rozkładów. Jeśli wszystkim wariantom cechy niezależnej odpowiadają identyczne rozkłady warunkowe cechy zależnej, oznacza to występowanie niezależności cech w sensie nieparametrycznym. W przypadku zróżnicowanych rozkładów warunkowych występuje zależność cech, a stopień ich „niepodobieństwa" pozwala na wstępną ocenę natężenia zależności.

Sformułowany wyżej w sposób opisowy warunek niezależności cech można sformalizować w sposób następujący:

jeśli dla wszystkich kombinacji wariantów cech zależnej i niezależnej (czyli wszystkich pól rozkładu łącznego w tablicy korelacyjnej) zachodzi relacja:

$$f_{ij} = f_i \cdot f_j \text{, tzn. } f_{ij} - f_i \cdot f_j = 0$$

wówczas występuje niezależność badanych cech.

Warunek ten jest symetryczny.

3.3.1. Współczynnik zbieżności Czuprowa

W oparciu o podany wyżej warunek skonstruowany został *współczynnik zbieżności Czuprowa* o postaci:

$$d^c = \sqrt{\frac{\sum_{i,j} \frac{\left(f_{ij} - f_i \cdot f_j\right)^2}{f_i \cdot f_j}}{\min(r,s) - 1}} \qquad 3.1.$$

gdzie: r , s – oznacza liczbę występujących w tablicy korelacyjnej wariantów cech X i Y

Jeśli przyjąć jako wyjściową sytuację, gdy materiał statystyczny stanowiący podstawę do badania zależności jest ujęty w tablicy korelacyjnej liczebności wówczas dla ustalenia współczynnika Czuprowa należy wykonać kolejno następujące działania:

1) przekształcić tablicę liczebności w tablicę częstości zawierającą wyrażenia:
   f_{ij}, f_i, f_j
2) ustalić dla każdego pola tablicy korelacyjnej iloczyny $f_i \cdot f_j$,
3) dla każdego pola tablicy sprawdzić warunek niezależności $f_{ij} - f_i \cdot f_j = 0$; jeśli jest on spełniony dla wszystkich pól stwierdzamy występowanie niezależności cech, w przeciwnym przypadku kontynuujemy kolejne czynności,
4) w każdym polu tablicy obliczyć kwadrat ustalonej różnicę $\left(f_{ij} - f_i \cdot f_j\right)^2$,

5) w każdym polu obliczyć iloraz kwadratu różnicy oraz iloczynu $f_i \cdot f_j$, tzn. $\dfrac{\left(f_{ij} - f_i \cdot f_j\right)^2}{f_i \cdot f_j}$

6) dokonać sumowania otrzymanych wielkości i uzyskaną wielkość wstawić do licznika podanego wzoru.

Powyższą procedurę postępowania ilustruje przykład 3.1.

Przykład 3.1.

Zebrane informacje dotyczące czasu pozostawania bez pracy oraz poziomu wykształcenia badanej grupy bezrobotnych ujęto w postaci następującej tablicy korelacyjnej:

Tablica 3.2. Bezrobotni miasta „K" według czasu pozostawania bez pracy (w miesiącach) oraz poziomu wykształcenia (stan na 30.06.2003 r.)

Wykształcenie (X) *Czas pozostawania bez pracy (Y)*	*podstawowe*	*średnie*	*wyższe*	n_i
0 - 6	10	15	10	35
6 - 12	42	30	3	75
12 - 18	33	15	2	50
18 – 24	25	15	-	40
n_j	110	75	15	200

Źródło: Dane umowne

Zbadać, czy występuje zależność czasu pozostawania bez pracy od poziomu wykształcenia bezrobotnych?

Rozwiązanie.

Badanie zależności zostanie dokonane zgodnie z podanym wyżej algorytmem postępowania, a wykonywanie działania umieszczono w poniższej tablicy roboczej (w pierwszym polu tablicy ponumerowano kolejne działania zgodnie z podaną wyżej kolejnością)

Suma ilorazów (występujących w ostatniej pozycji każdego pola) wynosi 0,1685., zaś minimalna liczba wariantów obu cech – 3. Wstawiając te wielkości do *formuły 3.1.* otrzymujemy:

$$d^c = \sqrt{\frac{0{,}1685}{3-1}} = 0{,}29$$

Uzyskana wielkość oznacza występowanie niewielkiej zależności czasu pozostawania bez pracy od poziomu wykształcenia bezrobotnych.

Wykształcenie(X) *Czas pozostawania bez pracy (Y)*	*podstawowe*	*średnie*	*wyższe*	n_i f_i
0 – 6	10 (1) 0,05 (2) 0,0963 (3) 0,0463 (4) 0,0021 0,0223	15 0,075 0,0656 0,0094 ~0,0001 0,0013	10 0,05 0,0131 0,0369 0,0014 0,1068	35 0,175
6 – 12	42 0,21 0,2063 0,0037 ~0,0001 0,0001	30 0,15 0,1406 0,0094 ~0,0001 0,0006	3 0,015 0,0281 -0,0131 0,0002 0,0061	75 0,375
12 - 18	33 0,165 0,1375 0,0275 0,0008 0,0055	15 0,075 0,0938 -0,0188 0,0004 0,0047	2 0,01 0,0188 -0,0088 0,0001 0,0041	50 0,25
18 – 24	25 0,125 0,1100 0,0150 0,0002 0,0020	15 0,075 0,0750 0 0 0	- - 0,0150 -0,015 0,0002 0,015	40 0,20
n_j f_j	110 0,55	75 0,375	15 0,075	200 1,00

Analizowana miara zależności posiada następujące własności:
- jest wielkością niemianowaną,
- przyjmuje wartości z unormowanego przedziału liczbowego < 0 ; 1>,
- wskazuje natężenie zależności,
- jest miarą symetryczną,
- może być stosowana do badania zależności w dowolnym układzie rodzajowym cech; należy ją stosować przede wszystkim do badania zależności między cechami opisowymi, bądź między opisowymi i liczbowymi.

3.3.2. Współczynnik zależności Hellwiga

Pewną modyfikację współczynnika zbieżności Czuprowa stanowi *współczynnik zależności Hellwiga*. Jego konstrukcja opiera się również na wykorzystaniu warunku niezależności cech w sensie nieparametrycznym. Procedura jego wyznaczania w porównaniu ze współczynnikiem Czuprowa jest mniej skomplikowana rachunkowo i wymaga wykonania pierwszych trzech

z podanych wyżej działań, w wyniku których otrzymujemy wyrażenia o postaci: $f_{ij} - f_i \cdot f_j$. Wśród otrzymanych wielkości wyodrębniamy dwa podzbiory:

$$G \text{, dla którego } f_{ij} - f_i \cdot f_j > 0$$

$$M \text{, dla którego } f_{ij} - f_i \cdot f_j < 0$$

Odpowiednio dla podanych dwóch podzbiorów zostały zaproponowane dwie postacie współczynnika zależności Hellwiga:

a) *dla podzbioru G:*

$$d_G^H = \sqrt{\frac{\sum_{i,j \in G} f_{ij} - \sum_{i,j \in G} f_i \cdot f_j}{1 - \frac{1}{\min(r,s)}}} \qquad 3.2.$$

b) *dla podzbioru M:*

$$d_M^H = \sqrt{\frac{\sum_{i,j \in M} f_i \cdot f_j - \sum_{i,j \in M} f_{ij}}{1 - \frac{1}{\min(r,s)}}} \qquad 3.3.$$

Należy zaznaczyć, że w obu przypadkach uzyskujemy tę samą wartość współczynnika Hellwiga. Współczynnik ten posiada własności identyczne jak współczynnik zbieżności Czuprowa, lecz jest mniej pracochłonny. Dla ilustracji procedury wyznaczania tego współczynnika wykorzystamy podany wyżej *przykład 3.1* i wykonane w tablicy roboczej obliczenia oznaczone jako *(1) – (3).*

Przyjmujemy, że w dalszym postępowaniu ograniczymy się do podzbioru G, czyli tych pól tablicy korelacyjnej, w których spełniony jest warunek $f_{ij} - f_i \cdot f_j > 0$, pozostałe pola tablicy pozostawimy puste (por. poniższa tablica robocza)

Wykształcenie (X) *Czas pozostawania bez pracy (Y)*	*podstawowe*	*średnie*	*wyższe*	f_i
0 - 6	X	(1) 0,075 (2) 0,0656 (3) 0,0094	0,05 0,0131 0,0369	0,175
6 - 12	0,21 0,2063 0,0037	0,15 0,1406 0,0094	X	0,375
12 - 18	0,165 0,1375 0,0275	X	X	0,25
18 – 24	0,125 0,1100 0,0150	X	X	0,20
f_j	0,55	0,375	0,075	1,00

Dla wyodrębnionych pól wyznaczamy odpowiednie sumy:

$$\sum_{i,j\in G} f_{ij} = 0{,}775 \qquad \text{i} \qquad \sum_{i,j\in G} f_i \cdot f_j = 0{,}6731$$

podstawiając je do *wzoru 3.2*. otrzymujemy:

$$d_G{}^H = \sqrt{\frac{0{,}775 - 0{,}6731}{1 - \frac{1}{3}}} = 0{,}39$$

Uzyskana wartość miary odbiega nieznacznie od otrzymanej dla współczynnika zbieżności Czuprowa. Oznacza ona występowanie wyraźnej zależności czasu pozostawania bez pracy od poziomu wykształcenia bezrobotnych.

3.4. METODA PARAMETRYCZNEGO BADANIA WSPÓŁZALEŻNOŚCI

Ta metoda badania współzależności cech, zwana również metodą korelacyjną, opiera się na analizie wzajemnych powiązań wartości cechy, a nie jedynie – jak w przypadku metody nieparametrycznej – ich rozkładów warunkowych (liczebności bądź częstości). Najczęściej wykorzystywane w ramach tej metody miary zależności to:

a) stosunek korelacyjny,
b) współczynnik korelacji liniowej Pearsona,
c) współczynnik korelacji rang Spearmana.

3.4.1. Stosunek korelacyjny

Istotą tej miary zależności jest badanie podobieństwa średnich warunkowych cechy zależnej (Y). Są to średnie arytmetyczne wartości tej cechy przyporządkowane występującym w tablicy korelacyjnej wariantom cechy niezależnej. Jeśli przyjmiemy je oznaczać jako $\bar{y}_{x_j}$ lub $\bar{y}_{x_i}$ (w zależności od układu tablicy korelacyjnej: w pierwszym przypadku cecha X będzie występowała w „główce" tablicy, w drugim – w „boczku" tablicy), to można sformułować następujący formalny warunek niezależności cech w sensie korelacyjnym:

jeśli spełniona jest relacja:

$\bar{y}_{x_1} = \bar{y}_{x_2} = \bar{y}_{x_3} = ... = \bar{y}_{x_k}$ *(gdzie: k – to liczba wariantów cechy niezależnej), to pomiędzy badanymi cechami występuje niezależność w sensie korelacyjnym (parametrycznym).*

Jeśli warunek ten nie jest spełniony wówczas występuje zależność w sensie parametrycznym. Z powyższego wynika, że sama analiza średnich warunkowych pozwala na wstępną ocenę zależności cech. Dodatkowo, gdy rosnącym wartościom cechy niezależnej towarzyszą rosnące wartości średnich warunkowych cechy zależnej, stwierdzamy występowanie zależności o

kierunku dodatnim; natomiast, gdy rosnącym wartościom cechy niezależnej odpowiadają malejące wartości średnich warunkowych cechy zależnej – kierunek zależności jest ujemny. W sytuacji, gdy średnie warunkowe nie wykazują jednoznacznej tendencji, nie określamy na ich podstawie kierunku zależności. Analiza średnich warunkowych pozwala zatem na wstępną ocenę zależności cech. Dla oceny natężenia zależności wykorzystywany jest *stosunek korelacyjny* o postaci:

$$r^k = \frac{s(\bar{y}_{x_j})}{s(y)} = \frac{\sqrt{\frac{\sum_j (\bar{y}_{x_j} - \bar{y})^2 \cdot n_j}{N}}}{\sqrt{\frac{\sum_i (y_i - \bar{y})^2 \cdot n_i}{N}}} \qquad 3.4.$$

lub

$$r^k = \frac{s(\bar{y}_{x_i})}{s(y)} = \frac{\sqrt{\frac{\sum_i (\bar{y}_{x_i} - \bar{y})^2 \cdot n_i}{N}}}{\sqrt{\frac{\sum_j (y_j - \bar{y})^2 \cdot n_j}{N}}} \qquad 3.5.$$

gdzie:

$s(\bar{y}_{x_j})$ lub $s(\bar{y}_{x_i})$ - odchylenie standardowe średnich warunkowych cechy zależnej,

$s(y)$ - odchylenie standardowe ogólne cechy zależnej,

$\bar{y}_{x_j}$ lub $\bar{y}_{x_i}$ - średnie warunkowe cechy zależnej,

$\bar{y}$ - średnia ogólna cechy zależnej.

Podana pierwsza wersja tej miary jest wykorzystywana, gdy wartości cechy zależnej (Y) występują w „boczku" tablicy, zaś druga, gdy są one umieszczone w główce" tablicy. Należy również dodać, iż występujące we wzorach liczebności mogą być zastąpione częstościami.

Algorytm obliczania tej miary można ująć w następujących punktach:

1) wyznaczenie średnich warunkowych i średniej ogólnej cechy zależnej (Y),
2) ustalenie wariancji, a następnie odchylenia standardowego średnich warunkowych cechy zależnej,
3) ustalenie wariancji, a następnie odchylenia standardowego ogólnego cechy zależnej,
4) wyznaczenie wartości miary r^k.

Podaną wyżej procedurę postępowania ilustruje przykład 3.2.

Przykład 3.2.

W poniższej tablicy zawarto wyniki badania warunków materialnych losowej grupy gospodarstw domowych miasta „L" uwzględniające wysokość dochodów na 1 członka gospodarstwa domowego (Y) oraz liczbę osób w gospodarstwie (X).

Tablica 3.3 Gospodarstwa domowe miasta „L" według dochodów na 1 osobę w zł oraz liczbę osób w gospodarstwie

Dochód na 1 osobę w zł (Y)	*Liczba osób w gospodarstwie (X)*			n_i
	1 – 3	*4 – 6*	*7 - 9*	
150 – 350	2	13	20	35
350 – 550	8	17	-	25
550 – 750	20	20	-	40
n_j	30	50	20	100

Źródło: Dane umowne.

Na podstawie analizy średnich warunkowych oraz przy wykorzystaniu stosunku korelacyjnego zbadać, czy w badanej zbiorowości występuje zależność wysokości dochodu na osobę od liczby osób w gospodarstwie domowym.

Rozwiązanie:

Średnie warunkowe cechy zależnej będą stanowiły średnie dochodów na osobę w wyodrębnionych trzech kategoriach wielkości gospodarstw, tj. 1 – 3 osób, 4 – 6 osób oraz 7 – 9 osób. Dla ich wyznaczenia należy ustalić środki przedziałów klasowych dla dochodów (ujęto je w nawiasach w pierwszej kolumnie poniższej tablicy roboczej).

Poszczególne średnie warunkowe wyznaczone zostaną w następujący sposób:

$$\bar{y}_{1-3} = \frac{2\cdot 250 + 8\cdot 450 + 20\cdot 650}{30} = \frac{17100}{30} = 570 \text{ zł}$$

$$\bar{y}_{4-6} = \frac{13\cdot 250 + 17\cdot 450 + 20\cdot 650}{50} = \frac{23900}{50} = 478 \text{ zł}$$

$$\bar{y}_{7-9} = \frac{20\cdot 250}{20} = 250 \text{ zł}$$

Na podstawie uzyskanych średnich warunkowych można stwierdzić:

- po pierwsze, występuje zależność wysokości dochodów na 1 osobę od liczby osób w gospodarstwie domowym, ponieważ średnie warunkowe są zróżnicowane,
- po drugie, występuje ujemna zależność wysokości dochodów od liczby osób, ponieważ wraz ze wzrostem liczby osób w gospodarstwie maleje średnia wysokość dochodu na 1 osobę.

Dla pomiaru natężenia zależności wykorzystamy stosunek korelacyjny, a obliczenia pomocnicze zostały wykonane w tablicy roboczej. Niezależnie

od ustalonych średnich warunkowych należy wyznaczyć średni dochód na 1 osobę we wszystkich gospodarstwach domowych. Dokonamy tego według wzoru:

$$\bar{y} = \frac{\sum_i \dot{y}_i \cdot n_i}{N} = \frac{35 \cdot 250 + 25 \cdot 450 + 40 \cdot 650}{100} = \frac{46000}{100} = 460 \text{ zł}$$

Dochód (środki przedziałów - $\dot{y}_i$*)*	*Liczba osób (*x_i*)*			n_i	$(\dot{y}_i - \bar{y})^2$	$(\dot{y}_i - \bar{y})^2 \cdot n_i$
	1 – 3	*4 – 6*	*7 - 9*			
250	2	13	20	35	44100	1543500
450	8	17	-	25	100	2500
650	20	20	-	40	36100	1444000
n_j	30	50	20	100	X	$\sum_i = 2990000$
$\bar{y}_{x_j}$	570	478	250	X	X	X
$(\bar{y}_{x_j} - \bar{y})^2$	12100	324	44100	X	X	X
$(\bar{y}_{x_j} - \bar{y})^2 \cdot n_j$	363000	16200	882000	$\sum_j = 1261200$	X	X

Wyniki obliczeń pomocniczych podstawiamy do *wzoru 3.4* i otrzymujemy:

$$r^k = \frac{\sqrt{\frac{1261200}{100}}}{\sqrt{\frac{2990000}{100}}} = \frac{112,3}{172,9} = 0,65$$

Uzyskana wartość miary oznacza, że występuje wyraźna zależność wysokości dochodów na 1 osobę od liczby osób w gospodarstwie domowym. Wykorzystana miara nie określa kierunku zależności, ale na podstawie średnich warunkowych stwierdziliśmy, że jest on ujemny.

Omawiana miara zależności charakteryzuje się następującymi własnościami:

1) jest niemianowana,
2) jest unormowana w przedziale <0 ; 1>,
3) nie wskazuje kierunku zależności; w szczególnych przypadkach może być on określony w oparciu o średnie warunkowe,
4) jest w zasadzie niesymetryczna (zjawisko symetrii zachodzi tylko w przypadku związku prostoliniowego),
5) jest wykorzystywana do badania zależności w związkach krzywoliniowych lub, gdy charakter związku nie jest znany,
6) stosuje się ją do badania zależności, gdy przynajmniej cecha zależna jest liczbowa,
7) spełnia warunek jednolitej preferencji wartości.

3.4.2. Współczynnik korelacji liniowej Pearsona

W przypadku, gdy badanie zależności dotyczy dwóch cech liczbowych, a związek między nimi jest prostoliniowy, wówczas należy do tego celu wykorzystać *współczynnik korelacji liniowej Pearsona*. Miara ta występuje w dwóch wersjach:
- opartej na danych ujętych *w postaci szeregów szczegółowych*,
- wykorzystującej informacje ujęte *w tablicy korelacyjnej*.

Dla pierwszego przypadku obliczana jest ona według wzoru:

$$r^P = \frac{c(x,y)}{s(x)\cdot s(y)} = \frac{\dfrac{\sum_i (x_i - \bar{x})\cdot(y_i - \bar{y})}{N}}{\sqrt{\dfrac{\sum_i (x_i - \bar{x})^2}{N}}\cdot\sqrt{\dfrac{\sum_i (y_i - \bar{y})^2}{N}}} \qquad 3.6.$$

natomiast dla drugiego:

$$r^P = \frac{c(x,y)}{s(x)\cdot s(y)} = \frac{\dfrac{\sum_{i,j} (x_j - \bar{x})\cdot(y_i - \bar{y})\cdot n_{ij}}{N}}{\sqrt{\dfrac{\sum_j (x_j - \bar{x})^2\cdot n_j}{N}}\cdot\sqrt{\dfrac{\sum_i (y_i - \bar{y})^2\cdot n_i}{N}}} \qquad 3.7.$$

gdzie:

$c(x,y)$ – kowariancja cech X i Y,

$s(x)$ – odchylenie standardowe cechy X,

$s(y)$ – odchylenie standardowe cechy Y.

Wzór o *postaci 3.7*. odpowiada układowi tablicy, w której wartości cechy X umieszczone są w jej „główce" , zaś cechy Y w 'boczku"; w przypadku odwrotnego układu cech występujące w podanym wzorze indeksy *i* oraz *j* winny być zamienione miejscami. Należy również dodać, że występujące we *wzorze 3.7* liczebności mogą być zastąpione częstościami.

Jeśli dla badania zależności wykorzystujemy wersję *współczynnika korelacji dla szeregów szczegółowych*, wówczas wykonujemy kolejno następujące działania:

1) wyznaczamy na podstawie szeregów szczegółowych wartości średnie dla obu cech,
2) ustalamy wielkości różnic poszczególnych wartości cech X i Y oraz ich wartości średnich, tj. $x_i - \bar{x}$ oraz $y_i - \bar{y}$,
3) ustalamy kwadraty uzyskanych różnic,
4) dokonujemy ich sumowania,
5) ustalamy iloczyny różnic uzyskanych w p. 2,

6) dokonujemy sumowania tych iloczynów,
7) uzyskane sumy podstawiamy do *wzoru 3.6.*

Procedurę tę ilustruje przykład 3.3.

Przykład 3.3.

Dla zbadania zależności między wielkością miesięcznych wydatków na cele kulturalne (Y) a liczbą osób w gospodarstwie domowym (X) zebrano informacje dla 20 wylosowanych gospodarstw gminy „Z". Otrzymano dane zawarte w poniższej tablicy:

Tablica 3.4. Gospodarstwa domowe gminy „Z" według liczby osób oraz wysokości miesięcznych wydatków na cele kulturalne

Liczba osób	3	5	7	2	6	4	5	3	2	1	4	6	5	3	3	7	4	2	4	3
Wydatki w zł	70	75	60	100	85	70	90	50	70	70	110	75	40	60	90	90	75	60	80	90

Źródło: Dane umowne

Przy pomocy współczynnika korelacji liniowej Pearsona zbadać, czy występuje zależność wysokości wydatków od liczby osób w gospodarstwie domowym?

Rozwiązanie:

Obliczenia pomocnicze – zgodnie z podaną procedurą zostaną wykonane w poniższej tablicy roboczej:

Liczba osób (x_i)	*Wydatki* (y_i)	$x_i-\bar{x}$	$(x_i-\bar{x})^2$	$y_i-\bar{y}$	$(y_i-\bar{y})^2$	$(x_i-\bar{x})\cdot(y_i-\bar{y})$
3	70	-1	1	-5,5	30,25	5,5
5	75	1	1	-0,5	0,25	-0,5
7	60	3	9	-15,5	240,25	-46,5
2	100	-2	4	24,5	600,25	-49,0
6	85	2	4	9,5	90,25	19
4	70	0	0	-5,5	30,25	0
5	90	1	1	14,5	210,25	14,5
3	50	-1	1	-25,5	650,25	25,5
2	70	-2	4	-5,5	30,25	11
2	70	-2	4	-5,5	30,25	0
4	110	0	0	34,5	1190,25	0
6	75	2	4	-0,5	0,25	1
5	40	1	1	-35,5	1260,25	-35,5
3	60	-1	1	-15,5	240,25	15,5
3	90	-1	1	14,5	210,25	-14,5
7	90	3	9	14,5	210,25	43,5
4	75	0	0	-0,5	0,25	0
2	60	-2	4	-15,5	240,25	31,0
4	80	0	0	4,5	20,25	0
3	90	-1	1	14,5	210,25	-14,5
$\sum_i x_i =$ 80	$\sum_i y_i =$ 1510	X	$\sum_i =$ 50	X	$\sum_i =$ 5495	$\sum_i =$ 15,0

Wartości średnie dla obu cech będą wynosiły odpowiednio:

$$\bar{x} = \frac{80}{20} = 4{,}0 \text{ osoby} \qquad \bar{y} = \frac{1510}{20} = 75{,}5 \text{ zł.}$$

Uzyskane w tablicy roboczej wyniki obliczeń podstawiamy do *wzoru 3.6.* i otrzymujemy:

$$r^P = \frac{\frac{15}{20}}{\sqrt{\frac{50}{20}} \cdot \sqrt{\frac{5495}{20}}} = \frac{0{,}75}{1{,}58 \cdot 16{,}58} = 0{,}03$$

Otrzymana wartość miary wskazuje na występowanie niewielkiej zależności wysokości wydatków na cele kulturalne od liczby osób w badanych gospodarstwach domowych.

W przypadku badania zależności na podstawie *danych ujętych w tablicy korelacyjnej* wykorzystujemy drugą wersję współczynnika korelacji i w związku z tym wykonujemy kolejno następujące działania:

1) wyznaczamy wartości średnie dla obu cech,
2) ustalamy wielkości różnic poszczególnych wariantów cech X i Y oraz ich wartości średnich, tj. $x_j - \bar{x}$ oraz $y_i - \bar{y}$,
3) obliczamy iloczyny kwadratów ustalonych wyżej różnic oraz odpowiadających im liczebności brzegowych, tj. $(x_j - \bar{x})^2 \cdot n_j$ oraz $(y_i - \bar{y})^2 \cdot n_i$
4) dokonujemy sumowania tych iloczynów,
5) dla poszczególnych pól rozkładu łącznego ustalamy iloczyny $(x_j - \bar{x}) \cdot (y_i - \bar{y}) \cdot n_{ij}$
6) dokonujemy sumowania tych iloczynów,
7) uzyskane sumy podstawiamy do *wzoru 3.7.*

W przypadku, gdy wartości cech ujęte są w postaci przedziałów klasowych w obliczeniach wykorzystujemy środki tych przedziałów. Procedurę tę ilustruje przykład 3.4.

Przykład 3.4.

Na podstawie danych z *przykładu 3.2* zbadać zależność wysokości dochodów na 1 osobę od wielkości gospodarstwa domowego zakładając, że między cechami występuje związek prostoliniowy.

Rozwiązanie:

W tablicy korelacyjnej wartości obu cech ujęto w postaci przedziałów klasowych, wobec czego w dalszych obliczeniach będą wykorzystywane ich środki (ujęto je w poniższej tablicy roboczej). Ustalona w *przykładzie 3.2*

średnia dochodu ($\bar{y}$) wynosi 460 zł, a średnią liczby osób w gospodarstwie domowym ustalimy według wzoru:

$$\bar{x} = \frac{30 \cdot 2 + 50 \cdot 5 + 20 \cdot 8}{100} = \frac{470}{100} = 4{,}7 \text{ osoby.}$$

Dalsze obliczenia pomocnicze wykonano w tablicy roboczej.

Dochód (środki przedziałów- $\dot{y}_i$)	*Liczba osób (środki przedziałów ($\dot{x}_j$)*			n_i	$(\dot{y}_i - \bar{y})$	$(\dot{y}_i - \bar{y})^2 \cdot n_i$
	2	*5*	*8*			
250	2 567 1134	13 - 63 -819	20 -693 -13860	35	-210	1543500
450	8 27 216	17 -3 -51	- - -	25	-10	2500
650	20 -513 -10260	20 57 1140	- - -	40	190	1444000
n_j	30	50	20	100	X	$\sum_i = 2990000$
$\dot{x}_j - \bar{x}$	-2,7	0,3	3,3	X	X	X
$(\dot{x}_j - \bar{x})^2 \cdot n_j$	218,7	4,5	217,8	$\sum_j = 441$	X	X

Na podstawie wykonanych obliczeń ustalamy wartość kowariancji $c(x,y)$:

$$c(x,y) = \frac{1134 + (-819) + (-13860) + 216 + (-51) + (-10260) + 1140}{100} = \frac{-22500}{100} = -225 \text{ (zł} \cdot \text{osoby)}$$

Podstawiając tę wielkość i wyniki pozostałych obliczeń pomocniczych do *wzoru 3.7* otrzymujemy:

$$r^P = \frac{-225}{\sqrt{\frac{441}{100}} \cdot \sqrt{\frac{2990000}{100}}} = \frac{-225}{2{,}1 \cdot 172{,}9} = -0{,}62$$

Uzyskana wartość miary wskazuje na występowanie wyraźnej, dodatniej zależności wysokości dochodu na osobę od liczby osób w gospodarstwie

domowym. Przypomnijmy, że obliczony wcześniej stosunek korelacyjny przyjął wartość 0,65 (zależność wyraźna). Powstała niewielka różnica między wartościami obu miar świadczy o występowaniu między badanymi cechami związku o charakterze zbliżonym do prostoliniowego (przyjęte dla współczynnika Pearsona założenie o związku prostoliniowym było wobec tego uprawnione).

Podstawowe własności współczynnika Pearsona można ująć w następujących punktach:

1) miara ta jest niemianowana,
2) miara jest unormowana w przedziale <-1 ; 1>,
3) wskazuje natężenie i kierunek zależności,
4) miara ta jest symetryczna,
5) może być stosowana do badania zależności w związkach prostoliniowych,
6) badane cechy muszą mieć charakter liczbowych,
7) spełnia własność jednolitej preferencji wartości.

3.4.3. Współczynnik korelacji rang Spearmana

Stanowi on najmniej pracochłonną miarę zależności ustalaną w oparciu o dane ujęte w szeregach szczegółowych. Istota badania zależności w tym przypadku polega na badaniu zgodności uporządkowań (rang) wartości obu cech. Obie cechy muszą być wobec tego mierzone przynajmniej na skali porządkowej. Dodatkowo wymaga się by związek między nimi był prostoliniowy. Porządkowanie (rangowanie) wartości obu cech może się odbywać w kierunku rosnącym lub malejącym (należy jednak pamiętać, by kierunek porządkowania obu cech był ten sam). W przypadku, gdy w szeregu występuje kilka identycznych wartości cechy, wówczas przyporządkowywana jest im średnia z rang odpowiadających tym wartościom cechy. Jeśli występuje pełna zgodność uporządkowań wartości obu cech, to związek między nimi ma charakter funkcyjny. Miara ta jest ustalana według formuły:

$$r^{Sp} = 1 - \frac{6\sum_i \left(d_{x_i} - d_{y_i}\right)^2}{N^3 - N} \qquad 3.8.$$

gdzie:

d_{x_i}, d_{y_i} - rangi przypisane wartościom cechy odpowiadającym i-tej jednostce,

N – liczebność badanej zbiorowości.

Procedurę obliczania tej miary zależności ilustruje przykład 3.5.

Przykład 3.5.

Na podstawie danych z *przykładu 3.3* zbadać zależność wysokości wydatków na cele kulturalne od liczby osób w gospodarstwie domowym za pomocą współczynnika korelacji rang Spearmana. Przyjąć założenie, że związek między cechami jest prostoliniowy.

Rozwiązanie:

Obliczenia pomocnicze zostaną wykonane w poniższej tablicy roboczej. W kolumnach 3. i 4. dokonano rangowania wartości cech X i Y od wartości najniższych do najwyższych.

Liczba osób (x_i)	*Wydatki* (y_i)	d_{x_i}	d_{y_i}	$\left(d_{x_i} - d_{y_i}\right)^2$
3	70	7	7,5	0,25
5	75	15	11	16
7	60	19,5	4	240,25
2	100	2,5	19	272,25
6	85	17,5	14	12,25
4	70	11,5	7,5	16
5	90	15	16,5	2,25
3	50	7	2	25
2	70	2,5	7,5	25
2	70	2,5	7,5	25
4	110	11,5	20	72,25
6	75	17,5	11	42,25
5	40	15	1	196
3	60	7	4	9
3	90	7	16,5	90,25
7	90	19,5	16,5	9
4	75	11,5	11	0,25
2	60	2,5	4	2,25
4	80	11,5	13	2,25
3	90	7	16,5	90,25
Razem				1139,0

Podstawiając uzyskaną sumę do *wzoru 3.8* otrzymujemy:

$$r^{Sp} = 1 - \frac{6 \cdot 1139,0}{20^3 - 20} = 1 - \frac{6834,0}{7980} = 1 - 0,86 = 0,14$$

Uzyskana wartość miary wskazuje na występowanie niewielkiej, dodatniej zależności wydatków na cele kulturalne od wielkości gospodarstw domowych. Niewielka różnica w wartościach miary Pearsona i Spearmana wiąże się z mniejszą precyzją tej drugiej. Współczynnik Spearmana nie uwzględnia bowiem stopnia zróżnicowania poszczególnych wartości cechy. Stosowany jest on z reguły do wstępnego badania zależności. Współczynnik ten charakteryzują następujące własności:

1) jest miarą niemianowaną,
2) przyjmuje wartości z unormowanego przedziału <-1; 1>,
3) wskazuje natężenie i kierunek zależności,
4) jest miarą symetryczna,

5) może być stosowany do badania zależności w związkach prostoliniowych,
6) badane cechy muszą być wyrażone przynajmniej na skali porządkowej,
7) spełnia własność jednolitej preferencji wartości.

3.5. ANALIZA REGRESJI

Prezentowana wyżej analiza współzależności pozwala na ocenę siły (i ewentualnie kierunku) zależności między cechami opisującymi określoną zbiorowość statystyczną. Uzyskany wynik badania nie pozwala jednak określić „liczbowo", jak zmiany wartości jednej z cech wpływają na wartości cechy drugiej. Pomocna w tym zakresie jest analiza regresji. Jej celem jest skonstruowanie poniższych funkcji regresji charakteryzujących związek cech X oraz Y:

a) funkcja regresji Y względem X, która określa, jakie zmiany Y powoduje wzrost X o jednostkę,
b) funkcja regresji X względem Y, która określa jakie zmiany X powoduje wzrost Y o jednostkę.

Funkcje te pozwalają zatem na oszacowanie wartości jednej ze zmiennych przy założonych wartościach drugiej z nich. Należy jednak dodać, że rzadko szacowane są obie funkcje; jest to zasadne tylko wówczas, gdy badany związek ma charakter dwustronny. W większości przypadków szacowana jest jedna funkcja regresji - Y względem X.

Procedura szacowania parametrów funkcji regresji musi być poprzedzona analizą zebranego materiału empirycznego (w postaci szeregów statystycznych lub tablicy korelacyjnej) dla dokonania wstępnej oceny charakteru związku między cechami. Można w tym celu wykorzystać wskazówki podane w *podrozdziale 3.1* (tabelaryczna i graficzna ocena zależności). Jej celem jest dobór odpowiedniej postaci funkcji regresji: liniowej bądź krzywoliniowej (np. hiperbolicznej, parabolicznej, wykładniczej).

Dla dalszych rozważań przyjmiemy, że związek między cechami jest w przybliżeniu prostoliniowy. Wówczas szacowane funkcje regresji mają postać prostych określonych następującymi formułami:

a) *regresja Y względem X:*

$$\hat{y} = a_y \cdot x + b \qquad 3.9.$$

b) *regresja X względem Y*:

$$\hat{x} = a_x \cdot y + b \qquad 3.10.$$

gdzie:

a, b – parametry równania regresji ustalane według wzorów podanych niżej.

Dla dalszych rozważań jako podstawowe przyjmiemy pierwsze z podanych równań regresji, bowiem opisuje zależność cechy Y (zależnej) od cechy X (niezależnej). Występujące w nim parametry ***a*** i ***b*** ustalane są według wzorów:

$$a_y = \frac{C(x,y)}{s^2(x)} = \frac{\frac{\sum_i (x_i - \bar{x})\cdot(y_i - \bar{y})}{N}}{\frac{\sum_i (x_i - \bar{x})^2}{N}} \qquad 3.11.$$

$$b = \bar{y} - a_y \cdot \bar{x} \qquad 3.12.$$

W przypadku drugiego równania parametry ***a*** i ***b*** będą wyznaczane z formuł:

$$a_x = \frac{C(x,y)}{s^2(y)} = \frac{\frac{\sum_i (x_i - \bar{x})\cdot(y_i - \bar{y})}{N}}{\frac{\sum_i (y_i - \bar{y})^2}{N}} \qquad 3.13$$

$$b = \bar{x} - a_x \cdot \bar{y} \qquad 3.14.$$

gdzie:

$c(x,y)$ – kowariancja cech X i Y,

$s^2(x)$ – wariancja cechy X,

$s^2(y)$ – wariancja cechy Y.

Występujący w podanych równaniach parametr ***a*** określany jest mianem współczynnika regresji i oznacza skalę wzrostu (lub spadku) zmiennej, dla której skonstruowano równanie przy wzroście wartości drugiej zmiennej o jedną jednostkę, zaś parametr ***b*** oznacza poziom tej zmiennej, gdy druga z nich przyjmuje wartość równą zero.

Z *wzorów 3.11 i 3.13* wynika oczywista zależność:

$$r^P = \pm\sqrt{a_y \cdot a_x} \qquad 3.15.$$

jak również, iż o znaku współczynnika regresji decyduje znak współczynnika korelacji liniowej Pearsona wyrażającego zależność obu cech. Dla potrzeb szacowania równań regresji może być wykorzystywany materiał staty-

styczny ujęty zarówno w postaci szeregów szczegółowych jak i tablic korelacyjnych.

Oszacowane równania regresji cechy Y względem X oraz X względem Y nanoszone są zwykle na układ współrzędnych, a analiza ich wzajemnego położenia pozwala na ocenę stopnia współzależności obu cech. I tak, jeśli obie linie położone są blisko siebie (kąt między nimi jest niewielki), to taka sytuacja oznacza występowanie zależności silnej; w przypadku pokrywania się obu linii jest to zależność funkcyjna. W drugim skrajnym przypadku, gdy kąt między liniami jest prosty, występuje niezależność cech.

W praktyce oszacowane funkcje regresji jedynie z pewnym przybliżeniem odzwierciedlają faktyczne związki występujące między badanymi cechami (zwłaszcza, gdy zakłada się prostoliniowość związku). Powoduje to konieczność badania stopnia „dopasowania" oszacowanych funkcji do danych empirycznych. W tym celu wykorzystywany może być *współczynnik* φ^2 (*fi kwadrat*) określający, jaka część zmiennej zależnej (objaśnianej) nie jest wynikiem oddziaływania zmiennej niezależnej (objaśniającej).

Dla funkcji regresji Y względem X współczynnik ten wyraża się wzorem:

$$\varphi^2{}_y = \frac{\sum_i (y_i - \hat{y}_i)^2}{\sum_i (y_i - \bar{y})^2} \qquad 3.16.$$

natomiast dla funkcji regresji X względem Y ma postać:

$$\varphi^2{}_x = \frac{\sum_i (x_i - \hat{x}_i)^2}{\sum_i (x_i - \bar{x})^2} \qquad 3.17.$$

Im niższą wartość przyjmuje ten współczynnik, tym lepsze jest „dopasowanie" równania regresji do danych empirycznych. Jako miarę „dopasowania" funkcji wykorzystuje się również tzw. *współczynnik determinacji* R^2, który stanowi przeciwieństwo współczynnika zbieżności:

$$R^2 = 1 - \varphi^2 \qquad 3.18.$$

On z kolei określa, jaka część zmienności badanej zmiennej objaśnianej (zależnej) jest zdeterminowana oddziaływaniem zmiennej objaśniającej (niezależnej). W tym przypadku wyższa wartość tego współczynnika oznacza lepsze dopasowanie funkcji regresji do danych empirycznych.

Wskazać należy również na związek tego współczynnika ze współczynnikiem korelacji liniowej Pearsona. Związek ten wyraża się wzorem:

$$R^2 = \left(r^P\right)^2 \qquad 3.19$$

Procedurę szacowania równań regresji ilustruje przykład 3.6.

Przykład 3.6.
Dla losowej grupy pracowników zebrano informacje dotyczące ich stażu pracy (X) - w latach oraz wydajności pracy (Y) - w szt./godz. Informacje te ujęto w postaci podanej tablicy:

Tabela 3.5. Pracownicy przedsiębiorstwa „Z" według stażu pracy i wydajności pracy

Staż	8	10	2	4	7	25	18	17	13	5	7	10	22	8	4
Wydajność	12	13	6	7	10	15	20	19	15	9	10	14	20	12	8

Źródło: Dane umowne.

Oszacować równania regresji wydajności pracy pracowników względem ich stażu pracy i stażu pracy względem ich wydajności pracy. Dokonać na podstawie ich przebiegu stopnia zależności badanych cech. Dokonać oceny stopnia dopasowania uzyskanych funkcji do danych empirycznych.

Rozwiązanie:
Oszacowane zostaną dwa równania regresji o *postaci 3.9 i 3.10*. Dla wyznaczenia ich parametrów obliczenia pomocnicze wykonano w poniższej tablicy roboczej.

Staż pracy (x_i)	*Wydajność pracy* (y_i)	$x_i-\bar{x}$	$(x_i-\bar{x})^2$	$y_i-\bar{y}$	$(y_i-\bar{y})^2$	$(x_i-\bar{x})\cdot(y_i-\bar{y})$	$\hat{y}_i$	$(y_i-\hat{y}_i)^2$
1	*2*	*3*	*4*	*5*	*6*	*7*	*8*	*9*
8	12	-3	9	-1	1	3	11,38	0,38
10	13	-1	1	0	0	0	12,46	0,29
4	8	-7	49	-5	25	35	9,22	1,49
7	10	-4	16	-3	9	12	10,84	0,71
7	10	-4	16	-3	9	12	10,84	0,71
25	15	14	196	2	4	28	20,56	30,91
18	20	7	49	7	49	49	16,78	10,37
17	19	6	36	6	36	36	16,24	7,62
13	15	2	4	2	4	4	14,08	0,85
5	9	-6	36	-4	16	24	9,76	0,58
7	10	-4	16	-3	9	12	10,84	0,71
10	14	-1	1	1	1	-1	12,46	2,37
22	20	11	121	7	49	77	18,94	1,12
8	12	-3	9	-1	1	3	11,38	0,38
4	8	-7	49	-5	25	35	9,22	1,49
$\sum_i x_i = 165$	$\sum_i y_i = 195$	X	608	X	238	329	X	59,98

Wartości średnie dla obu cech wynoszą:

$$\bar{x} = \frac{165}{15} = 11 \text{ lat} \qquad \bar{y} = \frac{195}{15} = 13 \text{ szt/godz.}$$

Na podstawie wykonanych obliczeń pomocniczych wyznaczamy parametry równań regresji:
- Y względem X:

$$a = \frac{\frac{329}{15}}{\frac{608}{15}} = \frac{21{,}93}{40{,}53} = 0{,}54$$

$$b = 13 - 0{,}54 \cdot 11 = 7{,}06$$

i w konsekwencji otrzymujemy równanie o postaci:

$$\hat{y} = 0{,}54x + 7{,}06$$

Parametry tego równania można interpretować następująco: wzrost stażu pracy pracownika o 1 rok powoduje średni wzrost wydajności o 0,54 szt/godz., zaś przy „zerowym" stażu pracy wydajność wyniosłaby 7,06 szt/godz.
- X względem Y:

$$a = \frac{\frac{329}{15}}{\frac{238}{15}} = \frac{21{,}93}{15{,}87} = 1{,}38$$

$$b = 11 - 1{,}38 \cdot 13 = -6{,}94$$

a równanie regresji ma postać:

$$\hat{x} = 1{,}38y - 6{,}94 .$$

Przebieg prostych oszacowanych powyższymi równaniami obrazuje rys. 3.2.

W tym przypadku uzyskane parametry oznaczają: wzrost wydajności pracy o 1 szt/godz. wywoływany jest wzrostem stażu o 1,38 roku, interpretacja drugiego z parametrów jest bezsensowna.

Podstawowe znaczenie posiada pierwsze z oszacowanych równań, bowiem trudno uznać badany związek za dwustronny.

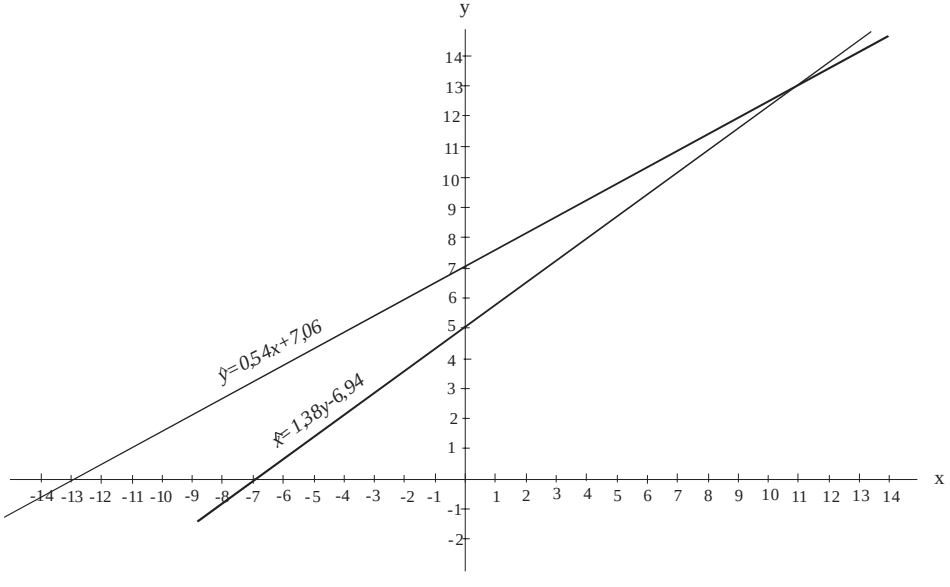
Na podstawie uzyskanego równania regresji można – zgodnie z *wzorem 3.15* - określić wartość współczynnika korelacji liniowej Pearsona mierzącego zależność między badanymi cechami:

$$r^P = +\sqrt{0{,}54 \cdot 1{,}38} \approx 0{,}86 .$$

Wskazuje on na silną dodatnią zależność wydajności pracy od stażu pracy.

Rys. 3.2

Proste regresji



Źródło: opracowanie własne

Na podstawie obliczeń pomocniczych zamieszczonych w tablicy roboczej określimy również stopień dopasowania funkcji regresji Y względem X do danych empirycznych. W kolumnie 8. ustalono wartości hipotetyczne $\hat{y}_i$ przy wykorzystaniu oszacowanego równania regresji, a w kolumnie 9. kwadraty odchyleń wartości empirycznych od odpowiadających im wartości hipotetycznych. Współczynnik zbieżności obliczony według *wzoru 3.16* wynosi:

$$\varphi^2_y = \frac{59{,}98}{238} = 0{,}25$$

Ustalony na jego podstawie współczynnik determinacji (por. *wzór 3.18*) będzie wynosił:

$$R^2 = 1 - 0{,}25 = 0{,}75$$

Wobec tego można stwierdzić, że zmienność wydajności pracy jest w 75 % zdeterminowana stażem pracy, a w 25% innymi czynnikami.

Rozdział IV

Analiza szeregów czasowych

4.1. WPROWADZENIE

Opisu rozwoju zjawiska bądź zbiorowości można dokonać również przy pomocy *szeregów czasowych (inaczej chronologicznych, dynamicznych).* Szeregi takie mogą mieć postać tablic dwukolumnowych bądź dwuwierszowych, w których wartości analizowanej zmiennej są uporządkowane według czasu. Zmienną tą będziemy w dalszym ciągu oznaczali jako Y, a jej realizacje y_{t_i}, zaś zmienną czasową -T i jej realizacje t_i. Czas jest zmienną ciągłą, której realizacjami mogą być poszczególne *momenty czasowe* bądź *przedziały czasowe (okresy).* Odpowiednio do tych dwóch przypadków można wyróżnić dwa typy szeregów prezentujących rozwój zjawiska bądź zbiorowości w czasie.

Szereg czasowy momentów prezentuje zwykle rozwój określonej zbiorowości (zjawiska), której jednostki istnieją w dłuższym okresie czasu, a szereg podaje jej stan w różnych momentach czasu (np. *ludność Jeleniej Góry w latach 1990 – 2000, stan na 31.12.)* Sumowanie wartości tak ujętej zmiennej nie ma sensu (oznaczałoby ono wielokrotne dodawanie tych samych wartości); można je natomiast odejmować od siebie, a różnice obrazują zmiany stanu badanej zbiorowości. Umożliwia to przejście z szeregów czasowych momentów na szeregi czasowe okresów. Informacje ujęte w takim szeregu obrazują przyrosty bądź spadki poziomu badanego zjawiska w określonym przedziale czasowym.

Przeciętny poziom zjawiska, którego wielkości ujęte są w szeregu czasowym momentów ustalamy przy wykorzystaniu *średniej chronologicznej* o postaci:

$$\bar{y}_{chr} = \frac{\frac{1}{2}y_1 + y_2 + ... + \frac{1}{2}y_n}{n-1} \qquad 4.1.$$

Szereg czasowy okresów zawiera informacje dotyczące zbiorowości lub zjawiska istniejącego w poszczególnych okresach, gdy jednostki tworzące je mogą być klasyfikowane rozłącznie według czasu. Rozwój zbiorowości w takim przypadku podawany jest jako liczba jednostek w poszczególnych okresach czasu (np. *liczba zawartych małżeństw w Polsce w latach 1991 - 2002*). Wartości ujęte w takim szeregu mogą być sumowane.

Przeciętny poziom zjawiska opisanego szeregiem czasowym okresów ustalamy (przy założeniu równości przedziałów czasowych) wykorzystując do tego celu *średnią arytmetyczną* w wersji przewidzianej dla szeregów szczegółowych (zob. *wzór 2.5*)

Analiza danych ujętych w szeregach czasowych okresów jest możliwa przy zapewnieniu ich porównywalności. Warunki tej porównywalności można ująć następująco:

- poziom zjawiska dla poszczególnych okresów musi być wyrażony w tych samych jednostkach miary,
- musi występować jednakowa rozpiętość poszczególnych przedziałów czasowych; w przypadku nierównych okresów należy badać natężenie zmiennej w przeliczeniu na przedziały o określonej rozpiętości.
- można porównywać między sobą informacje ujęte w szeregach tego samego typu (szeregi okresów bądź szeregi momentów) przy czym informacje muszą dotyczyć tych samych okresów bądź momentów,
- porównywalność danych jest możliwa, gdy dotyczą one tego samego obszaru (np. niemożliwe jest porównywanie danych dotyczących województw sprzed i po 1999 roku).

4.2. MIARY DYNAMIKI

4.2.1. Klasyfikacja miar dynamiki

Analiza dynamiki zjawiska opisanego szeregiem czasowym prowadzona jest przy wykorzystaniu miar dynamiki. Mogą być one klasyfikowane według dwóch kryteriów:

I. z uwagi na sposób wyznaczania wartości miary dynamiki możemy wyróżnić:

- *miary dynamiki różnicowe* (przyrosty)[8],
- *miary dynamiki ilorazowe* (indeksy, wskaźniki dynamiki, miary stosunkowe).

Miary dynamiki różnicowe są ustalane jako różnica poziomów zjawiska w okresach (momentach) badanym i porównawczym, natomiast *miary ilorazowe* ustalane są jako iloraz poziomów zjawiska w okresach (momentach) badanym i porównawczym.

II. ze względu na rodzaj przyjętej bazy porównawczej wyróżnia się:

- *miary dynamiki jednopodstawowe* (miary dynamiki o podstawie stałej),
- *miary dynamiki łańcuchowe* (miary dynamiki o podstawie zmiennej).

[8] Posługiwanie się kategorią „przyrostu" w przypadku malejącego poziomu zjawiska (czyli ujemnej wartości tej miary) wydaje się być niezręcznym określeniem

W przypadku *miar jednopodstawowych* dla celów analizy dynamiki jeden z okresów (momentów) przyjmowany jest jako stała baza porównawcza i poziom zjawiska każdego z pozostałych okresów (momentów) jest porównywany (na zasadzie różnicy bądź ilorazu) z poziomem zjawiska występującym w okresie (momencie) bazowym. Istotny problem stanowi tutaj kwestia wyboru bazy porównawczej; nie powinien jej stanowić okres (moment), w którym poziom zjawiska był wyjątkowo niski bądź wysoki. Najczęściej jako podstawę porównań przyjmuje się poziom zjawiska w pierwszym występującym w szeregu okresie (momencie).

Dokonując analizy dynamiki przy pomocy *miar łańcuchowych* poziom zjawiska każdego z okresów (momentów) porównywany jest (na zasadzie różnicy bądź ilorazu) z poziomem zjawiska występującym w okresie (momencie) poprzedzającym; badamy więc w tym przypadku zmiany poziomu zjawiska z okresu na okres.

Zarówno miary różnicowe jak i ilorazowe mogą mieć charakter miar jednopodstawowych jak i łańcuchowych.

4.2.2. Różnicowe miary dynamiki

Wśród miar różnicowych wyróżnia się dodatkowo:

- *różnicę absolutną*,
- *różnicę względną*.

Obie miary mogą mieć charakter miar jednopodstawowych lub łańcuchowych.

Różnicę absolutną jednopodstawową ($R^a_{r/s}$) obliczamy według wzoru:

$$R^a_{r/s} = y_r - y_s \qquad 4.2.$$

natomiast *różnicę absolutną łańcuchową* ($R^a_{r/r-1}$) według formuły:

$$R^a_{r/r-1} = y_r - y_{r-1} \qquad 4.3.$$

gdzie: y_r - oznacza poziom zjawiska w okresie (momencie) badanym,

y_s - oznacza poziom zjawiska w okresie (momencie) bazowym,

y_{r-1} - poziom zjawiska w okresie (momencie) poprzedzającym.

Wymienione miary mają charakter miar absolutnych (mianowanych); nie mogą więc być wykorzystywane w analizie porównawczej dynamiki różnych zjawisk bądź zbiorowości. Mogą one przyjmować wartości z przedziału $(-\infty ; +\infty)$.

Kolejne miary dynamiki to różnica względna jednopodstawowa ($R^w_{r/s}$) ustalana według wzoru:

$$R^{w}_{r/s} = \frac{R^{a}_{r/s}}{y_s} \cdot 100 = \frac{y_r - y_s}{y_s} \cdot 100 \qquad 4.4.$$

i *różnica względna łańcuchowa* ($R^{w}_{r/r-1}$) obliczana według wzoru:

$$R^{w}_{r/r-1} = \frac{R^{a}_{r/r-1}}{y_{r-1}} \cdot 100 = \frac{y_r - y_{r-1}}{y_{r-1}} \cdot 100 \qquad 4.5.$$

Wymienione wyżej miary o charakterze względnym są miarami niemianowanymi i mogą być wyrażone w postaci dziesiętnej lub w procentach (należy wówczas – jak w podanych formułach – ich wartość pomnożyć przez 100). Wartość miary (może być dodatnia bądź ujemna lub zerowa) informuje, jaki był wzrost lub spadek poziomu badanego zjawiska w okresie (momencie) badanym w stosunku do porównawczego.

4.2.3. Ilorazowe miary dynamiki (indeksy)

W literaturze wyróżnia się dwa typy indeksów (miar ilorazowych): *indywidualne* (proste) i *agregatowe* (zespołowe). *Indeksy indywidualne* są wykorzystywane w analizie dynamiki zjawisk prostych i mogą mieć charakter miar jednopodstawowych (o podstawie stałej) bądź łańcuchowych (o podstawie zmiennej) [9]. *Indeks indywidualny o podstawie stałej* ($w^{i}_{r/s}$) obliczany jest według wzoru :

$$w^{i}_{r/s} = \frac{y_r}{y_s} \cdot 100 \qquad 4.6.$$

zaś *indeks indywidualny o podstawie zmiennej* ($w^{i}_{r/r-1}$) według formuły:

$$w^{i}_{r/r-1} = \frac{y_r}{y_{r-1}} \cdot 100 \qquad 4.7.$$

gdzie: oznaczenia jak we *wzorach 4.2 i 4.3.*

Podane miary mogą być wyrażone w postaci ułamków dziesiętnych lub w procentach (wymaga to pomnożenia indeksów – jak w podanych wzorach

[9] Należy tu zwrócić uwagę na sposób zapisu tych indeksów we wszelkiego rodzaju rocznikach; zapis "rok poprzedni = 100" oznacza, iż podane indeksy mają charakter miar łańcuchowych, zaś zapis " rok s = 100" oznacza, iż podane indeksy mają charakter miar jednopodstawowych , gdzie jako stałą bazę porównawczą przyjęto rok s.

- przez 100). W praktyce częściej wykorzystywana jest ta druga możliwość wyrażania wartości indeksów z uwagi na łatwość ich interpretacji.

Spośród wymienionych miar dynamiki w praktyce częściej wykorzystywane są indeksy, co ma m. in. niewątpliwy związek z większymi możliwościami ich wzajemnego przekształcania bez konieczności sięgania do danych pierwotnych. I tak:

a) jeśli zachodzi potrzeba zmiany bazy porównawczej w szeregu indeksów jednopodstawowych należy dokonać dzielenia wszystkich indeksów występujących w takim szeregu przez indeks okresu wybranego jako bazowy,
b) w celu przekształcenia szeregu indeksów o podstawie stałej w ciąg indeksów łańcuchowych indeks każdego z okresów dzielimy przez indeks okresu poprzedzającego,
c) dla przekształcenia szeregu indeksów o podstawie zmiennej w ciąg indeksów o podstawie stałej:
   - dla okresów „wcześniejszych" od okresu obranego za bazowy odpowiednie indeksy uzyskujemy jako odwrotność iloczynu wszystkich indeksów łańcuchowych począwszy od indeksu występującego bezpośrednio po badanym okresie do indeksu dla okresu bazowego,
   - dla okresu przyjętego za bazowy indeks wyniesie 1 lub 100 %,
   - dla okresów „późniejszych" od okresu bazowego odpowiednie indeksy uzyskujemy jako iloczyn wszystkich kolejnych indeksów łańcuchowych począwszy od indeksu okresu występującego bezpośrednio po bazowym na indeksie dla okresu badanego kończąc.

Indeksy agregatowe (zespołowe) są wykorzystywane w analizie dynamiki zbiorowości (zjawisk) złożonych. W takich przypadkach występuje zwykle niejednorodność jednostek tworzących zbiorowość, w związku z czym analiza jej rozwoju w czasie wymaga znalezienia wspólnej jednostki miary. Najczęściej taką wspólną jednostką miary (czynnikiem agregującym) jest cena, a jej wykorzystanie umożliwia zastosowanie indeksów wartościowych w analizie dynamiki takiej populacji bądź zjawiska. Wykorzystywane w takiej sytuacji indeksy mogą mieć oczywiście również charakter miar jednopodstawowych bądź łańcuchowych. W analizie zjawisk socjologicznych indeksy te nie mają w zasadzie zastosowania.

Procedurę ustalania wyżej omówionych miar dynamiki jak również istniejących możliwości ich wzajemnego przekształcania ilustrują poniższe przykłady.

Przykład 4.1.

Poniższe zestawienie zawiera informacje dotyczące liczby zarejestrowanych bezrobotnych w mieście „K" w latach 1995 - 2002 (stan na 31.12.)

Rok	*1995*	*1996*	*1997*	*1998*	*1999*	*2000*	*2001*	*2002*
Bezrobotni (w tys. osób)	15,5	14,9	14,5	14,5	15,1	15,4	15,7	16,2

Źródło: Dane umowne

Na podstawie powyższych danych zbadać dynamikę liczby bezrobotnych przy pomocy:

a) różnicowych miar dynamiki,

b) ilorazowych miar dynamiki.
Dokonać interpretacji ustalonych wartości miar.

Rozwiązanie

ad. a) Dla zbadania dynamiki wykorzystamy następujące miary dynamiki: różnicę absolutną o podstawie stałej *(wzór4.2*), różnicę absolutną o podstawie zmiennej *(wzór 4.3),* różnicę względną jednopodstawową *(wzór 4.4)* oraz różnicę względną łańcuchową *(wzór 4.5).* Obliczeń pomocniczych dokonano w poniższej tabeli roboczej w kolumnach 3 - 6; w przypadku miar jednopodstawowych jako bazę porównawczą przyjęto liczbę bezrobotnych w 1995 roku.

Rok	*Liczba bezrobotnych*	$R^{a}_{r/1995}$	$R^{a}_{r/r-1}$	$R^{w}_{r/1995}$ *(w %)*	$R^{w}_{r/r-1}$ *(w %)*	$w^{i}_{r/1995}$ *(w %)*	$w^{i}_{r/r-1}$ *(w %)*
1	*2*	*3*	*4*	*5*	*6*	*7*	*8*
1995	15,5	0	-	0	-	100	-
1996	14,9	- 0,6	- 0,6	- 3,87	- 3,87	96,13	96,13
1997	14,5	-1,0	- 0,4	- 6,45	-2,68	93,55	97,32
1998	14,5	- 1,0	0	- 6,45	0	93,55	100
1999	15,1	- 0,4	0,6	- 2,58	4,14	97,42	104,14
2000	15,4	- 0,1	0,3	- 0,65	1,99	99,35	101,99
2001	15,7	0,2	0,3	1,29	1,94	101,29	101,94
2002	16,2	0,7	0,5	4,52	3,18	104,52	103,18

Przykładowo zinterpretujemy wartości uzyskanych wyżej miar dynamiki dla roku 2000:

- $R^{a}_{2000/1995}$ = - 0,1 oznacza, że liczba bezrobotnych (według stanu na 31.12.) w roku 2000 była o 100 osób niższa niż w roku 1995;
- $R^{a}_{2000/1999}$ = 0,3 oznacza, iż liczba bezrobotnych na koniec roku 2000 była o 300 osób wyższa niż w roku 1999;
- $R^{w}_{2000/1995}$ = - 0,65 oznacza, iż liczba bezrobotnych w roku 2000 była o 0,65 % niższa w relacji do stanu na koniec 1995 roku;
- $R^{w}_{2000/1999}$ = 1,99 oznacza wzrost liczby bezrobotnych na koniec roku 2000 o 1,99 % w stosunku do roku poprzedniego.

ad. b) Wyznaczymy tu wartości indeksów indywidualnych jednopodstawowych (*wzór 4.6*) oraz indeksów indywidualnych o podstawie zmiennej (*wzór 4.7*). Obliczenia pomocnicze zawarte są w powyższej tablicy roboczej w kolumnach 7 i 8.

Zinterpretujemy wartości tych miar również dla roku 2000:

- $w^{i}_{2000/1995}$ = 99,35 % oznacza, że liczba bezrobotnych na koniec 2000 roku stanowiła 99,35 % ich stanu z roku 1995, co jednocześnie świadczy o spadku ich liczby o 0,65 % w stosunku do roku bazowego (por. $R^{w}_{2000/1995}$ = - 0,65);

- $w^{i}_{2000/1999}$ = 101,99 % oznacza, z liczba bezrobotnych na koniec roku 2000 stanowiła 101,99 % ich liczby z roku poprzedzającego, czyli była wyższa o 1,99 % (por. $R^{w}_{2000/1999}$ = 1,99).

Z uzyskanych wartości miar wynikają oczywiste zależności zachodzące między różnicą względną a indeksem indywidualnym:

$$R^{w}_{r/s} + 100 = w^{i}_{r/s}$$
$$R^{w}_{r/r-1} + 100 = w^{i}_{r/r-1}$$

Przykład 4.2.

Poniższa tablica zawiera informacje dotyczące dynamiki stopy bezrobocia w powiecie „L" w latach 1995 - 2001

Rok	*1995*	*1996*	*1997*	*1998*	*1999*	*2000*	*2001*
$R^{w}_{r/1996}$ *(w %)*	5,5	0	8,2	10,5	15,0	12,0	10,0

Źródło: Dane umowne

Podany szereg przekształcić w ciąg:
a) indeksów jednopodstawowych o podstawie: rok 1996 =100,
b) indeksów łańcuchowych,
c) indeksów jednopodstawowych o podstawie: rok 1999 = 100.

Rozwiązanie:

Przykład ten ilustruje możliwości ustalania różnych miar dynamiki w przypadku posiadania jakiegokolwiek ciągu miar dynamiki, bez konieczności „sięgania" do danych pierwotnych. W tym celu korzystamy tu z podanych wcześniej możliwości wzajemnego przekształcania miar dynamiki. Poszczególne przekształcenia zostaną wykonane w kolejnych kolumnach poniższej tabeli roboczej

Rok	$R^{w}_{r/1996}$ *(w %)*	$w^{i}_{r/1996}$ *(w %)*	$w^{i}_{r/r-1}$ *(w %)*	$w^{i}_{r/1999}$ *(w %)*
1	*2*	*3*	*4*	*5*
1995	5,5	105,5	-	91,74
1996	0	100,0	94,79	86,96
1997	8,2	108,2	108,20	94,09
1998	10,5	110,5	102,13	96,09
1999	15,0	115,0	104,07	100
2000	12,0	112,0	97,39	97,39
2001	10,0	110,0	98,21	95,65

ad. a) Podany szereg charakteryzuje dynamikę stopy bezrobocia za pomocą ciągu różnic względnych przy założonym roku bazowym 1996. Dla uzyskania ciągu indeksów jednopodstawowych przy tym samym roku bazowym dodajemy 100 do wartości różnic względnych, np. dla

roku 2000: 100 + 12,0 = 112,0 %; uzyskane wielkości zawarto w kolumnie 3,

ad. b) Indeksy łańcuchowe (kolumna 4) dla poszczególnych lat ustalono dzieląc indeks jednopodstawowy danego roku przez indeks roku poprzedzającego (z kolumny 3), np. dla roku 2000: $\frac{112,0}{115,0} \cdot 100 = 97,39\,\%$ (mnożenie tego ilorazu przez 100 daje wynik działania wyrażony w %); indeks łańcuchowy dla roku 1995 nie istnieje,

ad. c) Dla ustalenia indeksów o podstawie stałej na poziomie roku 1999 dokonujemy zamiany bazy porównawczej w ciągu indeksów z kolumny 3. Odbywa się to poprzez dzielenie indeksów dla każdego roku (z kolumny 3) przez indeks dla roku 1999 z tej kolumny, tj. przez115 %, np. dla roku 1995: $\frac{105,5}{115,0} \cdot 100 = 91,74\,\%$

4.3. METODY WYODRĘBNIANIA TENDENCJI ROZWOJOWEJ

4.3.1. Metoda oceny wzrokowej i mechaniczna

Badanie tendencji rozwojowej wiąże się z analizą rozwoju zjawiska w dłuższych przedziałach czasowych: kilkunasto-, kilkudziesięcioletnich. Dokonując takiej analizy należy zwrócić uwagę, iż rozwój większości zjawisk w czasie uwarunkowany jest oddziaływaniem wielu czynników (przyczyn). Można wśród nich wyróżnić przyczyny główne; efektem ich oddziaływania - w dłuższym okresie czasu – jest występowanie określonej *tendencji rozwojowej* (zwanej również *trendem).* Niezależnie od występowania przyczyn głównych należy odnotować oddziaływanie przyczyn ubocznych. Powodują one występowanie - obok wspomnianej tendencji rozwojowej - różnego rodzaju wahań. Wyodrębnia się wśród nich:

- wahania okresowe (cykliczne), czyli wahania regularne o określonym cyklu,
- wahania sezonowe, tzn. wahania okresowe o cyklu rocznym, miesięcznym lub dobowym,
- wahania nieregularne, wynikające np. ze zdarzeń katastrofalnych bądź oddziaływania znacznej liczby przyczyn ubocznych.

W literaturze wyróżnia się kilka metod wyznaczania *tendencji rozwojowej.* Najprostsza z nich to *metoda odręczna (metoda oceny wzrokowej).* Dla potrzeb tej metody informacje zawarte w szeregu czasowym przenosimy na układ współrzędnych i na podstawie uzyskanego rozrzutu punktów dokonujemy oceny charakteru tendencji rozwojowej. Odbywa się to poprzez odręczne wykreślenie linii przebiegającej między naniesionymi punktami w ten sposób by w przybliżeniu była ona najmniej oddalona od naniesionych punktów. Tak wykreślona linia pomija oczywiście występujące w poszczególnych okresach wahania. Metoda ta pozwala na efektywne określenie charakteru tendencji w przypadkach, gdy punkty obrazujące poziom zjawi-

ska w poszczególnych okresach układają się w niezbyt szerokim „pasie". Najczęściej jednak metoda ta stanowi jedynie etap wstępny analitycznej metody wyznaczania tendencji rozwojowej.

Metoda mechaniczna wyodrębniania tendencji rozwojowej (zwana również metodą wygładzania szeregów czasowych) polega na zastępowaniu danych empirycznych występujących w szeregu wartościami średnimi. Występuje ona w dwóch odmianach. Pierwszą z nich określamy mianem *metody mechanicznej opartej na średnich ruchomych* liczonych dla określonej liczby okresów. Taka procedura wyodrębniania tendencji rozwojowej prowadzi do wygładzenia pierwotnego szeregu czasowego, ponieważ występujące wahania rozkładają się na odpowiednią liczbę okresów. Średnie takie mogą być one liczone dla danych z nieparzystej lub parzystej liczby okresów (w drugim przypadku średnie noszą nazwę centrowanych). W przypadku średnich liczonych z nieparzystej liczby okresów występuje łatwość ich przyporządkowania konkretnemu okresowi środkowemu, co powoduje, iż są one częściej wykorzystywane. Mogą to być średnie 3-, 5-, 7- okresowe, przy czym należy zwrócić uwagę, że „wygładzenie" szeregu będzie tym większe im średnia jest liczona z większej liczby wyrazów. Średnie te mają charakter średniej arytmetycznej i są liczonej według następujących wzorów

- *średnie ruchome 3-okresowe*:

$$\bar{y}^{(3)}{}_2 = \frac{y_1 + y_2 + y_3}{3}; \bar{y}^{(3)}{}_3 = \frac{y_2 + y_3 + y_4}{3};\ldots\ldots, \bar{y}^{(3)}{}_{n-1} = \frac{y_{n-2} + y_{n-1} + y_n}{3} \qquad 4.8.$$

Zauważmy, że pierwsza średnia ($\bar{y}^{(3)}{}_2$) jest przyporządkowana okresowi drugiemu, druga ($\bar{y}^{(3)}{}_3$)- trzeciemu, a ostatnia ($\bar{y}^{(3)}{}_{n-1}$) – przedostatniemu,

- *średnie 5-okresowe:*

$$\bar{y}_{3..}{}^{(5)} = \frac{y_1 + y_2 + y_3 + y_4 + y_5}{5}\ldots;\ldots \bar{y}_4{}^{(5)} = \frac{y_2 + y_3 + y_4 + y_5 + y_6}{5}$$

$$\bar{y}_{n-2}{}^{(5)} = \frac{y_{n-4} + y_{n-3} + y_{n-2} + y_{n-1} + y_n}{5} \qquad 4.9.$$

W tym przypadku pierwsza średnia ($\bar{y}^{(5)}{}_3$) jest przyporządkowana okresowi trzeciemu, druga ($\bar{y}^{(5)}{}_4$) – czwartemu, a ostatnia ($\bar{y}^{(5)}{}_{n-2}$) - trzeciemu od końca

Średnie ruchome dla większej liczby okresów będą liczone według podobnych zasad.

Obok zalety metody mechanicznej - tj. prostoty obliczeń - należy zwrócić uwagę na istotny jej mankament polegający na „skracaniu" szeregu czasowego. Im średnia jest wyznaczana dla większej liczby okresów, tym większemu skróceniu ulega szereg, ale jednocześnie uzyskujemy jego większe

„wygładzenie"; w przypadku średniej 3-okresowej tracimy informacje dla dwóch skrajnych okresów, przy 5-okresowej - dla czterech, 7-okresowej aż dla sześciu okresów. Długość okresu średniej ruchomej zależy głównie od długości szeregu czasowego. Powoduje to ograniczoną praktyczną przydatność tej metody, zwłaszcza w przypadku krótszych szeregów czasowych i dużej amplitudy wahań.

Druga odmiana *metody mechanicznej oparta* jest *na średnich podokresów*. Polega ona na wyznaczeniu średniego poziomu badanego zjawiska dla wyodrębnionych z szeregu czasowego dwóch podokresów (w miarę możliwości równych) i naniesieniu uzyskanych średnich na układ współrzędnych, a następnie wykreśleniu prostej przebiegającej przez te punkty. Taki sposób postępowania pozwala na jednoznaczne określenie charakteru tendencji rozwojowej. Metoda ta może być stosowana, gdy metoda oparta na średnich ruchomych nie pozwala wyodrębnić tendencji rozwojowej z uwagi na zbyt duże wahania poziomu zjawiska, a stosunkowo krótki szereg czasowy nie pozwala na wyznaczanie średnich ruchomych z dużej liczby okresów.

4.3.2. Metoda analityczna i badanie średniego tempa zmian

Istotą tej metody jest „dopasowanie" funkcji matematycznej (określanej w tym przypadku funkcją trendu) do ciągu danych ujętych w szeregu czasowym. Właściwe szacowanie parametrów takiej funkcji musi być poprzedzone doborem odpowiedniej klasy funkcji trendu. W tym celu można wykorzystać metodę oceny wzrokowej. Zalecany jest dobór prostej postaci funkcji o parametrach, którym można nadać logiczną interpretację. Wyznaczaną na podstawie danych ujętych w szeregu czasowym (przy założeniu liniowego charakteru tendencji rozwojowej) ogólną postać takiej funkcji trendu można zapisać następująco:

$$\hat{y}_t = a \cdot t + b \qquad 4.10.$$

gdzie: $\hat{y}_t$ - oszacowane teoretyczne, hipotetyczne wartości poziomu badanego zjawiska dla poszczególnych okresów (momentów),
t - zmienna czasowa,
a, b - oszacowane parametry funkcji trendu (b - oznacza poziom zjawiska w okresie „zerowym", zaś a - oznacza przeciętny przyrost bądź spadek poziomu zjawiska przypadający na jednostkę czasu).

Łatwo zauważyć, iż szacowane równanie trendu można traktować jako szczególny przypadek równania regresji, w którym rolę zmiennej niezależnej pełni czas (jej wartościami są kolejne numery okresów).

Parametry tego równania szacowane są według poniższych wzorów:

$$a = \frac{\sum_{i=1}^{n} (t_i - \bar{t}) \cdot y_{t_i}}{\sum_{i=1}^{n} (t_i - \bar{t})^2} \qquad 4.11.$$

$$b = \bar{y} - a \cdot \bar{t} \qquad 4.12.$$

W przypadku liniowej funkcji trendu jej obrazem graficznym jest linia prosta, co może powodować, że punkty na niej położone mogą dość istotnie odbiegać od danych empirycznych występujących w szeregu czasowym. W związku z powyższym dość często badana jest zgodność oszacowanych wielkości hipotetycznych wynikających z równania trendu z danymi empirycznymi. Do tego celu wykorzystywany jest poznany w rozdz. III współczynnik zbieżności, który po odpowiedniej modyfikacji przyjmie postać:

$$\varphi^2 = \frac{\sum_{i=1}^{n}(y_{t_i} - \hat{y}_{t_i})^2}{\sum_{i=1}^{n}(y_{t_i} - \bar{y})^2} \qquad 4.13.$$

Mierzy on poziom „niedopasowania" funkcji trendu do danych empirycznych, w związku z czym malejąca jego wartość oznacza wyższy poziom zgodności danych empirycznych i teoretycznych.

Charakter tendencji rozwojowej można również badać określając *przeciętne tempo zmian poziomu badanego zjawiska*. Jest ono wyznaczane jako średnia geometryczna ciągu indeksów łańcuchowych według wzoru:

$$\overline{w}^{i}_{r/r-1} = \sqrt[n-1]{w^{j}_{2/1} \cdot w^{j}_{3/2} \cdot w^{j}_{4/3} \cdot \ldots \cdot w^{j}_{n/n-1}} = \sqrt[n-1]{w^{j}_{n/1}} \qquad 4.14.$$

Korzystając z relacji zachodzących między ciągami indeksów łańcuchowych i jednopodstawowych zamiast iloczynu wszystkich indeksów łańcuchowych można w powyższej formule wykorzystać indeks jednopodstawowy okresu ostatniego w stosunku do okresu pierwszego. Wskazane jest tu wykorzystywanie indeksów wyrażonych w postaci dziesiętnej. Trudności pojawiają się w przypadku wyznaczania średniego tempa dla długiego szeregu czasowego; zachodzi wówczas konieczność określenia wartości pierwiastka wysokiego stopnia. W takiej sytuacji wykorzystuje się postać logarytmiczną podanego wyżej wzoru:

$$\log \overline{w}^{i}_{r/r-1} = \frac{1}{n-1}\sum_{i=2}^{n} \log w^{i}_{r/r-1} = \frac{1}{n-1} \log w^{i}_{n/1} \qquad 4.15.$$

Procedurę wykorzystania omówionych wyżej metod do wyodrębniania tendencji rozwojowej ilustruje poniższy przykład 4.3.

Przykład 4.3.

W pewnym badaniu statystycznym obejmującym obszar województwa „K" zebrano m. in. informacje dotyczące kształtowania się wskaźnika rozwodów (na 1000 mieszkańców) w latach 1991 – 2000. Informacje te ujęto w poniższej tablicy:

Rok	*1991*	*1992*	*1993*	*1994*	*1995*	*1996*	*1997*	*1998*	*1999*	*2000*
Wskaźnik rozwodów	2,2	2,5	2,0	2,3	2,8	3,2	3,5	3,9	4,2	4,4

Źródło: Dane umowne

Określić tendencję rozwojową wskaźnika rozwodów w badanym okresie:
a) metodą średnich ruchomych,
b) wyznaczając równanie trendu,
c) określając średnie tempo zmian.

Rozwiązanie:

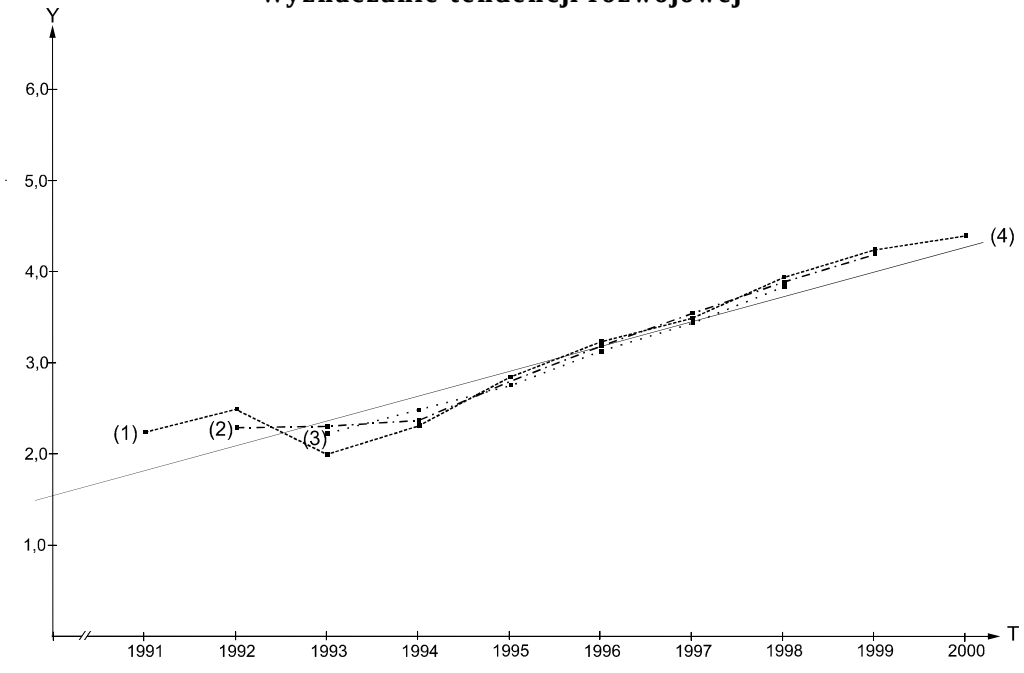
ad. a) Dla wykonania obliczeń pomocniczych przekształcamy powyższy szereg w tablicę dwukolumnową, w której w kolumnach 4 i 5 wyznaczymy odpowiednie średnie ruchome; przyjmujemy, iż będą to średnie 3 i 5-okresowe.

Rok	*Nr okresu* (t_i)	*Wskaźnik rozwodów* (y_i)	*Średnie 3-okresowe*	*Średnie 5-okresowe*	$t_i - \bar{t}$	$(t_i - \bar{t})^2$	$(t_i - \bar{t})y_{t_i}$	$w^i_{r/r-1}$
1	*2*	*3*	*4*	*5*	*6*	*7*	*8*	*9*
1991	1	2,2	-	-	- 4,5	20,25	- 9,9	-
1992	2	2,5	2,23	-	- 3,5	12,25	- 8,75	1,14
1993	3	2,0	2,27	2,36	- 2,5	6,25	- 5,0	0,80
1994	4	2,3	2,37	2,56	- 1,5	2,25	- 3,45	1,15
1995	5	2,8	2,77	2,76	- 0,5	0,25	- 1,4	1,22
1996	6	3,2	3,17	3,14	0,5	0,25	1,6	1,14
1997	7	3,5	3,53	3,52	1,5	2,25	5,25	1,09
1998	8	3,9	3,87	3,84	2,5	6,25	9,75	1,11
1999	9	4,2	4,17	-	3,5	12,25	14,7	1,08
2000	10	4,4	-	-	4,5	20,25	19,8	1,05
Razem	55	31	X	X	X	82,5	22,6	X

Wielkości podane w kolumnach 3, 4 i 5 nanosimy na układ współrzędnych. Linia łamana powstała z połączenia punktów obrazujących wielkości empiryczne (dane z kolumny 3) może służyć do oceny wzrokowej charakteru tendencji rozwojowej. Na jej podstawie można stwierdzić, iż wielkość wskaźnika rozwodów w badanym okresie wykazuje tendencję rosnącą (por. rys. 4.1; linia 1)

Rys. 4.1

Wyznaczanie tendencji rozwojowej



Źródło: opracowanie własne

Linia przebiegająca przez punkty wyznaczone przez średnie trzyokresowe (dane z kolumny 4) ma kształt bardziej „wygładzony" w stosunku do linii empirycznej. „Wygładzenie" to jednak odbyło się „kosztem" skrócenia szeregu czasowego o informacje dla lat 1991 i 2000 (patrz rys. 4.1, linia 2).

Linia, której przebieg wyznaczyły średnie 5-okresowe charakteryzuje się dalszym „wygładzeniem" przy kolejnej stracie informacji dla dwóch kolejnych okresów. Wszystkie one potwierdzają jednak występowanie tendencji rosnącej wskaźnika rozwodów na 1000 mieszkańców na badanym obszarze (por. rys. 4.1, linia 3).

ad. b) Wyznaczamy równanie trendu o *postaci 4.10*, a parametry tego równania według *wzorów 4.11 i 4.12*. Obliczenia pomocnicze wykonano w powyższej tablicy roboczej w kolumnach 6 -8. Wartości średnie dla badanych zmiennych wynoszą:

- dla zmiennej czasowej T: $\bar{t}$ = 5,5;
- dla zmiennej Y, tj. wskaźnika rozwodów: $\bar{y}$ = 3,1

Wstawiając odpowiednie wielkości uzyskane w tabeli roboczej do *wzorów 4.11. i 4.12* uzyskujemy następujące wartości parametrów równania trendu:

$$a = \frac{22,6}{82,5} = 0,27 \text{ (rozwodów na 1000 mieszkańców/ rok)}$$

$b = 3{,}1 - 0{,}27 \cdot 5{,}5 = 1{,}59$ (rozwodów na 1000 mieszkańców)
i w konsekwencji równanie trendu będzie miało postać:

$$\hat{y}_t = 0{,}27t + 1{,}59$$

Z oszacowanej funkcji wynika, że w badanym okresie wskaźnik rozwodów wykazywał tendencję rosnącą (parametr *a* jest dodatni) i wzrastał on średniorocznie o 0,27, zaś jego wielkość w roku „zerowym", tj. 1990 wynosiła 1,59. Obraz graficzny oszacowanej tendencji rozwojowej zaprezentowano również na powyższym wykresie (por. rys. 4.1, linia .4).

ad. c) W celu określenia średniorocznego tempa zmian wielkości wskaźnika rozwodów dokonamy wyznaczenia średniego indeksu łańcuchowego (indeksy te obliczono w tabeli roboczej w kolumnie 9). Konieczne jest zatem obliczenie średniej geometrycznej ustalonego ciągu indeksów łańcuchowych. Należy tu zauważyć, iż ich iloczyn jest identyczny jak indeks dla roku 2000 liczony w stosunku do roku 1991. Postępując zgodnie z *formułą 4.14* otrzymujemy:

$$\overline{w}^{i}_{r/r-1} = \sqrt[9]{1{,}14 \cdot 0{,}90 \cdot 1{,}15 \cdot 1{,}22 \cdot 1{,}14 \cdot 1{,}09 \cdot 1{,}11 \cdot 1{,}08 \cdot 1{,}05} = \sqrt[9]{2{,}0} = 1{,}079$$

Uzyskana wielkość oznacza:

- po pierwsze, z uwagi na uzyskaną wartość średniego indeksu łańcuchowego większą od jedności, iż wskaźnik rozwodów w badanym okresie wykazywał tendencję rosnącą,
- po drugie, średnie roczne tempo wzrostu wskaźnika rozwodów w badanym okresie wynosiło 7,9 %.

4.4. ANALIZA SEZONOWOŚCI

Badanie sezonowości stanowi - obok wyznaczania tendencji rozwojowej - kolejne istotne zagadnienie z zakresu analizy szeregów czasowych. Zadaniem analizy w tym przypadku jest *wyodrębnienie wahań sezonowych* w takim szeregu oraz określenie ich natężenia. Jest to zagadnienie szczególnie ważne w przypadku zjawisk społeczno-ekonomicznych, bowiem podlegają one oddziaływaniu wielu czynników wywołujących nie tylko ujawnienie się określonej tendencji rozwojowej, ale również różnego rodzaju wahań, w tym sezonowych. *Wahania sezonowe* są szczególnym przypadkiem wahań okresowych, najczęściej o rocznym cyklu wahań. Można wśród nich wyróżnić wahania o charakterze *miesięcznym, kwartalnym bądź półrocznym*. Przyczyną tego typu wahań może być cykliczny ruch Ziemi wokół Słońca i związane z tym występowanie pór roku. Można podać przykłady wielu zjawisk charakteryzujących się wahaniami sezonowymi związanymi z występowaniem pór roku, np. produkcja roślinna, zapotrzebowanie na określony typ odzieży (letnia, zimowa), popyt na napoje chłodzące, popyt na określony typ

sprzętu sportowego (zimowy, letni). Podłoże sezonowości mogą stanowić również czynniki o charakterze instytucjonalnym bądź prawnym nakładające wykonywanie określonych czynności w narzuconych terminach, np. organizacja roku szkolnego i związany z tym popyt na podręczniki i artykuły papiernicze, konieczność sporządzania niektórych sprawozdań i rozliczeń w ściśle określonych cyklach (miesięcznych, kwartalnych, rocznych). Następstwa występowania wahań sezonowych są na ogół negatywne. Powodują one np. nierytmiczny przebieg procesów produkcyjnych, nierównomierne wykorzystanie czynników produkcji, nierównomierne – w ciągu roku – wykorzystanie bazy dydaktycznej uczelni, szkół. W każdym przypadku skutkiem są ponoszone dodatkowe koszty działalności (związane np. z koniecznością gromadzenia zapasów środków rzeczowych bądź przechowywania wyrobów gotowych, utrzymania zbędnej bazy).

Prowadzona analiza sezonowości pozwala na jej identyfikację, określenie skali tego zjawiska i w konsekwencji podjęcie działań, których celem będzie przynajmniej częściowe złagodzenie negatywnych jej następstw np. poprzez właściwe planowanie dostaw surowców lub zbytu wyrobów gotowych, uwzględnienie zatrudnienia ewentualnych pracowników sezonowych, planowanie gospodarki remontowej. Skalę sezonowości ustala się poprzez wyznaczanie *wskaźników sezonowości*, które najczęściej obrazują odchylenia poziomu badanego zjawiska w poszczególnych wyodrębnionych okresach od poziomu średniego.

Wahania sezonowe mogą być wyodrębniane różnymi metodami. Najprostsza z nich to *metoda oparta na średnich okresów jednoimiennych***.** Dla potrzeb tej metody analizy sezonowości szereg czasowy musi ujmować materiał empiryczny w układzie umożliwiającym badanie określonego typu sezonowości (np. w układzie miesięcznym, kwartalnym bądź półrocznym). Wskaźnik sezonowości dla j-tego okresu (S_j) – w przypadku tej metody - wyznaczany jest według wzoru:

$$S_j = \frac{\bar{y}_j}{\bar{y}} \cdot 100 \qquad 4.16.$$

gdzie :

$\bar{y}_j$ - średni poziom zjawiska dla j-tego okresu,

$\bar{y}$ - średni poziom zjawiska dla wszystkich okresów.

Wskaźnik ten najczęściej wyrażany jest w %, co znacznie ułatwia jego interpretację, bowiem informuje on jak - przeciętnie biorąc - kształtował się poziom zjawiska w określonym okresie w stosunku do poziomu średniego. Metoda ta jest zalecana w przypadku, gdy badane zjawisko (a tym samym poziom wahań) nie wykazuje określonej tendencji rozwojowej.

Kolejna *metoda oparta jest na wyznaczaniu średnich ruchomych centrowanych***.** Zastosowanie średnich ruchomych powoduje „wygładzenie" wahań przypadkowych w szeregu czasowym, przy czym typ średniej ruchomej zdeterminowany jest charakterem badanej sezonowości; i tak w przypadku badania sezonowości miesięcznej wyznaczamy średnie 12-okresowe, kwartal-

nej - średnie 4-okresowe. Wyznaczone średnie posiadają ten mankament, iż istnieje problem z ich przyporządkowaniem konkretnemu okresowi z uwagi sposób ich liczenia (z parzystej liczby wyrazów szeregu czasowego). Tę niedogodność eliminuje się wyznaczając średnie ruchome centrowane 2-okresowe z wyznaczonych uprzednio średnich ruchomych. Centrowanie to powoduje przyporządkowanie uzyskanych w ten sposób średnich konkretnym wartościom empirycznym występującym w szeregu czasowym, a to z kolei umożliwia wyznaczenie wskaźników sezonowości według poniższego wzoru:

$$S_j = \frac{\sum_j y_{t_i}^{(j)}}{\sum_j \bar{y}_{c_i}^{(j)}} \cdot 100 \qquad 4.17.$$

gdzie:

$\sum_j y_{t_i}^{(j)}$ - oznacza sumę wartości empirycznych dla j-tego okresu,

$\sum_j \bar{y}_{c_i}^{(j)}$ - oznacza sumę średnich ruchomych centrowanych dla j-tego okresu.

Tak liczony wskaźnik sezonowości jest zatem ilorazem sumy wartości empirycznych i sumy średnich ruchomych centrowanych dla badanego okresu.

Należy tu zwrócić uwagę, iż podobnie jak w przypadku metody mechanicznej wyodrębniania tendencji rozwojowej zastępowanie wartości empirycznych średnimi ruchomymi powoduje „skracanie" szeregu informacji. W przypadku średnich 4-okresowych strata dotyczy dwóch pierwszych i dwóch ostatnich okresów, w przypadku średnich 12-okresowych - sześciu pierwszych i sześciu ostatnich okresów. Z uwagi na sposób liczenia wskaźnika sezonowości wielkości występujące w liczniku tej formuły muszą pomijać wartości y_{t_i}, dla których nie jest możliwe przyporządkowanie średnich ruchomych centrowanych. Wartości ustalonych wskaźników interpretowane są podobnie jak w przypadku pierwszej metody. Metodę tę zaleca się stosować, gdy ujęte w szeregu czasowym zjawisko charakteryzuje się określoną tendencją rozwojową.

Kolejna *metoda oparta jest na wykorzystaniu równania trendu* oszacowanego dla szeregu czasowego, dla którego będzie badana sezonowość. Procedura tej metody wymaga - w pierwszej fazie – oszacowania równania trendu dla analizowanego szeregu czasowego według zasad poznanych w *podrozdziale 4.3*. Następnie na podstawie oszacowanego równania wyznaczane są wartości teoretyczne ($\hat{y}_{t_i}$) dla każdego z wyodrębnionych okresów (t_i). Konstrukcja wskaźnika sezonowości zakłada porównywanie uzyskanych wartości teoretycznych ($\hat{y}_{t_i}$) z odpowiadającymi im wielkościami empirycznymi przy wykorzystaniu do tego celu następującej postaci wskaźnika sezonowości:

$$S_j = \frac{\sum_j y_{t_i}^{(j)}}{\sum_j \hat{y}_{t_i}^{(j)}} \cdot 100 \qquad 4.18.$$

gdzie:

$\sum_j y_{t_i}^{(j)}$ - suma wartości empirycznych dla j-tego okresu,

$\sum_j \hat{y}_{t_i}^{(j)}$ - suma wartości teoretycznych dla j–tego okresu oszacowanych za pomocą równania trendu

Wskaźnik sezonowości stanowi w tym przypadku iloraz sum wartości empirycznych i oszacowanych wartości teoretycznych dla badanego okresu. Interpretacja uzyskanych wskaźników może mieć podobny charakter jak wyżej.

Podsumowując należy wskazać na jeszcze jeden problem związany z badaniem sezonowości. Dotyczy on tzw. „oczyszczania" wskaźników sezonowości. Uzyskiwane wartości wskaźników sezonowości winny spełniać własność, którą można wyrazić następująco:

suma wskaźników sezonowości winna być równa iloczynowi liczby wyodrębnionych okresów i liczby 100.

W praktyce często zdarza się, iż warunek ten nie jest spełniony; wyznaczone wówczas wskaźniki traktujemy jako „surowe", które będą podlegały procedurze „oczyszczenia". W tym celu wyznaczany jest wskaźnik korygujący o postaci:

$$k = \frac{\sum_j S_j}{r \cdot 100} \qquad 4.19.$$

gdzie:

S_j - „surowe" wskaźniki sezonowości dla wyodrębnionych okresów,

r - liczba wyodrębnionych okresów.

Następnie dokonujemy korekty „surowych" wskaźników dzieląc każdy z nich przez ustalony wskaźnik korygujący.

Procedurę badania sezonowości przy wykorzystaniu omówionych metod ilustruje poniższy przykład.

Przykład 4.4.

W trakcie analizy zatrudnienia w firmie „Z" w Warszawie dokonano m. in. zestawienia liczby zatrudnionych w ujęciu kwartalnym w latach 1996 - 2000. Uzyskano następujące dane (przeciętne zatrudnienie w poszczególnych kwartałach):

Lata	*Kwartały*			
	I	*II*	*III*	*IV*
1996	171	191	297	223
1997	160	174	276	214
1998	154	168	249	202
1999	146	158	221	192
2000	143	148	213	186

Źródło: Dane umowne

Na podstawie powyższych danych zbadać, czy poziom zatrudnienia w tej firmie charakteryzuje się sezonowością kwartalną i ewentualnie określić jej skalę posługując się metodami opartymi na:
a) średnich okresów jednoimiennych,
b) średnich ruchomych centrowanych,
c) równaniu trendu.

Dokonać interpretacji uzyskanych wskaźników sezonowości. W przypadku konieczności dokonać „oczyszczenia" ustalonych wskaźników.

Rozwiązanie:

ad. a) Dla potrzeb tej metody wyznaczamy średnie dla poszczególnych kwartałów $(\bar{y}_j)$ oraz średnią liczoną dla wszystkich kwartałów $(\bar{y})$. Obliczenia pomocnicze wykonano w poniższej tabeli roboczej:

Lata	*Kwartały*			
	I	*II*	*III*	*IV*
1996	171	191	297	223
1997	160	174	276	214
1998	154	168	249	202
1999	146	158	221	192
2000	143	148	213	186
Suma	774	839	1256	1017
$\bar{y}_j$	154,8	167,8	251,2	203,4

Średnia liczona dla wszystkich kwartałów wynosi:

$$\bar{y} = \frac{154{,}8 + 167{,}8 + 251{,}2 + 203{,}4}{4} = 194{,}3$$

Zróżnicowane średnie dla poszczególnych kwartałów wskazują na występowanie sezonowości kwartalnej w poziomie zatrudnienia pracowników w badanej firmie. Skalę tego zjawiska określimy wyznaczając wskaźniki sezonowości dla poszczególnych kwartałów zgodnie z *wzorem 4.16:*

$$S_I = \frac{154,8}{194,3} \cdot 100 = 79,67\%$$

$$S_{II} = \frac{167,8}{194,3} \cdot 100 = 86,36\%$$

$$S_{III} = \frac{251,2}{194,3} \cdot 100 = 129,28\%$$

$$S_{IV} = \frac{203,4}{194,3} \cdot 100 = 104,68\%$$

Uzyskane wielkości wskaźników interpretujemy następująco:

- w kwartale pierwszym (przeciętny) poziom zatrudnienia kształtował się zwykle około 20 % poniżej średniego poziomu dla całego roku,
- w kwartale drugim poziom zatrudnienia kształtował się około 14 % poniżej poziomu średniego,
- w kwartale trzecim przeciętne zatrudnienie kształtowało się około 30 % powyżej poziomu średniego,
- w kwartale czwartym liczba zatrudnionych kształtowała się około 5 % powyżej średniej rocznej.

Suma obliczonych wskaźników sezonowości wynosi 399,99 %, tj. w przybliżeniu 400 %, a więc nie zachodzi konieczność ich korygowania.

ad. b) Dla potrzeb tej metody wyznaczamy - zgodnie z jej regułami - średnie ruchome 4-okresowe, a następnie poddajemy je centrowaniu. Dokonano tego w poniższej tablicy roboczej. w kolumnach 4 i 5.

Wykorzystując *wzór 4.17* wyznaczamy wskaźniki sezonowości dla poszczególnych kwartałów (dla kwartału I podano szczegółowo sposób liczenia wskaźnika sezonowości):

$$S_I = \frac{160+154+146+143}{210,875+199,625+185,25+175} \cdot 100 = \frac{603}{770,75} \cdot 100 = 78,24\%$$

$$S_{II} = \frac{648}{755,625} \cdot 100 = 85,76\%$$

$$S_{III} = \frac{1043}{795,5} \cdot 100 = 131,11\%$$

$$S_{IV} = \frac{831}{786,625} \cdot 100 = 105,64\%$$

Rok/ kwartał	*Nr okresu* (t_i)	*Zatrudnienie* (y_{t_i})	*Średnie ruchome 4-okresowe*	*Średnie ruchome 4-okresowe centrowane*	$t_i - \bar{t}$	$(t_i - \bar{t})^2$	$(t_i - \bar{t})y_{t_i}$	$\hat{y}_{t_i}$
1	*2*	*3*	*4*	*5*	*6*	*7*	*8*	*9*
1996/ I	1	171		-	- 9,5	90,25	- 1624,5	216,65
II	2	191	220,5	-	- 8,5	72,25	- 1623,5	214,3
III	3	297	217,75	219,125	- 7,5	56,25	- 2227,5	211,95
IV	4	223	213,5	215,625	- 6,5	42,25	- 1449,5	209,6
1997/ I	5	160	208,25	210,875	- 5,5	30,25	- 880	207,25
II	6	174	206	207,125	- 4,5	20,25	- 1749	204,9
III	7	276	204,5	205,25	-3,5	12,25	- 966	202,55
IV	8	214	203	203,75	- 2,5	6,25	- 535	200,2
1998/ I	9	154	196,25	199,625	- 1,5	2,25	- 231	197,85
II	10	168	193,25	194,75	- 0,5	0,25	- 84	195,5
III	11	249	191,25	192,25	0,5	0,25	124,5	193,15
IV	12	202	188,75	190,00	1,5	2,25	303	190,8
1999/ I	13	146	181,75	185,25	2,5	6,25	365	188,45
II	14	158	179,25	180,50	3,5	12,25	553	186,1
III	15	221	178,5	178,875	4,5	20,25	994,5	183,75
IV	16	192	176	177,25	5,5	30,25	1056	181,4
2000/ I	17	143	174	175,00	6,5	42,25	929,5	179,05
II	18	148	172,5	173,25	7,5	56,25	1110	176,7
III	19	213		-	8,5	72,25	1810,5	174,35
IV	20	186		-	9,5	90,25	1767	172
Suma	X	X	X	X	X	665	- 1565	X

Suma wyznaczonych wskaźników sezonowości wynosi 400,72 %, czyli nieznacznie przekracza wielkość teoretyczną. Oznacza to, iż w zasadzie nie jest konieczne dokonywanie korekty uzyskanych wskaźników. Jednak dla ilustracji tej procedury dokonamy ich korygowania. Wskaźnik korygujący ustalony zgodnie z *wzorem 4.19* wynosi:

$$k = \frac{400,72}{4 \cdot 100} = 1,001875$$

Przy jego pomocy dokonujemy korekty otrzymanych „surowych" wskaźników sezonowości:

$$S_I = \frac{78,24}{1,001875} = 78,09\%$$

$$S_{II} = \frac{85,76}{1,001875} = 85,60\%$$

$$S_{III} = \frac{131,11}{1,001875} = 130,86\%$$

$$S_{IV} = \frac{105,64}{1,001875} = 105,44\%$$

Interpretacja uzyskanych skorygowanych wskaźników odbywa się podobnie jak w przypadku poprzedniej metody.

ad. c) W celu zastosowania tej metody konieczne jest oszacowanie równania trendu o *postaci 4.10.* dla analizowanego szeregu czasowego okresów, gdzie okresami są poszczególne kwartały. Obliczenia pomocnicze dla określenia parametrów równania zgodnie z *wzorami 4.11 i 4.12* zawarto w powyższej tablicy roboczej w kolumnach 6 - 8. Wartości średnie wynoszą:

- dla zmiennej czasowej T : $\bar{t} = 10,5$,
- dla zmiennej Y: $\bar{y} = 194,3$.

Parametry równania wyznaczane według podanych wzorów będą wynosiły:

$$a = \frac{-1565}{665} = -2,35$$
$$b = 194,3 + 2,35 \cdot 10,5 = 219$$

a równanie trendu przyjmie postać

$$\hat{y}_t = -2,35t + 219$$

Wykorzystując oszacowane równanie wyznaczamy wartości teoretyczne ($\hat{y}_{t_i}$) dla poszczególnych okresów podstawiając do równania kolejne ich numery (por. kolumna 9 tabeli roboczej). Po podstawieniu do *wzoru 4.18* odpowiednich wartości empirycznych i odpowiadających im wartości teoretycznych otrzymujemy wskaźniki sezonowości dla poszczególnych kwartałów (dla kwartału I podano szczegółowo sposób wyznaczania wskaźnika):

$$S_I = \frac{171+160+154+146+143}{216{,}65+207{,}25+197{,}85+188{,}45+179{,}05}\cdot 100 = \frac{774 =}{989{,}25}\cdot 100 = 78{,}24\%$$

$$S_{Ii} = \frac{839}{977{,}5}\cdot 100 = 85{,}83\%$$

$$S_{III} = \frac{1256}{965{,}75}\cdot 100 = 130{,}05\%$$

$$S_{IV} = \frac{1017}{954}\cdot 100 = 106{,}60\%$$

Suma uzyskanych wskaźników nieznacznie przekracza wartość teoretyczną (wynosi 400,72 %), a więc nie zachodzi potrzeba ich korygowania.

Na podstawie przeprowadzonych badań uzyskano podobne wielkości wskaźników sezonowości przy wykorzystaniu wszystkich trzech metod ich wyodrębniania.

Rozdział V

Wnioskowanie statystyczne

5.1. ISTOTA I METODY WNIOSKOWANIA STATYSTYCZNEGO

Metody wnioskowania o całej zbiorowości statystycznej na podstawie informacji zebranych w trakcie badania próby statystycznej (reprezentacyjnej) są przedmiotem teorii statystyki matematycznej. Proponuje ona metody, które takie wnioskowanie umożliwiają. Wnioskowanie to może dotyczyć:

1) oceny, do jakiej klasy należy rozkład badanej zmiennej,
2) wartości parametrów badanej zmiennej w populacji generalnej,
3) występowania niezależności bądź zależności określonych zmiennych.

Podejście klasyczne wyróżnia dwie metody wnioskowania:

- *teorię szacowania (teorię estymacji*), która daje przede wszystkim możliwość szacowania określonych parametrów populacji generalnej na podstawie próby. W tym przypadku przystępując do wnioskowania nie posiadamy żadnych pochodzących spoza próby informacji o populacji generalnej, które mogłyby znacznie uściślić wyniki wnioskowania; w rezultacie jako wynik wnioskowania przyjmujemy taki rozkład lub takie wartości parametrów, które są najbardziej naturalne dla analizowanej próby statystycznej.
- *teorię weryfikacji hipotez statystycznych*, gdzie statystyk z góry formułuje pewne przypuszczenia dotyczące określonych własności populacji generalnej (np. typu rozkładu zmiennych, wartości parametrów, niezależności zmiennych). Metody statystyki matematycznej umożliwiają na podstawie wyników z próby wnioskowanie o przyjęciu bądź odrzuceniu tych przypuszczeń.

Z powyższego wynika, iż omawiane metody są ściśle związane z badaniami częściowymi. Okoliczności, w jakich takie badania są prowadzone, zostały wskazane w rozdziale I.

Teoria wnioskowania statystycznego opiera się na grupie twierdzeń zwanych granicznymi. Podstawowe wśród nich to *twierdzenie Lindeberga-Levy`ego: Jeśli zmienne losowe* $X_1, X_2, ..., X_n$ *są niezależne i posiadają jed-*

nakowy rozkład z wartością oczekiwaną E(x) i wariancją $\sigma^2(x)$ *to przy* $n \to \infty$ *rozkład średniej tych zmiennych ma rozkład asymptotycznie normalny o parametrach:* $N\left[E(x); \frac{\sigma(x)}{\sqrt{n}}\right]$. Twierdzenie to zasługuje na uwagę, ponieważ stwierdza „zmierzanie" rozkładu średniej z próby do rozkładu normalnego niezależnie od rozkładu populacji, z której próba została pobrana. Przyjmuje się jednak zastrzeżenie, że próba winna być dostatecznie liczna, tzn. winna liczyć powyżej 30 jednostek. Jest to reguła dość arbitralna. Większa minimalna liczebność jest wymagana, gdy rozkład w populacji daleko odbiega od rozkładu normalnego, gdy zaś jest zbliżony można przyjąć mniejszą próbę.

Szczególnym przypadkiem twierdzenia granicznego jest *twierdzenie Moivre'a - Laplace'a*. Zgodnie z nim rozkład normalny jest rozkładem granicznym dla rozkładu dwumianowego, gdy *n* rośnie nieograniczenie, co można ująć następująco: *zmienna o rozkładzie dwumianowym i parametrach* ***n*** *oraz* ***p*** *przy* $n \to \infty$ *ma asymptotycznie rozkład normalny o parametrach:* $E(x) = n \cdot p$ *oraz* $\sigma(x) = \sqrt{n \cdot p \cdot q}$. Twierdzenie to może być formułowane jako twierdzenie lokalne lub integralne; w pierwszym przypadku, przy dużych wartościach ***n*** prawdopodobieństwa rozkładu dwumianowego mogą być obliczone za pomocą funkcji gęstości rozkładu normalnego, natomiast w drugim, dla dużych ***n*** dystrybuanta rozkładu dwumianowego może być zastąpiona dystrybuantą rozkładu normalnego.

Wnioski wynikające z twierdzeń granicznych można ująć następująco: *jeśli zmienną losową traktować jako sumę znacznej liczby zmiennych losowych, z których żadna nie posiada dominującego wpływu na wielkość tej sumy, to posiada ona najczęściej charakter rozkładu normalnego.*

W konsekwencji losowaną próbę można również traktować jako sumę zmiennych i jej rozkład dla dużych *n* jest zbliżony do normalnego, co praktycznie oznacza możliwość przyjęcia odpowiednich parametrów ustalonych dla próby jako parametrów rozkładu normalnego dla populacji generalnej.

5.2. PRÓBA STATYSTYCZNA I SCHEMATY JEJ LOSOWANIA

Próba statystyczna, na podstawie której odbywa się wnioskowanie o populacji, może być z niej rozmaicie pobierana, a zasadniczym postulatem jest by miała ona charakter losowy. Nie może zatem mieć miejsca świadomy wybór jednostek do próby. Wnioski o populacji generalnej otrzymane na podstawie zbadanej próby są słuszne tylko wtedy, gdy próba jest podobna do populacji, z której pochodzi. O próbie, która dobrze odzwierciedla wszystkie interesujące nas własności populacji generalnej mówimy, że jest *próbą reprezentatywną*. Warunki, które musi spełniać taka próba, ujmuje się następująco:

- próba musi być podzbiorem populacji generalnej,
- elementy próby muszą być dobrane w sposób losowy,

- próba musi być dostatecznie liczna.

W oparciu o powyższe warunki Gliwienko sformułował następujące twierdzenie: *Jeżeli próba jest dostatecznie liczna to z prawdopodobieństwem bliskim 1 mamy prawo oczekiwać, że rozkład empiryczny cechy w próbie mało różni się od rozkładu teoretycznego w populacji generalnej.*

Jednym z warunków reprezentatywności próby jest losowy sposób jej pobierania, tzn. o wyborze jednostek do próby decyduje przypadek. Winien to gwarantować właściwy mechanizm doboru jednostek do próby zwany *mechanizmem* lub *schematem losowania*.

Dobry mechanizm losowania - gwarantujący uzyskanie takiej próby - winien spełniać następujące warunki:

- kryterium doboru elementów do próby winno być tak skonstruowane, aby było niezależne od tych zmiennych, które będą badane w próbie,
- kryterium doboru musi jednoznacznie określać, czy określony element populacji należy włączyć do próby, czy też z niej wykluczyć,
- kryterium doboru musi zapewniać każdemu elementowi populacji dodatnie prawdopodobieństwo trafienia do próby i wykluczać możliwość trafienia do niej elementu spoza populacji,
- w przypadku losowania bardzo dużej próby przy wykorzystaniu licznego zbioru badaczy konieczne jest ustalenie prostych zasad losowania by zapobiec pomyłkom.

Omawiane w literaturze schematy doboru jednostek do próby można podzielić na:

- arbitralne,
- losowe.

Dobór arbitralny – dobierający próbę przy jej wyborze kieruje się jedynie własną wiedzą i intuicją. Nie ma możliwości zweryfikowania słuszności doboru, a więc pobrana próba może mieć charakter tendencyjny. Jako nielosowy należy również określić dobór oparty na wiedzy ekspertów. Uzyskane w taki sposób próby mają charakter subiektywny i nie można w stosunku do wyników uzyskanych z takich prób stosować metod statystyki matematycznej, gdyż uzyskane zmienne nie muszą mieć charakteru losowego.

W praktyce najczęściej wykorzystywane są następujące *losowe schematy* kwalifikowania jednostek do próby:

- *losowanie indywidualne ze zwracaniem*, tzw. losowanie niezależne – po wylosowaniu elementu z populacji i zapisaniu wartości interesującej nas cechy powraca on do populacji; może więc on być powtórnie wylosowany; prawdopodobieństwo wylosowania każdego kolejnego elementu do próby jest stałe. Taki sposób losowania posiada następujące własności:
  - każdy element populacji ma tę samą szansę (czyli to samo prawdopodobieństwo) dostania się do próby,
  - wynik pierwszego i każdego kolejnego losowania nie ma wpływu na wynik następnego losowania.

  Schemat ten jest stosowany przy losowaniu próby z populacji mało licznej.

- *losowanie indywidualne bez zwracania* tzw. losowanie zależne – element raz wylosowany nie powraca do populacji generalnej, wobec czego skład populacji po każdym losowaniu jest inny (jej liczebność jest pomniejszona o jeden); zmienia się więc prawdopodobieństwo trafienia do próby każdego kolejnego elementu (jest to istotne w przypadku populacji mało licznych). Ponieważ schematy losowania mają gwarantować jednakową szansę dostania się do próby każdego elementu populacji generalnej, to zastosowanie tego schematu w przypadku populacji bardzo licznych lub nieskończonych nie ma praktycznego wpływu na prawdopodobieństwo trafienia do próby każdego kolejnego elementu. Ten schemat losowania jest zalecany dla populacji bardzo licznych.

Porównując obie metody losowania należy stwierdzić, iż przy losowaniu próby z populacji mało licznych metoda losowania bez zwracania jest bardziej efektywna, ponieważ pozwala na uzyskanie w oparciu o próbę większej ilości informacji o badanej populacji (w przypadku losowania ze zwracaniem elementy próby mogą się powtarzać, przy losowaniu bez zwracania nie jest to możliwe). W przypadku populacji bardzo licznych sposób pobierania próby nie ma znaczenia, chociaż należy zauważyć, że metody wnioskowania statystycznego zakładają przede wszystkim zwrotny sposób pobierania próby.

W przypadku populacji skończonych efektywną metodę pobierania próby gwarantuje *dobór przy wykorzystaniu tablic liczb losowych* cztero- pięcio- lub sześciocyfrowych. Tablice takie zawierają liczby 4, 5, 6 –cyfrowe zgrupowane w losowej kolejności w wierszach i kolumnach. W celu pobrania próby – po ponumerowaniu jednostek, tj. sporządzeniu tzw. operatu losowania i określeniu liczebności próby - wybieramy dowolny wiersz i dowolną kolumnę, w którym rozpoczynamy odczytywanie i kolejno (w zależności od liczebności próby: jeśli jej liczebność wyraża się w dziesiątkach – bierzemy pod uwagę dwie ostatnie cyfry odczytywanej liczby, a jeśli w setkach – trzy ostatnie cyfry) typujemy jednostki do próby stosując schemat losowania ze zwracaniem (bierzemy wówczas pod uwagę jednostki o numerach powtarzających się) lub bez zwracania (w tym przypadku numery powtarzające się pomijamy). Odczytane numery, którym nie odpowiadają żadne jednostki są pomijane. Odczytywanie kończymy, gdy próba zawiera żądaną liczbę elementów.

Niezależnie od powyższych schematów próba może być pobierana drogą losowania *warstwowego* lub *systematycznego.*

Dobór warstwowy winien być stosowany wówczas, gdy populacja generalna nie jest jednorodna. Dokonywany jest podział tej populacji na rozłączne części zwane warstwami. Poszczególne warstwy winny być w miarę jednorodne. Próba losowa jest pobierana z każdej warstwy oddzielnie, a jej skład jest proporcjonalny do liczebności poszczególnych warstw. W ten sposób każda z warstw ma zapewniony udział w wylosowanej próbie.

Schemat *losowania systematycznego* wymaga sporządzenia uporządkowanego wykazu wszystkich jednostek populacji generalnej (tzw. operatu losowania) i nadania każdej jednostce określonego numeru: 1, 2, ..., N. Dobór systematyczny polega na zakwalifikowaniu do próby co „*k* – tego" elementu poczynając od wylosowanego numeru pierwszej jednostki. Wielkość „*k*"

zwana jest interwałem losowania i jest ustalana jako iloraz liczebności populacji generalnej i losowanej próby.

5.3. ESTYMACJA PARAMETRYCZNA

5.3.1. Pojęcie i własności estymatora. Metody estymacji

Parametry populacji generalnej szacowane są przy wykorzystaniu statystyk z pobranej próby. Statystyka z próby wykorzystywana do oszacowania parametru populacji generalnej (tzw. *parametru estymowanego*) nosi nazwę *estymatora* tego parametru. *Estymatorem* Z_n parametru Q będziemy nazywali funkcję $Z_n = f(X_1, X_2, ..., X_n)$ określoną na próbie, która ma tę własność, że prawdopodobieństwo zdarzenia Z_n = Q jest tym bliższe jedności, im większa jest liczebność próby.

Parametr estymowany i estymator są najczęściej parametrami tego samego typu, np. średnia z próby jest estymatorem średniej w populacji generalnej. Poza średnią rolę estymatorów dla odpowiednich parametrów mogą pełnić również takie statystyki, jak np. wariancja, odchylenie standardowe - w przypadku cech liczbowych. W przypadku cech opisowych może interesować nas częstość (wskaźnik struktury bądź frakcja) występowania określonej kategorii elementów w populacji generalnej.

Od estymatorów oczekuje się by spełniały one określone własności. Zalicza się do nich przede wszystkim takie jak:

- *Zgodność* – jest oczywiste, że im mniejsza jest liczebność próby tym trudniej jest dokonać oszacowania na jej podstawie odpowiedniego parametru populacji generalnej (z całą pewnością niższe jest w tym przypadku zaufanie do wyników takiego oszacowania). Inna jest sytuacja w przypadku próby dość licznej. Estymator będziemy uważali za zgodny, jeżeli przy wzrastającej liczebności próby ($n \to \infty$) prawdopodobieństwo, że bezwzględna wartość różnicy między estymatorem Z_n, a parametrem estymowanym Q przekroczy dowolnie małą liczbę ε (gdzie ε > 0) zmierzać będzie do zera. Można to ująć następująco:

$$\lim_{n\to\infty} P\{|Z_n - Q| < \varepsilon\} = 1$$

  Praktycznie relacja ta oznacza, że ze wzrostem liczebności próby wartość estymatora będzie się zbliżała do wartości szacowanego parametru.
- *Nieobciążoność* – wykorzystując wartości estymatora uzyskiwane z różnych prób do oszacowania parametru estymowanego można uzyskać różne jego oceny (oszacowania). Pożądane jest by oszacowania te nie zawierały błędu systematycznego tzn. nie odchylały się przeciętnie ani poniżej ani powyżej wartości tego parametru. Estymator będziemy uważali za nieobciążony, jeżeli jego wartość oczekiwana jest równa parame-

trowi populacji, do oszacowania którego służy. Formalnie można to zapisać:

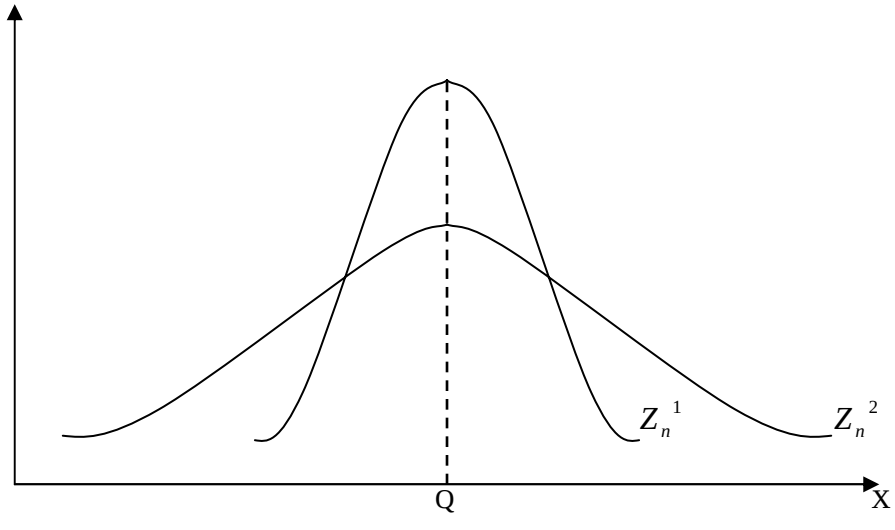
$$E(Z_n) = Q$$

Praktyczny aspekt tej relacji w przypadku szacowania wartości średniej oznacza, że jeżeli będziemy powtarzali wielokrotnie pobieranie próby z populacji i obliczali średnią dla kolejnych prób, to w końcowym efekcie wartość przeciętna z tych średnich będzie się pokrywała z interesującą nas średnią dla całej populacji, czyli nie wystąpi systematyczne odchylanie się wartości estymatora od szacowanego parametru.

- *Efektywność* – spośród nieobciążonych estymatorów Z_n parametru Q najlepszym będzie ten, który posiada najmniejszą wariancję, gdyż gwarantuje on, że uzyskane wyniki oszacowania będą mało różniły się od faktycznej wartości szacowanego parametru. Problem ten ilustruje rys. 5.1. Spośród dwóch zaprezentowanych estymatorów efektywniejszy jest estymator Z_n^1, ponieważ w porównaniu z estymatorem Z_n^2 charakteryzuje się mniejszą zmiennością. Estymator o najmniejszej wariancji spośród wszystkich możliwych nieobciążonych estymatorów parametru Q jest nazywany estymatorem najefektywniejszym. Miarą efektywności dowolnego estymatora jest iloraz wariancji estymatora najefektywniejszego i wariancji tego estymatora. Jeśli efektywność estymatora wzrasta wraz ze wzrostem liczebności próby, to o estymatorze mówimy, że jest *asymptotycznie najefektywniejszy*.

Rys. 5.1.

Efektywność estymatorów



Źródło: opracowanie własne

- *Dostateczność* - estymator jest dostateczny (wystarczający), jeżeli wykorzystuje wszystkie informacje o szacowanym parametrze, które są zawarte w próbie.

Teoria szacowania parametrów obejmuje dwie metody estymacji: *punktową i przedziałową. Estymacja punktowa* polega na tym, że jako ocenę nieznanego parametru Q populacji generalnej przyjmujemy uzyskaną z wylosowanej próby wartość estymatora Z_n. Szacowanie polega w tym przypadku na podaniu jednej konkretnej wartości liczbowej parametru estymowanego. Taki sposób postępowania oznacza, że jeśli z populacji będziemy pobierali kolejne próby, wyznaczali dla każdej z nich wartość estymatora, to można się spodziewać zróżnicowanych wartości liczbowych, a to z kolei może oznaczać, iż dla tej samej populacji istnieje kilka wartości tego samego parametru estymowanego (np. kilka wartości średnich tej samej zmiennej), co jest przecież niemożliwe. Prawdopodobieństwo zajścia zdarzenia, że uzyskana z dowolnej próby wartość estymatora jest identyczna jak faktyczna wartość szacowanego parametru jest praktycznie równe zero, co można zapisać następującą relacją:

$$P(Z_n = Q) = 0$$

Dyskwalifikuje ona tę metodę estymacji.

W przypadku *estymacji przedziałowej*, na podstawie wyników z wylosowanej próby, konstruowany jest przedział liczbowy, który z określonym z góry prawdopodobieństwem pokrywa wartość parametru estymowanego. Przedział ten jest określany mianem *przedziału ufności*, natomiast prawdopodobieństwo – *poziomem (współczynnikiem) ufności. Poziom ufności* (oznaczany dalej jako α) można zdefiniować jako prawdopodobieństwo, że skonstruowany przedział ufności zawiera wartość parametru estymowanego. Przyjmuje się, że prawdopodobieństwo to spełnia warunek: $\alpha \geq 0{,}90$. Istnieje określona relacja między wielkością poziomu ufności a precyzją szacowania parametru estymowanego: *im wyższy jest poziom ufności, tym mniejsza precyzja szacowania (większy błąd szacunku, większa rozpiętość przedziału ufności).*

Ogólny schemat postępowania w procedurze szacowania parametrów metodą przedziałową można ująć w następujących punktach:

1) z populacji generalnej losowana jest próba statystyczna,
2) na podstawie wyników uzyskanych z próby ustalana jest wartość estymatora odpowiedniego dla szacowanego parametru estymowanego,
3) zakładany jest poziom ufności α uwzględniający wynikające z tego faktu konsekwencje w postaci określonej precyzji szacowania parametru estymowanego,
4) z tablic statystycznych odpowiedniego rozkładu odczytywana jest właściwa dla przyjętego poziomu ufności wartość statystyki teoretycznej t_t ,

5) uzyskane dla próby wartości odpowiednich parametrów oraz odczytana z tablic wielkość statystyki teoretycznej wstawiane są do odpowiedniej formuły szacowania przedziału ufności dla określonego parametru estymowanego; przedział ten zostaje określony poprzez wyznaczenie jego dolnej i górnej granicy.

Poniżej zostaną omówione metody estymacji podstawowych parametrów statystycznych.

5.3.2. Estymacja przedziałowa wartości średniej

W literaturze wymienia się zazwyczaj dwa modele szacowania wartości średniej ściśle powiązane z liczebnością próby, na podstawie której jest ono dokonywane, tj. modele oparte na wynikach z małej i dużej próby.

Model dla małej próby

Jako małą przyjmuje się traktować próbę o liczebności $n \leq 30$. Estymatorem dla oszacowania wartości średniej w populacji generalnej $E(x)$ jest średnia z próby $\overline{x}$. Przyjmuje się założenie, że rozkład badanej zmiennej w populacji generalnej ma charakter rozkładu normalnego. Z populacji tej losowana jest próba i na podstawie uzyskanych z niej danych wyznaczana jest wartość średnia $\overline{x}$ i odchylenie standardowe $s(x)$. Z góry zakładany jest poziom ufności α. Przedział ufności dla wartości średniej $E(x)$ w populacji generalnej szacowany jest według wzoru:

$$\overline{x} - t_t \cdot \frac{s(x)}{\sqrt{n-1}} < E(x) < \overline{x} + t_t \cdot \frac{s(x)}{\sqrt{n-1}} \qquad 5.1$$

Występująca w powyższym wzorze wielkość t_t jest wartością statystyki odczytywaną z tablic rozkładu t Studenta dla $k = n-1$ oraz $1-\alpha$. Uzyskany przedział z prawdopodobieństwem równym poziomowi ufności pokrywa nieznaną wartość średnią w populacji generalnej. Warto zwrócić uwagę, iż otrzymany przedział jest symetryczny względem średniej z próby.

Należy zaznaczyć, iż błędna byłaby interpretacja, że szacowana średnia znajduje się w uzyskanym przedziale z prawdopodobieństwem równym α, ponieważ to przedział jest zmienny, a nie szacowana wartość średnia (ona jest wielkością stałą). Uwaga ta dotyczy estymacji wszelkich parametrów szacowanych metodą przedziałową.

Przykład 5.1.

W badaniach rozwoju czytelnictwa wśród młodzieży szkolnej dla losowej próby 15 uczniów klas I – III pewnej szkoły zebrano informacje dotyczące liczby przeczytanych książek w roku szkolnym. Otrzymano następujące informacje: *2; 6; 12; 10; 5; 4; 20; 22; 10; 15; 9; 8; 21; 14.*; 7; Zakładając, że rozkład przeczytanej liczby książek w całej populacji uczniów jest zbliżony

do normalnego - przy poziomie ufności 0,98 - oszacować metodą przedziałową średnią roczną liczbę przeczytanych książek dla tej populacji.

Rozwiązanie

Wylosowana próba jest mała, a więc dla oszacowania przedziału ufności wykorzystamy *formułę 5.1*. W pierwszej kolejności wymaga ona wyznaczenia średniej i odchylenia standardowego liczby przeczytanych książek w próbie. Korzystając z odpowiednich wzorów otrzymujemy:

$$\bar{x} = \frac{165}{15} = 11 \text{ książek}$$

$$s(x) = \sqrt{\frac{550}{15}} = 6{,}1 \text{ książki.}$$

Dla przyjętego poziomu ufności odczytujemy z tablic rozkładu t Studenta (*tablica 2. w Aneksie*) wartość statystyki teoretycznej t_t dla $k = 15-1 = 14$ oraz $1-\alpha = 1-0{,}98 = 0{,}02$. Wynosi ona 2,624. Uzyskane wielkości podstawiamy do podanej formuły :

$$11 - 2{,}624 \cdot \frac{6{,}1}{\sqrt{15-1}} < E(x) < 11 + 2{,}624 \cdot \frac{6{,}1}{\sqrt{15-1}}$$

$$11 - 4{,}30 < E(x) < 11 + 4{,}30$$

$$6{,}7 < E(x) < 15{,}3 \text{ książek}$$

Przedział ufności o końcach 6,7 i 15,3 książek z prawdopodobieństwem 0,98 zawiera nieznaną średnią roczną liczbę przeczytanych książek przez wszystkich uczniów klas I – III tej szkoły.

Zauważmy, że przedział ten jest symetryczny względem średniej z próby równej 11 książek; połowa jego rozpiętości, tj. $\frac{15{,}3-6{,}7}{2} = 4{,}3$ jest określana mianem maksymalnego błędu szacunku bądź tolerancją lub precyzją szacowania (oznaczana jest zwykle jako *d).*

Model dla dużej próby

Wylosowana próba winna posiadać liczebność przekraczającą 30 elementów. Przyjmuje się – podobnie jak w poprzednim modelu - założenie o normalnym rozkładzie populacji generalnej. Na podstawie wyników uzyskanych z próby ustalana jest średnia $\bar{x}$ i odchylenie standardowe $s(x)$. Z góry zakładany jest poziom ufności α. Przedział ufności dla średniej $E(x)$ w populacji generalnej szacowany jest według wzoru:

$$\bar{x} - t_t \cdot \frac{s(x)}{\sqrt{n}} < E(x) < \bar{x} + t_t \cdot \frac{s(x)}{\sqrt{n}} \qquad 5.2$$

gdzie: t_t jest wartością statystyki odczytywaną z tablic dystrybuanty rozkładu normalnego dla prawdopodobieństwa $\frac{\alpha}{2}$.

Przykład 5.2.

W badaniach struktury wydatków gospodarstw domowych zebrano m. in. informacje dotyczące wydatków na zakup artykułów przemysłowych. Dla losowej próby 200 gospodarstw uzyskano roczne kwoty wydatków na zakup tych artykułów podane w tablicy 5.1.

Tablica 5.1. Gospodarstwa domowe miasta „K" według rocznej kwoty *wydatków na zakup artykułów przemysłowych*

Kwota wydatków w zł	*Liczba gospodarstw*
500 - 1000	40
1000 – 1500	65
1500 - 2000	55
2000 – 2500	30
2500 - 3000	10

Źródło: Dane umowne

Zakładając, że w całej populacji gospodarstw wydatki te mają charakter rozkładu normalnego przy poziome ufności 0,99 oszacować metodą przedziałową średnie roczne wydatki na zakup artykułów przemysłowych w całej populacji gospodarstw domowych.

Rozwiązanie

Z uwagi na dużą próbę oszacowania przedziału ufności dla średniej dokonamy zgodnie z *wzorem 5.2*. W poniższej tablicy roboczej wykonano obliczenia pomocnicze dla ustalenia wartości średniej $\bar{x}$ i odchylenia standardowego $s(x)$ wydatków w wylosowanej próbie.

Kwota wydatków w zł (x_i)	*Liczba gospodarstw* (n_i)	$\dot{x}_i \cdot n_i$	$\dot{x}_i - \bar{x}$	$(\dot{x}_i - \bar{x})^2 \cdot n_i$
500 - 1000	40	30.000	- 762,5	23.255.487,5
1000 – 1500	65	81.250	- 262,5	4.478.643,75
1500 - 2000	55	96.250	237,5	3.102.581,25
2000 – 2500	30	67.500	737,5	16.317.925,0
2500 - 3000	10	27.500	1237,5	15.315.300,0
Razem	200	302.500	X	62.469.937,5

Otrzymujemy:

$$\bar{x} = \frac{302500}{200} = 1512{,}50 \text{ zł}$$

$$s(x) = \sqrt{\frac{62469937{,}5}{200}} = 558{,}9 \text{ zł}$$

Z tablic dystrybuanty rozkładu normalnego (*tablica 1. w Aneksie*) odczytujemy t_t dla $\frac{0{,}99}{2} = 0{,}495$; jako wartość najbardziej zbliżoną do tej wielkości przyjmujemy 0,4951, której odpowiada t_t=2,58. Podstawiając uzyskane wielkości do *wzoru 5.2* otrzymujemy:

$$1512{,}5 - 2{,}58 \cdot \frac{558{,}9}{\sqrt{200}} < E(x) < 1512{,}5 + 2{,}58 \cdot \frac{558{,}9}{200}$$

$$1512{,}5 - 101{,}98 < E(x) < 1512{,}5 + 101{,}98$$

$$1410{,}52 < E(x) < 1614{,}48 \text{ zł}$$

Otrzymany przedział z prawdopodobieństwem 0,99 pokrywa nieznaną średnią roczną kwotę wydatków na zakup artykułów przemysłowych przez wszystkie gospodarstwa domowe.

5.3.3. Wyznaczanie minimalnej liczebności próby w procedurze szacowania wartości średniej

Jest to problem często występujący w badaniach statystycznych. Pojawia się pytanie, jak liczną próbę należałoby zbadać, by uzyskać zadowalające wyniki oszacowania określonego parametru. W przypadku szacowania wartości średniej problem ten można ująć następująco: jaka winna być minimalna liczebność pobranej próby, by przy założonym poziomie ufności oszacować wartość średnią dla populacji generalnej z żądaną dokładnością (precyzją)? Proponowana procedura przyjmuje założenie, że rozkład populacji generalnej jest normalny, a jego parametry nieznane. Z populacji tej losowana jest wstępna mała próba o liczebności *n*. Na podstawie wyników z tej próby określana jest wariancja o postaci:

$$\hat{s}^2(x) = \frac{\sum_i (x_i - \bar{x})^2}{n-1} \quad \text{w przypadku szeregu szczegółowego} \qquad 5.3$$

lub o postaci

$$\hat{s}^2(x) = \frac{\sum_i (x_i - \bar{x})^2 \cdot n_i}{n-1} \quad \text{w przypadku szeregu rozdzielczego} \qquad 5.4$$

Zakładany jest poziom ufności α oraz żądana dokładność szacunku wartości średniej *d*. Minimalną liczebność próby wyznaczamy z wzoru:

$$n_{\min} = \frac{t_t^{\,2} \cdot \hat{s}^2(x)}{d^2} \qquad 5.5$$

Występującą w podanym wzorze wartość statystyki t_t odczytujemy z tablic rozkładu t Studenta dla $k = n-1$ oraz $1-\alpha$. Z uwagi na fakt, że liczebność próby musi być liczbą całkowitą w związku z tym – w przypadku konieczności - dokonujemy zawsze jej zaokrąglenia do pełnej jednostki w górę.

Przykład 5.3.

Traktując wylosowaną w *przykładzie 5.1* próbę uczniów jako próbę wstępną ustalić, jaka minimalna liczba uczniów pozwoliłaby oszacować średnią roczną liczbę przeczytanych książek dla wszystkich uczniów klas I – III z błędem maksymalnym 2 książki przy poziomie ufności 0,95.

Rozwiązanie

Na podstawie wyników z próby wstępnej ustalamy zgodnie z *wzorem 5.3* wariancję $\hat{s}^2(x)$ liczby przeczytanych książek:

$$\hat{s}^2(x) = \frac{550}{14} = 39{,}28 \text{ (książek)}^2$$

Z tablic rozkładu t Studenta odczytujemy wartość statystyki t_t dla $k = 15-1 = 14$ oraz $1-0{,}95 = 0{,}05$; wynosi ona 2,145. Założony błąd szacunku *d* = 2.

Podstawiając te wielkości do *wzoru 5.5* otrzymujemy:

$$n_{\min} = \frac{(2{,}145)^2 \cdot 39{,}28}{(2)^2} = \frac{180{,}7}{4} = 45{,}17 \text{ uczniów}$$

Oznacza to, że dla oszacowania średniej rocznej liczby przeczytanych książek z błędem maksymalnym 2 książek przy poziomie ufności 0,95 nale-

ży wylosować do próby co najmniej 46 uczniów (wynik zaokrąglamy w górę). Do próby wstępnej należy wobec tego „dolosować" jeszcze 31 uczniów.

5.3.4. Estymacja przedziałowa wskaźnika struktury

W przypadku cechy opisowej – gdy określanie typowych parametrów statystycznych jest niemożliwe – procedura szacowania może dotyczyć udziału określonego wariantu tej cechy w populacji generalnej. W tym celu z populacji tej losowana jest duża próba ($n > 100$), dla której określa się wskaźnik struktury o postaci $\frac{m}{n}$, gdzie *m* jest liczbą wyróżnionych w próbie elementów, a *n* jej liczebnością. Zakładany jest poziom ufności α. Przedział ufności dla wskaźnika struktury (*p*) w populacji generalnej wyznaczany jest według formuły:

$$\frac{m}{n} - t_t \cdot \sqrt{\frac{\frac{m}{n}\left(1-\frac{m}{n}\right)}{n}} < p < \frac{m}{n} + t_t \cdot \sqrt{\frac{\frac{m}{n}\left(1-\frac{m}{n}\right)}{n}} \qquad 5.6$$

Występującą w podanym wzorze wartość statystyki t_t odczytujemy z tablic dystrybuanty rozkładu normalnego dla $\frac{\alpha}{2}$.

Przykład 5.4.

W badaniach warunków socjalnych studentów pewnej uczelni zebrano między innymi informacje dotyczące miejsca ich zamieszkania w okresie studiów. Uzyskano dane ujęte w tablicy 5.2.

Tablica 5.2. Studenci Akademii Medycznej w „K" według miejsca zamieszkania w czasie studiów

Miejsce zamieszkania	*Liczba studentów*
Dom studencki	120
Stancja	60
Dom rodzinny	40
Razem	220

Źródło: Dane umowne

Przyjmując poziom ufności 0,95 oszacować metodą przedziałową:
a) udział studentów zamieszkujących w domu studenckim,
b) udział studentów zamieszkujących poza domem rodzinnym.

Rozwiązanie

ad. a) W celu oszacowania przedziału ufności dla wskaźnika struktury wykorzystamy *wzór 5.6*. Wymaga on wyznaczenia z próby wskaźnika struktury dla studentów zamieszkujących w domu studenckim. Wskaźnik ten wynosi

$$\frac{m}{n} = \frac{120}{220} = 0{,}5455$$

Z tablic dystrybuanty rozkładu normalnego odczytujemy wartość statystyki t_t dla $\frac{0{,}95}{2} = 0{,}475$; wynosi ona 1,96. Podstawiamy otrzymane wielkości do *wzoru 5.6* i otrzymujemy

$$0{,}5455 - 1{,}96 \cdot \sqrt{\frac{0{,}5455(1-0{,}5455)}{220}} < p < 0{,}5455 + 1{,}96 \cdot \sqrt{\frac{0{,}5455(1-0{,}5455)}{220}}$$

$$0{,}5455 - 0{,}0669 < p < 0{,}5455 + 0{,}0669$$

$$0{,}4786 < p < 0{,}6124$$

Wyrażając końce przedziału w procentach otrzymujemy:

$$47{,}86\% < p < 61{,}24\%.$$

Przedział liczbowy o końcach 47,86 % i 61,24 % z prawdopodobieństwem 0,95 zawiera nieznany udział studentów tej uczelni zamieszkujących w domu studenckim.

ad. b) W stosunku do *punktu a* zmianie ulegnie wskaźnik struktury dla próby i wyniesie on:

$$\frac{m}{n} = \frac{120+60}{220} = 0{,}8182$$

Wartość t_t będzie identyczna jak wyżej. Podstawiając otrzymane wielkości do *wzoru 5.6* otrzymujemy

$$0{,}8182 - 1{,}96 \cdot \sqrt{\frac{0{,}8182(1-0{,}8182)}{220}} < p < 0{,}8182 + 1{,}96 \cdot \sqrt{\frac{0{,}8182(1-0{,}8182)}{220}}$$

$$0{,}8182 - 0{,}0516 < p < 0{,}8182 + 0{,}0516$$

$$0{,}7666 < p < 0{,}8698,$$

a w ujęciu procentowym:

$$76{,}66\% < p < 86{,}98\%$$

Uzyskany wynik oznacza, że przedział o końcach 76,66 % i 86,98 % z ufnością 0,95 zawiera nieznany udział studentów tej uczelni zamieszkujących w czasie studiów poza domem rodzinnym.

5.3.5. Estymacja przedziałowa wariancji i odchylenia standardowego

Z uwagi na ścisłe powiązania obu parametrów ich szacowanie odbywa się zwykle łącznie. W zależności od wielkości próby, na podstawie której dokonywane jest ono, można wyróżnić dwa modele postępowania.

Model oparty na wynikach z małej próby

Zakłada się, że populacja generalna posiada rozkład normalny. Z populacji tej losowana jest mała próba ($n \leq 30$). Na jej podstawie ustalana jest wariancja $s^2(x)$ uzyskanych wyników. Stanowi ona estymator dla szacowanej wariancji populacji generalnej. Zakładany jest poziom ufności α. Przedział ufności dla wariancji populacji generalnej szacowany jest według wzoru:

$$\frac{n \cdot s^2(x)}{t_{t_1}} < \sigma^2(x) < \frac{n \cdot s^2(x)}{t_{t_2}} \qquad 5.7$$

gdzie: t_{t_1} i t_{t_2} są wartościami statystyki teoretycznej odczytywanymi z tablic rozkładu χ^2 (chi-kwadrat) przy założonym poziomie ufności odpowiednio:

- t_{t_1} dla $k = n-1$ oraz $\frac{1-\alpha}{2}$,
- t_{t_2} dla $k = n-1$ oraz $\frac{1+\alpha}{2}$.

W celu uzyskania przedziału ufności dla odchylenia standardowego wyznaczamy pierwiastki kwadratowe z końców przedziału oszacowanego dla wariancji (korzystamy tu z oczywistej relacji zachodzącej między tymi parametrami).

Przykład 5.5.

Na wylosowanej grupie 10 dzieci w wieku przedszkolnym przeprowadzono test pamięci. Otrzymano następujący rozkład liczby zapamiętanych przez nie elementów: *15; 34; 45; 32; 18; 52; 25; 50; 40; 29*. Zakładając, że w populacji generalnej rozkład liczby zapamiętanych elementów ma charakter rozkładu normalnego oszacować granice przedziału ufności dla wariancji i

odchylenia standardowego liczby zapamiętanych elementów przy poziomie ufności 0,96

Rozwiązanie

Ze względu na małą próbę korzystamy z podanej wyżej procedury postępowania. Na podstawie uzyskanych wyników z próby ustalamy w pierwszej kolejności średnią $\bar{x}$, a następnie wariancję $s^2(x)$ liczby zapamiętanych elementów. Wartość średnia wyniesie:

$$\bar{x} = \frac{340}{10} = 34 \text{ elementy}$$

zaś wariancja (wyznaczona według wzoru dla szeregu szczegółowego):

$$s^2(x) = \frac{1464}{10} = 146{,}4$$

Dla przyjętego poziomu ufności z tablic rozkładu χ^2 odczytujemy:

- t_{t_1} dla $k = 10 - 1 = 9$ oraz $\frac{1 - 0{,}96}{2} = 0{,}02$ i otrzymujemy 19,679
- t_{t_2} dla $k = 10 - 1 = 9$ oraz $\frac{1 + 0{,}96}{2} = 0{,}98$ i wynosi ono 2,532.

Uzyskane wielkości podstawiamy do *formuły 5.7* i otrzymujemy:

$$\frac{10 \cdot 146{,}4}{19{,}679} < \sigma^2(x) < \frac{10 \cdot 146{,}4}{2{,}532}$$

$$74{,}39 < \sigma^2(x) < 578{,}20 \text{ (elementów)}^2$$

Oszacowany przedział o końcach 74,39 i 578,2 (elementów)^2 zawiera wariancję liczby zapamiętanych elementów dla wszystkich dzieci w wieku przedszkolnym przy poziomie ufności 0,96.

Przedział ufności dla odchylenia standardowego liczby zapamiętanych elementów uzyskamy ustalając pierwiastki kwadratowe z końców oszacowanego powyżej przedziału. Otrzymujemy:

$$\sqrt{74{,}39} < \sigma(x) < \sqrt{578{,}2}$$

$$8{,}6 < \sigma(x) < 24{,}0 \text{ elementy.}$$

Przedział o końcach 8,6 i 24 elementy z prawdopodobieństwem 0,96 zawiera nieznane odchylenie standardowe liczby zapamiętanych elementów przez wszystkie dzieci w wieku przedszkolnym.

Model dla dużej próby

Model ten również zakłada, że populacja generalna ma rozkład co najmniej zbliżony do normalnego. W odróżnieniu od poprzedniego modelu losowana jest w tym przypadku duża próba ($n > 30$) i na jej podstawie ustalana jest wartość odchylenia standardowego $s(x)$. Zakładany jest poziom ufności α. Przedział ufności dla odchylenia standardowego populacji generalnej szacowany jest według formuły:

$$\frac{s(x)}{1+\frac{t_t}{\sqrt{2n}}} < \sigma(x) < \frac{s(x)}{1-\frac{t_t}{\sqrt{2n}}} \qquad 5.8$$

gdzie: t_t jest wartością statystyki odczytaną z tablic dystrybuanty rozkładu normalnego dla $\frac{\alpha}{2}$.

Korzystając z relacji zachodzącej między odchyleniem standardowym a wariancją przedział ufności dla wariancji populacji generalnej uzyskamy ustalając kwadraty końców przedziału oszacowanego dla odchylenia standardowego.

Przykład 5.6.

W badaniach dostępności pacjentów do lekarzy - specjalistów na terenie miasta „K" zebrano informacje dotyczące czasu ich oczekiwania na wizytę u lekarza. Otrzymano dane ujęte w poniższej tablicy.

Tablica 5.3. Pacjenci według czasu oczekiwania (w dniach) na wizytę u lekarza specjalisty w mieście „K".

Czas oczekiwania w dniach	*Liczba pacjentów*
0 – 5	20
5 - 15	30
15 - 30	25
Razem	75

Źródło: Dane umowne

Zakładając poziom ufności 0,90 oszacować metodą przedziałową odchylenie standardowe i wariancję czasu oczekiwania pacjentów na wizytę u lekarza specjalisty.

Rozwiązanie

Z uwagi na dużą próbę dla oszacowania przedziału ufności dla odchylenia standardowego i wariancji wykorzystamy *formułę 5.8*. Na podstawie danych zawartych w tablicy 5.3 obliczamy odchylenie standardowe $s(x)$ czasu oczekiwania z próby. Obliczenia pomocnicze zawarto w poniższej tablicy roboczej

Czas oczekiwania w dniach	*Liczba pacjentów*	$\dot{x}_i \cdot n_i$	$\dot{x}_i - \overline{x}$	$(\dot{x}_i - \overline{x})^2 \cdot n_i$
0 – 5	20	50	- 9,7	1872,1
5 - 15	30	300	- 2,2	145,2
15 - 30	25	562,5	10,3	2662,55
Razem	75	912,5	X	4679,85

Otrzymujemy $\overline{x} = \frac{912,5}{75} = 12,2$ dnia oraz $s(x) = \sqrt{\frac{4679,85}{75}} = 7,9$ dnia.

Z tablic dystrybuanty rozkładu normalnego odczytujemy wartość statystyki t_t dla $\frac{0,90}{2} = 0,45$; jako wartość najbliższą tej wielkości przyjmijmy 0,4505, co oznacza przyjęcie t_t = 1,65. Na podstawie *wzoru 5.8*, w pierwszej kolejności oszacujemy przedział ufności dla odchylenia standardowego. Będzie on wynosił:

$$\frac{7,9}{1+\frac{1,65}{\sqrt{2 \cdot 75}}} < \sigma(x) < \frac{7,9}{1-\frac{1,65}{\sqrt{2 \cdot 75}}}$$

$$\frac{7,9}{1+0,13} < \sigma(x) < \frac{7,9}{1-0,13}$$

$$7,0 < \sigma(x) < 9,1 \text{ dni}$$

Przedział liczbowy o końcach 7 i 9,1 dni z ufnością 0,90 pokrywa odchylenie standardowe czasu oczekiwania na wizytę u lekarza specjalisty dla wszystkich pacjentów.

Przedział ufności dla wariancji czasu oczekiwania otrzymamy ustalając kwadraty końców powyższego przedziału. Otrzymamy:

$$(7,0)^2 < \sigma^2(x) < (9,1)^2$$

$$49,0 < \sigma^2(x) < 82,8 \text{ (dni)}^2$$

Przedział liczbowy 49 – 82,8 (dni)^2 z ufnością 0,90 zawiera wariancję czasu oczekiwania na wizytę u lekarza specjalisty dla wszystkich pacjentów.

5.3.6. Estymacja przedziałowa współczynnika korelacji

Badanie współzależności cech statystycznych odbywa się najczęściej w warunkach badań częściowych, co oznacza, iż interpretacja uzyskanych wyników odnosi się do badanej próby. Zasadne jest w tej sytuacji postawienie pytania: Jakie – wobec tego - natężenie i kierunek współzależności występują pomiędzy badanymi zmiennymi w przypadku populacji generalnej? Odpowiedzi na to pytanie udzielimy dokonując oszacowania przedziału ufności dla współczynnika korelacji dla tej populacji.

Zakłada się, że rozkład badanych cech (koniecznie liczbowych) w populacji generalnej ma rozkład w przybliżeniu normalny, a związek między nimi jest prostoliniowy. Z populacji tej losowana jest duża próba (co najmniej kilkaset elementów), a wyniki dla niej uzyskane ujmowane są w formie tablicy korelacyjnej. Dla tak ujętych wyników ustalane jest natężenie i kierunek zależności między badanymi cechami za pomocą współczynnika korelacji liniowej Pearsona. Następnie przyjmowane jest założenie o wielkości poziomu ufności α. Przedział ufności dla współczynnika korelacji szacowany jest według wzoru:

$$r^P - t_t \cdot \frac{1-(r^P)^2}{\sqrt{n}} < \rho < r^P + t_t \cdot \frac{1-\left(r^P\right)^2}{\sqrt{n}} \qquad 5.9$$

gdzie: r^P – współczynnik korelacji liniowej Pearsona ustalony dla próby,

ρ (czytaj: ro) – współczynnik korelacji liniowej Pearsona dla populacji generalnej,

t_t - wartość statystyki odczytywana z tablic dystrybuanty rozkładu normalnego dla $\frac{\alpha}{2}$.

Oszacowany przedział z prawdopodobieństwem równym poziomowi ufności pokrywa nieznaną wartość współczynnika korelacji dla populacji generalnej.

Przykład 5.7.

W pewnym badaniu socjologicznym zebrano m. in. informacje dotyczące wieku kobiet i mężczyzn wstępujących w związek małżeński. Dla wylosowanych 200 par małżeńskich stwierdzono, iż pomiędzy badanymi cechami występuje zależność mierzona współczynnikiem korelacji liniowej Pearsona równa + 0,75. Przy poziomie ufności 0,99 oszacować metodą przedziałową współczynnik korelacji dla wieku wszystkich kobiet i mężczyzn zawierających związek małżeński.

Rozwiązanie

Przedział ufności dla współczynnika korelacji oszacujemy zgodnie z *formułą 5.9*. Dla przyjętego współczynnika ufności z tablic rozkładu normalnego odczytujemy $t_t = 2{,}58$.

Podstawiając odpowiednie dane do wzoru otrzymujemy

$$0{,}75 - 2{,}58 \cdot \frac{1-(0{,}75)^2}{\sqrt{200}} < \rho < 0{,}75 + 2{,}58 \cdot \frac{1-(0{,}75)^2}{\sqrt{200}}$$

$$0{,}75 - 0{,}08 < \rho < 0{,}75 + 0{,}08$$

$$0{,}67 < \rho < 0{,}83$$

Przedział liczbowy o końcach 0,67 i 0,83 z prawdopodobieństwem 0,99 zawiera współczynnik korelacji wieku kobiet i mężczyzn zawierających związek małżeński.

5.4. WERYFIKACJA HIPOTEZ STATYSTYCZNYCH

5.4.1. Istota procedury hipotez statystycznych

Weryfikacja hipotez statystycznych stanowi drugą metodę wnioskowania statystycznego. Mianem *hipotezy statystycznej* określa się jakiekolwiek przypuszczenie dotyczące rozkładu populacji generalnej. Dokonując weryfikacji postawionej hipotezy rozstrzygamy o jej słuszności. Procedura weryfikacji odbywa się przy wykorzystaniu narzędzi statystycznych zwanych testami. Szczególnie miejsce wśród nich zajmują *testy istotności*. Procedura tego typu testów pozwala, na podstawie wyników uzyskanych z próby losowej, na podjęcie jednej z dwóch alternatywnych decyzji:

a) o odrzuceniu hipotezy sprawdzanej,
b) o stwierdzeniu braku podstaw do jej odrzucenia.

Praktycznie oznacza to, że upoważniają one do odrzucenia sprawdzanej hipotezy, gdy jest ona fałszywa, natomiast nie dają podstaw do stwierdzenia, że postawiona hipoteza może być uznana za prawdziwą. Pierwsza z decyzji ma charakter jednoznaczny, należy jednak zauważyć, iż jest ona podejmowana jedynie w oparciu o wyniki uzyskane z próby. Zakładać więc należy możliwość podjęcia decyzji błędnej polegającej na odrzuceniu hipotezy pomimo, że jest ona prawdziwa (błąd ten określany jest mianem *błędu pierwszego rodzaju*). Prawdopodobieństwo popełnienia tego błędu określane jest mianem *poziomu istotności* i będziemy je oznaczali jako $1-\alpha$. Przyjmuje się, że jest to prawdopodobieństwo nie większe od 0,10, a jego wielkość jest ustalana przez prowadzącego badanie.

Algorytm postępowania w procedurze weryfikacji hipotez przy wykorzystaniu testu istotności można ująć w następujących punktach:

1) stawiamy hipotezę zerową i konkurencyjną wobec niej hipotezę alternatywną; w zależności od postaci hipotezy alternatywnej wykorzystywany jest test dwustronny bądź jednostronny (prawo- lub lewostronny); należy tu dodać, że w przypadku testu istotności hipoteza zerowa jest zawsze formułowana w postaci równości,
2) arbitralnie przyjmujemy poziom istotności $1-\alpha$,
3) z populacji generalnej losowana jest próba statystyczna i na podstawie wyników z tej próby ustalana jest wartość statystyki empirycznej t_{emp},
4) dla przyjętego poziomu istotności – z odpowiednich tablic – odczytywana jest wartość statystyki teoretycznej t_t określanej również mianem wartości krytycznej,
5) porównujemy wartości statystyki empirycznej i teoretycznej i w przypadku:

a) testu dwustronnego;
- jeśli $|t_{emp}| \geq t_t$ podejmujemy decyzję o odrzuceniu hipotezy zerowej,
- jeśli $|t_{emp}| < t_t$ stwierdzamy brak podstaw do odrzucenia hipotezy zerowej,

b) testu prawostronnego:
- jeśli $t_{emp} \geq t_t$ podejmujemy decyzję o odrzuceniu hipotezy zerowej,
- jeśli $t_{emp} < t_t$ stwierdzamy brak podstaw do odrzucenia hipotezy zerowej,

c) testu lewostronnego:
- jeśli $t_{emp} \leq -t_t$ podejmujemy decyzję o odrzuceniu hipotezy zerowej,
- jeśli $t_{emp} > -t_t$ stwierdzamy brak podstaw do odrzucenia hipotezy zerowej.

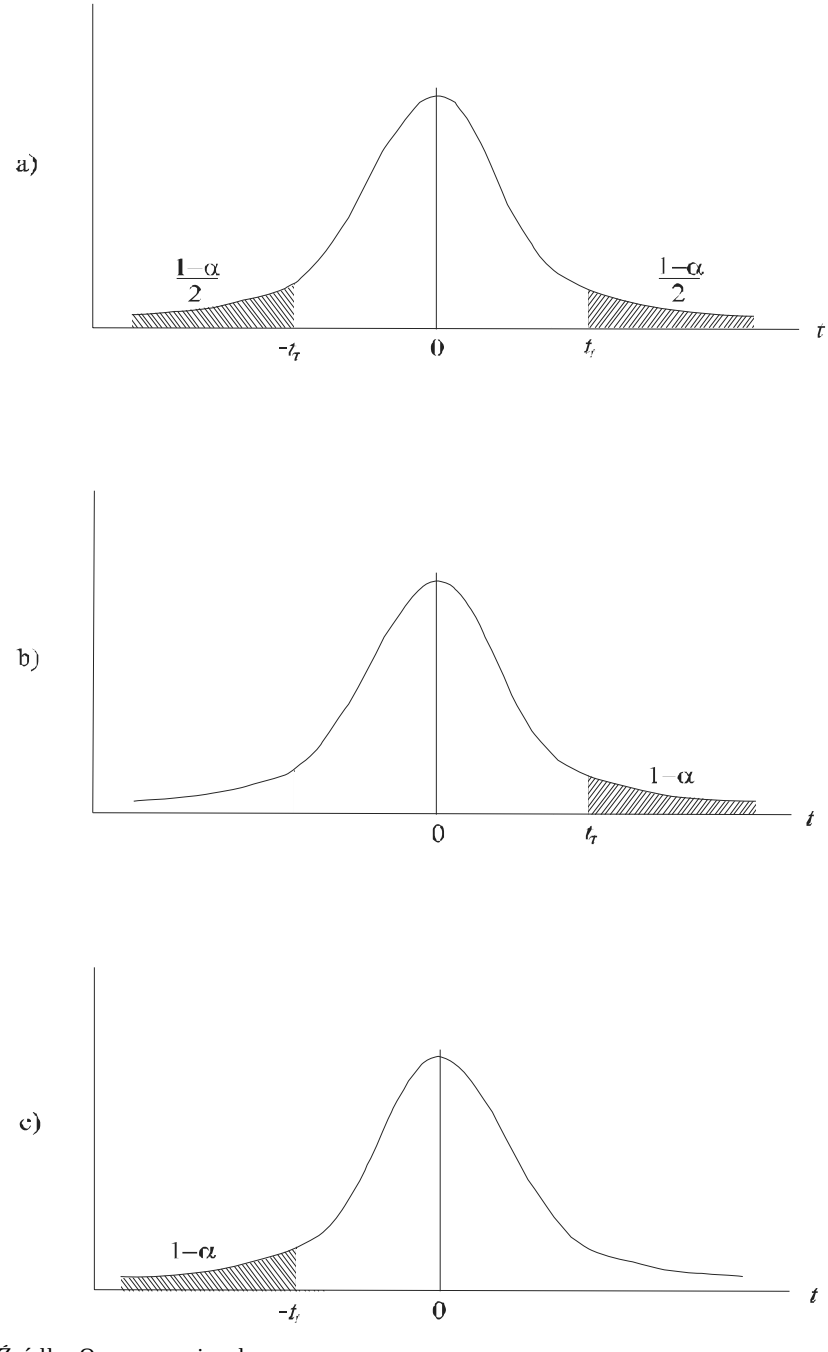
Procedura weryfikacji hipotez, a zwłaszcza ostatnia z wymienionych czynności może być również zilustrowana graficznie. Wówczas dla przyjętej postaci hipotezy alternatywnej konstruowany jest tzw. obszar krytyczny odpowiadający poziomowi istotności. Ilustruje to poniższy rys.5.2.

W przypadku (*a*) mamy do czynienia z testem dwustronnym i odpowiadającym mu położeniem obszaru krytycznego. Przypadek *(b*) odpowiada testowi prawostronnemu i takiemu również położeniu obszaru krytycznego, zaś przypadek (*c*) testowi lewostronnemu i odpowiedniemu położeniu obszaru krytycznego. Jeśli ustalona na podstawie próby wartość statystyki empirycznej „wpada" w obszar krytyczny wówczas podejmowana jest decyzja o odrzuceniu hipotezy zerowej, w przeciwnym przypadku brak jest podstaw do jej odrzucenia.

Jakkolwiek podany wyżej algorytm postępowania ma charakter ogólny, to jednak wymienione czynności są charakterystyczne dla wszystkich niżej omówionych przypadków weryfikacji hipotez.

Rys. 5.2

Relacje między postacią hipotezy alternatywnej a położeniem obszaru krytycznego



Źródło: Opracowanie własne

5.4.2. Weryfikacja hipotezy dla wartości średniej

Celem postępowania jest sprawdzenie hipotezy dotyczącej wartości średniej w populacji generalnej. Przyjmuje się założenie, że populacja ta ma charakter rozkładu normalnego o nieznanej średniej i odchyleniu standardowym. W zależności od wielkości losowanej z tej populacji próby wyróżnia się dwa modele postępowania.

Model dla małej próby

Kolejne czynności wykonujemy zgodnie z podanym wyżej algorytmem postępowania.

1) stawiamy hipotezę zerową o postaci:
$H_0 : E(x) = E_0(x)$
i jedną z niżej wymienionych hipotez alternatywnych:
   a) $H_1 : E(x) \neq E_0(x)$
   b) $H_1 : E(x) > E_0(x)$
   c) $H_1 : E(x) < E_0(x)$,
   gdzie: $E(x)$ – wartość średnia dla populacji generalnej,
   $E_0(x)$ – założona hipotetyczna wartość średnia.

W przypadku uwzględnienia pierwszej wersji hipotezy alternatywnej postępowanie będzie się odbywało przy wykorzystaniu *testu dwustronnego*, drugiej – *testu prawostronnego*, trzeciej – *lewostronnego*.

2) zakładamy poziom istotności $1-\alpha$,
3) z populacji generalnej losujemy małą próbę o liczebności $n \leq 30$ i na podstawie uzyskanych z niej wyników wyznaczamy wartość średnią $\bar{x}$ i odchylenie standardowe $s(x)$. Parametry te wykorzystujemy do wyznaczenia statystyki empirycznej zgodnie z wzorem:

$$t_{emp} = \frac{\bar{x} - E_0(x)}{s(x)} \cdot \sqrt{n-1} \qquad 5.10$$

4) z tablic rozkładu t Studenta odczytujemy wartość statystyki teoretycznej t_t według reguły:
   - w przypadku testu dwustronnego: dla $k = n-1$ oraz poziomu istotności $1-\alpha$,
   - w przypadku testu jednostronnego: dla $k = n-1$ oraz $2\cdot(1-\alpha)$.
5) zgodnie z podanymi zasadami podejmujemy decyzję odnośnie sprawdzanej hipotezy.

Przykład 5.8.

Dokonując analizy przestępczości nieletnich dla wylosowanej próby zgromadzono m. in. informacje dotyczące ich wieku. Uzyskano następujące

dane (wiek w latach): *17; 16; 18; 15; 17; 19; 16; 15; 17; 14; 13; 15; 16; 14; 18*. Zakładając, że rozkład wieku nieletnich przestępców ma charakter rozkładu normalnego przy poziomie istotności 0,01 zweryfikować hipotezę, iż średni wiek dla całej ich populacji jest równy 17 lat.

Rozwiązanie

Zgodnie z procedurą stawiamy hipotezy o postaci:

$H_0 : E(x) = 17$ lat

$H_1 : E(x) \neq 17$ lat

W treści zadania założono poziom istotności $1-\alpha$ = 0,01. Na podstawie wyników z próby ustalamy średnią i odchylenie standardowe wieku nieletnich przestępców:

$$\bar{x} = \frac{240}{15} = 16 \text{ lat}$$

$$s(x) = \sqrt{\frac{40}{15}} = 1{,}63 \text{ lat}$$

Na podstawie ustalonych parametrów wyznaczamy wartość statystyki empirycznej według *wzoru 5.10*:

$$t_{emp} = \frac{16-17}{1{,}63} \cdot \sqrt{15-1} = -0{,}61 \cdot 3{,}74 = -2{,}28$$

Przy założonym poziomie istotności odczytujemy z tablic rozkładu t Studenta (*tablica 2. w Aneksie*) wartość statystyki teoretycznej t_t dla k = 15 – 1 = 14 oraz $1-\alpha$ = 0,01, ponieważ test ma charakter dwustronny. Wynosi ona 2,977. Zachodzi zatem relacja $\left|t_{emp}\right| < t_t$, co oznacza, że nie ma podstaw do odrzucenia hipotezy zerowej. W tej sytuacji przy poziomie istotności 0,01 można twierdzić, że średni wiek nieletnich przestępców wynosi 17 lat.

Model dla dużej próby.

Czynności wstępne oznaczone wyżej jako *1 i 2* są identyczne jak w poprzednim modelu. W dalszej kolejności z populacji generalnej losowana jest duża próba o liczebności $n > 30$ i na podstawie uzyskanych danych ustalana jest wartość średnia $\bar{x}$ i odchylenie standardowe $s(x)$, a następnie wartość statystyki empirycznej według wzoru:

$$t_{emp} = \frac{\bar{x} - E_0(x)}{s(x)} \cdot \sqrt{n} \qquad 5.11$$

Dla przyjętego poziomu istotności z tablic dystrybuanty rozkładu normalnego ustalana jest wartość statystyki teoretycznej t_t zgodnie z regułą:

- w przypadku testu dwustronnego: t_t odczytywane jest dla $0{,}5-\frac{1-\alpha}{2}$,
- w przypadku testu jednostronnego: t_t odczytujemy dla $0{,}5-(1-\alpha)$.

Decyzja o odrzuceniu hipotezy zerowej bądź stwierdzeniu braku podstaw do takiej decyzji podejmowana jest jak w podanym algorytmie.

Przykład 5.9.

Zebrano informacje dla grupy kierowców, którzy w okresie ostatnich 8 lat na terenie miasta „K" spowodowali wypadek drogowy znajdując się pod wpływem alkoholu. Uzyskano następujące zestawienie:

Poziom alkoholu we krwi (w promilach)	*Liczba kierowców*
0,40 – 1,0	15
1,0 – 1,6	120
1,6 – 2,2	180
2,2 – 2,8	85

Zakładając, że badana populacja ma charakter rozkładu normalnego przy poziomie istotności 0,05 zweryfikować hipotezę, że średnie stężenie alkoholu we krwi w całej populacji nietrzeźwych kierowców, którzy spowodowali wypadek drogowy, jest większe od 2,3 promila.

Rozwiązanie

Stawiane hipotezy będą miały postać:

$H_0 : E(x) = 2{,}3$

$H_1 : E(x) > 2{,}3$

Założony poziom istotności wynosi 0,05. Dla wyznaczenia wartości statystyki empirycznej na podstawie uzyskanych danych ustalamy średnią $\overline{x}$ i odchylenie standardowe $s(x)$ stężenia alkoholu we krwi kierowców. Obliczenia pomocnicze zawarto w tablicy roboczej.

Poziom alkoholu we krwi (w promilach)	*Liczba kierowców*	$\dot{x}_i \cdot n_i$	$(\dot{x}_i - \overline{x})^2 \cdot n_i$
0,40 – 1,0	15	10,5	18,15
1,0 – 1,6	120	156,0	30,0
1,6 – 2,2	180	342,0	1,8
2,2 – 2,8	85	212,5	41,65
Razem	400	721	91,6

Otrzymujemy:

$$\bar{x} = \frac{721}{400} = 1{,}80 \text{ promila}$$

$$s(x) = \sqrt{\frac{91{,}6}{400}} = 0{,}48 \text{ promila}$$

Statystykę empiryczną obliczamy według *wzoru 5.11*:

$$t_{emp} = \frac{1{,}80 - 2{,}30}{0{,}48} \cdot \sqrt{400} = -20{,}83$$

Z tablic rozkładu normalnego odczytujemy wartość statystyki teoretycznej t_t dla $0{,}5 - 0{,}05 = 0{,}45$ (test ma charakter prawostronny). Wynosi ona 1,65. Zachodzi relacja:

$t_{emp} < t_t$, a więc nie ma podstaw do odrzucenia hipotezy zerowej, że średnie stężenie alkoholu we krwi nietrzeźwych kierowców, którzy spowodowali wypadek jest równe 2,3 promila.

5.4.3. Weryfikacja hipotezy dla dwóch średnich

Test dla dwóch średnich dotyczy weryfikacji hipotezy o równości średnich w dwóch populacjach o rozkładzie normalnym. W zależności od wielkości wylosowanych z tych populacji prób wyróżnia się dwa modele postępowania.

Model oparty na wynikach z dwóch małych prób.

Zakłada się, że rozkłady obu populacji są normalne o nieznanych wartościach średnich i nieznanych, ale jednakowych odchyleniach standardowych. Procedura weryfikacji odbywa się według następującego schematu:

1) stawiana jest hipoteza zerowa o postaci $H_0 : E_1(x) = E_2(x)$
   i jedna z niżej podanych postaci hipotezy alternatywnej:
   a) $H_1 : E_1(x) \neq E_2(x)$
   b) $H_1 : E_1(x) > E_2(x)$
   c) $H_1 : E_1(x) < E_2(x)$
   gdzie: $E_1(x)$ i $E_2(x)$ są hipotetycznymi wartościami średnimi dla pierwszej i drugiej populacji.
2) zakładany jest poziom istotności $1 - \alpha$,
3) z obu populacji generalnych losujemy dwie małe próby o liczebnościach n_1 i n_2; na ich podstawie wyznaczamy wartości średnie $\bar{x}_1$ i

$\bar{x}_2$ oraz wariancje $s_1^2(x)$ i $s_2^2(x)$, a w dalszej kolejności wartość statystyki empirycznej według wzoru

$$t_{emp} = \frac{\bar{x}_1 - \bar{x}_2}{\sqrt{\frac{n_1 s_1^2(x) + n_2 s_2^2(x)}{n_1 + n_2 - 2}\left(\frac{1}{n_1} + \frac{1}{n_2}\right)}} \qquad 5.12$$

Dla przyjętego poziomu istotności z tablic rozkładu t Studenta odczytujemy wartość statystyki teoretycznej według zasady:

a) dla testu dwustronnego: dla $k = n_1 + n_2 - 2$ oraz poziomu istotności $1-\alpha$,

b) dla testu jednostronnego: dla $k = n_1 + n_2 - 2$ oraz $2\cdot(1-\alpha)$.

Decyzję dotyczącą sprawdzanej hipotezy podejmujemy zgodnie z podanymi wskazówkami ogólnymi.

Przykład 5.10.

W badaniach absencji pracowniczej w pewnym przedsiębiorstwie w miesiącu lipcu zebrano informacje dla dwóch wylosowanych grup pracowników. Dla grupy 10 kobiet uzyskano następującą liczbę dni nieobecności w pracy: *0; 2; 3; 5; 7; 6; 8; 3; 5;1*, natomiast dla próby 12 mężczyzn odpowiednio: *0; 1; 2; 3; 2; 4; 3; 4; 7; 5; 6; 0*. Na poziomie istotności 0,05 zweryfikować hipotezę, że średnia dni nieobecności w pracy kobiet jest wyższa niż mężczyzn.

Rozwiązanie

Zgodnie z podanym schematem postępowania na wstępie stawiamy hipotezy o postaci

$$H_0 : E_1(x) = E_2(x)$$

$$H_1 : E_1(x) > E_2(x)$$

gdzie: subskryptem „*1*" oznaczono populację kobiet, natomiast „*2*" populację mężczyzn.

Zakładamy poziom istotności $1-\alpha$ = 0,05

Wartość statystyki empirycznej wyznaczamy według *wzoru 5.12*, co wymaga wyznaczenia średnich i wariancji absencji dla obu prób:

- dla kobiet: $\bar{x}_1 = \frac{40}{10} = 4$ dni i $s_1^2(x) = \frac{62}{10} = 6{,}2$ (dni)²

- dla mężczyzn: $\bar{x}_2 = \frac{37}{12} = 3{,}1$ dni i $s_2^2(x) = \frac{54{,}92}{12} = 4{,}58$ (dni)²

Podstawiając uzyskane wielkości do *wzoru 5.12* uzyskujemy wartość statystyki empirycznej:

$$t_{emp} = \frac{4{,}0 - 3{,}1}{\sqrt{\frac{10 \cdot 6{,}2 + 12 \cdot 4{,}58}{10+12-2}\left(\frac{1}{10}+\frac{1}{12}\right)}} = \frac{0{,}9}{\sqrt{\frac{116{,}92}{20} \cdot 0{,}18}} = \frac{0{,}9}{1{,}03} = 0{,}87$$

Z tablic rozkładu t Studenta odczytujemy wartość statystyki t_t dla $k = 10+12-2 = 20$ oraz 2*0,05 = 0,10 (z uwagi na fakt, że wykorzystujemy test jednostronny) i otrzymujemy t_t = 1,725.

Ponieważ zachodzi relacja $t_{emp} < t_t$ wobec tego przy poziomie istotności 0,05 nie ma podstaw do odrzucenia hipotezy zerowej, że średnia absencja kobiet jest identyczna jak absencja mężczyzn.

Model oparty na wynikach z dwóch dużych prób

Przyjmuje się, podobnie jak w poprzednim modelu, że obie populacje generalne posiadają rozkład normalny o nieznanych wariancjach. Po postawieniu hipotezy zerowej i alternatywnej i założeniu określonego poziomu istotności $1-\alpha$ z obu populacji generalnych losowane są dwie duże próby o liczebnościach n_1 i $n_2 > 30$. Na podstawie danych dla obu prób ustalamy średnie arytmetyczne $\bar{x}_1$ i $\bar{x}_2$ oraz wariancje $s_1^2(x)$ i $s_2^2(x)$. Parametry te wykorzystujemy do wyznaczenia statystyki empirycznej według wzoru

$$t_{emp} = \frac{\bar{x}_1 - \bar{x}_2}{\sqrt{\frac{s_1^2(x)}{n_1} + \frac{s_2^2(x)}{n_2}}} \qquad 5.13$$

Wartość statystyki teoretycznej t_t odczytujemy z tablic dystrybuanty rozkładu normalnego:

a) w przypadku testu dwustronnego - dla $0{,}5 - \frac{1-\alpha}{2}$,

b) w przypadku testu jednostronnego - dla $0{,}5 - (1-\alpha)$.

Końcowa czynność polegająca na podjęciu odpowiedniej decyzji odnośnie hipotezy zerowej jest podejmowana zgodnie z wcześniej podanymi zasadami.

Przykład 5.11.

W badaniach efektywności szkolenia zawodowego pracowników bezpośrednio produkcyjnych w pewnym przedsiębiorstwie dla losowo wybranej próby 60 pracowników dokonano pomiaru ich wydajności pracy

(w szt./zmianę) przed i po przejściu szkolenia. Uzyskano dane ujęte w tablicy 5.4.

Tablica 5.4. Pracownicy przedsiębiorstwa „Z" według wydajności pracy

Wydajność pracy w szt./zmianę	*Liczba pracowników*	
	przed szkoleniem	*po szkoleniu*
10 – 14	28	5
14 - 18	18	20
18 - 22	12	25
22 - 26	2	10

Źródło: Dane umowne

Zakładając, że w całej populacji pracowników wydajność pracy ma rozkład zbliżony do normalnego przy poziomie istotności 0,01 zweryfikować hipotezę, iż szkolenie zawodowe istotnie zwiększa wydajność pracy pracowników.

Rozwiązanie

Stawiane hipotezy będą miały postać:

$H_0 : E_1(x) = E_2(x)$

$H_1 : E_1(x) < E_2(x)$; (subskryptem „1" oznaczono populację przed odbyciem szkolenia, natomiast „2"- po jego odbyciu)

Przyjęty poziom istotności $1-\alpha$ wynosi 0,01. Wyznaczenie statystyki empirycznej wymaga obliczenia dla obu sytuacji (przed i po odbyciu szkolenia) średniej i wariancji wydajności pracy. Dokonamy tego w poniższej tablicy roboczej

Wydajność (x_i)	*Liczba pracowników*		$\dot{x}_i \cdot n_{1i}$	$\dot{x}_i \cdot n_{2i}$	$(\dot{x}_i - \bar{x}_1)^2 \cdot n_{1i}$	$(\dot{x}_i - \bar{x}_2)^2 \cdot n_{2i}$
	n_{1i}	n_{2i}				
10 - 14	28	5	336	60	286,72	224,45
14 - 18	18	20	288	320	11,52	145,8
18 - 22	12	25	240	500	276,48	42,25
22 – 26	2	10	48	240	163,68	280,9
Razem	60	60	912	1120	738,4	693,4

Otrzymujemy:

- przed odbyciem szkolenia: $\bar{x}_1 = \frac{912}{60} = 15{,}2$ szt., $s_1^2(x) = \frac{738{,}4}{60} = 12{,}3$ (szt.)²

- po odbyciu szkolenia: $\bar{x}_2 = \frac{1120}{60} = 18{,}7$ szt. i $s_2^2(x) = \frac{693{,}4}{60} = 11{,}6$ (szt.)²

Wartość statystyki empirycznej ustalimy według *wzoru 5.13*:

$$t_{emp} = \frac{15{,}2 - 18{,}7}{\sqrt{\frac{12{,}3}{60} + \frac{11{,}6}{60}}} = \frac{-3{,}5}{\sqrt{0{,}398}} = -5{,}56$$

Wartość statystyki teoretycznej t_t zostanie odczytana z tablic rozkładu normalnego dla $0{,}5 - 0{,}01 = 0{,}49$ (test ma charakter jednostronny) i wynosi ona 2,33.

Zachodzi relacja: $t_{emp} < -t_t$, co oznacza, że hipotezę zerową należy odrzucić, czyli przy poziomie istotności 0,01 można twierdzić, że szkolenie zawodowe istotnie wpływa na wzrost wydajności pracy

5.4.4. Weryfikacja hipotezy dla wskaźnika struktury

Ten typ hipotezy odnosi się najczęściej do przypadków badania populacji generalnej ze względu na cechę opisową (por. również estymacja przedziałowa wskaźnika struktury). Wnioskowanie dotyczy wówczas głównie jej struktury. Zakłada się, że populacja ta ma rozkład dwupunktowy o parametrach *p* i *q*, gdzie *p* jest wskaźnikiem struktury dla wyróżnionych elementów populacji. Stawiana jest hipoteza zerowa o postaci: $H_0 : p = p_0$, która oznacza, że wskaźnik struktury w populacji przyjmie pewną hipotetyczną wartość p_0. Wobec niej stawiana może być jedna z trzech postaci hipotezy alternatywnej:

a) $H_1: p \neq p_0$

b) $H_1: p > p_0$

c) $H_1: p < p_0$.

Zakładany jest poziom istotności $1-\alpha$. Z populacji losowana jest duża próba o liczebności przekraczającej 100 elementów. Na podstawie wyników z próby obliczana jest wartość statystyki empirycznej według wzoru:

$$t_{emp} = \frac{\frac{m}{n} - p_0}{\sqrt{\frac{p_0 \cdot q_0}{n}}} \qquad 5.14$$

gdzie: n – liczebność próby,

m – liczba wyróżnionych elementów w próbie,

p_0 – hipotetyczny wskaźnik struktury dla wyróżnionych elementów,

$q_0 = 1 - p_0$

Statystykę teoretyczną odczytujemy z tablic dystrybuanty rozkładu normalnego:

a) w przypadku testu dwustronnego - dla $0{,}5-\frac{1-\alpha}{2}$,

b) w przypadku testu jednostronnego - dla $0{,}5-(1-\alpha)$.

Decyzję dotyczącą postawionej hipotezy podejmujemy zgodnie z ogólnymi zasadami.

Przykład 5.12.

Wśród mieszkańców pewnego miasta przeprowadzono badanie ankietowe dotyczące ulubionego sposobu spędzania wolnego czasu. Uzyskano dane zawarte w poniższym zestawieniu:

Sposób spędzania wolnego czasu	*Liczba odpowiedzi*
Oglądanie telewizji, słuchanie radia	120
Czytanie prasy, książek	60
Czynny wypoczynek (zajęcia sportowe)	55
Sen	5

Przy poziomie istotności 0,10 zweryfikować hipotezę, że odsetek osób czynnie spędzających wolny czas wynosi 0,30.

Rozwiązanie

Stawiamy hipotezy o postaci:

$H_0 : p = 0{,}30$

$H_1 : p \neq 0{,}30$

Przyjęty poziom istotności wynosi 0,10.

Ustalamy wskaźnik struktury dla czynnie wypoczywających w próbie: $\frac{m}{n} = \frac{55}{240} = 0{,}229$. Wielkość tę podstawiamy do *wzoru 5.14* i otrzymujemy

$$t_{emp} = \frac{0{,}229 - 0{,}30}{\sqrt{\frac{0{,}30 \cdot 0{,}70}{240}}} = \frac{-0{,}071}{0{,}030} = -2{,}37$$

Statystykę teoretyczną t_t odczytujemy z tablic rozkładu normalnego dla $0{,}5-\frac{0{,}1}{2} = 0{,}45$ (test ma charakter dwustronny) i wynosi ona 1,65. Zachodzi relacja $|t_{emp}| > t_t$, co oznacza, że hipotezę zerową należy odrzucić, czyli odsetek osób czynnie wypoczywających jest różny od 0,30 (tj. 30 %).

5.4.5. Weryfikacja hipotezy dla współczynnika korelacji

Omawiany test służy weryfikacji hipotezy, że między dwiema cechami populacji generalnej występuje niezależność w sensie parametrycznym. Przyjmuje się założenie, że rozkład badanych cech jest przynajmniej zbliżony do normalnego. Stawiana jest hipoteza zerowa o postaci: $H_0 : \rho = 0$, zakładająca, że pomiędzy badanymi zmiennymi w populacje generalnej występuje niezależność w sensie parametrycznym wobec jednej z poniższych wersji hipotezy alternatywnej:

a) $H_1 : \rho \neq 0$ (występuje zależność w sensie parametrycznym),

b) $H_1 : \rho > 0$ (występuje zależność o kierunku dodatnim),

c) $H_1 : \rho < 0$ (występuje zależność o kierunku ujemnym).

Kolejna czynność dotyczy założenia określonego poziomu istotności $1-\alpha$. Następnie z populacji generalnej losowana jest mała próba. Na podstawie uzyskanych dla niej wyników przy pomocy współczynnika korelacji liniowej Pearsona r^P (w wersji dla szeregów szczegółowych) ustalana jest siła i kierunek zależności między badanymi cechami. Uzyskaną wartość współczynnika wykorzystujemy dla wyznaczenia statystyki empirycznej według wzoru:

$$t_{emp} = \frac{r^P}{\sqrt{1-\left(r^P\right)^2}} \cdot \sqrt{n-2} \qquad 5.15$$

Graniczną wartość statystyki teoretycznej odczytujemy z tablic rozkładu t Studenta:

a) w przypadku testu dwustronnego: dla $k = n-2$ oraz poziomu istotności $1-\alpha$,

b) w przypadku testu jednostronnego: dla $k = n-2$ oraz $2 \cdot (1-\alpha)$.

Decyzję dotyczącą prawdziwości hipotezy zerowej podejmujemy zgodnie z ogólnymi zasadami.

Należy podkreślić, że podana procedura może być wykorzystana jedynie w warunkach stosowalności współczynnika korelacji liniowej Pearsona, tzn. obie cechy muszą mieć charakter liczbowy, a związek między nimi prostoliniowy. W pozostałych przypadkach należy stosować prezentowany dalej test niezależności.

Przykład 5.13.

Dla losowej próby 20 małżeństw zebrano informacje dotyczące wieku współmałżonków w momencie zawierania przez nich związku małżeńskiego i przy pomocy współczynnika korelacji liniowej Pearsona zbadano zależność ich wieku. Uzyskano $r^P = 0{,}80$. Zweryfikować hipotezę, że istnieje istotna dodatnia zależność między wiekiem kobiet i mężczyzn wstępujących w związek małżeński. Przyjąć poziom istotności 0,01

Rozwiązanie

Stawiamy hipotezy o postaci:

$H_0 : \rho = 0$, tzn. między badanymi cechami występuje niezależność,

$H_0 : \rho > 0$,czyli między badanymi cechami występuje zależność dodatnia.

Założono poziom istotności $1-\alpha$ = 0,01. Wartość statystyki empirycznej ustalamy według *wzoru 5.15*:

$$t_{emp} = \frac{0{,}80}{\sqrt{1-(0{,}80)^2}} \cdot \sqrt{20-2} = 5{,}65$$

Statystykę teoretyczną odczytujemy z tablic rozkładu t Studenta dla k = 20 – 2 = 18 oraz $2 \cdot 0{,}01 = 0{,}02$; otrzymujemy $t_t = 2{,}552$. Zachodzi relacja: $t_{emp} > t_t$, co oznacza, że przy poziomie istotności 0,01 można twierdzić, iż występuje istotna dodatnia zależność między wiekiem osób zawierających związek małżeński.

5.4.6. Test niezależności χ^2 (chi-kwadrat)

Test ten służy weryfikacji hipotezy, że dwie zmienne opisujące populację generalną są niezależne. Stawiana hipoteza zerowa ma postać: $H_0 : f_{ij} = f_i \cdot f_j$ i zakłada niezależność badanych zmiennych. Zauważmy, iż został w niej wykorzystany warunek niezależności cech w sensie nieparametrycznym. Alternatywna wobec niej hipoteza zakłada występowanie zależności i ma postać: $H_1 : f_{ij} \neq f_i \cdot f_j$. Z populacji generalnej losowana jest duża próba, a wyniki dla niej uzyskane ujmowane są w postaci tablicy korelacyjnej o *l* wierszach i *s* kolumnach. Liczebność próby (przy uwzględnieniu liczby wariantów obu cech) winna być na tle duża, by każde n_{ij} było nie mniejsze od 8. Zakłada się poziom istotności $1-\alpha$. Na podstawie tablicy korelacyjnej wyznaczana jest wartość statystyki empirycznej według wzoru:

$$t_{emp} = \sum_{i,j} \frac{\left(n_{ij} - N \cdot f_i \cdot f_j\right)^2}{N \cdot f_i \cdot f_j} \qquad 5.16$$

Wartość statystyki teoretycznej t_t odczytywana jest z tablic rozkładu χ^2 dla $k = (l-1)\cdot(s-1)$ oraz poziomu istotności $1-\alpha$. Gdy zachodzi relacja $t_{emp} \geq t_t$ odrzucamy hipotezę zerową o niezależności cech w populacji generalnej; w przeciwnym przypadku występuje brak podstaw do jej odrzucenia.

Przykład 5.14.

Dla losowej próby bezrobotnych zarejestrowanych w Powiatowym Urzędzie Pracy w „K" zebrano informacje dotyczące ich poziomu wykształcenia (X) oraz czasu pozostawania bez pracy (Y). Wyniki badania ujęto w poniższej tablicy korelacyjnej.

Tablica 5.4. Bezrobotni zarejestrowani w Powiatowym Urzędzie Pracy w „K" według poziomu wykształcenia i czasu pozostawania bez pracy.

Czas pozostawania bez pracy w miesiącach	*Poziom wykształcenia*			n_i
	podstawowe	*średnie*	*wyższe*	
do 6	15	15	15	45
6 - 12	25	25	10	60
12 - 24	30	15	10	55
n_j	70	55	35	160

Źródło: Dane umowne

Na poziomie istotności 0,05 zweryfikować hipotezę o niezależności czasu pozostawania bez pracy od poziomu wykształcenia bezrobotnych.

Rozwiązanie

Stawiamy hipotezę zerową o niezależności czasu pozostawania bez pracy od poziomu wykształcenia bezrobotnych o postaci $H_0 : f_{ij} = f_i \cdot f_j$ i hipotezę wobec niej alternatywną $H_1 : f_{ij} \neq f_i \cdot f_j$ zakładającą, że taka zależność występuje.

Statystykę empiryczną obliczamy w poniższej tablicy roboczej zgodnie z *wzorem 5.16* wykonując następujące działania (ich kolejność ponumerowano w pierwszym wierszu poniższej tablicy roboczej):

1) przekształcenie rozkładów brzegowych liczebności w rozkłady częstości,
2) ustalenie iloczynów częstości brzegowych $f_i \cdot f_j$ dla każdego pola tablicy korelacyjnej,
3) określenie dla każdego pola tablicy liczebności hipotetycznych poprzez wyznaczenie iloczynów $N \cdot f_i \cdot f_j$,
4) ustalenie dla każdego pola tablicy wielkości różnic liczebności empirycznych i hipotetycznych , a następnie kwadratów tych różnic zgodnie z formułą $\left(n_{ij} - N \cdot f_i \cdot f_j\right)^2$,
5) określenie dla każdego pola tablicy ilorazu $\dfrac{\left(n_{ij} - N \cdot f_i \cdot f_j\right)^2}{N \cdot f_i \cdot f_j}$, a następnie ich sumy.

Czas pozostawania bez pracy w miesiącach	*Poziom wykształcenia*			f_i
	podstawowe	*średnie*	*wyższe*	
do 6	15 ***2)*** 0,123 ***3)*** 19,7 ***4)*** 22,09 ***5)*** 1,12	15 0,097 15,5 0,25 0,02	15 0,061 9,8 27,04 2,76	***1)*** 0,281
6 - 12	25 0,164 26,2 1,44 0,05	25 0,129 20,6 19,36 0,94	10 0,082 13,1 9,61 0,73	0,375
12 - 24	30 0,151 24,2 33,64 1,39	15 0,118 18,9 15,2 0,80	10 0,075 12 4 0,33	0,344
f_j	0,438	0,344	0,218	1,00

Na podstawie wykonanych obliczeń otrzymujemy zgodnie z *wzorem 5.16* $t_{emp} = 1{,}12 + 0{,}02 + 2{,}76 + 0{,}05 + 0{,}94 + 0{,}73 + 1{,}39 + 0{,}80 + 0{,}33 =$ 8,14.

Dla $k = (3-1)\cdot(3-1) =$ 4 oraz 1-$\alpha = 0{,}05$ z tablic rozkładu χ^2 odczytujemy wartość statystyki t_t = 9,488. Ponieważ zachodzi relacja $t_{emp} < t_t$ stwierdzamy brak podstaw do odrzucenia hipotezy zerowej, co oznacza, że przy poziomie istotności 0,05 można twierdzić, iż występuje niezależność czasu pozostawania bez pracy od poziomu wykształcenia bezrobotnych.

5.4.7. Test zgodności χ^2 (chi-kwadrat)

Może być on wykorzystywany do weryfikacji hipotez dwojakiego rodzaju:

a) populacja posiada określony typ rozkładu,
b) dwie wylosowane próby pochodzą z populacji o takim samym rozkładzie.

Rozważania ograniczymy do pierwszego przypadku. Stawiana jest w tym przypadku hipoteza zerowa, że dystrybuanta empiryczna $F(x)$, ustalana na podstawie wyników z wylosowanej dużej próby, jest zgodna z dystrybuantą teoretyczną $F_0(x)$ określonego typu rozkładu; można to wyrazić zapisem: $H_0 : F(x) = F_0(x)$. Wobec tak sformułowanej hipotezy stawiana jest hipoteza alternatywna: $H_1 : F(x) \neq F_0(x)$. Zgromadzony na podstawie wylosowanej próby materiał statystyczny ujmowany jest w postaci szeregu rozdzielczego punktowego bądź przedziałowego. Liczebność próby, przy

uwzględnieniu liczby klas, winna być tak dobrana, by liczebność każdej z klas była nie mniejsza niż 5. Następnie zakłada się poziom istotności $1-\alpha$. Wartość statystyki empirycznej ustalana jest według wzoru:

$$t_{emp} = \sum_i \frac{(n_i - N \cdot p_i)^2}{N \cdot p_i} \qquad 5.17$$

gdzie: n_i - oznacza liczebność i-tej klasy,

p_i - prawdopodobieństwo teoretyczne, że badana zmienna przyjmie wartości należące do i-tej klasy; prawdopodobieństwa te mogą być odczytywane z tablic odpowiedniego rozkładu teoretycznego.

Analizując powyższy wzór należy zauważyć, iż ma w tym przypadku miejsce porównywanie szeregu liczebności empirycznych z oszacowanymi liczebnościami hipotetycznymi (teoretycznymi). Statystykę teoretyczną t_t odczytujemy z tablic rozkładu χ^2 dla $k = r-1$ lub $k = r-l-1$ (gdzie: r – liczba klas w szeregu rozdzielczym, l – liczba szacowanych z próby parametrów) i poziomu istotności $1-\alpha$. Końcową decyzję podejmujemy zgodnie z ogólnymi zasadami.

Przykład 5.15.

W badaniach warunków życia mieszkańców pewnego miasta zebrano m. in. informacje o wysokości dochodów przypadających na 1 członka gospodarstwa domowego. Dla losowej próby 200 gospodarstw uzyskano następujące wyniki badań:

Dochód na 1 osobę w zł	*Liczba gospodarstw*
150 – 350	5
350 – 550	25
550 – 750	80
750 - 950	70
950 - 1150	15
1150 - 1350	5

Na poziomie istotności 0,01 zweryfikować hipotezę, że rozkład dochodów w gospodarstwach domowych ma charakter rozkładu normalnego.

Rozwiązanie

Stawiana jest hipoteza zerowa o postaci $H_0 : F(x) = F_0(x)$ zakładająca, że rozkład dochodów ma charakter rozkładu normalnego i przeciwstawna niej hipoteza alternatywna $H_1 : F(x) \neq F_0(x)$. Z uwagi na dużą próbę, wartość średnią i odchylenie standardowe dochodów ustalone z próby, możemy przyjąć jako parametry rozkładu normalnego. Otrzymujemy:

$$E(x) = \bar{x} = \frac{146000}{200} = 730 \text{ zł}$$

$$\sigma(x) = s(x) = \sqrt{\frac{6599000}{200}} = 181{,}6 \text{ zł}$$

W wyniku tych ustaleń hipotetyczny rozkład normalny posiadałby parametry: N(730 zł; 181,6 zł). Dalsze obliczenia pomocnicze dla wyznaczenia statystyki empirycznej zgodnie z *wzorem 5.17* zostaną wykonane w poniższej tablicy roboczej, w której:

- w kolumnie 1. poszczególne przedziały klasowe zastąpiono ich górnymi krańcami,
- w kolumnie 2. podano liczebności empiryczne poszczególnych klas,
- w kolumnie 3. dokonano standaryzacji górnych końców przedziałów klasowych według formuły: $t_i = \frac{x_i - E(x)}{\sigma(x)}$,
- w kolumnie 4. umieszczono wartości dystrybuanty teoretycznej rozkładu normalnego dla poszczególnych t_i odczytane z tablic rozkładu normalnego,
- w kolumnie 5. na podstawie odczytanych wartości dystrybuanty ustalono prawdopodobieństwa teoretyczne uzyskania dochodów mieszczących się w poszczególnych przedziałach klasowych,
- w kolumnie 6. ustalono teoretyczne liczebności dla poszczególnych klas,
- w kolumnie 7. dokonano obliczenia statystyki empirycznej.

Dochód na 1 osobę w zł (x_i)	*Liczba gospodarstw* (n_i)	t_i	$F_{0i}(x) = F(t_i)$	p_i	$N \cdot p_i$	$\frac{(n_i - N \cdot p_i)^2}{N \cdot p_i}$
1	*2*	*3*	*4*	*5*	*6*	*7*
350	5	- 2,09	0,0183	0,0183	3,7	1,76
550	25	- 0,99	0,1611	0,1428	28,6	0,45
750	80	0,11	0,5438	0,3827	76,5	0,16
950	70	1,21	0,8869	0,3431	68,6	0,03
1150	15	2,31	0,9896	0,1027	20,5	1,48
1350	5	3,41	~ 1,00	0,0104	2,1	4,00
Razem	200	X	X	1,0000	X	7,88

Wartość statystyki empirycznej t_{emp} wynosi 7,88. Statystykę teoretyczną t_t odczytujemy z tablic rozkładu χ^2 dla k = 6 – 2 – 1 = 3 i poziomu istotności $1-\alpha$ = 0,01. Otrzymujemy t_t = 11,345. Zachodzi relacja: $t_{emp} < t_t$, wobec czego przy poziomie istotności 0,01nie ma podstaw do odrzucenia

hipotezy, że rozkład dochodów na jedną osobę w gospodarstwach domowych ma charakter rozkładu normalnego

5.4.8. Test zgodności Kołmogorowa

Ma on podobny charakter do wyżej omawianego testu. Zadaniem testu Kołmogorowa jest weryfikacja hipotezy o zgodności rozkładu określonej populacji z rozkładem normalnym. Badanie zgodności odbywa się poprzez porównywanie wartości dystrybuanty empirycznej i dystrybuanty hipotetycznej rozkładu normalnego. Test ten ma zastosowanie do zmiennych typu ciągłego, dla innego typu zmiennych należy wykorzystać podany wyżej test zgodności χ^2.

Stawiana na wstępie hipoteza zerowa ma postać $H_0 : F(x) = F_0(x)$, gdzie dystrybuanta empiryczna F(x) ustalana na podstawie wyników z wylosowanej dużej próby, zaś $F_0(x)$ jest dystrybuantą teoretyczną rozkładu normalnego. Zakłada ona, że rozkład badanej zmiennej w populacji generalnej jest zgodny z rozkładem normalnym. Wobec tak sformułowanej hipotezy stawiana jest hipoteza alternatywna o postaci $H_1 : F(x) \neq F_0(x)$ o braku takiej zgodności. Z populacji generalnej losowana jest duża próba, a jej wyniki ujmowane są w szeregu rozdzielczym przedziałowym. Zalecane jest tworzenie dużej liczby klas, gdyż daje to możliwość badania zgodności w wielu punktach. Dla utworzonego szeregu wyznaczamy wartości dystrybuanty empirycznej $F(x)$. Tworzy je szereg częstości skumulowanej. Duża próba pozwala na przyjęcie jej średniej $(\bar{x})$ i odchylenia standardowego $[s(x)]$ jako parametrów rozkładu normalnego $E(x)$ i $\sigma(x)$. Z tablic dystrybuanty rozkładu normalnego dla górnych krańców poszczególnych przedziałów klasowych odczytujemy wartości dystrybuanty hipotetycznej $F_0(x)$. W dalszej kolejności porównujemy parami wartości obu dystrybuant i maksymalna różnica między nimi stanowi podstawę do ustalenia statystyki empirycznej zgodnie z wzorem:

$$t_{emp} = D \cdot \sqrt{n} \qquad 5.18$$

gdzie: $D = \max |F_i(x) - F_{0i}(x)|$ oznacza maksymalną różnicę odpowiadających sobie wartości dystrybuant empirycznej i teoretycznej,
n – liczebność wylosowanej próby.

Wartość statystyki teoretycznej t_t - przy założeniu poziomu istotności $1-\alpha$ - odczytujemy z tablic granicznego rozkładu Kołmogorowa dla α. Jeśli zachodzi relacja: $t_{emp} \geq t_t$ hipotezę zerową należy odrzucić, w przeciwnym przypadku brak jest podstaw do jej odrzucenia, co oznacza występowanie zgodności rozkładu badanej zmiennej w populacji generalnej z rozkładem

normalnym. Należy również dodać, że istnieje odmiana tego testu pozwalająca na weryfikację hipotezy o zgodności rozkładów dwóch populacji określana mianem testu zgodności Kołmogorowa – Smirnowa.

Przykład 5.16.

Na podstawie danych z *przykładu 5.15* - przy poziomie istotności 0,05 - zweryfikować hipotezę, że rozkład dochodów w całej populacji gospodarstw domowych jest normalny.

Rozwiązanie

Stawiane hipotezy mają postać identyczną jak w *przykładzie 5.15*, tj. $H_0 : F(x) = F_0(x)$ i $H_1 : F(x) \neq F_0(x)$. Z uwagi na dużą próbę – podobnie jak poprzednio - średnią i odchylenie standardowe z próby możemy przyjąć jako parametry rozkładu normalnego. Wobec tego hipotetyczny rozkład normalny posiadać będzie parametry: N(730 zł; 181,6 zł). Dalsze obliczenia pomocnicze dla wyznaczenia statystyki empirycznej zgodnie z *wzorem 5.18* zostały wykonane w poniższej tablicy roboczej, w której:

- w kolumnie 1.poszczególne przedziały klasowe zastąpiono ich górnymi krańcami,
- w kolumnie 2. podano liczebności empiryczne poszczególnych klas,
- w kolumnie 3. dokonano standaryzacji górnych końców przedziałów klasowych według formuły: $t_i = \dfrac{x_i - E(x)}{\sigma(x)}$,
- w kolumnie 4. umieszczono wartości dystrybuanty teoretycznej rozkładu normalnego dla poszczególnych t_i odczytane z tablic dystrybuanty rozkładu normalnego,
- w kolumnie 5. umieszczono wartości dystrybuanty empirycznej odpowiadające częstościom skumulowanym,
- w kolumnie 6. ustalono bezwzględne odchylenia wartości dystrybuant empirycznej i teoretycznej.

Dochód na 1 osobę w zł (x_i)	*Liczba gospodarstw* (n_i)	t_i	$F_{0i}(x) = F(t_i)$	$F_i(x) = cumf_i$	$\lvert F_{0i} - F_i(x) \rvert$
1	*2*	*3*	*4*	*5*	*6*
350	5	- 2,09	0,0183	0,025	0,0067
550	25	- 0,99	0,1611	0,15	0,0111
750	80	0,11	0,5438	0,55	0,0062
950	70	1,21	0,8869	0,90	0,0131
1150	15	2,31	0,9896	0,975	***0,0146***
1350	5	3,41	~ 1,00	1,00	0
Razem	200	X	X	X	X

Na podstawie obliczeń wykonanych w ostatniej kolumnie otrzymujemy $D = \max\left|F_{0i} - F_i(x)\right| = 0{,}0146$. Podstawiając tę wartość do *wzoru 5.18* uzyskujemy: $t_{emp} = 0{,}0146 \cdot \sqrt{200} = 0{,}206$. Statystykę teoretyczną odczytujemy z tablic granicznego rozkładu Kołmogorowa *(tablica 4 w Aneksie)* dla $1 - 0{,}05 = 0{,}95$. Wynosi ona 1,36. Zachodzi relacja $t_{emp} < t_t$, a więc przy poziomie istotności 0,05 można przyjąć, że rozkład dochodów w badanej populacji jest normalny.

5.4.9. Test serii

Ten rodzaj testu posiada szerokie zastosowanie w procedurach weryfikacji hipotez statystycznych. Może być stosowany do weryfikacji hipotez:

a) o losowości próby,
b) o liniowej postaci funkcji regresji,
c) że dwie populacje posiadają ten sam typ rozkładu.

Dalsze rozważania ograniczymy do pierwszego przypadku, ponieważ warunek losowości próby jest podstawą metod wnioskowania statystycznego. *Serią* określa się każdy podciąg kolejnych wyrazów ciągu *n*- elementowego, który ma identyczne wartości oraz który poprzedza, ewentualnie za którym występuje inna wartość niż w określonym podciągu. Jako szczególny przypadek serii można przyjąć ciąg elementów pobieranych do próby. Test ten jest szczególnie zalecany, gdy elementy te są pobierane w pewnych momentach czasowych, a w miarę upływu czasu istnieje możliwość zmiany rozkładu populacji bądź zmiany prawdopodobieństwa wylosowania kolejnych elementów.

Populacja generalna może mieć dowolny rozkład. Pobierana jest z niej próba licząca *n* elementów. Dla uzyskanych wyników z próby ujętych w szeregu szczegółowym wyznaczamy wartość mediany według zasad poznanych *w rozdziale II.* Następnie w szeregu pierwotnym (nieuporządkowanym) każdemu wynikowi x_i spełniającemu warunek $x_i < Me(x)$ przypisujemy symbol *„a”*, natomiast gdy $x_i > Me(x)$ - symbol *„b”*. W ten sposób pierwotny ciąg wyrazów $x_1, x_2, ..., x_n$ zostaje zastąpiony ciągiem symboli *„a” i „b”*. W ciągu tym ustalamy liczbę serii (podciągów składających się z jednakowych symboli) oznaczaną dalej jako k. Liczbę tę należy traktować jako statystykę empiryczną. Statystykę teoretyczną (hipotetyczną liczbę serii) wyznaczamy z tablic rozkładu serii określając dwie wielkości k_1 i k_2 w następujący sposób:

- k_1 dla n_1 i n_2 oraz $\dfrac{1-\alpha}{2}$,
- k_2 dla n_1 i n_2 oraz $1-\dfrac{1-\alpha}{2}$,

gdzie: n_1 i n_2 odpowiadają liczbie występujących w ciągu symboli *„a”* i *„b”*.

Jeśli spełniona jest relacja, że:

$$k_1 < k < k_2 \qquad 5.20$$

wówczas nie ma podstaw do odrzucenia hipotezy o losowości próby. W przeciwnym przypadku hipotezę taką należy odrzucić.

Przykład 5.17.

W badaniach wyników studiowania osiąganych przez studentów pewnej uczelni z ich populacji wylosowano próbę 25 studentów, dla której ustalono następujące średnie z całego toku studiów: *3,11; 4,05; 3,75; 3,33; 4,25; 3,15; 3,96; 4,02; 2,99; 3,28; 3,65; 4,12; 3,48; 3,73; 3,26; 2,87; 4,54; 3,24; 4,15; 3,66; 3,74; 4,28; 3,90; 3,45; 4,67*. Na poziomie istotności 0,10 zweryfikować hipotezę, że dobór próby był losowy.

Rozwiązanie

Dla uzyskanych wyników ustalamy wartość mediany $Me(x)$ zgodnie z zasadami obowiązującymi dla szeregów szczegółowych. W analizowanym przypadku medianą jest trzynasta w kolejności (po uprzednim uporządkowaniu) średnia i wynosi ona 3,73. Uzyskane wyniki zastępujemy symbolami: gdy $x_i < Me(x)$ przypisujemy symbol „*a*”, natomiast gdy $x_i > Me(x)$ - symbol „*b*”. W ten sposób otrzymujemy następujący ciąg symboli:

abbababbaaabaaababababbbab,

w którym liczba serii *k* wynosi 16. Liczba elementów „*a*” wynosi 12 i elementów „*b*” również 12. Z tablic rozkładu liczby serii (*tablica 5. w Aneksie)* odczytujemy:

- k_1 dla n_1=12 i n_2=12 oraz $\frac{0{,}10}{2} = 0{,}05$; wynosi ono 8
- k_2 dla n_1=12 i n_2=12 oraz $1 - \frac{0{,}10}{2} = 0{,}95$; otrzymujemy 17

Zachodzi relacja $k_1 < k < k_2$, co oznacza, że dobór próby był losowy.

BIBLIOGRAFIA

[1] Badania marketingowe. Podstawowe metody i obszary zastosowań. Pod red. K. Mazurek-Łopacińskiej. Wrocław: AE 2002

[2] Bazarnik J., Grabiński T. (i inni); Badania marketingowe. Metody i oprogramowanie komputerowe. Warszawa – Kraków: Canadian Consortium of Management Schools. AE w Krakowie

[3] Bąk I., Mankowicz I., Mojsiewicz M., Wawrzyniak K.: Statystyka w zadaniach. Cz. I i II. Warszawa: WN-T 2001

[4] Blalock H. M.: Statystyka dla socjologów. PWN Warszawa 1975

[5] Clauss G., Ebnmer H.: Podstawy statystyki dla psychologów, pedagogów i socjologów. Warszawa: PZWS 1972

[6] Forlicz S. Rozkłady asymetryczne zmiennej losowej. Prace Naukowe AE we Wrocławiu nr 348. Wrocław: AE 1986

[7] Góralski A.: Metody opisu i wnioskowania statystycznego w psychologii. Warszawa: PWN 1976

[8] Greń J.: Statystyka matematyczna. Modele i zadania. Warszawa: PWE 1974

[9] Gruziński R.: Opis statystyczny w badaniach prawoznawczych. Warszawa: Wydawnictwo Prawnicze 1981

[10] Krzysztofiak M., Urbanek D.: Metody statystyczne. Warszawa: PWE 1981

[11] Lange O., Banasiński A.: Teoria statystyki. Warszawa: PWE 1970

[12] Luszniewicz A.: Statystyka ogólna. Warszawa: PWE 1987

[13] Makać W., Urbanek-Krzysztofiak D.: Metody opisu statystycznego. Gdańsk: Wyd. Uniwersytetu Gdańskiego 1995

[14] Nowaczyk C.: Podstawy metod statystycznych dla pedagogów. Jelenia Góra: VIS 1995

[15] Ostasiewicz S., Rusnak Z., Siedlecka U.: Statystyka. Materiały do ćwiczeń. Książka ekonomiczna. Wrocław: AE 1994

[16] Rocznik Statystyczny RP. Warszawa: GUS 2001

[17] Rusnak Z.: Problemy badań ankietowych. w: Metody ilościowe w ekonomii. Praca zbiorowa pod red. W. Ostasiewicza. Wrocław: AE 1999

[18] Sadowski W.: Statystyka dla ekonomistów. Wnioskowanie statystyczne. Warszawa: PWE 1972

[19] Sadowski W.: Statystyka matematyczna. Warszawa: PWE 1969

[20] Sobczyk M.: Statystyka. Warszawa: PWE 1997

[21] Szulc B.: Statystyka dla ekonomistów. Opis statystyczny. Warszawa: PWE 1968

[22] Walter J., McLean M.: Statystyka dla każdego. Warszawa: WSiP 1994

[23] Wawrzynek J.: Wybrane metody opisu i wnioskowania statystycznego w biznesie. Wrocław: AE 1995

[24] Zając K.: Zarys metod statystycznych. Warszawa: PWE 1994

[25] Zastosowanie metod statystycznych. Praca zbiorowa pod red. A. Luszniewicza. Warszawa: PWE 1983

PODSTAWOWE WZORY

Wzór	*Zastosowanie*
$k \approx \sqrt{N}$ □	*Liczba klas*
$h = \frac{R(x)}{k}$	*Rozpiętość klas*
$M_1(x) = \frac{\sum_{i=1}^{l} x_i \cdot n_i}{N} = \bar{x}$	*Moment zwykły rzędu pierwszego*
$m_2(x) = \frac{\sum_{i=1}^{k} (x_i - \bar{x})^2 \cdot n_i}{N}$	*Moment centralny rzędu drugiego*
$m_3(x) = \frac{\sum_{i=1}^{k} (x_i - \bar{x})^3 \cdot n_i}{N}$	*Moment centralny rzędu trzeciego*
$m_4(x) = \frac{\sum_{i=1}^{k} (x_i - \bar{x})^4 \cdot n_i}{N}$	*Moment centralny rzędu czwartego*
$\bar{x} = \frac{\sum_{i=1}^{N} x_i}{N}$	*Średnia arytmetyczna w szeregu szczegółowym*
$\bar{x} = \frac{\sum_{i=1}^{k} x_i \cdot n_i}{N}$; $\bar{x} = \frac{\sum_{i=1}^{k} x_i \cdot f_i}{\sum_{i=1}^{k} f_i}$	*Średnia arytmetyczna w szeregu rozdzielczym punktowym*
$\bar{x} = \frac{\sum_{i=1}^{k} \dot{x}_i \cdot n_i}{N}$; $\bar{x} = \frac{\sum_{i=1}^{k} \dot{x}_i \cdot f_i}{\sum_{i=1}^{k} f_i}$	*Średnia arytmetyczna w szeregu rozdzielczym z przedziałami klasowymi*
$x_g = \sqrt[N]{x_1 \cdot x_2 \cdot \ldots \cdot x_N}$	*Średnia geometryczna*
$\bar{y}_{chr} = \frac{\frac{1}{2} y_1 + y_2 + \ldots + \frac{1}{2} y_n}{n-1}$	*Średnia chronologiczna*

$D(x) = x_0 + \frac{n_0 - n_{0-1}}{(n_0 - n_{0-1}) + (n_0 - n_{0+1})} \cdot h_0$	*Dominanta w szeregu rozdzielczym z przedziałami klasowymi*
$Me(x) = x_0 + \frac{\frac{N}{2} - cum_{n_{0-1}}}{n_0} \cdot h_0$	*Mediana w szeregu rozdzielczym z przedziałami klasowymi*
$Q_1(x) = x_0 + \frac{\frac{N}{4} - cum_{n_{0-1}}}{n_0} \cdot h_0$	*Kwartyl pierwszy w szeregu rozdzielczym z przedziałami klasowymi*
$Q_3(x) = x_0 + \frac{\frac{3N}{4} - cum_{n_{0-1}}}{n_0} \cdot h_0$	*Kwartyl trzeci w szeregu rozdzielczym z przedziałami klasowymi*
$R(x) = x_{\max} - x_{\min}$	*Obszar zmienności*
$O_c(x) = \frac{Q_3(x) - Q_1(x)}{2}, O_c(x) = \frac{[Q_3(x) - Me(x)] + [Me(x) - Q_1(x)]}{2}$	*Odchylenie ćwiartkowe*
$s^2(x) = \frac{\sum_{i=1}^{N} (x_i - \bar{x})^2}{N}$	*Wariancja w szeregu szczegółowym*
$s^2(x) = \frac{\sum_{i=1}^{k} (x_i - \bar{x})^2 \cdot n_i}{N}$	*Wariancja w szeregu rozdzielczym punktowym*
$s^2(x) = \frac{\sum_{i=1}^{k} (\dot{x}_i - \bar{x})^2 \cdot n_i}{N}$	*Wariancja w szeregu rozdzielczym z przedziałami klasowymi*
$s^2(x) = \bar{s}_j^{\,2}(x) + s^2(\bar{x}_j)$	*Równość wariancyjna*
$s(x) = \sqrt{\frac{\sum_{i=1}^{N} (x_i - \bar{x})^2}{N}}$	*Odchylenie standardowe w szeregu szczegółowym*

$s(x)=\sqrt{\frac{\sum_{i=1}^{k}(x_i-\bar{x})^2\cdot n_i}{N}}$	*Odchylenie standardowe w szeregu rozdzielczym punktowym*
$s(x)=\sqrt{\frac{\sum_{i=1}^{k}(\dot{x}_i-\bar{x})^2\cdot n_i}{N}}$	*Odchylenie standardowe w szeregu rozdzielczym z przedziałami klasowymi*
$d(x)=\frac{\sum_{i=1}^{N}\lvert x_i-\bar{x}\rvert}{N}$	*Odchylenie przeciętne w szeregu szczegółowym*
$d(x)=\frac{\sum_{i=1}^{k}\lvert x_i-\bar{x}\rvert\cdot n_i}{N}$	*Odchylenie przeciętne w szeregu rozdzielczym punktowym*
$d(x)=\frac{\sum_{i=1}^{k}\lvert \dot{x}_i-\bar{x}\rvert\cdot n_i}{N}$	*Odchylenie przeciętne w szeregu rozdzielczym z przedziałami klasowymi*
$W_z(x)=\frac{s(x)}{\bar{x}}$	*Klasyczny współczynnik zmienności*
$W_z(x)=\frac{O_c(x)}{Me(x)};\qquad W_z(x)=\frac{Q_3(x)-Q_1(x)}{Q_3(x)+Q_1(x)}$	*Kwartylowy współczynnik zmienności*
$M_s(x)=\bar{x}-D(x);\quad M_s(x)=3[\bar{x}-Me(x)]$	*Absolutna miara skośności*
$W_{s1}(x)=\frac{\bar{x}-D(x)}{s(x)};\quad W_{s1}(x)=\frac{3[\bar{x}-Me(x)]}{s(x)}$	*Współczynnik skośności*
$W_{s2}(x)=\frac{[Q_3(x)-Me(x)]-[Me(x)-Q_1(x)]}{[Q_3(x)-Me(x)]+[Me(x)-Q_1(x)]}=\frac{[Q_3(x)-Me(x)]-[Me(x)-Q_1(x)]}{2O_c(x)}$	*Kwartylowy współczynnik skośności*
$m_3^{*}(t)=\frac{m_3(x)}{\lvert m_3(x)\rvert+s^3(x)}$	*Trzeci moment centralny standaryzowany*

$k = \frac{a}{5000}$	*Współczynnik koncentracji*
$m_4(t) = \frac{m_4(x)}{s^4(x)} = \frac{\frac{\sum_{i=1}^{k}(x_i - \bar{x})^4 \cdot n_i}{N}}{s^4(x)}$	*Miara kurtozy*
$e(t) = m_4(t) - 3$	*Miara ekscecsu*
$d^c = \sqrt{\frac{\sum_{i,j}\frac{(f_{ij} - f_i \cdot f_j)^2}{f_i \cdot f_j}}{\min(r,s)-1}}$	*Współczynnik Czuprowa*
$d_G{}^H = \sqrt{\frac{\sum_{i,j\in G} f_{ij} - \sum_{i,j\in G} f_i \cdot f_j}{1 - \frac{1}{\min(r,s)}}}$; $d_M{}^H = \sqrt{\frac{\sum_{i,j\in M} f_i \cdot f_j - \sum_{i,j\in M} f_{ij}}{1 - \frac{1}{\min(r,s)}}}$	*Współczynnik Hellwiga*
$r^k = \frac{s(\bar{y}_{x_j})}{s(y)} = \frac{\sqrt{\frac{\sum_j (\bar{y}_{x_j} - \bar{y})^2 \cdot n_j}{N}}}{\sqrt{\frac{\sum_i (y_i - \bar{y})^2 \cdot n_i}{N}}}$; $r^k = \frac{s(\bar{y}_{x_i})}{s(y)} = \frac{\sqrt{\frac{\sum_i (\bar{y}_{x_i} - \bar{y})^2 \cdot n_i}{N}}}{\sqrt{\frac{\sum_j (\bar{y}_{x_i} - \bar{y})^2 \cdot n_j}{N}}}$	*Stosunek korelacyjny*
$r^P = \frac{c(x,y)}{s(x)\cdot s(y)} = \frac{\frac{\sum_i (x_i - \bar{x})\cdot(y_i - \bar{y})}{N}}{\sqrt{\frac{\sum_i (x_i - \bar{x})^2}{N}} \cdot \sqrt{\frac{\sum_i (y_i - \bar{y})^2}{N}}}$	*Współczynnik korelacji liniowej Pearsona dla szeregów szczegółowych*
$r^P = \frac{c(x,y)}{s(x)\cdot s(y)} = \frac{\frac{\sum_{i,j} (x_j - \bar{x})\cdot(y_i - \bar{y})\cdot n_{ij}}{N}}{\sqrt{\frac{\sum_j (x_j - \bar{x})^2 \cdot n_j}{N}} \cdot \sqrt{\frac{\sum_i (y_i - \bar{y})^2 \cdot n_i}{N}}}$	*Współczynnik korelacji liniowej Pearsona dla tablicy korelacyjnej*
$r^{Sp} = 1 - \frac{6\sum_i (d_{x_i} - d_{y_i})^2}{N^3 - N}$	*Współczynnik korelacji rang Spearmana*
$\hat{y} = a_y x + b$	*Równanie regresji Y względem X*

$a_y = \frac{C(x,y)}{s^2(x)} = \frac{\frac{\sum_i (x_i - \bar{x}) \cdot (y_i - \bar{y})}{N}}{\frac{\sum_i (x_i - \bar{x})^2}{N}}$	*Parametr równania regresji (współczynnik regresji)*
$b = \bar{y} - a_y \cdot \bar{x}$	*Parametr równania regresji*
$\varphi^2{}_y = \frac{\sum_i (y_i - \hat{y}_i)^2}{\sum_i (y_i - \bar{y})^2}$	*Współczynnik zbieżności*
$R^2 = 1 - \varphi^2$	*Współczynnik determinacji*
$R^a{}_{r/s} = y_r - y_s$	*Różnica absolutna jednopodstawowa*
$R^a{}_{r/r-1} = y_r - y_{r-1}$	*Różnica absolutna łańcuchowa*
$R^w{}_{r/s} = \frac{R^a{}_{r/s}}{y_s} \cdot 100 = \frac{y_r - y_s}{y_s} \cdot 100$	*Różnica względna jednopodstawowa*
$R^w{}_{r/r-1} = \frac{R^a{}_{r/r-1}}{y_{r-1}} \cdot 100 = \frac{y_r - y_{r-1}}{y_{r-1}} \cdot 100$	*Różnica względna łańcuchowa*
$w^i{}_{r/s} = \frac{y_r}{y_s} \cdot 100$	*Indeks indywidualny jednopodstawowy*
$w^i{}_{r/r-1} = \frac{y_r}{y_{r-1}} \cdot 100$	*Indeks indywidualny łańcuchowy*
$\bar{y}_2^{(3)} = \frac{y_1 + y_2 + y_3}{3}; \bar{y}_3^{(3)} = \frac{y_2 + y_3 + y_4}{3}; \ldots\ldots,$ $\bar{y}_{n-1}^{(3)} = \frac{y_{n-2} + y_{n-1} + y_n}{3}$	*Średnie ruchome trzyokresowe*
$\bar{y}_{3..}{}^{(5)} = \frac{y_1 + y_2 + y_3 + y_4 + y_5}{5} \ldots; \ldots \bar{y}_4{}^{(5)} = \frac{y_2 + y_3 + y_4 + y_5 + y_6}{5}$ $\bar{y}_{n-2}{}^{(5)} = \frac{y_{n-4} + y_{n-3} + y_{n-2} + y_{n-1} + y_n}{5}$	*Średnie ruchome pięciookresowe*
$\hat{y}_t = a \cdot t + b$	*Równanie trendu*

$a = \frac{\sum_{i=1}^{n}(t_i - \bar{t}) \cdot y_{t_i}}{\sum_{i=1}^{n}(t_i - \bar{t})^2}$	*Parametr równania trendu*
$b = \bar{y} - a \cdot \bar{t}$	*Parametr równania trendu*
$\varphi^2 = \frac{\sum_{i=1}^{n}(y_{t_i} - \hat{y}_{t_i})^2}{\sum_{i=1}^{n}(y_{t_i} - \bar{y})^2}$	*Współczynnik zbieżności dla równania trendu*
$\overline{w}^i_{r/r-1} = \sqrt[n-1]{w^j_{2/1} \cdot w^j_{3/2} \cdot w^j_{4/3} \cdot \ldots \cdot w^j_{n/n-1}} = \sqrt[n-1]{w^j_{n/1}}$	*Średnie tempo zmian*
$S_j = \frac{\bar{y}_j}{\bar{y}} \cdot 100$	*Wskaźnik sezonowości oparty na średnich okresów jednoimiennych*
$S_j = \frac{\sum_j y_{t_i}^{(j)}}{\sum_j \bar{y}_{c_i}^{(j)}} \cdot 100$	*Wskaźnik sezonowości oparty na średnich ruchomych centrowanych*
$S_j = \frac{\sum_j y_{t_i}^{(j)}}{\sum_j \hat{y}_{t_i}^{(j)}} \cdot 100$	*Wskaźnik sezonowości oparty na równaniu trendu*
$k = \frac{\sum_j S_j}{r \cdot 100}$	*Współczynnik korygujący surowe wskaźniki sezonowości*
$\bar{x} - t_t \cdot \frac{s(x)}{\sqrt{n-1}} \langle E(x) \langle \bar{x} + t_t \cdot \frac{s(x)}{\sqrt{n-1}}$	*Przedział ufności dla średniej (mała próba)*
$\bar{x} - t_t \cdot \frac{s(x)}{\sqrt{n}} \langle E(x) \langle \bar{x} + t_t \cdot \frac{s(x)}{\sqrt{n}}$	*Przedział ufności dla średniej (duża próba)*
$\hat{s}^2(x) = \frac{\sum_i (x_i - \bar{x})^2}{n-1}$	*Wariancja dla małej próby*

$n_{\min} = \frac{t_t^{\,2} \cdot \hat{s}^2(x)}{d^2}$	*Minimalna liczebność próby dla szacowania średniej*
$\frac{m}{n} - t_t \cdot \sqrt{\frac{\frac{m}{n}\left(1-\frac{m}{n}\right)}{n}} < p < \frac{m}{n} + t_t \cdot \sqrt{\frac{\frac{m}{n}\left(1-\frac{m}{n}\right)}{n}}$	*Przedział ufności dla wskaźnika struktury*
$\frac{n \cdot s_2^{\,2}(x)}{t_{t_1}} \langle \sigma^2(x) \langle \frac{n \cdot s_2^{\,2}(x)}{t_{t_2}}$	*Przedział ufności dla wariancji/ odchylenia standardowego (mała próba)*
$\frac{s(x)}{1+\frac{t_t}{\sqrt{2n}}} \langle \sigma(x) \langle \frac{s(x)}{1-\frac{t_t}{\sqrt{2n}}}$	*Przedział ufności dla odchylenia standardowego / wariancji (duża próba)*
$r^P - t_t \cdot \frac{1-(r^P)^2}{\sqrt{n}} \langle \rho \langle r^P + t_t \cdot \frac{1-\left(r^P\right)^2}{\sqrt{n}}$	*Przedział ufności dla współczynnika korelacji*
$t_{emp} = \frac{\bar{x} - E_0(x)}{s(x)} \cdot \sqrt{n-1}$	*Statystyka empiryczna przy weryfikacji hipotezy dla średniej (mała próba)*
$t_{emp} = \frac{\bar{x} - E_0(x)}{s(x)} \cdot \sqrt{n}$	*Statystyka empiryczna przy weryfikacji hipotezy dla średniej (duża próba)*
$t_{emp} = \frac{\bar{x}_1 - \bar{x}_2}{\sqrt{\frac{n_1 s_1^{\,2}(x) + n_2 s_2^{\,2}(x)}{n_1 + n_2 - 2}\left(\frac{1}{n_1} + \frac{1}{n_2}\right)}}$	*Statystyka empiryczna przy weryfikacji hipotezy dla dwóch średnich (mała próba)*
$t_{emp} = \frac{\bar{x}_1 - \bar{x}_2}{\sqrt{\frac{s_1^{\,2}(x)}{n_1} + \frac{s_2^{\,2}(x)}{n_2}}}$	*Statystyka empiryczna przy weryfikacji hipotezy dla dwóch średnich (duża próba)*

$t_{emp} = \dfrac{\dfrac{m}{n} - p_0}{\sqrt{\dfrac{p_0 \cdot q_0}{n}}}$	*Statystyka empiryczna przy weryfikacji hipotezy dla wskaźnika struktury*
$t_{emp} = \dfrac{r^P}{\sqrt{1-\left(r^P\right)^2}} \cdot \sqrt{n-2}$	*Statystyka empiryczna przy weryfikacji hipotezy dla współczynnika korelacji*
$t_{emp} = \sum_{ij} \dfrac{\left(n_{ij} - N \cdot f_i \cdot f_j\right)^2}{N \cdot f_i \cdot f_j}$	*Statystyka empiryczna dla testu niezależności chi-kwadrat*
$t_{emp} = \sum_{i} \dfrac{\left(n_i - N \cdot p_i\right)^2}{N \cdot p_i}$	*Statystyka empiryczna dla testu zgodności chi-kwadrat*
$t_{emp} = D \cdot \sqrt{n}\ ;\ D = \max\left\|F_i(x) - F_{0i}(x)\right\|$	*Statystyka empiryczna dla testu zgodności Kołmogorowa*

ANEKS

Tablica 1. Dystrybuanta rozkładu normalnego (dla $0{,}00 \le t < 3{,}10$)[1]

t_t	*0,00*	*0,01*	*0,02*	*0,03*	*0,04*	*0,05*	*0,06*	*0,07*	*0,08*	*0,09*
0,0	0,00000	0,00399	0,00798	0,01197	0,01595	0,01994	0,02392	0,02790	0,03188	0,03586
0,1	0,03983	0,04380	0,04776	0,05172	0,05567	0,05962	0,06356	0,06749	0,07142	0,07535
0,2	0,07926	0,08317	0,08706	0,09095	0,09483	0,09871	0,10257	0,10642	0,11026	0,11409
0,3	0,11791	0,12172	0,12552	0,12930	0,13307	0,13683	0,14058	0,14431	0,14803	0,15173
0,4	0,15542	0,15910	0,16276	0,16640	0,17003	0,17364	0,17724	0,18082	0,18439	0,18793
0,5	0,19146	0,19497	0,19847	0,20194	0,20540	0,20884	0,21226	0,21566	0,21904	0,22240
0,6	0,22575	0,22907	0,23237	0,23565	0,23891	0,24215	0,24537	0,24857	0,25175	0,25490
0,7	0,25804	0,26115	0,26424	0,26730	0,27035	0,27337	0,27637	0,27935	0,28230	0,28524
0,8	0,28814	0,29103	0,29389	0,29673	0,29955	0,30234	0,30511	0,30785	0,31057	0,31327
0,9	0,31594	0,31859	0,32121	0,32381	0,32639	0,32894	0,33147	0,33398	0,33646	0,33891
1,0	0,34134	0,34375	0,34614	0,34849	0,35083	0,35314	0,35543	0,35769	0,35993	0,36214
1,1	0,36433	0,36650	0,36864	0,37076	0,37286	0,37493	0,37698	0,37900	0,38100	0,38298
1,2	0,38493	0,38686	0,38877	0,39065	0,39251	0,39435	0,39617	0,39796	0,39973	0,40147
1,3	0,40320	0,40490	0,40658	0,40824	0,40988	0,41149	0,41308	0,41466	0,41621	0,41774
1,4	0,41924	0,42073	0,42220	0,42364	0,42507	0,42647	0,42785	0,42922	0,43056	0,43189
1,5	0,43319	0,43448	0,43574	0,43699	0,43822	0,43943	0,44062	0,44179	0,44295	0,44408
1,6	0,44520	0,44630	0,44738	0,44845	0,44950	0,45053	0,45154	0,45254	0,45352	0,45449
1,7	0,45543	0,45637	0,45728	0,45818	0,45907	0,45994	0,46080	0,46164	0,46246	0,46327
1,8	0,46407	0,46485	0,46562	0,46638	0,46712	0,46784	0,46856	0,46926	0,46995	0,47062
1,9	0,47128	0,47193	0,47257	0,47320	0,47381	0,47441	0,47500	0,47558	0,47615	0,47670
2,0	0,47725	0,47778	0,47831	0,47882	0,47932	0,47982	0,48030	0,48077	0,48124	0,48169
2,1	0,48214	0,48257	0,48300	0,48341	0,48382	0,48422	0,48461	0,48500	0,48537	0,48574
2,2	0,48610	0,48645	0,48679	0,48713	0,48745	0,48778	0,48809	0,48840	0,48870	0,48899
2,3	0,48928	0,48956	0,48983	0,49010	0,49036	0,49061	0,49086	0,49111	0,49134	0,49158
2,4	0,49180	0,49202	0,49224	0,49245	0,49266	0,49286	0,49305	0,49324	0,49343	0,49361
2,5	0,49379	0,49396	0,49413	0,49430	0,49446	0,49461	0,49477	0,49492	0,49506	0,49520
2,6	0,49534	0,49547	0,49560	0,49573	0,49585	0,49598	0,49609	0,49621	0,49632	0,49643
2,7	0,49653	0,49664	0,49674	0,49683	0,49693	0,49702	0,49711	0,49720	0,49728	0,49736
2,8	0,49744	0,49752	0,49760	0,49767	0,49774	0,49781	0,49788	0,49795	0,49801	0,49807
2,9	0,49813	0,49819	0,49825	0,49831	0,49836	0,49841	0,49846	0,49851	0,49856	0,49861
3,0	0,49865	0,49873	0,049878	0,49882	0,49886	0,49889	0,49893	0,49896	0,49898	0,4990

[1] – *dla ujemnych wartości t należy wykorzystywać własność symetrii rozkładu normalnego*

Tablica 2. Rozkład t Studenta

k	$1-\alpha$								
	0,8	***0,6***	***0,4***	***0,2***	***0,1***	***0,05***	***0,02***	***0,01***	***0,001***
1	0,325	0,727	1,376	3,078	6,314	12,706	31,821	63,656	636,578
2	0,289	0,617	1,061	1,886	2,920	4,303	6,965	9,925	31,600
3	0,277	0,584	0,978	1,638	2,353	3,182	4,541	5,841	12,924
4	0,271	0,569	0,941	1,533	2,132	2,776	3,747	4,604	8,610
5	0,267	0,559	0,920	1,476	2,015	2,571	3,365	4,032	6,869
6	0,265	0,553	0,906	1,440	1,943	2,447	3,143	3,707	5,959
7	0,263	0,549	0,896	1,415	1,895	2,365	2,998	3,499	5,408
8	0,262	0,546	0,889	1,397	1,860	2,306	2,896	3,355	5,041
9	0,261	0,543	0,883	1,383	1,833	2,262	2,821	3,250	4,781
10	0,260	0,542	0,879	1,372	1,812	2,228	2,764	3,169	4,587
11	0,260	0,540	0,876	1,363	1,796	2,201	2,718	3,106	4,437
12	0,259	0,539	0,873	1,356	1,782	2,179	2,681	3,055	4,318
13	0,259	0,538	0,870	1,350	1,771	2,160	2,650	3,012	4,221
14	0,258	0,537	0,868	1,345	1,761	2,145	2,624	2,977	4,140
15	0,258	0,536	0,866	1,341	1,753	2,131	2,602	2,947	4,073
16	0,258	0,535	0,865	1,337	1,746	2,120	2,583	2,921	4,015
17	0,257	0,534	0,863	1,333	1,740	2,110	2,567	2,898	3,965
18	0,257	0,534	0,862	1,330	1,734	2,101	2,552	2,878	3,922
19	0,257	0,533	0,861	1,328	1,729	2,093	2,539	2,861	3,883
20	0,257	0,533	0,860	1,325	1,725	2,086	2,528	2,845	3,850
21	0,257	0,532	0,859	1,323	1,721	2,080	2,518	2,831	3,819
22	0,256	0,532	0,858	1,321	1,717	2,074	2,508	2,819	3,792
23	0,256	0,532	0,858	1,319	1,714	2,069	2,500	2,807	3,768
24	0,256	0,531	0,857	1,318	1,711	2,064	2,492	2,797	3,745
25	0,256	0,531	0,856	1,316	1,708	2,060	2,485	2,787	3,725
26	0,256	0,531	0,856	1,315	1,706	2,056	2,479	2,779	3,707
27	0,256	0,531	0,855	1,314	1,703	2,052	2,473	2,771	3,689
28	0,256	0,530	0,855	1,313	1,701	2,048	2,467	2,763	3,674
29	0,256	0,530	0,854	1,311	1,699	2,045	2,462	2,756	3,660
30	0,256	0,530	0,854	1,310	1,697	2,042	2,457	2,750	3,646
31	0,256	0,530	0,853	1,309	1,696	2,040	2,453	2,744	3,633
32	0,255	0,530	0,853	1,309	1,694	2,037	2,449	2,738	3,622
33	0,255	0,530	0,853	1,308	1,692	2,035	2,445	2,733	3,611
34	0,255	0,529	0,852	1,307	1,691	2,032	2,441	2,728	3,601
35	0,255	0,529	0,852	1,306	1,690	2,030	2,438	2,724	3,591
36	0,255	0,529	0,852	1,306	1,688	2,028	2,434	2,719	3,582
37	0,255	0,529	0,851	1,305	1,687	2,026	2,431	2,715	3,574
38	0,255	0,529	0,851	1,304	1,686	2,024	2,429	2,712	3,566
39	0,255	0,529	0,851	1,304	1,685	2,023	2,426	2,708	3,558
40	0,255	0,529	0,851	1,303	1,684	2,021	2,423	2,704	3,551

Tablica 3. Rozkład χ^2 (chi–kwadrat)

k	*0,99*	*0,98*	*0,95*	*0,90*	*0,80*	*0,70*	*0,50*	*0,30*	*0,20*	*0,10*	*0,05*	*0,02*	*0,01*	*0,001*
1	0,0002	0,0006	0,004	0,016	0,064	0,148	0,455	1,074	1,642	2,706	3,841	5,412	6,635	10,827
2	0,020	0,040	0,103	0,211	0,446	0,713	1,386	2,408	3,219	4,605	5,991	7,824	9,210	13,815
3	0,115	0,185	0,352	0,584	1,005	1,424	2,366	3,665	4,642	6,251	7,815	9,837	11,345	16,266
4	0,297	0,429	0,711	1,064	1,649	2,195	3,357	4,878	5,989	7,779	9,488	11,668	13,277	18,466
5	0,554	0,752	1,145	1,610	2,343	3,000	4,351	6,064	7,289	9,236	11,070	13,388	15,086	20,515
6	0,872	1,134	1,635	2,204	3,070	3,828	5,348	7,231	8,558	10,645	12,592	15,033	16,812	22,457
7	1,239	1,564	2,167	2,833	3,822	4,671	6,346	8,383	9,803	12,017	14,067	16,622	18,475	24,321
8	1,647	2,032	2,733	3,490	4,594	5,527	7,344	9,524	11,030	13,362	15,507	18,168	20,090	26,124
9	2,088	2,532	3,325	4,168	5,380	6,393	8,343	10,656	12,242	14,684	16,919	19,679	21,666	27,877
10	2,558	3,059	3,940	4,865	6,179	7,267	9,342	11,781	13,442	15,987	18,307	21,161	23,209	29,588
11	3,053	3,609	4,575	5,578	6,989	8,148	10,341	12,899	14,631	17,275	19,675	22,618	24,725	31,264
12	3,571	4,178	5,226	6,304	7,807	9,034	11,340	14,011	15,812	18,549	21,026	24,054	26,217	32,909
13	4,107	4,765	5,892	7,041	8,634	9,926	12,340	15,119	16,985	19,812	22,362	25,471	27,688	34,527
14	4,660	5,368	6,571	7,790	9,467	10,821	13,339	16,222	18,151	21,064	23,685	26,873	29,141	36,124
15	5,229	5,985	7,261	8,547	10,307	11,721	14,339	17,322	19,311	22,307	24,996	28,259	30,578	37,698
16	5,812	6,614	7,962	9,312	11,152	12,624	15,338	18,418	20,465	23,542	26,296	29,633	32,000	39,252
17	6,408	7,255	8,672	10,085	12,002	13,531	16,338	19,511	21,615	24,769	27,587	30,995	33,409	40,791
18	7,015	7,906	9,390	10,865	12,857	14,440	17,338	20,601	22,760	25,989	28,869	32,346	34,805	42,312
19	7,633	8,567	10,117	11,651	13,716	15,352	18,338	21,689	23,900	27,204	30,144	33,687	36,191	43,819
20	8,260	9,237	10,851	12,443	14,578	16,266	19,337	22,775	25,038	28,412	31,410	35,020	37,566	45,314
21	8,897	9,915	11,591	13,240	15,445	17,182	20,337	23,858	26,171	29,615	32,671	36,343	38,932	46,796
22	9,542	10,600	12,338	14,041	16,314	18,101	21,337	24,939	27,301	30,813	33,924	37,659	40,289	48,268
23	10,196	11,293	13,091	14,848	17,187	19,021	22,337	26,018	28,429	32,007	35,172	38,968	41,638	49,728
24	10,856	11,992	13,848	15,659	18,062	19,943	23,337	27,096	29,553	33,196	36,415	40,270	42,980	51,179
25	11,524	12,697	14,611	16,473	18,940	20,867	24,337	28,172	30,675	34,382	37,652	41,566	44,314	52,619
26	12,198	13,409	15,379	17,292	19,820	21,792	25,336	29,246	31,795	35,563	38,885	42,856	45,642	54,051
27	12,878	14,125	16,151	18,114	20,703	22,719	26,336	30,319	32,912	36,741	40,113	44,140	46,963	55,475
28	13,565	14,847	16,928	18,939	21,588	23,647	27,336	31,391	34,027	37,916	41,337	45,419	48,278	56,892
29	14,256	15,574	17,708	19,768	22,475	24,577	28,336	32,461	35,139	39,087	42,557	46,693	49,588	58,301
30	14,953	16,306	18,493	20,599	23,364	25,508	29,336	33,530	36,250	40,256	43,773	47,962	50,892	59,702

Tablica 4. Rozkład graniczny Kołmogorowa

t	α	t	α	t	α	t	α
1,01	0,740566	***1,41***	0,962486	***1,81***	0,997146	***2.21***	0,999886
1,02	0,750826	***1,42***	0,964552	***1,82***	0,997346	***2,22***	0,999896
1,03	0,760780	***1,43***	0,966516	***1,83***	0,997533	***2,23***	0,999904
1,04	0,770434	***1,44***	0,968382	***1,84***	0,997707	***2,24***	0,999912
1,05	0,779794	***1,45***	0,970158	***1,85***	0,997870	***2,25***	0,999920
1,06	0,788860	***1,46***	0,971846	***1,86***	0,998023	***2,26***	0,999926
1,07	0,797636	***1,47***	0,973448	***1,87***	0,998145	***2,27***	0,999934
1,08	0,806128	***1,48***	0,974970	***1,88***	0,998297	***2,28***	0,999940
1,09	0,814342	***1,49***	0,976412	***1,89***	0,998421	***2,29***	0,999944
1,10	0,822282	***1,50***	0,977782	***1,90***	0,998536	***2,30***	0,999949
1,11	0,829950	***1,51***	0,979080	***1,91***	0,998644	***2,31***	0,999954
1,12	0,837356	***1,52***	0,980310	***1,92***	0,998744	***2,32***	0,999958
1,13	0,844502	***1,53***	0,981476	***1,93***	0,998837	***2,33***	0,999962
1,14	0,851394	***1,54***	0,982578	***1,94***	0,998924	***2,34***	0,999965
1,15	0,858038	***1,55***	0,983622	***1,95***	0,999004	***2,35***	0,999968
1,16	0,864442	***1,56***	0,984610	***1,96***	0,999079	***2,36***	0,999970
1,17	0,870612	***1,57***	0,985544	***1,97***	0,999149	***2,37***	0,999973
1,18	0,876548	***1,58***	0,986426	***1,98***	0,999213	***2,38***	0,999976
1,19	0,882258	***1,59***	0,987260	***1,99***	0,999273	***2,39***	0,999978
1,20	0,887750	***1,60***	0,988048	***2,00***	0,999329	***2,40***	0,999980
1,21	0,893030	***1,61***	0,988791	***2,01***	0,999380	***2,41***	0,999982
1,22	0,898104	***1,62***	0,989492	***2,02***	0,999428	***2,42***	0,999984
1,23	0,902972	***1,63***	0,990154	***2,03***	0,999474	***2,43***	0,999986
1,24	0,907648	***1,64***	0,990777	***2,04***	0,999516	***2,44***	0,999987
1,25	0,912132	***1,65***	0,991364	***2,05***	0,999552	***2,45***	0,999988
1,26	0,916432	***1,66***	0,991917	***2,06***	0,999588	***2,46***	0,999989
1,27	0,920556	***1,67***	0,992928	***2,07***	0,999620	***2,47***	0,999990
1,28	0,924505	***1,68***	0,992928	***2,08***	0,999650	***2,48***	0,999991
1,29	0,928288	***1,69***	0,993389	***2,09***	0,999680	***2,49***	0,999992
1,30	0,931908	***1,70***	0,993828	***2,10***	0,999705	***2,50***	0,99993
1,13	0,935370	***1,71***	0,994230	***2,11***	0,999723	***2.55***	0,999995
1,32	0,938682	***1,72***	0,994612	***2,12***	0,999750	***2,60***	0,999974
1,33	0,941848	***1,73***	0,994972	***2,13***	0,999770	***2,65***	0,999998
1,34	0,944872	***1,74***	0,995309	***2,14***	0,999790	***2,70***	0,999999
1,35	0,947756	***1,75***	0,995625	***2,15***	0,999806	***2,75***	0,9999994
1,36	0,950512	***1,76***	0,995922	***2,16***	0,999822	***2,80***	0,9999997
1,37	0,953142	***1,77***	0,996200	***2,17***	0,999838	***2,85***	0,9999998
1,38	0,955650	***1,78***	0,996460	***2,18***	0,999852	***2,90***	0,9999999
1,39	0,958040	***1,79***	0,996704	***2,19***	0,999864	***2,95***	0,99999994
1,40	0,960318	***1,80***	0,996932	***2,20***	0,999874	***3,00***	0,99999997

Tablica 5. Rozkład serii

dla $\alpha = 0{,}05$

n_2	n_1																		
	2	***3***	***4***	***5***	***6***	***7***	***8***	***9***	***10***	***11***	***12***	***13***	***14***	***15***	***16***	***17***	***18***	***19***	***20***
2	2																		
3	2	2																	
4	2	2	2																
5	2	2	2	3															
6	2	2	3	3	3														
7	2	2	3	3	4	4													
8	2	2	3	3	4	4	5												
9	2	2	3	4	4	5	5	6											
10	2	3	3	4	5	5	6	6	6										
11	2	3	3	4	5	5	6	6	7	7									
12	2	3	4	4	5	6	6	7	7	8	8								
13	2	3	4	4	5	6	6	7	8	8	9	9							
14	2	3	4	5	5	6	7	7	8	8	9	9	10						
15	2	3	4	5	6	6	7	8	8	9	9	10	10	11					
16	2	3	4	5	6	6	7	8	8	9	10	10	11	11	11				
17	2	3	4	5	6	7	7	8	9	9	10	10	11	11	12	12			
18	2	3	4	5	6	7	8	8	9	10	10	11	11	12	12	13	13		
19	2	3	4	5	6	7	8	8	9	10	10	11	12	12	13	13	14	14	
20	2	3	4	5	6	7	8	8	9	10	11	11	12	12	13	13	14	14	15

dla $\alpha = 0{,}95$

n_2	n_1																		
	2	***3***	***4***	***5***	***6***	***7***	***8***	***9***	***10***	***11***	***12***	***13***	***14***	***15***	***16***	***17***	***18***	***19***	***20***
2	4																		
3	5	6																	
4	5	6	7																
5	5	7	8	8															
6	5	7	8	9	10														
7	5	7	8	9	10	11													
8	5	7	9	10	11	12	12												
9	5	7	9	10	11	12	13	13											
10	5	7	9	10	11	12	13	14	15										
11	5	7	9	11	12	13	14	14	15	16									
12	5	7	9	11	12	13	14	15	16	16	17								
13	5	7	9	11	12	13	14	15	16	17	17	18							
14	5	7	9	11	12	13	15	16	16	17	18	19	19						
15	5	7	9	11	13	14	15	16	17	18	18	19	20	20					
16	5	7	9	11	13	14	15	16	17	18	19	20	20	21	22				
17	5	7	9	11	13	14	15	16	17	18	19	20	21	21	22	23			
18	5	7	9	11	13	14	15	17	18	19	20	20	21	22	23	23	24		
19	5	7	9	11	13	14	15	17	18	19	20	21	22	22	23	24	24	25	
20	5	7	9	11	13	14	16	17	18	19	20	21	22	23	24	24	25	26	26