KRAKOWSKA SZKOŁA WYŻSZA
IM. ANDRZEJA FRYCZA MODRZEWSKIEGO

Michał Major, Janusz Niezgoda

Elementy Statystyki

Część I. Statystyka opisowa

Kraków 2003

ISBN 83-918302-5-X

Na zlecenie: Krakowskiej Szkoły Wyższej im. Andrzeja Frycza Modrzewskiego

Wydawca: Krakowskie Towarzystwo Edukacyjne sp. z o.o., Kraków 2003

Skład i łamanie: Maciej Major

Druk i oprawa: Multipol II

Spis treści

Od autorów

Drodzy Czytelnicy!

Przekazujemy na Wasze ręce podręcznik z nadzieją, że ułatwi on proces studiowania przedmiotu statystyka oraz dyscyplin pokrewnych, takich jak ekonometria, marketing, finanse czy rachunkowość.

Skrypt zawiera podstawowe zagadnienia związane z teorią i zastosowaniem statystyki opisowej, która – obok wnioskowania statystycznego – stanowi jeden z dwóch filarów szeroko rozumianej statystyki.

Zagadnienia omawiane w tej pracy zostały podzielone na pięć rozdziałów. Rozdział pierwszy przeznaczono na opis podstawowych pojęć statystycznych, takich jak – między innymi – zbiorowość i jednostka statystyczna, cecha i zmienna statystyczna. Opisano tutaj także metodologię badań statystycznych. W rozdziale drugim przedstawiono etapy badań statystycznych, ze szczególnym zwróceniem uwagi na sposoby opracowywania i prezentacji materiału statystycznego. W kolejnym rozdziale – trzecim – dokonano przeglądu metod opisu struktury zbiorowości statystycznej, przy wykorzystaniu tzw. miar położenia, zmienności, asymetrii i koncentracji. Rozdział czwarty jest przeglądem metod badania współzależności dwóch i więcej zmiennych statystycznych. Opisano tutaj różnego rodzaju współczynniki korelacji, oraz sposób szacowania parametrów tzw. funkcji regresji przy zastosowaniu metody najmniejszych kwadratów. Ostatni z rozdziałów – piąty – poświęcono zagadnieniom związanym z analizą szeregów czasowych. Opisane w nim metody pozwalają na badanie dynamiki zjawisk lub na określanie tendencji rozwojowych.

Podczas pisania podręcznika staraliśmy się zamieścić jak najwięcej przykładów liczbowych oraz zadań zalecanych do samodzielnego rozwiązania. Dla lepszej czytelności miejsca w tekście, gdzie kończą się przykłady, a zaczyna się tekst o charakterze ogólnym, zaznaczyliśmy znakiem #. Znaku tego nie stosowaliśmy natomiast w sytuacji, gdy bezpośrednio po przykładzie zaczyna się kolejny rozdział, czy też podrozdział.

Podręcznik zawiera także liczne przypisy i odwołania do literatury, której studiowanie zalecamy osobom pragnącym lepiej poznać zagadnienia statystyki i dyscyplin z nią powiązanych.

Wiadomości wstępne

Rozdział 1

1.1. Geneza i istota przedmiotu statystyka

Słowo statystyka wywodzi się od łacińskiego słowa *status*, co oznacza stan rzeczy lub państwo[*]. W piśmiennictwie, po raz pierwszy, słowo to zostało użyte przez G. Achenwalda dla oznaczenia zbioru informacji o państwie[**]. Z czasem obok informacji opisowych dotyczących państwa zaczęły pojawiać się dane liczbowe ujmowane tabelarycznie. Proces gromadzenia i prezentacji tabelarycznej zaczęto nazywać statystyką, a ich autorów tabelarystami.

Do ukształtowania zakresu przedmiotu statystyki przyczynili się również J. Graunt i W. Petty, przedstawiciele tzw. arytmetyków politycznych, którzy dostrzegali w statystyce metodę umożliwiającą wyodrębnienie spośród pozornie chaotycznych zjawisk masowych pewnych prawidłowości.

Do dalszego rozwoju statystyki przyczynili się również B. Pascal i P. Fermat, żyjący w XVII wieku, których uważa się za prekursorów teorii rachunku prawdopodobieństwa[***]. Dzięki rachunkowi prawdopodobieństwa rozwinęła się **statystyka matematyczna**, której głównym celem jest wyodrębnianie i uogólnianie wyników otrzymanych z próby losowej na całą populację, z której ta próba pochodzi. Proces taki nosi nazwę **wnioskowania statystycznego**. Każde wnioskowanie musi być jednak poprzedzone wnikliwym i rzetelnym opisem losowych prób i cech statystycznych. Służyć temu mają metody opisowe określane mianem **statystyki opisowej**.

Obecnie pod pojęciem **statystyki** rozumie się **naukę traktującą o ilościowych metodach badania zjawisk (procesów) masowych**. Zaliczyć można do niej wspomnianą wcześniej statystykę opisową i statystykę matematyczną. Zaznaczyć jednak należy, że nie jest to jedyny sposób interpretacji słowa statystyka. W potocznym rozumieniu słowa statystyka używa się często

[*] Zob. np. *Słownik wyrazów obcych* pod redakcją Jana Tokarskiego, PWN 1980.

[**] M. Sobczyk, *Statystyka*, PWN 1998, s. 5.

[***] Ibidem.

do oznaczenia czynności polegających na prostym zbieraniu a następnie opracowywaniu danych liczbowych, lub też określa ono zbiór informacji liczbowych (danych) dotyczących jakiegoś zjawiska.

Statystyka i jej metody znalazły szerokie zastosowanie w wielu dziedzinach wiedzy – między innymi – w naukach społecznych, w antropologii, biologii, medycynie oraz geografii.

1.2. Podstawowe pojęcia statystyczne

Jak już zostało zaznaczone powyżej przedmiotem statystyki nie są zjawiska jednostkowe lecz tzw. zjawiska masowe, czyli takie o których można powiedzieć, że często się powtarzają. Badając zjawiska masowe statystyka dąży do wykrywania i określania pewnych zachodzących w nich prawidłowości. Innymi słowy można powiedzieć, że statystyka nie zajmuje się pojedynczymi zdarzeniami ani pojedynczymi obiektami, lecz zbiorowością osób, rzeczy lub zjawisk. Zbiorowość taką określa się mianem **zbiorowości statystycznej** (lub **populacją**, **masą statystyczną** lub **zbiorowością generalną**) i definiuje jako **zbiór elementów (osób, przedmiotów, zdarzeń) podobnych, lecz nie identycznych pod względem określonej cechy, poddanych badaniom statystycznym**[*)].

Elementy wchodzące w skład zbiorowości statystycznej nazywane są **jednostkami statystycznymi** a ich liczba **liczebnością zbiorowości** lub **liczebnością całkowitą, (generalną)**. Jednostki statystyczne charakteryzują się pewnymi właściwościami określanymi mianem **cech statystycznych**. Cechy te mogą być kwalifikujące i badane. **Cechy kwalifikujące**[**)] pozwalają jednoznacznie określić jednostki statystyczne i zbiorowość statystyczną pod względem rzeczowym lub przedmiotowym (co?), terytorialnym (gdzie?) i czasowym (kiedy?). Cechy te nie podlegają badaniu, lecz pozwalają na przyporządkowanie jednostek do zbiorowości generalnej. **Cechy badane** natomiast, to te własności, ze względu na które różnią się jednostki statystyczne. W odróżnieniu od cech kwalifikujących podlegają one badaniu i decydują o zakresie prowadzonych badań. Jednostki statystyczne zaliczane do pewnej zbiorowości powinny posiadać przynajmniej jedną cechę wspólną (stałą) oraz cechy różniące je pomiędzy sobą.

Przykład 1.1. Dnia 31 grudnia 2001 roku przeprowadzono badanie stopy bezrobocia (wyrażonej w procentach) w powiatach województwa małopolskiego.

*)Zob. np. M. Woźniak, *Statystyka ogólna*, AE w Krakowie, Kraków 1997.

**)Zob. A. Iwasiewicz, Z. Paszek, *Statystyka z elementami metod sterowania jakością*, AE w Krakowie, Kraków 2000, s. 71.

Na podstawie powyższego sformułowania, zdefiniować zbiorowość statystyczną i jednostkę statystyczną. Określić liczebność zbiorowości oraz badaną cechę statystyczną.

Zbiorowość statystyczna, to zbiór powiatów województwa małopolskiego. Jednostką statystyczną jest tu powiat. Liczebność zbiorowości jest równa liczbie powiatów badanego województwa, czyli 19. Data 31. 12. 2001, stanowi określenie czasowe. Nazwa województwa umiejscawia badanie pod względem terytorialnym. Cechą statystyczną jest procentowa stopa bezrobocia. #

Cechy badane można podzielić na **cechy jakościowe** (niemierzalne) oraz **ilościowe** (mierzalne). Cechy jakościowe (np. płeć, kolor włosów, wykształcenie) to te cechy, ze względu na które, każdą jednostkę statystyczną można zakwalifikować do jednej z wyróżnionych kategorii, nie przypisując jej określonej miary. Natomiast cechy ilościowe (np. ciężar, wzrost, temperatura, liczba dzieci w rodzinie itp.) to te, w odniesieniu do których wyróżnia się zbiór rozróżnialnych stanów, oraz każdemu stanowi – na drodze pomiaru – przyporządkowuje się określoną liczbę.

Pomiar jest procesem empirycznym, który polega na przyporządkowaniu wartości liczbowych stanom obserwowanych cech, zrealizowanych w badanych obiektach lub w powtórzeniach zjawisk*). Dziedziną funkcji pomiaru jest zbiór obiektów (powtórzeń zjawiska) lub ich stanów, ze względu na badaną cechę, a przeciwdziedziną podzbiór zbioru liczb rzeczywistych. Podstawowym wymaganiem sformułowanym w stosunku do pomiaru jest warunek, aby relacje zachodzące między liczbami uzyskiwanymi w trakcie pomiaru były odbiciem relacji między badanymi obiektami lub powtórzeniami zjawiska.

Wyróżnić można następujące skale pomiarowe**):

— nominalną,
— porządkową,
— przedziałową oraz
— ilorazową.

Skale te tworzą układ hierarchiczny, od skal najsłabszych do najmocniejszych.

Skala nominalna (mianowa) jest najprostszą a zarazem najsłabszą spośród wszystkich skal pomiarowych. Stosowana jest wówczas, gdy stany badanej cechy rozróżniane przez metodę badawczą są rozłącznymi kategoriami

*)Ibidem, s. 85.

**)Podczas omówienia wymienionych skal pomiarowych zostały wykorzystane informacje zawarte w pracach: A. Iwaszkiewicz, Z. Paszek, *Statystyka z elementami statystycznych metod sterowania jakością*, op. cit., s. 85-88; B. Niemierko, *Testy osiągnięć szkolnych*, WSiP Warszawa, 1975, s. 110-114; M. Łobocki, *Metody badań pedagogicznych*, PWN Warszawa 1984, s. 38-43.

jakościowymi. Funkcja pomiarowa przyporządkowuje jednakowym obiektom lub powtórzeniom zjawiska jednakowe wartości liczbowe, a różnym obiektom (powtórzeniom zjawiska) przypisuje różne wartości liczbowe. Przyporządkowane liczby pełnią rolę przysłowiowych etykiet tożsamości (znaków rozpoznawczych), są ich oznaczeniami lub nazwami, pozwalającymi na ich jednoznaczną identyfikację i klasyfikację. Skali nominalnej można użyć numerując np. autobusy, tramwaje, telefony, a także studentów, których nazwiska znajdują się w protokołach ocen itp. Niewielka jest liczba operacji matematyczno – statystycznych, które można wykonać dla sklasyfikowanych w ten sposób obiektach lub powtórzeniach zjawisk. Należy tutaj wymienić: wyznaczanie liczebności, procentów i frakcji, modalnych i współczynnika skojarzenia Yule'a*).

Jeżeli stany badanych cech są uporządkowanymi rozłącznymi, a także uporządkowanymi malejąco lub rosnąco kategoriami jakościowymi, to wówczas stosuje się **skalę porządkową**. Określa ona pozycję, jaką zajmuje każdy z badanych przedmiotów lub każda z badanych osób, a także każde z badanych zjawisk w odpowiednio uporządkowanym i uszeregowanym zbiorze, zgodnie z przyjętymi kryteriami oceny. Wyznaczona w ten sposób pozycja jest względna i niedokładna. Wiemy bowiem, że jeden z badanych obiektów poprzedza lub następuje po innych, nie znamy jednak wielkości dzielącego ich dystansu. Skalami porządkowymi są np. skale stopni szkolnych, przy czym w polskim systemie szkolnictwa, bardziej preferowanemu stanowi przypisuje się większą liczbę. Na skali porządkowej oparta jest także większość wyników badań testowych dotyczących poziomu osiągnięć szkolnych, inteligencji, zdolności i osobowości czy badań marketingowych. Skala porządkowa, obok operacji statystycznych stosowanych w przypadku skali nominalnej, dopuszcza także takie operacje jak: ustalanie wartości środkowych (median), centyli i współczynników korelacji rangowej**).

Kolejna ze skal – **skala przedziałowa** (interwałowa) – zachowuje wszystkie możliwości pomiarowe skal nominalnej i porządkowej, dodając do nich możliwość pomiaru dystansu pomiędzy dwoma dowolnymi stanami badanej cechy. Określenie wspomnianego dystansu stało się możliwe dzięki temu, że operuje ona równymi jednostkami pomiaru (równymi interwałami) i tzw. umownym zerem. Takim umownym zerem może być np. narodzenie Chrystusa, w chronologii dziejów lub temperatura topnienia lodu w skali temperatur Celsjusza. Od nich można odliczać jednostki miary (lata, stopnie), w kierunku dodatnim lub w kierunku ujemnym. Do wyników pomiaru opartych na skali porządko-

*) Pojęcia te zostaną dokładniej omówione w dalszej części podręcznika.

**) Omówienie powyższych pojęć można znaleźć w dalszej części podręcznika.

wej można stosować – oprócz wymienionych wcześniej operacji statystycznych – również takie statystyki jak: średnie arytmetyczne, odchylenia standardowe i korelacje według momentu iloczynowego Pearsona. Miary te omówimy szczegółowo w dalszej części podręcznika. Skala przedziałowa ze, względu na brak tzw. zera absolutnego, nie daje jednak możliwości oceny stosunku mierzonych wielkości. Zmiana położenia umownego zera na osi badanej zmiennej powoduje zmianę stosunków między liczbami otrzymanymi w rezultacie pomiaru, pomimo braku zmian pomiędzy odpowiednimi stanami badanej cechy.

Jeżeli zostanie ustalony naturalny punkt zerowy skali, to wówczas możliwe staje się określenie stosunków między wynikami pomiaru. Skala, która to umożliwia nosi nazwę **skali ilorazowej** (stosunkowej). Przykładem takiej skali może być skala metryczna długości przedmiotów lub skala termometryczna Kelvina. Skala ilorazowa jest najsilniejszą spośród omówionych powyżej skal pomiarowych. W niektórych podręcznikach z zakresu teorii pomiaru *) można znaleźć jeszcze jedną – piątą – najsilniejszą skalę pomiarową określaną mianem **skali absolutnej**. Wyniki pomiarów uzyskuje się wówczas na drodze zliczania obiektów lub powtórzeń zjawisk. W przypadku tej skali niedopuszczalna jest żadna transformacja pierwotnego wyniku pomiaru. Jako przykład takiego pomiaru można podać zliczanie klientów kupujących określony produkt, zliczanie głosów w wyborach parlamentarnych itp.

Odwzorowując zbiór rozróżnialnych stanów w zbiór liczb rzeczywistych za pomocą określonej funkcji pomiarowej otrzymujemy zbiór wartości odpowiedniej **zmiennej**. Powyższy proces określa się mianem **kwantyfikacji cechy**, czyli przekształcenia cechy w odpowiednie zmienne (obrazy liczbowe). Z uwagi na to, że poziom natężenia badanych cech zależy zarówno od czynników systematycznych jak i losowych, otrzymane w wyniku kwantyfikacji zmienne nazywa się **zmiennymi losowymi**.

Zbiór rozróżnialnych stanów cechy badanej oraz zbiór wartości zmiennej może być zbiorem skończonym lub przeliczalnym. Liczba rozróżnialnych stanów, a następnie zbiór wartości zmiennej zależy od tzw. czułości metody badawczej. Jeżeli określona metoda posiada zdolność rozróżniania bardzo mało różniących się stanów, to wówczas zbiór wartości jest traktowany jako przedział na osi liczb rzeczywistych, a zmienną określa się mianem **zmiennej ciągłej**. Przykładem takich zmiennych mogą być: temperatura mierzona w stopniach Celsjusza lub zużycie prądu mierzone w kWh. Jeżeli natomiast zbiór wartości zmiennej składa się tylko z niektórych liczb rzeczywistych z określonych przedziałów, najczęściej liczb całkowitych nieujemnych, to wówczas zmienną

*) Zob. np. K. Walenta, *Podstawowe pojęcia teorii pomiaru*, w: *Problemy psychologii matematycznej* (red. J. Kozielecki), PWN, Warszawa 1971. Zob. także A. Iwasiewicz, *Zarządzanie jakością*, PWN, Warszawa-Kraków 1999, s. 124.

taką nazywa się **zmienną skokową** (dyskretną). Przykładem takiej zmiennej może być liczba dzieci w rodzinach, liczba domów pomocy społecznej w wybranym województwie, liczba wadliwych jednostek produktu w partii znajdującej się w magazynie określonego sklepu itp. Jeżeli metoda badawcza pozwala na wyróżnienie tylko dwóch stanów cechy (np. produkt: wykonany zgodnie z wymaganiami jakościowymi lub wadliwy), to wówczas zbiór wartości zmiennej jest zbiorem dwuelementowym. Jeśli stanowi pierwszemu przypiszemy wartość 0 a drugiemu wartość 1, to wówczas taką zmienną nazywać będziemy zmienną **zero–jedynkową** (dychotomiczną).

W praktyce podział zmiennych na ciągłe oraz skokowe nie jest zawsze wyraziste. Powodem tego są głównie ograniczenia wnoszone przez instrumenty pomiarowe i ich dokładność. Liczne zaokrąglenia wyników pomiarów (np. do jednego czy dwóch miejsc po przecinku) powodują, że otrzymujemy jedynie pewien skończony zbiór danych bez wartości pośrednich, sprawiając tym samym, że badana zmienna jest ciągła tylko na płaszczyźnie teoretycznej *).

Pojęcie zmiennej jest, więc pojęciem wtórnym w stosunku do pojęcia cechy. Zauważyć także należy, że jednej cesze można przyporządkować kilka zmiennych, których liczba jest zależna od liczby metod badania cechy oraz różnorodności funkcji pomiarowych. Cecha zużycie paliwa w samochodzie osobowym może być np. opisywana przez zmienną: ilość spalanej benzyny na 100 km lub zmienną: ilość przejechanych kilometrów na jednym litrze paliwa. Podobnie rzecz ma się z temperaturą, która może być mierzona na różnych skalach: w stopniach Celsjusza, Fahrenheita lub Kelvina.

1.3. Metody badań statystycznych

Każde badanie statystyczne wymaga ustalenia określonej metody badania. Rodzaj wybranej metody uzależniony jest od szeregu czynników takich jak: cel przeprowadzanego badania, charakter zbiorowości generalnej, liczność badanej zbiorowości, dokładność badania, budżet projektu badawczego, wielkość zespołów badawczych itp. W literaturze przedmiotu **) wyróżnia się zwykle:

— **badania pełne** (wyczerpujące, całkowite),

— **badania niepełne** (częściowe, wyrywkowe),

Badaniem pełnym nazywamy badanie obejmujące wszystkie jednostki statystyczne, zaliczane do określonej zbiorowości statystycznej. Badanie takie

*)W takim przypadku zmienne określa się jako *quasi*-ciągłe.

**)Zob. np. A. Komosa, J. Musiałkiewicz, *Statystyka*, Ekonomik 2001; N. Sobczyk, *Statystyka*, op. cit.

przeprowadza się najczęściej wówczas, gdy zbiorowość statystyczna nie jest zbyt liczna, koszt badania jednostki jest niski a badania nie mają charakteru niszczącego. Jeżeli z jakichś powodów (np. finansowych, organizacyjnych) badanie całościowe staje się niemożliwe lub bezcelowe, wówczas stosuje się badanie niepełne (częściowe), podczas którego badaniu podlega tylko pewien podzbiór zbiorowości generalnej nazywany **próbą** lub **zbiorowością próbną**[*)], przy czym elementy mogą być dobierane do próby w sposób celowy lub losowy.

Wśród technik pozyskiwania informacji wyróżnia się badania **ankietowe**, **monograficzne** i **reprezentacyjne**. Badania ankietowe polegają na skierowaniu do określonej grupy osób (tzw. respondentów) zaproszenia do dobrowolnego wypowiedzenia się na określony temat. Ankieta taka może być przesłana pocztą, zamieszczona w prasie lub wyłożona w miejscu publicznym. W pewnych sytuacjach informacje pochodzące od respondentów są zbierane i spisywane przez przedstawiciela instytucji badającej. W takim przypadku mówi się o tzw. wywiadzie i kwestionariuszu wywiadu. Badania ankietowe należą do technik pozyskiwania informacji, które można stosować zarówno w badaniach częściowych jak i wyczerpujących.

Badania monograficzne obejmują najczęściej niewielką grupę jednostek statystycznych, które są typowe dla danej zbiorowości albo wyróżniają się w charakterze pozytywnym lub negatywnym na tle pozostałych jednostek. Opis jednostek typowych pozwala na uogólnienie wniosków na całą zbiorowość generalną, natomiast badanie jednostek wyróżniających się pozwala na poznanie przyczyn ich odrębności. Badania monograficzne są najczęściej badaniami dokładnymi i szczegółowymi, a ich wyniki są w postaci liczbowej oraz opisowej.

Trzecia spośród technik stosowana w badaniach częściowych – metoda reprezentacyjna[**)] polega na wyciąganiu wniosków dotyczących zbiorowości generalnej na podstawie pobranej próby losowej. Próba taka ma być reprezentatywna w stosunku do zbiorowości generalnej, co oznacza, że powinna być ona miniaturą całej zbiorowości. Jednostki, które się w niej znajdą, muszą być reprezentatywne dla całej populacji. Reprezentatywność próby oznacza, że struktura próby i struktura populacji powinny być niemal identyczne. Na plus metody reprezentacyjnej należy także zapisać, że w odróżnieniu od ankiety czy badań monograficznych, podczas wnioskowania czy szacowania statystycznego, można określić wielkość popełnionego błędu szacunku.

Zarówno badania częściowe, jak i pełne mogą być prowadzone **sporadycznie** (jednorazowo), **okresowo** lub w sposób **ciągły**.

[*)] Więcej na temat próby i sposobu jej doboru zob. np. A. Iwasiewicz, Z. Paszek, *Statystyka z elementami statystycznych metod sterowania jakością*, op. cit., s. 75.

[**)] Szerzej o metodzie reprezentacyjnej zob. np. J. Steczkowski, *Reprezentacyjne badania jakości wyrobów, kontrola odbiorcza*, Kraków 1993.

Badaniem ciągłym może być np. ewidencja wydatków na reklamę, czy na wynagrodzenia w pewnej firmie, ewidencja zawieranych związków małżeńskich, urodzeń i zgonów, które odbywają się w sposób ciągły. Badaniami okresowymi mogą być np. odbywające się, co 10 lat, powszechne spisy ludności lub coroczne spisy ludności. Badania sporadyczne (jednorazowe) dokonywane są najczęściej w związku z zajściem jakiejś wyjątkowej sytuacji (np. klęski żywiołowej, czy klęski epidemiologicznej) i wiążą się koniecznością zdobycia niezbędnych informacji do podjęcia szybkiej decyzji.

Etapy badań statystycznych

Rozdział

2

Zwykle wymienia się kilka etapów badania statystycznego*). Są to:

1. Przygotowanie badania.
2. Zebranie materiału statystycznego (obserwacja statystyczna).
3. Przygotowanie, opracowanie i prezentacja materiału statystycznego.
4. Opis statystyczny badanego zjawiska lub wnioskowanie statystyczne.

2.1. Przygotowanie badania

Do podstawowych czynności występujących podczas przygotowania badania należy:

a. określenie celu i metody badania,

b. określenie zbiorowości statystycznej i cech podlegających badaniu,

c. zdefiniowanie jednostki sprawozdawczej,

d. określenie harmonogramu i budżetu projektu badawczego.

Celem badania może być np. ustalenie siły i kierunku współzależności pomiędzy stażem a wydajnością pracy, zbadanie częstotliwości i przyczyn wypadków na pewnym odcinku drogi, ustalenie potencjalnej liczby osób zainteresowanych wyjazdem na wycieczkę do Francji itp. We wszystkich tych przypadkach należy zdecydować czy badanie będzie pełne, czy częściowe.

Zbiorowość statystyczna jak i jednostki statystyczne – czyli przedmiot badania – powinny być dokładnie zdefiniowane pod względem rzeczowym, czasowym i przestrzennym. Np. w przypadku badania liczby klientów zainteresowanych wyjazdem do Francji zbiorowość statystyczną mogą tworzyć klienci pewnego biura podróży w mieście X, którzy w okresie od 1. 01. 1998 – 31. 12. 2001

*)Zob. M. Sobczyk, *Statystyka*, PWN 1998, s. 15; A. Komosa, J. Musiałkiewicz, *Statystyka*, Ekonomik 2001, s. 22.

skorzystali z jego usług. Badaną cechą statystyczną (cechą jakościową) może być tutaj kraj, do którego klient biura podróży zdecydował się wyjechać.

Jednostką sprawozdawczą może być osoba fizyczna lub prawna, która dysponuje danymi źródłowymi potrzebnymi do badania. W pewnych sytuacjach jednostką sprawozdawczą może być sama jednostka statystyczna. W omawianym powyżej przykładzie jednostką sprawozdawczą może być biuro rozważane podróży, jeżeli prowadzi ono bieżącą ewidencję obsługiwanych klientów lub klienci tego biura, w przypadku braku dokładnej ewidencji kierunków wyjazdów.

Określenie harmonogramu pracy i budżetu projektu badawczego pozwala na sprawne przeprowadzenie i ukończenie zaplanowanych badań lub ewentualną korektę zakresu i terminów otrzymania wyników końcowych.

2.2. Obserwacja statystyczna

Drugim etapem badań statystycznych jest obserwacja statystyczna. Polega ona na przyporządkowaniu wartości liczbowych cechom ilościowym oraz wariantów słownych cechom jakościowym u wszystkich jednostek wchodzących w skład zbiorowości generalnej lub w skład próby. Przyporządkowanie wartości cechom odbywa się na drodze pomiaru lub zbierania informacji od jednostek sprawozdawczych. Zebrane w ten sposób dane tworzą tzw. **szereg statystyczny nieuporządkowany** (szereg pierwotny).

Dane empiryczne mogą być obciążone pewnymi błędami zarówno o charakterze **systematycznym** jak i **przypadkowym**. Źródłem błędów systematycznych jest zwykle jednokierunkowa tendencja do zniekształcenia badanej rzeczywistości, co powoduje przy dużej liczbie powtórzeń znaczne zawyżenie lub zaniżenie końcowych rezultatów. Błędy o charakterze przypadkowym powstają zwykle z winy osób zbierających informacje. Błędy przypadkowe w odróżnieniu od błędów systematycznych mają zwykle różny kierunek (zawyżający lub zaniżający badaną wartość rzeczywistą), a ich wpływ na zniekształcenie badania jest zwykle mniejszy niż błędu systematycznego.

Czynnikiem przeciwdziałającym błędom (systematycznym i przypadkowym) są kontrole formalne i merytoryczne. Kontrola formalna ma za zadanie sprawdzić kompletność, pełność i zupełność zebranego materiału, natomiast celem kontroli merytorycznej jest sprawdzenie materiału pod względem logicznym i arytmetycznym[*)].

[*)] Szerzej o kontroli materiału statystycznego zob. J. Kordos, *Jakość danych statystycznych*, PWE, Warszawa, 1988. Zob. także: A. Komosa, J. Musiałkiewicz, *Statystyka*, Ekonomik 2001, s. 35.

2.3. Opracowywanie materiału statystycznego i jego prezentacja

Następnym etapem jest opracowywanie materiału statystycznego i jego prezentacja. Jedną z podstawowych czynności odbywających się w trakcie opracowywania materiału wyróżnia się tzw. **grupowanie i zliczanie**.

Grupowanie polega na wyodrębnieniu spośród całej badanej zbiorowości statystycznej określonych w miarę jednorodnych grup (części). Celem grupowania jest przejście od informacji dotyczących jednostek statystycznych do informacji dotyczących całej populacji lub jej części (próby).

Grupowanie odbywa się według określonych kryteriów. Najczęściej kryteriami tymi są stany (warianty) cech statystycznych. W przypadku cech o charakterze naturalnym takich jak np. płeć, podział odbywa się również w sposób naturalny (np. podział studentów określonego kierunku na grupę mężczyzn i kobiet). W innych przypadkach decyzja dotycząca liczby wyodrębnionych grup należy do prowadzącego badania. I tak np. pracowników uczelni można podzielić według pełnionych funkcji na pracowników naukowo-dydaktycznych i innych lub też naukowo-dydaktycznych, dydaktycznych, administracyjnych i technicznych.

Biorąc za kryterium podziału cel, jakiemu ma służyć grupowanie, możemy podzielić je na tzw. **typologiczne** i **wariancyjne** (oparte na zmienności). Grupowanie typologiczne opiera swój podział na wariantach cechy jakościowej (np. grupowanie ludności według wykształcenia na: podstawowe, zasadnicze zawodowe, średnie, wyższe). Grupowanie wariancyjne dotyczy zwykle cechy mierzalnej. Przykładem takiego grupowania może być podział pracowników określonej firmy ze względu na wielkość zarobków. Można wówczas wyróżnić przykładowe grupy (tzw. **przedziały klasowe**): $(700; 900]$, $(900; 1100]$, $(1100; 1300]$, $(1300; 1500]$, $(1500; 1700]$, $(1700; 1900]$, $(1900; 2100]$ itd. Przedziały klasowe mogą mieć jednakową lub różną długość, a przedziały skrajne mogą być zamknięte lub otwarte. Na przykład, jeżeli nie można określić najmniejszej teoretycznej stawki płacowej, to wówczas przedział klasowy zapisujemy 700 i mniej. Podobnie należy postąpić, gdy istnieje trudność z wyznaczeniem płacy maksymalnej wówczas ostatni z przedziałów można zapisać np. 2100 i więcej. Należy jednak zaznaczyć, że takie postępowanie w znaczny sposób ogranicza możliwość stosowania ilościowych metod analizy zebranego materiału. Dlatego też zaleca się, jeżeli jest to tylko możliwe, tworzenie przedziałów klasowych zamkniętych o równych długościach.

W literaturze przedmiotu można się spotkać z wieloma sposobami tworzenia przedziałów klasowych (formalnymi i mniej formalnymi). Sposoby te zostaną opisane w dalszej części tego podręcznika, podczas omawiana zagadnienia prezentacji materiału statystycznego w postaci szeregów statystycznych.

Po określeniu grup w obrębie zbiorowości statystycznej następuje **zliczanie danych** przypadających na wyodrębnione grupy. Jeżeli zbiorowość nie jest zbyt liczna, to zliczanie odbywa się ręcznie, natomiast w przypadku zbiorowości licznych do zliczania stosuje się technikę komputerową. Dość powszechnym sposobem zliczania jest tzw. sposób kreskowy, w którym pionowymi kreskami zaznacza się wystąpienie określonego wariantu cechy. Kreski te najczęściej grupowane są w „pęczki" po 5 sztuk, przy czym piąta kreska ułożona jest poziomo i przecina pozostałe 4 kreski. Innym sposobem zliczania kreskowego jest budowa z kresek figury kwadratu z przekątną. Poniżej został zamieszczony przykład zapisu kreskowego (w formie „pęczku" i „kwadratu z przekątną") liczb 12 i 16.

liczba 12 𝍸 𝍸 || liczba 12 ⊠⊠⌞

liczba 16 𝍸 𝍸 𝍸 | liczba 16 ⊠⊠⊠|

Zebrany i pogrupowany materiał musi być odpowiednio zaprezentowany na przykład w postaci **szeregów statystycznych** przedstawionych tabelarycznie i graficznie.

Szeregiem statystycznym nazywamy zbiór wyników obserwacji jednostek według pewnej cechy. Wyróżnić można następujące rodzaje szeregów statystycznych:

1. Szereg szczegółowy:
   - szereg szczegółowy nieuporządkowany (pierwotny),
   - szereg szczegółowy uporządkowany (pozycyjny).
2. Szereg rozdzielczy (strukturalny):
   - szereg rozdzielczy (strukturalny) cechy jakościowej,
   - szereg rozdzielczy (strukturalny) cechy ilościowej,
     - szereg rozdzielczy punktowy,
     - szereg rozdzielczy przedziałowy,
3. Szereg przestrzenny (geograficzny).
4. Szereg czasowy (dynamiczny):
   - szereg czasowy (dynamiczny) momentów,
   - szereg czasowy (dynamiczny) okresów.

Jeżeli jednostkowe wartości cechy mierzalnej lub niemierzalnej, zostaną spisane według kolejności badania jednostek statystycznych, to otrzymamy szereg szczegółowy nieuporządkowany (szereg pierwotny). Poniżej przedstawiono przykłady takich szeregów.

Przykład 2.1. W grupie 38 studentów studiów uzupełniających magisterskich, posiadających telefony komórkowe, przeprowadzono ankietę. Celem ankiety było zebranie informacji, z jakiej sieci telefonii komórkowej korzystają badani studenci. Otrzymano następujące wyniki przedstawione w postaci szeregu szczegółowego nieuporządkowanego:

Tab. 2.1. Rodzaj sieci telefonii komórkowej - szereg szczegółowy nieuporządkowany

ii	1	2	3	4	5	6	7	8	9	10
sieć	ERA	IDEA	IDEA	ERA	ERA	PLUS	PLUS	PLUS	ERA	ERA
ii	11	12	13	14	15	16	17	18	19	20
sieć	ERA	IDEA	IDEA	IDEA	ERA	IDEA	PLUS	PLUS	ERA	IDEA
ii	21	22	23	24	25	26	27	28	29	30
sieć	IDEA	ERA	PLUS	ERA	PLUS	IDEA	IDEA	PLUS	ERA	IDEA
ii	31	32	33	34	35	36	37	38		
sieć	ERA	IDEA	PLUS	PLUS	ERA	ERA	PLUS	ERA	—	—

Legenda: ERA sieć Era GSM; PLUS – sieć Plus GSM; IDEA – sieć IDEA
Źródło: badania własne.

Przykład 2.2. W jednym z miast województwa podkarpackiego zbadano liczbę osób korzystających z usług Biblioteki Miejskiej, w ciągu 100 kolejnych dni roboczych. Wyniki badania prezentuje tablica 2.2.

Tab. 2.2. Liczba osób korzystających z usług Biblioteki Miejskiej

ii	1	2	3	4	5	6	7	8	9	10
Liczba korzystających	103	88	72	46	93	88	60	62	45	62
ii	11	12	13	14	15	16	17	18	19	20
Liczba korzystających	72	79	63	52	57	97	78	65	61	63
ii	21	22	23	24	25	26	27	28	29	30
Liczba korzystających	57	72	67	80	55	77	82	52	64	69
ii	31	32	33	34	35	36	37	38	39	40
Liczba korzystających	84	85	71	84	95	67	68	53	46	73
ii	41	42	43	44	45	46	47	48	49	50
Liczba korzystających	73	72	56	56	77	104	79	76	85	53
ii	51	52	53	54	55	56	57	58	59	60
Liczba korzystających	77	110	72	73	47	82	95	79	65	87
ii	61	62	63	64	65	66	67	68	69	70
Liczba korzystających	66	83	88	75	61	102	58	85	86	60
ii	71	72	73	74	75	76	77	78	79	80
Liczba korzystających	61	70	65	101	94	111	70	79	38	119
ii	81	82	83	84	85	86	87	88	89	90
Liczba korzystających	101	72	74	63	52	89	91	102	77	80
ii	91	92	93	94	95	96	97	98	99	100
Liczba korzystających	44	50	96	115	98	125	131	82	98	73

Źródło: badania własne.

Jeżeli szereg pierwotny zostanie uporządkowany według określonego kryterium to wówczas nazwiemy go szeregiem szczegółowym uporządkowanym. Porządkowanie może odbywać się według różnego „klucza". W przypadku cechy jakościowej porządkowanie może odbywać się np. alfabetycznie lub według innego kryterium celowego. Natomiast wartości przypisane stanom cechy mierzalnej porządkuje się w sposób rosnący (od najmniejszej do największej) lub malejący (od największej do najmniejszej).

Jeżeli porządkowania dokonamy alfabetycznie, to wówczas powyższy szereg będzie przedstawiał się następująco:

Tab. 2.3. Rodzaj sieci telefonii komórkowej – szereg szczegółowy uporządkowany

i	1	2	3	4	5	6	7	8	9	10
ii	1	4	5	9	10	11	15	19	22	24
sieć	ERA	ERA	ERA	ERA	ERA	ERA	ERA	ERA	ERA	ERA
i	11	12	13	14	15	16	17	18	19	20
ii	29	31	35	36	38	2	3	12	13	14
sieć	ERA	ERA	ERA	ERA	ERA	IDEA	IDEA	IDEA	IDEA	IDEA
i	21	22	23	24	25	26	27	28	29	30
ii	16	20	21	26	27	30	32	6	7	8
sieć	IDEA	IDEA	IDEA	IDEA	IDEA	IDEA	IDEA	PLUS	PLUS	PLUS
i	31	32	33	34	35	36	37	38	—	—
ii	17	18	23	25	28	33	34	37	—	—
sieć	PLUS	PLUS	PLUS	PLUS	PLUS	PLUS	PLUS	PLUS	—	—

Źródło: tablica 2.1.

Pierwszy z indeksów (i) wskazuje, które miejsce zajmuje element w szeregu uporządkowanym. Natomiast drugi (ii) określa miejsce, które zajmował porządkowany element w szeregu pierwotnym (szczegółowym nieuporządkowanym). #

Równie łatwo i szybko można utworzyć szereg szczegółowy uporządkowany (pozycyjny) biorąc dane z przykładu 2.2. (Tab. 2.4.). Podobnie jak powyżej „i" jest indeksem porządkowym, wskazującym na miejsce (pozycję) kolejnych wartości w szeregu. Zmienną – liczbę korzystających z usług biblioteki – oznaczono symbolem X.

Szereg szczegółowy, zwłaszcza w przypadku, gdy zawiera dużą liczbę obserwacji, jest mało czytelny. Dlatego też najczęściej przekształca się go w szereg rozdzielczy. Szereg ten jest zbiorem wartości liczbowych uporządkowanych według stanów badanej cechy mierzalnej lub niemierzalnej, przy czym poszczególnym wariantom cechy przyporządkowane są odpowiadające im liczebności (f_j). W sytuacji, gdy rozpatrywana cecha jest niemierzalna (tak jak np. w przykładzie 2.1) tworzy się tak zwany szereg rozdzielczy cechy niemierzalnej (szereg jakościowy), natomiast w odniesieniu do cech mierzalnych konstruuje

się szeregi rozdzielcze punktowe i przedziałowe. Najczęściej szeregi rozdzielcze punktowe buduje się, gdy zmienna ma charakter skokowy natomiast, gdy zmienna przyjmuje wartości w sposób ciągły, do prezentacji materiału statystycznego stosuje się szereg rozdzielczy przedziałowy. Dopuszcza się także, zwłaszcza w przypadku dużego zagęszczenia zbioru wartości badanej zmiennej, konstrukcję szeregu rozdzielczego przedziałowego dla wartości zmiennej skokowej.

Tab. 2.4. Liczba osób korzystających z usług Biblioteki Miejskiej w 100 kolejnych dniach pracy zestawiona w szeregu szczegółowym uporządkowanym (pozycyjnym)

i	1	2	3	4	5	6	7	8	9	10
ii	79	91	9	4	39	55	92	14	28	85
Liczba korzystających (x_i)	38	44	45	46	46	47	50	52	52	52
i	11	12	13	14	15	16	17	18	19	20
ii	38	50	25	43	44	15	21	67	7	70
Liczba korzystających (x_i)	53	53	55	56	56	57	57	58	60	60
i	21	22	23	24	25	26	27	28	29	30
ii	19	65	71	8	10	13	20	84	29	18
Liczba korzystających (x_i)	61	61	61	62	62	63	63	63	64	65
i	31	32	33	34	35	36	37	38	39	40
ii	59	73	61	23	36	37	30	72	77	33
Liczba korzystających (x_i)	65	65	66	67	67	68	69	70	70	71
i	41	42	43	44	45	46	47	48	49	50
ii	3	11	22	42	53	82	40	41	54	100
Liczba korzystających (x_i)	72	72	72	72	72	72	73	73	73	73
i	51	52	53	54	55	56	57	58	59	60
ii	83	64	48	26	45	51	89	17	12	47
Liczba korzystających (x_i)	74	75	76	77	77	77	77	78	79	79
i	61	62	63	64	65	66	67	68	69	70
ii	58	78	24	90	27	56	98	62	31	34
Liczba korzystających (x_i)	79	79	80	80	82	82	82	83	84	84
i	71	72	73	74	75	76	77	78	79	80
ii	32	49	68	69	60	2	6	63	86	87
Liczba korzystających (x_i)	85	85	85	86	87	88	88	88	89	91
i	81	82	83	84	85	86	87	88	89	90
ii	5	75	35	57	93	16	95	99	74	81
Liczba korzystających (x_i)	93	94	95	95	96	97	98	98	101	101
i	91	92	93	94	95	96	97	98	99	100
ii	66	88	1	46	52	76	94	80	96	97
Liczba korzystających (x_i)	102	102	103	104	110	111	115	119	125	131

Źródło: opracowanie własne.

Sposób konstrukcji szeregów rozdzielczych, dla danych z przykładu 2.1 oraz 2.2, prezentują tablice 2.5 i 2.6.

Tab. 2.5. Rodzaj sieci telefonii komórkowej – szereg rozdzielczy

Sieć	Zliczanie danych metodą kreskową	Liczba abonentów (f_j)	Udział abonentów w %
ERA	𝍸 𝍸 𝍸	15	39%
IDEA	𝍸 𝍸 ||	12	32%
PLUS	𝍸 𝍸 |	11	29%
suma	xxx	38	100%

Źródło: obliczenia własne.

Tab. 2.6. Struktura liczby korzystających z biblioteki w 100 dniach pracy

j	Liczba korzyst. z biblioteki (x_j)	Zliczanie metodą kreskową	Liczba dni (f_j), w których było (x_j) korzystaj.	$f_{j.skum}$	Częstość względna (frakcja) v_j	v_j[%]
1	**2**	**3**	**4**	**5**	**6**	**7**
1	38	|	1	1	0,01	1,00
2	44	|	1	2	0,01	1,00
3	45	|	1	3	0,01	1,00
4	46	||	2	5	0,02	2,00
5	47	|	1	6	0,01	1,00
6	50	|	1	7	0,01	1,00
7	52	|||	3	10	0,03	3,00
8	53	||	2	12	0,02	2,00
9	55	|	1	13	0,01	1,00
10	56	||	2	15	0,02	2,00
11	57	||	2	17	0,02	2,00
12	58	|	1	18	0,01	1,00
13	60	||	2	20	0,02	2,00
14	61	|||	3	23	0,03	3,00
15	62	||	2	25	0,02	2,00
16	63	|||	3	28	0,03	3,00
17	64	|	1	29	0,01	1,00
18	65	|||	3	32	0,03	3,00
19	66	|	1	33	0,01	1,00
20	67	||	2	35	0,02	2,00
21	68	|	1	36	0,01	1,00
22	69	|	1	37	0,01	1,00
23	70	||	2	39	0,02	2,00
24	71	|	1	40	0,01	1,00
25	72	𝍸 |	6	46	0,06	6,00
26	73	||||	4	50	0,04	4,00
27	74	|	1	51	0,01	1,00
28	75	|	1	52	0,01	1,00
29	76	|	1	53	0,01	1,00
30	77	||||	4	57	0,04	4,00

Tab. 2.6. Struktura liczby korzystających z biblioteki w 100 dniach pracy cd.

j	Liczba korzyst. z biblioteki (x_j)	Zliczanie metodą kreskową	Liczba dni (f_j), w których było (x_j) korzystaj.	$f_{j.skum}$	Częstość względna (frakcja) v_j	v_j[%]
1	**2**	**3**	**4**	**5**	**6**	**7**
31	78	|	1	58	0,01	1,00
32	79	||||	4	62	0,04	4,00
33	80	||	2	64	0,02	2,00
34	82	|||	3	67	0,03	3,00
35	83	|	1	68	0,01	1,00
36	84	||	2	70	0,02	2,00
37	85	|||	3	73	0,03	3,00
38	86	|	1	74	0,01	1,00
39	87	|	1	75	0,01	1,00
40	88	|||	3	78	0,03	3,00
41	89	|	1	79	0,01	1,00
42	91	|	1	80	0,01	1,00
43	93	|	1	81	0,01	1,00
44	94	|	1	82	0,01	1,00
45	95	||	2	84	0,02	2,00
46	96	|	1	85	0,01	1,00
47	97	|	1	86	0,01	1,00
48	98	||	2	88	0,02	2,00
49	101	||	2	90	0,02	2,00
50	102	||	2	92	0,02	2,00
51	103	|	1	93	0,01	1,00
52	104	|	1	94	0,01	1,00
53	110	|	1	95	0,01	1,00
54	111	|	1	96	0,01	1,00
55	115	|	1	97	0,01	1,00
56	119	|	1	98	0,01	1,00
57	125	|	1	99	0,01	1,00
58	131	|	1	100	0,01	1,00
Suma			100		1,00	100,00

Źródło: obliczenia własne.#

Podczas budowy szeregu rozdzielczego, dla danych z przykładu 2.2, celowo zostały pominięte wszystkie wartości zmiennej, które nie wystąpiły w szeregu szczegółowym i ich liczebność empiryczna f_j wynosiła 0. Zmienna X (liczba korzystających z biblioteki) mogła teoretycznie przyjąć wartości: $0, 1, 2, \ldots, 8, 9, 10, \ldots, +\infty$, lecz w praktyce przyjmowała tylko wybrane wartości z tego przedziału i wypisywanie ich wszystkich podczas tworzenia szeregu rozdzielczego punktowego pogorszyłoby jego przejrzystość. Ostatnia z kolumn zawiera wielkości nazywane **frakcjami** lub **licznościami (częstościami) względ-**

nymi (v_j). Liczności te obliczamy według wzoru:

$$v_j = \frac{f_j}{\sum_{j=1}^{k} f_j} = \frac{f_j}{N} \tag{2.1}$$

lub

$$v_j = \frac{f_j}{\sum_{j=1}^{k} f_j} \cdot 100\ [\%] = \frac{f_j}{N} \cdot 100\ [\%], \tag{2.2}$$

gdzie N jest sumą wszystkich liczebności f_j, przyporządkowanych wartościom zmiennej $x_1, x_2, \ldots, x_j, \ldots, x_k$.

Wzór (2.2) wykorzystuje się, gdy chcemy otrzymać częstości względne wyrażone w procentach. Dzięki obliczonym częstościom względnym można stwierdzić, jak często w badanej zbiorowości występują jednostki posiadające określony wariant cechy. Na przykład na podstawie tablicy 2.5, obrazującej szereg rozdzielczy jakościowy, można zauważyć, że najwięcej (39%) spośród badanych studentów korzysta z usług sieci ERA GSM, 32% badanych z sieci Plus GSM, oraz 29% z sieci IDEA.
Suma wszystkich częstości jest równa 1 lub 100, gdy częstość wyrażona jest w procentach.

Jednym ze sposobów polepszenia czytelności szeregu statystycznego – w przypadku cechy mierzalnej – jest transformacja szeregu szczegółowego do postaci szeregu rozdzielczego przedziałowego.

W etapie pierwszym, na podstawie przedziału zmienności realizacji badanej zmiennej, ustala się dwa podstawowe parametry szeregu rozdzielczego przedziałowego, jakimi są **liczba przedziałów klasowych** (k) oraz ich **rozpiętość (długość)** (l). Poprzez rozpiętość klasy należy rozumieć różnicę pomiędzy górną i dolną granicą określonego przedziału klasowego. Liczba wyodrębnionych klas jest zależna od różnicy pomiędzy maksymalną i minimalną zrealizowaną wartością zmiennej, od liczebności zbiorowości oraz od celu badania. Ogólnie można powiedzieć, że im liczniejsza jest badana zbiorowość i im większy jest przedział zmienności zbioru realizacji zmiennej, tym większa powinna być liczba przedziałów klasowych. Nie można jednak przesadzać z nadmiernym podziałem zbiorowości gdyż prowadzi to do nadmiernej szczegółowości, a tym samym utrudnia opis i wyciąganie wniosków.

Studiując literaturę przedmiotu można natrafić na szereg zaleceń i sposobów wyznaczania parametru, jakim jest liczba przedziałów klasowych. Na przykład K. Zając proponuje, aby zbiorowość liczącą 40 - 60 jednostek podzie-

lić na 6 - 8 klas, przy liczebności 60 - 100 jednostek wyodrębnić 7 - 10 klas, przy 100 - 200 jednostkach 9 - 12 klas, natomiast przy 200 - 500 12 - 17 klas [*].

Można się również spotkać z sposobem, że przybliżoną wartość liczby przedziałów ustala się w oparciu o zasadę $k = \sqrt{N}$, gdzie N jest liczebnością zbiorowości. Niekiedy też przyjmuje się następującą regułę postępowania [**]:

Tab. 2.7.

N	k
50	8
100	10
500	13
1000	15
10000	20

Wartość k można wyznaczyć również według następujących wzorów [***]:

$$k \approx 1 + 3,3 \lg N, \tag{2.3}$$

i jednocześnie

$$0,5\sqrt{N} \leqslant k \leqslant \sqrt{N}. \tag{2.4}$$

Znak $\approx$ we wzorze 2.3 oznacza, że wartość k uzyskuje się poprzez zaokrąglenie obliczonej wielkości do najbliższej liczby całkowitej.

Drugim z koniecznych parametrów, niezbędnym do zbudowania szeregu statystycznego rozdzielczego jest rozpiętość przedziału klasowego (l), którą ustala się dzieląc rozstęp badanej zmiennej (R) przez liczbę przedziałów klasowych. Rozstęp jest to różnica pomiędzy maksymalną (x_{max}) i minimalną wartością (x_{min}) realizacji badanej zmiennej X. Można więc zapisać:

$$l = \frac{R}{k} = \frac{x_{max} - x_{min}}{k}. \tag{2.5}$$

Jeżeli powyższy iloraz ma wartość utrudniającą dalsze obliczenia numeryczne, to należy dokonać korekty końców przedziału zmienności. Kres dolny x_{min} należy obniżyć do poziomu x'_{min} takiego, że $x'_{min} < x_{min}$, natomiast kres górny x_{max} podwyższyć do poziomu x'_{max} takiego, że $x'_{max} > x_{max}$. Należy pamiętać, aby korekta była możliwie najmniejsza w stosunku do pierwotnych granic zmienności. Korekcie można poddać, również tylko jeden z kresów zmienności –

[*] K. Zając, *Zarys metod statystycznych*, PWE, Warszawa, 1988, s. 92.

[**] Zob. J. Bielecki, B. Jurkiewicz, Z. Szymanowska, *Zbiór zadań ze statystyki ogólnej i matematycznej*, PWN, Warszawa 1975, s. 9.

[***] Zob. np. A. Iwasiewicz, Z. Paszek, *Statystyka z elementami statystycznych metod kontroli jakości*, AE w Krakowie, Kraków 2000, s. 94.

dolny lub górny. Po dokonanej korekcie należy ponownie obliczyć wartość parametru l, podstawiając tym razem w liczniku wyrażenia (2.5) $R = x'_{max} - x'_{min}$. Czynność tę powtarzamy tak długo aż powyższy iloraz przyjmie wartość, która nie będzie utrudniać dalszych obliczeń liczbowych. Po ustaleniu parametrów k i l, kolejne przedziały klasowe będą miały postać:

$$(x_{d.j}; x_{g.j}], \tag{2.6}$$

przy czym: $j = 1, 2, \ldots, k$; $x_{d.1} = x'_{min}$; $x_{g.k} = x'_{max}$; $x_{g.j} - x_{d.j} = l$.
W następnym etapie budowy szeregu rozdzielczego przedziałowego, następuje zliczanie liczebności obserwacji przypadających na wyszczególnione przedziały, przy czym realizacja x_i będzie należeć do j-tego przedziału klasowego, jeżeli:

$$x_{d.j} < x_i \leqslant x_{g.j}, \tag{2.7}$$

Budowę szeregu rozdzielczego przedziałowego dla danych z przykładu 2.2, przedstawia poniższa tablica 2.8. Podczas konstrukcji powyższego szeregu statystycznego liczba klas $k = 1 + 3,3 \lg 100 = 7,6 \approx 8$. Spełniona jest również zależność: $0,5\sqrt{100} = 5 \leqslant k = 8 \leqslant \sqrt{100} = 10$.
Pierwotna długość przedziału: $l = \frac{R}{k} = \frac{x_{max} - x_{min}}{k} = \frac{131-38}{8} = 11,65$. Ponieważ 11,65 jest wartością, która może utrudniać dalsze obliczenia, zdecydowano się dokonać korekty kresów zmienności i założono, że: $x'_{min} = 36$ i $x'_{max} = 132$. Nowa długość przedziałów klasowych $l = \frac{R}{k} = \frac{x'_{max} - x'_{min}}{k} = \frac{131-36}{8} = 12$.

Tab. 2.8. Struktura liczby korzystających z usług biblioteki w 100 dniach pracy

j	$(x_{d.j}; x_{g.j}]$	Zliczanie metodą kreskową	Liczba dni w których było (f_j) odwiedzaj.	Częstość względna (v_j)	Częstość względna v_j[%]
1	**2**	**3**	**4**	**5**	**6**
1	(36;48]	𝍸 |	6	0,06	6,00
2	(48;60]	𝍸 𝍸 ||||	14	0,14	14,00
3	(60;72]	𝍸 𝍸 𝍸 𝍸 |	26	0,26	26,00
4	(72;84]	𝍸 𝍸 𝍸 𝍸 |||	24	0,24	24,00
5	(84;96]	𝍸 𝍸 ||||	15	0,15	15,00
6	(96;108]	𝍸 |||	9	0,09	9,00
7	(108;120]	|||	4	0,04	4,00
8	(120;132]	||	2	0,02	2,00
SUMA		**XXX**	**100**	**1,00**	**100,00**

Źródło: obliczenia własne.

Podobnie – jak w przypadku szeregu rozdzielczego punktowego – obok liczności bezwzględnych (f_j) można umieścić częstości względne (v_j). Z powyższego szeregu rozdzielczego wyraźnie wynika, że najwięcej obserwacji 26%

przypada na przedział klasowy (60; 72], a najmniej, 2%, na skrajny przedział (120; 132]. Zatem można stwierdzić, że z biblioteki korzystało najczęściej od 60 do 72 osób dziennie.

Tworzenie szeregów rozdzielczych przedziałowych ma jednak i swoje wady. Jedną z nich jest utrata pewnej ilości informacji. Wada ta nie występuje w przypadku szeregu szczegółowego oraz w przypadku szeregu rozdzielczego punktowego. Zauważmy, że mając szereg szczegółowy uporządkowany możemy go łatwo transformować do postaci szeregu rozdzielczego punktowego. Tak samo łatwo, rozpisując szereg rozdzielczy punktowy, można otrzymać szereg szczegółowy uporządkowany. A zatem w przypadku tych dwóch typów szeregów transformacja może przebiegać dwukierunkowo. Natomiast, jeżeli weźmiemy szereg szczegółowy uporządkowany i szereg rozdzielczy przedziałowy, to wówczas omawiane przekształcenie może być tylko jednokierunkowe – od szeregu szczegółowego uporządkowanego do szeregu rozdzielczego przedziałowego. Nie można jednak wychodząc od szeregu rozdzielczego przedziałowego dojść do szeregu szczegółowego uporządkowanego. W przypadku szeregu rozdzielczego przedziałowego znana jest tylko liczba realizacji badanej zmiennej zawartych w poszczególnych przedziałach. Nieznane są natomiast ich dokładne wartości. Zatem można stwierdzić, że decydując się na tworzenie szeregu rozdzielczego, świadomie wprowadzamy pewien błąd. Błąd ten nosi nazwę **błędu grupowania**, a jego wielkość – jak zobaczymy w dalszej części tego podręcznika – wpływa na wartości charakterystyk opisujących badaną zbiorowość.

Dość często, podczas analizowania danych zawartych w szeregu rozdzielczym punktowym i przedziałowym, istnieje potrzeba określenia sumy liczebności (lub częstości) określonego wariantu zmiennej i liczebności (częstości) wariantów poprzedzających. W przypadku szeregu rozdzielczego przedziałowego sumowaniu podlegają liczebności (częstości) określonej klasy i klas, które ją poprzedzają. Działanie takie prowadzi do zbudowania tzw. **szeregu skumulowanego (kumulacyjnego)**. Technikę konstrukcji tego szeregu zilustrowano wykorzystując dane w postaci szeregu rozdzielczego z tablicy 2.8.

Pierwsza pozycja zapisu w szeregu skumulowanym jest identyczna jak w szeregu rozdzielczym, druga pozycja oznacza, że w 20 przypadkach (dniach pracy biblioteki) liczba osób korzystających z jej usług nie przekroczyła 60 osób, co stanowi 20% $(20/(N = 100) = 0,20)$. W podobny sposób należy interpretować pozostałe kolejne zapisy w przedostatniej i ostatniej kolumnie tablicy 2.9.

Szereg skumulowany można również tworzyć bezpośrednio w oparciu o szereg szczegółowy uporządkowany. Aby zilustrować technikę jego tworzenia posłużymy się danymi pochodzącymi z tego samego przykładu 2.2, ale zebranymi w postaci szeregu szczegółowego uporządkowanego (zob. tablica 2.4).

Tab. 2.9. Struktura liczby korzystających z usług biblioteki w 100 dniach pracy

j	$(x_{d\cdot j}; x_{g\cdot j}]$	f_j	v_j	Kumulacja liczebności obliczenia pomocnicze	Liczebności skumulowane $f_{j\cdot skum}$	Częst. względ. skumulowane $v_{j\cdot skum}$
1	2	3	4	5	6	7
1	(36;48]	6	0,06	6	6	0,06
2	(48;60]	14	0,14	6+14	20	0,20
3	(60;72]	26	0,26	6+14+26	46	0,46
4	(72;84]	24	0,24	6+14+26+24	70	0,70
5	(84;96]	15	0,15	6+14+26+24+15	85	0,85
6	(96;108]	9	0,09	6+14+26+24+15+9	94	0,94
7	(108;120]	4	0,04	6+14+26+24+15+9+4	98	0,98
8	(120;132]	2	0,02	6+14+26+24+15+9+4+2	100	1,00
SUMA		100	1,00	XXX	XXX	XXX

Źródło: obliczenia własne.

Szereg skumulowany przedstawia się wówczas następująco:

{(38; 1), (44; 2), (45; 3), (46; 5), (47; 6), (50; 7), (52; 10), (53; 12), (55; 13), (56; 15), (57; 17), (58; 18), (60; 20),(61; 23), (62; 25), (63; 28), (64; 29), (65; 32), (66; 33), (67; 35), (68; 36), (69; 37), (70; 39), (71; 40), (72; 46), (73; 50), (74; 51), (75; 52), (76; 53), (77; 57), (78; 58), (79; 62), (80; 64), (82; 67), (83; 68), (84; 70), (85; 73), (86; 74), (87; 75), (88; 78), (89; 79), (91; 80), (93; 81), (94; 82), (95; 84), (96; 85), (97; 86), (98; 88), (101; 90), (102; 92), (103; 93), (104; 94), (110; 95), (111; 96), (115; 97), (119; 98), (125; 99), (131; 100)}.

Wyrazy znajdujące się w tym szeregu składają się z dwóch współrzędnych $(x_i; i)$. Pierwsza z nich (x_i) odpowiada zrealizowanej wartości zmiennej, natomiast druga pokrywa się z numerem indeksu (i) w którym ostatni raz zrealizowała się x_i. Np. para (38; 1) oznacza, że był tylko taki jeden dzień, w którym bibliotekę odwiedziło zaledwie 38 osób, (44; 2) informuje nas, że w dwóch przypadkach liczba odwiedzających bibliotekę nie przekroczyła 44 osób; (45; 3) oznacza, że w co najwyżej 3 przypadkach liczba korzystających wynosiła 45 osób lub mniej. Podobnie należy interpretować pozostałe pary występujące w tym szeregu.

Jeżeli każdą drugą współrzędną podzielimy przez sumę wszystkich obserwacji (N), to wówczas otrzymamy szereg skumulowany częstości względnych. W analizowanym przykładzie $N = 100$, a szereg skumulowany częstości względnych będzie się przedstawiał się następująco:

{(38; 1/100), (44; 2/100), (45; 3/100), (46; 5/100), (47; 6/100), (50; 7/100), (52; 10/100), (53; 12/100), (55; 13/100), (56; 15/100),(57; 17/100), (58; 18/100), (60; 20/100), (61; 23/100), (62; 25/100), (63; 28/100), (64; 29/100), (65; 32/100), (66; 33/100), (67; 35/100), (68; 36/100), (69; 37/100), (70; 39/100), (71; 40/100),

(72;46/100), (73;50/100), (74;51/100), (75;52/100), (76;53/100), (77;57/100), (78;58/100), (79;62/100), (80;64/100), (82;67/100), (83;68/100), (84;70/100), (85;73/100), (86;74/100), (87;75/100), (88;78/100), (89;79/100), (91;80/100), (93;81/100), (94;82/100), (95;84/100), (96;85/100), (97;86/100), (98;88/100), (101;90/100), (102;92/100), (103;93/100), (104;94/100), (110;95/100), (111;96/100), (115;97/100), (119;98/100), (125;99/100), (131;100/100)}.

Uważny czytelnik może łatwo zauważyć, że identyczne szeregi skumulowane można utworzyć biorąc za punkt wyjścia dane przedstawione w postaci szeregu rozdzielczego punktowego (zob. np. dane w tablicy 2.6, kolumna 5.). Postępowanie będzie wówczas identyczne jak w przypadku szeregu rozdzielczego przedziałowego.

Kolejną grupę szeregów statystycznych stanowią szeregi przestrzenne i czasowe. Szeregi przestrzenne, nazywane niekiedy geograficznymi lub terytorialnymi, przedstawiają rozmieszczenie danych statystycznych na tle jednostek administracyjnych (województw, powiatów), części świata regionów gospodarczych i przemysłowych. Przykład takiego szeregu został zaprezentowany w tablicy 2.10.

Tab. 2.10. Przykład szeregu przestrzennego

Województwa	Ludność – stan w dniu 31. III. 2002 r. (w tys.)		
	Ogółem	Miasta	Wieś
Polska	38627,8	23839,7	14788,1
Dolnośląskie	2968,9	2119,5	849,4
Kujawsko-pomorskie	2101,6	1304,9	796,7
Lubelskie	2226,0	1043,4	1182,6
Lubuskie	1024,6	661,4	363,2
Łódzkie	2630,4	1703,3	927,1
Małopolskie	3235,8	1622,9	1612,9
Mazowieckie	5080,8	3265,5	1815,3
Opolskie	1079,6	564,5	515,1
Podkarpackie	2131,2	871,4	1259,8
Podlaskie	1219,4	715,5	503,9
Pomorskie	2205,7	1501,9	703,8
Śląskie	4834,5	3834,8	999,7
Świętokrzyskie	1318,5	605,1	713,4
Warmińsko-mazurskie	1469,5	883,9	585,6
Wielkopolskie	3366,9	1936,5	1430,4
Zachodniopomorskie	1734,4	1205,2	529,2

Źródło: www.stat.gov.pl

Szereg czasowy (chronologiczny, rozwojowy, dynamiczny) ukazuje natomiast rozwój określonych zjawisk w czasie. Składa się one zwykle z dwóch ko-

lumn (wierszy). W pierwszej kolumnie (wierszu) ujęte są momenty czasu (szereg czasowy momentów) lub okresy (szereg czasowy okresów) natomiast w drugiej kolumnie (wierszu) wielkości badanego zjawiska korespondujące z określonym momentem lub okresem czasu. Momentem czasu może być, np. 31 grudnia lub 1 stycznia każdego roku, natomiast okresem pewien przedział jak lata, miesiące czy kwartały. Tablica 2.11 prezentuje przykładowe dane będące szeregiem czasowym momentów, natomiast tablica 2.12 to przykład szeregu czasowego okresów.

Tab. 2.11. Kurs akcji spółki giełdowej Agora S.A.

t	Data (rr-mm-dd)	Kurs zamknięcia
1	01-01-02	86,40
2	01-01-03	85,00
3	01-01-04	87,50
4	01-01-05	83,70
5	01-01-08	80,40
6	01-01-09	82,10
7	01-01-10	79,00
8	01-01-11	76,00
9	01-01-12	77,60
10	01-01-15	74,80
11	01-01-16	71,20
12	01-01-17	74,00
13	01-01-18	75,30
14	01-01-19	77,00
15	01-01-22	75,60
16	01-01-23	74,30
17	01-01-24	74,30
18	01-01-25	74,80
19	01-01-26	73,20
20	01-01-29	73,10

Źródło: www://penetrator.pl

Tab. 2.12. Przychody budżetowe wynikające z prywatyzacji w latach 1992 – 2001

Rok	Przychód w mld. zł.
1992	0.3
1993	0,4
1994	0,8
1995	1,7
1996	1,9
1997	6,6
1998	7,0
1999	13,3
2000	27,2
2001	6.8

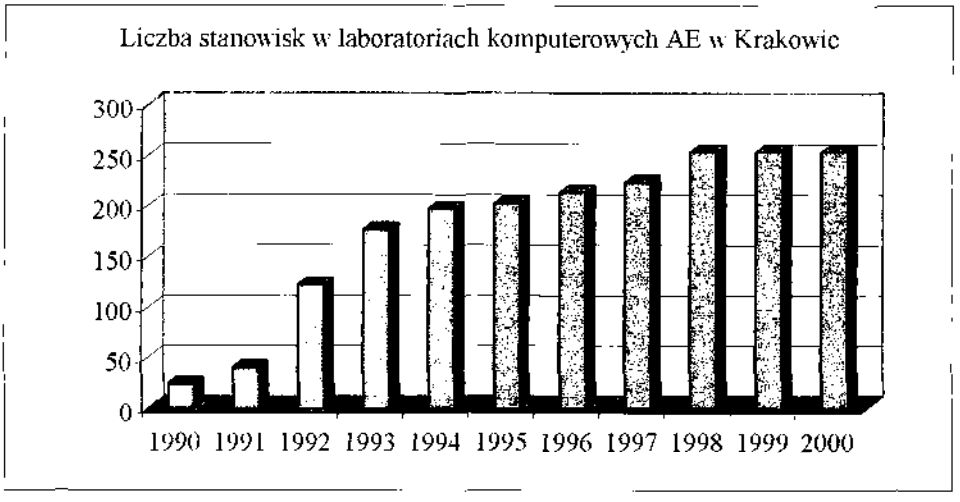
Źródło: Gazeta Wyborcza – Gospodarka, 12 września 2002 r.

Dane statystyczne można również prezentować graficznie za pomocą wykresów. Głównym zadaniem wykresów jest wizualizacja uogólnionych informacji statystycznych. Głównym źródłem, wykorzystywanym podczas tworzenia wykresów są szeregi i tablice statystyczne. Każdy wykres, podobnie jak tablica statystyczna powinien posiadać swój tytuł oraz źródło pochodzenia danych w oparciu, o które został on sporządzony. Oprócz tego obok wykresu, jeżeli wymaga tego jego charakter, powinna znajdować się legenda z objaśnieniem symboli, kolorów czy przyjętej skali – użytych podczas tworzenia wykresu.

Najczęściej stosowane rodzaje wykresów to*):

— liniowe,
— powierzchniowe,
— słupkowe,
— bryłowe,
— punktowe,
— mapowe,
— kombinowane i specjalne.

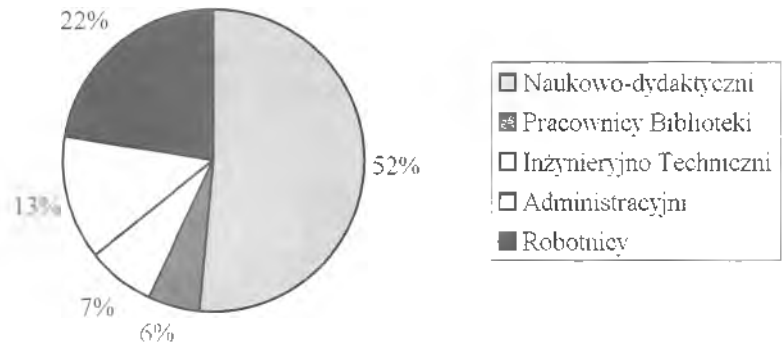
Sporządzenie większości z tych typów wykresów umożliwiają powszechnie stosowane arkusze kalkulacyjne np. Excel czy Lotus. Poniżej przedstawiono kilka przykładów różnych typów wykresów:



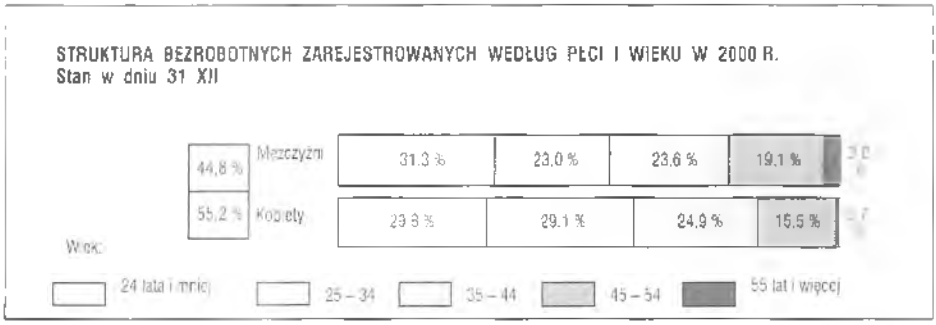
Rys. 2.1. Przykład wykresu słupkowego
Źródło: folder pt. *Akademia Ekonomiczna w Krakowie w liczbach*, AE w Krakowie, Kraków 2001.

*)Obszerny opis typów i sposobów tworzenia wykresów statystycznych można znaleźć w: *Wykresy i mapy statystyczne*, Główny Urząd Statystyczny, Warszawa 1977.

Rys. 2.2. Przykład wykresu mapowego
Źródło: www.stat.gov.pl



Rys. 2.3. Przykład wykresu kołowego – Struktura zatrudnienia w AE w Krakowie w 2000 roku
Źródło: folder pt. *Akademia Ekonomiczna w Krakowie w liczbach*, op. cit.



Rys. 2.4. Przykład wykresu powierzchniowego
Źródło: www.stat.gov.pl

Odrębną grupę wykresów stanowią **wykresy sporządzone w prostokątnym układzie współrzędnych.** W grupie tej należy wyróżnić:

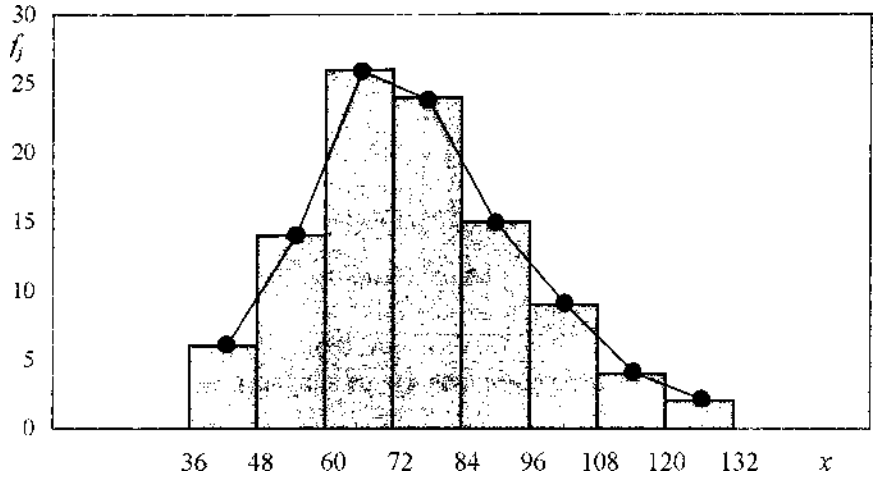
— wykresy strukturalne służące do opisu szeregów rozdzielczych (takie jak: histogram, wielobok liczebności),

— wykresy dynamiczne służące do opisu szeregów dynamicznych (czasowych),

— wykresy korelacyjne służące do zobrazowanie rodzaju współzależności pomiędzy cechami.

Histogram (zob. rys. 2.5) tworzymy w ten sposób, że na osi odciętych odkładamy granice przedziałów klasowych, a na osi rzędnych liczebności (f_j) lub częstości względne (v_j) odpowiadające poszczególnym przedziałom. Jest to zbiór przylegających prostokątów, których podstawy są równe długości przedziałów

klasowych a wysokości są liczebnościami lub częstościami tych przedziałów. W przypadku, gdy przedziały klasowe nie są równe, na osi rzędnych odkłada się wartości wskaźnika natężenia wyznaczonego w następujący sposób: wskaźnik natężenia = (liczebność danej klasy · interwał klasy najwęższej lub klasy najszerszej)/interwał danej klasy.

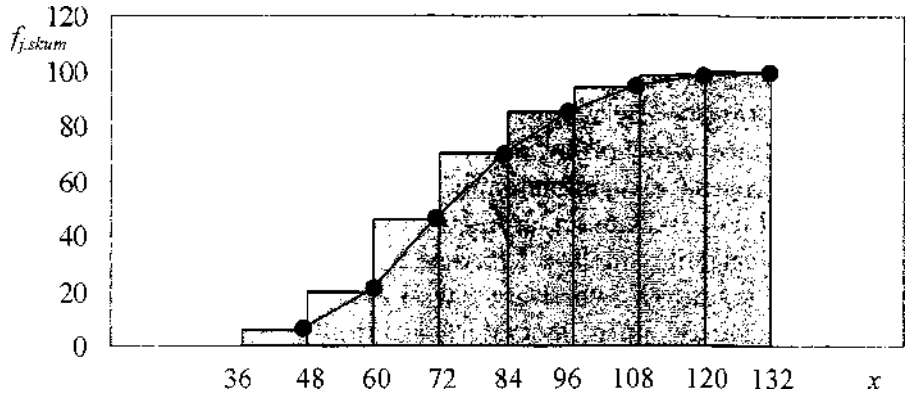


Rys. 2.5. Histogram oraz wielobok liczebności przedstawiający liczbę dni f_i, w których obsłużono $[x_{di} - x_{gi}]$ osób
Źródło: tablica 2.8.

Dopuszczalna jest także konstrukcja histogramów dla wartości skumulowanych. Wówczas, na osi rzędnych, zamiast liczebności (f_j) lub częstości (v_i) odkładane są liczebności skumulowane $(f_{j.skum})$ lub odpowiednie częstości skumulowane $(v_{j.skum})$. Wielobok liczebności lub częstości jest to łamana powstała poprzez połączenie punktów, których pierwsza współrzędna jest środkiem przedziału a druga liczebnością lub częstością względną. Jest to więc łamana łącząca środki wierzchołków prostokątów.

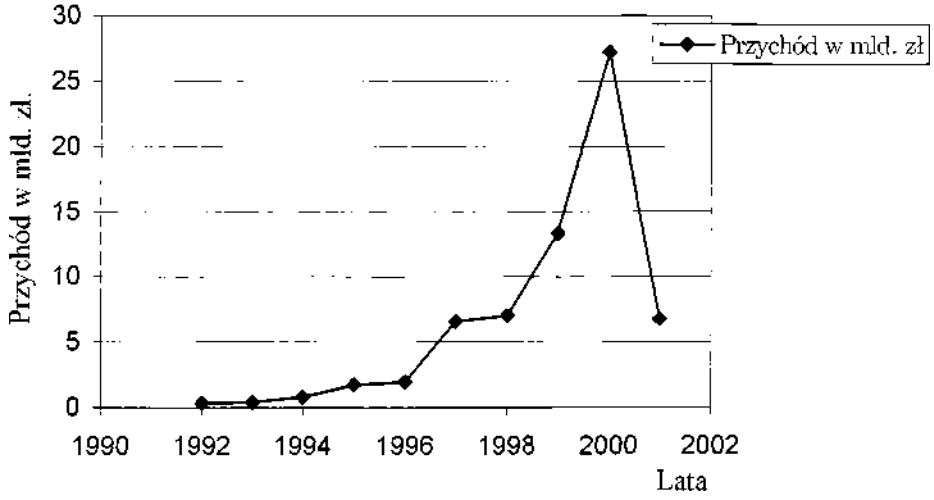
W podobny sposób można otrzymać histogram dla liczebności skumulowanych (wielobok liczebności, wykres empirycznej dystrybuanty) łącząc tym razem punkty o współrzędnych: górne granice klas przedziałowych i odpowiadające im liczebności (częstości) skumulowane. Przykład takiego histogramu prezentuje rys. 2.6.

Im większa jest liczba przedziałów klasowych i im mniejszy jest interwał klasowy, tym połączenie punktów w diagramie (liczebności lub kumulacyjnym) staje się gładsze, a łamana upodabnia się do krzywej. Będziemy ją nazywać krzywą liczebności lub krzywą częstości. Z powyższego wynika, że zwiększanie liczby przedziałów i zmniejszanie ich interwałów stanowi jeden ze sposobów przejścia od diagramu do krzywej liczebności.



Rys. 2.6. Graficzna prezentacja szeregu kumulacyjnego
Źródło: tablica 2.9.

Drugim rodzajem wykresu sporządzanego w układzie współrzędnych jest wykres dynamiczny, przedstawiany najczęściej w postaci łamanej łączącej punkty o współrzędnych: moment czasowy (okres) i wartość zmiennej badanej korespondującej z danym momentem czasowym (okresem). Przykład takiego wykresu prezentuje rys. 2.7.



Rys. 2.7. Przychody budżetowe wynikające z prywatyzacji w latach 1992 - 2001
Źródło: tablica 2.12.

Trzeci z typów wykresów wykres korelacyjny służy do graficznej prezentacji współzależności pomiędzy dwoma zmiennymi. Sposób jego budowy i interpretacji zostanie omówiony w rozdziale 4.

2.4. Opis lub wnioskowanie statystyczne

Ostatnim etapem badań statystycznych jest opis lub wnioskowanie statystyczne. Opis statystyczny odnosi się tylko do danej zbiorowości statystycznej lub pochodzącej z niej próby. Ma on charakter sumaryczny i uogólniający. Opis taki posiłkuje się różnymi miarami, spośród których wyróżniają się miary położenia (średnie), zmienności, asymetrii i koncentracji oraz miary współzależności (współczynniki korelacji i funkcje regresji). Metody wykorzystywane do opisów statystycznych wchodzą w zakres statystyki opisowej.

W odróżnieniu od opisu statystycznego wnioskowanie statystyczne ma miejsce wówczas, gdy wykorzystując wiadomości zebrane w drodze badania reprezentatywnej próby staramy się ekstrapolować wnioski na całą zbiorowość, z której próba ta pochodzi. Działanie takie nazywa się wnioskowaniem statystycznym i opiera się w głównej mierze na rachunku prawdopodobieństwa, który stanowi jego teoretyczną podstawę.

Metody wnioskowania statystycznego zaliczane są do drugiego działu statystyki nazywanego statystyką matematyczną. Spośród metod statystyki matematycznej wyróżnia się najczęściej teorię estymacji oraz teorię weryfikacji hipotez statystycznych.

Opisowa analiza struktury zbiorowości statystycznej

Rozdział 3

3.1. Rozkłady empiryczne zmiennej losowej

Rozkładem empirycznym zmiennej losowej jest funkcja, która wartościom zmiennej (x_j) przyporządkowuje liczebności (f_j). Rozkład empiryczny charakteryzuje strukturę badanej zbiorowości ze względu na wyróżniona cechę, opisywaną przez zmienną (X).

Analiza szeregów rozdzielczych oraz ich graficznych prezentacji w postaci histogramów pozwala na rozpoznanie typu rozkładu empirycznego. Zwykle stosuje się kilka kryteriów decydujących o typie rozkładu empirycznego *):

— ze względu na charakter badanej cechy (zmiennej) wyróżnia się rozkłady zmiennej skokowej i ciągłej,

— ze względu na liczbę maksimów wyróżniamy rozkłady jednomodalne (posiadające jedno maksimum), dwumodalne (o dwóch maksimach) i wielomodalne (o więcej niż dwóch maksimach),

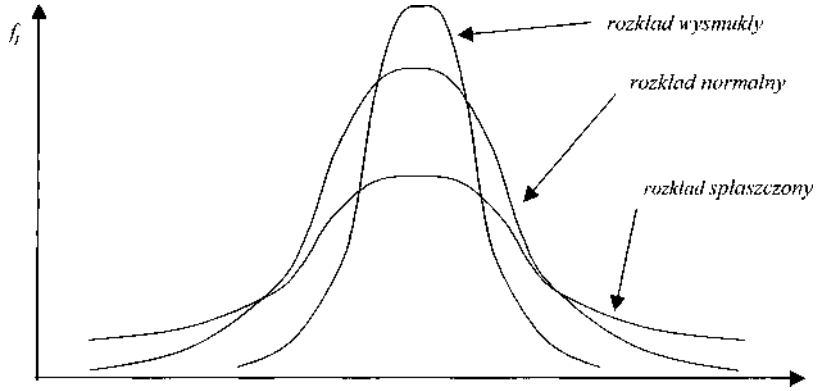
— ze względu na symetryczność rozłożenia wartości w stosunku do wartości centralnej, rozkłady symetryczne (dla których obserwacje rozłożone są symetrycznie po obu stronach osi symetrii) i rozkłady asymetryczne (dla których obserwacje rozłożone są niesymetrycznie).

Jeżeli rozkład empiryczny jest jednomodalny to wówczas punktem odniesienia podczas badania symetryczności/asymetryczności rozkładu jest punkt będący maksimum rozkładu. Wśród rozkładów symetrycznych zmiennej ciągłej, o jednym maksimum, wyróżnia się rozkład normalny **), rozkład leptykurtyczny (wysmukły) i rozkład plaktykurtyczny (spłaszczony). Przykłady

*) Zob. M. Sobczyk, *Statystyka*, op. cit., s. 27; W. Makać. D. Urbanek-Krzysztofiak, *Metody opisu statystycznego*, Wydawnictwo Uniwersytetu Gdańskiego, Gdańsk 2001, s. 51.

**) Rozkład normalny jest jednym z ważniejszych rozkładów znajdujących zastosowanie w statystyce matematycznej.

takich rozkładów przedstawia rys. 3.1. Rozkłady empiryczne o charakterze symetrycznym spotykane są jednak bardzo rzadko. Najczęściej rozkłady empiryczne są asymetryczne, przy czym asymetria może być umiarkowana lub też skrajna. Ze względu na kierunek asymetrii wyróżnia się natomiast asymetrię prawostronną i lewostronną. Jeżeli większa powierzchnia pod krzywą liczebności znajduje się po prawej stronie punktu maksimum to wówczas rozkład jest prawostronnie asymetryczny (prawoskośny). W przeciwnym razie rozkład jest lewostronnie asymetryczny (lewoskośny). Jeżeli rozkład posiada tylko jedno „ramię" to wówczas o takim rozkładzie mówimy, że jest rozkładem skrajnie asymetrycznym.



Rys. 3.1. Rozkład normalny oraz rozkład wysmukły i rozkład spłaszczony
Źródło: opracowanie własne.

Szczególnym przypadkiem wśród rozkładów jest tak zwany rozkład U – kształtny (siodłowy) powstały, z połączenia dwóch rozkładów asymetrycznych, który w pewnych przypadkach może być również rozkładem symetrycznym z osią symetrii przebiegającą przez punkt minimum. Na rys. 3.2 i 3.3 przedstawiono graficznie przykłady wymienionych powyżej rozkładów.

3.2. Charakterystyki liczbowe rozkładów empirycznych

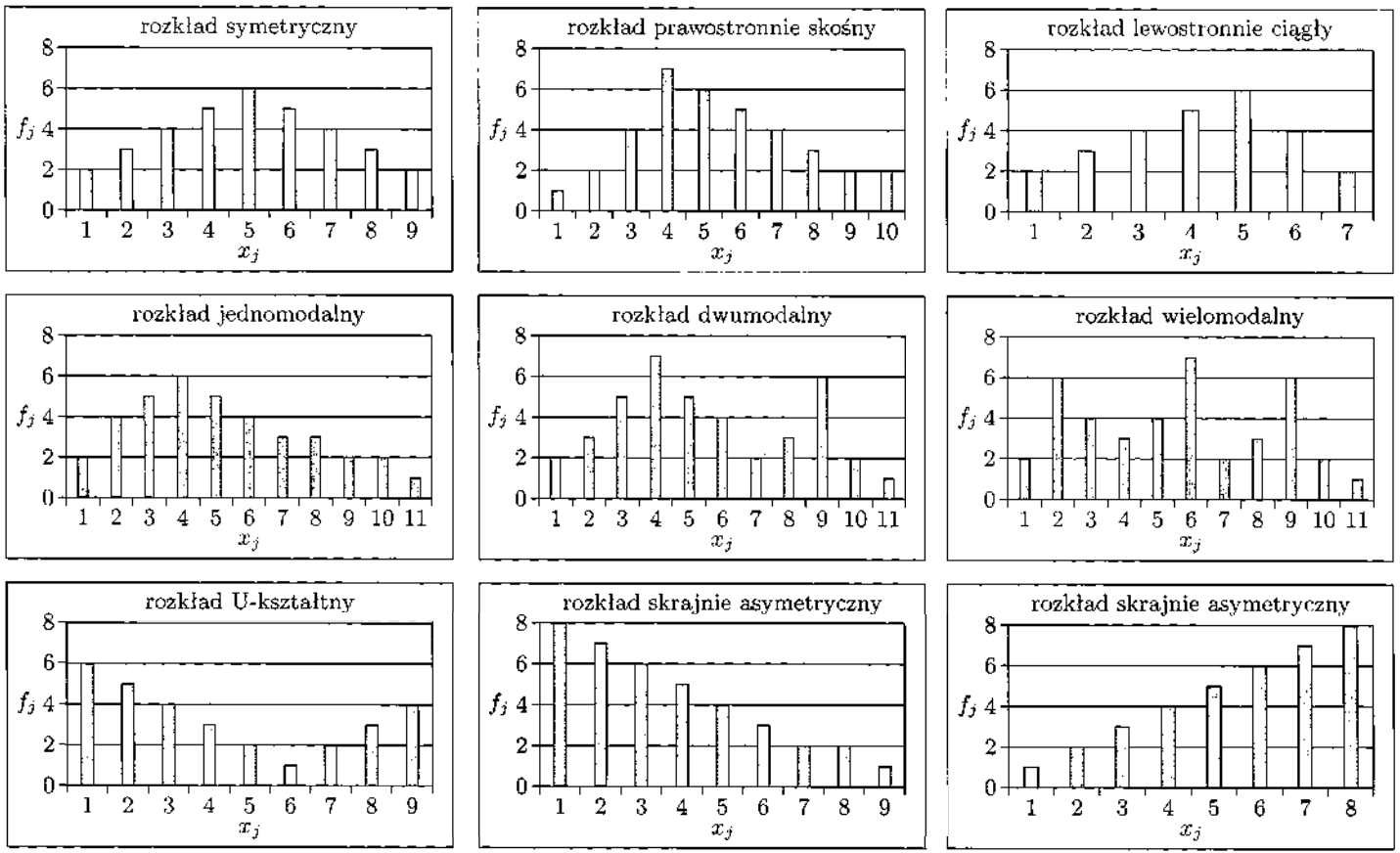
Wśród charakterystyk liczbowych wyróżnić można cztery zasadnicze grupy miar:

— miary położenia (średnie, przeciętne),

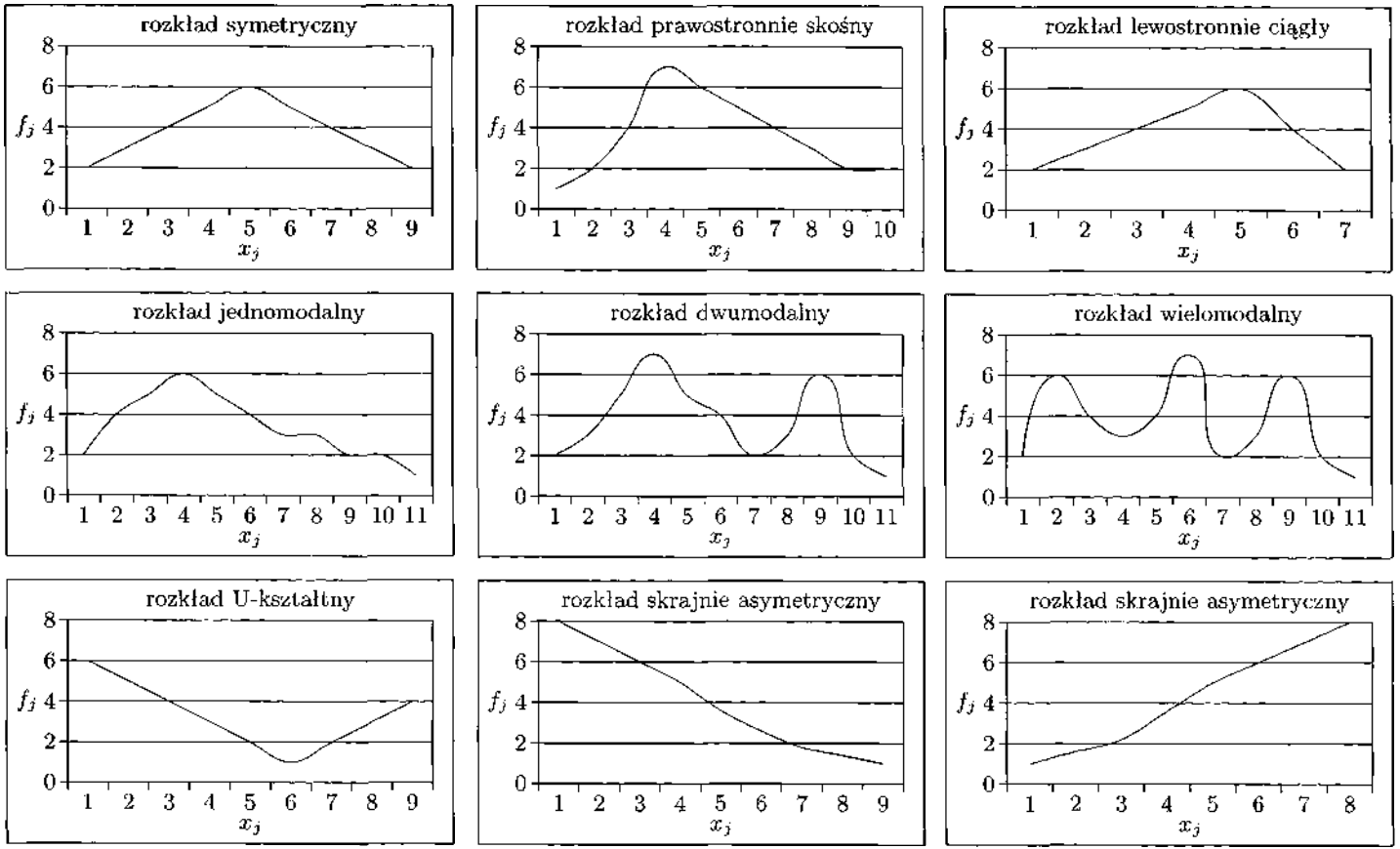
— miary zmienności (rozproszenia, dyspersji, zróżnicowania),

— miary asymetrii (skośności),

— miary koncentracji (kurtozy).



Rys. 3.2. Typy rozkładów empirycznych zmiennej skokowej
Źródło: opracowanie własne.



Rys. 3.3. Typy rozkładów zmiennej ciągłej
Źródło: opracowanie własne.

3.2.1. Miary położenia

Wśród miar położenia wyróżnia się tzw. średnie klasyczne i średnie pozycyjne. Do średnich klasycznych zalicza się **średnią arytmetyczną**, średnią harmoniczną i średnią geometryczną.

Najczęściej wykorzystywanymi średnimi pozycyjnymi są **mediana** (wartość środkowa) oraz modalna (dominanta, wartość najczęstsza). Obliczanie wartości przeciętnej badanej zmiennej ma sens tylko wówczas, gdy zbiorowość jest jednorodna. Wartość przeciętna charakteryzuje zbiorowość jako całość i informuje, jaki jest jej przeciętny poziom ze względu na badaną zmienną.

Jedną z najczęściej stosowanych średnich klasycznych jest średnia arytmetyczna, która jest ilorazem sumy poszczególnych wartości badanej zmiennej i liczby obserwacji. Sposób obliczania (wzór) średniej arytmetycznej zależy od rodzaju, wykorzystywanego w tym celu, szeregu statystycznego. Można wyróżnić tutaj trzy przypadki:

1. Jeżeli realizacje obserwowanej zmiennej X dane są w postaci szeregu szczegółowego (uporządkowanego lub nieuporządkowanego), to wówczas średnią arytmetyczną obliczamy według wzoru:

$$\overline{x} = \frac{1}{N}\sum_{i=1}^{N} x_i = \frac{x_1 + x_2 + \ldots + x_N}{N}, \tag{3.1}$$

gdzie:

$\overline{x}$ – średnia arytmetyczna zmiennej X,
N – liczebność jednostek ststystycznych badanej zbiorowości,
x_i – i-ta realizacja badanej zmiennej, przy czym $i = 1, 2, 3, \ldots, N$.

W celu ilustracji sposobu obliczania średniej arytmetycznej na podstawie szeregu szczegółowego posłużymy się danymi z przykładu 2.2, zestawionymi w postaci szeregu szczegółowego uporządkowanego (zob. tablica 2.4). Po podstawieniu wartości do wzoru (3.1) otrzymamy:

$$\overline{x} = \frac{1}{N}\sum_{i=1}^{100} x_i = \frac{38 + 44 + 45 + \ldots + 119 + 125 + 131}{100} = 76,1.$$

Z powyższego obliczenia wynika, że z usług badanej biblioteki korzysta średnio (przeciętnie) 76,1 osób dziennie.

2. Jeżeli realizacje obserwowanej zmiennej X dane są w postaci szeregu rozdzielczego punktowego, to wówczas do obliczenia średniej arytmetycznej

stosujemy wzór:

$$\overline{x} = \frac{1}{\sum_{j=1}^{k} f_j} \cdot \sum_{j=1}^{k} x_j f_j = \frac{\sum_{j=1}^{k} x_j f_j}{N} = \frac{x_1 f_1 + x_2 f_2 + \ldots + x_k f_k}{N}, \qquad (3.2)$$

gdzie: f_j jest liczebością, z jaką występowała j-ta wartość zmiennej X. Średnią arytmetyczną można również obliczyć wykorzystując częstości względne (v_j).

Możemy tego dokonać transformując wzór (3.2) do postaci:

$$\overline{x} = \sum_{j=1}^{k} x_j v_j = x_1 v_1 + x_2 v_2 + \ldots + x_k v_k, \qquad (3.3)$$

gdzie: v_j jest częstością względną występowania j-tej wartości zmiennej X.

Podobnie jak poprzednio, do zilustrowania sposobu obliczania średniej, wykorzystamy dane z przykładu 2.2, przedstawione tym razem w postaci szeregu rozdzielczego punktowego (zob. tablica 2.6).

W tablicy 3.1, pokazano obliczenia niezbędne do wyznaczenia średniej arytmetycznej według wzoru (3.2). Kolumny 1-3 oraz 5 pochodzą z tablicy 2.6. Podstawiając sumy kolumn 3 i 4 do wzoru (3.2) otrzymamy:

$$\overline{x} = \frac{1}{\sum_{j=1}^{58} f_j} \cdot \sum_{j=1}^{58} x_j f_j = \frac{7610}{100} = 76,1.$$

Identyczny rezultat otrzymamy, jeżeli do obliczenia średniej zostanie użyty wzór (3.3). Wówczas suma kolumny 5 jest poszukiwaną średnią arytmetyczną. Mamy mianowicie:

$$\overline{x} = \sum_{j=1}^{58} x_j v_j = 76,1.$$

Tab. 3.1. Obliczanie średniej arytmetycznej z szeregu rozdzielczego punktowego

j	Liczba korzyst. z bibl. x_j	f_j	$x_j f_j$	v_j	$v_j f_j$	j	Liczba korzyst. z bibl. x_j	f_j	$x_j f_j$	v_j	$v_j f_j$
1	**2**	**3**	**4**	**5**	**6**	**1**	**2**	**3**	**4**	**5**	**6**
1	38	1	38	0,01	0,38	30	77	4	308	0,04	3,08
2	44	1	44	0,01	0,44	31	78	1	78	0,01	0,78
3	45	1	45	0,01	0,45	32	79	4	316	0,04	3,16
4	46	2	92	0,02	0,92	33	80	2	160	0,02	1,60
5	47	1	47	0,01	0,47	34	82	3	246	0,03	2,46
6	50	1	50	0,01	0,50	35	83	1	83	0,01	0,83
7	52	3	156	0,03	1,56	36	84	2	168	0,02	1,68
8	53	2	106	0,02	1,06	37	85	3	255	0,03	2,55
9	55	1	55	0,01	0,55	38	86	1	86	0,01	0,86
10	56	2	112	0,02	1,12	39	87	1	87	0,01	0,87
11	57	2	114	0,02	1,14	40	88	3	264	0,03	2,64
12	58	1	58	0,01	0,58	41	89	1	89	0,01	0,89
13	60	2	120	0,02	1,20	42	91	1	91	0,01	0,91
14	61	3	183	0,03	1,83	43	93	1	93	0,01	0,93
15	62	2	124	0,02	1,24	44	94	1	94	0,01	0,94
16	63	3	189	0,03	1,89	45	95	2	190	0,02	1,90
17	64	1	64	0,01	0,64	46	96	1	96	0,01	0,96
18	65	3	195	0,03	1,95	47	97	1	97	0,01	0,97
19	66	1	66	0,01	0,66	48	98	2	196	0,02	1,96
20	67	2	134	0,02	1,34	49	101	2	202	0,02	2,02
21	68	1	68	0,01	0,68	50	102	2	204	0,02	2,04
22	69	1	69	0,01	0,69	51	103	1	103	0,01	1,03
23	70	2	14	0 0,0	21,40	52	104	1	104	0,01	1,04
24	71	1	71	0,01	0,71	53	110	1	110	0,01	1,10
25	72	6	432	0,06	4,32	54	111	1	111	0,01	1,11
26	73	4	292	0,04	2,92	55	115	1	115	0,01	1,15
27	74	1	74	0,01	0,74	56	119	1	119	0,01	1,19
28	75	1	75	0,01	0,75	57	125	1	125	0,01	1,25
29	76	1	76	0,01	0,76	58	131	1	131	0,01	1,31
							SUMA	100	7610	1,00	76,10

Źródło: obliczenia własne.

Można zatem stwierdzić, że wartość średniej arytmetycznej obliczonej na postawie szeregu szczegółowego jest identyczna jak wartość średniej arytmetycznej wyznaczona z szeregu rozdzielczego punktowego.

3. Jeżeli realizacje obserwowanej zmiennej X dane są w postaci szeregu rozdzielczego przedziałowego, to wówczas średnią arytmetyczną obliczamy

stosując wzór:

$$\overline{x} = \frac{1}{\sum_{j=1}^{k} f_j} \cdot \sum_{j=1}^{k} x_j^{\circ} f_j = \frac{\sum_{j=1}^{k} x_j^{\circ} f_j}{N} = \frac{x_1^{\circ} f_1 + x_2^{\circ} f_2 + \ldots + x_k^{\circ} f_k}{N}, \qquad (3.4)$$

gdzie: x_j° jest środkiem j-tego przedziału klasowego obliczonego zgodnie z regułą

$$x_j^{\circ} = \frac{x_{dj} + x_{gj}}{2}, \qquad (3.5)$$

gdzie: $x_{d.j}$ i $x_{g.j}$ są dolną i górną granicą j-tego przedziału klasowego.

Podobnie jak w przypadku szeregu rozdzielczego punktowego wzór (3.4) można stransformować do postaci:

$$\overline{x} = \sum_{j=1}^{k} x_j^{\circ} v_j = x_1^{\circ} v_1 + x_2^{\circ} v_2 + \ldots + x_k^{\circ} v_k. \qquad (3.6)$$

W tablicy 3.2, przedstawiono obliczenia niezbędne do wyznaczenia średniej arytmetycznej według wzoru (3.4). Kolumny 1-3 oraz kolumna 6, pochodzą z tablicy 2.9. W kolumnie 4 zestawiono środki przedziałów obliczone zgodnie z regułą (3.5). Podobnie jak we wcześniejszym przykładzie sumy kolumn 3 i 5 podstawiamy do wzoru (3.4), dzięki czemu otrzymujemy:

$$\overline{x} = \frac{1}{\sum_{j=1}^{8} f_j} \cdot \sum_{j=1}^{8} x_j^{\circ} f_j = \frac{7572}{100} = 75,72.$$

Jeżeli do obliczenia średniej arytmetycznej wykorzystamy wzór (3.6), to wówczas wystarczy podsumować ostatnią siódmą kolumnę w tablicy 3.1. Otrzymamy wówczas:

$$\overline{x} = \sum_{j=1}^{8} x_j^{\circ} v_j = 75,72.$$

Tab. 3.2. Obliczanie średniej arytmetycznej z szeregu rozdzielczego punktowego

j	$(x_{d,j}; x_{g,j}]$	fj	Środki przedziałów x_j°	$x_j^\circ f_j$	v_j	$x_j^\circ v_j$
1	2	3	4	5	6	7
1	(36;48]	6	42	252	0,06	2,52
2	(48;60]	14	54	756	0,14	7,56
3	(60;72]	26	66	1716	0,26	17,16
4	(72;84]	24	78	1872	0,24	18,72
5	(84;96]	15	90	1350	0,15	13,50
6	(96;108]	9	102	918	0,09	9,18
7	(108;120]	4	114	456	0,04	4,56
8	(120;132]	2	126	252	0,02	2,52
SUMA		100		7572	1,00	75,72

Źródło: obliczenia własne.

Różnica pomiędzy wartościami średniej obliczonej na podstawie szeregu szczegółowego i rozdzielczego punktowego, a szeregu rozdzielczego przedziałowego jest efektem wspomnianego już wcześniej błędu grupowania. W omawianym przypadku średnia arytmetyczna obliczona na podstawie szeregu rozdzielczego przedziałowego jest obciążona błędem grupowania wynoszącym $-0,38$ $(75,72 - 76,1 = -0,38)$.

Średnia arytmetyczna posiada kilka ważnych własności:

1. Wartość średniej arytmetycznej zawsze zawiera się w przedziale pomiędzy x_{min} i x_{max}, co zapiszemy:

$$x_{min} \leqslant \overline{x} \leqslant x_{max} \tag{3.7}$$

2. Suma realizacji (wszystkich wartości) zmiennej X równa się iloczynowi średniej arytmetycznej i liczebności N, co można ująć zapisem:

$$\overline{x}N = \sum_{i=1}^{N} x_i \tag{3.8}$$

w przypadku szeregu szczegółowego, lub

$$\sum_{j=1}^{k} f_j = \sum_{j=1}^{k} x_j f_j \tag{3.9}$$

w przypadku szeregu rozdzielczego punktowego, lub

$$\sum_{j=1}^{k} f_j = \sum_{i=1}^{k} x_j^\circ f_j, \tag{3.10}$$

gdy rozpatrujemy szereg rozdzielczy przedziałowy.

3. Suma odchyleń realizacji zmiennej X od średniej arytmetycznej jest zawsze równa zero. Powyższe twierdzenie można zapisać:

$$\sum_{i=1}^{N}(x_i - \overline{x}) = 0 \tag{3.11}$$

w przypadku szeregu szczegółowego, lub

$$\sum_{j=1}^{k} f_j(x_j - \overline{x}) = 0 \tag{3.12}$$

w przypadku szeregu rozdzielczego punktowego, lub

$$\sum_{j=1}^{k} f_j(x_j^{\circ} - \overline{x}) = 0, \tag{3.13}$$

gdy analizie podlega szereg rozdzielczy przedziałowy.

4. Suma kwadratów odchyleń realizacji zmiennej X od średniej arytmetycznej jest zawsze mniejsza lub równa niż suma kwadratów odchyleń realizacji zmiennej X od dowolnej liczby rzeczywistej a. Podobnie jak w przypadku własności 1 – 3, własność 4 (dla szeregu szegółowego i rozdzielczego) można zapisać symbolicznie w następujący sposób:

$$\sum_{i=1}^{N}(x_i - \overline{x})^2 \leqslant \sum_{i=1}^{N}(x_i - a)^2 \tag{3.14}$$

lub

$$\sum_{j=1}^{k} f_j(x_j - \overline{x}) \leqslant \sum_{j=1}^{k} f_j(x_j - a)^2 \tag{3.15}$$

lub

$$\sum_{j=1}^{k} f_j(x_j^{\circ} - \overline{x})^2 \leqslant \sum_{j=1}^{k} f_j(x_j^{\circ} - a)^2 \tag{3.16}$$

Średnia harmoniczna ($\overline{x}_h$) jest odwrotnością średniej arytmetycznej, obliczonej z odwrotności wartości cechy. Jeżeli średnią harmoniczną obliczamy z szeregu szczegółowego, to wówczas należy zastosować wzór:

$$\overline{x}_h = \frac{N}{\sum_{i=1}^{N}\frac{1}{x_i}}. \tag{3.17}$$

Natomiast, gdy do obliczeń wykorzystujemy szereg rozdzielczy punktowy, to wówczas średnią harmoniczną obliczamy stosując wzór:

$$\overline{x}_h = \frac{\sum_{j=1}^{k} f_j}{\sum_{j=1}^{k} \frac{f_j}{x_j}} = \frac{N}{\sum_{j=1}^{k} \frac{f_j}{x_j}} = \frac{1}{\sum_{j=1}^{k} \frac{v_j}{x_j}}. \tag{3.18}$$

W przypadku szeregu rozdzielczego przedziałowego powyższy wzór zastąpimy:

$$\overline{x}_h = \frac{\sum_{j=1}^{k} f_j}{\sum_{j=1}^{k} \frac{f_j}{x_j}} = \frac{N}{\sum_{j=1}^{k} \frac{f_j}{x_j^\circ}} = \frac{1}{\sum_{j=1}^{k} \frac{v_j}{x_j^\circ}}. \tag{3.19}$$

Przykład 3.1. Zmierzono powierzchnię (w m^2), przypadającą na osobę w pięciu gospodarstwach domowych w standardowych mieszkaniach o powierzchni 64 m^2. Wyniki badania zestawiono w tablicy 3.3.

Tab. 3.3. Dane liczbowe do przykładu 3.1

i	Metraż [m^2]	Liczba osób [os.]	[m^2/os.]
1	64	4	16
2	64	1	64
3	64	2	32
4	64	4	16
5	64	8	8
suma	320	19	136

Źródło: badania własne.

Obliczyć przeciętną liczbę metrów kwadratowych przypadających na osobę. Jeżeli podczas obliczeń zastosujemy średnią arytmetyczną, to wówczas otrzymamy:

$$\overline{x} = \frac{136}{5} = 27,2 \qquad [\mathrm{m}^2/\mathrm{os}].$$

Wynik ten jest jednak nieprawidłowy, gdyż jeżeli łączny metraż badanych mieszkań wynosi 320 m^2, a łączna liczba badanych osób wynosi 19, to średnia ilość metrów kwadratowych przypadających na osobę wyniesie:

$$\frac{320}{19} = 16,84211 \qquad [\mathrm{m}^2/\mathrm{os}].$$

Zauważmy, że wynik ten otrzymamy, obliczając średnią harmoniczną:

$$\overline{x}_h = \frac{5}{\frac{1}{16} + \frac{1}{64} + \frac{1}{32} + \frac{1}{16} + \frac{1}{8}} = 16,84211 \qquad [\mathrm{m}^2/\mathrm{os}].$$

Zastosowanie średniej arytmetycznej do obliczenia przeciętnej metrów kwadratowych przypadających na 1 osobę, jest prawidłowe tylko wówczas, gdy liczba osób w badanych gospodrstwach jest stała. W pozostałych przypadkach do obliczeń należy stować średnią harmoniczną.#

Trzecią spośród średnich klasycznych – średnią geometryczną – omówimy w rozdziale piątym tego podręcznika.

Mediana (wartość środkowa) zaliczana jest do tzw. kwantyli (fraktyli). Kwantyl stopnia q $(0 < q < 1)$ to taka liczba x_q, która dzieli N-elementowy zbiór realizacji badanej zmiennej X na N_q-elementowy podzbiór wartości $x_i \leqslant x_q$ oraz na $N(1-q)$-elementowy podzbiór wartości $x_i > x_q$ *). Wśród kwantyli wyróżnia się:

— kwartyle (dzielące zbiorowość na cztery części); $q = 0,25; 0,5; 0,75$,

— kwintyle (na pięć części); $q = 0,2; 0,4; 0,6; 0,8$,

— decyle (na dziesięć części); $q = 0,1; 0,2; 0,3; \ldots; 0,9$,

— centyle (na sto części); $q = 0,01; 0,02; 0,03; \ldots; 0,99$.

Mediana (x_{me}) jest drugim kwartylem. Zgodnie z powyższą definicją będzie ona taką wartością x_{me}, która podzieli badaną zbiorowość w ten sposób, że połowa realizacji zmiennej X będzie posiadać wartości nie przekraczające x_{me}, a druga połowa będzie miała wartości większe niż x_{me}.

Sposób wyznaczania wartości mediany zależy od rodzaju szeregu statystycznego. Jeżeli, jest to szereg statystyczny szczegółowy, to w pierwszej kolejności należy go transformować do postaci szeregu szczegółowego uporządkowanego (pozycyjnego). Jeżeli liczba elemetów w szeregu (N) jest nieparzysta, to wówczas mediana jest równa wartości znajdującej się dokładnie w środku tego szeregu. W przypadku, gdy N jest liczbą parzystą mediana jest równa średniej arytmetycznej wyznaczonej z dwóch centralnie położonych wartości w szeregu pozycyjnym. Sposób wyznaczania mediany można opisać następującym wzorem:

$$x_{me} = \begin{cases} x_{\left(\frac{N+1}{2}\right)}, & \text{gdy } N \text{ nieparzyste,} \\ \frac{1}{2}\left[x_{\left(\frac{N}{2}\right)} + x_{\left(\frac{N}{2}+1\right)}\right], & \text{gdy } N \text{ parzyste.} \end{cases} \tag{3.20}$$

*) Zob. A. Iwasiewicz, Z. Paszek, *Statystyka z elementami statystycznych metod ...*, op. cit.

Indeksy we wzorze (3.20) zapisane w nawiasach okrągłych, dotyczą szeregu szczegółowego uporządkowanego.

Do wyznaczenia mediany wykorzystamy dane z przykładu 2.2, zestawione w postaci szeregu pozycyjnego (zob. tablica 2.4). Ponieważ $N = 100$, zatem należy wykorzystać drugi wariant wzoru (3.20) i wówczas po podstawieniu wartości otrzymamy:

$$x_{me} = \frac{x_{(50)} + x_{(51)}}{2} = \frac{73 + 74}{2} = 73,5.$$

Z powyższych obliczeń wynika, że 50% stanowią przypadki, gdy dzienna liczba osób korzystających z usług biblioteki była mniejsza od 73,5, a 50% przypadki, gdy liczba osób była większa niż 73,5.

Nieco bardziej skomplikowany jest sposób wyznaczania mediany na podstawie szeregu rozdzielczego przedziałowego. W pierwszym kroku postępowania należy utworzyć szereg skumulowany (zob. tablica 3.4).

Tab. 3.4. Wyznaczanie przedziału mediany i modalnej

j	$(x_{d.i}; x_{g.i}]$	f_j	Liczebności skumulowane $f_{j.skum}$	Częstości względne skumulowane $v_{j.skum}$
1	**2**	**3**	**4**	**5**
1	(36;48]	6	6	0,06
2	(48;60]	14	20	0,20
3	(60;72]	26	46	0,46
4	(72;84]	24	70	0,70
5	(84;96]	15	85	0,85
6	(96;108]	9	94	0,94
7	(108;120]	4	98	0,98
8	(120;132]	2	100	1,00
SUMA		100	XXX	XXX

przedział, w którym znajduje się modalna

przedział, w którym znajduje się mediana

Źródło: obliczenia własne.

W szeregu tym należy wskazać przedział mediany. Przedział, w którym znajduje się mediana jest to pierwszy przedział klasowy, dla którego suma liczebności skumulowanych jest większa od $\frac{N}{2}$. W omawianym przypadku jest to przedział o końcach (72; 84].

W drugim kroku postępowania wartość mediany wyznacza się stosując wzór:

$$x_{me} = x_{d,r} + \frac{\left(\frac{N}{2} - \sum_{j=1}^{r-1} f_j\right) \cdot l}{f_r}, \tag{3.21}$$

gdzie:

$x_{d,r}$ – dolna granica przedziału, w którym znajduje się mediana,

f_r – liczebność przedziału, w którym znajduje się mediana,

l – długość przedziału, w którym znajduje się mediana,

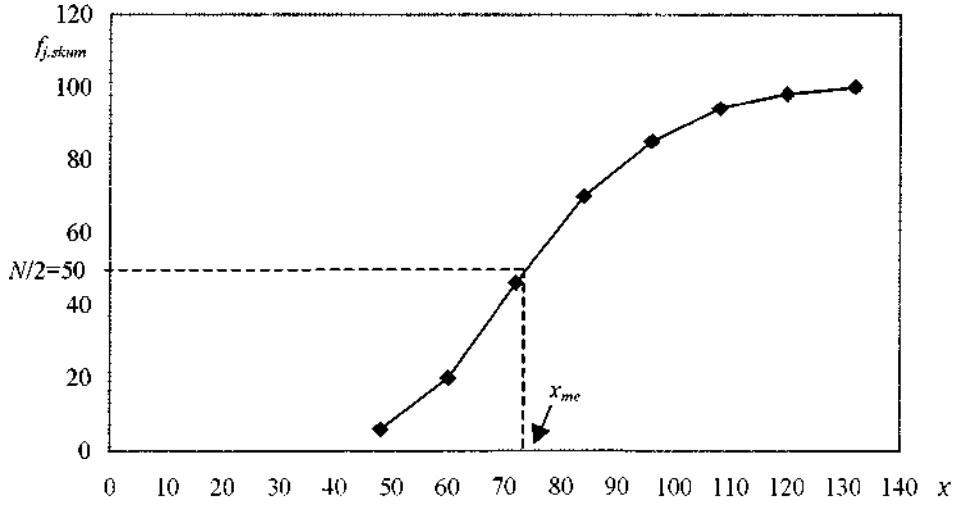
$\sum_{j=1}^{r-1} f_j$ – suma liczebności przedziałów poprzedzających przedział, w którym znajduje się mediana.

Uwzględniając dane zawarte w tablicy 3.4, mediana wynosi:

$$x_{me} = x_{d,r} + \frac{\left(\frac{N}{2} - \sum_{j=1}^{r-1} f_j\right) \cdot l}{f_r} = 72 + \frac{\left(\frac{100}{2} - 46\right) \cdot 12}{24} = 74.$$

Różnica pomiędzy medianą obliczoną na podstawie szeregu rozdzielczego przedziałowego i szeregu szczegółowego uporządkowanego jest niewielka i wynosi zaledwie 0,5.

Graficzna metoda wyznaczania mediany została przedstawiona na rysunku 3.4. W tym celu w układzie współrzędnych wykreślamy krzywą liczności (lub częstości) skumulowanych. Następnie z punktu $\frac{N}{2}$ prowadzimy prostą równoległą do osi odciętych, punkt przecięcia prostej z krzywą liczebności skumulowanych rzutujemy na oś odciętych. Otrzymana w ten sposób odcięta jest równa medianie.



Rys. 3.4. Graficzna metoda wyznaczania mediany
Źródło: opracowanie własne.

Kolejną ważną miarą położenia jest **modalna**. Do oznaczenia modalnej użyjemy symbolu x_{mo}. **Modalna** jest wartością najczęstszą w szeregu staty-

stycznym. W przypadku szeregu szczegółowego lub przedziałowego punktowego wyznaczenie modalnej sprowadza się do określenia, która wartość wystąpiła najczęściej, to znaczy, dla której wartości x_j liczebność f_j jest największa.

Najwięcej, bo aż 6 razy zdarzyła się sytuacja, że bibliotekę odwiedziły w ciągu dnia 72 osoby. A zatem wartość modalnej, wyznaczona w oparciu o szereg (pozycyjny lub rozdzielczy punktowy), to $x_{mo} = 72$.

Nieco inaczej postępujemy, gdy chcemy wyznaczyć modalną na podstawie szeregu rozdzielczego przedziałowego.

W pierwszym kroku określamy przedział, w którym znajdować się będzie modalna. Jest to przedział klasowy o największej liczbie obserwacji empirycznych. W rozważanym przykładzie (zob. tablica 3.4), to przedział (60; 72].

Następnie wyznaczamy wartość x_{mo}:

$$x_{mo} = x_{d,r} + \frac{(f_r - f_{r-1}) \cdot l}{(f_r - f_{r-1}) + (f_r - f_{r+1})}, \tag{3.22}$$

gdzie:

$x_{d,r}$ – dolna granica przedziału, w którym znajduje się modalna,
f_r – liczebność przedziału, w którym znajduje się modalna,
f_{r-1} – liczebność przedziału poprzedzającego przedział, w którym znajduje się modalna,
f_{r+1} – liczebność przedziału następującego po przedziale, w którym znajduje się modalna,
l – długość przedziału, w którym znajduje się modalna.

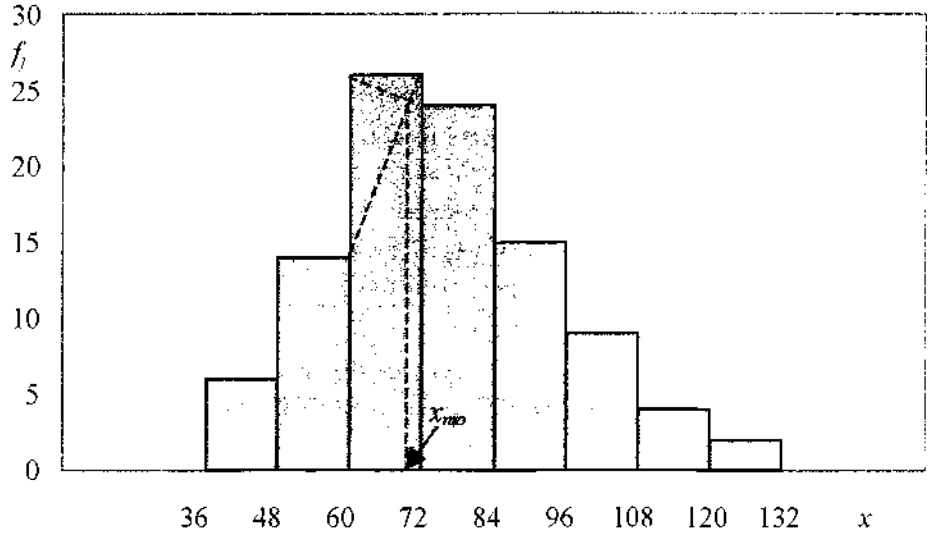
Po podstawieniu wartości liczbowych do wzoru 3.22 otrzymamy:

$$x_{mo} = x_{d,r} + \frac{(f_r - f_{r-1}) \cdot l}{(f_r - f_{r-1}) + (f_r - f_{r+1})} = 60 + \frac{(26 - 14) \cdot 12}{(26 - 14) + (26 - 24)} = 70,29.$$

W tym przypadku, uzyskany wynik jest mniejszy od wyniku uzyskanego na podstawie szeregu szczegółowego lub szeregu rozdzielczego punktowego.

Istnieje również graficzny sposób wyzanczenia wartości modalnej. Sprowadza się on do wykreślenia histogramu liczebności dla przedziału, w którym znajduje się modalna oraz dwóch przedziałów sąsiednich (dopuszczalne jest również wykreślanie histogramu, dla wszystkich przedziałów klasowych). Z górnej podstawy najwyższego prostokąta prowadzi się dwie linie pomocnicze, łączące najbliżej położone punkty górnych podstaw sąsiednich prostokątów. W punkcie przecięcia się tych linii wykreślamy prostą prostopadłą do osi odciętych. Punkt, w którym prosta ta przetnie oś odciętych jest przybliżoną wartością modalnej.

Na rys. 3.5 pokazano sposób graficznego wyznaczania wartość modalnej w oparciu o dane z analizowanego przykładu.



Rys. 3.5. Graficzna metoda wyznaczania modalnej
Źródło: opracowanie własne.

3.2.2. Wariancja i odchylenie standardowe

Podstawowymi miarami zmienności są wariancja (s_x^2) oraz odchylenie standardowe (s_x). Indeks (x) znajdujący się w dolnej części symbolu oznacza zmienną charakteryzującą badaną zbiorowość.
Poniższe wzory (3.23), (3.24) i (3.25) przedstawiają kolejno sposób wyznaczania wariancji w przypadku szeregu szczegółowego, rozdzielczego punktowego i rozdzielczego przedziałowego:

$$s_x^2 = \frac{1}{N}\sum_{i=1}^{N}(x_i - \overline{x})^2, \tag{3.23}$$

$$s_x^2 = \frac{1}{\sum_{j=1}^{k} f_j} \cdot \sum_{j=1}^{k} f_j(x_j - \overline{x})^2, \tag{3.24}$$

$$s_x^2 = \frac{1}{\sum_{j=1}^{k} f_j} \cdot \sum_{j=1}^{k} f_j(x_j^{\circ} - \overline{x})^2. \tag{3.25}$$

Podobnie jak w przypadku średniej arytmetycznej wzory (3.24) i (3.25) można przekształcić do postaci:

$$s_x^2 = \sum_{j=1}^{k} v_j (x_j - \overline{x})^2, \tag{3.26}$$

$$s_x^2 = \sum_{j=1}^{k} v_j (x_j^\circ - \overline{x})^2. \tag{3.27}$$

Odchylenie standardowe uzyskuje się biorąc dodatni pierwiastek kwadratowy wariancji, co zapiszemy:

$$s_x = \sqrt{s_x^2}. \tag{3.28}$$

Wartość odchylenia standardowego informuje badacza o ile przeciętnie (in plus lub in minus) odchylają się realizacje badanej zmiennej od warości średniej. Natomiast wariancja nie posiada interpretacji. Podczas obliczania wariancji, a następnie odchylenia standardowego można skorzystać również z zależności głoszącej, że wariancja to średni kwadrat pomniejszony o kwadrat średniej, co zapiszemy:

$$s_x^2 = \overline{x^2} - (\overline{x})^2. \tag{3.29}$$

Technikę obliczania wariancji zilustrujemy wykorzystując – tak jak uprzednio – materiał statystyczny z przykładu 2.2. Obliczymy kolejno wariancję i odchylenie standardowe z szeregu rozdzielczego punktowego i przedziałowego*). Kolejno zastosujemy wzory (3.24) i (3.25).

W tablicy 3.5 zamieszczono niezbędne obliczenia do wyznaczenia wariancji i odchylenia standardowego na podstawie szeregu rozdzielczego punktowego.

Przypomnijmy, że wartość wyznaczonej – na podstawie tego szeregu – średniej arytmetycznej wyniosła

$$\overline{x} = \frac{1}{\sum_{j=1}^{24} f_j} \cdot \sum_{j=1}^{24} f_j = \frac{7610}{100} = 76,1.$$

*)Czytelnikowi zalecamy obliczenie wariancji i odchylenia standardowego na podstawie szeregu szczegółowego pozycyjnego.

Tab. 3.5. Obliczanie wariancji z szeregu rozdzielczego punktowego

j	Liczba odwiedz. bibliotekę (x_j)	f_j	$x_i - \overline{x}$	$(x_i - \overline{x})^2$	$f_j(x_i - \overline{x})^2$	$v_j(x_i - \overline{x})^2$
1	**2**	**3**	**4**	**5**	**6**	**7**
1	38	1	-38,10	1451,61	1451,61	14,5161
2	44	1	-32,10	1030,41	1030,41	10,3041
3	45	1	-31,10	967,21	967,21	9,6721
4	46	2	-30,10	906,01	1812,02	18,1202
5	47	1	-29,10	846,81	846,81	8,4681
6	50	1	-26,10	681,21	681,21	6,8121
7	52	3	-24,10	580,81	1742,43	17,4243
8	53	2	-23,10	533,61	1067,22	10,6722
9	55	1	-21,10	445,21	445,21	4,4521
10	56	2	-20,10	404,01	808,02	8,0802
11	57	2	-19,10	364,81	729,62	7,2962
12	58	1	-18,10	327,61	327,61	3,2761
13	60	2	-16,10	259,21	518,42	5,1842
14	61	3	-15,10	228,01	684,03	6,8403
15	62	2	-14,10	198,81	397,62	3,9762
16	63	3	-13,10	171,61	514,83	5,1483
17	64	1	-12,10	146,41	146,41	1,4641
18	65	3	-11,10	123,21	369,63	3,6963
19	66	1	-10,10	102,01	102,01	1,0201
20	67	2	-9,10	82,81	165,62	1,6562
21	68	1	-8,10	65,61	65,61	0,6561
22	69	1	-7,10	50,41	50,41	0,5041
23	70	2	-6,10	37,21	74,42	0,7442
24	71	1	-5,10	26,01	26,01	0,2601
25	72	6	-4,10	16,81	100,86	1,0086
26	73	4	-3,10	9,61	38,44	0,3844
27	74	1	-2,10	4,41	4,41	0,0441
28	75	1	-1,10	1,21	1,21	0,0121
29	76	1	-0,10	0,01	0,01	0,0001
30	77	4	0,90	0,81	3,24	0,0324
31	78	1	1,90	3,61	3,61	0,0361
32	79	4	2,90	8,41	33,64	0,3364
33	80	2	3,90	15,21	30,42	0,3042
34	82	3	5,90	34,81	104,43	1,0443
35	83	1	6,90	47,61	47,61	0,4761
36	84	2	7,90	62,41	124,82	1,2482
37	85	3	8,90	79,21	237,63	2,3763
38	86	1	9,90	98,01	98,01	0,9801
39	87	1	10,90	118,81	118,81	1,1881

Tab. 3.5. Obliczanie wariancji z szeregu rozdzielczego punktowego cd.

j	Liczba odwiedz. bibliotekę (x_j)	f_j	$x_i - \overline{x}$	$(x_i - \overline{x})^2$	$f_j(x_i - \overline{x})^2$	$v_j(x_i - \overline{x})^2$
1	**2**	**3**	**4**	**5**	**6**	**7**
40	88	3	11,90	141,61	424,83	4,2483
41	89	1	12,90	166,41	166,41	1,6641
42	91	1	14,90	222,01	222,01	2,2201
43	93	1	16,90	285,61	285,61	2,8561
44	94	1	17,90	320,41	320,41	3,2041
45	95	2	18,90	357,21	714,42	7,1442
46	96	1	19,90	396,01	396,01	3,9601
47	97	1	20,90	436,81	436,81	4,3681
48	98	2	21,90	479,61	959,22	9,5922
49	101	2	24,90	620,01	1240,02	12,4002
50	102	2	25,90	670,81	1341,62	13,4162
51	103	1	26,90	723,61	723,61	7,2361
52	104	1	27,90	778,41	778,41	7,7841
53	110	1	33,90	1149,21	1149,21	11,4921
54	111	1	34,90	1218,01	1218,01	12,1801
55	115	1	38,90	1513,21	1513,21	15,1321
56	119	1	42,90	1840,41	1840,41	18,4041
57	125	1	48,90	2391,21	2391,21	23,9121
58	131	1	54,90	3014,01	3014,01	30,1401
Suma		100			35107,00	351,0700

Źródło: obliczenia własne.

Trzy ostatnie kolumny przedstawiają kolejne etapy postępowania, podczas obliczania wartości sumy kwadratów odchyleń (wartość w prawym dolnym rogu). Mając tę wartość można obliczyć wariancję:

$$s_x^2 = \frac{1}{\sum_{j=1}^{k} f_j} \cdot \sum_{j=1}^{k} f_j (x_j - \overline{x})^2 = \frac{35107}{100} = 351,07,$$

a następnie odchylenie standardowe:

$$s_x = \sqrt{S_x^2} = \sqrt{351,07} = 18,74.$$

Tablica 3.6 prezentuje sposób postępowania przy wyznaczaniu wariancji na podstawie szeregu rozdzielczego przedziałowego. Średnia arytmetyczna obliczona na podstawie tego szeregu szczegółowego*) wyniosła $\overline{x} = 76,1$.

*)W sytuacji, gdy nie jest dostępna dokładna wartość średniej arytmetycznej, podczas obliczania wariancji należy posłużyć się wartością średniej arytmetycznej obliczonej na podstawie szeregu rozdzielczego przedziałowego.

Tab. 3.6. Obliczanie wariancji z szeregu rozdzielczego przedziałowego

j	$(x_{d,j}; x_{g,j}]$	f_j	Środki przedziałów x_j°	$x_j - \overline{x}$	$(x_j^\circ - \overline{x})^2$	$f_j(x_j^\circ - \overline{x})^2$	$v_j(x_j^\circ - \overline{x})^2$
1	**2**	**3**	**4**	**5**	**6**	**7**	**8**
1	(36;48]	6	42	-34,1000	1162,8100	6976,8600	69,7686
2	(48;60]	14	54	-22,1000	488,4100	6837,7400	68,3774
3	(60;72]	26	66	-10,1000	102,0100	2652,2600	26,5226
4	(72;84]	24	78	1,9000	3,6100	86,6400	0,8664
5	(84;96]	15	90	13,9000	193,2100	2898,1500	28,9815
6	(96;108]	9	102	25,9000	670,8100	6037,2900	60,3729
7	(108;120]	4	114	37,9000	1436,4100	5745,6400	57,4564
8	(120;132]	2	126	49,9000	2490,0100	4980,0200	49,8002
SUMA		100	XXX	XXX	XXX	36214,6000	362,1460

Źródło: obliczenia własne.

Po podzieleniu wartości sumy kwadratów odchyleń (prawy dolny róg) przez liczebność $N = 100$, otrzymamy wariancję równą:

$$s_x^2 = \frac{1}{\sum_{j=1}^{k} f_j} \cdot \sum_{j=1}^{k} f_j(x_j - \overline{x})^2 = \frac{36214,6000}{100} = 362,1460$$

oraz odchylenie standardowe:

$$s_x = \sqrt{s_x^2} = \sqrt{362,1460} = 19,03.$$

Róznica pomiędzy odchyleniem standardowym obliczonym na podstawie tych dwóch typów szeregów jest niewielka i wynosi $19,03 - 18,74$ czyli 0,29.

3.2.3. Inne miary zmienności

Oprócz wariancji i odchylenia standardowego, do miar zmienności zaliczyć można również współczynnik zmienności i odchylenie przeciętne.
Zarówno wariancja, jak i odchylenie standardowe są wielkościami mianowanymi, a ich miano zależy od jednostek miary obserwacji empirycznych. Uniemożliwia to bezpośrednie stosowanie tych miar w celu porównywania dwóch lub więcej zbiorowości pod względem stopnia zmienności. W takich sytuacjach zaleca się stosowanie współczynnika zmienności postaci:

$$v_x = \frac{s_x}{\overline{x}}. \tag{3.30}$$

Jest on stosunkiem (ilorazem) odchylenia standardowego zmiennej do jej wartości średniej.

W omawianym przykładzie wartość tego współczynnika wynosi:

$$v_x = \frac{s_x}{\overline{x}} = \frac{19,03}{76,10} = 0,25,$$

w przypadku szeregu szczegółowego i rozdzielczego punktowego, lub

$$v_x = \frac{s_x}{\overline{x}} = \frac{18,74}{76,10} = 0,2462,$$

dla szeregu rozdzielczego przedziałowego.#

Odchylenie przeciętne (d_x) jest to średnia arytmetyczna z bezwzględnych wartości odchyleń realizacji zmiennej X od średniej arytmetycznej. Poniższe wzory (3.31), (3.32), (3.33), przedstawiają kolejno sposób obliczania odchylenia przeciętnego na podstawie szeregu szczegółowego, szeregu rozdzielczego punktowego i szeregu rozdzielczego przedziałowego.

$$d_x = \frac{1}{N}\sum_{i=1}^{n}|x_i - \overline{x}|, \tag{3.31}$$

$$d_x = \frac{1}{\sum_{j=1}^{k} f_j} \cdot \sum_{j=1}^{k} f_j|x_j - \overline{x}|, \tag{3.32}$$

$$d_x = \frac{1}{\sum_{j=1}^{k} f_j} \cdot \sum_{j=1}^{k} f_j|x_j^\circ - \overline{x}|. \tag{3.33}$$

Podobnie jak w przypadku wariancji, wzory (3.32) i (3.33), można zapisać:

$$d_x = \sum_{j=1}^{k} v_j|x_j - \overline{x}|, \tag{3.34}$$

$$d_x = \sum_{j=1}^{k} v_j|x_j^\circ - \overline{x}|. \tag{3.35}$$

W celu zilustrowania zasady obliczania odchylenia przeciętnego wykorzystamy dane z przykładu 2.2 zestawione w szeregu rozdzielczym przedziałowym. Kolejne etapy postępowania zostały zaprezentowane w tablicy 3.7. Przypomnij-

my, że wartość średniej arytmetycznej $\overline{x} = 76,1$. Sumy kolumn 3 i 6 podstawiamy do wzoru (3.33) i otrzymujemy:

$$d_x = \frac{1}{\sum_{j=1}^{k} f_j} \cdot \sum_{j=1}^{k} f_j |x_j^\circ - \overline{x}| = \frac{1515,2000}{100} = 15,1520.$$

Tab. 3.7. Obliczanie odchylenia przeciętnego z szeregu rozdzielczego przedziałowego

j	$(x_{d\_j}; x_{g\_j}]$	f_j	v_j	Środki przedziałów x_j°	$x_j^\circ - \overline{x}$	$\lvert x_j^\circ - \overline{x}\rvert$	$\lvert x_j^\circ - \overline{x}\rvert f_j$	$\lvert x_j^\circ - \overline{x}\rvert v_j$
1	2	3	4	5	6	7	8	9
1	(36;48]	6	0,06	42	-34,1000	34,1000	204,6000	2,0460
2	(48;60]	14	0,14	54	-22,1000	22,1000	309,4000	3,0940
3	(60;72]	26	0,26	66	-10,1000	10,1000	262,6000	2,6260
4	(72;84]	24	0,24	78	1,9000	1,9000	45,6000	0,4560
5	(84;96]	15	0,15	90	13,9000	13,9000	208,5000	2,0850
6	(96;108]	9	0,09	102	25,9000	25,9000	233,1000	2,3310
7	(108;120]	4	0,04	114	37,9000	37,9000	151,6000	1,5160
8	(120;132]	2	0,02	126	49,9000	49,9000	99,8000	0,9980
SUMA		100	1,00	XXX	XXX	XXX	1515.2000	15,1520

Źródło: obliczenia własne.

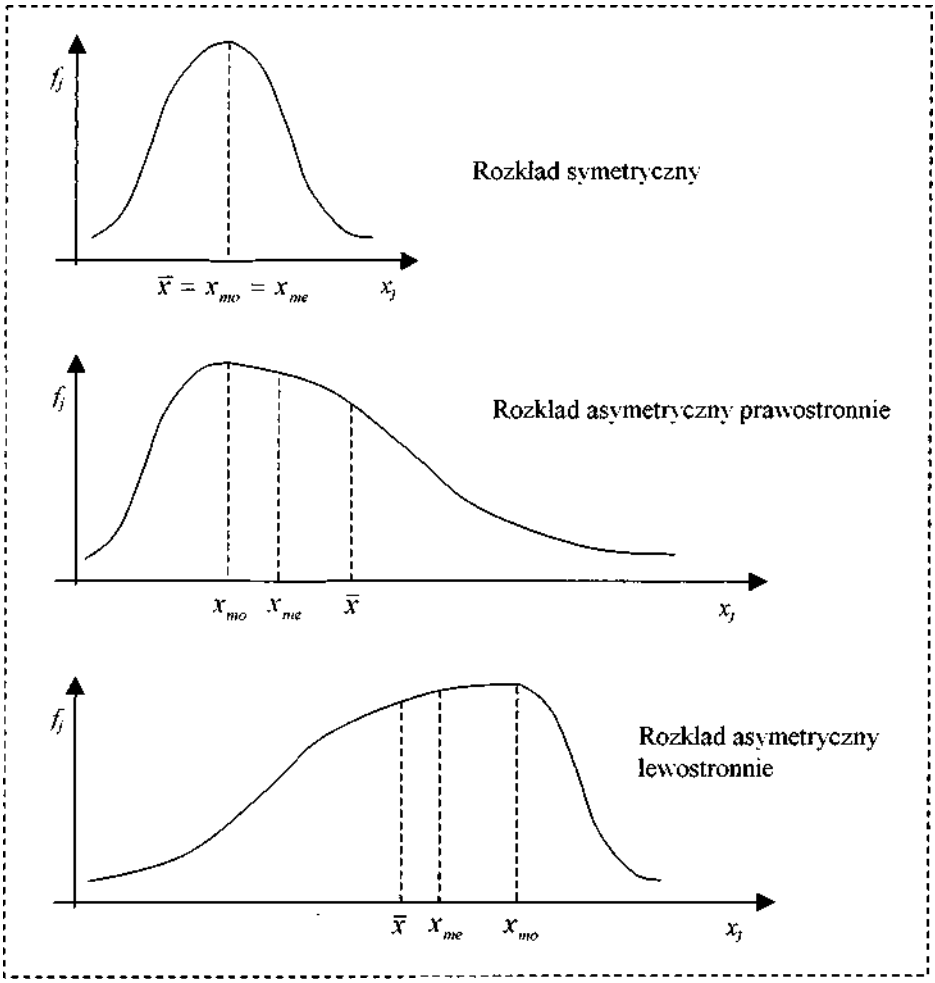
Obliczona wartość oznacza, że liczba osób korzystających, w kolejnych dniach, z usług biblioteki różni się przeciętnie od średniej liczby osób (76,1) o ±15,15.

Czytelnikowi polecamy, w ramach ćwiczeń, obliczenie i porównanie wartości odchyleń przeciętnych na podstawie szeregu szczegółowego i szeregu rozdzielczego punktowego.

3.2.4. Miary asymetrii

Miary asymetrii służą do zbadania skośności rozkładu badanej zmiennej. Jeżeli wartość średniej arytmetycznej jest równa medianie i modalnej ($\overline{x} = x_{me} = x_{mo}$), to wówczas rozkład badanej zmiennej jest symetryczny. Jeżeli modalna jest większa od mediany, a mediana większa od średniej arytmetycznej ($x_{mo} > x_{me} > \overline{x}$), to wówczas rozkład zmiennej jest asymetryczny lewostronnie. Natomiast w przypadku, gdy średnia jest większa od mediany, a mediana od modalnej($x_{mo} < x_{me} < \overline{x}$), to wówczas rozkład zmiennej charakteryzuje się asymetrią prawostronną. Położenie średnich w rozkładach symetrycznych i asymetrycznych zilustrowano graficznie na rys. 3.6.

Ponieważ w rozkładach asymetrycznych mediana leży pomiędzy średnią arytmetyczną i medianą, zatem korzystając z prostej zasady przechodniości, do stwierdzenia rodzaju asymetrii wystarczy porównać jedynie średnią arytmetyczną i modalną. Jeżeli średnia jest większa niż modalna ($\overline{x} > x_{mo}$), to wówczas o rozkładzie zmiennej mówimy, że jest asymetryczny prawostronnie, natomiast jeśli średnia jest mniejsza niż modalna ($\overline{x} < x_{mo}$), to wówczas rozkład zmiennej jest asymetryczny lewostronnie. W przypadku rozkładów symetrycznych spełniona jest relacja ($\overline{x} = x_{mo}$).



Rys. 3.6. Położenie średnich w rozkładach symetrycznych i asymetrycznych
Źródło: opracowanie własne.

W celu obliczenia siły asymetrii stosuje się współczynniki asymetrii. Jednym z nich jest współczynnik asymetrii (A_s) obliczany na podstawie średniej

arytmetycznej, modalnej i odchylenia standardowego*). Współczynnik ten oblicza się stosując wzór:

$$A_s = \frac{\overline{x} - x_{mo}}{s_x}. \tag{3.36}$$

Jeżeli $A_s = 0$, to rozkład jest symetryczny. Gdy $A_s > 0$ rozkład charakteryzuje się asymetrią prawostronną. Jeśli natomiast $A_s < 0$, to o rozkładzie można powiedzieć, że jest asymetryczny lewostronnie. Podczas analizy asymetrii rozkładu zmiennej X opisującej liczbę przyjętych pacjentów wykorzystamy obliczone wcześniej parametry. Jeżeli za podstawę obliczeń posłuży nam szereg szczegółowy lub rozdzielczy punktowy, współczynnik asymetrii wyniesie:

$$A_s = \frac{\overline{x} - x_{mo}}{s_x} = \frac{76,10 - 72,00}{18,7368} = 0,2188.$$

W przypadku szeregu rozdzielczego przedziałowego, współczynnik asymetrii będzie równy:

$$A_s = \frac{\overline{x} - x_{mo}}{s_x} = \frac{76,10 - 70,29}{19,0301} = 0,3.$$

W obydwu przypadkach współczynnik asymetrii jest większy od zera, co oznacza, że rozkład zmiennej X charakteryzuje się umiarkowaną prawostronną asymetrią. Błąd grupowania spowodował, w tym przypadku, zawyżenie współczynnika asymetrii.

3.2.5. Miary koncentracji (kurtoza)

Koncentrację można rozumieć jako skupienie zbiorowości wokół średniej albo nierównomierny podział zjawiska w zbiorowości. Pierwsza kwestia dotyczy jedynie rozkładów symetrycznych**) i nosi nazwę kurtozy. Do wytłumaczenia tego typu koncentracji niezbędne staje się wprowadzenie pojęcia **momentu rozkładu**.

Momentem r-tego rzędu (stopnia) nazywamy średnią arytmetyczną odchyleń poszczególnych wartości zmiennej X od dowolnej liczby A podniesionych do r-tej potęgi***). Jeżeli zbiór realizacji zmiennej X ma postać szeregu szczegółowego, to wówczas moment r-tego rzędu ($M_{x,r}$) będzie miał postać:

$$M_{x,r} = \frac{1}{N}\sum_{i=1}^{N}(x_i - A)^r. \tag{3.37}$$

*)Oprócz omawianego tutaj współczynnika jako miary asymetrii używa się również współczynniki zbudowane na podstawie kwartyli lub oparte o analizę tzw. momentu centralnego rzędu trzeciego. Zob. np. W. Makać, D. Urbanek-Krzysztofiak. *Metody opisu statystycznego*, Wydawnictwo Uniwersytetu Gdańskiego, Gdańsk 2001, s. 90.

**)Zob. podrozdział 3.1.

***)Zob. M. Woźniak, *Statystyka ogólna*, AE w Krakowie, Kraków 1999, s. 55.

W przypadku szeregu rozdzielczego punktowego i przedziałowego do wyznaczenia momentu wykorzystamy odpowiednio wzory (3.38) i (3.39) w postaci:

$$M_{x,r} = \frac{1}{\sum_{j=1}^{k} f_j} \cdot \sum_{j=1}^{k} f_j (x_j - A)^r, \tag{3.38}$$

$$M_{x,r} = \frac{1}{\sum_{j=1}^{k} f_j} \cdot \sum_{j=1}^{k} f_j (x_j^\circ - A)^r. \tag{3.39}$$

Jeśli A jest równe zero, to wówczas otrzymujemy moment zwykły ($M'_{x,r}$):

$$M'_{x,r} = \frac{1}{N} \sum_{i=1}^{N} x_i^r \tag{3.40}$$

lub

$$M'_{x,r} = \frac{1}{\sum_{j=1}^{k} f_j} \cdot \sum_{i=1}^{k} x_j^r f_j \tag{3.41}$$

lub

$$M'_{x,r} = \frac{1}{\sum_{j=1}^{k} f_j} \cdot \sum_{i=1}^{k} [(x_j^\circ)^r f_j]. \tag{3.42}$$

W przypadku, gdy A jest równe średniej arytmetycznej ($A = \overline{x}$), to mamy moment centralny $M''_{x,r}$:

$$M'_{x,r} = \frac{1}{N} \sum_{i=1}^{N} (x_i - \overline{x})^r \tag{3.43}$$

lub

$$M''_{x,r} = \frac{1}{\sum_{j=1}^{k} f_j} \cdot \sum_{j=1}^{k} f_j (x_j - \overline{x})^r \tag{3.44}$$

lub

$$M'_{x,r} = \frac{1}{\sum_{j=1}^{k} f_j} \cdot \sum_{j=1}^{k} f_j (x_j^\circ - \overline{x})^r. \tag{3.45}$$

Uważny czytelnik z łatwością zauważy, że omawiane wcześniej miary położenia i zmienności mogą być identyfikowane jako momenty. Na przykład średnia arytmetyczna, to moment zwykły rzędu pierwszego ($r = 1$), wariancja jest natomiast momentem centralnym rzędu drugiego ($r = 2$). Również odchylenie przeciętne może być identyfikowane jako bezwzględny moment centralny rzędu pierwszego.

Do badania natężenia koncentracji (skupienia) poszczególnych obserwacji wokół średniej wykorzystuje się moment centralny rzędu czwartego ($M''_{x,r=4}$) lub współczynnik koncentacji będący stosunkiem tegoż momentu do odchylenia standardowego podniesionego do potęgi czwartej. Współczynnik koncentracji (K_x) można zapisać:

$$K_x = \frac{M''_{x,r=4}}{s_x^4}. \tag{3.46}$$

Im wyższa wartość $M''_{x,r=4}$ oraz współczynnika K_x, tym realizacje zmiennej X charakteryzują się większym skupieniem wokół średniej, a krzywa liczebności jest bardziej wysmukła. Współczynnik koncentracji stanowi podstawę oceny stopnia spłaszczenia krzywej liczebności za pomocą współczynnika ekscesu:

$$E_x = K_x - 3. \tag{3.47}$$

Jeżeli $K_x = 3$ i w konsekwencji $E_x = 0$, to przyjmuje się, że krzywa liczebności jest zbliżona do krzywej tzw. rozkładu normalnego.
Jeżeli $K_x > 3$ i $E_x > 0$, to badany rozkład zmiennej jest bardziej wysmukły niż rozkład normalny.
W przypadku, gdy $K_x < 3$ i $E_x < 0$, to rozkład jest spłaszczony w stosunku do normalnego. Omawiane powyżej rozkłady i zależności pomiędzy nimi zostały przedstawione na rys. 3.1 w podrozdziale 3.1.

W analizowanym przykładzie, współczynnik asymetrii jest względnie niski ($A_s = 0,22$), a zatem zasadne będzie zbadanie stopnia koncentracji wartości zmiennej wokół średniej. W tym celu obliczamy moment centralny rzędu czwartego. Technikę obliczeń przedstawiono w tablicy 3.8.

W obliczeniach wykorzystano, że: $\overline{x} = 76,10$ oraz $S_x = \sqrt{S_x^2} = \sqrt{351,07} = 18,74$. Moment centralny rzędu czwartego oraz współczynnik koncentracji wynoszą odpowiednio:

$$M''_{x,r=4} = \frac{1}{\sum_{j=1}^{k} f_j} \cdot \sum_{j=1}^{k} f_j (x_j^\circ - \overline{x})^4 = \frac{36986473,4}{100} = 369864,734,$$

$$k_x = \frac{M''_{x,r=4}}{s_x^4} = \frac{369864,734}{18,74^4} = 2,9989 \approx 3.0.$$

Obliczona wartość współczynnika koncentracji oznacza, że krzywa rozkładu kształtem jest zbliżona do rozkładu normalnego.

Tab. 3.8. Obliczanie momentu centralnego rzędu czwartego

j	$(x_{d\_j}; x_{g\_j}]$	Liczba dni f_j	Środki przedziałów x_j°	$x_j^\circ - \overline{x}$	$(x_j^\circ - \overline{x})^4$	$f_j(x_j^\circ - \overline{x})^4$
1	**2**	**3**	**4**	**5**	**6**	**7**
1	(36;48]	2	8	-34,1000	1352127,10	8112762,58
2	(48;60]	9	12	-22,1000	238544,33	3339620,59
3	(60;72]	28	16	-10,1000	10406,04	270557,04
4	(72;84]	33	20	1,9000	13,03	312,77
5	(84;96]	18	24	13,9000	37330,10	559951,56
6	(96;108]	8	28	25,9000	449986,06	4049874,50
7	(108;120]	1	32	37,9000	2063273,69	8253094,75
8	(120;132]	1	36	49,9000	6200149,80	12400299,60
SUMA		100	XXX	XXX	XXX	36986473,40

Źródło: obliczenia własne.#

Termin koncentracja jest używany także do określenia nierównomiernego podziału łącznego funduszu wartości zmiennej ($\sum_{j=1}^{k} x_j f_j$) pomiędzy poszczególne klasy jednostek zbiorowości. Przykładem może być tutaj nierównomierny podział liczby pacjentów pomiędzy lekarzy określonej jednostki leczniczej lub nierównomierny podział dochodu pomiędzy indywidualne osoby. Koncentracja, w tym znaczeniu, jest ściśle powiązana ze zmiennością i asymetrią. Im dana zbiorowość charakteryzuje się większym zróżnicowaniem i większą asymetrią, tym koncentracja jest silniejsza. W skrajnych przypadkach koncentracja może być zupełna lub zerowa. O koncentracji zupełnej mówimy wówczas, gdy łączny fundusz zmiennej przypada na daną jednostkę zbiorowości (np. wszyscy pacjenci przychodni zdrowia przypisani są do jednego lekarza). Z koncentracją zerową spotykamy się natomiast wtedy, gdy na każdą jednostkę zbiorowości przypada jednakowa frakcja ogólnej sumy wartości (np. każdy lekarz przychodni ma przypisaną identyczną liczbę pacjentów).

Do badania tego typu koncentracji stosuje się dwie metody: graficzną i analityczną.

Metoda graficzna sprowadza się do wykreślenia w prostokątnym układzie współrzędnych tzw. **wieloboku koncentracji Lorenza**. W tym celu, na osi odciętych odkładamy wyrażone w procentach skumulowane częstości względne, natomiast na osi rzędnych – procentowe skumulowane częstości względne łącznego funduszu zmiennej. Po połączeniu otrzymanych w ten sposób punktów, wykreślamy tzw. krzywą Lorenza. Im bardziej krzywa Lorenza, odbiega od przekątnej kwadratu o bokach 100% na 100% (tzw. linii

równomiernego podziału), tym większą koncentracją charakteryzuje się rozkład zmiennej. Jeżeli krzywa pokryje się z przekątną kwadratu, to będziemy wnioskować, że nastąpił równomierny podział łącznego funduszu zmiennej pomiędzy wszystkie jednostki zbiorowości. Natomiast, w przypadku, gdy krzywa Lorenza pokryje się z dolną i prawą krawędzią kwadratu, wystąpi drugi ze skrajnych przypadków koncentracji – koncentracja zupełna.

Stosunek pola zawartego między linią równomiernego rozdziału a krzywą koncentracji Lorenza (a) do pola połowy kwadratu nosi nazwę współczynnika koncentracji Lorenza:

$$K_{x,L} = \frac{a}{5000}. \tag{3.48}$$

Wartość 5000 występująca w mianowniku powyższego wzoru jest wynikiem następującego działania: $(100 \cdot 100)/2 = 5000$. Jeżeli częstości względne oraz skumulowane częstości względne łącznego funduszu zmiennej zostaną wyrażone liczbami z przedziału [0;1], to wówczas w mianowniku wzoru (3.48) należy wprowadzić liczbę $(1 \cdot 1)/2$ tj. $0,5$. Powyższy współczynnik przybiera wartości z przedziału [0;1], przy czym, jeżeli $K_{x,L} = 0$, to wówczas mówimy o braku koncentracji, natomiast, jeżeli $K_{x,L} = 1$, koncentracja ma charakter zupełny. Pole powierzchni a wyznacza się jako różnicę pola połowy kwadratu o boku 100 i pola znajdującego się pod krzywą Lorenza. Do dokładnego określenia pola powierzchni pod krzywą Lorenza, niezbędna jest znajomość analitycznej postaci krzywej funkcji koncentracji. W praktyce powierzchnię tą określa się w sposób przybliżony jako sumę odpowiednich pól trójkąta i trapezów. Można zapisać, że:

$$a = 5000 - (P_1 - \sum_{j=2}^{k} P_j), \tag{3.49}$$

gdzie: P_1 – pole trójkąta, P_j – pole j-tego trapezu.

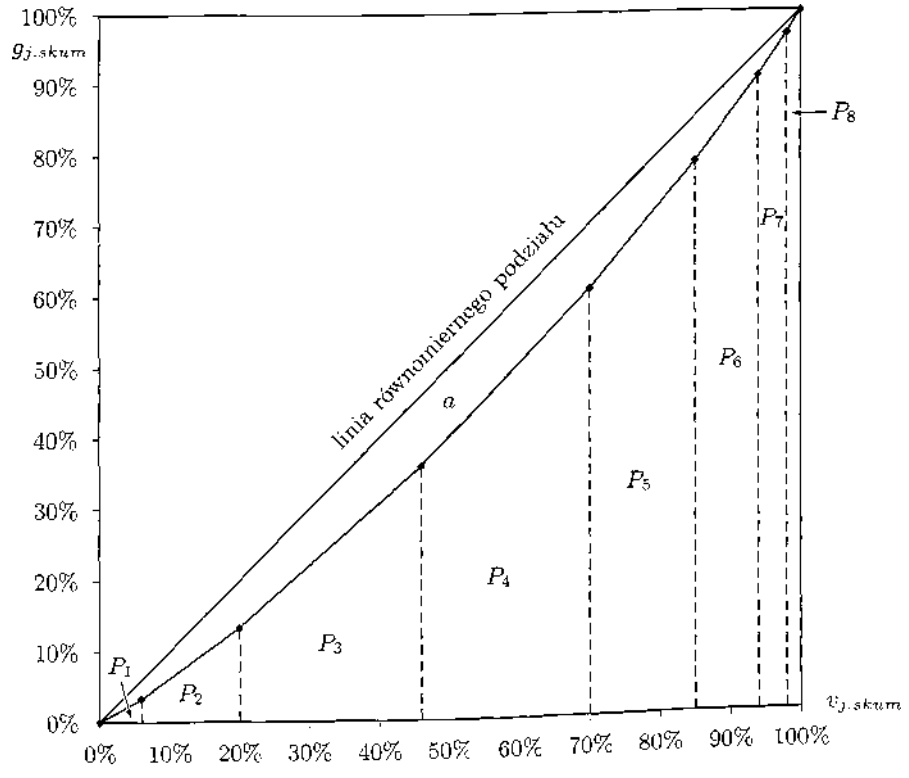
Poniżej zaprezentowano praktyczny sposób postępowania podczas pomiaru siły koncentracji metodą graficzną i analityczną. Do obliczeń wykorzystano dane z przykładu 2.2 zestawione w postaci szeregu rozdzielczego przedziałowego.

Wartości w kolumnach 8 i 9, stanowią współrzędne punktów niezbędnych do wyznaczenia krzywej koncentracji Lorenza. Kształt krzywej koncentracji oraz wielobok koncentracji prezentuje rysunek 3.7.

Tab. 3.9. Wykreślanie krzywej koncentracji Lorenza i obliczanie współczynnika koncentracji

j	$(x_{d.j};x_{g.j}]$	f_j	Środki przedziałów x_j°	$x_j^\circ f_j$	$v_j=\frac{f_j}{\sum_{j=1}^{k} f_j}$ (%)	$g_j=\frac{x_j^\circ f_j}{\sum_{j=1}^{k} x_j^\circ f_j}$ (%)	$v_{j.skum}$ (%)	$g_{j.skum}$ (%)
1	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**
1	(36;48]	6	42	252	6,00%	3,33%	6,00%	3,33%
2	(48;60]	14	54	756	14,00%	9,98%	20,00%	13,31%
3	(60;72]	26	66	1716	26,00%	22,66%	46,00%	35,97%
4	(72;84]	24	78	1872	24,00%	24,72%	70,00%	60,70%
5	(84;96]	15	90	1350	15,00%	17,83%	85,00%	78,53%
6	(96;108]	9	102	918	9,00%	12,12%	94,00%	90,65%
7	(108;120]	4	114	456	4,00%	6,02%	98,00%	96,67%
8	(120;132]	2	126	252	2,00%	3,33%	100,00%	100,00%
SUMA		100	XXX	7572	100,00%	100,00%	XXX	XXX

Źródło: obliczenia własne.



Rys. 3.7. Wielobok koncentracji
Źródło: opracowanie własne.

Korzystając w wzoru (3.48) można wyznaczyć współczynnik koncentracji. W tym celu obliczamy pole powierzchni a jako różnicę pomiędzy polem połowy kwadratu, a sumą pól figur geometrycznych znajdujących się pod krzywą koncentracji Lorenza. W analizowanym przykładzie będą to: pole trójkąta i pole siedmiu trapezów. Przypomnijmy, że pole trójkąta jest równe 1/2 długości podstawy razy długość jego wysokości opuszczonej na tą podstawę, czyli:

$$P_1 = \frac{g_{1.skum}}{2} \cdot v_{1.skum}. \tag{3.50}$$

Zatem:

$$P_1 = \frac{6 \cdot 3,33}{2} = 9,99.$$

Natomiast pole trapezu, to iloczyn średniej z podstaw i wysokości. Podstawy trapezu, to skumulowane odsetki $g_{j.skum}$, a wysokości są równe odsetkom v_j. Kolejne pola P_j dla $j > 1$ będą zatem równe:

$$P_j = \frac{g_{j.skum} + g_{j-1.skum}}{2} \cdot v_j. \tag{3.51}$$

Po podstawieniu wartości zamieszczonych w tablicy 3.9 otrzymamy:

$$P_2 = \frac{13,31 + 3,33}{2} \cdot 14 = 116,48; \qquad P_3 = \frac{35,97 + 13,31}{2} \cdot 26 = 640,73;$$

$$P_4 = \frac{60,7 + 35,97}{2} \cdot 24 = 1160,06; \qquad P_5 = \frac{78,53 + 60,7}{2} \cdot 15 = 1044,18;$$

$$P_6 = \frac{90,65 + 78,53}{2} \cdot 9 = 761,29; \qquad P_7 = \frac{96,67 + 90,65}{2} \cdot 4 = 374,64;$$

$$P_8 = \frac{100 + 96,67}{2} \cdot 2 = 196,67.$$

Stąd pole

$$\begin{aligned} a &= 5000 - (9,98 + 116,48 + 640,73 + 1160,06 \\ &\qquad + 1044,18 + 761,29 + 374,64 + 196,67) \\ &= 5000 - 4304,04 = 695,96 \end{aligned}$$

oraz współczynnik koncentracji:

$$K_{x,L} = \frac{a}{5000} = \frac{695,96}{5000} = 0,139.$$

Wartość współczynnika wskazuje na niezbyt silną koncentrację badanej zbiorowości. Potwierdziła się również wcześniejsza uwaga, dotycząca zależności pomiędzy współczynnikiem asymetrii i współczynnikiem koncentracji. Przypomnijmy, że omawiany rozkład charakteryzował się dość niskim współczynnikiem asymetrii $A_s = 0,22$.#

Analiza współzależności zmiennych

Rozdział 4

W dotychczasowych rozważaniach zbiorowość statystyczna była poddawana opisowi i analizie ze względu na jedną zmienną. Obecnie zajmiemy się sposobem analizy współzależności pomiędzy dwiema i więcej zmiennymi opisującymi tą samą zbiorowość. Podobnie, jak we wcześniejszych rozdziałach do oznaczania zmiennych używać będziemy dużych liter tj. X, Y lub X_1, X_2 itd.

4.1. Współzależność liniowa dwóch zmiennych

Nasze rozważania ograniczymy do analizy współzależności pomiędzy dwoma zmiennymi. Podobnie jak we wcześniejszych rozdziałach zakładać będziemy, że badania mają charakter wyczerpujący. Przedmiotem badań statystycznych będzie populacja składająca się z N jednostek statystycznych badana pod względem zmiennej dwuwymiarowej oznaczonej symbolem $(X;Y)$. Każda i-ta jednostka statystyczna jest opisana przez parę liczb $(x_i;y_i)$. Podobnie jak w przypadku jednej zmiennej, materiał statystyczny opisujący badaną zbiorowość, ze względu na dwie zmienne, może być przedstawiony w postaci szeregu szczegółowego lub w postaci dwuwymiarowego szeregu rozdzielczego nazywanego tablicą korelacyjną.

Przykład 4.1. Zbiorowość 20 studentów pewnej szkoły wyższej kierunku „Gospodarka przestrzenna" analizowano pod względem ocen uzyskanych na koniec semestru z przedmiotów makroekonomia (zmienna X) i statystyka (zmienna Y). W wyniku przeprowadzonych badań uzyskano następujący ciąg 20 par liczb:
(4;4), (3;4), (3;5), (3;4,5), (2;5), (4;4), (5;4), (5;3), (3;4), (4,5;4), (3,5;4), (3;4), (4,5;4), (3;4,5), (3;5), (3;4,5), (3;4,5), (3;4,5), (3,5;4), (4;4).
Źródło: badania własne.#

Powyższe dane przedstawione w postaci szeregu szczegółowego możemy pogrupować i transformować do postaci dwuwymiarowego szeregu rozdzielczego.

Zestawienie tego typu prezentuje tablica 4.1.

Tab. 4.1. Oceny 20 studentów z makroekonomii i ze statystyki

$X = x_i$ \ $Y = y_j$	2	3	3.5	4	4.5	5	suma
2	0	0	0	0	0	1	1
3	0	0	0	3	5	2	10
3.5	0	0	0	1	1	0	2
4	0	0	0	3	0	0	3
4.5	0	0	0	2	0	0	2
5	0	1	0	1	0	0	2
suma	0	1	0	10	6	3	20

Źródło: opracowanie własne.

Wartości w ostatnim wierszu nazywa się liczebnościami brzegowymi zmiennej Y. Wskazują one liczbę przypadków f_j, w których zmienna Y przyjęła wartość y_j. Odpowiednio wartości w ostatniej kolumnie są liczebnościami brzegowymi zmiennej X, które wskazują, w ilu przypadkach f_i zmienna X przyjęła wartość x_i. Suma liczebności brzegowych, zarówno dla zmiennej X, jak i zmiennej Y, jest równa sumie wszystkich realizacji zmiennych N. Wartości wewnątrz tablicy oznaczają, że w f_{ij} przypadkach zmienna X przyjęła wartość x_i, a zmienna Y wartość y_j.

Obserwując dane statystyczne zestawione w postaci tablicy podobnej do powyższej, można w przybliżeniu określić, czy istnieje związek pomiędzy opisującymi badaną zbiorowość zmiennymi. Jednak w większości przypadków samo stwierdzenie faktu występowania związku pomiędzy zmiennymi bywa niewystarczające. Z reguły oprócz tego, czy pomiędzy zmiennym występuje zależność interesuje nas także siła i kierunek tej współzależności. Należy zatem znaleźć odpowiedź na pytanie: czy i w jakim stopniu wzrostowi wartości jednej zmiennej towarzyszy wzrost wartości drugiej zmiennej, czy też wzrost wartości jednej zmiennej wpływa na spadek wartości drugiej.

Natężenie współzależności między zmiennymi w badanej zbiorowości wyraża się stopniem zgodności ich zmienności. Podczas badania współzależności poszukuje się odpowiedzi na pytanie, w jakim stopniu, zmienność jednej zmiennej jest uzależniona od zmienności drugiej. Do najczęściej stosowanych miar współzmienności zalicza się **kowariancję**, **współczynnik korelacji** oraz **współczynnik determinacji**. Jeżeli oprócz pytania o siłę i kierunek współzależności pytamy również, co się stanie z wartościami jednej ze zmiennych, jeżeli druga zmienna wzrośnie lub zmaleje o jednostkę, to wówczas zachodzi konieczność wyznaczenia parametrów tzw. **funkcji regresji**.

W dalszej części podręcznika skupimy się na sposobie wyznaczania współczynnika korelacji i funkcji regresji na podstawie szeregu szczegółowego dwóch zmiennych *).

4.1.1. Współczynnik korelacji linowej Pearsona

W badaniach wyczerpujących współczynnik korelacji liniowej (r_{xy}) definiuje się następująco:

$$r_{xy} = \frac{\operatorname{cov}(x,y)}{s_x \cdot s_y}, \tag{4.1}$$

gdzie:
s_x i s_y są odchyleniami standardowymi zmiennej X i zmiennej Y, natomiast $\operatorname{cov}(x,y)$ jest kowariancją (wspólną wariancją) zmiennych X i Y.
Kowariancja jest to mieszany moment centralny rzędu drugiego dwóch zmiennych. Można ją także określić jako średnią arytmetyczną wartości charakteryzujących współzmienność obu zmiennych. Kowariancja wyraża się wzorem:

$$\operatorname{cov}(x,y) = \frac{1}{N}\sum_{i=1}^{N}(x_i - \overline{x})(y_i - \overline{y}). \tag{4.2}$$

Przy założeniu, że odchylenia standardowe zmiennych X i Y, są odpowiednio równe:

$$s_x = \sqrt{\frac{1}{N}\sum_{i=1}^{N}(x_i - \overline{x})^2}, \quad s_y = \sqrt{\frac{1}{N}\sum_{i=1}^{N}(y_i - \overline{y})^2},$$

wzór (4.2) można zapisać w postaci:

$$r_{xy} = \frac{\frac{1}{N}\sum_{i=1}^{N}(x_i - \overline{x})(y_i - \overline{y})}{\sqrt{\frac{1}{N}\sum_{i=1}^{N}(x_i - \overline{x})^2}\sqrt{\frac{1}{N}\sum_{i=1}^{N}(y_i - \overline{y})^2}} \tag{4.3}$$

lub po pomnożeniu licznika i mianownika przez N w postaci:

$$r_{xy} = \frac{\sum_{i=1}^{N}(x_i - \overline{x})(y_i - \overline{y})}{\sqrt{\sum_{i=1}^{N}(x_i - \overline{x})^2 \cdot \sum_{i=1}^{N}(y_i - \overline{y})^2}}. \tag{4.4}$$

*) Czytelnika zainteresowanego sposobem wyznaczania korelacji w oparciu o szeregi rozdzielcze odsyłamy np. do pracy K. Zając, *Wykłady ze statystyki*, AE w Krakowie, Kraków 1995.

Współczynnik korelacji może przyjmować wartości z przedziału [-1; 1]. W sytuacji, gdy wartość współczynnika wynosi 1, to korelacja pomiędzy zmiennymi jest doskonała dodatnia, gdy -1 o korelacji mówimy, że jest doskonała ujemna. Jeżeli natomiast współczynnik korelacji przybierze wartość 0, to wówczas zmienność jednej ze zmiennych nie zależy od zmienności drugiej zmiennej.

Współczynnik korelacji liniowej stanowi podstawę do wyznaczenia współczynnika determinacji (R^2_{xy}) i współczynnika indeterminacji (φ^2_{xy}). Współczynnik determinacji wynosi:

$$R^2_{xy} = r^2_{xy}, \tag{4.5}$$

natomiast współczynnik indeterminacji:

$$\varphi^2_{xy} = 1 - R^2_{xy} = 1 - r^2_{xy}. \tag{4.6}$$

Zarówno współczynnik determinacji, jak i indeterminacji po przemnożeniu przez 100 można wyrazić w procentach. Wówczas współczynnik determinacji informuje nas, w jakim procencie zmienność jednej zmiennej (X lub Y) można wyjaśnić zmiennością drugiej zmiennej (Y lub X). Obliczając współczynnik indeterminacji dowiemy się, w jakim procencie zmienność jednej ze zmiennych (Y lub X) nie zależy od zmienności drugiej (Y lub X), lecz od innych czynników.

Przykład 4.2. Poniższe dane ukazują liczbę studentów Akademii Ekonomicznej w Krakowie oraz liczbę stanowisk komputerowych w latach 1990 – 1999.

Tab. 4.2. Liczba studentów Akademii Ekonomicznej w Krakowie oraz liczba stanowisk komputerowych w latach 1990 – 1999

Lata	i	x_i (liczba studentów)	y_i (liczba stanowisk komputerowych)
1990	1	4780	24
1991	2	4975	40
1992	3	7732	122
1993	4	9701	177
1994	5	12154	198
1995	6	14300	203
1996	7	15766	214
1997	8	17516	224
1998	9	18355	254
1999	10	18633	254

Źródło: *Akademia Ekonomiczna w liczbach*, op. cit.

Na podstawie powyższych informacji ustalić siłę i kierunek współzależności obu zjawisk. W tablicy poniżej zamieszczone zostały wszystkie pomocnicze

obliczenia konieczne do wyznaczenia współczynnika korelacji na podstawie wzoru (4.4).

Tab. 4.3. Obliczanie współczynnika korelacji liniowej

i	x_i (l. stud.)	y_i (l. st. komp.)	$x_i-\overline{x}$	$y_i-\overline{y}$	$(x_i-\overline{x})\cdot(y_i-\overline{y})$	$(x_i-\overline{x})^2$	$(y_i-\overline{y})^2$
1	**2**	**3**	**4**	**5**	**6**	**7**	**8**
1	4780	24	-7611,20	-147,00	1118846,40	57930365,44	21609,00
2	4975	40	-7416,20	-131,00	971522,20	55000022,44	17161,00
3	7732	122	-4659,20	-49,00	228300,80	21708144,64	2401,00
4	9701	177	-2690,20	6,00	-16141,20	7237176,04	36,00
5	12154	198	-237,20	27,00	-6404,40	56263,84	729,00
6	14300	203	1908,80	32,00	61081,60	3643517,44	1024,00
7	15766	214	3374,80	43,00	145116,40	11389275,04	1849,00
8	17516	224	5124,80	53,00	271614,40	26263575,04	2809,00
9	18355	254	5963,80	83,00	494995,40	35566910,44	6889,00
10	18633	254	6241,80	83,00	518069,40	38960067,24	6889,00
Σ	**123912**	**1710**	**0,00**	**0,00**	**3787001,00**	**257755317,60**	**61396,00**

Źródło: obliczenia własne.

Wartości średniej liczby studentów oraz średniej liczby stanowisk komputerowych obliczamy stosując znany nam wzór na średnią arytmetyczną szeregu szczegółowego:

$$\overline{x} = \frac{1}{N}\sum_{i=1}^{10} x_i = \frac{123912}{10} = 12391,2$$

oraz

$$\overline{y} = \frac{1}{N}\sum_{i=1}^{10} y_i = \frac{1710}{10} = 171.$$

Podstawiając do licznika wzoru (4.4) sumę wartości z kolumny 6 oraz do mianownika pierwiastek kwadratowy z iloczynów sum kolumny 7 i 8 otrzymamy:

$$r_{xy} = \frac{\sum_{i=1}^{10}(x_i-\overline{x})(y_i-\overline{y})}{\sqrt{\sum_{i=1}^{10}(x_i-\overline{x})^2 \cdot \sum_{i=1}^{10}(y_i-\overline{y})^2}} = \frac{3787001}{\sqrt{257755317,6 \cdot 61396}} = 0,951966.$$

Obliczona wartość współczynnika korelacji wskazuje, że pomiędzy badanymi zmiennymi (liczbą studentów i liczbą stanowisk komputerowych) istnieje wysoka dodatnia współzależność. Wartość współczynnika determinacji otrzymamy podnosząc do kwadratu współczynnik korelacji:

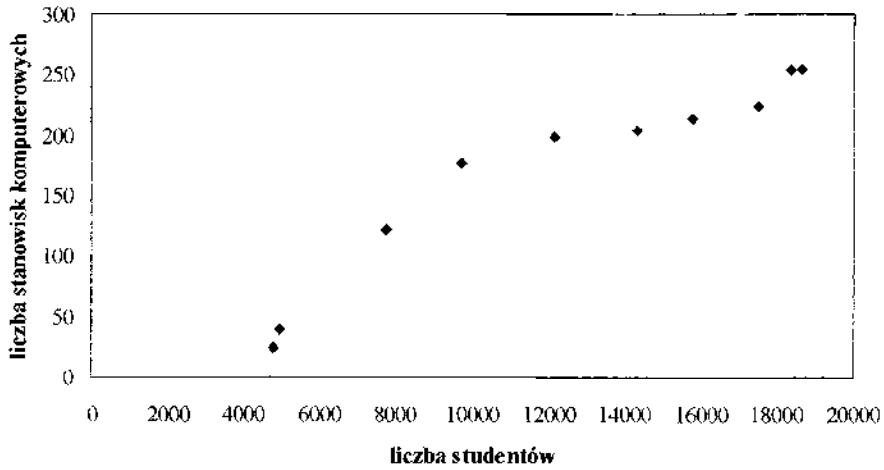
$$R_{xy}^2 = r_{xy}^2 = 0,951966^2 = 0,90624 \quad (90,624\%).$$

Po odjęciu od jedności współczynnika determinacji mamy współczynnik indeterminacji:

$$\varphi_{xy}^2 = 1 - r_{xy}^2 = 1 - 0,90624 = 0,09376 \quad (9,376\%).$$

Z powyższych obliczeń wynika, że zmienność jednej ze zmiennych (liczby studentów lub liczby komputerów) jest zależna od zmienności drugiej (liczby komputerów lub liczby studentów) w około 90,624%, natomiast w 9,376% zależy od innych czynników nie ujętych w powyższym badaniu.

Zbiór realizacji obserwowanych zmiennych (X,Y) można przedstawić również graficznie, wykreślając tzw. diagram współzależności. Na rys. 4.1 został przedstawiony wykres korelacyjny dla wartości X i Y zestawionych w tablicy 4.3.



Rys. 4.1. Wykres współzależności pomiędzy średnim liczbą studentów a liczbą stanowisk komputerowych
Źródło: tablica 4.2.

4.1.2. Funkcja regresji dwóch zmiennych

Często, oprócz badania siły i kierunku współzależności pomiędzy zmiennymi, wyznacza się funkcję za pomocą, której można aproksymować zależność pomiędzy nimi. Funkcję taką nazywa się funkcją regresji, a jej parametry współczynnikami regresji. Badanie może prowadzić do określenia zależności Y względem X lub też X względem Y. Jeżeli rozważamy regresję Y względem X, to wówczas Y jest zmienną objaśnianą (endogeniczną), natomiast X zmienną objaśniająca (egzogeniczną). Jeżeli zostaną odwrócone role zmiennych i będziemy badać regresję X względem Y, wówczas X określać będziemy mianem

zmiennej objaśnianej (endogenicznej), a Y zmiennej objaśniającej (egzogenicznej). Jednym z ważniejszych punktów w procesie tworzenia funkcji regresji jest wybór jej postaci analitycznej. W przypadku modelu z jedną zmienną objaśniającą, wybór postaci funkcyjnej modelu jest stosunkowo łatwy. Wówczas podstawową rolę odgrywa wizualna ocena wykresu współzależności.

Sprawa wyboru postaci funkcji komplikuje się, gdy w modelu występuje więcej niż jedna zmienna objaśniająca. Wówczas, podczas wyboru postaci funkcji, oprócz algorytmów numerycznych, należy uwzględnić również wiedzę płynącą z innych źródeł, takich jak np. makro- i mikroekonomia. Do najczęściej stosowanych w badaniach ekonomicznych funkcji należą: funkcja liniowa jednej zmiennej, funkcja liniowa wielu zmiennych, funkcja wykładnicza, funkcja potęgowa, funkcja logarytmiczna, funkcje Törnquista, wielomiany stopnia drugiego i trzeciego oraz funkcje logistyczne*).

W dalszej części tego rozdziału poprzestaniemy na opisie sposobu wyznaczenia liniowej funkcji regresji postaci:

$$y = ax + b. \tag{4.7}$$

Do wyznaczenia wartości parametrów funkcji a i b wykorzystamy metodę najmniejszych kwadratów (MNK). Jej nazwa wiąże się z tym, że podczas dopasowywania linii do wartości empirycznych, dąży się do tego, aby suma kwadratów odchyleń wartości empirycznych y_i od wartości hipotetycznych $(\hat{y}_i)$ wynikających z funkcji regresji $\hat{y}_i = ax_i + b$, była jak najmniejsza. Można, zatem zapisać:

$$\sum_{i=1}^{N}(y_i - \hat{y}_i)^2 = \min. \tag{4.8}$$

Rzecz sprowadza się zatem do wyznaczenia minimum funkcji $f(a, b)$ postaci:

$$f(a, b) = \sum_{i=1}^{N}(y_i - \hat{y}_i)^2 = \sum_{i=1}^{N}(y_i - ax_i - b)^2. \tag{4.9}$$

W celu wyznaczenia minimum powyższej funkcji należy ją zróżniczkować względem a i b. Pierwsza pochodna funkcji względem a i b wynosi odpowiednio:

$$\frac{\partial f(a, b)}{\partial b} = -2\sum_{i=1}^{N} y_i + 2Nb + 2a\sum_{i=1}^{N} x_i, \tag{4.10}$$

*) Większość tych funkcji powinna być znana czytelnikowi ze szkoły średniej. Funkcje te zostały opisane także np. w: A. Goryl, Z. Jędrzejczyk, K. Kukuła, J. Osiewalski, A. Walkosz, *Wprowadzenie do Ekonometrii w przykładach i zadaniach*, Wydawnictwo Naukowe PWN, Warszawa 1996, s. 23.

$$\frac{\partial f(a,b)}{\partial b} = -2\sum_{i=1}^{N} x_i y_i + 2b\sum_{i=1}^{N} x_i + 2a\sum_{i=1}^{N} x_i^2. \tag{4.11}$$

Po przyrównaniu powyższych funkcji do zera i po odpowiednich przekształceniach otrzymamy układ równań:

$$\begin{cases} a_{yx}\sum_{i=1}^{N} x_i + b_{yx}N = \sum_{i=1}^{N} y_i, \\ a_{yx}\sum_{i=1}^{N} x_i^2 + b_{yx}\sum_{i=1}^{N} x_i = \sum_{i=1}^{N} x_i y_i. \end{cases} \tag{4.12}$$

Z układu tego, stosując dowolną metodę rozwiązywania (np. metodę podstawiania), należy wyznaczyć parametry a_{yx} i b_{yx}:

$$a_{yx} = \frac{\sum_{i=1}^{N} x_i y_i - \frac{\sum_{i=1}^{N} x_i \sum_{i=1}^{N} y_i}{N}}{\sum_{i=1}^{N} x_i^2 - \frac{\left(\sum_{i=1}^{N} x_i\right)^2}{N}}, \tag{4.13}$$

$$b_{yx} = \frac{\sum_{i=1}^{N} y_i}{N} - a_{yx}\frac{\sum_{i=1}^{N} x_i}{N}. \tag{4.14}$$

Wyrażenia (4.13) i (4.14) można też przedstawić w formie:

$$a_{yx} = \frac{\sum_{i=1}^{N}(x_i - \overline{x})(y_i - \overline{y})}{\sum_{i=1}^{N}(x_i - \overline{x})^2} = r_{xy}\frac{s_y}{s_x}, \tag{4.15}$$

$$b_{yx} = \overline{y} - a_{yx}\overline{x}. \tag{4.16}$$

Postępując analogicznie w przypadku liniowej funkcji regresji X względem Y otrzymamy parametry a_{xy} i b_{yx} funkcji regresji $\hat{x}_i = a_{xy}y_i + b_{xy}$:

$$a_{xy} = \frac{\sum_{i=1}^{N}(x_i - \overline{x})(y_i - \overline{y})}{\sum_{i=1}^{N}(y_i - \overline{y})^2} = r_{xy}\frac{s_x}{s_y}, \tag{4.17}$$

$$b_{xy} = \overline{x} - a_{xy}\overline{y}. \tag{4.18}$$

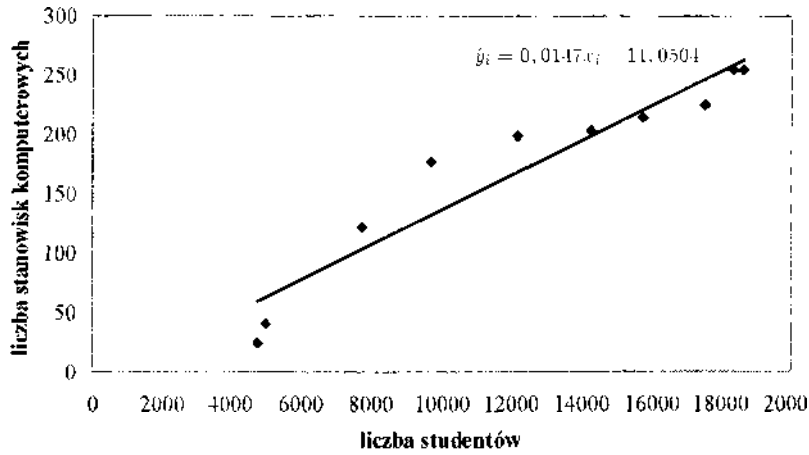
Przykład 4.3. Korzystając z danych ujętych w przykładzie 4.2, oszacować parametry liniowej funkcji regresji Y względem X oraz X względem Y. Położenie punktów na diagramie korelacyjnym (zob. rys. 4.1) wskazuje na możliwość dopasowania liniowej funkcji regresji. Wykorzystując obliczenia zawarte w tablicy 4.4, oraz wzory (4.15) i (4.16), otrzymujemy:

$$a_{yx} = \frac{\sum_{i=1}^{10}(x_i - \overline{x})(y_i - \overline{y})}{\sum_{i=1}^{10}(x_i - \overline{x})^2} = \frac{3787001}{257755317,6} = 0,0147$$

oraz

$$b_{yx} = \overline{y} - a_{yx}\overline{x} = 171 - 0,0147 \cdot 12391,2 = -11,0544.$$

Zatem, funkcja regresji Y względem X ma postać: $\hat{y}_i = 0,01469223x_i - 11,05438874$, a jej wykres przedstawia rys. 4.2. W analizowanym przykładzie parametr a informuje nas, że jeżeli liczba studentów wzrośnie o jednostkę (1 osobę), to wówczas liczba komputerów wzrośnie średnio o 0,01469. Wynik ten można zinterpretować również inaczej i stwierdzić, że jeżeli liczba studentów wzrośnie o około 68 osób, to wówczas liczba komputerów powinna wzrosnąć o jedno stanowisko.



Rys. 4.2. Funkcja regresji $\hat{y}_i = a_{yx}x_i + b_{yx}$
Źródło: opracowanie własne.

Natomiast parametr przesunięcia b_{yx}, wskazuje, ile wyniosłaby teoretycznie liczba stanowisk komputerowych, gdyby liczba studentów uczelni wynosiła 0.

Z uwagi na fakt, że b w tym przykładzie jest mniejsze od zera jego interpretacja staje się niemożliwa.

Aby oszacować parametry funkcji regresji X względem Y postaci $\hat{x}_i = a_{xy}y_i + b_{xy}$, należy skorzystać z wzorów (4.17) i (4.18) i wówczas otrzymamy:

$$a_{xy} = \frac{\sum_{i=1}^{N}(x_i - \overline{x})(y_i - \overline{y})}{\sum_{i=1}^{N}(y_i - \overline{y})^2} = \frac{3787001}{61396} = 61,68156,$$

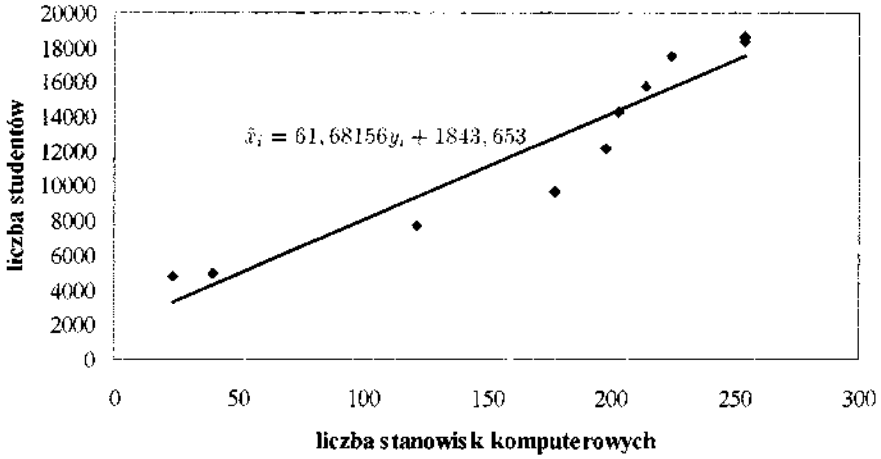
$$b_{xy} = \overline{x} - a_{xy}\overline{y} = 12391,2 - 61,68156 \cdot 171 = 1843,653.$$

Zatem funkcja regresji ma postać:

$$\hat{x}_i = 61,68156y_i + 1843,653.$$

Wartość współczynnika regresji informuje nas, że jeżeli liczba komputerów wzrośnie o jedno stanowisko, to wówczas liczba studentów zwiększy się średnio o ok. 62 osoby. Drugi z parametrów, wyraz wolny, oznacza, że teoretycznie w przypadku, gdyby uczelnia nie dysponowała komputerami, liczba studentów powinna wynieść ok. 1844 osoby.

Wykres wyznaczonej funkcji został przedstawiony na rysunku 4.3.#



Rys. 4.3. Funkcja regresji $\hat{x}_i = a_{xy}y_i + b_{xy}$
Źródło: opracowanie własne.

Po wyznaczeniu funkcji regresji należy zbadać, jak dobrze funkcja hipotetyczna jest dopasowana do danych empirycznych. W tym celu należy wyznaczyć wariancję resztową (s_ε^2) oraz odchylenie standardowe składnika resztowego (s_ε).

Jeżeli rozpatrujemy funkcję regresji Y względem X, to wówczas wariancję resztową obliczamy według wzoru*):

$$s^2_{\varepsilon(y/x)} = \frac{\sum_{i=1}^{N}(y_i - \hat{y})^2}{N}, \tag{4.19}$$

natomiast odchylenie standardowe składnika resztowego:

$$s_{\varepsilon(y/x)} = \sqrt{s^2_{\varepsilon(y/x)}} = \sqrt{\frac{\sum_{i=1}^{N}(y_i - \hat{y})^2}{N}}. \tag{4.20}$$

Odchylenie standardowe składnika resztowego informuje, o ile średnio (in plus lub in minus) odchylają się wartości empiryczne Y od wartości hipotetycznych określonych na podstawie funkcji regresji.
Dla funkcji regresji X względem Y wariancja resztowa oraz odchylenie standardowe składnika resztowego, będą miały postać:

$$s^2_{\varepsilon(x/y)} = \frac{\sum_{i=1}^{N}(x_i - \hat{x})^2}{N} \tag{4.21}$$

oraz

$$s_{\varepsilon(x/y)} = \sqrt{s^2_{\varepsilon(x/y)}} = \sqrt{\frac{\sum_{i=1}^{N}(x_i - \hat{x})^2}{N}}. \tag{4.22}$$

Dokładność oszacowania funkcji regresji można również ocenić na podstawie, wspomnianych już wcześniej, współczynników determinacji i indeterminacji. Współczynnik indeterminacji (φ^2_{yx}) funkcji regresji Y względem X jest równy:

$$\varphi^2_{yx} = \frac{\sum_{i=1}^{N}(y_i - \hat{y})^2}{\sum_{i=1}^{N}(y_i - \overline{y})^2}, \tag{4.23}$$

*)W przypadku, gdy badania mają charakter częściowy, w mianowniku wyrażenia (4.19), zamiast N, podstawia się $n - k$, gdzie n jest licznością próby, k – liczbą szacowanych parametrów.

natomiast współczynnik determinacji:

$$R_{yx}^2 = 1 - \varphi_{yx}^2 = 1 - \frac{\sum_{i=1}^{N}(y_i - \hat{y})^2}{\sum_{i=1}^{N}(y_i - \overline{y})^2}. \tag{4.24}$$

Analogicznie można zapisać powyższe współczynniki dla funkcji regresji X względem Y. Otrzymamy wówczas:

$$\varphi_{xy}^2 = \frac{\sum_{i=1}^{N}(x_i - \hat{x})^2}{\sum_{i=1}^{N}(x_i - \overline{x})^2}, \tag{4.25}$$

natomiast współczynnik determinacji:

$$R_{xy}^2 = 1 - \varphi_{xy}^2 = 1 - \frac{\sum_{i=1}^{N}(x_i - \hat{x})^2}{\sum_{i=1}^{N}(x_i - \overline{x})^2}. \tag{4.26}$$

Współczynnik determinacji informuje nas, jaka część zmian wartości zmiennej objaśnianej została wyjaśniona, przez oszacowaną funkcję regresji. Im współczynnik determinacji jest bliższy jedności, tym funkcja regresji jest lepiej dopasowana do danych empirycznych.

Współczynnik indeterminacji wskazuje natomiast, jaka część zmienności zmiennej objaśnianej nie jest wyjaśniona przez zmienność zmiennej objaśniającej (zmiennych objaśniających) występujących w funkcji regresji. Im wartość współczynnika jest bliższa zeru, tym funkcja regresji jest lepiej dopasowana do zmiennych empirycznych.

W celu zbadania dobroci dopasowania funkcji regresji $\hat{y}_i = 0,01469223x_i - 11,05438874$, w pierwszej kolejności należy podstawić do wzoru kolejne wartości zmiennej X i wyznaczyć wartości hipotetyczne (zob. tab. 4.5, kolumna 4). W kolejnym kroku od wartości empirycznych odejmujemy wartości teoretyczne i podnosimy do kwadratu (kolumny 5 i 6).

Tab. 4.4. Badanie dobroci dopasowania funkcji regresji Y względem X

i	x_i (liczba stud.)	y_i (liczba stan. komp.)	$\hat{y}_i$	$y_i - \hat{y}$	$(y_i - \hat{y})$	$(y_i - \overline{y})^2$
1	**2**	**3**	**4**	**5**	**6**	**7**
1	4780	24	59,17	-35,17	1237,24	21609,00
2	4975	40	62,04	-22,04	485,74	17161,00
3	7732	122	102,55	19,45	378,46	2401,00
4	9701	177	131,47	45,53	2072,53	36,00
5	12154	198	167,52	30,48	929,34	729,00
6	14300	203	199,04	3,96	15,65	1024,00
7	15766	214	220,58	-6,58	43,34	1849,00
8	17516	224	246,29	-22,29	497,06	2809,00
9	18355	254	258,62	-4,62	21,36	6889,00
10	18633	254	262,71	-8,71	75,79	6889,00
sumy	**123912**	**1710**	**XXX**	**XXX**	**5756,50**	**61396,00**

Źródło: obliczenia własne.

Suma kolumny 6 po podzieleniu przez $N = 10$, pozwala na wyznaczenie wariancji resztowej. Wyciągając pierwiastek kwadratowy z wariancji otrzymamy odchylenie standardowe składnika resztowego. Mamy wówczas:

$$s^2_{\varepsilon(y/x)} = \frac{1}{N}\sum_{i=1}^{N}(y_i - \hat{y})^2 = \frac{5756,50}{10} = 575,65$$

oraz

$$s_{\varepsilon(y/x)} = \sqrt{s^2_{\varepsilon(y/x)}} = 23,99271.$$

Powyższy wynik oznacza, że wartości empiryczne liczby komputerów odchylają się od wartości hipotetycznych średnio o $\pm 23,99 \approx 24$ osoby.

Do wyznaczenia współczynnika determinacji i indeterminacji potrzebna jest również suma kwadratów odchyleń wartości empirycznych zmiennej Y od wartości średniej $\overline{y}$ (zob. kolumna 7 tablicy 4.5). Po podstawieniu do wzorów (4.23) i (4.24) otrzymamy:

$$\varphi^2_{yx} = \frac{\sum_{i=1}^{N}(y_i - \hat{y})^2}{\sum_{i=1}^{N}(y_i - \overline{y})^2} = \frac{5756,5}{61396} = 0,09376 \quad (9,376\%),$$

$$R^2_{yx} = 1 - \varphi^2_{yx} = 1 - \frac{\sum_{i=1}^{N}(y_i - \hat{y})^2}{\sum_{i=1}^{N}(y_i - \overline{y})^2} = 1 - \frac{5756,5}{61396} = 0,90624 \quad (90,624\%).$$

Z powyższych obliczeń wynika, że w przeszło 90,6% zmienność zmiennej objaśnianej (liczba stanowisk komputerowych) wynika ze zmienności zmiennej objaśniającej (liczby studentów), natomiast 9,4% zmienności zmiennej objaśnianej jest uzależniona od innych czynników nie objętych badaniem. #

W podobny sposób można zbadać dobroć dopasowania funkcji regresji X względem Y, wykorzystując tym razem wzory (4.22), (4.25) i (4.26). Analizę taką zalecamy sumiennemu czytelnikowi jako jedno z zadań domowych.

4.2. Inne miary współzależności

4.2.1. Współczynnik korelacji dwuseryjnej

Jeżeli jedna ze zmiennych jest zmienną ciągłą (np. zmienna Y), a druga zmienna jest zmienną zero – jedynkową (np. zmienna X), to do określenia siły współzależności pomiędzy zmiennymi można wykorzystać tzw. **współczynnik korelacji dwuseryjnej** ($r_{d.xy}$) postaci:

$$r_{d.xy} = \frac{\overline{y}_1 - \overline{y}_0}{s_y} \sqrt{\frac{N_1 \cdot N_0}{N(N-1)}}, \tag{4.27}$$

gdzie:

$\overline{y}_0$ – średnia arytmetyczna realizacji zmiennej Y, skojarzonych z realizacjami zmiennej X o wartości 0,

$\overline{y}_1$ – średnia arytmetyczna realizacji zmiennej Y, skojarzonych z realizacjami zmiennej X o wartości 1,

s_y – odchylenie standardowe zmiennej Y,

N_0 – liczebność podzbioru zer,

N_1 – liczebność podzbioru jedynek,

$N = N_0 + N_1$.

Przykład 4.4. W celu zbadania wpływu uczestnictwa na wykładzie na wyniki otrzymane ze sprawdzianu ze statystyki, poddano badaniu grupę 10 studentów. Pierwszą z cech oceniano na skali dwupunktowej w następujący sposób:

$$X = \begin{cases} 1, & \text{gdy student był obecny,} \\ 0, & \text{gdy student był nieobecny,} \end{cases}$$

natomiast druga z cech była oceniana na skali punktowej od 0 do 25 punktów. Otrzymane wyniki zawarto w tablicy 4.5..

Czy istnieje współzależność pomiędzy wynikami ze sprawdzianu a obecnością na wykładzie?

Tab. 4.5. Zależność między obecnością na wykładzie a liczbą uzyskanych punktów

Student (i)	Obecność (x_i)	Liczba punktów (y_i)
1	1	15,5
2	0	12
3	0	13
4	1	20
5	0	8
6	0	10
7	1	20,5
8	0	14
9	1	19
10	1	18

Źródło: badania własne.

W poniższej tablicy zostały ujęte rachunki pomocnicze potrzebne do wyznaczenia wariancji i odchylenia standardowego zmiennej Y. Średnia liczba punktów wyniosła:

$$\overline{y} = \frac{1}{N}\sum_{i=1}^{10} y_i = \frac{150}{10} = 15.$$

Tab. 4.6. Obliczanie współczynnika korelacji dwuseryjnej

Student i	Obecność x_i	Liczba punktów y_i	$y_i - \overline{y}$	$(y_i - \overline{y})^2$
1	**2**	**3**	**4**	**5**
1	1	15,5	0,5	0,25
2	0	12	-3	9
3	0	13	-2	4
4	1	20	5	25
5	0	8	-7	49
6	0	10	-5	25
7	1	20,5	5,5	30,25
8	0	14	-1	1
9	1	19	4	16
10	1	18	3	9
suma	**5**	**150**		**168,5**

Źródło: obliczenia własne.

Wariancja i odchylenie standardowe zmiennej Y wyniosły:

$$s_y^2 = \frac{1}{N}\sum_{i=1}^{10}(y_i - \overline{y})^2 = \frac{168,5}{10} = 16,85,$$

$$s_y = \sqrt{16,85} \approx 4,1.$$

Wartości $\overline{y}_0$ i $\overline{y}_1$, N, N_0, N_1, obliczamy następująco:

$$\overline{y}_0 = \frac{12+12+8+10+14}{5} = 11,4,$$

$$\overline{y}_1 = \frac{15,5+20+20,5+19+18}{5} = 18,6.$$

$$N_0 = N_1 = 5 \Rightarrow N = 10.$$

Zatem współczynnik korelacji dwuseryjnej wyniesie:

$$r_{d.xy} = \frac{\overline{y}_0 - \overline{y}_1}{s_y}\sqrt{\frac{N_1 \cdot N_0}{N(N-1)}} = \frac{18,6-11,4}{4,1}\sqrt{\frac{5\cdot 5}{10\cdot 9}} = 0,92.$$

Uzyskany wynik świadczy o bardzo dużym związku pomiędzy obecnością podczas wykładu, a liczbą uzyskanych punktów podczas sprawdzianów kontrolnych.

4.2.2. Współczynnik skojarzenia

Załóżmy obecnie, że obydwie zmienne X i Y, to zmienne zero – jedynkowe. Rozkład wartości zmiennych oraz ich liczebności przedstawia tablica 4.7.

Tab. 4.7. Rozkład wartości zmiennych

		Y		Σ
		0	1	
X	0	$f(0;0)$	$f(0;1)$	$f(0;0)+f(0;1)$
	1	$f(1;0)$	$f(1;1)$	$f(1;0)+f(1;1)$
Σ		$f(0;0)+f(1;0)$	$f(0;1)+f(1;1)$	N

Źródło: opracowanie własne.

W ostatniej kolumnie i w ostatnim wierszu tablicy zapisuje się liczebności brzegowe zmiennej X i zmiennej Y.
Współczynnik skojarzenia (Q_{xy}) oblicza się stosując wzór:

$$Q_{xy} = \frac{f(0;0)\cdot f(1;1) - f(0;1)\cdot f(1;0)}{f(0;0)\cdot f(1;1) + f(0;1)\cdot f(1;0)} \tag{4.28}$$

Przykład 4.5. 60 studentów regularnie przygotowywało się do zajęć ze „Statystyki", a 40 nieregularnie. W grupie pierwszej egzaminy poprawkowe zdarzyły się 10 razy w ciągu studiów, a w drugiej aż 30. Czy istnieje związek pomiędzy solidnością pracy i koniecznością poprawkowych egzaminów? Proszę uzasadnić odpowiedź posługując się odpowiednim miernikiem.

Przyjmijmy następujące oznaczenia:
$X = 0$, jeżeli student zdał egzamin w pierwszym terminie,
$X = 1$, jeżeli student miał egzamin poprawkowy,
$Y = 0$, jeżeli student uczył się regularnie,
$Y = 1$, jeżeli student uczył się nieregularnie.

Tablica korelacji do przykładu 4.5 ma postać:

Tab. 4.8. Rozkład wartości zmiennych do przykładu 4.5

		Y		Σ
		0	1	
X	0	50	10	60
	1	10	30	40
Σ		60	40	100

Źródło: obliczenia własne.

Po podstawieniu do wzoru (4.28), otrzymamy:

$$Q_{xy} = \frac{f(0;0) \cdot f(1;1) - f(0;1) \cdot f(1;0)}{f(0;0) \cdot f(1;1) + f(0;1) \cdot f(1;0)} = \frac{50 \cdot 30 - 10 \cdot 10}{50 \cdot 30 + 10 \cdot 10} = 0,875.$$

Otrzymany wynik świadczy o wysokiej dodatniej współzależności pomiędzy solidnością pracy, a terminem zdania egzaminu.

4.2.3. Współczynnik korelacji rang Spearmana

Współczynnik korelacji rang stosuje się wówczas, gdy wartości cech mierzalnych opisanych przez odpowiednie zmienne (ciągłe lub skokowe) lub warianty cechy niemierzalnej, zostały zastąpione rangami, czyli kolejnymi liczbami. W sytuacji, gdy w rangowanym ciągu pojawią się takie same wartości dla kilku jednostek, to wówczas wszystkim tym jednostkom nadaje się taką samą rangę wyznaczoną jako średnia arytmetyczna z rang przypadających na te jednostki*). Na przykład, jeśli kolejne wartości zmiennej opisującej badane jednostki statystyczne, wynoszą odpowiednio: 15; 16; 16; 16; 23; 25, to wówczas, przy założeniu, że 1 oznacza najniższą rangę, otrzymamy następujący ciąg rang: 1, 3, 3, 3, 5, 6. Ranga 3 została wyznaczona jako średnia arytmetyczna z rangi 2, 3 i 4, co daje $(2 + 3 + 4)/3 = 9/3 = 3$.

*)W literaturze (zob. np. Steczkowski J., Zeliaś A., *Statystyczne metody analizy cech jakościowych*, PWE 1981, s. 164) uzyskane tą metodą rangi nazywa się „rangami powiązanymi" (ang. *tied ranks*).

Rangowanie służy głównie porządkowaniu jednostek statystycznych, czyli ich przedstawieniu na skali porządkowej. Jedyna relacja, jaka wówczas zachodzi pomiędzy porządkowanymi obiektami to relacja poprzedzania i następstwa. Jednym z klasycznych przykładów wykorzystywania skali porządkowej są oceny uczniów lub studentów. Jeżeli porządkowanie odbywa się ze względu na dwie cechy opisane przez dwie zmienne (X i Y), to wówczas do zbadania zgodności nadanych ocen (rang) możemy wykorzystać następujący współczynnik*):

$$r_{s.xy} = 1 - \frac{6\sum_{i=1}^{N} d_i^2}{N(N^2-1)}, \tag{4.29}$$

gdzie:

d – różnica pomiędzy rangami zmiennej X i Y,

N – liczba par obserwacji zmiennej X i Y.

Podobnie jak klasyczny współczynnik korelacji liniowej, współczynnik korelacji rang przyjmuje wartości z przedziału [-1;1]. Jeżeli $r_{s.xy} = -1$, to oznacza to pełną przeciwstawność uporządkowań, gdy $r_{s.xy} = 1$, uporządkowania są w pełni zgodne, natomiast, gdy $r_{s.xy} = 0$, to mówimy o całkowitym braku uporządkowań.

Przykład 4.6. Wykorzystując dane z przykładu 4.2 wyznaczyć współczynnik korelacji rang Spearmana pomiędzy liczbą studentów i liczbą stanowisk komputerowych.

W pierwszym kroku realizacjom zmiennych X i Y, przypisujemy rangi w ten sposób, aby największa wartość otrzymała rangę 1, a najmniejsza wartość rangę najwyższą z możliwych. W drugim kroku postępowania, obliczamy kwadraty różnic pomiędzy rangami. Sposób przypisania rang oraz sposób obliczania poszczególnych składowych wzoru (4.29), prezentuje tablica 4.9.

Na podstawie wzoru (4.29) otrzymamy

$$r_{s.xy} = 1 - \frac{6\sum_{i=1}^{N} d_i^2}{N(N^2-1)} = 1 - \frac{6 \cdot 0}{10 \cdot 990} = 1 - 0 = 1.$$

Zauważmy, że współczynnik korelacji obliczony na podstawie powyższego wzoru oznacza korelację zupełną. Różnica pomiędzy obliczonym wcześniej

*)W przypadku wystąpienia „rang powiązanych", zaleca się uwzględnienie podczas obliczeń poprawek modyfikujących wzór 4.29. Sposób postępowania w takiej sytuacji został opisany, np. w pracy wymienionej w poprzednim przypisie.

współczynnikiem korelacji liniowej, a wyznaczonym obecnie wynosi zaledwie niecałe 0,05. Powodem takiej sytuacji jest to, że podczas obliczania współczynnika korelacji rang Spearmana wykorzystuje się dane przestawione na skali porządkowej. Skala ta, co już stwierdzono w rozdziale pierwszym podręcznika, należy do grupy słabych skal pomiarowych.#

Tab. 4.9. Obliczanie współczynnika korelacji rang Spearmana

i	x_i (liczba stud.)	y_i (liczba. stan. komp.)	Rangi zmiennej X	Rangi zmiennej Y	d_i	d_i^2
1	**2**	**3**	**4**	**5**	**6**	**7**
1	4780	24	1	1	0	0
2	4975	40	2	2	0	0
3	7732	122	3	3	0	0
4	9701	177	4	4	0	0
5	12154	198	5	5	0	0
6	14300	203	6	6	0	0
7	15766	214	7	7	0	0
8	17516	224	8	8	0	0
9	18355	254	9	9	0	0
10	18633	254	10	10	0	0
sumy	**123912**	**1710**	**XXX**	**XXX**	**XXX**	**0**

Źródło: obliczenia własne.

4.3. Współzależność liniowa wielu zmiennych

4.3.1. Równanie regresji wielu zmiennych i korelacja wieloraka

W dotychczasowych rozważaniach dotyczących współzależności zakładaliśmy występowanie dwóch zmiennych. Obecnie założenie to zostanie rozszerzone do przypadku, gdy badaniu poddajemy k zmiennych, z których jedna jest zmienną objaśnianą a pozostałe zmiennymi objaśniającymi. Jeżeli zmienną objaśnianą oznaczymy przez Y, a zmienne objaśniające przez $X_1, X_2, \ldots, X_k$, to wówczas liniowa funkcja regresji będzie miała postać:

$$\hat{y}_i = a_0 + a_1 x_{1i} + a_2 x_{2i} + \ldots + a_k x_{ki}. \tag{4.30}$$

Jeżeli liczbę zmiennych ograniczymy do dwóch X_1 i X_2, to wówczas równanie regresji zapiszemy następująco:

$$\hat{y}_i = a_0 + a_1 x_{1i} + a_2 x_{2i}. \tag{4.31}$$

Po zastosowaniu metody najmniejszych kwadratów, otrzymamy następujący układ równań z trzema niewiadomymi:

$$\begin{cases} a_0 N + a_1 \sum_{i=1}^{N} x_{1i} + a_2 \sum_{i=1}^{N} x_{2i} = \sum_{i=1}^{N} y_i, \\ a_0 \sum_{i=1}^{N} x_{1i} + a_1 \sum_{i=1}^{N} x_{i1}^2 + a_2 \sum_{i=1}^{N} x_{1i} x_{2i} = \sum_{i=1}^{N} x_{1i} y_i, \\ a_0 \sum_{i=1}^{N} x_{2i} + a_1 \sum_{i=1}^{N} x_{1i} x_{2i} + a_2 \sum_{i=1}^{N} x_{1i} x_{2i}^2 = \sum_{i=1}^{N} x_{2i} y_i. \end{cases} \tag{4.32}$$

Po rozwiązaniu powyższego układu równań otrzymamy parametry równania: a_0, a_1 i a_2.

Wygodnym sposobem wyznaczania parametrów funkcji regresji jest metoda macierzowa, w której wektor parametrów funkcji regresji ($\boldsymbol{a}$) wyznacza się w oparciu o wzór:

$$\boldsymbol{a} = (\boldsymbol{X}^T \boldsymbol{X})^{-1} \boldsymbol{X}^T \boldsymbol{Y}, \tag{4.33}$$

gdzie $\boldsymbol{Y}$ to wektor realizacji zmiennej objaśnianej Y postaci:

$$\boldsymbol{Y} = \begin{bmatrix} y_1 \\ y_1 \\ \vdots \\ y_i \\ \vdots \\ y_N \end{bmatrix}, \tag{4.34}$$

$\boldsymbol{X}$ zmodyfikowana macierz realizacji zmiennych objaśniających:

$$\boldsymbol{X} = \begin{bmatrix} 1 & x_{11} & x_{21} \\ 1 & x_{12} & x_{22} \\ \vdots & \vdots & \vdots \\ 1 & x_{1i} & x_{2i} \\ \vdots & \vdots & \vdots \\ 1 & x_{1N} & x_{2N} \end{bmatrix}, \tag{4.35}$$

natomiast $\boldsymbol{a}$ jest wektorem parametrów funkcji regresji:

$$\boldsymbol{a} = \begin{bmatrix} a_0 \\ a_1 \\ a_2 \end{bmatrix}. \tag{4.36}$$

Zauważmy, że lewa strona układu równań (4.32) odpowiada iloczynowi macierzy $\boldsymbol{X}^T\boldsymbol{X}$ i wektora parametrów $\boldsymbol{a}$, natomiast prawa strona układu, to $\boldsymbol{X}^T\boldsymbol{X}$. Korzystając ze wzoru (4.24), można wyznaczyć współczynnik determinacji $R^2_{y.x_1x_2}$, z którego po wyciągnięciu pierwiastka kwadratowego otrzymamy współczynnik korelacji ($R_{y.x_1x_2}$) pomiędzy zmienną Y, a zmiennymi objaśniającymi X_1 i X_2. Współczynnik ten określa się mianem współczynnika korelacji wielorakiej.

Przykład 4.7. Poniższe dane ukazują liczbę studentów (Y), liczbę pracowników naukowo-dydaktycznych (X_1) oraz liczbę stanowisk komputerowych (X_2) Akademii Ekonomicznej w Krakowie w latach 1990 – 1999.

Tab. 4.10. Liczba studentów, Liczba pracowników naukowo-dydaktycznych oraz liczba stanowisk komputerowych w dziesięciu badanych okresach

i	Liczba studentów y_i	Liczba pracowników x_{1i}	Liczba stanowisk komputerowych x_{2i}
1	4780	417	24
2	4975	463	40
3	7732	433	122
4	9701	487	177
5	12154	506	198
6	14300	518	203
7	15766	554	214
8	17516	563	224
9	18355	575	254
10	18633	594	254
SUMA	**123912**	**5110**	**1710**

Źródło: *Akademia Ekonomiczna w liczbach*, op. cit.

Oszacować liniową funkcję regresji $\hat{y}_i = a_0 + a_1x_{1i} + a_2x_{2i}$.
W analizowanym przykładzie macierze $\mathbf{X}$ i $\mathbf{Y}$ będą miały postać:

$$\boldsymbol{X} = \begin{bmatrix} 1 & 417 & 24 \\ 1 & 463 & 40 \\ 1 & 433 & 122 \\ 1 & 487 & 177 \\ 1 & 506 & 198 \\ 1 & 518 & 203 \\ 1 & 554 & 214 \\ 1 & 563 & 224 \\ 1 & 575 & 254 \\ 1 & 594 & 254 \end{bmatrix}, \quad \boldsymbol{Y} = \begin{bmatrix} 4780 \\ 4975 \\ 7732 \\ 9701 \\ 12154 \\ 1430 \\ 15766 \\ 17516 \\ 18355 \\ 18633 \end{bmatrix}.$$

Macierze $\boldsymbol{X}^T\boldsymbol{X}$ oraz $\boldsymbol{X}^T\boldsymbol{Y}$ będą przedstawiać się następująco:

$$\boldsymbol{X}^T\boldsymbol{X} = \begin{bmatrix} 10 & 5110 & 1710 \\ 5110 & 2644622 & 914489 \\ 1710 & 914489 & 353806 \end{bmatrix}, \quad \boldsymbol{X}^T\boldsymbol{Y} = \begin{bmatrix} 123912 \\ 66144351 \\ 24975953 \end{bmatrix}.$$

Macierz odwrotna do macierzy $\boldsymbol{X}^T\boldsymbol{X}$ wynosi*):

$$(\boldsymbol{X}^T\boldsymbol{X})^{-1} = -17622,98.$$

Mamy zatem:

$$\begin{aligned} \boldsymbol{a} &= (\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{X}^T\boldsymbol{Y} \\ &= \begin{bmatrix} 25,062401 & -0,061569 & 0,038009 \\ -0,061569 & 0,000155 & -0,000103 \\ 0,038009 & -0,000103 & 0,000084 \end{bmatrix} \cdot \begin{bmatrix} 123912 \\ 66144351 \\ 24975953 \end{bmatrix} = \begin{bmatrix} -17622,98 \\ 48,95 \\ 29,25 \end{bmatrix}. \end{aligned}$$

Równanie funkcji regresji można, więc zapisać następująco:

$$\hat{y}_i = -17622,98 + 48,95x_{1i} + 29,25x_{2i}.$$

Mając funkcję regresji można wyznaczyć współczynnik determinacji oraz współczynnik korelacji wielorakiej. W tym celu musimy dokonać kilku nieodzownych obliczeń, które prezentuje tablica 4.11.

Współczynnik determinacji będzie wynosił:

$$R^2_{y.x_1x_2} = 1 - \frac{\sum_{i=1}^{10}(y_i - \hat{y})^2}{\sum_{i=1}^{10}(y_i - \overline{y})^2} = 1 - \frac{8691088}{257755317,6} = 0.966282 \quad (96,6282\%).$$

Zatem, współczynnik korelacji wielorakiej pomiędzy badanymi zmiennymi wyniesie:

$$R_{y.x_1x_2} = \sqrt{R^2_{y.x_1x_2}} = \sqrt{0,966282} = 0.982996.$$

Wartość obliczonego współczynnika wskazuje na istnienie silnego związku pomiędzy zmienną Y (liczbą studentów), oraz zmiennymi ją objaśniającymi X_1 i X_2 (liczbą pracowników naukowo-dydaktycznych i liczbą stanowisk komputerowych).#

*) Do wyznaczenia macierzy odwrotnej można użyć np. metody wyznacznikowej. Zob. np. J. Kubala, E. Smaga, T. Stanisz, *Elementy algebry liniowej*. PWN, Warszawa 1983, s. 114. Można posłużyć się również jedną ze standardowych funkcji programu Excel: MACIERZ.ODW.

Tab. 4.11. Wyznaczanie współczynnika determinacji

i	Liczba studentów y_i	$\hat{y}_i$	$y_i - \hat{y}_i$	$(y_i - \hat{y}_i)^2$	$y_i - \overline{y}_i$	$(y_i - \overline{y}_i)^2$
1	**2**	**3**	**4**	**5**	**6**	**7**
1	4780	3490,31	1289,69	1663305,0	-7611,2	57930365,44
2	4975	6209,92	-1234,92	1525020,0	-7416,2	55000022,44
3	7732	7140,00	592,00	350464,5	-4659,2	21708144,64
4	9701	11391,95	-1690,95	2859316,0	-2690,2	7237176,04
5	12154	12936,22	-782,22	611863,6	-237,2	56263,84
6	14300	13669,84	630,16	397097,2	1908,8	3643517,44
7	15766	15753,72	12,28	150,8	3374,8	11389275,04
8	17516	16486,76	1029,24	1059345,0	5124,8	26263575,04
9	18355	17951,64	403,36	162700,9	5963,8	35566910,44
10	18633	18881,65	-248,65	61826,1	6241,8	38960067,24
SUMA	**123912**	**XXX**	**XXX**	**8691088,00**	**XXX**	**257755317,60**

Źródło: opracowanie własne.

4.3.2. Korelacja cząstkowa

Dla uproszczenia zapisu załóżmy, że zmienną objaśnianą Y zastąpimy symbolem 0, natomiast zmienne objaśniające X_1 i X_2, odpowiednio symbolami 1 i 2. Uwzględniając nowe założenie możemy zapisać, że:

$$r_{yx_1} = r_{x_1y} = r_{01} = r_{10}, \tag{4.37}$$

$$r_{yx_2} = r_{x_2y} = r_{02} = r_{20}, \tag{4.38}$$

$$r_{x_1x_2} = r_{x_2x_1} = r_{12} = r_{21}, \tag{4.39}$$

$$r_{x_1x_1} = r_{x_2x_2} = r_{yy} = r_{11} = r_{22} = r_{00} = 1. \tag{4.40}$$

Do wyznaczenia wartości współczynników korelacji cząstkowej niezbędna jest znajomość macierzy współczynników korelacji postaci:

$$\boldsymbol{R} = \begin{bmatrix} r_{00} & r_{01} & r_{02} \\ r_{10} & r_{11} & r_{12} \\ r_{20} & r_{21} & r_{22} \end{bmatrix} = \begin{bmatrix} 1 & r_{01} & r_{02} \\ r_{10} & 1 & r_{12} \\ r_{20} & r_{21} & 1 \end{bmatrix}. \tag{4.41}$$

Wyznaczając wartości współczynników korelacji cząstkowej wygodnie jest posłużyć się metodami rachunku macierzowego. W przypadku, gdy analizujemy tylko trzy zmienne mamy:

$$r_{ij.k} = \frac{-R_{ij}}{\sqrt{R_{ii}R_{jj}}}, \tag{4.42}$$

gdzie:
$r_{ij.k}$ – współczynnik korelacji cząstkowej pomiędzy i-tą i j-tą zmienną, przy eliminacji wpływu k-tej zmiennej,
R_{ij}, R_{ii}, R_{jj} – dopełniena algebraicze macierzy $\boldsymbol{R}$,
i, j, k – może przyjmować wartości 1, 2 i 3.

Przypomnijmy, że dopełnieniem algebraicznym elementu r_{ij} macierzy kwadratowej $\boldsymbol{R}$ nazywamy iloczyn wyznacznika macierzy powstałej z macierzy $\boldsymbol{R}$ przez skreślenie w niej i-tego wiersza i j-tej kolumny oraz liczby $(-1)^{i+j}$ (zob. [10], s. 109). W analizowanym przypadku trzech zmiennych, po wykonaniu odpowiednich działań możemy zapisać, że:

$$r_{01.2} = \frac{-R_{01}}{\sqrt{R_{00}R_{11}}} = \frac{r_{01} - r_{02}r_{12}}{\sqrt{(1-r_{02}^2)(1-r_{12}^2)}}, \tag{4.43}$$

$$r_{02.1} = \frac{-R_{02}}{\sqrt{R_{00}R_{22}}} = \frac{r_{02} - r_{01}r_{12}}{\sqrt{(1-r_{01}^2)(1-r_{12}^2)}}, \tag{4.44}$$

$$r_{12.0} = \frac{-R-12}{\sqrt{R_{11}R_{22}}} = \frac{r_{12} - r_{01}r_{02}}{\sqrt{(1-r_{01}^2)(1-r_{02}^2)}}. \tag{4.45}$$

Korzystając z macierzy korelacji można również wyznaczyć wartość współczynnika korelacji wielorakiej ($R_{0.12}$). Stosuje się wówczas wzór:

$$R_{0.12} = \sqrt{1 - \frac{|\boldsymbol{R}|}{R_{11}}} = \sqrt{\frac{r_{01}^2 + r_{02}^2 - 2r_{01}r_{02}r_{12}}{1 - r_{12}^2}}, \tag{4.46}$$

gdzie $|\boldsymbol{R}|$ jest wyznacznikiem macierzy korelacji.

Przykład 4.8. Korzystając z danych z przykładu 4.7, wyznaczyć wartości korelacji cząstkowych oraz wartość korelacji wielorakiej. Aby zbudować macierz korelacji, musimy dokonać kilku niezbędnych obliczeń. Obliczenia te prezentuje tablica robocza 4.12 na stronie 94. Korzystając z nich, otrzymujemy:

$$r_{x_1x_2} = r_{x_2x_1} = r_{12} = r_{21} = \frac{\operatorname{cov}(x_1, x_2)}{S_{x_i} \cdot S_{x_2}} = \frac{4067,9}{57,8031 \cdot 78,3556} = 0,8982,$$

$$r_{yx_1} = r_{x_1y} = r_{01} = r_{10} = \frac{\operatorname{cov}(x_1, y)}{S_{x_i} \cdot S_y} = \frac{282531,9}{57,8031 \cdot 5076,9609} = 0,9627,$$

$$r_{yx_2} = r_{x_2y} = r_{02} = r_{20} = \frac{\operatorname{cov}(x_2, y)}{S_{x_2} \cdot S_y} = \frac{378700,1}{78,3556 \cdot 5076,9609} = 0,9520.$$

Zatem, macierz korelacji będzie miała postać:

$$\boldsymbol{R} = \begin{bmatrix} r_{00} & r_{01} & r_{02} \\ r_{10} & r_{11} & r_{12} \\ r_{20} & r_{21} & r_{22} \end{bmatrix} = \begin{bmatrix} 1,0000 & 0,9627 & 0,9520 \\ 0,9627 & 1,0000 & 0,8982 \\ 0,9520 & 0,8982 & 1,0000 \end{bmatrix}.$$

Z macierzy tej wyznaczamy wartości korelacji cząstkowych:

$$r_{01.2} = \frac{r_{01} - r_{02}r_{12}}{\sqrt{(1 - r_{02}^2)(1 - r_{12}^2)}} = \frac{0,9627 - (0,9520 \cdot 0,8982)}{\sqrt{(1 - 0,9520^2) \cdot (1 - 0,8982^2)}} = 0,8002,$$

$$r_{02.1} = \frac{r_{02} - r_{01}r_{12}}{\sqrt{(1 - r_{01}^2)(1 - r_{12}^2)}} = \frac{0,9520 - (0,9627 \cdot 0,8982)}{\sqrt{(1 - 0,9627^2) \cdot (1 - 0,8982^2)}} = 0,7341,$$

$$r_{12.0} = \frac{r_{12} - r_{01}r_{02}}{\sqrt{(1 - r_{01}^2)(1 - r_{02}^2)}} = \frac{0,8982 - (0,9627 \cdot 0,9520)}{\sqrt{(1 - 0,9627^2) \cdot (1 - 0,9520^2)}} = -0,2217.$$

Korelacja pomiędzy liczbą studentów uczelni, a liczbą jej pracowników, przy eliminacji wpływu liczby stanowisk komputerowych jest wysoka i wynosi 0,8. Podobną wysoką korelację, wynoszącą 0,7341, zauważyć można pomiędzy liczbą studentów i liczbą stanowisk komputerowych przy eliminacji wpływu liczby pracowników naukowo-dydaktycznych. Niska korelacja ujemna ($-0,277$) występuje natomiast pomiędzy liczbą pracowników i liczbą stanowisk komputerowych, przy eliminacji wpływu liczby studentów.
Współczynnik korelacji wielorakiej będzie wynosił:

$$\begin{aligned} R_{0.12} &= \sqrt{\frac{r_{01}^2 + r_{02}^2 - 2r_{01}r_{02}r_{12}}{1 - r_{12}^2}} \\ &= \sqrt{\frac{0,9627^2 + 0,9520^2 - 2 \cdot (0,9627 \cdot 0,9520 \cdot 0,8982)}{1 - 0,8982^2}} \\ &= 0,982996. \end{aligned}$$

Współczynnik ten świadczy o wysokiej współzależności pomiędzy liczbą studiujących studentów, a zespołem zmiennych ją objaśniających (liczbą pracowników naukowo-dydaktycznych i liczbą stanowisk komputerowych). #

Tab. 4.12. Wyznaczenie macierzy korelacji

i	Liczba studentów y_i	Liczba pracowników x_{1i}	Liczba stan. komp. x_{2i}	$x_{1i}-\overline{x}_1$	$x_{2i}-\overline{x}_2$	$y_i-\overline{y}$	$(x_{1i}-\overline{x}_1)^2$	$(x_{2i}-\overline{x}_2)^2$	$(y_i-\overline{y})^2$	$(x_{1i}-\overline{x}_1)\cdot(x_{2i}-\overline{x}_2)$	$(x_{1i}-\overline{x}_1)\cdot(y_i-\overline{y})$	$(x_{2i}-\overline{x}_2)\cdot(y_i-\overline{y})$
1	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**	**10**	**11**	**12**	**13**
1	4780	417	24	-94	-147	-7611,2	8836	21609	57930365,44	13818	715452,8	1118846,0
2	4975	463	40	-48	-131	-7416,2	2304	17161	55000022,44	6288	355977,6	971522,2
3	7732	433	122	-78	-49	-4659,2	6084	2401	21708144,64	3822	363417,6	228300,8
4	9701	487	177	-24	6	-2690,2	576	36	7237176,04	-144	64564,8	-16141,2
5	12154	506	198	-5	27	-237,2	25	729	56263,84	-135	1186,0	-6404,4
6	14300	518	203	7	32	1908,8	49	1024	3643517,44	224	13361,6	61081,6
7	15766	554	214	43	43	3374,8	1849	1849	11389275,04	1849	145116,4	145116,4
8	17516	563	224	52	53	5124,8	2704	2809	26263575,04	2756	266489,6	271614,4
9	18355	575	254	64	83	5963,8	4096	6889	35566910,44	5312	381683,2	494995,4
10	18633	594	254	83	83	6241,8	6889	6889	38960067,24	6889	518069,4	518069,4
Σ	123912	5110	1710	0	0	0,00	33412	61396	257755317,6	40679	2825319,0	3787001,0

Źródło: obliczenia własne.

Analiza szeregów czasowych

Rozdział 5

Każde zjawisko: społeczne, ekonomiczne czy fizyczne występuje zawsze w określonym czasie. Dość częstą kwestią badań statystycznych jest analiza określonego zjawiska na tle upływającego czasu. Zmienną opisującą czas najczęściej oznacza się symbolem T, a jej wartości odpowiednio $t_1, t_2, t_3, \ldots$

Przypomnijmy (z pierwszego rozdziału podręcznika) że ten typ szeregu statystycznego, w którym momentom czasu lub jego okresom przyporządkowane są realizacje zmiennej (zmiennych) opisującej badane zjawisko, nazywa się szeregiem czasowym (chronologicznym, rozwojowym lub dynamicznym).

Przedmiotem dalszej części tego rozdziału będzie sposób opisu i analizy szeregów czasowych.

5.1. Przyrosty i indeksy indywidualne

Wśród przyrostów rozróżnia się:

— przyrosty absolutne,

— przyrosty względne.

Przyrosty absolutne i względne dzielą się na:

— jednopodstawowe (o stałej podstawie),

— łańcuchowe.

Przyrost absolutny jednopodstawowy obliczamy według wzoru:

$$\Delta_{t/t_b} = y_t - y_{t_b}, \tag{5.1}$$

gdzie:
y_t – wartość zmiennej Y w okresie badanym t,
y_{t_b} – wartość zmiennej Y w okresie bazowym.

Okres bazowy może być dowolnie wybranym okresem t, dla $t = 1, 2, .., N$. Przyrost absolutny łańcuchowy wyznacza się w oparciu o wzór:

$$\Delta_{t/(t-1)} = y_t - y_{t-1}, \tag{5.2}$$

gdzie:
y_t i y_{t-1} – wartości zmiennej Y odpowiednio w okresach t i $t-1$.

Przyrosty absolutne informują, o ile jednostek zmienił się (wzrósł lub zmalał) poziom zjawiska w okresie badanym t w stosunku do okresu przyjętego za podstawę (t_b lub $t-1$).

Jeżeli przyrost absolutny jednopodstawowy obliczany w okresie t, podzielimy przez wartość zmiennej Y w okresie bazowym t_b, to wówczas otrzymamy odpowiednio przyrost względny jednopodstawowy:

$$\delta_{t/t_b} = \frac{y_t - y_{t_b}}{y_{t_b}} \cdot 100\ [\%] = \frac{\Delta_{t/t_b}}{y_{t_b}} \cdot 100\ [\%]. \tag{5.3}$$

Natomiast, jeśli przyrost absolutny łańcuchowy w okresie t, podzielimy przez wartość zmiennej Y w okresie $t-1$, to wówczas dostaniemy przyrost względny łańcuchowy:

$$\delta_{t/t-1} = \frac{y_t - y_{t-1}}{y_{t-1}} \cdot 100\ [\%] = \frac{\Delta_{t/t-1}}{y_{t-1}} \cdot 100\ [\%]. \tag{5.4}$$

Badając dynamikę zjawisk można posłużyć się także indeksami indywidualnymi. Podobnie jak w przypadku przyrostów wyróżnić można dwa rodzaje indeksów indywidualnych:

— indeksy indywidualne jednopodstwowe,

— indeksy indywidualne łańcuchowe.

Pierwszy z nich – indeks indywidualny jednopodstawowy – obliczamy ze wzoru:

$$\gamma_{t/t_b} = \frac{y_t}{y_{t_b}} \cdot 100\ [\%], \tag{5.5}$$

natomiast do wyznaczenia indeksu indywidualnego łańcuchowego używa się wzoru:

$$\gamma_{t/t-1} = \frac{y_t}{y_{t-1}} \cdot 100\ [\%]. \tag{5.6}$$

Symbole użyte w powyższych wzorach mają identyczne znaczenie, jak w przypadku wzorów przyrostów. Zauważmy, że pomiędzy przyrostem względnym

a indeksem indywidualnym, wyrażonych w procentach, zachodzą następujące relacje:

$$\gamma_{t/t_b} = \delta_{t/t_b} + 100\ [\%], \tag{5.7}$$

oraz

$$\gamma_{t/t-1} = \delta_{t/t-1} + 100\ [\%]. \tag{5.8}$$

Wartość indeksu większa od 100%, odpowiada wyższemu poziomowi zjawiska (procesu) w okresie badanym w porównaniu z okresem podstawowym, natomiast wartość mniejsza od 100%, oznacza spadek wartości poziomu procesu w okresie badanym w porównaniu z okresem podstawowym. Na przykład, jeżeli wartość indeksu indywidualnego łańcuchowego wynosi 115%, to oznacza to, że w okresie t w stosunku do okresu $t-1$ nastąpił przyrost badanego zjawiska o 15%. Jeżeli, natomiast indeks ten wynosiłby 95%, to oznacza to, że wartość badanego zjawiska zmalała w okresie badanym w stosunku do okresu poprzedniego o 5%.

Jeżeli zachodzi konieczność określenia przeciętnego tempa zmian badanego zjawiska w całym objętym badaniem przedziale czasowym, to wówczas można posłużyć się średnią geometryczną z indeksów łańcuchowych. Do tego celu używa się wzoru:

$$\overline{\gamma}_g = \sqrt[N-1]{\gamma_{2/1} \cdot \gamma_{3/2} \cdot \gamma_{4/3} \cdot \ldots \cdot \gamma_{N/N-1}} \tag{5.9}$$

lub (po jego przekształceniu) wzoru:

$$\log \overline{\gamma}_g = \frac{\log \gamma_{2/1} + \log \gamma_{3/2} + \log \gamma_{4/3} + \ldots + \log \gamma_{N/N-1}}{N-1}, \tag{5.10}$$

gdzie: N – liczba obserwacji.

Średnia ta jest stosowana przede wszystkim w sytuacji, gdy zachodzi konieczność uśrednienia wartości charakteryzujących stosunki pomiędzy dwiema wielkościami. Przykładem takim są właśnie wartości indeksów indywidualnych łańcuchowych. Średnia wartość $(\overline{\gamma}_g)$ indeksu (γ) powinna posiadać następującą własność:

$$y_N = y_1 \cdot \gamma_g^{N-1}, \tag{5.11}$$

której nie posiada średnia arytmetyczna.

Przykład 5.1. Liczba zarejestrowanych bezrobotnych [w tys.] w Polsce, w kolejnych miesiącach 2001 roku wynosiła odpowiednio:

Tab. 5.1. Dane robocze do przykładu 5.1

miesiąc	I	II	III	IV	V	VI	VII	VIII	IX	X	XI	XII
liczba bezrobotnych	2836	2877	2899	2878	2841	2849	2872	2893	2920	2944	3022	3115

Źródło: Rocznik Statystyczny 2002.

Wyznaczyć wartości przyrostów absolutnych i względnych oraz obliczyć wartości indeksów indywidualnych. Wyznaczyć średnie miesięczne tempo przyrostu bezrobocia.

Tab. 5.2. Tablica robocza

			Przyrosty absolutne		Przyrosty względne		Indeksy indwywidualne		
Miesiąc	t	Bezrobotni w tys. y_i	Δ_{t/t_b} $t_b = t_1$	$\Delta_{t/(t-1)}$	δ_{t/t_b}	$\delta_{t/t-1}$	γ_{t/t_b}	$\gamma_{t/t-1}$	$\log(\gamma_{t/t-1})$
1	2	3	4	5	6	7	8	9	10
I	1	2836	–	–	–	–	100,00%	–	–
II	2	2877	41	41	1,45%	1,45%	101,45%	101,45%	2,0062
III	3	2899	63	22	2,22%	0,76%	102,22%	100,76%	2,0033
IV	4	2878	42	-21	1,48%	-0,72%	101,48%	99,28%	1,9968
V	5	2841	5	-37	0,18%	-1,29%	100,18%	98,71%	1,9944
VI	6	2849	13	8	0,46%	0,28%	100,46%	100,28%	2,0012
VII	7	2872	36	23	1,27%	0,81%	101,27%	100,81%	2,0035
VIII	8	2893	57	21	2,01%	0,73%	102,01%	100,73%	2,0032
IX	9	2920	84	27	2,96%	0,93%	102,96%	100,93%	2,0040
X	10	2944	108	24	3,81%	0,82%	103,81%	100,82%	2,0036
XI	11	3022	186	78	6,56%	2,65%	106,56%	102,65%	2,0114
XII	12	3115	279	93	9,84%	3,08%	109,84%	103,08%	2,0132
Źródło: obliczenia własne.									$\Sigma = 22,0408$

Z powyższych obliczeń wynika, że liczba bezrobotnych w grudniu w stosunku do stycznia wzrosła o 279 tys. osób (co stanowi 9,84%). Największy wzrost bezrobocia przypadł na przełom listopada i grudnia. W grudniu liczba bezrobotnych w stosunku do listopada wzrosła o 93 tys. osób (2,99%). Spadek liczby bezrobotnych o 21 osób zanotowano na przełomie marca i kwietnia oraz o 37 osób, w maju w stosunku do kwietnia. Średni miesięczny przyrost (tempo) bezrobocia wyznaczamy w oparciu o wzór (5.10) i otrzymujemy:

$$\log \overline{\gamma}_g = \frac{22,0408}{11} = 2,003705,$$

stąd

$$\overline{\gamma}_g = 10^{2,003705} = 100,8567.$$

Średnie miesięczne tempo wzrostu bezrobocia wynosi, zatem ok. 0,86%. Wykorzystując wartość obliczonej średniej można sprawdzić, czy prawdziwa jest własność (5.11). Otrzymamy wówczas:

$$\tilde{y}_{12} = y_1 \cdot \overline{\gamma}_g^{11} = 2836 \cdot 1,00856^{11} = 3114,7656.$$

Otrzymana wartość $\tilde{y}_{12}$ jest oceną wartości y_{12}. Ocena ta jest obarczona błędem wynoszącym:

$$\Delta y_{12} = \tilde{y}_{12} - y_{12} = 3114,7656 - 3115 = -0,2344.$$

Wielkość błędu oceny jest na pewno mniejsza niż w sytuacji, gdyby do uzyskania średniego przyrostu użyć średniej arytmetycznej. Średnia arytmetyczna indeksów indywidualnych łańcuchowych będzie wynosić:

$$\overline{\gamma} = \frac{1,0145 + 1,0076 + \ldots + 1,0265 + 1,0308}{11} = 1,008639,$$

natomiast ocena i błąd oceny wyniosą:

$$\tilde{y}_{12} = y_1 \cdot \overline{\gamma}_g^{11} = 2836 \cdot 1,008639^{11} = 3117,44,$$

$$\Delta y_{12} = \tilde{y}_{12} - y_{12} = 3117,4400 - 3115 = 2,44.$$

Z powyższych obliczeń wynika, że średnia arytmetyczna jest gorszą oceną średniego poziomu dynamiki badanego w tym przykładzie zjawiska.

5.2. Indeksy agregatowe

W wielu przypadkach interesujemy się zmianami określonych grup (agregatów) zjawisk. Przykładem takich agregatów mogą być koszyki dóbr czy usług zaspokajających określoną grupę potrzeb. Wśród agregatowych wskaźników dynamiki wyróżnia się zwykle:

— agregatowy indeks wartości,

— agregatowy indeks masy towarowej (ilości),

— agregatowy indeks cen.

Agregatowy indeks wartości jest stosunkiem sumy wartości określonego agregatu dóbr w okresie badanym do sumy określonego agregatu dóbr w okresie podstawowym:

$$I_w = \frac{\sum_{i=1}^{N} q_{1.i} \cdot p_{1.i}}{\sum_{i=1}^{N} q_{0.i} \cdot p_{0.i}} \cdot 100 \ [\%], \tag{5.12}$$

gdzie:
$q_{1.i}$ – ilość dobra i w okresie badanym,
$q_{0.i}$ – ilość dobra i w okresie podstawowym,
$p_{1.i}$ – cena dobra i w okresie badanym,
$p_{0.i}$ – cena dobra i w okresie podstawowym.

Wartość tego indeksu informuje badacza, o ile wzrosła wartość określonego agregatu dóbr w okresie badanym w stosunku do okresu podstawowego.

Aby dowiedzieć się, w jakiej części zmiana wartości była spowodowana zmianą cen, a w jakiej zmianą ilości dóbr należy wyznaczyć agregatowe indeksy cen i ilości.

Jednym z najczęściej stosowanych indeksów agregatowych cen jest indeks cen Laspeyresa obliczany przy założeniu, że ilość dóbr pozostaje ustalona na poziomie z okresu podstawowego. Oznacza to, że ceny dóbr w obydwu jednostkach czasu są ważone ilościami (masą towarową) dóbr z okresu podstawowego. Agregatowy indeks cen Laspeyresa wyznaczamy stosując wzór:

$$I_{P.L} = \frac{\sum_{i=1}^{N} q_{0.i} \cdot p_{1.i}}{\sum_{i=1}^{N} q_{0.i} \cdot p_{0.i}} \cdot 100\ [\%]. \tag{5.13}$$

Wartość powyższego indeksu informuje nas, o ile procent wzrośnie lub zmaleje – w okresie badanym – wartość dóbr, na wskutek zmiany cen, przy założeniu, że ilość dóbr pozostanie na poziomie okresu podstawowego.

Inną propozycję indeksu agregatowego cen sformułował Paasche zakładając, że ilość sprzedanych lub skonsumowanych dóbr kształtować się będzie na poziomie okresu badanego. Agregatowy indeks cen Paascheg opisuje następujący wzór:

$$I_{P.P} = \frac{\sum_{i=1}^{N} q_{1.i} \cdot p_{1.i}}{\sum_{i=1}^{N} q_{1.i} \cdot p_{0.i}} \cdot 100\ [\%]. \tag{5.14}$$

Wartość indeksu wskazuje, w jakim stopniu zmiana cen spowodowała zmianę wartość dóbr w okresie badanym, jeżeli ilość dóbr zostanie ustalona na poziomie z okresu badanego.

Laspeyres i Paasche zdefiniowali również indeksy agregatowe mierzące wpływ zmiany masy towarowej na wartość sprzedawanych lub konsumowanych dóbr. Jeżeli cena dóbr zostanie ustalona na poziomie okresu podstawowego, to

wówczas otrzymamy agregatowy indeks ilości Laspeyresa:

$$I_{q.L} = \frac{\sum_{i=1}^{N} q_{1.i} \cdot p_{0.i}}{\sum_{i=1}^{N} q_{0.i} \cdot p_{0.i}} \cdot 100 \ [\%]. \tag{5.15}$$

Indeks ten informuje badacza, w jakim stopniu zmiana wartości dóbr w okresie badanym zależy od zmiany ilości dóbr, przy założeniu, że ceny tych dóbr będą na poziomie z okresu podstawowego.

Jeżeli założymy stałość cen na poziomie okresu badanego, to wówczas otrzymamy agregatowy indeks ilości Paaschego:

$$I_{q.P} = \frac{\sum_{i=1}^{N} q_{1.i} \cdot p_{1.i}}{\sum_{i=1}^{N} q_{0.i} \cdot p_{1.i}} \cdot 100 \ [\%], \tag{5.16}$$

który informuje nas, w jakim stopniu zmiana ilości dóbr spowoduje zmianę wartości dóbr w okresie badanym, jeżeli założymy, że ceny dóbr są na poziomie z okresu badanego.

Mając indeksy Laspeyresa i Paaschego agregatowe można wyznaczyć indeksy cen i ilości Fishera:

$$I_{P.F} = \sqrt{I_{P.L} \cdot I_{P.P}}, \tag{5.17}$$

oraz

$$I_{q.F} = \sqrt{I_{q.L} \cdot I_{q.P}}. \tag{5.18}$$

Wyznaczając wartości powyższych indeksów dowiemy się, ile wynosił średni wzrost lub spadek cen oraz ilości rozważanego agregatu dóbr.

Przykład 5.2. Ceny oraz liczby słuchaczy na czterech kierunkach studiów podyplomowych, jednej z wyższych uczelni, w roku akademickim t_0 (2000/2001) oraz w roku t_1 (2001/2002) przedstawia tablica 5.3.. Na podstawie tych danych obliczyć i zinterpretować agregatowy indeks wartości oraz agregatowe indeksy Laspeyresa, Paaschego i Fischera.

Tab. 5.3. Liczba słuchaczy oraz cena studiów w okresie podstawowym t_0 i okresie badanym t_1

	2000/2001 (t_0)		2001/2002 (t_1)	
Kierunek studiów (i)	$q_{0.i}$	$p_{0.i}$	$q_{1.i}$	$p_{1.i}$
1	31	4 000,00	40	4 200,00
2	37	4 000,00	41	4 300,00
3	73	4 100,00	102	4 400,00
4	36	4 000,00	39	4 100,00

Źródło: badania własne.

Tab. 5.4. Obliczenia pomocnicze do przykładu 5.2

	2000/2001 (t_0)		2001/2002 (t_1)					
Kierunek studiów (i)	$q_{0.i}$	$p_{0.i}$	$q_{1.i}$	$p_{1.i}$	$q_{1.i}p_{1.i}$	$q_{0.i}p_{0.i}$	$q_{0.i}p_{1.i}$	$q_{1.i}p_{0.i}$
1	31	4 000,00	40	4 200,00	168000	124000	130200	160000
2	37	4 000,00	41	4 300,00	176300	148000	159100	164000
3	73	4 100,00	102	4 400,00	448800	299300	321200	418200
4	36	4 000,00	39	4 100,00	159900	144000	147600	156000
suma	**XXX**	**XXX**	**XXX**	**XXX**	**953000**	**715300**	**758100**	**898200**

Źródło: obliczenia własne.

Wykorzystując obliczenia pomocnicze zestawione w tablicy 5.4. oraz odpowiednie wzory otrzymamy:

agregatowy indeks wartości:

$$I_w = \frac{\sum_{i=1}^{4} q_{1.i} \cdot p_{1.i}}{\sum_{i=1}^{4} q_{0.i} \cdot p_{0.i}} \cdot 100\ [\%] = \frac{953000}{715300} \cdot 100\ [\%] = 133,23\ [\%].$$

Wartość usługi wzrosła w okresie badanym, w stosunku do okresu podstawowego o 33,23%.

— agregatowe indeksy cen Laspeyresa:

$$I_{P.L} = \frac{\sum_{i=1}^{4} q_{0.i} \cdot p_{1.i}}{\sum_{i=1}^{4} q_{0.i} \cdot p_{0.i}} \cdot 100\ [\%] = \frac{758100}{715300} \cdot 100\ [\%] = 105,98\ [\%].$$

Przy przyjęciu liczby słuchaczy z okresu podstawowego na wskutek zmiany cen – wartość usługi wzrosłaby w okresie badanym średnio o 5,98%.

— agregatowe indeksy cen Paaschego:

$$I_{P.P} = \frac{\sum_{i=1}^{4} q_{1.i} \cdot p_{1.i}}{\sum_{i=1}^{4} q_{1.i} \cdot p_{0.i}} \cdot 100\ [\%] = \frac{953000}{898200} \cdot 100\ [\%] = 106,10\ [\%].$$

Przy przyjęciu liczby słuchaczy z okresu badanego na wskutek zmiany cen – wartość usługi wzrosłaby w okresie badanym średnio o 6,10%.

— agregatowy indeks ilości Laspeyresa:

$$I_{q.L} = \frac{\sum_{i=1}^{4} q_{1.i} \cdot p_{0.i}}{\sum_{i=1}^{4} q_{0.i} \cdot p_{0.i}} \cdot 100\ [\%] = \frac{898200}{715300} \cdot 100\ [\%] = 125,57\ [\%].$$

Jeżeli przyjmiemy ceny za studia z okresu podstawowego, to wówczas na wskutek wzrostu liczby studentów – wartość usługi wzrośnie średnio o 25,57%.

— agregatowy indeks ilości Paaschego:

$$I_{q.P} = \frac{\sum_{i=1}^{4} q_{1.i} \cdot p_{1.i}}{\sum_{i=1}^{4} q_{0.i} \cdot p_{1.i}} \cdot 100\ [\%] = \frac{953000}{758100} \cdot 100\ [\%] = 125,71\ [\%].$$

Jeżeli przyjmiemy ceny z okresu badanego, to wówczas na wskutek wzrostu liczby studentów – wartość usługi wzrośnie średnio o 25,71%

— agregatowe indeksy Fishera:

$$I_{P.F} = \sqrt{I_{P.L} \cdot I_{P.P}} = \sqrt{105,98 \cdot 106,10} = 106,04\ [\%]$$

oraz

$$I_{q.F} = \sqrt{I_{q.L} \cdot I_{q.P}} = \sqrt{125,57 \cdot 125,71} = 125,64\ [\%].$$

W badanym okresie ceny studiów charakteryzowały się średnio wzrostem o 6,04%, natomiast liczba studentów wzrosła średnio o 25,64%.

Zauważmy również, że pomiędzy indeksem wartości oraz agregatowymi indeksami cen i ilości zachodzą następujące związki:

$$I_w = I_{P.P} \cdot I_{q.L} = 1,0610 \cdot 1,2557 = 1,3323,$$

$$I_w = I_{P.L} \cdot I_{q.P} = 1,0598 \cdot 1,2571 = 1,3323.$$

5.3. Wyznaczanie tendencji rozwojowych

Wyznaczenie tendencji rozwojowych sprowadza się do eliminacji (stłumienia) wpływu wahań przypadkowych (losowych) oraz wahań okresowych. Proces ten nazywa się wygładzaniem szeregów czasowych. Do podstawowych metod wygładzania szeregów czasowych należą: metoda średnich ruchomych (metoda mechaniczna) oraz metoda najmniejszych kwadratów (metoda analityczna).

5.3.1. Metoda średnich ruchomych

Metoda średnich ruchomych sprowadza się do zastąpienia wartości empirycznych $y_1, y_2, \ldots, y_t, \ldots, y_N$ odpowiednimi średnimi $\overline{y}_t$. Sposób wyznaczania średniej zależy od liczby okresów, z których jest ona wyznaczana oraz od tego, czy liczba ta jest parzysta czy nieparzysta. Średnią ruchomą dla nieparzystej liczby okresów – na przykład trzech – oblicza się następująco:

$$\begin{aligned} \overline{y}_2 &= \frac{y_1 + y_2 + y_3}{3}, \\ \overline{y}_3 &= \frac{y_2 + y_3 + y_4}{3}, \\ &\vdots \\ \overline{y}_t &= \frac{y_{t-1} + y_t + y_{t+1}}{3}. \end{aligned} \tag{5.19}$$

Nieco trudniej wyznacza się średnią ruchomą w sytuacji, gdy liczba okresów jest parzysta. W przypadku parzystej liczby okresów pojawia się problem z przypisaniem uzyskanej wartości średniej. Na przykład, jeżeli liczba okresów wynosi 4, to wówczas pierwszy wynik obliczonej średniej należałoby przypisać nie istniejącemu punktowi środkowemu między t_2 i t_3. Aby uniknąć takiej sytuacji stosuje się zabieg centrowania średnich. Średnią ruchomą czterookresową centrowaną obliczamy wówczas następująco:

$$\begin{aligned} \overline{y}_3 &= \frac{1}{4}\left[\frac{1}{2}y_1 + y_2 + y_3 + y_4 + \frac{1}{2}y_5\right], \\ \overline{y}_4 &= \frac{1}{4}\left[\frac{1}{2}y_2 + y_3 + y_4 + y_5 + \frac{1}{2}y_6\right], \\ &\vdots \\ \overline{y}_t &= \frac{1}{4}\left[\frac{1}{2}y_{t-2} + y_{t-1} + y_t + y_{t+1} + \frac{1}{2}y_{t+2}\right]. \end{aligned} \tag{5.20}$$

Przykład 5.3. Kurs akcji Banku Handlowego S.A. w 21 kolejnych sesjach giełdowych prezentuje tablica 5.5.

Tab. 5.5. Kurs (cena zamknięcia) akcji Banku Handlowego S.A.

t	Data	Kurs zamknięcia	t	Data	Kurs zamknięcia
1	16.11.01	58,20	12	03.12.01	60,30
2	19.11.01	58,60	13	04.12.01	61,00
3	20.11.01	58,80	14	05.12.01	60,60
4	21.11.01	59,00	15	06.12.01	60,50
5	22.11.01	59,10	16	07.12.01	61,50
6	23.11.01	58,50	17	10.12.01	62,00
7	26.11.01	59,50	18	11.12.01	61,20
8	27.11.01	60,00	19	12.12.01	61,00
9	28.11.01	60,00	20	13.12.01	61,00
10	29.11.01	60,10	21	14.12.01	61,00
11	30.11.01	60,20	–	–	–

Źródło: www://penetrator.pl

Na podstawie powyższych danych należy:

— wyznaczyć tendencję rozwojową za pomocą średniej ruchomej 3-okresowej, 5-okresowej oraz 4-okresowej,

— przedstawić szereg empiryczny i szeregi wygładzone na wykresie.

Średnią ruchomą dla trzech okresów obliczamy następująco:

$$\begin{aligned}
\overline{y}_2 &= \frac{58,20+58,60+58,80}{3} = 58,53,\\
\overline{y}_3 &= \frac{58,60+58,80+59,00}{3} = 58,80,\\
\overline{y}_4 &= \frac{58,80+59,00+59,10}{3} = 58,97,\\
&\vdots\\
\overline{y}_{20} &= \frac{61,00+61,00+61,00}{3} = 61,00.
\end{aligned}$$

Obliczamy średnią ruchomą dla pięciu okresów:

$$\begin{aligned}
\overline{y}_3 &= \frac{58,20+58,60+58,80+59,00+59,10}{5} = 58,74,\\
\overline{y}_4 &= \frac{58,60+58,80+59,00+59,10+58,50}{5} = 58,80,\\
&\vdots\\
\overline{y}_{19} &= \frac{62,00+61,20+61,00+61,00+61,00}{5} = 61,24.
\end{aligned}$$

Natomiast, średnią ruchomą 4-okresową centrowaną wyznaczmy następująco:

$$\overline{y}_3 = \frac{\frac{1}{2}\cdot 58,20 + 58,60 + 58,80 + 59,00 + \frac{1}{2}\cdot 59,10}{4} = 58,76,$$

$$\overline{y}_4 = \frac{\frac{1}{2}\cdot 58,60 + 58,80 + 59,00 + 59,10 + \frac{1}{2}\cdot 58,50}{4} = 58,86,$$

$$\vdots$$

$$\overline{y}_{19} = \frac{\frac{1}{2}\cdot 62,00 + 61,20 + 61,00 + 61,00 + \frac{1}{2}\cdot 61,00}{4} = 61,24.$$

Te i pozostałe wartości średnich ruchomych zostały zamieszczone w tablicy 5.6.

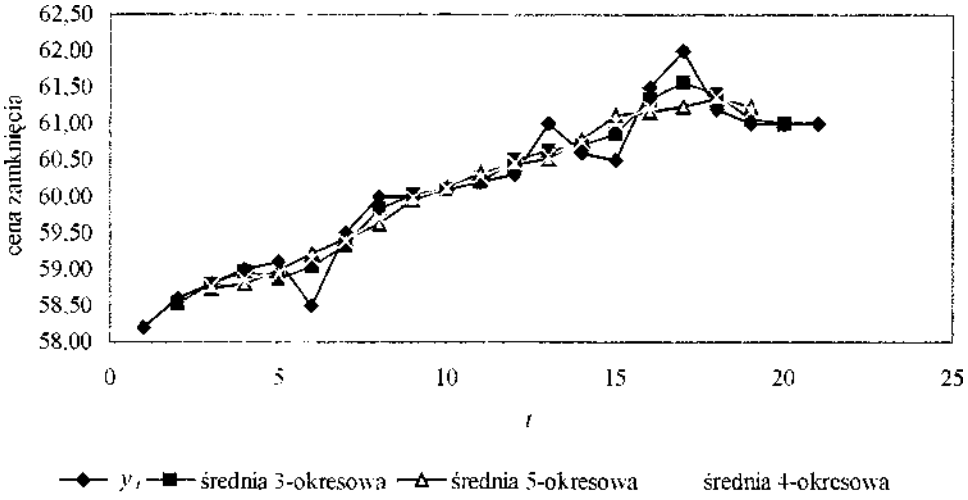
Tab. 5.6. Wygładzanie szeregu metodą średnich ruchomych

t	Data (dd.mm.rr)	Cena zamknięcia y_t	Średnia. 3-okresowa	Średnia 4-okresowa	Średnia 4-okresowa centrowana
1	**2**	**3**	**4**	**5**	**6**
1	16.11.01	58,20	–	–	–
2	19.11.01	58,60	58,53	–	–
3	20.11.01	58,80	58,80	58,74	58,76
4	21.11.01	59,00	58,97	58,80	58,86
5	22.11.01	59,10	58,87	58,98	58,94
6	23.11.01	58,50	59,03	59,22	59,15
7	26.11.01	59,50	59,33	59,42	59,39
8	27.11.01	60,00	59,83	59,62	59,70
9	28.11.01	60,00	60,03	59,96	59,99
10	29.11.01	60,10	60,10	60,12	60,11
11	30.11.01	60,20	60,20	60,32	60,28
12	03.12.01	60,30	60,50	60,44	60,46
13	04.12.01	61,00	60,63	60,52	60,56
14	05.12.01	60,60	60,70	60,78	60,75
15	06.12.01	60,50	60,87	61,12	61,03
16	07.12.01	61,50	61,33	61,16	61,23
17	10.12.01	62,00	61,57	61,24	61,36
18	11.12.01	61,20	61,40	61,34	61,36
19	12.12.01	61,00	61,07	61,24	61,18
20	13.12.01	61,00	61,00	–	–
21	14.12.01	61,00	–	–	–

Źródło: obliczenia własne.

Zauważmy, że im średnia ruchoma obliczana jest z dłuższego okresu, tym nowy szereg wygładzony jest krótszy. I tak, przy średniej ruchomej 3-okresowej tracimy 2 obserwacje, 4- i 5-okresowej – 4 obserwacje, 7-okresowej, 6 obserwacji itd.

Szeregi empiryczne i wyrównane można również przedstawić graficznie.



Rys. 5.1. Graficzna prezentacja wyznaczania tendencji rozwojowej metodą średnich ruchomych
Źródło: opracowanie własne.

Analiza powyższego rysunku nasuwa wniosek, że tendencja kursu badanych akcji jest rosnąca i zbliżona do liniowej. Nieznaczne wahania świadczą o oddziaływaniu czynnika losowego oraz o innych przyczynach zakłócających względnie równomierny przebieg rozwoju zjawiska w czasie. Można zatem próbować oszacować postać funkcji liniowej opisującej badane zjawisko. Funkcja taka nosi nazwę funkcji trendu.

5.3.2. Metoda analityczna

Najprostszą spośród metod umożliwiających oszacowanie parametrów funkcji trendu jest metoda najmniejszych kwadratów (poznana we wcześniejszym rozdziale). Załóżmy, że obrazem tendencji rozwojowej będzie funkcja liniowa postaci:

$$\hat{y}_t = at + b,$$

gdzie:

t - zmienna czasowa $t = 1, \ldots, N$,

a, b – parametry funkcji trendu.

Stosując metodę najmniejszych kwadratów otrzymamy układ równań:

$$\begin{cases} a\sum_{i=1}^{N} t + bN = \sum_{i=1}^{N} y_i, \\ a\sum_{i=1}^{N} t_i^2 + b\sum_{i=1}^{N} t = \sum_{i=1}^{N} x_i y_i, \end{cases} \tag{5.21}$$

który należy rozwiązać ze względu na parametry a i b. Otrzymamy wówczas:

$$a = \frac{N\sum_{i=1}^{N} y_t t - \sum_{i=1}^{N} y_t \sum_{i=1}^{N} t}{N\sum_{i=1}^{N} t^2 - \left(\sum_{i=1}^{N} t\right)^2}, \tag{5.22}$$

$$b = \frac{\sum_{i=1}^{N} y_t - a\sum_{i=1}^{N} t}{N}. \tag{5.23}$$

Parametr a – nazywany **współczynnikiem trendu** – wskazuje, jaki jest przeciętny wzrost lub spadek badanego zjawiska w jednostce czasu t. Natomiast parametr b, określa poziom badanego zjawiska w okresie (momencie) $t = 0$.

Aby zmierzyć dobroć dopasowania funkcji trendu można wykorzystać odchylenie standardowe składnika resztowego postaci:

$$s_\varepsilon = \sqrt{\frac{\sum_{i=1}^{N} (y_t - \hat{y}_t)^2}{N}}, \tag{5.24}$$

którego wartość informuje nas o ile średnio będą się odchylać wartości empiryczne szeregu od wartości hipotetycznych, wyznaczonych w oparciu o funkcję trendu.

Przykład 5.4. Korzystając z danych z przykładu 5.3 dopasować liniową funkcję trendu oraz zbadać dobroć dopasowania.

Tab. 5.7. Wyznaczanie liniowej funkcji trendu

Data (dd.mm.rr)	t	Cena zamknięcia y_t	$y_t t$	t^2
1	**2**	**3**	**4**	**5**
16.11.01	1	58,20	58,2	1
19.11.01	2	58,60	117,2	4
20.11.01	3	58,80	176,4	9
21.11.01	4	59,00	236	16
22.11.01	5	59,10	295,5	25
23.11.01	6	58,50	351	36
26.11.01	7	59,50	416,5	49
27.11.01	8	60,00	480	64
28.11.01	9	60,00	540	81
29.11.01	10	60,10	601	100
30.11.01	11	60,20	662,2	121
03.12.01	12	60,30	723,6	144
04.12.01	13	61,00	793	169
05.12.01	14	60,60	848,4	196
06.12.01	15	60,50	907,5	225
07.12.01	16	61.50	984	256
10.12.01	17	62,00	1054	289
11.12.01	18	61,20	1101,6	324
12.12.01	19	61,00	1159	361
13.12.01	20	61,00	1220	400
14.12.01	21	61,00	1281	441
sumy	$\sum_{t=1}^{21} t = 231$	$\sum_{t=1}^{21} y_t = 1262,1$	$\sum_{t=1}^{21} y_t t = 14006,1$	$\sum_{t=1}^{21} y_t t^2 = 3311$

Źródło: obliczenia własne.

Podstawiając do wzorów (5.22) i (5.23) otrzymamy, że:

$$a = \frac{N\sum_{i=1}^{N} y_t t - \sum_{i=1}^{N} y_t \sum_{i=1}^{N} t}{N\sum_{i=1}^{N} t^2 - \left(\sum_{i=1}^{N} t\right)^2} = \frac{21 \cdot 14006,1 - 1262,1 \cdot 231}{21 \cdot 3311 - 231^2} = 0,15974.$$

$$b = \frac{\sum_{i=1}^{N} y_t - a\sum_{i=1}^{N} t}{N} = \frac{1225,2}{21} = 58,34286.$$

Funkcja trendu będzie miała postać:

$$\hat{y}_t = 0,15974t + 58,34286.$$

Interpretując parametry funkcji trendu należy stwierdzić, że kurs akcji Banku Handlowego S.A. charakteryzował się przeciętnym wzrostem 0,16 zł. To znaczy, jeżeli przesuniemy się w czasie o jedną sesję giełdową, to wówczas kurs akcji wzrośnie średnio o 0,16zł. Parametr przesunięcia (58,34 zł) wskazuje, ile wyniesie średni kurs akcji, dla $t = 0$ (tzn. 15.11.01).

W celu zbadania dobroci dopasowania wyznaczonej funkcji trendu wyznaczymy wartość odchylenia standardowego składnika resztowego. Wszystkie niezbędne obliczenia zawiera poniższa tablica robocza.

Tab. 5.8. Wyznaczanie liniowej funkcji trendu

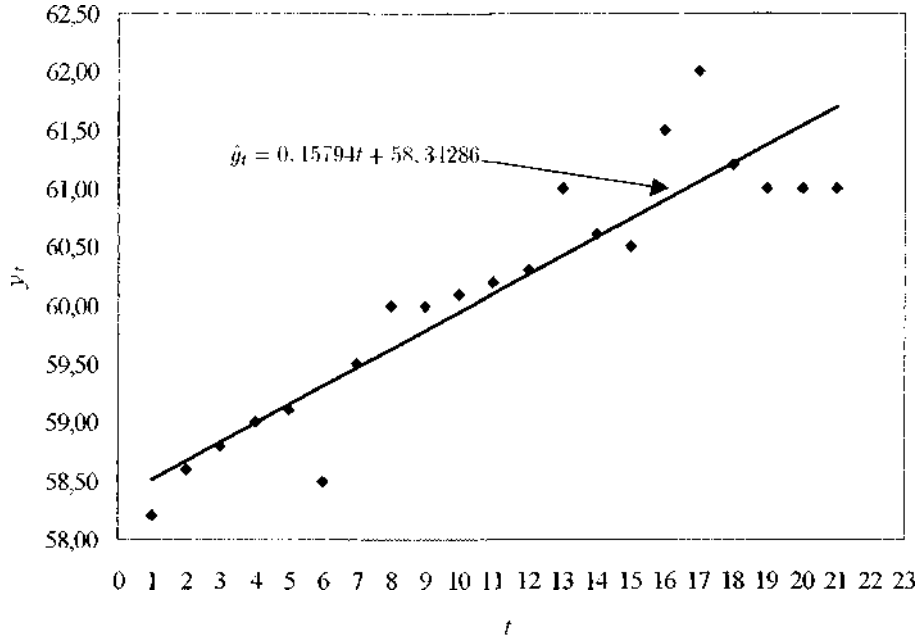
Data (dd.mm.rr)	t	Kurs zamknięcia y_t	$\hat{y}_t$	$y_t - \hat{y}_t$	$(y_t - \hat{y}_t)^2$
1	**2**	**3**	**4**	**5**	**6**
16.11.01	1	58,20	58,5026	-0,30	0,09
19.11.01	2	58,60	58,66234	-0,06	0,00
20.11.01	3	58,80	58,82208	-0,02	0,00
21.11.01	4	59,00	58,98182	0,02	0,00
22.11.01	5	59.10	59,14156	-0,04	0,00
23.11.01	6	58,50	59,3013	-0,80	0,64
26.11.01	7	59,50	59,46104	0,04	0,00
27.11.01	8	60,00	59,62078	0,38	0,14
28.11.01	9	60,00	59,78052	0,22	0,05
29.11.01	10	60,10	59,94026	0,16	0,03
30.11.01	11	60,20	60,1	0,10	0,01
03.12.01	12	60,30	60,25974	0,04	0,00
04.12.01	13	61,00	60,41948	0,58	0,34
05.12.01	14	60,60	60,57922	0,02	0,00
06.12.01	15	60,50	60,73896	-0,24	0,06
07.12.01	16	61,50	60,8987	0,60	0,36
10.12.01	17	62,00	61,05844	0,94	0,89
11.12.01	18	61,20	61,21818	-0,02	0,00
12.12.01	19	61,00	61,37792	-0,38	0,14
13.12.01	20	61,00	61,53766	-0,54	0,29
14.12.01	21	61,00	61,6974	-0,70	0,49
suma	**231**	**1262,10**	**XXX**	**XXX**	**3,53**

Źródło: obliczenia własne.

Wartości empiryczne i funkcja trendu zostały zaprezentowane na rysunku 5.2.

Wartość odchylenia standardowego składnika resztowego wynosi:

$$s_\varepsilon = \sqrt{\frac{\sum_{t=1}^{21}(y_t - \hat{y}_t)^2}{N}} = \sqrt{\frac{3,531948}{21}} = \sqrt{0,168188} = 0,410107.$$



Rys. 5.2. Graficzna prezentacja wyznaczania tendencji rozwojowej metodą średnich ruchomych
Źródło: opracowanie własne.

Z powyższego wynika, że wartości empiryczne notowań odchylają się od wartości hipotetycznych, wyznaczonych w oparciu o oszacowaną funkcję trendu przeciętnie o $\pm 0,41$ zł. #

Spośród metod służących wygładzaniu szeregów czasowych w literaturze przedmiotu spotkać można również tzw. metodę wyrównywania wykładniczego. Zainteresowanych tą metodą czytelników odsyłamy np. do pracy: J. Jóźwiak, J. Podgórski, *Statystyka od podstaw*, [6] s. 312. Odrębną kwestią, istotną zwłaszcza podczas analizy szeregów czasowych długookresowych, jest analiza wahań okresowych. Zagadnienia te są jednak dość szeroko omawiane w większości podręczników z dziedziny statystyki, cytowanych i wymienianych przez nas w bibliografii. Jedną z pozycji może być np. przytoczona powyżej praca: J. Jóźwiak i A. Podgórskiego. Spośród innych na uwagę zasługują też prace: I. Bąk, I. Mankowicz, M. Mojsiewicz, K. Wawrzyniak, *Statystyka w zadaniach, cz. I statystyka opisowa*, [1] s. 209; W. Makuć, D. Urbanek-Krzysztofiak, *Metody opisu statystycznego*, [12] s. 239.

Zadania do samodzielnego rozwiązania

Zadanie 1. W jednym z miast województwa Podkarpackiego zbadano liczbę wypożyczeń w Bibliotece Miejskiej w ciągu stu kolejnych dni roboczych. Otrzymano następujące wyniki zestawione w tablicy:

Tab. 1

Dzień (i)	1	2	3	4	5	6	7	8	9	10
Liczba wypożyczeń (x_i)	199	176	175	117	196	266	141	145	98	164
Dzień (i)	11	12	13	14	15	16	17	18	19	20
Liczba wypożyczeń (x_i)	160	234	196	138	136	236	224	212	172	161
Dzień (i)	21	22	23	24	25	26	27	28	29	30
Liczba wypożyczeń (x_i)	178	200	201	163	134	203	227	142	245	168
Dzień (i)	31	32	33	34	35	36	37	38	39	40
Liczba wypożyczeń (x_i)	162	168	164	194	232	147	176	116	97	171
Dzień (i)	41	42	43	44	45	46	47	48	49	50
Liczba wypożyczeń (x_i)	160	170	118	141	156	251	155	134	169	100
Dzień (i)	51	52	53	54	55	56	57	58	59	60
Liczba wypożyczeń (x_i)	128	193	148	141	104	160	180	164	170	136
Dzień (i)	61	62	63	64	65	66	67	68	69	70
Liczba wypożyczeń (x_i)	141	156	175	154	128	254	85	148	159	124
Dzień (i)	71	72	73	74	75	76	77	78	79	80
Liczba wypożyczeń (x_i)	90	130	122	178	178	188	132	175	73	243
Dzień (i)	81	82	83	84	85	86	87	88	89	90
Liczba wypożyczeń (x_i)	230	126	143	125	107	201	174	161	144	121
Dzień (i)	91	92	93	94	95	96	97	98	99	100
Liczba wypożyczeń (x_i)	124	105	206	176	220	280	224	164	215	165

Źródło: badania własne.

Na podstawie powyższych danych:
a) zbuduj szereg pozycyjny,
b) zbuduj szereg rozdzielczy punktowy,
c) zbuduj szereg rozdzielczy przedziałowy.
Zaprezentuj powyższe szeregi graficznie i tabelarycznie. Zinterpretuj wybrane wartości otrzymanych szeregów.

Zadanie 2. W celu ustalenia średnich miesięcznych wydatków związanych z użytkowaniem telefonów komórkowych w grupie 36 abonamentów przeprowadzono badania ankietowe. Otrzymae dane zestawiono w tablicy II. Na podstawie powyższych danych:
a) zbuduj szereg pozycyjny,
b) zbuduj szereg rozdzielczy punktowy,
c) zbuduj szereg rozdzielczy przedziałowy.

Zaprezentuj powyższe szeregi graficznie i tabelarycznie. Zinterpretuj wybrane wartości otrzymanych szeregów.

Tab. II

i	1	2	3	4	5	6	7	8	9	10
Wydatki w zł (x_i)	50	50	150	30	40	150	80	50	70	120
i	11	12	13	14	15	16	17	18	19	20
Wydatki w zł (x_i)	80	30	60	100	20	35	50	40	50	50
i	21	22	23	24	25	26	27	28	29	30
Wydatki w zł (x_i)	50	60	50	70	80	75	110	150	150	35
i	31	32	33	34	35	36	–			–
Wydatki w zł (x_i)	40	80	90	50	40	30	–	–	–	–

Źródło: badania własne.

Zadanie 3. Badaniem objęto 50 punktów małej gastronomii obserwując liczbę wydawanych posiłków. Otrzymane wyniki – w kolejności ich zebrania – zestawiono w tablicy III.

Tab. III

41	52	46	42	46	36	44	68	58	44
49	48	48	65	52	50	45	72	45	43
47	49	57	44	48	49	45	47	48	43
45	56	61	54	51	47	42	53	41	45
58	55	43	63	38	42	43	46	49	47

Źródło: dane umowne.

Na podstawie powyższych danych:

a) zbuduj szereg pozycyjny,

b) zbuduj szereg rozdzielczy punktowy,

c) zbuduj szereg rozdzielczy przedziałowy.

Zaprezentuj powyższe szeregi graficznie i tabelarycznie. Zinterpretuj wybrane wartości otrzymanych szeregów.

Zadanie 4. Tablica IV prezentuje liczbę punktów otrzymanych ze sprawdzianu ze statystyki, przeprowadzonego w trzech grupach kierunku Towaroznawstwo dnia 22. 11. 2002. Wykorzystując zebrane dane:

a) przekształć szereg do postaci szeregu rozdzielczego punktowego,

b) na podstawie otrzymanego szeregu rozdzielczego wyznacz i zinterpretuj:
   – średnią arytmetyczną i odchylenie standardowe,
   – medianę i modalną,

c) zbadaj symetryczność rozkładu ocen.

Tab. IV

i	1	2	3	4	5	6	7	8	9	10
Punkty (x_i)	23	19	16	10	22	18	24	24	8	15
i	11	12	13	14	15	16	17	18	19	20
Punkty (x_i)	20	18	29	13	29	20	9	28	15	17
i	21	22	23	24	25	26	27	28	29	30
Punkty (x_i)	25	10	25	26	28	17	29	21	20	23
i	31	32	33	34	35	36	37	38	39	40
Punkty (x_i)	29	18	23	22	24	19	22	15	26	22
i	41	42	43	44	45	46	47	48	49	50
Punkty (x_i)	23	15	25	16	17	27	29	20	30	24
i	51	52	53	54	55	56	57	58	59	60
Punkty (x_i)	28	23	17	23	22	13	29	27	9	18
i	61	62	63	64						
Punkty (x_i)	26	23	29	15						

Źródło: badania własne.

Zadanie 5. Oceny z przedmiotu Mikroekonomia 20 studentów kierunku Gospodarka Przestrzenna AE w Krakowie w roku akademickim 2002/2003 przedstawiały się następująco:

Tab. V

Student (i)	1	2	3	4	5	6	7	8	9	10
Ocena (x_i)	4	3	3	3	2	4	5	5	3	4.5
Student (i)	11	23	13	14	15	16	17	18	19	20
Ocena (x_i)	3,5	3	4,5	3	3	3	3	3	3,5	4

Źródło: badania własne.

Korzystając z danych:

a) przedstaw powyższe dane w postaci szeregu szczegółowego uporządkowanego,
b) przekształć szereg do postaci szeregu rozdzielczego punktowego,
c) na podstawie otrzymanego szeregu rozdzielczego wyznacz i zinterpretuj:
   – średnią arytmetyczną i odchylenie standardowe,
   – medianę i modalną,
d) zbadaj symetryczność rozkładu ocen.

Zadanie 6. Tablica VI przedstawia strukturę liczby punktów otrzymanych podczas sprawdzianu ze statystyki. Na podstawie danych zawartych w tablicy wyznacz i zinterpretuj:

a) średnią arytmetyczną i odchylenie standardowe,
b) medianę i modalną (analitycznie oraz graficznie).

Czy dla badanego przykładu można wyznaczyć współczynnik koncentracji?

Tab. VI

j	$(x_{d.j}; x_{g.j}]$	Liczba studentów, którzy otrzymali więcej niż $x_{d.j}$, lecz najwyżej $x_{g.j}$ punktów (f_j)
1	(4; 8]	1
2	(8; 12]	4
3	(12; 16]	9
4	(16; 20]	14
5	(20; 24]	17
6	(24; 28]	11
7	(28; 32]	8
suma	xxx	64

Źródło: badania własne.

Zadanie 7. Strukturę wydatków związanych z użytkowaniem telefonów komórkowych w grupie 36 abonamentów przedstawia tablica VII.

Tab. VII

j	$(x_{d.j}; x_{g.j}]$	v_j [%]
1	(19; 41]	27,78
2	(41; 63]	30,56
3	(63; 85]	19,44
4	(85; 107]	5,56
5	(107; 129]	5,56
6	(129; 151]	11,11

Źródło: badania własne.

Na podstawie danych:

a) zaprezentuj powyższy szereg w postaci graficznej,

b) wyznacz średnią arytmetyczną i odchylenie standardowe.

Zadanie 8. W listopadzie 1999 roku zanotowano w Krakowie powierzchnie dziesięciu mieszkań (typu M1) (x_i [m^2]) oraz odpowiadające im ceny (y_i [tys. zł]). Dane zebrano w tablicy VIII.

Na podstawie danych:

a) narysuj wykres współzależności,

b) oceń kierunek i siłę związku pomiędzy badanymi cechami,

c) wyznacz równanie regresji liniowej i na jego podstawie oszacuj, ile będzie wynosiła cena mieszkania o powierzchni 32 m^2,

d) oceń dopasowanie modelu do danych rzeczywistych,

e) skomentuj otrzymane wyniki.

Tab. VIII

i	Powierzchnia mieszkania x_i [m^2]	Cena y_i [tys. zł]
1	16	69
2	18	73
3	20	59
4	24	77
5	25	74
6	29	82
7	35	81
8	39	83
9	40	103
10	50	129

Źródło: badania własne.

Zadanie 9. Liczby szkół wyższych i wartości stopy bezrobocia w siedmiu wybranych miastach odnotowane na koniec grudnia 2001 roku przedstawiały się następująco:

Tab. IX

MIASTA	i	Liczba szkół wyższych x_i	Stopa bezrobocia y_i
Warszawa	1	62	5,10
Białystok	2	8	13,30
Bydgoszcz	3	8	11,70
Gdańsk	4	12	10,10
Gorzów Wlkp.	5	3	15,40
Katowice	6	10	7,00
Kielce	7	8	15,00
Kraków	8	18	8,10
Lublin	9	10	12,70
Łódź	10	17	17,80
Olsztyn	11	4	12,50
Opole	12	3	9,60
Poznań	13	20	5,60
Rzeszów	14	4	9,50
Szczecin	15	14	11,00
Toruń	16	3	13,70
Wrocław	17	17	9,70
Zielona Góra	18	1	12,30

Źródło: Główny Urząd Statystyczny (www.stat.gov.pl)

Wykorzystując powyższy materiał statystyczny:
a) oceń, czy istnieje zależność pomiędzy tymi wielkościami? Jeżeli tak, to podaj jej siłę,
b) określ, o ile w przybliżeniu zmieni się stopa bezrobocia, jeżeli w mieście powstanie jeszcze jedna szkoła wyższa?
c) oceń dopasowanie modelu do danych rzeczywistych,
d) skomentuj uzyskane wyniki.

Zadanie 10. Emisja przemysłowych zanieczyszczeń powietrza (w tonach) oraz liczba przestępstw stwierdzonych o zakończonych postępowaniach przygotowawczych w 2001r. w szesnastu podregionach kształtowała się następująco:

Tab. X

Podregion	i	Emisja przemysłowych zanieczyszczeń powietrza	Przestępstwa stwierdzone w zakończonych postępowaniach przygotowawczych
Dolnośląskie	1	20239	117002
Kujawsko - Pomorskie	2	12476	79154
Lubelskie	3	7196	64022
Lubuskie	4	4491	43656
Łódzkie	5	10339	85032
Małopolskie	6	14432	104470
Mazowieckie	7	13748	195651
Opolskie	8	7431	34397
Podkarpackie	9	4146	44802
Podlaskie	10	1969	34022
Pomorskie	11	4952	116646
Śląskie	12	32805	178189
Świętokrzyskie	13	6230	36356
Warmińsko - Mazurskie	14	2233	60461
Wielkopolskie	15	12433	116742
Zachodniopomorskie	16	7080	79487

Źródło: Główny Urząd Statystyczny (www.stat.gov.pl).

Wykorzystując powyższy materiał statystyczny:
a) oceń, czy istnieje zależność pomiędzy tymi wielkościami? Jeżeli tak, to podaj jej siłę,
b) zbuduj model regresji liniowej,
c) oceń dopasowanie modelu do danych rzeczywistych,
d) skomentuj uzyskane wyniki.

Zadanie 11. W 1999 roku zanotowano w podregionach następujące wielkości emisji przemysłowych gazowych zanieczyszczeń powietrza (w tonach) oraz ogólną liczbę zgonów wywołanych zapaleniem płuc lub oskrzeli:

Tab. XI

Województwo	i	Emisja przemysłowych gazowych zanieczyszczeń powietrza	Liczba zgonów wywołanych zapaleniem płuc i oskrzeli
Dolnośląskie	1	109446	1089
Kujawsko - Pomorskie	2	78699	1017
Lubelskie	3	44349	902
Lubuskie	4	26490	345
Łódzkie	5	380587	1813
Małopolskie	6	197381	1218
Mazowieckie	7	205280	2305
Opolskie	8	53864	259
Podkarpackie	9	31917	683
Podlaskie	10	16563	613
Pomorskie	11	54107	868
Śląskie	12	590133	1867
Świętokrzyskie	13	78116	666
Warmińsko - Mazurskie	14	17460	473
Wielkopolskie	15	189647	1289
Zachodniopomorskie	16	95124	400

Źródło: Główny Urząd Statystyczny (www.stat.gov.pl).

Zbadaj istnienie współzależności pomiędzy obiema wielkościami. Opisz liczbę zgonów jako liniową funkcję emisji przemysłowych gazowych zanieczyszczeń powietrza.

Zadanie 12. Liczba szkół wyższych oraz liczba mieszkań oddanych do użytku w 2001 roku w wybranych miastach kształtowała się następująco:

Tab. XII

Miasto	i	Liczba szkół wyższych	Mieszkania oddane do użytku
M.st. Warszawa	1	62	16278
Białystok	2	8	1825
Bydgoszcz	3	8	1563
Gdańsk	4	12	1088
Gorzów Wlkp.	5	3	755
Katowice	6	10	521
Kielce	7	8	732
Kraków	8	18	5517
Lublin	9	10	2573
Łódź	10	17	1652
Olsztyn	11	4	1245
Opole	12	3	322
Poznań	13	20	3305

Rzeszów	14	4	434
Szczecin	15	14	2259
Toruń	16	3	1358
Wrocław	17	17	5571
Zielona Góra	18	1	738

Źródło: Główny Urząd Statystyczny (www.stat.gov.pl).

Zbadaj, czy pomiędzy zmiennymi występuje zależność wykorzystując:
a) współczynnik korelacji liniowej Persona,
b) współczynnik korelacji rang Spearmana,
c) porównaj otrzymane wyniki.

Zadanie 13. Poniższe dane prezentują „wiek" samochodu osobowego pewnej marki x_i (w latach) oraz średnią cenę y_i (w tys. zł).

Tab. XIII

x_i	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17
y_i	50	45	35,1	32,3	27,1	26,2	23,2	20,1	16,2	14	12,2	11,5	10,5	11,5	8	5

Źródło: http://whww.trader.pl/pl/cenniki_uzywane

Wykorzystując powyższe dane:
a) zbadaj siłę i kierunek współzależności analizowanych zmiennych,
b) oszacuj parametry liniowej funkcji regresji opisującej wpływ rocznika samochodu na jego cenę.

Zadanie 14. Postanowiono zbadać, czy istnieje zależność pomiędzy powierzchnią mieszkania (typu M1), a zainstalowaniem w nim telefonu. W tym celu zebrano dane, które zaprezentowano w tablicy XIV:

Tab. XIV

i	Powierzchnia mieszkania y_i [m^2]	Telefon
1	16	N
2	18	N
3	20	N
4	24	T
5	25	N
6	29	N
7	35	T
8	39	N
9	40	T
10	50	T

Źródło: badania własne.

Posługując się odpowiednią miarą statystyczną oceń stopień zależności badanych zmiennych. Skomentuj otrzymane wyniki.

Zadanie 15. Sześćdziesiąt samochodów poddawanych było regularnym przeglądom technicznym, a w czterdziestu samochodach przegląd był wykonywany bardzo rzadko. W grupie pierwszej awarie zdarzyły się 10 razy w ciągu roku, w drugiej aż 30. Czy istnieje związek pomiędzy przeglądem technicznym i liczbą awarii? Uzasadnij swoją odpowiedź posługując się odpowiednim miernikiem.

Zadanie 16. Poniższa tablica przedstawia liczbę bydła (w czerwcu 2002, w tys. sztuk), liczbę trzody chlewnej (w lipcu 2002, w tys. sztuk) oraz liczbę loch przeznaczonych na chów (w lipcu 2002, w tys. sztuk) w poszczególnych województwach.

Tab. XVI

Województwa	Liczba loch (y_i) w tys. szt.	Trzoda chlewna (x_{1i}) w tys. szt.	Bydło (x_{2i}) w tys. szt.
Dolnośląskie	52,1	502,7	133,5
Kujawsko-Pomorskie	250,8	2358,9	410,5
Lubelskie	139,5	1434,8	438,1
Lubuskie	29,1	271,2	66,6
Łódzkie	132,7	1389,1	450,8
Małopolskie	65,3	519,9	288,1
Mazowieckie	209,2	2007,1	913,6
Opolskie	75,8	797,1	132,7
Podkarpackie	35,4	347,9	204,0
Podlaskie	95,4	963,4	681,2
Pomorskie	104,8	1028,4	199,6
Śląskie	38,8	370,5	140,9
Świętokrzyskie	51,7	391,4	208,7
Warmińsko-Mazurskie	82,9	813,1	392,3
Wielkopolskie	474,0	4866,7	723,1
Zachodniopomorskie	72,8	645,2	117,8
suma	**1910,30**	**18707,40**	**5501,50**

Źródło: Główny Urząd Statystyczny (www.stat.gov.pl).

Na podstawie powyższych danych wyznacz:

a) współczynniki korelacji cząstkowej,

b) parametry równania regresji wielorakiej typu: $y_i = a_0 + a_1 x_{1i} + a_2 x_{2i}$,

c) współczynnik korelacji wielorakiej.

Zinterpretuj uzyskane wyniki liczbowe.

Zadanie 17. W okresie od 1990 roku do 2000 roku Bezpośrednie inwestycje zagraniczne (X_1), Eksport ogółem (X_2) oraz Nakłady inwestycyjne ogółem (X_3) kształtowały się, w cenach stałych z roku 1990, następująco:

Tab. XVII

Rok	t	x_1 [mln \$]	x_2 [mln ZLP]	x_3 [mln ZLP]
1990	1	111,00	14321,60	11581,00
1991	2	249,00	14903,40	9914,09
1992	3	1236,00	13186,60	8252,03
1993	4	1619,00	14143,10	7482,86
1994	5	1702,00	17240,10	7749,45
1995	6	2862,00	22894,90	8468,60
1996	7	6239,00	24439,80	9831,01
1997	8	6560,00	25751,30	11791,10
1998	9	10064,00	28228,90	13156,09
1999	10	8262,00	27407,40	13689,21
2000	11	10479,00	28796,20	13471,32

Źródło: Państwowa Agencja Inwestycji Zagranicznych.

Wyznacz i zinterpretuj współczynniki korelacji cząstkowej.

Zadanie 18. W okresie od stycznia do września 2002 roku liczba zarejestrowanych bezrobotnych w Polsce wynosiła odpowiednio:

Tab. XVIII

Miesiąc	Bezrobotni (w tys.)
I	3253
II	3278
III	3260
IV	3204
V	3065
VI	3091
VII	3105
VIII	3106
IX	3113

Źródło: Główny Urząd Statystyczny (www.stat.gov.pl).

W oparciu o powyższe dane:

a) wyznacz przyrosty bezwzględne łańcuchowe,

b) wyznacz przyrosty bezwzględne jednopodstawowe, przyjmując za okres bazowy styczeń 2002,

c) wyznacz przyrosty względne łańcuchowe,

d) wyznacz przyrosty względne jednopodstawowe, przyjmując za okres bazowy styczeń 2002,

e) wygładź szereg czasowy wykorzystując średnią ruchomą trzyokresową,
f) wygładź szereg czasowy wykorzystując liniowe równanie trendu.

Zadanie 19. W poszczególnych miesiącach 2001 roku oddano do użytku następującą liczbę mieszkań:

Tab. XIX

2001	Mieszkania oddane do użytku [tys.] (x_t)
1	9,54
2	7,78
3	8,44
4	8,10
5	7,06
6	6,70
7	8,85
8	7,96
9	7,44
10	9,07
11	10,31
12	14,73
SUMA	105,97

Źródło: Główny Urząd Statystyczny (www.stat.gov.pl).

Wykorzystując powyższy szereg czasowy:
a) wyznacz indeksy indywidualne łańcuchowe,
b) wyznacz indeksy indywidualne jednopodstawowe, przyjmując za okres bazowy styczeń,
c) wygładź szereg metodą mechaniczną wykorzystując do tego celu średnią ruchomą czterookresową,
d) wygładź szereg metodą analityczną (równaniem regresji).

Zadanie 20. W poszczególnych miesiącach 2001 roku przeciętne miesięczne wynagrodzenie nominalne brutto, w sektorze przedsiębiorstw łącznie z obligatoryjną składką na ubezpieczenia społeczne (w zł) prezentuje tablica XX.
Wykorzystując szereg czasowy:
a) wyznacz przyrosty bezwzględne jednopodstawowe, przyjmując za okres bazowy styczeń,
b) wyznacz przyrosty względne łańcuchowe,
c) wyznacz indeksy indywidualne jednopodstawowe, przyjmując za okres bazowy styczeń,
d) wygładź szereg metodą mechaniczną wykorzystując do tego celu średnią ruchomą czterookresową,
e) wygładź szereg metodą analityczną dopasowując liniową funkcję trendu.

Tab. XX

2001	Przeciętne miesięczne wynagrodzenie nominalne brutto w sektorze przedsiębiorstw [zł]
I	2069,29
II	2074,91
III	2149,13
IV	2175,55
V	2163,44
VI	2148,44
VII	2198,50
VIII	2192,41
IX	2217,55
X	2252,16
XI	2302,46
XII	2474,11

Źródło: Główny Urząd Statystyczny (www.stat.gov.pl).

Zadanie 21. W ciągu 30 kolejnych sesji giełdowych zanotowano następujące kursy akcji Banku Handlowego:

Tab. XXI

t	Data (dd-mm-rr)	Kurs	t	Data (dd-mm-rr)	Kurs
1	06-11-01	54,50	16	27-11-01	60,00
2	07-11-01	55,00	17	28-11-01	60,00
3	08-11-01	55,50	18	29-11-01	60,10
4	09-11-01	54,90	19	30-11-01	60,20
5	12-11-01	56,10	20	03-12-01	60,30
6	13-11-01	57,30	21	04-12-01	61,00
7	14-11-01	58,20	22	05-12-01	60,60
8	15-11-01	58,00	23	06-12-01	60,50
9	16-11-01	58,20	24	07-12-01	61,50
10	19-11-01	58,60	25	10-12-01	62,00
11	20-11-01	58,80	26	11-12-01	61,20
12	21-11-01	59,00	27	12-12-01	61,00
13	22-11-01	59,10	28	13-12-01	61,00
14	23-11-01	58,50	29	14-12-01	61,00
15	26-11-01	59,50	30	17-12-01	62,50

Źródło: www.penetrator.com.pl

Przedstaw powyższy szereg graficznie oraz:

a) wygładź szereg wykorzystując średnią ruchomą pięciookresową,

b) wygładź szereg wykorzystując średnią ruchomą ośmiookresową,

c) wyznacz parametry liniowej funkcji trendu typu $y = at + b$,
d) odpowiedz na pytanie, o ile przyrosły ceny akcji w okresie od 1 do 10 sesji,
e) skomentuj otrzymane wyniki.

Zadanie 22. W okresie od 1990 roku do 2000 roku nakłady inwestycyjne ogółem (x_t) kształtowały się, w cenach stałych z roku 1990, następująco:

Tab. XXII

Rok	t	x_t [mln ZLP]
1990	1	11581,00
1991	2	9914,09
1992	3	8252,03
1993	4	7482,86
1994	5	7749,45
1995	6	8468,60
1996	7	9831,01
1997	8	11791,10
1998	9	13156,09
1999	10	13689,21
2000	11	13471,32

Źródło: Państwowa Agencja Inwestycji Zagranicznych.

Przedstaw szereg graficznie i oszacuj parametry liniowej funkcji trendu. Zinterpretuj uzyskane wyniki.

Zadanie 23. Przeciętne ceny trzech artykułów spożywczych (w zł) oraz wielkość dostaw (w tys. ton) w Polsce w 1995 i 2000 roku wynosiły odpowiednio:

Tab. XXIII

		1995 (t_0)		2000 (t_1)	
i	Produkt	Wielkość dostaw $q_{0,i}$	Cena $p_{0,i}$	Wielkość dostaw $q_{1,i}$	cena $p_{1,i}$
1	mąka	1823	1,07	2080	1,62
2	makaron	64,2	1,68	102	3,59
3	czekolada	49,9	1,60	112	2,45

Źródło: Mały rocznik statystyczny Polski 2002.

Przyjmując jako okres podstawowy rok 1995, a okres badany rok 2000, oblicz i zinterpretuj:
a) agregatowe indeksy cen i wielkości dostaw (Paaschego i Laspeyresa),
b) agregatowe indeksy Fishera,
c) agregatowy indeks wartości.

Bibliografia

[1] I. Bąk, I. Mankowicz, M. Mojsiewicz, K. Wawrzyniak, *Statystyka w zadaniach, cz. I statystyka opisowa*, Wydawnictwo Naukowo-Techniczne, Warszawa 2001.

[2] J. Bielecki, B. Jurkiewicz, Z. Szymanowska, *Zbiór zadań ze statystyki ogólnej i matematycznej*, PWN, Warszawa 1975.

[3] A. Goryl, Z. Jędrzejczyk, K. Kukuła, J. Osiewalski, A. Walkosz, *Wprowadzenie do Ekonometrii w przykładach i zadaniach*, Wydawnictwo Naukowe PWN, Warszawa 1996.

[4] A. Iwasiewicz, Z. Paszek, *Statystyka z elementami statystycznych metod sterowania jakością*, AE w Krakowie, Kraków 2000.

[5] A. Iwasiewicz, *Zarządzanie jakością*, Wydawnictwo Naukowe PWN, Warszawa-Kraków 1999.

[6] J. Jóźwiak, J. Podgórski, *Statystyka od podstaw*, PWE, Warszawa 1994 (2001).

[7] A. Komosa, J. Musiałkiewicz, *Statystyka*, Ekonomik 2001.

[8] A. Komosa, J. Musiałkiewicz, *Statystyka – ćwiczenia*, Ekonomik 2001.

[9] J. Kordos, *Jakość danych statystycznych*, PWE, Warszawa 1988.

[10] J. Kubala, E. Smaga, T. Stanisz, *Elementy algebry liniowej*, PWN, Warszawa 1983.

[11] M. Łobocki, *Metody badań pedagogicznych*, PWN Warszawa 1984.

[12] W. Makać, D. Urbanek-Krzysztofiak, *Metody opisu statystycznego*, Wydawnictwo Uniwersytetu Gdańskiego, Gdańsk 2001.

[13] B. Niemierko, *Testy osiągnięć szkolnych*, Wydawnictwo Szkolne i Pedagogiczne, Warszawa 1975.

[14] *Słownik wyrazów obcych* pod redakcją Jana Tokarskiego, PWN 1980.

[15] M. Sobczyk, *Statystyka*, PWN, Warszawa 1998 (2000, 2002).

[16] J. Steczkowski, *Reprezentacyjne badania jakości wyrobów, kontrola odbiorcza*, Wydawnictwo PLATAN, Kraków 1993.

[17] Steczkowski J., Zeliaś A., *Statystyczne metody analizy cech jakościowych*, PWE 1981.

[18] K. Walenta, *Podstawowe pojęcia teorii pomiaru*, w: Problemy psychologii matematycznej (red. J. Kozielecki), PWN, Warszawa 1971.

[19] M. Woźniak, *Statystyka ogólna*, AE w Krakowie, Kraków 1999.

[20] *Wykresy i mapy statystyczne*, Główny Urząd Statystyczny, Warszawa 1977.

[21] K. Zając, *Wykłady ze statystyki*, AE w Krakowie, Kraków 1995.

[22] K. Zając, *Zarys metod statystycznych*, PWE, Warszawa 1988.