Krakowska Szkoła Wyższa
im. Andrzeja Frycza Modrzewskiego

Jolanta Kurkiewicz, Marcin Stonawski

# Podstawy statystyki

Kraków 2005

Rada Wydawnicza:
Klemens Budzowski, Andrzej Kapiszewski,
Jacek Majchrowski, Zbigniew Maciąg

Recenzja:
Prof. dr hab. Andrzej Iwasiewicz

Redaktor prowadzący:
Halina Baszak Jaroń

Adiustacja i korekta:
Mariusz Warchol




ISBN 83-89823-95-0



Na zlecenie:
Krakowskiej Szkoły Wyższej im. Andrzeja Frycza Modrzewskiego
www.ksw.edu.pl

Wydawca:
Krakowskie Towarzystwo Edukacyjne sp. z o.o.,
Oficyna Wydawnicza AFM, Kraków 2005

Skład i łamanie:
Wojciech Prażuch

Druk i oprawa: Zakład Poligraficzny „Cenzus"

# Spis treści

# Wstęp

Studenci, podejmując studia na wybranym kierunku i zapoznając się z ich programem, często zastanawiają się, w jakich sytuacjach będzie im przydatna wiedza z zakresu dyscyplin naukowych występujących w planie studiów. Przemyślenia te odnoszą się również do statystyki. W podręczniku przedstawimy wzorcowe przykłady odnoszące się do sytuacji, w których posiadanie wiadomości z tego zakresu może okazać się użyteczne.

Wybierając kierunek studiów, zastanawiamy się, jaka jest szansa znalezienia dobrej pracy oraz jakich zarobków możemy oczekiwać po zdobyciu wyższego wykształcenia. Kierownictwo telewizji w celu pozyskania jak najszerszej widowni obserwuje oglądalność emitowanych programów. Badania opinii publicznej dostarczają informacji między innymi o tym, ilu widzów ogląda programy informacyjne, sportowe, publicystyczne, jaka jest oglądalność filmów akcji, komedii, dramatów, w jakich godzinach widownia jest największa, jakie są preferencje widzów w zależności od wieku, wykształcenia, aktywności zawodowej, miejsca zamieszkania. Projektanci domu mody pracują nad nową kolekcją, którą będą chcieli wprowadzić na rynek w nadchodzącym sezonie. Dział marketingu przeprowadza więc badania mające na celu ustalenie zapotrzebowania, biorąc pod uwagę, iż na wydatki na odzież, oprócz gustów, preferencji i mody, mają wpływ takie czynniki jak: poziom dochodów, wiek, płeć, wykształcenie przyszłych nabywców. Klient banku zwraca się do swojego doradcy o informację, jak najlepiej zainwestować posiadane środki finansowe. Doradca musi wziąć pod uwagę obecny i przewidywany poziom inflacji, okres, na jaki klient decyduje się zainwestować swoje środki. Planujemy zakup mieszkania i decydujemy się zaciągnąć kredyt. Rozpatrując wniosek o jego przyznanie, bank wymaga dostarczenia informacji o poziomie dochodów, stanie zadłużenia, udzielonych i spłaconych dotychczas kredytach, wieku oraz stanie rodzinnym klienta. Dane te są niezbędne do ustalenia zdolności kredytowej klienta oraz własnego ryzyka kredytowego.

W podanych przykładach zaprezentowano różne obszary rzeczywistości. Student interesuje się przyszłą sytuacją na rynku pracy. Bank gromadzi dane o swoich klientach, aby na tej podstawie podejmować decyzje w celu zaspokojenia ich potrzeb oraz

zapewnić sobie niezakłócone funkcjonowanie. Dział marketingu prowadzi badania rynku, aby zapewnić zbyt produkowanych wyrobów.

We wszystkich tych przypadkach, chcąc uzyskać odpowiedź na postawione pytania, musimy zebrać odpowiednie informacje. Będą to obszerne zbiory danych w postaci liczb. Liczby te odzwierciedlają istotne cechy obserwowanej rzeczywistości. Zgromadzone dane muszą być poddane analizie. W tym właśnie zakresie pomocne będą metody statystyczne.

Podręcznik zatytułowany „Podstawy statystyki" jest przeznaczony dla studentów, którzy w przyszłej pracy będą podejmować tego rodzaju decyzje. Problematykę ujęto w siedmiu rozdziałach.

W rozdziale 1 zaprezentowano statystykę jako naukę. Zdefiniowano podstawowe pojęcia, takie jak: zjawiska i procesy masowe, zespół przyczyn głównych i przypadkowych oraz generowany przez nie systematyczny i przypadkowy składnik badanych procesów. Przedstawiono podstawowe typy prawidłowości, na które należy zwracać uwagę, obserwując zjawiska masowe.

Rozdział 2 poświęcono praktycznym zagadnieniom opracowywania materiału statystycznego. Przedstawiono zasady grupowania oraz otrzymywane jako rezultat szeregi statystyczne. Omówiono metody prezentacji danych w formie tabelarycznej i graficznej. Zwrócono uwagę na znaczenie tych prac w kontekście adekwatności uzyskanych wyników do badanej rzeczywistości.

Kolejne trzy rozdziały są poświęcone statystycznym metodom opisu prawidłowości występujących odpowiednio w rozkładach liczebności, we współzależności zjawisk oraz w dynamice.

W rozdziale 3 omówiono charakterystyki opisowe rozkładu liczebności mierzące przeciętny poziom cech ujmujących badane zjawiska oraz stopień ich zróżnicowania. Ukazano, w jakich warunkach wskazane jest zastosowanie omawianych charakterystyk oraz sposób prawidłowego interpretowania otrzymanych wyników.

Przedmiot rozdziału 4 stanowią metody badania związków pomiędzy zjawiskami. Przedstawiono zasady analizy korelacji i regresji. Ograniczono się do najprostszej ich postaci a mianowicie do powiązań liniowych. Najwięcej uwagi poświęcono metodom badania współzależności między dwiema zmiennymi. W celu zwrócenia uwagi na to, że rzeczywistość jest bardziej złożona przedstawiono przykładowe metody ustalania siły i kierunku powiązań między wieloma zmiennymi. Wykład zilustrowano przykładami, które ukazują zastosowania omawianych procedur oraz interpretację wyników.

W rozdziale 5 omówiono metody wykrywania prawidłowości występujących w dynamice zjawisk, czyli w ich rozwoju w czasie. Rozpoczęto od procedur o najprostszej konstrukcji, a mianowicie od indeksów indywidualnych, a następnie wprowadzono indeksy agregatowe. Przedstawiono mechaniczne i analityczne metody wyodrębniania trendu. Wskazano zakres praktycznego wykorzystania wiedzy o tendencji rozwojowej badanych zjawisk.

Przedstawiony wyżej zakres podręcznika (rozdziały 2–5) poświęcony jest wnioskowaniu metodą dedukcyjną. Gdybyśmy zatrzymali prezentację metod statystycz-

nych na tym etapie, podręcznik miałby postać zbioru przepisów prezentujących, jakie czynności należy wykonać, aby ustalić na przykład przewidywane średnie zarobki absolwentów uczelni ekonomicznej. Dlatego w rozdziale 6 wyłożono podstawy rachunku prawdopodobieństwa. Badanie statystyczne musi opierać się na rachunku prawdopodobieństwa, ponieważ jeśli formułujemy sądy o prawidłowości wyprowadzone z niepełnej informacji pochodzącej z próby statystycznej, to prawdziwość tych sądów jest obciążona niepewnością. Niezbędne są zatem takie reguły postępowania, które pozwolą przyjąć odpowiednio wysokie prawdopodobieństwo prawdziwości naszych wniosków lub niskie prawdopodobieństwo wniosków fałszywych. W wykładzie ograniczono się tylko do zmiennych losowych, przyjmując, że rachunek prawdopodobieństwa zdarzeń losowych studenci poznali już w szkole średniej.

Rozdział 7 jest wstępem do wnioskowania statystycznego. I na tym wstępie pozostajemy. Najpierw przedstawiono zasady estymacji parametrów a następnie weryfikację hipotez statystycznych. Ograniczono się tylko do dwóch podstawowych parametrów, a mianowicie do wartości przeciętnej i wariancji.

***

Podręcznik „Podstawy statystyki" został opracowany z myślą o studentach przygotowujących się do prowadzenia badań empirycznych w zakresie nauk społecznych, do których należą również nauki ekonomiczne. Jest adresowany do studiujących na takich kierunkach, jak na przykład: zarządzanie, ekonomia, nauki polityczne, stosunki międzynarodowe. Został on przygotowany w taki sposób, aby mógł stanowić pomoc zarówno do wykładów, jak i ćwiczeń ze statystyki. Autorzy pomyśleli również o tych, którzy samodzielnie chcieliby poznać zasady badań zjawisk ekonomiczno-społecznych z zastosowaniem metod statystycznych. Z myślą o nich podane zostały liczne przykłady ukazujące zastosowanie omawianych procedur.

*Autorzy*

# Podstawowe pojęcia i problemy

Rozdział 1

## 1.1. Statystyka jako nauka

Wprowadzenie pojęcia statystyki jako nauki rozpoczniemy od przykładu z zakresu demografii[1]. Przedmiotem badań demograficznych są między innymi zachowania matrymonialne. Zachowania te są rezultatem podjęcia życiowych decyzji poprzedzonych przemyśleniami. Człowiek, najczęściej młody, będący w stanie wolnym może stawiać sobie następujące pytania:

- czy kiedykolwiek w życiu chcę zawrzeć związek małżeński?
- jeśli tak, to w jakim okresie życia chciałbym to uczynić?
- jakie warunki powinny być spełnione, aby zdecydować się na zawarcie małżeństwa?

Poszczególne osoby udzielają sobie różnych odpowiedzi i podejmują indywidualne decyzje. W podanym przykładzie istotne jest to, że małżeństwo jest zdarzeniem, którego realizacja jest w znacznym zakresie wynikiem indywidualnych przemyśleń i decyzji. Zajście tego zdarzenia jest udokumentowane odpowiednim aktem, a zatem może podlegać rejestracji. Rozpatrywać będziemy tylko pierwsze małżeństwa kobiet, czyli te związki, które mogą zawrzeć tylko panny. Jeśli będziemy rozpatrywać wyodrębnioną w ten sposób zbiorowość kobiet, to powiemy, że interesujemy się zjawiskiem, którym jest zawieranie pierwszych małżeństw. Okazuje się bowiem, że indywidualne zachowania posiadają wspólne cechy możliwe do wykrycia tylko na poziomie populacji. Podejmiemy zatem badania w celu poznania cech charakteryzujących zawieranie pierwszych małżeństw w wyodrębnionej zbiorowości kobiet.

Przystępując do badań, posiadamy już pewne wiadomości pochodzące albo z obserwacji zachowań ludzkich w tej materii, albo z nagromadzonej wcześniej wiedzy. Wiadomo na przykład, że nie wszystkie kobiety wstępują w związki małżeńskie oraz

---

[1] Jak wyodrębniały się te obie dyscypliny przedstawimy w podręczniku do demografii.

że związki te zawierane są w znacznej większości w ustalonym okresie życia zwanym w demografii wiekiem matrymonialnym. Możemy zapytać:

1) jaka w danej populacji jest częstość zawierania małżeństw przynajmniej raz w życiu,
2) jak zmienia się częstość zawierania małżeństw w zależności od wieku,
3) jakie uwarunkowania i w jaki sposób wpływają na zachowania w zakresie zawierania małżeństw i ich zróżnicowanie.

Jako przykładowe odpowiedzi możemy podać następujące charakterystyki, które należy traktować jako umowne, ponieważ nie odnoszą się do żadnej istniejącej zbiorowości:

(1) Częstość zawierania małżeństw w pewnej zbiorowości kobiet wynosi 0,9. Oznacza to, że 90% spośród nich zawiera małżeństwo przynajmniej raz w życiu.
(2) Pierwsze związki małżeńskie są najczęściej zawierane przez kobiety w wieku 23 lat.
(3) Decyzja o zawarciu małżeństwa jest odkładana aż do osiągnięcia wykształcenia przynajmniej średniego, uzyskania stałej pracy, pozwalającej zapewnić rodzinie warunki życia na odpowiednim poziomie.

Odpowiedzi (1) i (2) mają postać liczbową, będą przez wszystkich jednakowo rozumiane. Odpowiedź (3) ma charakter werbalny. Może być zatem różnie interpretowana. Najbardziej jednoznaczna jest odpowiedź „uzyskanie przynajmniej średniego wykształcenia". Mniej jednoznaczne może być rozumienie pojęcia „stała praca". Najbardziej zróżnicowanie będzie pojmowane „zapewnienie rodzinie odpowiedniego poziomu życia". W celu uzyskania bardziej obiektywnej interpretacji będziemy się starać przypisać tym pojęciom charakterystyki liczbowe.

Interesujemy się zjawiskiem (zawieranie małżeństw) obserwowanym w zbiorowości (w populacji) kobiet i staramy się wykryć jego istotne cechy. Jeśli potrafimy rozważane zjawisko scharakteryzować za pomocą liczb, to możemy posłużyć się metodami określanymi mianem statystycznych. Metody te stanowią przedmiot statystyki.

Statystyka jest nauką o **ilościowych metodach badania prawidłowości występujących w zjawiskach masowych scharakteryzowanych za pomocą liczb**[2].

Niektóre pojęcia występujące w sformułowanej wyżej definicji wymagają wyjaśnienia. Przedstawimy je teraz kolejno.

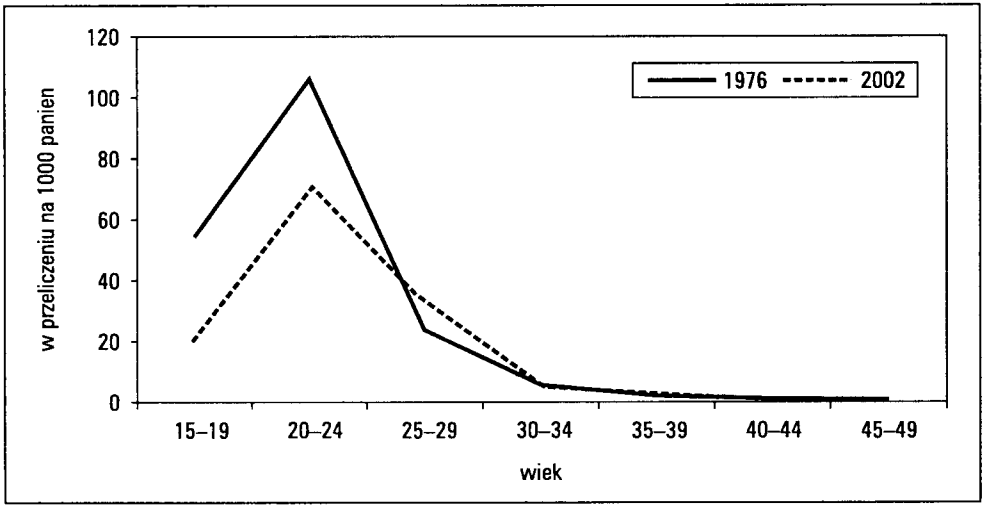
**Metoda** jest to sposób postępowania prowadzący do osiągnięcia wyznaczonego celu. Przeprowadzając badania naukowe, musimy zdawać sobie sprawę, że nie każde postępowanie jest dopuszczalne. Istnieją kryteria, które umożliwiają ocenę, czy dane

[2] Por. np.: K. Zając, *Zarys metod statystycznych*, Warszawa 1988.

postępowanie można określić mianem metody naukowej. Ustalanie tych kryteriów wchodzi między innymi w zakres filozofii, a dokładniej mieści się w ramach ontologii (filozofia bytu) oraz epistemologii, czyli teorii poznania.

Jako prawidłowość rozumieć będziemy względnie trwałe relacje charakteryzujące badane zjawisko. Gdyby relacje te miały charakter trwały, to mówilibyśmy o prawach. Tak jest na przykład w przypadku prawa powszechnego ciążenia lub prawa Archimedesa. Rozróżnienie prawa i prawidłowości łączy się przede wszystkim z możliwością kontrolowania czynników wpływających na kształtowanie się badanego zjawiska. Możliwość tej kontroli jest znacznie większa w naukach przyrodniczych niż w społecznych. W naukach społecznych najczęściej rozpatrujemy prawidłowości[3].

Wcześniej podaliśmy przykład mówiący, jakimi prawidłowościami charakteryzuje się zawieranie pierwszych małżeństw. Teraz wyjaśnimy, na czym polega względna trwałość badanego zjawiska. Na rysunku 1.1 przedstawiono liczbę pierwszych małżeństw według wieku zawartych przez kobiety w Polsce w roku 1976 i 2002 przeliczonych na 1000 kobiet w danej grupie wieku.




Rys. 1.1. Zawieranie pierwszych małżeństw wśród kobiet w Polsce w roku 1976 i 2002

Źródło: opracowanie własne na podstawie danych pochodzących z www.stat.gov.pl

Na uwagę zasługują następujące cechy:

- krzywe ilustrujące zawieranie związków małżeńskich w zależności od wieku kobiet mają charakterystyczny kształt; można wskazać wiek, któremu przypisana jest najwyższa częstość wstępowania w związek małżeński (20–24 lata),

[3] Por. np.: H. M. Blalock, *Statystyka dla socjologów*, Warszawa 1975.

- krzywe te mają podobny kształt w obydwu porównywanych latach kalendarzowych,
- w 2002 roku obniżyła się częstotliwość związków zawieranych przed 30 rokiem życia, a wzrosła po przekroczeniu tego wieku w porównaniu do 1976 roku.

Niektóre charakterystyki wstępowania w związki małżeńskie nie zmieniły się (kształt krzywej), a inne uległy przemianom i dlatego mówimy tylko o względnej trwałości zachowań matrymonialnych.

Dla ujawnienia przedstawionych prawidłowości konieczne jest posiadanie dużej liczby obserwacji. Tylko w takiej sytuacji mogliśmy wykreślić krzywe ilustrujące zawieranie pierwszych małżeństw według wieku kobiet przedstawione na rysunku 1.1. Zjawiska takie określamy mianem masowych.

Prawidłowości obserwowane w zjawiskach masowych kształtują się w efekcie oddziaływania dwóch grup przyczyn. Jedną stanowią przyczyny główne, a drugą uboczne.

**Przyczyny główne** oddziałują w sposób jednokierunkowy i sprawiają ujawnienie się prawidłowości. Kształtują tak zwany składnik **systematyczny zjawiska masowego**. Jest to charakterystyczna dla niego prawidłowość.

Dążąc do określenia zespołu przyczyn głównych w przypadku zawierania małżeństw, należy wziąć pod uwagę uwarunkowania biologiczne, społeczne oraz tradycję. Łączenie się ludzi w pary dla stworzenia warunków do wydania na świat i wychowania potomstwa pojawiło się na pewnym etapie rozwoju ludzkości. Później nadano temu związkowi charakter formalno-prawny. Okres życia, w którym małżeństwa są zawierane pozostaje w związku ze zdolnością rozrodczą gatunku ludzkiego. Najwyższa częstość związków przypada na ten wiek, w którym zdolność ta jest najwyższa. Częstość zawierania małżeństw kształtuje się w znacznej mierze pod wpływem społecznej oceny przypisywanej stanowi wolnemu (szczególnie kawalerskiemu i panieńskiemu). Jeśli ocena ta jest niska, to częstość małżeństw jest wysoka i na odwrót.

**Przyczyny uboczne** występują w dużej ilości i działają różnokierunkowo. Traktujemy je jako szeroko rozumiane uwarunkowania środowiskowe. W obserwowanej zbiorowości wywołują one odchylenia od prawidłowości. Odchylenia te ujawniają się w indywidualnych przypadkach. Rezultatem oddziaływania przyczyn ubocznych jest **przypadkowy składnik zjawiska masowego**.

W odniesieniu do zawierania małżeństw, jako rezultat oddziaływania przyczyn ubocznych można na przykład interpretować odchylenie wieku, w jakim była dana osoba w momencie pierwszego małżeństwa od wieku, w którym związki te są zawierane najczęściej. Odchylenie to jest rezultatem oddziaływania wielu czynników często trudnych do określenia.

Zadaniem statystyki jest dostarczenie odpowiednich metod, które prowadzą do liczbowego oszacowania zarówno składnika systematycznego, jak i przypadkowego rozpatrywanego zjawiska masowego.

## 1.2. Zbiorowości, jednostki, cecha statystyczna

Zjawiska masowe są obserwowane w zbiorowościach nazywanych również populacjami. Jednym z warunków poprawności wniosków sformułowanych w rezultacie przeprowadzonych badań, to znaczy ich adekwatności do badanej rzeczywistości, jest między innymi: dokładne zdefiniowanie populacji, w której realizuje się zjawisko, jednoznaczne przyporządkowanie jej jednostek oraz ustalenie cech kwalifikujących.

W badaniach statystycznych istotne znaczenie ma wyodrębnienie zbiorowości (populacji) generalnej. „**Zbiorowość generalna** albo **populacja generalna** jest to zbiór obiektów materialnych lub potencjalnych powtórzeń zjawiska, który jest przedmiotem zainteresowań badacza. Poddana badaniu zbiorowość musi być jednorodna ze względu na **cechy kwalifikujące**"[4].

Chcemy określić wiek kobiet w chwili pierwszego małżeństwa zawartego w Polsce w 2002 roku. Spośród ogółu mieszkańców Polski interesują nas zatem tylko osoby płci żeńskiej, które pierwszy raz wstąpiły w związek małżeński w 2002 roku. Do populacji generalnej zaliczymy więc osoby posiadające następujące cechy kwalifikujące[5]:

- miejsce zamieszkania w Polsce,
- płeć żeńska,
- zawarcie pierwszego małżeństwa w 2002 roku.

W tak wyróżnionej zbiorowości możemy przeprowadzać różne badania. Na początku musimy więc określić cel przedsięwzięcia. W rozważanym przypadku może nim być ustalenie wieku, w którym pierwsze małżeństwa są zawierane najczęściej. Wiek kobiet w chwili zawarcia pierwszego małżeństwa jest wówczas **cechą badaną.** Jeśli ponadto przypuszczamy, że wiek ten jest zróżnicowany w zależności od zamieszkania w mieście lub na wsi, od poziomu wykształcenia, od aktywności zawodowej itp., i chcemy sprawdzić słuszność tej hipotezy, to musimy odpowiednio rozszerzyć zbiór cech badanych. Pozwoli nam to zbadać uwarunkowania zachowań matrymonialnych panien zamieszkałych w Polsce w 2002 roku.

Jeśli interesuje nas struktura wydatków gospodarstw domowych uzyskujących dochody z pracy w województwie małopolskim w 2003 roku, to do populacji generalnej zaliczamy gospodarstwa domowe spełniające następujące warunki (posiadające cechy kwalifikujące):

- istnieją w 2003 roku na terenie województwa małopolskiego,
- uzyskują dochody z pracy.

Cechą badaną są wydatki ponoszone w celu nabycia dóbr i usług. Jeśli chcemy określić związek wydatków z poziomem dochodów, to dochody również wejdą do zbioru cech badanych.

---

[4] A. Iwasiewicz, Z. Paszek, *Statystyka z elementami statystycznych metod sterowania jakością*, Kraków 2004.

[5] Formalną definicję cechy kwalifikującej podają A. Iwasiewicz, Z. Paszek, *ibidem*.

Zbiorowość (populacja) generalna jest więc zbiorem jednostek nieidentycznych ze względu na badaną cechę[6]. Zbiorowość taką stanowią na przykład gospodarstwa domowe różniące się, a więc nieidentyczne pod względem wysokości dochodów i wydatków.

**Zbiorowość (populacja)** generalna składa się z **jednostek statystycznych**. Jednostka jest jej podstawowym elementem. W przykładzie dotyczącym zawierania małżeństw jednostką jest kobieta. W przypadku badania struktury wydatków jednostką jest gospodarstwo domowe. Możemy zatem mieć do czynienia z jednostkami, będącymi zespołami elementów, które w trakcie przeprowadzanego badania nie są traktowane indywidualnie. Gospodarstwo składa się z osób ujmowanych w badaniach jako jedność.

Rozważając strukturę wydatków gospodarstw domowych, możemy brać pod uwagę między innymi następujące cechy badane:

1) wydatki na wyróżnione dobra i usługi,
2) wysokość dochodów,
3) ilość dzieci,
4) ilość pracujących osób,
5) wiek głowy gospodarstwa domowego,
6) płeć głowy gospodarstwa domowego,
7) wykształcenie głowy gospodarstwa domowego,
8) obecność osób starszych.

Cechy: wydatki na wyróżnione dobra i usługi, wysokość dochodów, wiek głowy gospodarstwa domowego, liczba dzieci i liczba pracujących osób w gospodarstwie domowym można bezpośrednio wyrazić za pomocą liczb. Określamy je więc mianem **cech ilościowych**. Nazywamy je również zmiennymi. Zmienne: liczba dzieci i liczba pracujących osób w gospodarstwie domowym mogą przyjąć tylko niektóre wartości z danego przedziału liczbowego. Zbiór tych wartości jest skończony lub przeliczalny. Takie zmienne nazywamy zmiennymi typu skokowego lub zmiennymi skokowymi (dyskretnymi).

Wydatki, dochody, wiek głowy gospodarstwa domowego mogą przyjąć każdą wartość z przedziału $(0, \infty)$. Nazywamy je **zmiennymi typu ciągłego** lub zmiennymi ciągłymi. Ogólnie zmienne ciągle mogą przyjmować wartości z przedziału $(-\infty, \infty)$. Wtedy zbiór ich wartości jest równoliczny ze zbiorem liczb rzeczywistych. Jest więc mocy continuum.

Inna grupę stanowią cechy: płeć, wykształcenie głowy gospodarstwa domowego, obecność osób w starszym wieku. Nazywamy je **cechami jakościowymi**. Kategorie ich zostały określone werbalnie. Zamiast opisu słownego można wyróżnionym kategoriom przyporządkować umowne wartości. Najłatwiej jest to uczynić w przypadku

[6] Por. np.: S. Ostasiewicz, Z. Rusnak, U. Siedlecka, *Statystyka. Elementy teorii i zadania*, Wrocław 1999.

wykształcenia. Poziomowi wykształcenie: podstawowe, średnie, wyższe można odpowiednio przyporządkować liczbę lat nauki. Płeć występuje w dwóch przeciwnych stanach, którym można umownie przypisać wartości: 0 – mężczyzna, 1 – kobieta (lub na odwrót). Obecności osób w starszym wieku można przyporządkować 1, a nieobecność zakodować jako 0. Mamy wówczas do czynienia ze zmienną zerojedynkową.

Cechy statystyczne rozpatrujemy jako rezultat pomiaru[7]. Wykształcenie (podstawowe, wyższe, średnie), płeć głowy gospodarstwa domowego, obecność osób w starszym wieku mogą stanowić podstawę klasyfikacji gospodarstw domowych. Klasyfikacja jest podstawową operacją w nauce. Jest ona pomiarem na najniższym poziomie. Dzielimy wówczas jednostki zbiorowości ze względu na posiadanie cechy. Dążymy przy tym do uzyskania takiej klasyfikacji, aby w obrębie wyróżnionych kategorii znalazły się elementy bardziej do siebie podobne niż elementy zaliczone do różnych kategorii.

Dzielimy zbiorowość gospodarstw domowych na takie, w których głową jest mężczyzna i na te, w których jest nią kobieta. Klasyfikując gospodarstwa według wyróżnionych poziomów wykształcenia głowy gospodarstwa, wyróżnimy na przykład trzy grupy określone przez wykształcenie podstawowe, średnie i wyższe. Nazwy wyróżnionych kategorii są przyjmowane arbitralnie. Mogą być określone werbalnie lub umownie za pomocą liczb.

**Skalę najniższego poziomu pomiaru nazywamy skalą nominalną.** Należy zwrócić uwagę na to, że na tym poziomie pomiaru podział na kategorie nie stanowi podstawy do ich porządkowania. Jeśli nawet słowne określenie kategorii zastąpimy liczbami, to nadal nie upoważnia nas to do wykonywania na nich jakichkolwiek działań arytmetycznych.

Warunkiem stosowania procedur statystycznych jest uzyskanie klasyfikacji wyczerpującej oznaczającej, że każdy element został zaliczony do odpowiedniej kategorii oraz rozłącznej, a więc takiej, w wyniku której każdy element należy tylko do jednej kategorii.

Skala nominalna jest symetryczna i przechodnia. Symetryczność oznacza, że jeśli relacja zachodzi między A i B, to zachodzi również między B i A. Na przykład, jeśli syn jest podobny do ojca, to ojciec jest podobny do syna. Przechodniość występuje, jeśli A = B i B = C, to A = C. Jeśli Jan jest absolwentem tego samego uniwersytetu, co Maria, a Maria tego samego, co Anna, to Jan ukończył ten sam uniwersytet co Anna.

**Wyższym poziomem pomiaru niż skala nominalna jest skala porządkowa.** Oprócz klasyfikacji umożliwia ona porządkowanie kategorii pod względem stopnia natężenia danej cechy, ale bez możliwości określenia odległości. Skala ta może ustalać porządek słaby lub mocny. Słabe uporządkowanie uzyskujemy, gdy dopuszczamy kryterium „mniejszy lub równy", „większy lub równy". Taka skala charakteryzuje się

[7] Por. np.: H. M. Blalock, *op. cit.*, Warszawa 1975.

symetrycznością oraz przechodniością. W wypadku porządku mocnego, wyznaczonego przez nierówność „większy niż", „mniejszy niż" pojawia się asymetria. Na przykład, jeśli Paweł jest wyższy od Piotra, to nie może być na odwrót. Jeśli natomiast Jan jest wyższy od Piotra, a Piotr jest wyższy od Pawła, to Jan jest wyższy od Pawła. Skala porządkowa zawsze jest przechodnia. W skali tej można elementy porządkować, ale nie można porównywać różnic między nimi. Oznacza to, że nawet jeśli pomiar na skali porządkowej jest wyrażony liczbowo, to podobnie jak w przypadku skali nominalnej na liczbach tych nie możemy wykonywać żadnych działań arytmetycznych. Skala porządkowa posiada wszystkie własności skali nominalnej, a ponadto możliwość porządkowania. Od skali porządkowej można przejść do nominalnej, ale nie na odwrót.

**Skala nominalna i porządkowa są pomiarem w szerszym sensie, który pozwala klasyfikować i rangować obiekty. Pomiar w sensie wąskim umożliwia: klasyfikowanie, rangowanie oraz określanie odległości między obiektami. Warunki te spełnia skala interwałowa. Jeśli ponadto możliwe jest niearbitralne ustalenie punktu zerowego, to taką skalę nazywamy ilorazową lub stosunkową.** W tym przypadku można bowiem porównywać stosunki. Przykładem skali interwałowej jest dochód gospodarstwa domowego. Rozpatrujemy na przykład cztery gospodarstwa domowe posiadające miesięczne dochody w przeliczeniu na jedną osobę równe: 500 zl, 750 zł, 1500 zł i 1750 zł. Możemy je klasyfikować i porządkować według wysokości dochodu. Możemy porównać różnice bezwzględne wynoszące odpowiednio: 250 zl, 750 zl, 250 zł. Mówimy, że różnica pomiędzy dochodami pierwszego i drugiego oraz trzeciego i czwartego jest jednakowa i wynosi 250 zl. Skala, w jakiej jest mierzony dochód, posiada niearbitralne zero. Możemy więc obliczać stosunki. Powiemy wówczas, że dochód trzeciego gospodarstwa jest dwa razy większy od dochodu gospodarstwa drugiego.

**Wyróżnione poziomy pomiarów tworzą skalę kumulatywną.** Najwyższy poziom pomiaru posiada wszystkie własności skal poziomów niższych. Możliwe jest schodzenie z góry w dół; od najwyższego do najniższego poziomu, ale nie na odwrót. Dochody wyrażone w skali interwałowej (ilorazowej) możemy zastąpić rangami odpowiednio: 500 zł – 1; 750 zł – 2; 1500 zł – 3; 1750 zl – 4 (skala porządkowa) albo kategoriami: 500 zł – dochód niski; 750 zł – dochód średni; 1500 zl – dochód wysoki; 1750 zł – dochód bardzo wysoki (skala nominalna).

Rozważane dotychczas określenia i definicje (jednostka, cecha) odnosiliśmy do populacji generalnej. Jeśli jednak chcemy zbadać prawidłowość charakteryzującą interesujące nas zjawisko, to nie musimy obserwować wszystkich jednostek. Możemy uzyskać zadowalające rezultaty, ograniczając się do badania wyodrębnionego w odpowiedni sposób podzbioru populacji generalnej. **Ten podzbiór nazywamy zbiorowością (populacją) próbną lub krótko – próbą.** W tym miejscu należy zwrócić uwagę na to, że przedmiotem zainteresowania jest prawidłowość w zbiorowości generalnej i to ona podlega obserwacji. Próba musi być więc wybrana w odpowiedni sposób. Najogólniej rozróżniamy dobór celowy i losowy. W przypadku **doboru celowego** przeprowadzający badania decyduje, która jednostka wejdzie w skład próby. **Dobór jest**

losowy, jeśli to przypadek decyduje o tym, który element dostanie się do próby. Taką próbę nazywamy próbą losową lub próbą statystyczną[8]. Próbę tę poddajemy badaniu statystycznemu, a na podstawie uzyskanych rezultatów wnioskujemy o prawidłowości w populacji generalnej.

Wnioskowanie o prawidłowości w populacji generalnej na podstawie próby statystycznej jest uzasadnione tylko wówczas, gdy struktura próby jest podobna do struktury populacji generalnej. **Jest to próba reprezentatywna.** Dla zapewnienia reprezentatywności staramy się wybrać próbę, posługując się odpowiednimi technikami. Techniki te określamy mianem **schematów losowania.** Schematy losowania są szczegółowo rozpatrywane w ramach wyodrębnionego działu statystyki, którym jest metoda reprezentacyjna[9]. Jednym z prostszych schematów jest losowanie indywidualne. Do próby wybierane są jednostki statystyczne. Możemy przy tym zastosować **losowanie ze zwracaniem,** zgodnie z którym wylosowany element wraca do zbiorowości generalnej przed następnym wyborem. Otrzymujemy w ten sposób próbę **prostą, niezależną.** Jest to próba z **powtórzeniami.** Jeśli po wylosowaniu element wybrany nie wraca do populacji generalnej, a więc nie może ponownie wejść do próby, to schemat taki jest schematem losowania bez zwracania. W rezultacie otrzymujemy **zależną,** która jest **próbą bez powtórzeń.**

Innym schematem jest losowanie zespołowe. Na przykład, jeśli chcemy zbadać warunki mieszkaniowe w dużym mieście, zamiast losować indywidualnie poszczególne mieszkania, możemy wylosować naturalne ich zespoły, jakimi są bloki mieszkalne. W przypadku losowania indywidualnego jednostka losowania pokrywa się z jednostką badania. W losowaniu zespołowym jednostka losowania jest zespołem jednostek badania.

**Losowy dobór próby zwiększa prawdopodobieństwo uzyskania próby reprezentatywnej. Prawdopodobieństwo to wzrasta, gdy próba jest dostatecznie liczna. Jeśli zatem próba jest losowa i dostatecznie liczna, to jest wysoce prawdopodobne, że jest ona reprezentatywna.**

Badania statystyczne mogą przyjąć charakter badań całkowitych, wyczerpujących, jeśli obserwacji podlega cała populacja generalna. Gdy ograniczamy się do przeprowadzenia badań na podstawie próby, to badania te określamy mianem częściowych, niewyczerpujących. Rezygnujemy z badań całkowitych, ponieważ wymagają one znacznych nakładów finansowych. Długi jest również okres oczekiwania na rezultaty. Badania wyczerpujące przeważnie są kosztowne i czasochłonne. Innym powodem prze-

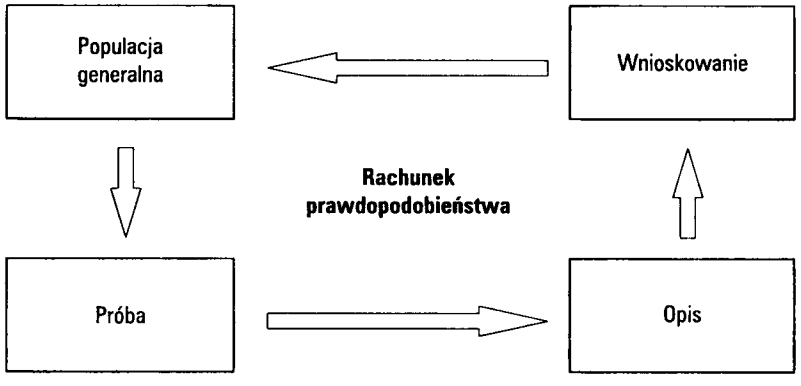
---

[8] Formalną definicję próby losowej można znaleźć na przykład M. Fisz, *Rachunek prawdopodobieństwa i statystyka matematyczna*, Warszawa 1965.

[9] Różne schematy losowania są rozpatrywane na przykład w pracach: C. Bracha, *Teoretyczne podstawy metody reprezentacyjnej*, Warszawa 1996, J. Steczkowskiego, *Metoda reprezentacyjna w poglądach jej twórców*, PN 1990, nr 513; idem, *Metoda reprezentacyjna w badaniach zjawisk ekonomiczno-społecznych*, Warszawa–Kraków 1995, R. Zasępa, *Zarys metody reprezentacyjnej*, „Biblioteka Wiadomości Statystycznych", t. 30, Warszawa 1991.

prowadzania badań metodą reprezentacyjną jest konieczność zniszczenia obiektu w trakcie badania. Sytuacja taka występuje na przykład w badaniu wytrzymałości elementu konstrukcyjnego w ramach statystycznej kontroli jakości. Jeśli rozpatrujemy populacje nieskończone, to wyniki badań również mogą być traktowane jako niewyczerpujące. Posłużymy się tutaj innym przykładem z zakresu demografii. Umieralność populacji można przedstawić za pomocą tablicy trwania życia[10]. Jedną z zawartych w niej charakterystyk jest prawdopodobieństwo zgonu osoby, która dożyła do określonego wieku (na przykład ukończyła 45 lat). Wartości prawdopodobieństwa obliczono na podstawie danych odnoszących się do dokładnie wyodrębnionej ze względu na czas i miejsce populacji. Prawdopodobieństwo to możemy również odnieść do zbiorowości, której życie przebiega w podobnych warunkach do tej zbiorowości, dla której skonstruowano tablicę.

Na rysunku 1.2 przedstawiono schemat postępowania badawczego w przypadku wnioskowania o prawidłowości w populacji generalnej na podstawie wyników zaobserwowanych w próbie statystycznej.




Rys. 1.2. Schemat badania statystycznego

Źródło: opracowanie własne.

Badania statystyczne rozpoczynamy od wyznaczenia celu. Dla rozważanych tutaj jako przykładowe badania z zakresu demografii oraz z analizy budżetów domowych cel można ująć jako:

- określenie uwarunkowań zawierania pierwszych związków małżeńskich przez kobiety w Polsce w okresie transformacji ekonomiczno-społecznej,
- poznanie czynników kształtujących strukturę wydatków gospodarstw domowych województwa małopolskiego w 2003 roku.

[10] Tablice trwania życia są dokładnie przedstawione w każdym podręczniku do demografii, jak np.: J. Z. Holzer, *Demografia*, Warszawa 2003, J. Kurkiewicz, *Podstawowe metody analizy demograficznej*, Warszawa 1992.

Dobrze wyznaczony cel pozwoli:
1) sformułować odpowiednie hipotezy,
2) poprawnie wyodrębnić populację generalną,
3) ustalić listę zmiennych odzwierciedlających badaną prawidłowość.

Jeśli zdecydujemy się na badania częściowe (niewyczerpujące), musimy wylosować próbę[11]. Próbę tę poddajemy bezpośredniemu badaniu. Opisujemy ją za pomocą odpowiednich charakterystyk. Informacje o każdej jednostce statystycznej zastępujemy wówczas odpowiednimi miarami, takimi jak na przykład wartości przeciętne. Postępowanie badawcze opiera się tutaj na rozumowaniu dedukcyjnym. Ten dział statystyki nazywamy **statystyką opisową**. Mając jednak na uwadze, że cały czas interesuje nas prawidłowość w populacji generalnej, nie możemy pozostać na poziomie próby. Uzyskane wyniki stanowią podstawę do wnioskowania o tej prawidłowości w zbiorowości generalnej. Metody tego rodzaju wnioskowania stanowią przedmiot **statystyki matematycznej**, określanej również mianem **statystyki indukcyjnej**. Od rozumowania dedukcyjnego przechodzimy do rozumowania indukcyjnego. W badaniach metodami statystycznymi opiera się ono na rachunku prawdopodobieństwa[12].

Badanie statystyczne musi opierać się na rachunku prawdopodobieństwa, ponieważ w rezultacie wnioskowania formułujemy sądy o prawidłowości w populacji generalnej wyprowadzone z niepełnej informacji. Pochodzi ona z próby statystycznej. W tych warunkach prawdziwość naszych sądów nie jest pewna. Niezbędne są zatem takie reguły postępowania, które pozwolą nam przyjąć odpowiednio wysokie prawdopodobieństwo prawdziwości naszych wniosków lub niskie prawdopodobieństwo wniosków fałszywych.

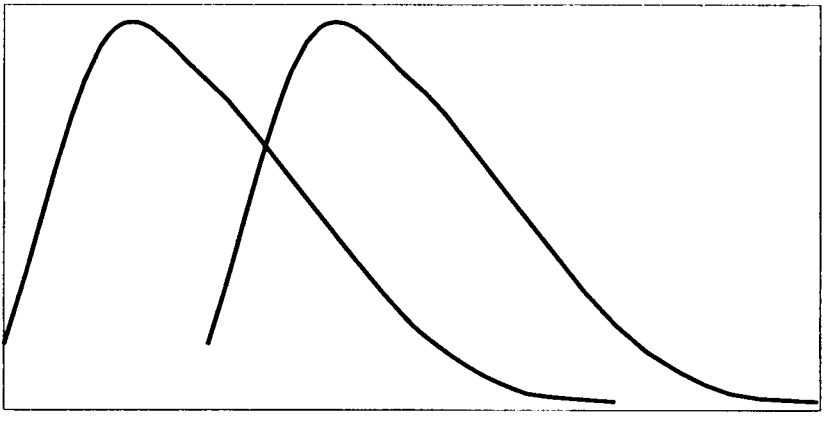
Konieczność uwzględniania prawdopodobieństwa łączy się nie tylko z samym wnioskowaniem na podstawie próby. Prawdopodobieństwo pojawia się również przypadku badania całkowitego (wyczerpującego). W rzeczywistości mamy bowiem do czynienia wyłącznie ze zdarzeniami losowymi, którym przypisana jest – zgodnie z regułami rachunku prawdopodobieństwa – szansa realizacji.

## 1.3. Typy prawidłowości

Prawidłowości, jakimi charakteryzują się zjawiska masowe, możemy rozpatrywać jako prawidłowości, które występują:
1) w rozkładzie liczebności populacji według wartości zmiennej,
2) w związkach między zjawiskami,
3) w dynamice, czyli w rozwoju zjawiska w czasie.

[11] Szczegółowe omówienie postępowania w przypadku losowania próby można znaleźć w pracy R. Zasępa, *Zarys metody reprezentacyjnej*, *op. cit.*
[12] H. M. Blalock, *op. cit.*

Aby wyjaśnić prawidłowości występujące w rozkładach liczebności, rozważymy wyniki, jakie uzyskali po zaliczeniu pierwszego roku studenci dwóch wydziałów pewnej uczelni. Ponieważ nie możemy porównywać wszystkich studentów, posłużymy się pewnymi charakterystykami liczbowymi. Poziom nauczania na wydziałach możemy porównać, posługując się średnią ocen, zakładając że studentom stawiane są jednakowe wymagania. Sytuację tę zilustrowano na rysunku 1.3a.



**Rys. 1.3a.** Rozkłady liczebności o różnym położeniu

W tym przypadku mówimy, że rozkłady liczebności mają różne położenie w układzie współrzędnych. Rozkład przesunięty w prawo (w stronę plus nieskończoności) charakteryzuje się wyższą wartością przeciętną. Jedną z nich jest średnia arytmetyczna. **Wartości przeciętne nazywamy miarami położenia**[13].

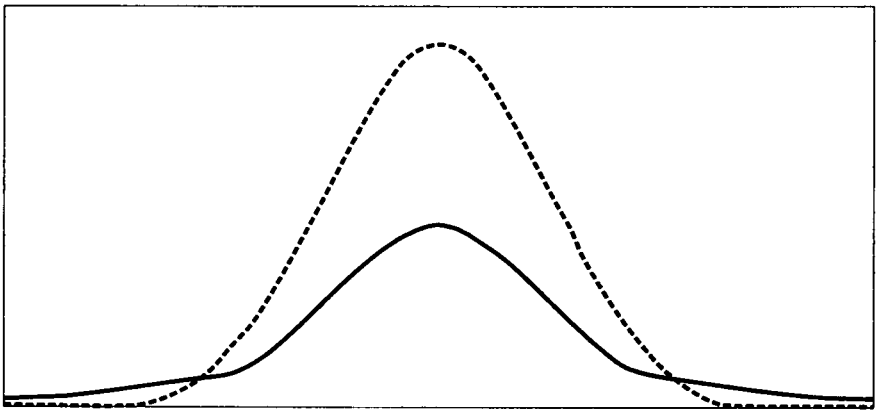
Średnia ocen nie jest jedyną charakterystyką różnicującą wyniki nauczania studentów porównywanych wydziałów. Oceny mogą być mniej lub bardziej zróżnicowane w porównaniu z wartością przeciętną. W tym przypadku mówimy, że rozkłady mogą mieć różną zmienność, a charakterystyki stosowane do jej pomiaru nazywamy **miarami zmienności**. Tę cechę rozkładów liczebności przedstawiono na rysunku 1.3b.

Rozkład nakreślony linią przerywaną charakteryzuje się mniejszą zmiennością od rozkładu narysowanego linią ciągłą.

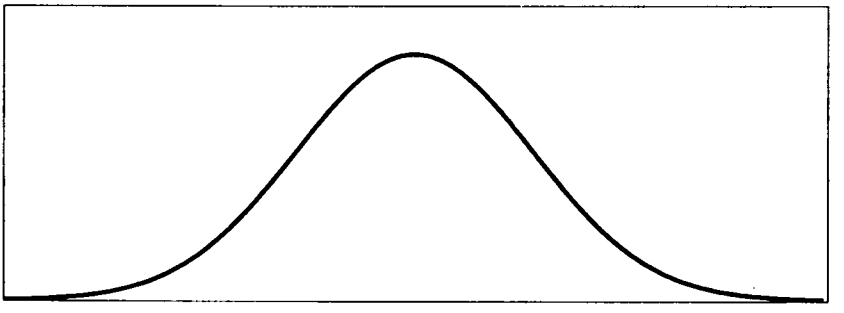
Możemy zapytać, czy na porównywanych wydziałach przeważają studenci osiągający oceny wyższe od średniej, czy też więcej jest osób z ocenami poniżej średniej charakteryzującej dany wydział. Mówimy wówczas o asymetrii rozkładu liczebności. Tę właściwość przedstawiono na rysunkach 1.3c–1.3e. Jeśli liczba studentów, którzy uzyskali oceny niższe od średniej, jest równa liczbie studentów z ocenami wyższymi

[13] Poszczególne miary będą szczegółowo omówione w rozdziale 3.

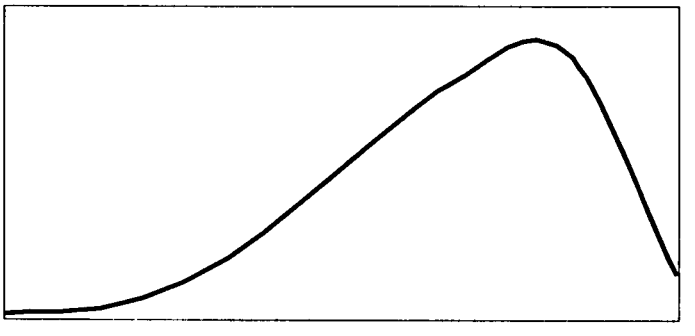
niż średnia, to mamy do czynienia z rozkładem symetrycznym. Jeśli w danej zbiorowości przeważają jednostki o wartościach zmiennej większych od średniej, to mówimy o asymetrii lewostronnej.



Rys. 1.3b. Rozkłady liczebności o różnej zmienności

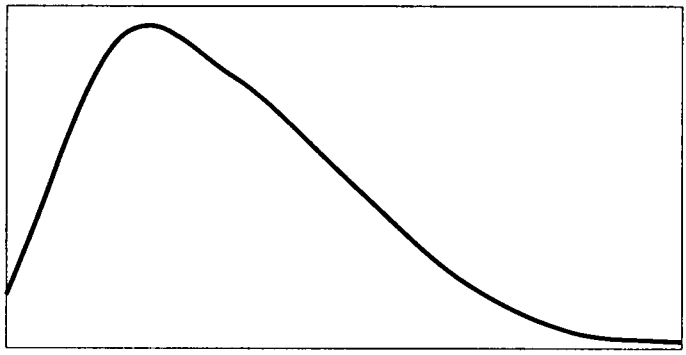


Rys. 1.3c. Symetryczny rozkład liczebności



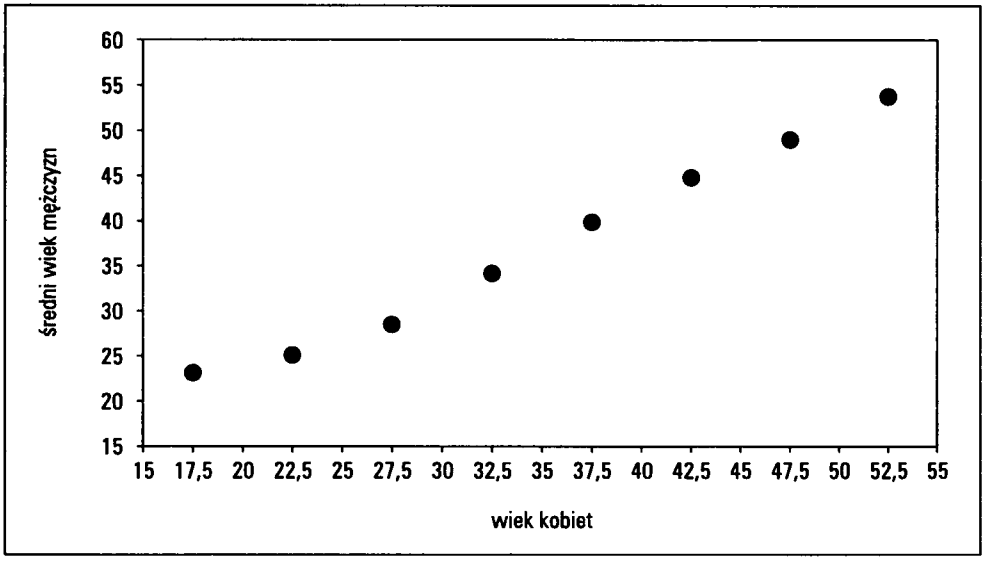
Rys. 1.3d. Asymetria lewostronna

Jeśli natomiast częściej występują jednostki o wartościach zmiennej mniejszych od średniej, to mamy do czynienia z asymetrią prawostronną.



Rys. 1.3e. Asymetria prawostronna

Na rysunku 1.4 przedstawiono prawidłowości przejawiające się w związkach między zjawiskami[14]. Wzięto pod uwagę wiek mężczyzn i kobiet w chwili zawarcia małżeństwa.



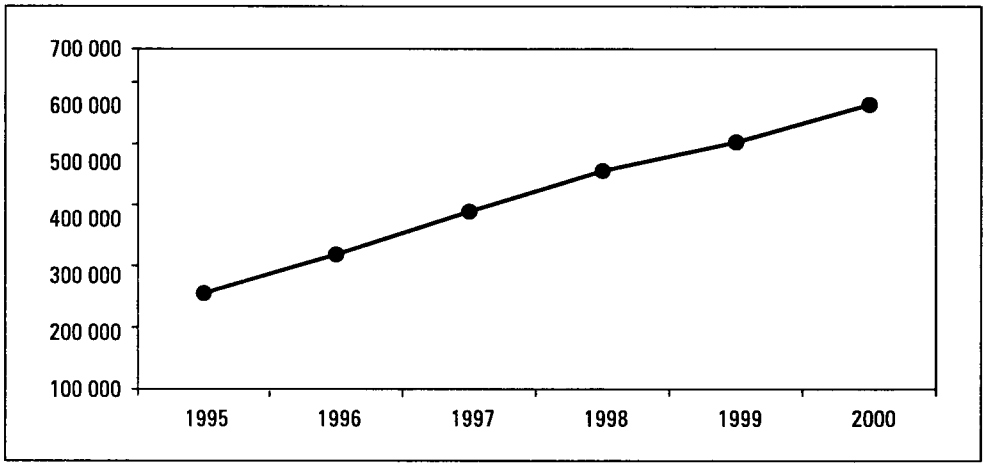

Rys. 1.4. Wiek mężczyzn i kobiet w chwili zawarcia małżeństwa

Źródło: opracowanie własne na podstawie danych GUS.

[14] Problemy badania współzależności i różne typy związków między zjawiskami stanowią przedmiot rozdziału 4.

Układ punktów świadczy o tym, że istnieje związek między porównywanymi zmiennymi. Podwyższeniu wieku kobiet odpowiada podwyższenie wieku mężczyzn w chwili zawarcia małżeństwa.

Rysunek 1.5 jako przykład prawidłowości przejawiającej się w rozwoju zjawisk w czasie przedstawia dynamikę Produkt Krajowego Brutto w Polsce w latach 1995–2000.




Rys. 1.5. Produkt Krajowy Brutto w Polsce w latach 1995–2000 (w mln zł)

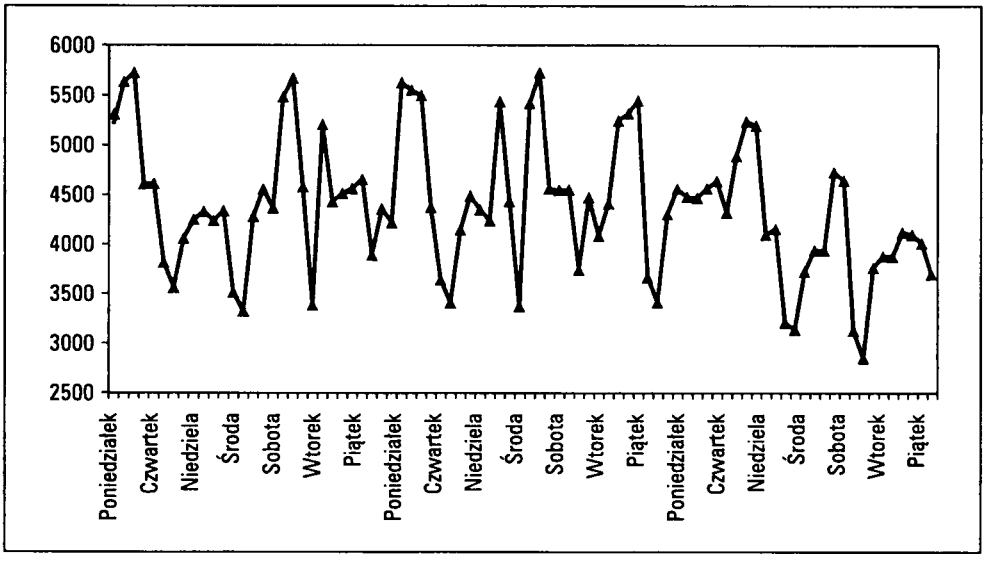
Źródło: opracowanie własne na podstawie danych GUS.

Produkt Krajowy Brutto w Polsce w latach 1995–2000 charakteryzował się rosnącą tendencją rozwojową. W rozpatrywanym przykładzie jednostkę obserwacji stanowi rok. Niektóre zjawiska podlegają wahaniom okresowym (na przykład sezonowym). Obserwujemy je w okresach krótszych (na przykład w dniach, tygodniach, kwartałach, miesiącach, półroczach).

Śledząc zmiany liczby urodzeń, stwierdzamy, że w rozważanym okresie cyklicznie powtarzają się dni tygodnia, w których liczba urodzeń żywych jest wysoka (początek tygodnia: wtorek, środa) oraz takie, w których jest ona niska (koniec tygodnia: sobota, niedziela).

Po dokładnym określeniu celu badania, sformułowaniu hipotez, wyodrębnieniu populacji generalnej wraz z tworzącymi ją jednostkami oraz po ustaleniu cech istotnych z punktu widzenia celu badania, przystępujemy do zgromadzenia niezbędnych danych statystycznych. Gromadzenie danych nazywamy obserwacją statystyczną. Jest to pierwszy etap badania statystycznego. Przystępując do gromadzenia danych, należy najpierw zorientować się, czy istnieją już źródła zawierające interesujące nas dane, nawet jeśli zostały one zgromadzone w innym celu niż ten, który chcemy osiągnąć w rezultacie naszych badań. W sytuacji gdy żadne znane nam i dostępne źródło nie zawiera takich danych, zdecydujemy się na samodzielne ich zbieranie.

Rysunek 1.6 przedstawia liczbę urodzeń żywych według dni tygodnia w poszczególnych miesiącach w 2002 roku[15].




**Rys. 1.6.** Liczba urodzeń żywych według dni tygodnia w miesiącach w 2002 roku

Źródło: opracowanie własne na podstawie danych GUS.

Spośród istniejących źródeł danych najczęściej wykorzystujemy zasoby Głównego Urzędu Statystycznego (GUS). Takim źródłem jest Narodowy Spis Powszechny (NSP) oraz bieżąca rejestracja danych o zdarzeniach, takich jak: urodzenia, małżeństwa i zgony, które są zapisywane w specjalnych księgach. Główny Urząd Statystyczny gromadzi również dane o procesach gospodarczych i społecznych. Na szczególne zwrócenie uwagi zasługują systematycznie prowadzone badania budżetów gospodarstw domowych. O ile informacje uzyskane w rezultacie Spisu Powszechnego i rejestracji bieżącej odnoszą się do całej populacji wyodrębnionej pod względem czasowym i przestrzennym, to badania budżetów rodzinnych są przykładem badań reprezentacyjnych, którymi objęta jest wybrana w sposób losowy subpopulacja nazywana próbą statystyczną.

[15] Na osi odciętych nie mogły być zaznaczone wszystkie dni tygodnia ze względu na ograniczony rozmiar rysunku.

# Opracowanie materiału statystycznego

Rozdział 2

## 2.1. Grupowanie statystyczne

Zebrany materiał statystyczny, sprawdzony pod względem kompletności i poprawności poddajemy grupowaniu. Grupowanie, czyli klasyfikacja jednostek według kategorii rozpatrywanych cech, jest pierwszym etapem opracowania materiału statystycznego. W rezultacie otrzymujemy szeregi statystyczne. **Szereg statystyczny** jest to zbiór uporządkowanych (rosnąco lub malejąco) wartości zmiennej. Jest to empiryczny (zaobserwowany) rozkład ogólnej liczebności populacji pomiędzy wyróżnione wartości zmiennej. Odzwierciedla on strukturę populacji według rozważanej cechy.

### 2.1.1. Grupowanie jednostek według cech jakościowych

Problem grupowania jest mniej skomplikowany, jeśli jako kryterium przyjmiemy cechę jakościową (pomiar w skali nominalnej) niż wówczas, gdy grupujemy je według cech ilościowych. Szereg składa się z dwóch kolumn. W pierwszej wymienione są wyróżnione kategorie, a w drugiej – przyporządkowana im liczba jednostek (liczebność).

Kategorie cech mogą być określone werbalnie lub można im przyporządkować umowne wartości liczbowe. Rezultaty grupowania zestawiamy w tablicach.

W tablicy 2.1 podano jako przykład klasyfikację (strukturę) ludności Polski ze względu na stan cywilny w 2002 roku. Każdej kategorii stanu cywilnego przyporządkowano zaobserwowaną liczbę osób. Są to liczebności bezwzględne. Liczebności te zależą od ogólnej liczebności populacji, a więc są nieporównywalne. Jeśli chcemy porównać na przykład rozkłady liczebności według stanu cywilnego dla Polski ogółem oraz dla zbiorowości mężczyzn i kobiet, lepiej jest posłużyć się liczebnościami względnymi. Mogą one przybierać postać częstości względnej. Uzyskamy ją, dzieląc liczbę jednostek przyporządkowanych poszczególnym kategoriom przez liczebność ogólną zgodnie z podanym niżej wzorem:

$$f_i' = \frac{f_i}{\sum_{i=1}^{k} f_i} = \frac{f_i}{n} \tag{2.1}$$

gdzie:

$f_i$ – liczba jednostek przyporządkowanych wyróżnionej ($i$-tej) kategorii cechy statystycznej,

$k$ – liczba wyróżnionych kategorii cechy,

$\sum_{i=1}^{k} f_i = n$ – ogólna liczebność populacji.

**Tablica 2.1.** Ludność w wieku przynajmniej 15 lat według płci i stanu cywilnego faktycznego w Polsce w 2002 roku* (I)

| Lp. | Stan cywilny | Ogółem | Mężczyźni | Kobiety |
|---|---|---|---|---|
| | | w liczbach bezwzględnych ($f_i$) | | |
| 1. | kawalerowie/panny | 8 732 027 | 4 862 997 | 3 869 030 |
| 2. | żonaci/zamężne | 18 099 824 | 9 021 417 | 9 078 407 |
| 3. | pozostający w związku małżeńskim | 17 703 770 | 8 823 388 | 8 880 382 |
| 4. | pozostający w związku partnerskim | 396 054 | 198 029 | 198 025 |
| 5. | wdowcy/wdowy | 2 871 010 | 424 696 | 2 446 314 |
| 6. | rozwiedzeni/rozwiedzione | 1 030 031 | 394 202 | 635 829 |
| 7. | separowani/separowane | 309 154 | 133 961 | 175 193 |
| 8. | nieustalony | 246 382 | 124 833 | 121 549 |
| | OGÓŁEM | 31 288 428 | 14 962 106 | 16 326 322 |

* Uwaga: Pozycje (3) i (4) dają w sumie pozycję (2).

Źródło: Tablice wynikowe Narodowego Spisu Ludności i Mieszkań 2002, www.stat.gov.pl

W tablicy 2.2 podano rozważany szereg statystyczny, w którym liczebności są podane jako częstości względne uzyskane zgodnie ze wzorem (2.1).

Interpretacja częstości względnych może nastręczać trudności[16]. Na przykład częstość kawalerów oznacza, że na jednego mężczyznę przypada 0,325 kawalerów. Dlatego wygodniej jest częstość tę odnieść do umownej liczebności populacji np.: 100 (procenty), 1000 (promile) itd. W tablicy 2.3 podano procenty, a więc liczebności względne w przeliczeniu na 100. (odpowiednio: ogółem, mężczyzn, kobiet). Kawalerowie stanowili więc 32,5% ogółu mężczyzn w Polsce w 2002 roku.

[16] Częstości względne są wykorzystywane do szacowania prawdopodobieństwa wystąpienia zdarzenia losowego.

**Tablica 2.2.** Ludność w wieku przynajmniej 15 lat według płci i stanu cywilnego faktycznego w Polsce w 2002 roku* (II)

| Lp. | Stan cywilny | Ogółem | Mężczyźni | Kobiety |
|---|---|---|---|---|
| | | częstości względne ($f_i'$) | | |
| 1. | kawalerowie/panny | 0,279 | 0,325 | 0,237 |
| 2. | żonaci/zamężne | 0,579 | 0,603 | 0,556 |
| 3. | pozostający w związku małżeńskim | 0,566 | 0,590 | 0,544 |
| 4. | pozostający w związku partnerskim | 0,013 | 0,013 | 0,012 |
| 5. | wdowcy/wdowy | 0,092 | 0,028 | 0,150 |
| 6. | rozwiedzeni/rozwiedzione | 0,033 | 0,026 | 0,039 |
| 7. | separowani/separowane | 0,010 | 0,009 | 0,011 |
| 8. | nieustalony | 0,008 | 0,008 | 0,007 |
| | OGÓŁEM | 1,000 | 1,000 | 1,000 |

* Uwaga: Pozycje (3) i (4) dają w sumie pozycję (2).

Źródło: Tablice wynikowe Narodowego Spisu Ludności i Mieszkań 2002, www.stat.gov.pl

**Tablica 2.3.** Ludność w wieku przynajmniej 15 lat według płci i stanu cywilnego faktycznego w Polsce w 2002 roku* (w %)

| Lp. | Stan cywilny | Ogółem | Mężczyźni | Kobiety |
|---|---|---|---|---|
| | | w procentach | | |
| 1. | kawalerowie/panny | 27,9 | 32,5 | 23,7 |
| 2. | żonaci/zamężne | 57,9 | 60,3 | 55,6 |
| 3. | pozostający w związku małżeńskim | 56,6 | 59,0 | 54,4 |
| 4. | pozostający w związku partnerskim | 1,3 | 1,3 | 1,2 |
| 5. | wdowcy/wdowy | 9,2 | 2,8 | 15,0 |
| 6. | rozwiedzeni/rozwiedzione | 3,3 | 2,6 | 3,9 |
| 7. | separowani/separowane | 1,0 | 0,9 | 1,1 |
| 8. | nieustalony | 0,8 | 0,8 | 0,7 |
| | OGÓŁEM | 1,000 | 1,000 | 1,000 |

* Uwaga: Pozycje (3) i (4) dają w sumie pozycję (2).

Źródło: Tablice wynikowe Narodowego Spisu Ludności i Mieszkań 2002, www.stat.gov.pl

Innym przykładem grupowania jest klasyfikacja według jednostek terytorialnych (regiony geograficzne, jednostki administracyjne) lub czasowych (lata, kwartały, miesiące, dni). Przykładem klasyfikacji według jednostek terytorialnych jest podany w tablicy 2.4 rozkład liczby ludności w Polsce według województw w 2002 roku.

**Tablica 2.4.** Liczba ludności Polski według województw w 2002 roku

| Województwo | Ogółem | Miasta | Wieś | Ogółem | Miasta | Wieś |
|---|---|---|---|---|---|---|
| | w liczbach bezwzględnych | | | w procentach | | |
| Dolnośląskie | 2 907 212 | 2 076 121 | 831 091 | 7,60 | 8,79 | 5,68 |
| Kujawsko-pomorskie | 2 069 321 | 1 288 519 | 780 802 | 5,41 | 5,46 | 5,34 |
| Lubelskie | 2 199 054 | 1 025 566 | 1 173 488 | 5,75 | 4,34 | 8,03 |
| Lubuskie | 1 008 954 | 651 045 | 357 909 | 2,64 | 2,76 | 2,45 |
| Łódzkie | 2 612 890 | 1 697 745 | 915 145 | 6,83 | 7,19 | 6,26 |
| Małopolskie | 3 232 408 | 1 626 865 | 1 605 543 | 8,46 | 6,89 | 10,98 |
| Mazowieckie | 5 124 018 | 3 312 618 | 1 811 400 | 13,40 | 14,03 | 12,39 |
| Opolskie | 1 065 043 | 560 064 | 504 979 | 2,79 | 2,37 | 3,45 |
| Podkarpackie | 2 103 837 | 853 053 | 1 250 784 | 5,50 | 3,61 | 8,56 |
| Podlaskie | 1 208 606 | 711 572 | 497 034 | 3,16 | 3,01 | 3,40 |
| Pomorskie | 2 179 900 | 1 484 838 | 695 062 | 5,70 | 6,29 | 4,75 |
| Śląskie | 4 742 874 | 3 751 393 | 991 481 | 12,41 | 15,89 | 6,78 |
| Świętokrzyskie | 1 297 477 | 595 388 | 702 089 | 3,39 | 2,52 | 4,80 |
| Warmińsko-mazurskie | 1 428 357 | 860 229 | 568 128 | 3,74 | 3,64 | 3,89 |
| Wielkopolskie | 3 351 915 | 1 934 790 | 1 417 125 | 8,77 | 8,19 | 9,69 |
| Zachodniopomorskie | 1 698 214 | 1 180 559 | 517 655 | 4,44 | 5,00 | 3,54 |
| Ogółem | 38 230 080 | 23 610 365 | 14 619 715 | 100 | 100 | 100 |

Źródło: Tablice wynikowe Narodowego Spisu Ludności i Mieszkań 2002, www.stat.gov.pl

W tablicy 2.5 zamieszczono szereg czasowy przedstawiający liczbę urodzeń żywych, liczbę zgonów i przyrost naturalny w przeliczeniu na 1000 mieszkańców w Polsce w latach 1990–2002.

Tablica 2.5. Liczba urodzeń żywych, liczba zgonów i przyrost naturalny w Polsce w latach 1990–2002

| Rok | Urodzenia żywe | Zgony | Przyrost naturalny |
|---|---|---|---|
| | na 1000 mieszkańców | | |
| 1990 | 14,3 | 10,2 | 4,1 |
| 1991 | 14,3 | 10,6 | 3,7 |
| 1992 | 13,5 | 10,3 | 3,2 |
| 1993 | 12,8 | 10,2 | 2,6 |
| 1994 | 12,5 | 10,0 | 2,5 |
| 1995 | 11,2 | 10,0 | 1,2 |
| 1996 | 11,1 | 10,0 | 1,1 |
| 1997 | 10,7 | 9,8 | 0,9 |
| 1998 | 10,2 | 9,7 | 0,5 |
| 1999 | 9,9 | 9,9 | 0,0 |
| 2000 | 9,8 | 9,5 | 0,3 |
| 2001 | 9,5 | 9,4 | 0,1 |
| 2002 | 9,3 | 9,4 | -0,1 |

Źródło: www.stat.gov.pl

## 2.1.2. Grupowanie według cech ilościowych (zmiennych)

W przypadku cech ilościowych zamiast określonych słownie kategorii występują wyrażone liczbowo warianty zmiennej. W tym przypadku zadanie jest bardziej skomplikowane. Najczęściej sami musimy zadecydować, jakie warianty zmiennej wyróżnimy.

Przedstawimy procedurę konstruowania szeregu statystycznego dla zmiennych typu skokowego i ciągłego. Zilustrujemy to odpowiednimi przykładami.

### Przykład 2.1
### Zmienna typu skokowego

Przeprowadzono analizę wielkości gospodarstw domowych w populacji Z w 2003 roku. Wylosowano w tym celu 50 gospodarstw i zebrano informacje o liczbie osób. Po uporządkowaniu ich rosnąco uzyskano podany niżej szereg statystyczny, zawierający indywidualne dane o każdej jednostce, którą jest gospodarstwo domowe. **Szereg taki nazywamy szeregiem szczegółowym.**

Gospodarstwa domowe według liczby osób w rodzinie ($x_j$):

| | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|
| 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 |
| 1 | 1 | 1 | 2 | 2 | 2 | 2 | 2 | 2 | 2 |
| 2 | 2 | 2 | 2 | 2 | 3 | 3 | 3 | 3 | 3 |
| 3 | 3 | 3 | 3 | 3 | 4 | 4 | 4 | 4 | 4 |
| 4 | 4 | 4 | 4 | 5 | 5 | 8 | 8 | 9 | 12 |

Źródło: dane umowne.

Przyjmujemy następujące oznaczenia:
$X$ – obserwowana zmienna, którą w danym przypadku jest liczba osób w rodzinie,
$x_j$ – wartość zmiennej przyporządkowana $j$-tej jednostce statystycznej.

Tutaj jest to liczba osób w $j$-tym gospodarstwie domowym. Na przykład, $j = 3$ oznacza trzecie w kolejności gospodarstwo o liczbie osób $x_3 = 1$; $n$ – liczebność populacji; $n = 50$, a zatem $j = 1, 2, \ldots, 50$.

W szeregu szczegółowym podane są wprawdzie indywidualne informacje, ale na ich podstawie trudno jest sformułować syntetyczną charakterystykę badanej zbiorowości. Obraz byłby jeszcze mniej wyraźny, gdyby liczba obserwacji była jeszcze większa, a tym bardziej gdybyśmy chcieli porównać wielkość gospodarstw domowych na przykład w mieście i na wsi. Przeprowadzimy więc grupowanie gospodarstw według liczby osób. Otrzymujemy wówczas **szereg rozdzielczy punktowy lub przedziałowy**, który składa się z dwóch kolumn. W jednej wyróżniono warianty zmiennej, a w drugiej przyporządkowane im liczebności. Punktowy szereg rozdzielczy gospodarstw domowych według liczby osób podano w tablicy 2.6.

**Tabela 2.6.** Gospodarstwa domowe według liczby osób w populacji
Z – szereg rozdzielczy punktowy

| $x_i$ – liczba osób w gospodarstwie domowym | $f_i$ – liczba gospodarstw domowych |
|---|---|
| 1 | 13 |
| 2 | 12 |
| 3 | 10 |
| 4 | 9 |
| 5 i więcej | 6 |
| $\sum_{i=1}^{k} f_i$ | 50 |

Źródło: dane umowne.

Jeśli uwzględnimy wszystkie wartości zmiennej, to uzyskamy szereg, w którym dla $x_6, x_7, x_{10}, x_{11}$ otrzymujemy liczebności ($f_i$) równe zeru. Przyjmujemy klasę zbiorczą ($x_5$) określoną jako „5 i więcej osób w rodzinie".

Ogólną liczebność populacji $n = 50$ uzyskujemy jako $\sum_{i=1}^{k} f_i$, gdzie $k$ oznacza liczbę wariantów zmiennej.

W przypadku szeregu, w którym uwzględniono wszystkie wartości zmiennej, jest ich $k = 12$, a w szeregu z klasą zbiorczą jest ich $k = 5$.

### Przykład 2.2
### Zmienna typu ciągłego

W przeprowadzonej analizie oprócz wielkości gospodarstw domowych w populacji Z w 2003 roku wzięto również pod uwagę wysokość dochodu przypadającego na jedną osobę (zmienna $X$).

Otrzymano następujące dane statystyczne:

| | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|
| 1,65 | 1,50 | 2,10 | 1,25 | 2,32 | 1,78 | 1,90 | 1,15 | 1,45 | 1,70 |
| 1,65 | 1,50 | 2,15 | 1,30 | 2,36 | 1,80 | 1,94 | 1,15 | 1,45 | 1,76 |
| 1,70 | 1,50 | 2,18 | 1,35 | 2,40 | 1,80 | 1,98 | 1,20 | 1,46 | 1,60 |
| 1,70 | 1,50 | 2,20 | 1,40 | 2,50 | 1,84 | 2,00 | 1,20 | 1,48 | 1,62 |
| 1,72 | 1,51 | 2,30 | 1,40 | 2,55 | 1,84 | 2,10 | 1,24 | 1,48 | 1,64 |

Po uporządkowaniu powstał następujący szereg szczegółowy gospodarstw domowych według wysokości dochodu przypadającego na jedną osobę ($x_j$):

| | | | | |
|---|---|---|---|---|
| 1,15 | 1,45 | 1,60 | 1,78 | 2,10 |
| 1,15 | 1,45 | 1,62 | 1,80 | 2,15 |
| 1,20 | 1,46 | 1,64 | 1,80 | 2,18 |
| 1,20 | 1,48 | 1,65 | 1,84 | 2,20 |
| 1,24 | 1,48 | 1,65 | 1,84 | 2,30 |
| 1,25 | 1,50 | 1,70 | 1,90 | 2,32 |
| 1,30 | 1,50 | 1,70 | 1,94 | 2,36 |
| 1,35 | 1,50 | 1,70 | 1,98 | 2,40 |
| 1,40 | 1,50 | 1,72 | 2,00 | 2,50 |
| 1,40 | 1,51 | 1,76 | 2,10 | 2,55 |

W celu lepszego ukazania prawidłowości występujących w rozkładzie liczebności gospodarstw ze względu na wysokość dochodów skonstruujemy szereg rozdzielczy. W przypadku zmiennej typu ciągłego, która może przyjąć każdą wartość z danego przedziału liczbowego, obszar jej zmienności musi być podzielony na odcinki, które nazywać będziemy przedziałami klasowymi lub klasami.

Musimy podjąć decyzję, ile będzie tych przedziałów lub, co jest równoznaczne, jaka będzie długość przedziału. Wygodnie jest, jeśli długości te są jednakowe. W literaturze podawane są procedury, według których można ustalić liczbę przedzia-

łów klasowych lub określić ich długość. A. Iwasiewicz i Z. Paszek[17] proponują następujące postępowanie:

1. liczbę przedziałów klasowych $k$ ustalamy według wzoru:

$$k' = 1 + 3{,}3 \cdot \lg n \tag{2.2}$$

$$0{,}5 \cdot \sqrt{n} \leq k \leq \sqrt{n} \tag{2.3}$$

Wartość $k$ uzyskujemy, zaokrąglając $k'$ do najbliższej liczby całkowitej.

2. ustalamy długość przedziału jako:

$$i = \frac{R}{k} = \frac{x_{\text{maks}} - x_{\text{min}}}{k} \tag{2.4}$$

gdzie:

$x_{\text{min}}$ – minimalna wartość zmiennej $X$

$x_{\text{maks}}$ – maksymalna wartość zmiennej $X$

W przykładzie 2.2, odnoszącym się do rozkładu liczebności gospodarstw domowych, otrzymujemy:

$$k' = 1 + 3{,}3 \cdot \lg 50 = 1 + 3{,}3 \cdot 1{,}699 = 6{,}61,$$

$$\sqrt{50} = 7{,}07,$$

$$0{,}5 \cdot \sqrt{50} = 0{,}5 \cdot 7{,}07 = 3{,}54.$$

Liczba przedziałów klasowych powinna mieścić się w przedziale liczbowym $3{,}54 \leq k \leq 7{,}07$. Biorąc pod uwagę wartość $k' = 6{,}61$, powinniśmy przyjąć $k = 7$ przedziałów klasowych o długości obliczonej zgodnie ze wzorem (2.4) jako:

$$i = \frac{2{,}3 - 1{,}1}{7} = \frac{1{,}4}{7} \approx 0{,}2.$$

Przedziały klasowe zapiszemy w sposób podany w tablicy 2.7. Dolną granicę pierwszego przedziału przyjmiemy jako $x^*_{\text{min}} = 1{,}0$ [tys. zł]. Granice przedziałów klasowych ustalamy tak, że górna granica klasy poprzedniej jest równa dolnej granicy klasy następnej. Aby uniknąć problemów przy zaliczaniu jednostek do klas, musimy przyjąć jednoznaczne kryterium. Do danego przedziału zaliczać będziemy jednostki, dla których zaobserwowane wartości zmiennej spełniają warunek:

$$x_i^d \leq x_i < x_i^g \tag{2.5}$$

gdzie:

$x_i^d$ – dolna granica $i$-tego przedziału klasowego,

$x_i^g$ – górna granica $i$-tego przedziału klasowego.

[17] A. Iwasiewicz, Z. Paszek, *op. cit.*

Przyjęte przedziały są lewostronnie domknięte i prawostronnie otwarte.

W kolumnie (1) tablicy 2.7 podano przedziały, których długość $i$ = 0,2 [tys. zł]. Zgodnie z przyjętą procedurą szereg powinien składać się z 7 przedziałów, ale w rozważanym przykładzie klasyfikacja nie będzie zupełna. Ostatni przedział [2,2–2,4) nie pozwala bowiem uwzględnić 3 gospodarstw o najwyższych dochodach. Dlatego ponownie ustalamy liczbę przedziałów klasowych, przyjmując $x^*_{\min}$ = 1,0 [tys. zł], a $x^*_{\text{maks}}$ = 2,6 [tys. zł].

Wówczas $k = \frac{2,6-1,0}{7} = \frac{1,6}{7} = 0,24$ [tys. zł]. W kolumnie (2) podano nowe przedziały klasowe. W kolumnie (4) wartości zmiennej przyporządkowane poszczególnym przedziałom. Kolumna (5) zawiera ich liczebności.

**Tablica 2.7.** Empiryczny rozkład liczebności gospodarstw domowych według wysokości dochodu przypadającego na jedną osobę w województwie Z w 2003 roku

| $x_i$ | Przedziały | | $x_i$ – zaobserwowane wartości zmiennej $X$ | $f_i$ |
|---|---|---|---|---|
| (1) | (2) | (3) | (4) | (5) |
| 1,0–1,2 | 1,00–1,24 | $1,00 \le x_i < 1,24$ | 1,15; 1,15; 1,20; 1,20 | 4 |
| 1,2–1,4 | 1,24–1,48 | $1,24 \le x_i < 1,48$ | 1,24; 1,25; 1,30; 1,35; 1,40; 1,40; 1,45; 1,45; 1,46 | 9 |
| 1,4–1,6 | 1,48–1,72 | $1,48 \le x_i < 1,72$ | 1,48; 1,48; 1,50; 1,50; 1,50; 1,50; 1,51; 1,60; 1,62; 1,64; 1,65; 1,65; 1,70; 1,70; 1,70 | 15 |
| 1,6–1,8 | 1,72–1,96 | $1,72 \le x_i < 1,96$ | 1,72; 1,76; 1,78; 1,80; 1,80; 1,84; 1,84; 1,90; 1,94; | 9 |
| 1,8–2,0 | 1,96–2,20 | $1,96 \le x_i < 2,20$ | 1,98; 2,00; 2,10; 2,10; 2,15; 2,18 | 6 |
| 2,0–2,2 | 2,20–2,44 | $2,20 \le x_i < 2,44$ | 2,20; 2,30; 2,32; 2,36; 2,43 | 5 |
| 2,2–2,4 | 2,44–2,68 | $2,44 \le x_i < 2,68$ | 2,50; 2,55 | 2 |
| | | | Suma: | 50 |

Źródło: dane umowne.

**W szeregu rozdzielczym przedziałowym** wartości zmiennej podane są jako przedziały. Opisując prawidłowości występujące w rozkładzie liczebności za pomocą różnych charakterystyk nazywanych dalej miarami, będziemy musieli posłużyć się jedną tylko liczbą, która reprezentuje daną klasę. Reprezentantem przedziału klasowego jest jego środek. Obliczamy go jako średnią arytmetyczną dolnej i górnej granicy.

$$x_i' = \frac{x_i^d + x_i^g}{2}. \tag{2.6}$$

Skonstruowany szereg rozdzielczy gospodarstw domowych według wysokości dochodu wraz ze środkami przedziałów podano w tablicy 2.8. Zamieszczono tam

również liczebności w postaci bezwzględnej ($f_i$), jako częstości względne ($f_i'$) zdefiniowane wzorem (2.1) oraz w wyrażaniu procentowym ($f_i' \cdot 100\%$).

**Tablica 2.8.** Rozkład liczebności (szereg rozdzielczy) gospodarstw domowych według wysokości dochodu

| $x_i$ | $f_i$ | $x_i'$ | $f_i'$ | $f_i' \cdot 100\%$ |
|---|---|---|---|---|
| 1,00–1,24 | 4 | 1,12 | 0,08 | 8,0 |
| 1,24–1,48 | 9 | 1,36 | 0,18 | 18,0 |
| 1,48–1,72 | 15 | 1,60 | 0,30 | 30,0 |
| 1,72–1,96 | 9 | 1,84 | 0,18 | 18,0 |
| 1,96–2,20 | 6 | 2,08 | 0,12 | 12,0 |
| 2,20–2,44 | 5 | 2,32 | 0,10 | 10,0 |
| 2,44–2,68 | 2 | 2,56 | 0,04 | 4,0 |
| Suma | 50 | × | 1,00 | 100 |

Źródło: dane umowne.

Możemy zastanawiać się, ile gospodarstw domowych posiada dochody niższe od górnej granicy przedziału klasowego. W celu uzyskania odpowiedzi musimy skumulować przedziały klasowe i zsumować liczebności odpowiednich przedziałów klasowych. Sumy te są przyporządkowane górnym granicom klas.

Skumulowane liczebności oznaczamy jako cum$f_i$. Symbole $f_i$ oraz cum$f_i$ oznaczają odpowiednio liczebności bezwzględne, a symbole $f_i'$ oraz cum$f_i'$ liczebności względne (częstości). Skumulowany szereg gospodarstw domowych według wysokości dochodu przedstawiono w tablicy 2.9.

**Tablica 2.9.** Skumulowane liczebności gospodarstw domowych według wysokości dochodu

| $x_i^g$ | cum$f_i$ | cum$f_i'$ | $f_i' \cdot 100\%$ |
|---|---|---|---|
| (–∞; 1,24] | 4 | 0,08 | 8 |
| (–∞; 1,48] | 13 | 0,26 | 26 |
| (–∞; 1,72] | 28 | 0,56 | 56 |
| (–∞; 1,96] | 37 | 0,74 | 74 |
| (–∞; 2,20] | 43 | 0,86 | 86 |
| (–∞; 2,44] | 48 | 0,96 | 96 |
| (–∞; 2,68] | 50 | 1,00 | 100 |

Źródło: dane umowne.

## 2.1.3. Grupowanie wielocechowe

Dotychczas rozpatrywaliśmy rozkłady liczebności (szeregi statystyczne) jednostek charakteryzowanych za pomocą jednej cechy (zmiennej). Grupowanie wielocechowe występuje wtedy, gdy każdej jednostce przyporządkowane są przynajmniej dwie cechy. Poniżej podajemy odpowiednie przykłady.

Obserwujemy wydatki na usługi w ich związku z liczbą pracujących kobiet w rodzinie. Oznaczamy jako $Y$ – wydatki na usługi, zaś jako $X$ – liczbę pracujących kobiet w rodzinie. Otrzymane wyniki dotyczące 6 gospodarstw domowych podano w tablicy 2.10.

**Tablica 2.10.** Wydatki na usługi względem liczby kobiet w rodzinie

| Nr | $y_i$ | $x_i$ |
|---|---|---|
| 1 | 5 | 0 |
| 2 | 20 | 1 |
| 3 | 15 | 2 |
| 4 | 37 | 3 |
| 5 | 27 | 4 |
| 6 | 45 | 5 |

Źródło: dane umowne.

W przypadku gdy liczba obserwacji jest duża, konstruujemy dwuwymiarowy szereg rozdzielczy. W tablicy 2.11 podano rezultat grupowania wielocechowego gospodarstw domowych sklasyfikowanych według typów rodzin i liczby dzieci w Polsce

**Tablica 2.11.** Rodziny z dziećmi w gospodarstwach domowych według typów rodzin i liczby dzieci w 2002 roku

| Liczba dzieci | Typ rodziny | | | | Ogółem |
|---|---|---|---|---|---|
| | Małżeństwa z dziećmi | Partnerzy z dziećmi | Matki z dziećmi | Ojcowie z dziećmi | |
| 1 | 2 063 078 | 53 683 | 661 589 | 73 972 | 2 852 322 |
| 2 | 1 886 387 | 27 978 | 264 504 | 24 203 | 2 203 072 |
| 3 | 634 283 | 10 795 | 66 512 | 5 836 | 717 426 |
| z 4 i więcej | 271 055 | 6 626 | 26 681 | 2 244 | 306 606 |
| Ogółem | 5 860 264 | 110 687 | 1 798 331 | 231 808 | 8 001 090 |

Źródło: Tablice wynikowe Narodowego Spisu Ludności i Mieszkań 2002, www.stat.gov.pl

w 2002 roku. Zestawiono tutaj cechę jakościową: typ rodziny (małżeństwa z dziećmi, partnerzy z dziećmi, matki z dziećmi, ojcowie z dziećmi) oraz zmienną typu skokowego – liczba dzieci w rodzinie.

W tablicy 2.12 podano łączny rozkład liczebności mężczyzn i kobiet według wieku w chwili zawarcia małżeństwa. Zestawiono tutaj dwa szeregi rozdzielcze ze zmiennymi typu ciągłego.

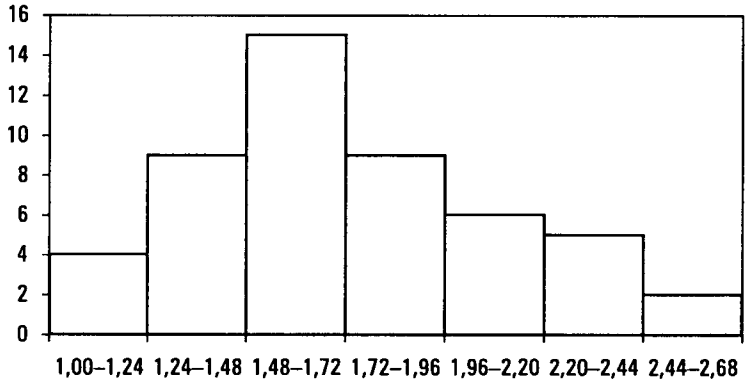
Tablica 2.12. Rozkład liczby mężczyzn i kobiet według wieku w chwili zawarcia małżeństwa w Polsce w 2000 roku

| Wiek mężczyzn | | Wiek kobiet ($Y$) | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|
| | $x_i'$ | 15–19 | 20–24 | 25–29 | 30–34 | 35–39 | 40–44 | 45–49 | 50-54 |
| | | $y_i'$ | | | | | | | |
| | | 17,5 | 22,5 | 27,5 | 32,5 | 37,5 | 42,5 | 47,5 | 52,5 |
| 15–19 | 17,5 | 3498 | 1848 | 79 | 11 | 1 | 0 | 0 | 0 |
| 20–24 | 22,5 | 16714 | 60412 | 8335 | 444 | 70 | 7 | 2 | 7 |
| 25–29 | 27,5 | 4577 | 41370 | 25630 | 2265 | 354 | 83 | 27 | 12 |
| 30–34 | 32,5 | 653 | 5813 | 8108 | 3529 | 804 | 265 | 75 | 15 |
| 35–39 | 37,5 | 142 | 1311 | 2452 | 2043 | 1273 | 572 | 211 | 57 |
| 40–49 | 45 | 73 | 470 | 1114 | 1412 | 1682 | 2028 | 1449 | 646 |
| 50–59 | 55 | 11 | 51 | 133 | 200 | 398 | 851 | 1684 | 5879 |

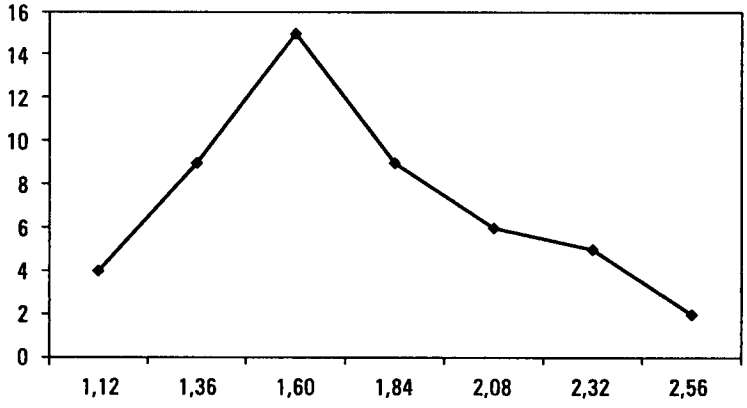
Źródło: „Rocznik Demograficzny", GUS, Warszawa 2001.

## 2.2. Prezentacja graficzna szeregów statystycznych

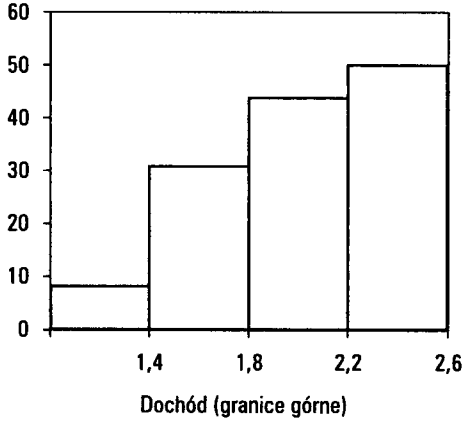
Otrzymane w rezultacie grupowania rozkłady liczebności podawaliśmy dotychczas w postaci tabelarycznej. Charakterystyczne dla nich prawidłowości będą łatwiejsze do spostrzeżenia, jeśli rozkłady te zaprezentujemy graficznie. Dla przedstawienia prawidłowości istotne znaczenie ma dobór odpowiedniego typu wykresu. Szeregi rozdzielcze z cechami ilościowymi przedstawiamy najczęściej w postaci histogramu i wieloboku liczebności. Rysunki 2.1 i 2.2 prezentują odpowiednio histogram i wielobok liczebności szeregu rozdzielczego gospodarstw domowych ze względu na wysokość dochodów. Na rysunkach 2.3 i 2.4 zaprezentowano szeregi skumulowanych liczebności. Jeśli wyobrazimy sobie, że możemy dowolnie zwiększać dokładność pomiaru a wówczas długość przedziału $i \to 0$, to wielobok liczebności przekształci się w wykres w postaci krzywej liczebności. Ten typ wykresu przedstawiono na rysunkach 2.5 i 2.6.



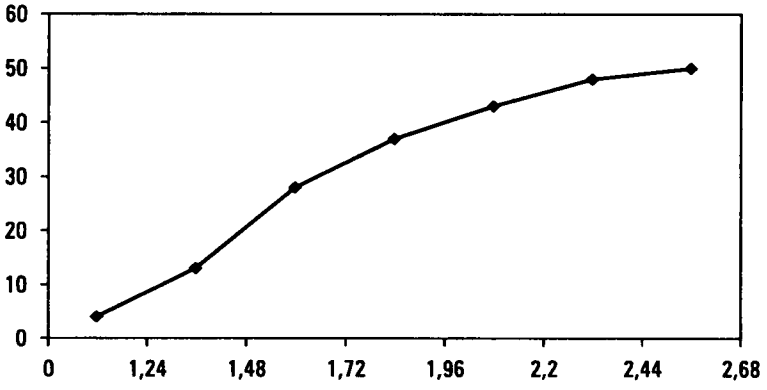

Rys. 2.1. Histogram liczebności gospodarstw domowych według wysokości dochodów



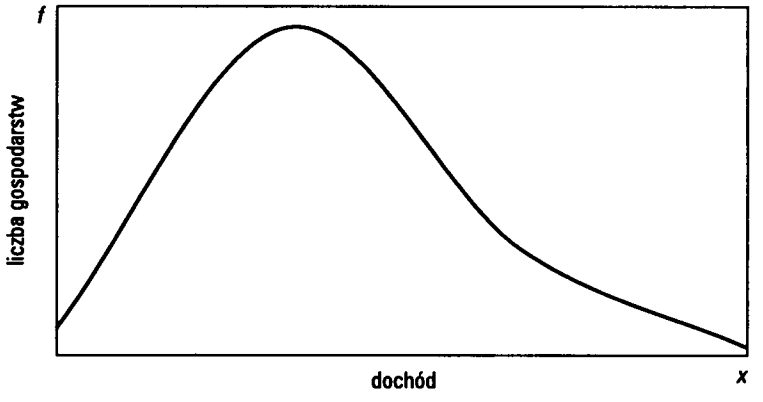

Rys. 2.2. Wielobok liczebności gospodarstw domowych według wysokości dochodów



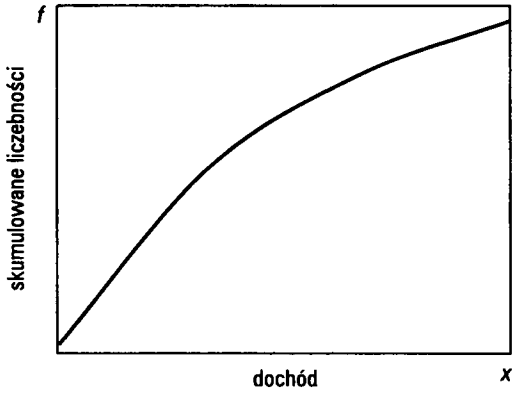

Rys. 2.3. Histogram skumulowanego szeregu gospodarstw domowych według wysokości dochodów




Rys. 2.4. Wielobok skumulowanych liczebności gospodarstw domowych według wysokości dochodów

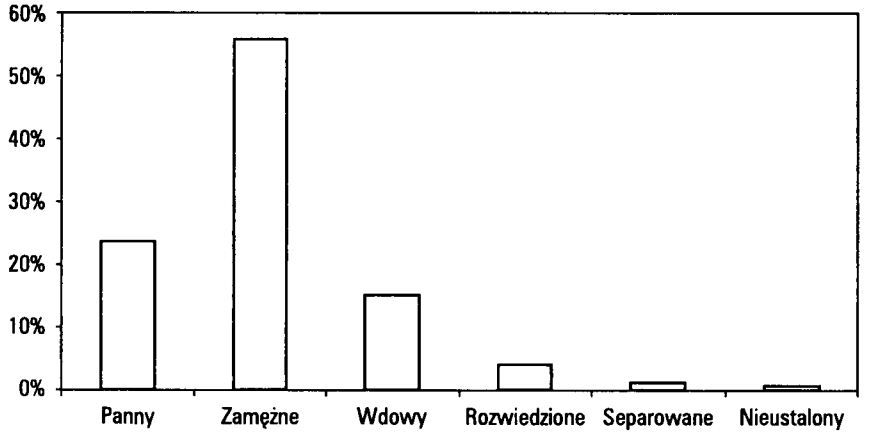



Rys. 2.5. Krzywa liczebności gospodarstw domowych według wysokości dochodów

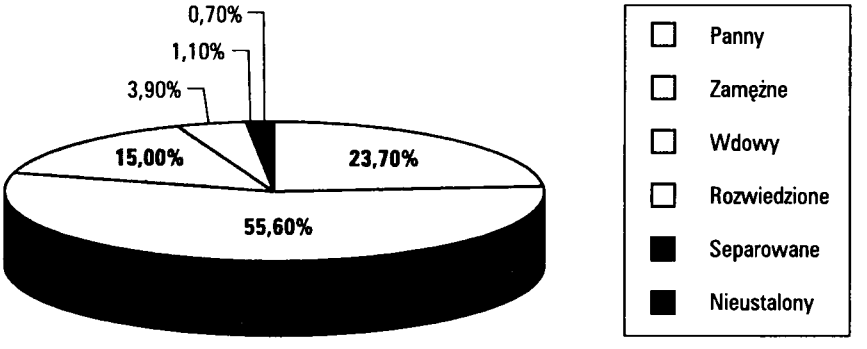



Rys. 2.6. Krzywa liczebności skumulowanej gospodarstw domowych według wysokości dochodów

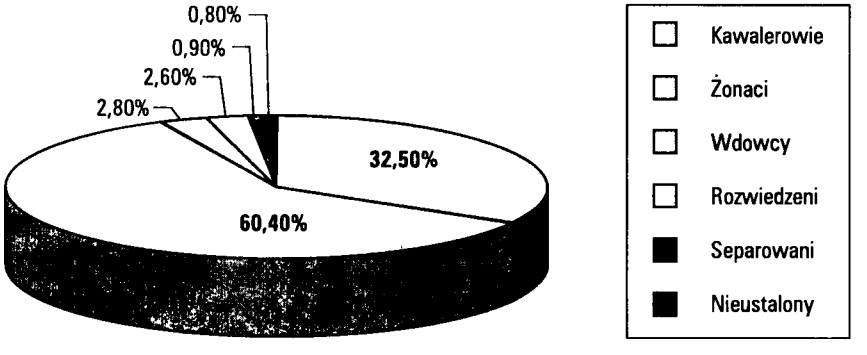
Szeregi statystyczne zmiennych jakościowych możemy przedstawiać za pomocą różnych typów wykresów. Na rysunkach 2.7–2.9 zaprezentowano jako przykłady zastosowanie histogramu (rys. 2.7) oraz wykresów kołowych (rys. 2.8–2.9).




Rys. 2.7. Struktura populacji kobiet według stanu cywilnego w Polsce w 2002 roku

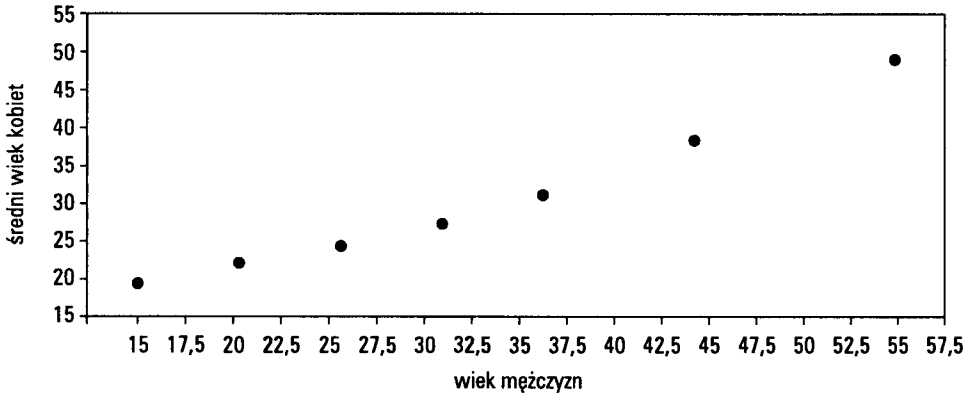



Rys. 2.8. Struktura populacji kobiet według stanu cywilnego Polsce w 2002 roku



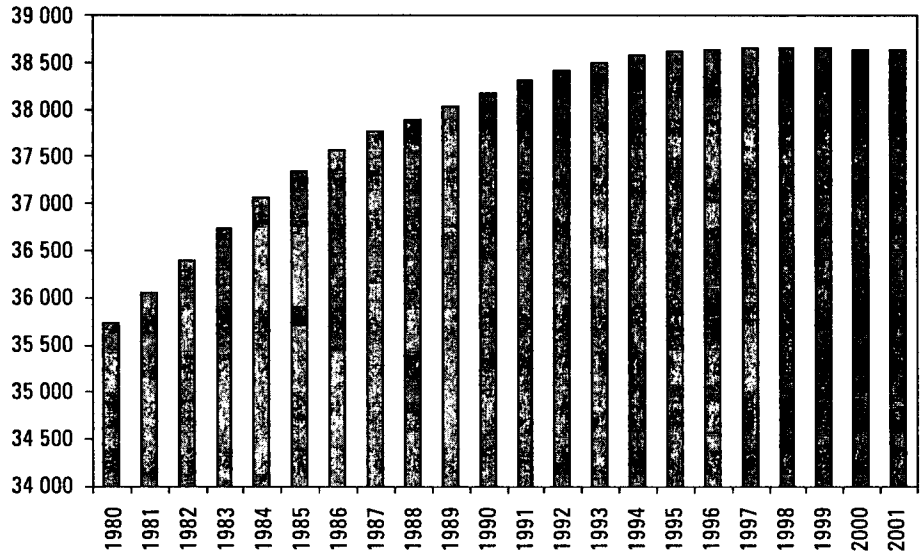

Rys. 2.9. Struktura populacji mężczyzn według stanu cywilnego w Polsce w 2002 roku

Szeregi uzyskane w rezultacie grupowania wielocechowego przedstawiamy w postaci diagramu korelacyjnego. Rysunek 2.10 przedstawia związek, jaki występuje między wiekiem mężczyzn i kobiet w chwili zawarcia związku małżeńskiego.



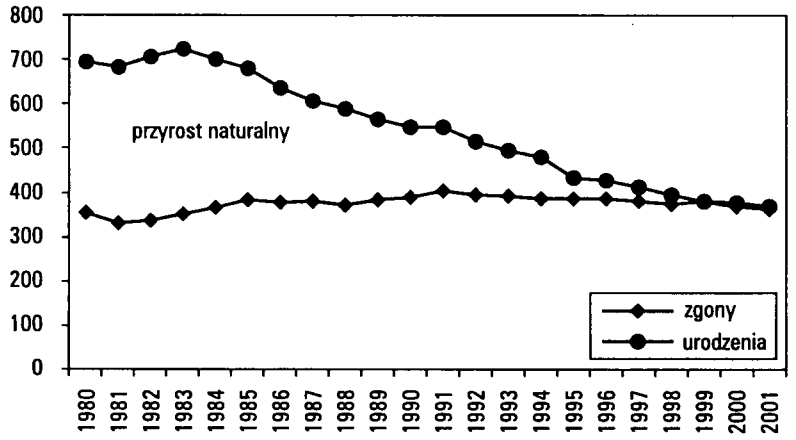

Rys. 2.10. Średni wiek kobiet w chwili małżeństwa względem wieku mężczyzn w Polsce w 2000 roku

Prezentując graficznie szeregi czasowe, należy rozróżnić szeregi momentów i okresów. Szeregi czasowe momentów przedstawiające stan w określonym momencie, na przykład 31.12 danego roku kalendarzowego, prezentujemy za pomocą wykresów słupkowych. Na rysunku 2.11 przedstawiono stany ludności w Polsce w dniu 31.12 w latach 1980–2001.




Rys. 2.11. Liczba ludności w Polsce w latach 1980–2001 (stan w dniu 31.12)

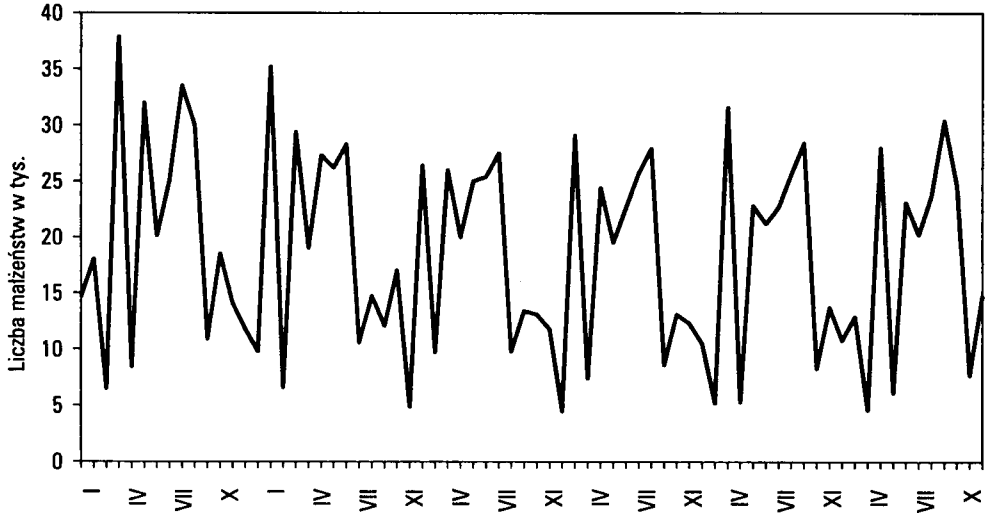
Na rysunku 2.12 przedstawiono dynamikę liczby urodzeń, zgonów i przyrostu naturalnego w Polsce w latach 1980–2001. Jest to szereg czasowy okresów. Szeregi tego typu ilustrujemy za pomocą wykresów w postaci linii.




Rys. 2.12. Liczba urodzeń, zgonów i przyrost naturalny w Polsce w latach 1980–2001

W rozpatrywanych szeregach czasowych jednostkę stanowił rok. Niektóre zjawiska, szczególnie te, które podlegają wahaniom okresowym (na przykład sezonowym), rozważamy w okresach krótszych (na przykład: w dniach, tygodniach, kwartałach, miesiącach, półroczach).

Rysunek 2.13 przedstawia liczbę zawieranych małżeństw w poszczególnych miesiącach okresu 1990–1995.




Rys. 2.13. Małżeństwa według miesięcy w latach 1990–1995

Sporządzając wykresy szeregów statystycznych, posługujemy się odpowiednimi programami komputerowymi, takimi jak arkusze kalkulacyjne (np. Microsoft Excel) oraz pakiety statystyczne (np. Statistica, SPSS).

Dotychczas przedstawiliśmy dwie możliwości ujmowania szeregów statystycznych, a mianowicie w postaci tablic oraz w formie graficznej. W obydwu przypadkach rozpatrujemy cały rozkład liczebności. Innym podejściem jest syntetyczna charakterystyka za pomocą parametrów opisowych.

# Charakterystyki opisowe rozkładu liczebności

Rozdział 3

Wyobraźmy sobie, że samoloty odbywają rejsy na trasie Kraków–Londyn. Od początku istnienia linii odbyło się 12 lotów. Przewoźnik gromadzi informacje o wykonanych rejsach. Między innymi posiada dane o liczbie pasażerów w każdym locie. Informacje te podano w tablicy 3.1.

**Tablica 3.1.** Liczba pasażerów na trasie Kraków–Londyn

| Numer lotu | Liczba pasażerów |
|---|---|
| 1 | 124 |
| 2 | 130 |
| 3 | 128 |
| 4 | 115 |
| 5 | 128 |
| 6 | 121 |
| 7 | 140 |
| 8 | 141 |
| 9 | 130 |
| 10 | 128 |
| 11 | 112 |
| 12 | 135 |

Źródło: dane umowne.

Na podstawie takich danych nie jesteśmy w stanie stwierdzić żadnych prawidłowości. Dysponujemy jedynie szeregiem liczb. W rzeczywistości może się zdarzyć, że

lotów (obserwacji) mogą być setki, tysiące w zależności od czasu istnienia linii, liczby wykonywanych rejsów i okresu objętego obserwacją. Kolejny problem może powstać, gdy przewoźnik zechce porównać swoją działalność z wynikami właścicieli innych linii. Jak zatem poradzimy sobie z taką ilością danych? Co należy w takiej sytuacji uczynić?

Możemy rozkład liczebności opisać za pomocą miar określanych mianem charakterystyk opisowych. Odpowiednio do prawidłowości, które chcemy analizować rozróżniamy miary:

1) położenia (tendencji centralnej),
2) zmienności (rozproszenia),
3) asymetrii,
4) koncentracji.

## 3.1. Miary położenia

Miary położenia dostarczają informacji o przeciętnym poziomie rozpatrywanej zmiennej. Dzielą się one na:

- średnie klasyczne: średnia arytmetyczna, geometryczna, harmoniczna,
- przeciętne pozycyjne: na przykład modalna, kwantyle itp.

### 3.1.1. Średnia arytmetyczna

Spośród charakterystyk opisowych należących do średnich klasycznych najwięcej uwagi poświęcimy **średniej arytmetycznej**, która jest miarą najbardziej znaną i najczęściej używaną, nawet w życiu codziennym.

Na przykład gdy kupujemy samochód, interesujemy się, jakie jest jego zużycie paliwa. Stwierdzenie, że wynosi ono 7 litrów na 100 kilometrów rozumiemy jako zużycie średnie. Śledząc rozgrywki ligowe piłki nożnej, otrzymujemy informację o liczbie bramek przypadających na jeden mecz w danej rundzie rozgrywek. Jest to również wartość średnia. Po każdej sesji egzaminacyjnej studenci obliczają średnią ocen, a jeśli jest ona odpowiednio wysoka, to mają prawo do otrzymania stypendium.

We wszystkich podanych przykładach wartość średniej arytmetycznej uzyskujemy w ten sam sposób. **Średnia ta jest bowiem sumą wszystkich wartości zmiennej podzieloną przez liczbę obserwacji.** W podanych wyżej przykładach dzielimy odpowiednio:

- całkowitą ilość paliwa zużytego na danej trasie przez liczbę przebytych kilometrów,
- sumę wszystkich bramek strzelonych w poszczególnych meczach przez liczbę meczów w danej kolejce ligowej,
- sumę ocen uzyskanych w czasie sesji przez liczbę zdawanych egzaminów.

Postępowanie prowadzące do uzyskania wartości średniej arytmetycznej możemy przedstawić w postaci następującego wzoru definicyjnego:

$$\bar{x} = \frac{\sum_{j=1}^{n} x_j}{n} = \frac{x_1 + x_2 + \ldots + x_n}{n} \tag{3.1}$$

gdzie:
$x_j$ – wartość zmiennej $X$ zaobserwowana dla $j$-tej jednostki statystycznej,
$n$ – liczba jednostek danej zbiorowości.

W tym miejscu należy zwrócić uwagę na niektóre ważne własności średniej wyróżniające ją spośród innych miar przeciętnych, a mianowicie:

1. Średnia arytmetyczna stałej jest równa tej stałej. Jeśli bowiem $x_j = c$ = constans ($j = 1, 2, \ldots, n$), to:

$$\bar{x} = \frac{c + c + \ldots + c}{n} = \frac{n \cdot c}{n} = c.$$

Oznacza to, że gdyby wszystkie jednostki posiadały jednakową wartość zmiennej, to każda z nich miałaby wartość równą średniej arytmetycznej.

2. Jeśli rozpatrujemy zmienną $Z$, która powstała przez pomnożenie zmiennej $X$ przez stałą $c$, to wówczas średnią arytmetyczną zmiennej $Z$ można uzyskać, mnożąc średnią zmiennej $X$ przez stałą $c$:

$$\bar{z} = \frac{\sum_{j=1}^{n} c \cdot x_j}{n} = \frac{c \cdot \sum_{j=1}^{n} x_j}{n} = c \cdot \frac{\sum_{j=1}^{n} x_j}{n} = c \cdot \bar{x}.$$

W podobny sposób można wykazać, że jeśli zmienna $Z$ powstała przez podzielenie zmiennej $X$ przez stałą $c$, to średnią arytmetyczną zmiennej $Z$ można uzyskać, dzieląc średnią zmiennej $X$ przez stalą $c$.

3. Jeśli rozpatrujemy zmienną $Y$, która jest sumą dwóch innych zmiennych oznaczonych jako $X_1$ oraz $X_2$, to wówczas:

$$\bar{y} = \frac{\sum_{j=1}^{n} y_j}{n} = \frac{\sum_{j=1}^{n} (x_{1j} + x_{2j})}{n} = \frac{\sum_{j=1}^{n} x_{1j}}{n} + \frac{\sum_{j=1}^{n} x_{2j}}{n} = \bar{x}_1 + \bar{x}_2.$$

W podobny sposób można wykazać, że jeśli zmienna $Y$ jest różnicą zmiennych $X_1$ oraz $X_2$, to wówczas średnia arytmetyczna zmiennej $Y$ jest różnicą średnich zmiennych $X_1$ oraz $X_2$.

4. Suma odchyleń zaobserwowanych wartości zmiennej od ich średniej arytmetycznej jest równa zeru:

$$\sum_{j=1}^{n} (x_j - \bar{x}) = \sum_{j=1}^{n} x_j - \sum_{j=1}^{n} \bar{x} = n \cdot \bar{x} - n \cdot \bar{x} = 0.$$

5. Suma kwadratów odchyleń zaobserwowanych wartości zmiennej od ich średniej arytmetycznej jest najmniejsza:

$$\sum_{j=1}^{n}(x_j - \bar{x})^2 = \min.$$

Dla wykazania tej własności należy udowodnić, że podana funkcja osiąga minimum w punkcie $\bar{x}$. W tym celu znajdujemy pierwszą pochodną ze względu na $x$ i przyrównujemy ją do zera.

$$\left[\sum_{j=1}^{n}(x_j - \bar{x})^2\right]' = \left[\sum_{j=1}^{n}(x_j^2 - 2 \cdot x_j \cdot \bar{x} + \bar{x}^2)\right]' = 2 \cdot \sum_{j=1}^{n} x_j - 2 \cdot n \cdot \bar{x} = 0.$$

$$\sum_{j=1}^{n} x_j - n \cdot \bar{x} = 0.$$

$$\sum_{j=1}^{n} x_j = n \cdot \bar{x} \Rightarrow \bar{x} = \frac{\sum_{j=1}^{n} x_j}{n}.$$

6. Średnia arytmetyczna nie jest przypisana żadnej jednostce statystycznej. Charakteryzuje całą populację. Dlatego w przypadku zmiennej typu skokowego średnia arytmetyczna może przyjmować wartość, która w rzeczywistości nie może wystąpić. Na przykład średnia liczba osób w rodzinie może być równa 3,7.
7. Średnia podaje w sposób sumaryczny informacje o wszystkich jednostkach zbiorowości:

$$n\bar{x} = \sum_{j=1}^{n} x_j.$$

Uwzględnione są w niej wszystkie informacje zawarte w zbiorze danych. Własność ta ma niestety niekorzystne konsekwencje, gdy pojawiają się obserwacje nietypowe, to znaczy takie, które w znacznym stopniu odbiegają od wartości, jakie przyjmują pozostałe jednostki zbiorowości. Średnia arytmetyczna pozostaje pod silnym wpływem wartości krańcowych. Wówczas traktowanie jej jako wielkości ujmującej sumarycznie informację o całej zbiorowości może okazać się nieadekwatne do rzeczywistości.

Na przykład obserwujemy zbiorowość małżeństw ze względu na liczbę posiadanych dzieci. Dla sześciu par otrzymano następujące liczby dzieci: 2, 1, 3, 2, 3, 1. Dla tej zbiorowości średnia arytmetyczna jest równa 2 dzieci przypadających na jedno małżeństwo. Jeżeli jednak do badanej zbiorowości dołączyłoby dodatkowe małżeństwo z 9 dzieci, to średnia przyjęłaby wówczas wartość równą 3 dzieci. Jak z tego wynika, pojawienie się jednej nietypowej wartości zmiennej w znacznym stopniu wpływa na wnioskowanie o całej zbiorowości.

8. Każdy rozkład liczebności ma tylko jedną średnią arytmetyczną.

W obliczeniach wzór (3.1) może przyjąć różną postać w zależności od tego, jakimi danymi posługujemy się w analizie. Rozróżnimy trzy następujące sytuacje:

1) posiadamy indywidualne dane o każdej jednostce statystycznej, a więc rozważamy szereg szczegółowy,
2) badaniu poddaliśmy dużą zbiorowość i pogrupowaliśmy dane, otrzymując szereg rozdzielczy,
3) mamy dostęp tylko do danych pogrupowanych w postaci szeregu rozdzielczego.

Sytuacja (1) występuje w przykładzie dotyczącym liczby pasażerów w poszczególnych lotach samolotów na trasie Kraków–Londyn. Znamy bowiem liczbę pasażerów w każdym rejsie, a więc dysponujemy szeregiem szczegółowym. Średnią liczbę pasażerów podróżujących tą linią obliczamy, podstawiając do wzoru (3.1) odpowiednie wartości z tablicy 3.1 i uzyskujemy:

$$\bar{x} = \frac{124+130+128+115+128+121+140+141+130+128+112+135}{12},$$

a zatem:

$$\bar{x} = \frac{1532}{12} = 127{,}66 \approx 128\,[\text{pasażerów}].$$

Rozważana tutaj zmienna typu skokowego przyjmuje wartości wyłącznie ze zbioru liczb całkowitych, a średnia arytmetyczna jako wynik dzielenia przyjmuje wartości ze zbioru liczb rzeczywistych i dlatego nie należy się dziwić, że na jeden rejs na trasie Kraków–Londyn przypadało 127,66 pasażerów (w przybliżeniu 128 osób). Interpretacja ta będzie łatwiejsza do zaakceptowania, jeśli pamiętać będziemy o tym, że średnia charakteryzuje populację rejsów, a nie wykonany w rzeczywistości rejs.

Sytuacje wymienione w punkcie (2) i (3) omówimy równocześnie, ponieważ nie jest tutaj istotne, czy szereg rozdzielczy zbudowaliśmy sami, czy pochodzi on na przykład z rocznika statystycznego. Ważne jest, że posługujemy się wówczas tzw. średnią arytmetyczną ważoną, gdzie wagami są liczebności w poszczególnych przedziałach klasowych. W tym przypadku musimy jednak wziąć pod uwagę, jaki jest typ zmiennej, a mianowicie, czy jest ona skokowa, czy ciągła (por. punkt 2.1.2).

W przypadku szeregu rozdzielczego ze zmienną skokową (punktowego szeregu rozdzielczego) średnią arytmetyczną obliczymy jako:

$$\bar{x} = \frac{\sum_{i=1}^{k} f_i \cdot x_i}{\sum_{i=1}^{k} f_i}, \tag{3.2}$$

gdzie:
$x_i$ – $i$-ta wartość zmiennej $X$ w szeregu rozdzielczym zmiennej skokowej,
$f_i$ – liczba jednostek, które przyjęły $i$-tą wartość zmiennej,
$k$ – liczba wartości zmiennej występujących w szeregu rozdzielczym.

Jeśli rozpatrujemy zmienną typu ciągłego, to posługujemy się następującym wzorem:

$$\bar{x} = \frac{\sum_{i=1}^{k} f_i \cdot x_i'}{\sum_{i=1}^{k} f_i}, \tag{3.3}$$

gdzie:
$x_i'$ – środek przedziału klasowego (środkowa wartość klasy),
$f_i$ – liczebność $i$-tego przedziału klasowego,
$k$ – liczba przedziałów klasowych.

### Przykład 3.1

Przeprowadzono badania jakości wyrobów produkowanych przez firmę Z. Pobrano w tym celu 60 próbek liczących po 5 wyrobów, które sklasyfikowano ze względu na przyjęte normy jakości. Zmienną $X$ jest liczba wyrobów uznanych za wybrakowane w poszczególnych próbkach. Jest to zmienna typu skokowego. W tablicy 3.2 podano rozkład liczby braków w poszczególnych próbach.

**Tablica 3.2.** Rozkład liczebności wyrobów uznanych za wybrakowane w poszczególnych próbkach

| $i$ | $x_i$ | $f_i$ | $x_i \cdot f_i$ |
|---|---|---|---|
| 1 | 0 | 7 | 0 |
| 2 | 1 | 10 | 10 |
| 3 | 2 | 19 | 38 |
| 4 | 3 | 11 | 33 |
| 5 | 4 | 9 | 36 |
| 6 | 5 | 4 | 20 |
| | Razem | 60 | 137 |

Źródło: dane umowne.

Posługując się wzorem (3.2), otrzymujemy:

$$\bar{x} = \frac{137}{60} = 2{,}28\,[\text{szt.}].$$

Średnia liczba wyrobów wybrakowanych w poszczególnych próbkach jest równa 2,28 [szt.].

**Przykład 3.2**

Właściciel sklepu „Zdrowa Żywność" jest zainteresowany tym, ile przeciętnie wydają klienci robiący u niego zakupy. Na podstawie wydruków z kasy fiskalnej otrzymał dane dotyczące 30 osób. Przedstawił je w postaci szeregu rozdzielczego, który został podany w tablicy 3.3.

**Tablica 3.3.** Wydatki klientów sklepu „Zdrowa Żywność"

| $i$ | Wydatkowana kwota $[x_i^d, x_i^g)$ | Liczba klientów $f_i$ | $x_i'$ | $f_i \cdot x_i'$ |
|---|---|---|---|---|
| 1 | 10–15 | 4 | 12,5 | 50 |
| 2 | 15–20 | 8 | 17,5 | 140 |
| 3 | 20–25 | 12 | 22,5 | 270 |
| 4 | 25–30 | 6 | 27,5 | 165 |
| | Razem | 30 | × | 625 |

Źródło: dane umowne.

W tym przypadku mamy do czynienia ze zmienną typu ciągłego. Średnie wydatki 30 klientów obliczamy zgodne ze wzorem (3.3) i uzyskujemy:

$$\bar{x} = \frac{625}{30} = 20{,}83\,[\text{zł}].$$

W sklepie „Zdrowa żywność" klienci wydawali średnio 20 złotych i 83 grosze.

## 3.1.2. Średnia geometryczna

Przedstawimy teraz inną średnią klasyczną, która jest znacznie rzadziej używana niż średnia arytmetyczna. Miarą tą jest średnia geometryczna. Najczęściej znajduje zastosowanie, gdy:

1) badamy zjawiska ujmowane dynamicznie, czyli w postaci szeregów czasowych. Możemy na przykład interesować się średnim tempem wzrostu Produktu Krajowego Brutto w wybranym kraju w rozważanym okresie (np. w Polsce w latach 1990–2003),

2) zmienna wyrażona jest w postaci częstości względnych (procenty, promile),
3) w zbiorze wartości zmiennych występują wartości nietypowe.

Średnia geometryczna jest zdefiniowana następującym wzorem:

$$G = \sqrt[n]{x_1 \cdot x_2 \cdot \ldots \cdot x_n} = \sqrt[n]{\prod_{j=1}^{n} x_j}. \tag{3.4}$$

Wzorem (3.4) posługujemy się, gdy posiadamy dane w postaci szeregu szczegółowego. W przypadku szeregów rozdzielczych wartość średniej geometrycznej obliczamy jako średnią ważoną:

$$G = \sqrt[n]{x_2^{f_2} \cdot x_2^{f_2} \cdot \ldots \cdot x_k^{f_k}} = \sqrt[n]{\prod_{i=1}^{k} x_i^{f_i}}, \tag{3.5}$$

gdzie: $n = \sum_{i=1}^{k} f_i$.

Obliczanie pierwiastka wysokiego stopnia jest bardzo uciążliwe i dlatego zamiast zaobserwowanych wartości zmiennej wprowadzamy ich logarytmy i wówczas otrzymujemy:

- dla szeregu szczegółowego:

$$\log G = \frac{\sum_{j=1}^{n} \log x_j}{n}, \tag{3.6}$$

- dla szeregu rozdzielczego:

$$\log G = \frac{\sum_{i=1}^{k} f_i \cdot \log x_i}{\sum_{i=1}^{k} f_i}. \tag{3.7}$$

Jeśli rozpatrujemy szereg rozdzielczy ze zmienną ciągłą, to wówczas zamiast wartości $x_i$ pojawią się środki przedziałów klasowych $x_i'$.

Logarytm średniej geometrycznej jest średnią arytmetyczną logarytmów wartości zmiennej. Posiada więc wymienione wcześniej własności średniej arytmetycznej wyliczone w punktach (1)–(7). Średnia geometryczna jest mniej niż średnia arytmetyczna czuła na wartości krańcowe, nietypowe. Obliczenie średniej geometrycznej jest niemożliwe, jeśli w szeregu pojawią się obserwacje o wartościach zmiennej równych zeru[18].

[18] Przykłady zastosowania średniej geometrycznej będą przedstawione w rozdziale 5, poświęconym metodom analizy dynamiki zjawisk.

### 3.1.3. Średnia harmoniczna

W praktyce często spotykamy zmienne, które, charakteryzując natężenie zjawiska, są ilorazem dwóch innych zmiennych. Do takich należą na przykład: cena, będąca stosunkiem wartości towaru do jego ilości, wydajność pracy uzyskana przykładowo jako iloraz czasu poświęconego na wytworzenie danej wielkości produkcji i ilości wyrobów, gęstość zaludnienia, czyli stosunek liczby mieszkańców danego terytorium do jego powierzchni. Posługiwanie się średnią harmoniczną zilustrujemy podanym poniżej przykładem 3.3.

**Przykład 3.3**

W województwie małopolskim wyodrębniono trzy podregiony (NATS 3), które mają liczbę ludności i gęstość zaludnienia podaną w tablicy 3.4. Należy obliczyć średnią gęstość zaludnienia województwa małopolskiego. Rozważaną zmienną jest gęstość zaludnienia. Oznaczymy ją jako $X$. Liczba ludności w podregionach stanowi wagi ($f_i$).

**Tablica 3.4.** Powierzchnia i gęstość zaludnienia podregionów województwa małopolskiego w 2002 roku

| Podregion | Ludność $f_i$ | Gęstość zaludnienia $x_i$ |
|---|---|---|
| Krakowsko-tarnowski | 13 951 731 | 190 |
| Nowosądecki | 1 105 018 | 148 |
| m. Kraków | 740 737 | 2266 |
| Suma | 15 797 486 | × |

Źródło: „Rocznik Demograficzny 2002", GUS, Warszawa 2003.

Średnią gęstość zaludnienia zdefiniujemy jako:

$$\overline{g} = \frac{\sum_{i=1}^{k} f_i}{\sum_{i=1}^{k} p_i},$$

gdzie $p_i$ oznacza powierzchnię podregionów.

W rozważanym przykładzie dana jest liczba ludności i gęstość zaludnienia w podregionach. Brak jest natomiast informacji o powierzchniach podregionów ($p_i$). Muszą być one obliczone jako:

$$p_i = \frac{f_i}{x_i}.$$

Średnią gęstość zaludnienia obliczamy zatem jako średnią harmoniczną:

$$\bar{g} = \frac{13951731 + 1105018 + 740737}{\frac{13951731}{190} + \frac{1105018}{148} + \frac{740737}{2266}},$$

$$\bar{g} = \frac{13951731 + 1105018 + 740737}{73430 + 7466 + 327} = \frac{15797486}{81223} = 194 \text{ [osoby na km}^2\text{]}.$$

Posługując się symbolami ogólnymi, otrzymujemy następujący wzór na średnią harmoniczną:

$$H = \frac{\sum_{i=1}^{k} f_i}{\sum_{i=1}^{k} \frac{f_i}{x_i}}. \tag{3.8}$$

Objaśnienia symboli występujących we wzorze podano już wcześniej.

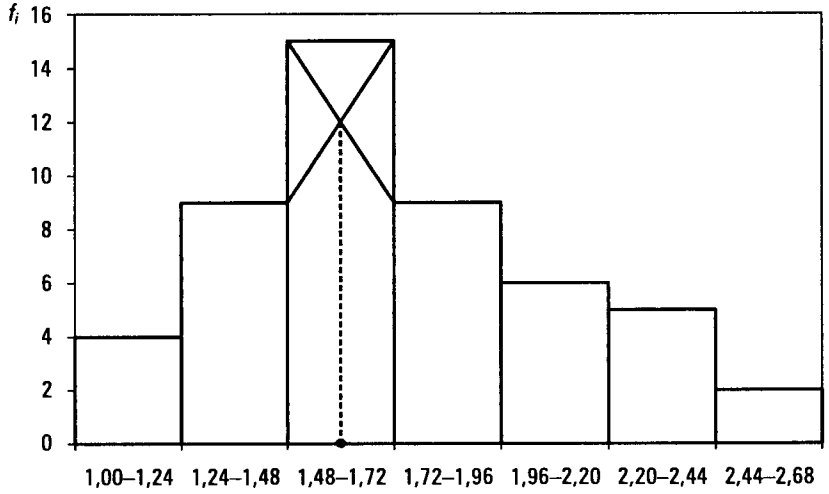
## 3.1.4. Przeciętne pozycyjne

Przeciętne pozycyjne są miarami wyznaczanymi z szeregu statystycznego. W odróżnieniu od średniej arytmetycznej i geometrycznej i harmonicznej są to wartości konkretnych jednostek statystycznych, a mianowicie tych, które w szeregu zajmują szczególną pozycję. Jedną z takich miar jest modalna.

**Modalna** (moda, dominanta, wartość najczęstsza, typowa) jest to wartość zmiennej, która w danej zbiorowości występuje najczęściej. Wyznaczenie jej z szeregu szczegółowego nie nastręcza większych trudności. Wystarczy tylko sprawdzić, która wartość zmiennej powtarza się największą ilość razy. Tę właśnie wartość wskazujemy jako modalną.

W przykładowych danych zawartych w tablicy 3.1 zauważamy, że były 3 rejsy, w czasie których na pokładzie samolotu podróżowało 128 pasażerów. Jest to zatem liczba, która pojawia się najczęściej w analizowanej zbiorowości. Jest to wartość modalna. Najczęściej na trasie Kraków–Londyn podróżowało 128 pasażerów.

Wyznaczanie modalnej z szeregu rozdzielczego z przedziałami klasowymi przedstawimy najpierw w wersji graficznej (rys. 3.1). Wykorzystamy w tym celu histogram przedstawiający rozkład liczebności gospodarstw domowych według wysokości dochodów rozważany w przykładzie 2.2. Jest on podany w tablicy 2.7 i przedstawiony na rysunku 2.1.




Rys. 3.1. Graficzne wyznaczanie modalnej

Aby wyznaczyć modalną, rozpatrujemy trzy przedziały klasowe: przedział posiadający największą liczebność, gdyż tutaj znajduje się modalna oraz przedziały z nim sąsiadujące (poprzedni i następny). Łączymy odcinkiem wierzchołek odpowiadający górnej granicy klasy poprzedzającej przedział z modalną z wierzchołkiem odpowiadającym górnej granicy klasy z modalną. Drugi odcinek prowadzimy od dolnej granicy przedziału z modalną do dolnej granicy przedziału następnego. Odcinki te przecinają się w punkcie, którego rzut na oś odciętych wskazuje przybliżoną wartość modalnej. Tę przybliżoną wartość możemy określić liczbowo za pomocą następującego wzoru interpolacyjnego:

$$M_o = l_m + i_m \cdot \frac{f_m - f_{m-1}}{(f_m - f_{m-1}) + (f_m - f_{m+1})}, \tag{3.9}$$

gdzie:
$l_m$ – dolna granica klasy, w której znajduje się modalna,
$i_m$ – długość przedziału, w którym znajduje się modalna,
$fl_m$ – liczebność klasy, w której znajduje się modalna,
$f_{m-1}$ – liczebność klasy poprzedzającej klasę, w której znajduje się modalna,
$f_{m+1}$ – liczebność klasy następującej po klasie z modalną.

### Przykład 3.4

W Wydziale Administracyjnym Urzędu Miasta przeprowadzono badania czasu obsługi petentów. W sposób losowy wybrano próbę liczącą 53 osoby, którym zmierzono czas, w jakim zostali obsłużeni. Uzyskane dane zawarto w tablicy 3.5. Interesuje nas najczęstszy czas obsługi petentów.

Tablica 3.5. Czas obsługi petentów

| Ilość czasu w minutach $[x_i^d, x_i^g)$ | Liczba petentów $f_i$ |
|---|---|
| 5–10 | 5 |
| 10–15 | 13 |
| 15–20 | 24 |
| 20–25 | 11 |

Źródło: dane umowne.

Zauważamy, że w przedziale 15–20 minut obsłużono najwięcej petentów (24), a zatem w tym przedziale znajdzie się modalna.

$$l_m = 15 \text{ minut}, i_m = 5 \text{ minut}, f_{m-1} = 13 \text{ osób}, f_{m+1} = 11 \text{ osób}.$$

Podstawiając do wzoru (3.4), otrzymujemy:

$$M_o = 15 + 5 \cdot \frac{24-13}{(24-13)+(24-11)} = 15 + 5 \cdot \frac{11}{11+13} = 15 + 2{,}29 = 17{,}29 \text{ [minut]}.$$

W Wydziale Administracyjnym Urzędu Miasta najczęstszy czas obsługi petentów wynosi 17,29 minut, to jest 17 minut i 17 sekund. Minuta liczy 60 sekund, a zatem 0,29 · 60 sekund = 17,4 [sekund].

### Przykład 3.5

Powrócimy do przykładu z rozdziału 2, w którym rozpatrywaliśmy rozkład liczebności gospodarstw domowych według wysokości dochodu podany w postaci szeregu szczegółowego, na podstawie którego skonstruowaliśmy szereg rozdzielczy zapisany w tablicy 2.8. Biorąc pod uwagę szereg szczegółowy, stwierdzamy, że dominujący dochód wynosi 1,5 tysiąca złotych. Teraz dla tej samej zbiorowości gospodarstw wyznaczymy modalną z danych pogrupowanych i posłużymy się wzorem interpolacyjnym. Dla przypomnienia szereg rozdzielczy podano w tablicy 3.6.

Otrzymujemy następującą wartość modalnej:

$$M_o = 1{,}48 + 0{,}24 \cdot \frac{15-9}{(15-9)+(15-9)} = 1{,}48 + 0{,}24 \cdot \frac{6}{6+6} =$$

$$= 1{,}48 + \frac{6 \cdot 0{,}24}{12} = 1{,}48 + 0{,}12 = 1{,}6 \text{ [tys.zł]}.$$

Wyznaczony za pomocą wzoru interpolacyjnego najczęstszy (typowy, dominujący) dochód gospodarstw domowych jest równy 1600 złotych.

**Tablica 3.6.** Rozkład liczebności gospodarstw domowych według wysokości dochodu

| $x_i$ | $f_i$ | $cum\, f_i$ |
|---|---|---|
| 1,00–1,24 | 4 | 4 |
| 1,24–1,48 | 9 | 13 |
| 1,48–1,72 | 15 | 28 |
| 1,72–1,96 | 9 | 37 |
| 1,96–2,20 | 6 | 43 |
| 2,20–2,44 | 5 | 45 |
| 2,44–2,68 | 2 | 50 |
| Suma | 50 | × |

Źródło: tablica 2.8 i 2.9.

W tym miejscu należy zwrócić uwagę na to, że zawsze musimy się liczyć z występowaniem różnic pomiędzy rezultatami uzyskanymi na podstawie szeregu szczegółowego i rozdzielczego. Konstruując szeregi rozdzielcze, zawsze tracimy pewne informacje. Przyjmujemy na przykład, że wszystkie wartości zmiennej jednostek statystycznych zaliczonych do danego przedziału klasowego są równe środkowi tego przedziału. Dlatego grupowanie danych jest ważnym etapem badań statystycznych. Musimy zatem dołożyć wszelkich starań, aby tak go przeprowadzić, by otrzymane wyniki jak najlepiej odzwierciedlały rozważaną rzeczywistość.

Charakterystykę rozkładu liczebności możemy wzbogacić, wyznaczając wartości miar opisowych określanych mianem wartości ćwiartkowych, nazywanych również kwartylami.

**Wartości ćwiartkowe (kwartyle)** dzielą badaną zbiorowość na ćwiartki. W centrum znajduje się mediana, czyli wartość środkowa. Jest ona przeciętną pozycyjną, która dzieli uporządkowany szereg statystyczny na dwie części w taki sposób, że połowa jednostek zbiorowości ma wartość zmiennej nie większą (mniejszą lub równą) od mediany, a połowa nie mniejszą od niej (większą lub równą). Z szeregu szczegółowego medianę wyznaczamy w następujący sposób:

- gdy liczebność zbiorowości $n$ jest nieparzysta, to:

$$M_e = x_{\frac{n+1}{2}}, \tag{3.10}$$

- gdy liczebność zbiorowości $n$ jest parzysta

$$M_e = \frac{x_{\frac{n}{2}} + x_{\frac{n}{2}+1}}{2}. \tag{3.11}$$

### Przykład 3.6

Powrócimy do przykładu dotyczącego linii lotniczej. Poniższe liczby przedstawiają liczbę pasażerów w poszczególnych rejsach. Chcemy wyznaczyć medianę dla tej zbiorowości.

124, 130, 128, 115, 128, 121, 140, 141, 130, 128, 112, 135

W pierwszej kolejności uporządkujemy jednostki statystyczne (rejsy) według wartości zmiennej (liczba pasażerów) i otrzymujemy następujący szereg:

112, 115, 121, 124, 128, 128, 128, 130, 130, 135, 140, 141

Zbiorowość liczy 12 obserwacji ($n = 12$), a więc $n$ jest parzyste. Szukamy jednostek zbiorowości, których wartości zmiennej podstawimy do wzoru (3.11):

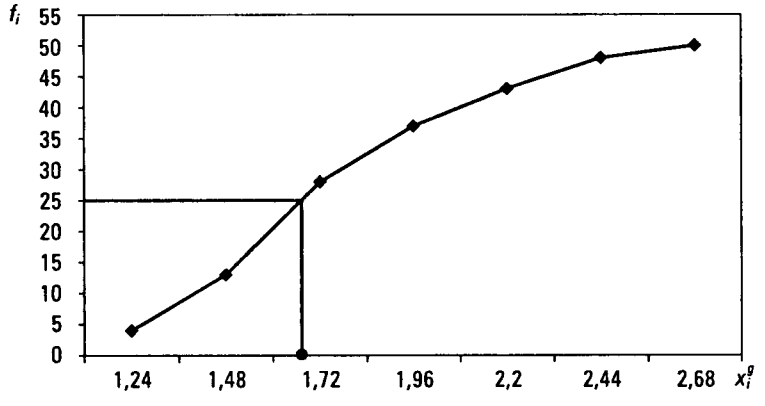
$$x_{\frac{n}{2}} = x_{\frac{12}{2}} = x_6,$$

$$x_{\frac{n}{2}+1} = x_{\frac{12}{2}+1} = x_7.$$

Do wzoru podstawimy wartości szóstej i siódmej jednostki w uporządkowanym szeregu. Odliczając od początku zbiorowości, odczytujemy $x_6 = 128$ oraz $x_7 = 128$ [osób]. Wartości te podstawiamy do wzoru (3.11) i otrzymujemy:

$$M_e = \frac{128+128}{2} = 128 \text{[pasażerów]}.$$

W pięćdziesięciu procentach rejsów na pokładzie samolotów podróżowało nie więcej niż 128 pasażerów i w pięćdziesięciu procentach rejsów leciało nie mniej niż 128 pasażerów. Sposób wyznaczania mediany w przypadku szeregu rozdzielczego zaprezentujemy najpierw graficznie. Wykorzystamy w tym celu diagram szeregu skumulowanego gospodarstw domowych przedstawiony na rysunku 2.4. Na rysunku 3.2 przedstawiono odpowiedni wykres skumulowanej liczebności gospodarstw domowych według wysokości dochodu rozważany w przykładzie 2.2.




Rys. 3.2. Graficzne wyznaczanie mediany dochodów gospodarstw domowych

Na osi rzędnych odnajdujemy punkt odpowiadający $\frac{n}{2} = 25$ jednostkom. Prostą prostopadłą do tej osi prowadzimy aż do przecięcia z krzywą wykresu. Rzut punktu przecięcia na oś odciętych wyznacza wartość środkową (Me).

Wartość mediany uzyskamy, korzystając z następującego wzoru interpolacyjnego:

$$Me = l_{me} + \frac{i_{me}}{f_{me}} \cdot \left( \frac{n}{2} - \sum_{i=1}^{me-1} f_i \right), \tag{3.12}$$

gdzie:

$l_{me}$ – dolna granica klasy, w której znajduje się mediana,
$i_{me}$ – długość przedziału, w którym znajduje się mediana,
$f_{me}$ – liczebność klasy, w której znajduje się mediana,
$\sum_{i=1}^{me-1} f_i$ – suma liczebności od klasy pierwszej do klasy poprzedzającej przedział, w którym znajduje się mediana.

### Przykład 3.7

Obliczenie mediany dla szeregu rozdzielczego zilustrujemy na przykładzie, w którym interesujemy się czasem obsługi petentów urzędu miejskiego. Odpowiednie dane zawiera tablica 3.7 (patrz: przykład 3.4, s. 55).

**Tablica 3.7.** Czas obsługi petentów oraz obliczenia pomocnicze

| Czas w minutach | Liczba petentów $f_i$ | Liczebność kumulacyjna $cum\, f_i$ |
|---|---|---|
| 5–10 | 5 | 5 |
| 10–15 | 13 | 18 |
| 15–20 | 24 | 42 |
| 20–25 | 11 | 53 |
| Suma | 53 | × |

Źródło: dane umowne.

Najpierw określimy, w którym przedziale znajduje się mediana. Dzielimy liczebność zbiorowości $n$ przez 2 i otrzymujemy:

$$\frac{n}{2} = \frac{52}{2} = 26{,}5.$$

Następnie sprawdzamy, w którym przedziale zawiera się 26 i 27 jednostka zbiorowości. Pomiędzy nimi znajduje się mediana. W tym celu musimy utworzyć liczebność kumulacyjną ($cum\, f_i$), dodając liczebności poszczególnych klas, poczynając od

klasy pierwszej (patrz tablica 3.7). Szukana wartość znajduje się w przedziale 15–20. Jest to przedział z medianą. Podstawiając do wzoru (3.12) odpowiednie wartości, otrzymujemy:

$$M_e = 15 + \frac{5}{24} \cdot \left( \frac{53}{2} - 18 \right) = 15 + 0{,}21 \cdot 8{,}5 = 15 + 1{,}78 = 16{,}78 \text{ [minut]}.$$

Zatem połowa petentów była obsługiwana nie dłużej niż 16 minut i 47 sekund, a połowa nie krócej niż 16 minut i 47 sekund.

### Przykład 3.8

Na podstawie danych z przykładu 3.6 możemy wyznaczyć medianę dochodów 50 gospodarstw domowych zarówno z szeregu szczegółowego, jak i z rozdzielczego. W tym przypadku również liczba obserwacji jest również parzysta $n = 50$. Medianą jest zatem dochód gospodarstwa wyznaczony jako:

$$x_{\frac{n}{2}} = x_{\frac{50}{2}} = x_{25} \qquad x_{\frac{n}{2}+1} = x_{\frac{50}{2}+1} = x_{26},$$

$$M_e = \frac{x_{25} + x_{26}}{2} = \frac{1{,}65 + 1{,}70}{2} = \frac{3{,}35}{2} = 1{,}675 \text{ [zł]}.$$

Zatem 50% gospodarstw domowych posiada dochody nie większe od 1645 zł, a dochody 50% gospodarstw są nie mniejsze niż 1675 zł. Posługując się wzorem interpolacyjnym i danymi w postaci szeregu rozdzielczego, otrzymujemy następujący dochód środkowy (zob. tablica 3.6):

$$M_e = 1{,}48 + \frac{0{,}24}{15} \cdot (25 - 13) = 1{,}48 + \frac{0{,}24 \cdot 12}{15} = 1{,}48 + 0{,}192 = 1{,}672 \text{ [zł]}.$$

Różnice w otrzymanych wynikach są rezultatem grupowania danych.

### Przykład 3.9

W magazynie „BusinessWeek"[19] zamieszczono artykuł na temat płac informatyków w niektórych krajach europejskich. Dane zilustrowano wykresem, który zamieszczamy na rysunku 3.3.

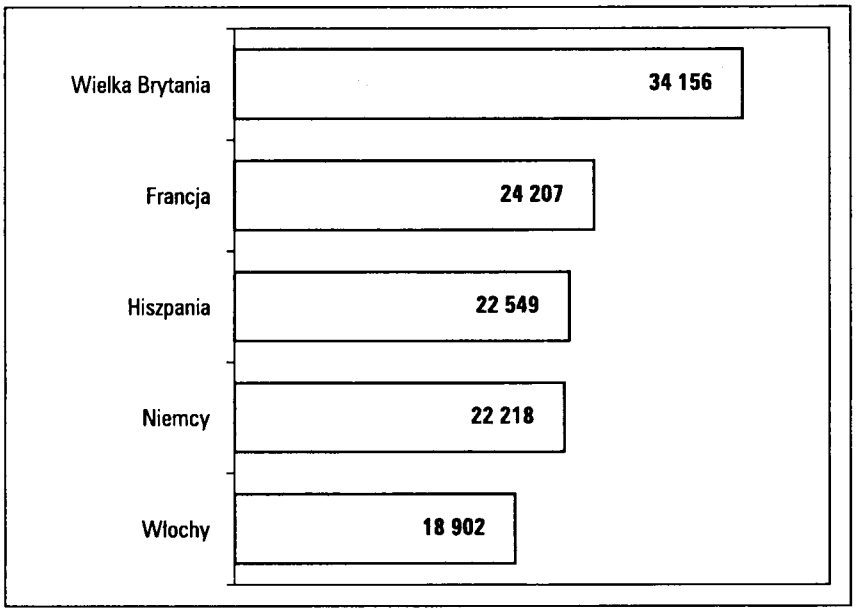
Mediana płac dyrektora we Francji wynosząca 24 207 złotych oznacza, że połowa dyrektorów ds. informatyki w tym kraju zarabia nie więcej niż 24 207 złotych, a połowa nie mniej niż 24 207 złotych[20].

Mediana jest jednym z trzech kwartyli. Zaprezentowano ją w pierwszej kolejności, ponieważ należy do miar tendencji centralnej i wraz ze średnią arytmetyczną

[19] „BusinessWeek" nr 10 (2003).

[20] Interpretację pozostałych wyników oraz porównania międzynarodowe pozostawiamy czytelnikowi.

i modalną jest najczęściej wykorzystywana do szacowania wartości przeciętnej w populacji generalnej.




Rys. 3.3. Mediana miesięcznych wynagrodzeń netto* dyrektora ds. informatyki w wybranych krajach europejskich w 2002 roku

*wg średniego kursu NBP z 31 października 2002 r.

Źródło: „BusinessWeek" nr 10 (2003), s. 81, AWR „Wprost".

**Kwartyle** są to wartości zmiennej tych jednostek statystycznych, które dzielą uporządkowaną zbiorowość na ćwiartki, odpowiednio:

- kwartyl pierwszy (dolny) dzieli zbiorowość na dwie części w taki sposób, że 25% jednostek zbiorowości ma wartości zmiennej nie większe od wartości kwartyla, a 75% ma wartości nie mniejsze niż wartość kwartyla,
- kwartyl drugi (mediana) został już przedstawiony,
- kwartyl trzeci (górny) dzieli zbiorowość na dwie części w taki sposób, że 75% jednostek zbiorowości ma wartości zmiennej nie większe od wartości kwartyla, a 25% ma wartości nie mniejsze niż wartość kwartyla.

Z szeregu szczegółowego kwartyle wyznaczamy w podobny sposób jak medianę. Dla liczebności parzystej w pierwszej kolejności wyznaczamy medianę zgodnie ze wzorem (3.11). Następnie dla każdej z wyodrębnionych części wyznaczamy wartości środkowe, które będą równe odpowiednio dla pierwszej części – kwartylowi pierwszemu, a dla drugiej części – kwartylowi trzeciemu.

**Przykład 3.10**

Wyznaczymy kwartyle rozkładu liczebności pasażerów linii lotniczej Kraków–Londyn. Mediana (kwartyl drugi) jest równa 128 pasażerów i znajduje się między 6 a 7 jednostką uporządkowanej zbiorowości. Dzielimy więc zbiorowość rejsów na trasie Kraków–Londyn na dwie równe części:

część I: 112, 115, 121, 124, 128, 128,

cześć II: 128, 130, 130, 135, 140, 141.

Wyznaczamy medianę dla każdej z nich:
Wartość środkowa dla części I, będąca kwartylem pierwszym, jest równa:

$$Q_1 = \frac{121+124}{2} = 122{,}5 \text{ [pasażera]}.$$

Wartość środkowa dla części II, będąca kwartylem trzecim, jest równa:

$$Q_3 = \frac{130+135}{2} = 133{,}5 \text{ [pasażera]}.$$

Zinterpretujemy teraz otrzymane wyniki.
W 25% rejsów na trasie Kraków–Londyn podróżowało nie więcej niż 122,5 pasażerów, a w 75% rejsów latało nie mniej niż 122,5 pasażerów. W 75% rejsów na pokładzie samolotów znajdowało się nie więcej niż 133,5 pasażerów, a w 25% rejsów latało nie mniej niż 133,5 pasażerów.
W przypadku liczebności nieparzystej najpierw wyznaczamy medianę (wzór (3.10)), a następnie wyodrębniamy dwie podzbiorowości w taki sposób, że jednostkę zbiorowości, dla której wartość jest medianą, zaliczamy do obydwu części. Kolejnym krokiem jest wyznaczenie mediany dla każdej z tak wyodrębnionych części.

**Przykład 3.11**

Tym razem rozpatrujemy wyniki uzyskane przez studentów na testowym egzaminie z ekonomii. Studenci otrzymali podaną niżej liczbę punktów.

34, 35, 42, 44, 45, 46, 51, 55, 66, 77, 80

Poszukujemy wartości wszystkich przedstawionych kwartyli.

1. Wyznaczamy medianę dla $n = 11$.

$$Me = x_{\frac{n+1}{2}} = x_6 = 46 \text{ [punktów]}.$$

2. Dzielimy zbiorowość na dwie części i dla każdej z nich wyznaczamy mediany:

część I: 34, 35, 42, 44, 45, **46**

część II: **46**, 51, 55, 66, 77, 80

$$Q_1 = \frac{42+44}{2} = 43 \text{ [punkty]}.$$

$$Q_3 = \frac{55+66}{2} = 60{,}5 \text{ [punkty]}.$$

Kwartyl pierwszy wskazuje, że 25% studentów otrzymało nie więcej niż 43 punkty, a 75% studentów nie mniej niż 43 punkty. Na podstawie kwartyla trzeciego stwierdzany, że 75% studentów otrzymało nie więcej niż 60,5 punktu, a 25% studentów nie mniej niż 60,5 punktu.

Dla szeregu rozdzielczego idea wyznaczania kwartyli jest podobna do wyznaczania mediany. W pierwszej kolejności sprawdzamy, w którym przedziale znajduje się dany kwartyl, określając przedział, na który przypada odpowiednio:

- dla kwartyla pierwszego: $\frac{n}{4}$ jednostek,
- dla kwartyla trzeciego: $\frac{3 \cdot n}{4}$ jednostek.

Wzory interpolacyjne mają następującą postać:

$$Q_1 = l_{q_1} + \frac{i_{q_1}}{f_{q_1}} \cdot \left( \frac{n}{4} - \sum_{i=1}^{q_1 - 1} f_i \right) \tag{3.13}$$

$$Q_2 = Me$$

$$Q_3 = l_{q_3} + \frac{i_{q_3}}{f_{q_3}} \cdot \left( \frac{3 \cdot n}{4} - \sum_{i=1}^{q_3 - 1} f_i \right). \tag{3.14}$$

**Przykład 3.12**

W celu zilustrowania sposobu wyznaczania kwartyli $Q_1$ oraz $Q_3$ posłużmy się danymi z przykładu 3.4. dotyczącego czasu obsługi petentów w urzędzie miejskim. Obliczenia pomocnicze zamieszczono w tablicy 3.8.

**Tablica 3.8.** Czas obsługi petentów oraz obliczenia pomocnicze

| Przedział kwartyla | Czas obsługi w minutach | Liczba petentów $f_i$ | Liczebność kumulacyjna *cum* $f_i$ |
|---|---|---|---|
| | 5–10 | 5 | 5 |
| $Q_1$ | 10–15 | 13 | 18 |
| $Q_3$ | 15–20 | 24 | 42 |
| | 20–25 | 11 | 53 |
| | Suma | 53 | × |

Źródło: dane umowne.

1. Najpierw obliczamy wartość $\frac{n}{4}$ (dla $Q_1$) i $\frac{3 \cdot n}{4}$ (dla $Q_3$):

$$Q_1 : \frac{53}{4} = 13{,}25,$$

$$Q_3 : \frac{3 \cdot 53}{4} = 39{,}75.$$

2. Następnie, korzystając z liczebności skumulowanej, sprawdzamy, w których przedziałach znajdują się otrzymane wartości.
3. Korzystając ze wzorów (3.13) i (3.14), wyznaczamy odpowiednie kwartyle.

$$Q_1 = 10 + \frac{5}{13} \cdot \left( \frac{53}{4} - 5 \right) = 10 + 0{,}38 \cdot 8{,}25 = 10 + 3{,}14 = 13{,}14 \text{[minut]},$$

$$Q_3 = 15 + \frac{5}{24} \cdot \left( \frac{3 \cdot 53}{4} - 18 \right) = 15 + 0{,}21 \cdot 21{,}75 = 15 + 4{,}57 = 19{,}57 \text{[minut]}.$$

Interpretacja:

- 25% petentów było obsługiwanych nie dłużej niż 13,14 minuty, a 75% petentów było obsługiwanych nie krócej niż 13,14 minuty.
- 75% petentów było obsługiwanych nie dłużej niż 19,57 minuty, a 25% petentów było obsługiwanych nie krócej niż 19,57 minuty.

W celu wyznaczenia kwartyli ograniczamy się do środkowych 50% obserwacji. Wartości tych miar nie zależą więc od wartości krańcowych. W związku z tym nie są one, w przeciwieństwie do średniej arytmetycznej, wrażliwe na nietypowe dla danej zbiorowości wyniki.

Zilustrujemy to przykładem dotyczącym notowań giełdowych.

**Przykład 3.13**

Uszeregowane rosnąco ceny akcji (w złotych) z jednego tygodnia (pięć notowań) wynosiły:

2,50; 2,52; 2,54; 2,60; 2,65

Mediana dla tej zbiorowości wynosi 2,54 złotych, a średnia arytmetyczna jest równa 2,56 złote. Nawet jeśli w miejsce wyniku 2,65 złotych pojawiłaby się cena akcji równa 7 złotych, to mediana nadal wynosiłaby 2,54 złotych. Wartość średniej wzrosłaby natomiast do 3,43 złotych.

Kwartyle stanowią szczególny przypadek większej grupy miar określanej jako **kwantyle**. Są one wartościami jednostek zbiorowości, które dzielą ją na określone części pod względem liczby jednostek tej zbiorowości. Na przykład, kwantyl rzędu 0,6 dzieli zbiorowość w taki sposób, że 60% jednostek zbiorowości jest nie większa niż wartość tego kwantyla, a 40% jest nie mniejsza od jego wartości.

Jeśli dzielimy zbiorowość na 10 części, to otrzymujemy miary nazywane **decylami**. Dla podziału populacji na 100 części posługujemy się centylami. Z kolei **percentyle** dzielą zbiorowość na 1000 części.

## 3.2. Miary zmienności (rozproszenia)

Zadaniem miar zmienności jest dostarczenie informacji o tym, w jakim stopniu wartości zmiennej poszczególnych jednostek statystycznych są zróżnicowane w porównaniu z wartością przeciętną. Wprowadzenie tych miar rozpoczniemy od przykładu.

**Przykład 3.14**

Dyrekcja banku rozważa przydzielenie swoich klientów (przedsiębiorstw) zlokalizowanych na terenie Małopolski do dwóch nowopowstałych centrów obsługi. Jako kryterium ma posłużyć wielkość obrotów przedsiębiorstw. Dokonano podziału, który przedstawiono w tablicy 3.9.

**Tablica 3.9.** Klienci centrów obsługi banku na terenie Małopolski

| Centrum I | | Centrum II | |
|---|---|---|---|
| **Klient** | **Obroty [mln zł]** | **Klient** | **Obroty [mln zł]** |
| Box-pol | 7 | Handlogum | 5 |
| Biuro Opis | 2 | Vitalex | 6 |
| FHU Wax | 3 | Olpis | 4 |
| Farmako | 11 | Saturn | 7 |
| Jowisz | 5 | PWL | 6 |
| Rab | 9 | Nordex | 7 |
| Polmięs | 6 | FirmaWD-KAN | 5 |
| Market II | 1 | Marbo | 6 |
| Dawex | 8 | Złompias | 8 |
| Ornament | 6 | Vived | 5 |
| Q-WER | 10 | Laboz | 7 |
| Pakor | 4 | Viol-Met | 6 |

Źródło: dane umowne.

Interesuje nas, czy pomiędzy wyodrębnionymi centrami scharakteryzowanymi ze względu na wysokość obrotów przedsiębiorstw istnieje zróżnicowanie oraz czy zbiorowości klientów obu centrów można traktować jako podobne. Scharakteryzujemy położenie i zmienność rozkładu liczebności klientów centrów obsługi banku ze względu na wysokość obrotów. Do opisu zbiorowości wykorzystano poznane już miary, takie jak: średnia arytmetyczna, modalna i mediana. Po wykonaniu odpowiednich obliczeń otrzymano rezultaty[21] zamieszczone w tablicy 3.10.

Tablica 3.10. Wartości przeciętne obrotów w wyodrębnionych centrach

| Miara | Centrum I | Centrum II |
|---|---|---|
| średnia arytmetyczna | 6 000 0000 | 6 000 0000 |
| modalna | 6 000 0000 | 6 000 0000 |
| mediana | 6 000 0000 | 6 000 0000 |

Źródło: obliczenia własne.

Dla obydwu centrów otrzymaliśmy identyczne wyniki. W dodatku liczba jednostek zbiorowości jest taka sama i wynosi 12. Czy możemy twierdzić, że obydwie zbiorowości klientów pod względem wysokości obrotów są jednakowe?

Już na pierwszy rzut oka widzimy, że dane opisujące centra różnią się. Obroty klientów Centrum I są bardziej zróżnicowane niż klientów Centrum II. W celu oszacowania poziomu rozproszenia stosujemy miary zwane miarami zmienności (rozproszenia). Rozróżniamy wśród nich miary[22]:

- bezwzględne: rozstęp, odchylenie przeciętne, odchylenie standardowe,
- względne: współczynnik zmienności.

Bezwzględne miary zmienności są wielkościami mianowanymi, wyrażonymi w takich jednostkach, jak rozpatrywana zmienna. Względne miary zmienności są najczęściej podawane w procentach. Dzięki temu umożliwiają porównanie zróżnicowania zarówno w odniesieniu do tej samej zmiennej w różnych populacjach (wiek mężczyzn i wiek kobiet w chwili zawarcia pierwszego małżeństwa), jak i ze względu na różne zmienne (na przykład: dochody, aktywność zawodowa, bezrobocie).

Miary te przedstawimy kolejno, ilustrując odpowiednimi przykładami sposób ich obliczania i interpretację.

[21] Jako ćwiczenie można samodzielnie wykonać odpowiednie obliczenia.
[22] A. Zeliaś, *Metody statystyczne*, Warszawa 2000.

## 3.2.1. Rozstęp

**Rozstęp** jest to różnica między największą i najmniejszą wartością zmiennej w analizowanej zbiorowości[23]. Wartość tej miary obliczamy za pomocą wzoru:

$$R = X_{maks} - X_{min}. \tag{3.15}$$

Rozstęp jest najprostszą i najłatwiejszą do obliczenia miarą zmienności. Przedstawia zakres zmienności wartości zmiennej w analizowanej zbiorowości.

Rozpatrujemy obroty centrów obsługi banków, o których informacje podano w tablicy 3.9. Ustalamy zakres (obszar) zmienności tych obrotów.

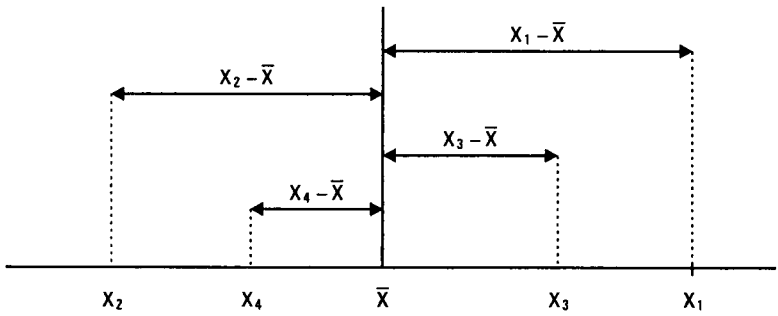
Na podstawie wzoru (3.15) otrzymujemy:

- Centrum I: R = 11 – 1 = 10 [mln złotych],
- Centrum II: R = 8 – 4 = 4 [mln złotych].

Z obliczeń wynika, że zbiorowość klientów Centrum I charakteryzuje się większą zmiennością. Obroty klientów tej placówki znajdują się w przedziale od 1 do 11 mln złotych, podczas gdy w Centrum II mieszczą się one w przedziale 4 do 8 mln złotych.

Rozstęp, opierając się na dwóch krańcowych wartościach, jest wrażliwy na obserwacje nietypowe, które będą przyjmować albo wartości maksymalne, albo minimalne, czyli te, na podstawie których wartość tej miary jest obliczana. Nie odzwierciedla on zróżnicowania wartości zmiennej cechy jednostek zbiorowości. Ta sama wartość rozstępu może charakteryzować różne rozkłady liczebności pomiędzy minimalną i maksymalną wartością. Rozstęp możemy obliczyć na podstawie szeregu szczegółowego i szeregu rozdzielczego punktowego.

Zaprezentujemy zatem miary zmienności, które opierają się na odchyleniach poszczególnych wartości zmiennej od ich średniej arytmetycznej. Będziemy ustalać poszczególne różnice pomiędzy daną wartością zmiennej ($x_j$) a średnią $\bar{x}$. Graficznie możemy to przedstawić tak, jak na rysunku 3.4.




Rys. 3.4. Odchylenia wartości zmiennej od ich średniej arytmetycznej

Źródło: opracowanie własne.

[23] Por. też: M. Woźniak (red.), *Statystyka ogólna*, Kraków 2002; oraz A. Zeliaś, *Metody statystyczne*, *op. cit.*

Chcemy określić średnie odchylenie poszczególnych wartości zmiennej od ich średniej arytmetycznej. Czy nic nie stoi na przeszkodzie, aby posłużyć się średnią arytmetyczną otrzymanych różnic? Napotykamy tu pewną trudność, ponieważ suma odchyleń zaobserwowanych wartości zmiennej od ich średniej arytmetycznej jest zawsze równa zeru[24]. W celu wyeliminowania tej przeszkody musimy pozbyć się znaków algebraicznych („+" lub „–") stojących przed obliczonymi różnicami. Możemy tak postąpić, ponieważ interesują nas jedynie wartości poszczególnych odchyleń, a nie ich kierunek.

Wyeliminowanie znaków można uzyskać dwiema metodami:

- posługując się wartościami bezwzględnymi różnic $|x_i - \bar{x}|$,
- podnosząc te różnice do kwadratu $(x_i - \bar{x})^2$.

Biorąc pod uwagę bezwzględne wartości różnic, zdefiniowano odchylenie przeciętnie. Kwadraty różnic bierzemy pod uwagę, definiując wariancję i odchylenie standardowe.

## 3.2.2. Odchylenie przeciętne

**Odchylenie przeciętne** jest średnią arytmetyczną bezwzględnych wartości odchyleń zmiennej od ich średniej arytmetycznej[25]. Miara ta pozwala określić, o ile wartości zmiennej zaobserwowane dla poszczególnych jednostek statystycznych odbiegają przeciętnie od średniej arytmetycznej. Jest ona zdefiniowana następującym wzorem:

$$d = \frac{\sum_{j=1}^{n} |x_j - \bar{x}|}{n}. \tag{3.16}$$

Według wzoru (3.16) obliczamy odchylenie przeciętne dla szeregu szczegółowego.

**Przykład 3.15**

Wykorzystamy teraz odchylenia przeciętne dla określenia stopnia zróżnicowania obrotów klientów centrów z przykładu 3.14. Średnie obroty klientów obydwu centrów wynosiły 6 mln złotych (por. tablica 3.10). Przebieg obliczeń niezbędnych do uzyskania wartości odchylenia przeciętnego przedstawiono w tablicy 3.11.

Podstawiając wyniki odpowiednich obliczeń do wzoru (3.16), otrzymujemy:

- dla Centrum I: $d_{C1} = \frac{30}{12} = 2{,}5$ [mln złotych].
- dla Centrum II: $d_{C2} = \frac{10}{12} = 0{,}83$ [mln złotych].

[24] Porównaj własność (4) średniej arytmetycznej.

[25] Por.: A. Zeliaś, *Metody statystyczne*, *op. cit.*

**Tablica 3.11.** Obroty klientów centrów obsługi banku oraz obliczenia pomocnicze

| Centrum I | | Centrum II | |
|---|---|---|---|
| Obroty [mln zł] $x_j$ | $\lvert x_j - \bar{x} \rvert$ | Obroty [mln zł] $x_j$ | $\lvert x_j - \bar{x} \rvert$ |
| 7 | 1 | 5 | 1 |
| 2 | 4 | 6 | 0 |
| 3 | 3 | 4 | 2 |
| 11 | 5 | 7 | 1 |
| 5 | 1 | 6 | 0 |
| 9 | 3 | 7 | 1 |
| 6 | 0 | 5 | 1 |
| 1 | 5 | 6 | 0 |
| 8 | 2 | 8 | 2 |
| 6 | 0 | 5 | 1 |
| 10 | 4 | 7 | 1 |
| 4 | 2 | 6 | 0 |
| $\sum_{j=1}^{n} \lvert x_j - \bar{x} \rvert$ | 30 | $\sum_{j=1}^{n} \lvert x_j - \bar{x} \rvert$ | 10 |

Źródło: dane umowne.

Obroty poszczególnych klientów Centrum I różniły się od średniego obrotu wynoszącego 6 mln złotych przeciętnie o 2,5 mln złotych. Obroty poszczególnych klientów Centrum II odbiegały od średniego obrotu wynoszącego 6 mln złotych przeciętnie o 0,83 mln złotych. W obydwu przypadkach wartości średniej arytmetycznej są jednakowe. Możemy stwierdzić, że zróżnicowanie obrotów klientów centrum I jest znacznie większe niż Centrum II.

Jeśli analizujemy zmienność rozkładu liczebności podanego w postaci szeregu rozdzielczego punktowego, wówczas posługujemy się następującym wzorem:

$$d = \frac{\sum_{i=1}^{k} f_i \lvert x_i - \bar{x} \rvert}{\sum_{i=1}^{k} f_i}. \tag{3.17}$$

W przypadku szeregu rozdzielczego zmiennej ciągłej obliczenia odchylenia przeciętnego wykonujemy według wzoru:

$$d = \frac{\sum_{i=1}^{k} f_i \left| x_i' - \bar{x} \right|}{\sum_{i=1}^{k} f_i}, \tag{3.18}$$

gdzie:

$x_i'$ – środki przedziałów klasowych,

$f_i$ – liczebności w poszczególnych przedziałach klasowych,

$\sum_{i=1}^{k} f_i = n$ – ogólna liczebność zbiorowości.

### Przykład 3.16

Urząd gminy pewnej miejscowości stara się o dotację z funduszy Unii Europejskiej na projekt związany z aktywizacją małych przedsiębiorstw. Jednym z wymogów wniosku jest zaprezentowanie dotychczasowego zatrudnienia w małych firmach tej miejscowości. Dane uzyskane przez urząd zostały podzielone na 6 klas pod względem wielkości zatrudnienia (tablica 3.12). Dla rozpatrzenia wniosku wymagana jest charakterystyka zróżnicowania zatrudnienia w małych przedsiębiorstwach.

**Tablica 3.12.** Zatrudnienie w małych firmach w miejscowości oraz obliczenia pomocnicze

| $i$ | Zatrudnienie | Liczba firm $f_i$ | $x_i'$ | $x_i' \cdot f_i$ | $x_i' - \bar{x}$ | $\lvert x_i' - \bar{x} \rvert \cdot f_i$ |
|---|---|---|---|---|---|---|
| 1 | 0–5 | 7 | 2,5 | 17,5 | 7,97 | 55,79 |
| 2 | 5–10 | 10 | 7,5 | 75 | 2,97 | 29,7 |
| 3 | 10–15 | 8 | 12,5 | 100 | 2,03 | 16,24 |
| 4 | 15–20 | 4 | 17,5 | 70 | 7,03 | 28,12 |
| 5 | 20–25 | 2 | 22,5 | 45 | 12,03 | 24,06 |
| 6 | 25–30 | 1 | 27,5 | 27,5 | 17,03 | 17,03 |
| | Razem | 32 | x | 335 | 49,06 | 170,94 |

Źródło: dane umowne.

Rozpoczynamy od obliczenia zatrudnienia średniego. Użyjemy w tym celu średniej arytmetycznej:

$$\bar{x} = \frac{\sum_{i=1}^{k} f_i \cdot x_i'}{\sum_{i=1}^{k} f_i} = \frac{335}{32} = 10{,}47 \text{ [pracowników]}.$$

Odchylenie przeciętne obliczymy według wzoru (3.18) jako:

$$d = \frac{\sum_{i=1}^{k} f_i \left| x_i' - \bar{x} \right|}{\sum_{i=1}^{k} f_i} = \frac{170{,}94}{32} = 5{,}34 \text{ [pracowników]}.$$

Zatrudnienie w poszczególnych firmach w rozważanej miejscowości różni się od średniego zatrudnienia wynoszącego 10,47 pracowników przeciętnie o 5,34 pracowników.

## 3.2.3. Wariancja i odchylenie standardowe

**Wariancja** jest miarą rozproszenia wartości zmiennej wokół średniej arytmetycznej. Jest ona średnią arytmetyczną kwadratów odchyleń poszczególnych wartości zmiennej od ich średniej arytmetycznej. Miarę tę definiujemy następującym wzorem:

$$s^2 = \frac{\sum_{j=1}^{n} (x_j - \bar{x})^2}{n}. \tag{3.19}$$

Własności wariancji są następujące:

1. Wariancja jest miarą mianowana wyrażoną w jednostkach zmiennej podniesionych do kwadratu [$j^2$].
2. Jeśli oznaczymy $(x_j - \bar{x})^2 = z_j$, to $s^2 = \frac{\sum_{j=1}^{k} z_j}{n}$ jest średnią arytmetyczną, a zatem wariancja posiada wszystkie jej własności.
3. Wariancja stałej jest równa zeru. Pamiętając że średnia arytmetyczna stałej jest równa tej stałej, otrzymujemy:

$$s^2 = \frac{\sum_{j=1}^{n} (c - c)^2}{n} = 0.$$

4. Jeśli wszystkie wartości zmiennej pomnożymy przez stałą, to[26]:

$$s^2 = \frac{\sum_{j=1}^{n} (x \cdot c - c \cdot \bar{x})^2}{n} = c^2 \cdot s^2.$$

5. Jeśli rozpatrujemy zmienną $Y$, która jest sumą dwóch innych zmiennych oznaczonych jako $X_1$ oraz $X_2$, to wówczas wariancja zmiennej $Y$ jest sumą warian-

[26] Wykazanie tej własności pozostawiamy Czytelnikowi.

cji zmiennych $X_1$ oraz $X_2$. Tak samo jeśli zmienna $Y$ jest różnicą zmiennych $X_1$ oraz $X_2$, to wówczas wariancja zmiennej $Y$ jest sumą wariancji zmiennych $X_1$ oraz $X_2$.

W rezultacie prostego algebraicznego przekształcenia wzoru (3.19) uzyskujemy wzór, który nie wymaga obliczania odchyleń wartości zmiennej od ich od średniej arytmetycznej. Przyjmuje ona postać:

$$s^2 = \frac{\sum_{j=1}^{n} x_j^2}{n} - \left( \frac{\sum_{j=1}^{n} x_j}{n} \right)^2 = \frac{\sum_{j=1}^{n} x_j^2}{n} - \bar{x}^2. \tag{3.20}$$

Wzór ten jest wygodny w użyciu, ponieważ nie wymaga obliczania odchyleń wartości zmiennej od ich średniej arytmetycznej.

Podobnie jak dla średniej arytmetycznej podajemy wzory, którymi posługujemy się, gdy dysponujemy szeregiem szczegółowym ważonym oraz szeregiem rozdzielczym.

W przypadku szeregu rozdzielczego punktowego posługujemy się wzorem:

$$s^2 = \frac{\sum_{i=1}^{k} f_i \cdot (x_i - \bar{x})^2}{\sum_{i=1}^{k} f_i}. \tag{3.21}$$

Dla szeregu rozdzielczego z przedziałami klasowymi wariancję obliczamy według wzoru:

$$s^2 = \frac{\sum_{i=1}^{k} f_i \cdot (x_i' - \bar{x})^2}{n}. \tag{3.22}$$

Wariancja jest trudna do interpretacji, ponieważ jej mianem jest kwadrat jednostki miary danej zmiennej (np. zł$^2$, szt$^2$). Dla uzyskania miary zmienności wyrażonej w jednostkach zgodnych z mianem rozważanej zmiennej obliczymy pierwiastek kwadratowy z wariancji. Tak uzyskaną miarę nazywamy odchyleniem standardowym, które jest dane jako:

$$s = \sqrt{s^2}. \tag{3.23}$$

**Odchylenie standardowe** jest średnią kwadratową odchyleń poszczególnych wartości zmiennej od ich średniej arytmetycznej. Informuje ono, o ile średnio wartości zmiennej poszczególnych jednostek danej zbiorowości różnią się od ich średniej arytmetycznej.

### Przykład 3.17

Wykorzystamy wariancję i odchylenie standardowe do pomiaru zróżnicowania obrotów klientów banku podzielonych na dwa centra (zob. tabl. 3.9) Odpowiednie obliczenia przedstawiono w tablicy 3.13.

**Tablica 3.13.** Wielkość obrotów klientów centrów obsługi banku oraz obliczenia pomocnicze

| Centrum I | | | Centrum I | | |
|---|---|---|---|---|---|
| Obroty [mln zł] $x_j$ | $x_j' - \bar{x}$ | $(x_j' - \bar{x})^2$ | Obroty [mln zł] $x_j$ | $x_j' - \bar{x}$ | $(x_j' - \bar{x})^2$ |
| 7 | 1 | 1 | 5 | –1 | 1 |
| 2 | –4 | 16 | 6 | 0 | 0 |
| 3 | –3 | 9 | 4 | –2 | 4 |
| 11 | 5 | 25 | 7 | 1 | 1 |
| 5 | –1 | 1 | 6 | 0 | 0 |
| 9 | 3 | 9 | 7 | 1 | 1 |
| 6 | 0 | 0 | 5 | –1 | 1 |
| 1 | –5 | 25 | 6 | 0 | 0 |
| 8 | 2 | 4 | 8 | 2 | 4 |
| 6 | 0 | 0 | 5 | –1 | 1 |
| 10 | 4 | 16 | 7 | 1 | 1 |
| 4 | –2 | 4 | 6 | 0 | 0 |
| Ogółem | × | 110 | Ogółem | × | 14 |

Źródło: dane umowne.

Uzyskane wyniki podstawiamy do wzorów na wariancję (3.19) i odchylenie standardowe (3.23) i otrzymujemy:

dla Centrum I:

$$s_{C1}^2 = \frac{110}{12} = 9{,}17 \text{ [mln złotych}^2\text{]}$$

$$s_{C1} = \sqrt{9{,}17} = 3{,}03 \text{ [mln złotych]}$$

dla Centrum II:

$$s_{C2}^2 = \frac{14}{12} = 1{,}17 \text{ [mln złotych}^2\text{]}$$

$$s_{C2} = \sqrt{1{,}17} = 1{,}08 \text{ [mln złotych].}$$

Wielkości obrotów poszczególnych klientów Centrum I odbiegały od średniego obrotu wynoszącego 6 mln złotych przeciętnie o 3,03 mln złotych. W Centrum II obroty poszczególnych klientów różniły się od średniego obrotu, wynoszącego również 6 mln złotych, przeciętnie o 1,08 mln złotych.

Procedurę obliczania wariancji i odchylenia standardowego dla szeregu rozdzielczego przedstawiamy na następującym przykładzie.

### Przykład 3.18

Powracamy do przykładu 3.16, w którym rozpatrywaliśmy zatrudnienie w małych firmach. Na podstawie wcześniej wykonanych obliczeń ustaliliśmy, że średnie zatrudnienie $\bar{x} = 10{,}47$ pracowników. Obliczenia prowadzące do uzyskania wariancji i odchylenia standardowego zamieszczamy w tablicy 3.14.

**Tablica 3.14.** Zatrudnienie w małych firmach oraz obliczenia pomocnicze

| Obroty [mln zł] $x_i$ | Liczba firm $f_i$ | $x_i'$ | $x_i' - \bar{x}$ | $(x_i' - \bar{x})^2$ | $f_i \cdot (x_i' - \bar{x})^2$ |
|---|---|---|---|---|---|
| 0–5 | 7 | 2,5 | –7,97 | 63,52 | 444,65 |
| 5–10 | 10 | 7,5 | –2,97 | 8,82 | 88,21 |
| 10–15 | 8 | 12,5 | 2,03 | 4,12 | 32,97 |
| 15–20 | 4 | 17,5 | 7,03 | 49,42 | 197,68 |
| 20–25 | 2 | 22,5 | 12,03 | 144,72 | 289,44 |
| 25–30 | 1 | 27,5 | 17,03 | 290,02 | 290,02 |
| Razem | 32 | × | × | × | 1342,97 |

Źródło: dane umowne.

Wariancja jest równa:

$$s^2 = \frac{1342{,}97}{32} = 41{,}97 \text{[pracowników}^2\text{]}.$$

Odchylenie standardowe równa się:

$$s = \sqrt{41{,}97} = 6{,}48 \text{[pracowników]}.$$

Zatrudnienie w poszczególnych firmach różni się od przeciętnego zatrudnienia wynoszącego 10,47 osób średnio o 6,48 osób.

W tym miejscu należy zauważyć, że między wartościami odchylenia standardowego oraz odchylenia przeciętnego zachodzi nierówność[27]:

$$s \geq d$$

$s = d$ tylko wtedy, gdy brak jest odchyleń od wartości średniej, co oznacza, że obydwie miary przyjmują wartość równą zeru.

## 3.2.4. Współczynnik zmienności

**Współczynnik zmienności** jest względną miarą zmienności. Jest on ilorazem bezwzględnej miary zmienności i odpowiedniej średniej. Najczęściej stosuje się współczynniki oparte na takich miarach bezwzględnych, jak odchylenie standardowe i odchylenie przeciętne. Jest miarą niemianowaną. Wyrażamy go w procentach. Współczynnik zmienności jest zdefiniowany za pomocą podanych następujących wzorów:

- współczynnik zmienności oparty na odchyleniu przeciętnym:

$$V_d = \frac{d}{\bar{x}} \cdot 100\% \quad \bar{x} \neq 0. \tag{3.24}$$

Obliczona wartość informuje, jaki procent średniej arytmetycznej stanowi odchylenie przeciętne.

- współczynnik zmienności oparty na odchyleniu standardowym:

$$V_s = \frac{s}{\bar{x}} \cdot 100\% \quad \bar{x} \neq 0. \tag{3.25}$$

Obliczona wartość informuje, jaki procent średniej arytmetycznej stanowi odchylenie standardowe.

Współczynnik zmienności może być używany do porównań jednej zbiorowości ze względu na zmienność kilku cech o różnych mianach albo kilku zbiorowości pod względem tej samej cechy. Przyjmuje się, że jeżeli współczynnik zmienności $V_s$ przekracza 20%, to cechy wykazują znaczne zróżnicowanie[28].

### Przykład 3.19

Chcemy zakupić akcje jednej ze spółek giełdowych A lub B. Przyjętym przez nas kryterium jest stabilność ceny. Przeanalizowaliśmy ceny akcji dwóch interesujących nas spółek giełdowych, otrzymując następujące odchylenia standardowe:

$$s_A = 20\,[\text{zł}] \text{ oraz } s_B = 100\,[\text{zł}].$$

Na podstawie analizy odchyleń standardowych moglibyśmy przypuszczać, że ceny akcji B mają większą zmienność, ponieważ przeciętne ceny odchylają się od swojej

[27] Zależność tę możemy wykorzystać do sprawdzenia prawidłowości wykonanych obliczeń.
[28] Por.: M. Woźniak (red.), *op. cit.*

średniej o 100 złotych, a ceny akcji A tylko o 20 złotych. Na tej podstawie wybralismy akcję spółki A.

Czy rzeczywiście dokonaliśmy dobrego wyboru ze względu na przyjęte kryterium? Najpierw sprawdzimy, ile wynosiła średnia cena akcji:

$$\bar{x}_A = 40\,[\text{zł}] \quad \bar{x}_B = 500\,[\text{zł}],$$

a następnie policzymy współczynniki zmienności:

$$V_A = \frac{20}{40} \cdot 100\% = 50\%$$

$$V_B = \frac{100}{500} \cdot 100\% = 20\%.$$

Teraz sytuacja przedstawia się inaczej. Widzimy, że zmienność cen akcji B jest mniejsza od zmienności cen akcji A. Odchylenie standardowe dla cen akcji B stanowi 20% średniej ceny akcji, a w przypadku ceny akcji A wynosi aż 50%. Zatem popełniono błąd, podejmując decyzję o zakupieniu akcji spółki A. Dopiero współczynnik zmienności daje nam możliwość porównania dwóch zbiorowości – w tym przypadku cen dwóch różnych spółek giełdowych.

**Przykład 3.20**

Mierzymy rozproszenie obrotów klientów centrów obsługi banku (por. przykład 3.17). Wybieramy współczynnik zmienności na podstawie odchylenia standardowego:

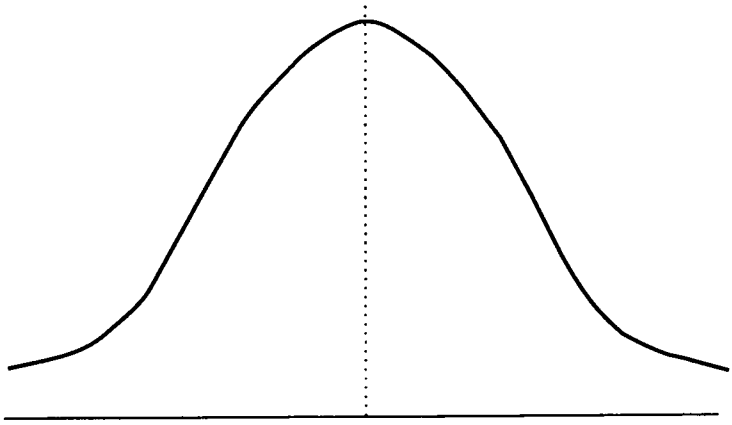
- dla Centrum I: $V_S = \frac{3{,}03}{6} \cdot 100\% = 50{,}5\%$,
- dla Centrum II: $V_S = \frac{1{,}08}{6} \cdot 100\% = 18\%$.

Zmienność obrotów klientów Centrum I jest znacznie większa, współczynnik zmienności wynosi 50,5%, niż obrotów klientów Centrum II, dla którego współczynnik ten jest równy 18%.

## 3.3. Miary asymetrii

Statystyczny opis zbiorowości nie powinien ograniczać się jedynie do określenia położenia oraz zmienności rozkładu liczebności. Może bowiem okazać się, że miary położenia oraz zmienności w porównywanych populacjach będą przyjmować podobne wartości, a mimo to zbiorowości te mogą posiadać odmienny rozkład liczebności. Rozkład ten może być symetryczny albo asymetryczny lewostronnie albo prawostron-

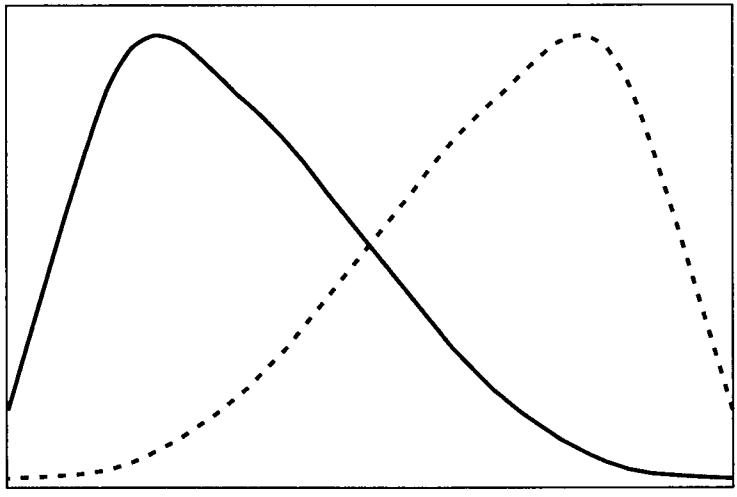
nie. Do określenia asymetrii rozkładu możemy użyć miar położenia. W rozkładach symetrycznych jedna połowa rozkładu wyników obserwacji jest lustrzanym odbiciem drugiej. Jeśli rozkład jest symetryczny, to średnia arytmetyczna, mediana i modalna są sobie równe (por. rys. 3.5).



$\bar{x} = M_o = M_e$

Rys. 3.5. Rozkład symetryczny

Jeśli wartości średniej, mediany i modalnej przyjmują różne wartości, to wówczas mówimy o asymetrii. Jej typy przedstawiono na rysunku 3.6



Rys. 3.6. Typy asymetrii rozkładu liczebności

Gdy krzywa ilustrująca rozkład wznosi się stromo do góry a następnie opada łagodnie w dół, to mamy do czynienia z asymetrią prawostronną (średnia znajduje się

na prawo od modalnej). Jeśli krzywa ta wznosi się łagodnie w górę a opada stromo w dół, mówimy o asymetrii lewostronnej (średnia znajduje się na lewo od modalnej).

Przedstawimy teraz dwie proste miary, na podstawie których będziemy badali asymetrię rozkładu. **Wskaźnik asymetrii** jest najprostszą miarą. Opiera się na różnicy między średnią arytmetyczną i modalną. Jest on zdefiniowany w następujący sposób[29]:

$$W_A = \bar{x} - M_O. \tag{3.26}$$

Możemy wyróżnić następujące przypadki:

1. Jeśli rozkład jest symetryczny, to $W_A = 0$.
   W rozkładzie symetrycznym jednakowy jest udział jednostek o wartościach zmiennej nie większych i nie mniejszych od średniej.
2. Jeśli rozkład jest asymetryczny lewostronnie, to $W_A < 0$.
   Asymetria lewostronna (ujemna) oznacza, że w badanej zbiorowości przeważają jednostki o wartościach zmiennej większych od średniej.
3. Jeśli rozkład jest asymetryczny prawostronnie, to $W_A > 0$.
   Asymetria prawostronna (dodatnia) oznacza, że w badanej zbiorowości przeważają jednostki o wartościach zmiennej mniejszych od średniej.

Wskaźnik asymetrii jest miarą bezwzględną, mianowaną i nie pozwala porównywać różnych rozkładów. Ponadto określa jedynie kierunek asymetrii, nie mówiąc nic o jej sile. Niedogodności te można wyeliminować, dzieląc wskaźnik asymetrii przez odchylenie standardowe lub odchylenie przeciętne. W ten sposób otrzymujemy **współczynnik asymetrii**, który jako miara względna, niemianowana umożliwia porównania asymetrii rozkładów w różnych populacjach scharakteryzowanych różnymi zmiennymi. Możemy posługiwać się współczynnikiem asymetrii zdefiniowanym wzorem (3.27) lub (3.28).

$$A_s = \frac{\bar{x} - M_o}{s}, \tag{3.27}$$

$$A_s^d = \frac{\bar{x} - M_o}{d}, \tag{3.28}$$

- gdy rozkład jest symetryczny, to $A_s = 0$,
- gdy rozkład jest asymetryczny lewostronnie, to $A_s < 0$,
- gdy rozkład jest asymetryczny prawostronnie, to $A_s > 0$.

Współczynnik asymetrii określa zarówno kierunek (asymetria prawo- albo lewostronna), jak i jej siłę. Im większą przyjmie wartość, tym asymetria jest silniejsza.

---

[29] Por. też: M. Woźniak (red.), *op. cit.*; oraz A. Zeliaś, *Metody statystyczne*, *op. cit.*

### Przykład 3.21

Ministerstwo posiada 10 samochodów służbowych. Dane zamieszczone w tablicy 3.15 przedstawiają roczny przebieg pojazdów w tysiącach kilometrów.

**Tablica 3.15.** Przebieg samochodów służbowych ministerstwa oraz obliczenia pomocnicze

| Nr pojazdu | Przebieg [tys. km] | $(x_j' - \bar{x})^2$ |
|---|---|---|
| 1 | 26 | 0,36 |
| 2 | 12 | 213,16 |
| 3 | 32 | 29,16 |
| 4 | 18 | 73,96 |
| 5 | 16 | 112,36 |
| 6 | 30 | 11,56 |
| 7 | 34 | 54,76 |
| 8 | 30 | 11,56 |
| 9 | 24 | 6,76 |
| 10 | 44 | 302,76 |
| Razem | 266 | 816,4 |

Źródło: dane umowne.

Należy ocenić asymetrię rozkładu pojazdów ministerstwa ze względu na ich roczny przebieg.

Do uzyskania współczynnika asymetrii musimy obliczyć:

- średnią arytmetyczną:

$$\bar{x} = \frac{266}{10} = 26{,}6 \text{ [tys. km]}.$$

Średni przebieg samochodów ministerstwa wynosił 26,6 tysięcy kilometrów.

- modalną:

$$Mo = 30 \text{ [tys. km]}.$$

Wśród 10 pojazdów najczęstszy przebieg wynosił 30 tysięcy kilometrów.

- odchylenie standardowe:

$$s = \sqrt{\frac{816{,}4}{10}} = 9{,}04 \text{ [tys. km]}.$$

Przebiegi poszczególnych samochodów różnią się od średniego przebiegu przeciętnie o 9,04 tysięcy kilometrów.

Obliczone wartości podstawiamy do wzoru (3.27) i otrzymamy:

- współczynnik asymetrii:

$$As = \frac{26,6 - 30}{9,04} = -0,37.$$

Rozkład przebiegu samochodów służbowych w Ministerstwie w badanym roku charakteryzuje się asymetrią lewostronną. Oznacza to, że przeważają pojazdy o przebiegu większym od średniego, który wynosi 26,6 tysięcy kilometrów.

## 3.4. Miary koncentracji

Kolejną grupą miar są miary opisujące koncentrację jednostek zbiorowości wokół wartości średniej. Należy tutaj dodać, że koncentracja ma ścisły związek ze zróżnicowaniem wartości zmiennej. Im większe jest zróżnicowanie, tym mniejsza jest koncentracja.

Jedną z miar koncentracji jest **współczynnik koncentracji**. Jest on zdefiniowany następującym wzorem:

$$K = \frac{\mu_4}{s^4}, \tag{3.29}$$

gdzie:

$s^4$ – odchylenie standardowe poniesione do czwartej potęgi,

$\mu_4$ – moment centralny czwartego rzędu obliczany w następujący sposób:

- dla szeregu szczegółowego:

$$\mu_4 = \frac{\sum_{j=1}^{n}(x_j - \bar{x})^4}{n}, \tag{3.30}$$

- dla szeregu rozdzielczego z przedziałami klasowymi:

$$\mu_4 = \frac{\sum_{i=1}^{k} f_i \cdot (x_i' - \bar{x})^4}{\sum_{i=1}^{k} f_i}, \tag{3.31}$$

Miara ta informuje o skupieniu wartości zmiennej poszczególnych jednostek statystycznych wokół średniej arytmetycznej. Punktem odniesienia dla porównywania spłaszczenia rozkładów jest koncentracja w rozkładzie normalnym[30]. Dla rozkładu normalnego współczynnik ten wynosi 3. Zatem im K jest większe od 3, tym badany rozkład jest bardziej wysmukły niż rozkład normalny (większa koncentracja). Im bardziej K jest

[30] Rozkład normalny będzie scharakteryzowany w rozdziale 6.

mniejsze od 3, tym rozkład jest bardziej spłaszczony niż rozkład normalny. Oznacza to mniejszą koncentrację[31]. Współczynnik koncentracji jest miarą względną, niemianowaną, co sprawia, że dzięki niemu możemy porównywać badane zbiorowości.

Modyfikacją współczynnika koncentracji jest **współczynnik ekscesu** określany wzorem:

$$K_u = \frac{\mu_4}{s^4} - 3. \tag{3.32}$$

Odjęcie liczby 3 od współczynnika koncentracji ułatwia interpretację, ponieważ wartość współczynnika ekscesu dla rozkładu normalnego wynosi zero.

Gdy:

$K_u = 0$ – to rozkład ma koncentrację charakterystyczną dla rozkładu normalnego,

$K_u < 0$ – to rozkład jest bardziej spłaszczony niż normalny (mniejsza koncentracja),

$K_u > 0$ – to rozkład jest bardziej wydłużony niż normalny (większa koncentracja).

### Przykład 3.22

Powróćmy do przykładu 3.21, w którym analizowaliśmy przebieg samochodów służbowych w ministerstwie. Teraz zbadamy koncentrację rozkładu za pomocą współ-

**Tablica 3.16.** Przebieg samochodów ministerstwa oraz obliczenia pomocnicze

| Nr pojazdu $j$ | Przebieg [tys. km] | $x_j - \bar{x}$ | $(x_j - \bar{x})^4$ |
|---|---|---|---|
| 1 | 26 | –0,6 | 0,13 |
| 2 | 12 | –14,6 | 45 437,19 |
| 3 | 32 | 5,4 | 850,31 |
| 4 | 18 | –8,6 | 5 470,08 |
| 5 | 16 | –10,6 | 12 624,77 |
| 6 | 30 | 3,4 | 133,63 |
| 7 | 34 | 7,4 | 2 998,66 |
| 8 | 30 | 3,4 | 133,63 |
| 9 | 24 | –2,6 | 45,70 |
| 10 | 44 | 17,4 | 91 663,62 |
| Razem | 266 | × | 159 357,71 |

Źródło: dane umowne.

[31] Por. A. Zeliaś, *Metody statystyczne*, *op. cit.*

czynnika ekscesu, korzystając ze wzoru (3.32). Odpowiednie obliczenia zamieszczono w tablicy 3.16.

Najpierw obliczamy wartość momentu centralny czwartego rzędu (wzór (3.30)):

$$\mu_4 = \frac{159357{,}71}{10} = 15935{,}77,$$

następnie podnosimy obliczone wcześniej odchylenie standardowe do potęgi czwartej.

$$s^4 = (9{,}04)^4 = 6678{,}42.$$

Uzyskane wyniki podstawiamy do wzoru (3.32).

$$K_u = \frac{15935{,}77}{6678{,}42} - 3 = -0{,}61.$$

Współczynnik ekscesu przyjął wartość mniejszą od 0. Oznacza to, że rozkład przebiegu pojazdów jest bardziej spłaszczony niż wówczas, gdy podlegałby rozkładowi normalnemu.

### Przykład 3.23

W rozdziale 2 konstruowaliśmy szereg rozdzielczy przedstawiający rozkład liczebności 50 gospodarstw domowych. W tablicy 3.17 przedstawiono obliczenia prowadzące do uzyskania wartości odchylenia przeciętnego, wariancji i współczynnika ekscesu. W tablicy 3.18 zebrano wartości wszystkich przedstawionych charakterystyk opisowych rozkładu liczebności gospodarstw domowych według wysokości dochodu.

**Tablica 3.17.** Rozkład liczebności gospodarstw domowych według wysokości dochodów oraz obliczenia pomocnicze

| Dochody | $f_i$ | $x_i'$ | $x_i' \cdot f_i$ | $f_i \lvert x_i' - \bar{x} \rvert$ | $f_i \cdot (x_i - \bar{x})^2$ | $f_i \cdot (x_i - \bar{x})^4$ |
|---|---|---|---|---|---|---|
| 1,00–1,24 | 4 | 1,12 | 4,48 | 2,44 | 1,486 | 0,552 |
| 1,24–1,48 | 9 | 1,36 | 12,24 | 3,33 | 1,229 | 0,168 |
| 1,48–1,72 | 15 | 1,60 | 24,00 | 1,94 | 0,252 | 0,004 |
| 1,72–1,96 | 9 | 1,84 | 16,56 | 0,99 | 0,110 | 0,001 |
| 1,96–2,20 | 6 | 2,08 | 12,48 | 2,10 | 0,737 | 0,090 |
| 2,20–2,44 | 5 | 2,32 | 11,60 | 2,95 | 1,743 | 0,608 |
| 2,44–2,68 | 2 | 2,56 | 5,12 | 1,66 | 1,379 | 0,951 |
| Suma | 50 | × | 86,48 | 15,419 | 6,936 | 2,375 |

Źródło: obliczenia własne.

**Tablica 3.18.** Charakterystyki rozkładu liczebności gospodarstw domowych według wysokości dochodów

| Parametr | Wartość parametru |
|---|---|
| średnia arytmetyczna | $\bar{x} = \frac{86,48}{50} = 1,730$ [tys. zł] |
| modalna | $Mo = 1,48 + 0,24 \cdot \frac{15-9}{(15-9)+(15-9)} = 1,60$ [tys. zł] |
| mediana | $Me = 1,48 + \frac{0,24}{15} \cdot (25-13) = 1,672$ [tys. zł] |
| kwartyl pierwszy | $Q_1 = 1,24 + \frac{0,24}{9} \cdot (12,5-4) = 1,467$ [tys. zł] |
| kwartyl trzeci | $Q_3 = 1,72 + \frac{0,24}{9} \cdot (3,75-28) = 1,973$ [tys zł] |
| odchylenie przeciętne | $d = \frac{15,418}{50} = 0,308$ [tys zł] |
| wariancja | $s^2 = \frac{6,936}{50} = 0,139$ [tys. zł]$^2$ |
| odchylenie standardowe | $s = \sqrt{0,139} = 0,372$ [tys. zł] |
| współczynnik asymetrii | $As = \frac{1,744-1,600}{0,365} = 0,347$ |
| moment centralny rzędu czwartego | $\mu_4 = \frac{2,375}{50} = 0,047$ [tys. zł]$^4$ |
| współczynnik ekscesu | $K = \frac{0,047}{(0,372)^4} = \frac{0,047}{0,019} - 3 = 2,480 - 3 = -0,480$ |

Źródło: obliczenia własne.

Średni dochód badanej grupy gospodarstw domowych wynosi 1 730 złotych. Najczęstszy dochód jest równy 1 600 złotych. Połowa gospodarstw domowych posiada dochody nie wyższe niż 1 672 złotych, a połowa nie niższe od tej wartości. 25% gospodarstw osiąga dochód nie wyższy niż 1 467 złotych, a 75% nie niższy. Kwartyl górny ($Q_3$) wskazuje, iż 75% gospodarstw posiada dochody nie wyższe niż 1 973 złotych, a 25 nie niższe od tej kwoty. Odchylenie przeciętne wskazuje, że dochody poszcze-

gólnych gospodarstw różnią się od dochodu średniego przeciętnie o 308 złotych. Odchylenie standardowe informuje, że to średnie zróżnicowanie wynosi 372 złotych. Rozkład charakteryzuje się asymetrią prawostronną (dodatnią), a więc w badanej zbiorowości przeważają gospodarstwa o dochodach niższych od dochodu średniego. Rozkład liczebności gospodarstw według wysokości dochodów jest bardziej smukły niż rozkład normalny. Wartości zmiennej są skoncentrowane wokół średniej arytmetycznej bardziej, niż wówczas gdybyśmy mieli do czynienia z normalnym rozkładem rozważanej zmiennej.

# Analiza współzależności zjawisk

Rozdział 4

## 4.1. Wprowadzenie

Odkrywanie związków zachodzących pomiędzy zjawiskami jest jedną z podstawowych umiejętności człowieka. Zdolność ta jest wykorzystywana w każdym rodzaju działalności ludzkiej (życie prywatne, zawodowe, nauka itp.). Na przykład uczeń w szkole powinien odkryć zależność między uczeniem się a uzyskiwaniem dobrych stopni. Pracownik powinien być wynagradzany w taki sposób, aby rozumiał, że lepsza praca powoduje uzyskiwanie wyższego wynagrodzenia. Makler chciałby poznać zależność między stanami gospodarki a ceną papierów wartościowych na giełdzie. Podczas kampanii wyborczej sztab danego kandydata chciałby poznać różne zależności, np. z iloma wyborcami powinien się spotkać kandydat, żeby uzyskać dobry wynik wyborczy, jak akcja rozdawania ulotek wpływa na ostateczny wynik wyborów itp. Podane przykłady świadczą o tym, że poszukiwanie zależności między zjawiskami jest powszechne i stanowi bardzo ważny problem.

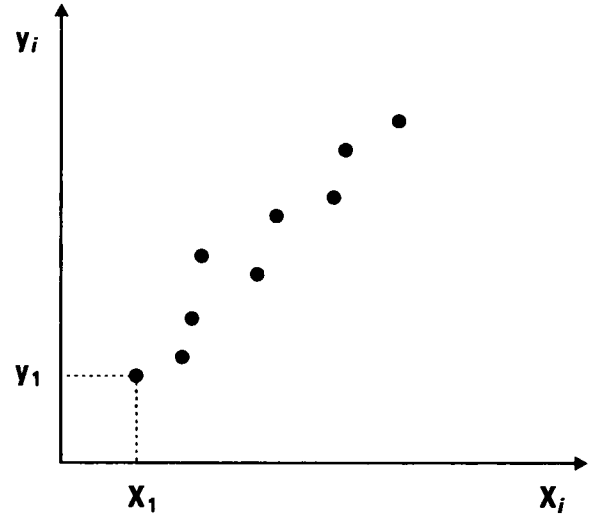
Przedstawimy teraz statystyczne metody badania związków między zmiennymi odzwierciedlającymi różnego rodzaju zjawiska. Badania te noszą nazwę **analizy współzależności zjawisk**. Analiza może dotyczyć zarówno zjawisk, które możemy zmierzyć i wyrazić w postaci liczbowej odpowiednich jednostkach (np. w metrach, sztukach itp.) lub wartościowo w jednostkach pieniężnych, jak i zjawisk, które są niemierzalne i najczęściej określa się je słownie, tak jak na przykład: preferencje, uczucia, wykształcenie[31] itp.

W badaniach współzależności możemy interesować się związkami między dwoma zjawiskami, które możemy wyrazić przez zmienne, które oznaczymy jako $X$ i $Y$. Związek między nimi możemy przedstawić graficznie w postaci **diagramu korelacyjnego**. W układzie współrzędnych zaznaczamy kolejne punkty odpowiadające wynikom obserwacji. Każda obserwacja jest opisana przez parę liczb: $x_i$ oraz $y_i$ będą-

[31] A. Zeliaś, *Metody statystyczne*, *op. cit.*

cych realizacjami zmiennych $X$ i $Y$, gdzie symbol $i$ oznacza numer jednostki statystycznej ($i = 1, 2,..., n$).

Na przykład badając związek między wzrostem i masą ciała, przeprowadzimy obserwację, w której jednostkami są osoby. Każdej z nich przypisano parę liczb $(x_i, y_i)$ będących odpowiednio wartościami zmiennej $X$ (wzrost) oraz zmiennej $Y$ (masa ciała). Punkty na wykresie są ilustracją wyników obserwacji. Przykładową postać związku przedstawiono na rysunku 4.1.




Rys. 4.1. Diagram korelacyjny między zmiennymi $X$ i $Y$

Źródło: opracowanie własne.

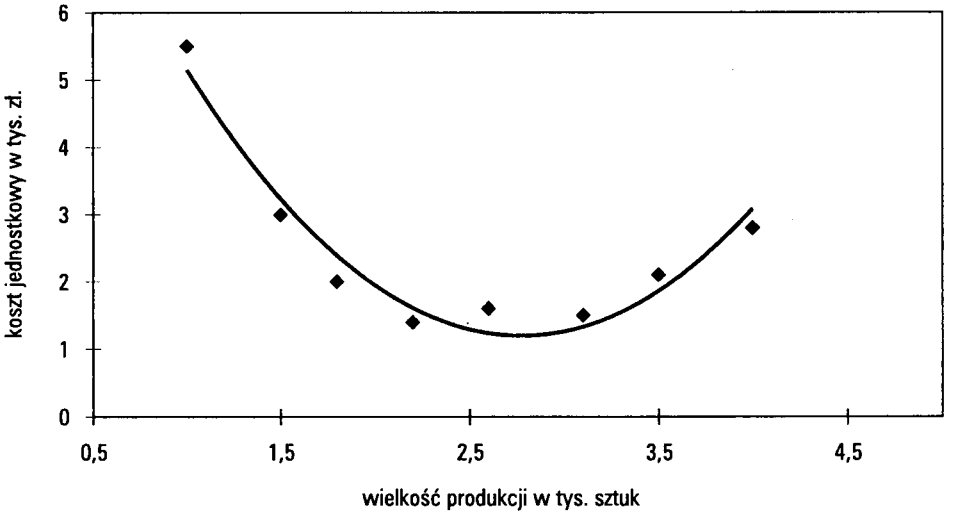
Na podstawie diagramu korelacyjnego możemy uzyskać odpowiedź na następujące pytania:

1) czy między rozpatrywanymi zmiennymi występuje związek?
2) jaka jest jego postać; liniowa czy nieliniowa?
3) jaki jest kierunek powiązań; dodatni czy ujemny?
4) czy możemy traktować te powiązania jako silne?

Na podstawie wizualnej analizy możemy uzyskać ważne wskazówki co do dalszego postępowania badawczego. Biorąc pod uwagę układ punktów, możemy powiedzieć, że pomiędzy rozpatrywanymi zmiennymi występuje związek, ponieważ punkty nie są rozrzucone chaotycznie po całej powierzchni układu, ale tworzą smugę, która sugeruje występowanie związku. Punkty obrazujące obserwacje empiryczne układają się w linię prostą, dlatego możemy przypuszczać, że mamy do czynienia ze związkiem o postaci liniowej. Kierunek powiązań zmiennej $Y$ i zmiennej $X$ jest dodatni, ponieważ wzrastającym wartościom zmiennej $X$ odpowiada wzrost średnich wartości zmien-

nej $Y$. Możemy przypuszczać, że związek ten jest dość silny, ponieważ rozproszenie punktów nie jest duże.

Inną postać związku przedstawia diagram korelacyjny – rys. 4.2. Jest to związek **nieliniowy**, w tym wypadku paraboliczny. Może on występować między zmienną $Y$ (koszty jednostkowe) oraz zmienną $X$ (wielkość produkcji). Związek nieliniowy może mieć również inną postać, na przykład: funkcji hiperbolicznej, wykładniczej, potęgowej, logistycznej.



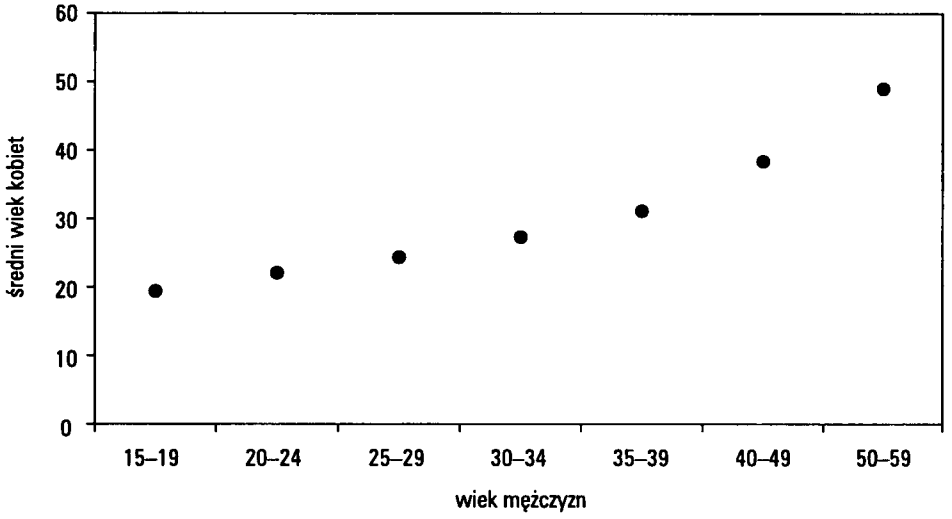

Rys. 4.2. Współzależność między kosztami jednostkowymi i wielkością produkcji

Na podstawie diagramu korelacyjnego możemy ustalić, że występuje dość silny związek między długością serii i kosztami jednostkowymi. Paraboliczny kształt oznacza zmianę kierunku powiązań z ujemnego na dodatni. Oznacza to, że istnieje pewna optymalna długość serii, po przekroczeniu której należy oczekiwać wzrostu kosztów jednostkowych.

Stwierdzenie związku między zmiennymi tylko na podstawie diagramu korelacyjnego nie może zadowolić statystyka. Prawidłowości te są bowiem sformułowane nieprecyzyjnie. Będziemy zatem poszukiwać odpowiednich mierników, które pozwolą zmierzyć siłę i ustalić kierunek powiązań między rozpatrywanymi zmiennymi. W dalszym ciągu będziemy starać się ująć te relacje w postaci liczbowej, wykorzystując w tym celu metody statystyczne. W tym zakresie możemy zastosować dwa ściśle ze sobą powiązane ujęcia, a mianowicie analizę korelacji i analizę regresji.

Związek korelacyjny występuje wówczas, gdy wartościom jednej zmiennej są przyporządkowane warunkowe rozkłady drugiej zmiennej. Jeśli warunkowy rozkład ujmiemy sumarycznie za pomocą wartości średniej, to możemy powiedzieć, że w związku korelacyjnym wartościom jednej zmiennej przyporządkowane są średnie warunkowe drugiej zmiennej.

W tablicy 4.1 podano przykładowy rozkłady liczby mężczyzn i kobiet według wieku w chwili zawarcia małżeństwa. Tablicę tę nazywamy tablicą korelacyjną. Zestawiono w niej dwa szeregi rozdzielcze. Rozkład liczby mężczyzn (kolumny) i liczby kobiet (wiersze) według wieku w chwili zawarcia małżeństwa. Wiekowi mężczyzny [kolumna (1)] jest przyporządkowany średni wiek kobiet [kolumna (11)]. Związek ten przedstawiono na rysunku 4.3.




Rys. 4.3. Związek między wiekiem mężczyzn i średnim wiekiem kobiet w chwili zawarcia małżeństwa w Polsce w 2000 roku

## 4.2. Współczynnik korelacji liniowej Pearsona

Współczynnik korelacji liniowej Pearsona jest jednym z najczęściej stosowanych mierników powiązania między zmiennymi[32]. Współczynnik ten znajduje zastosowanie, gdy zmienne są rezultatem pomiaru w skali przynajmniej interwałowej. Jest on zdefiniowany jako:

$$r(X,Y) = \frac{\text{cov}(X,Y)}{\sqrt{\text{var}(X) \cdot \text{var}(Y)}}, \tag{4.1}$$

gdzie:
var $(X)$, var $(Y)$ – wariancje zmiennych[33] $X$ i $Y$,
cov $(X, Y)$ – kowariancja między zmiennymi $X$ i $Y$.

[32] Por. np.: M. Woźniak (red.), *op.cit.* oraz A. Zeliaś, *Metody statystyczne*, *op. cit.*; W. Starzyńska, *Statystyka praktyczna*, Warszawa 2000.
[33] Miary te zostały przedstawione w punkcie 3.2.3.

**Tablica 4.1.** Rozkład liczby mężczyzn i kobiet według wieku w chwili zawarcia małżeństwa w Polsce w 2000 roku

| Wiek mężczyzn ($Y$) | Wiek kobiet ($X$) | | | | | | | | | Średni wiek kobiet $\bar{x}_j$ |
|---|---|---|---|---|---|---|---|---|---|---|
| | | 15–19 | 20–24 | 25–29 | 30–34 | 35–39 | 40–44 | 45–49 | 50–54 | |
| | $x_i'$ | 17,5 | 22,5 | 27,5 | 32,5 | 37,5 | 42,5 | 47,5 | 52,5 | |
| (1) | (2) | (3) | (4) | (5) | (6) | (7) | (8) | (9) | (10) | (11) |
| 15–19 | 17,5 | 3498 | 1848 | 79 | 11 | 1 | 0 | 0 | 0 | 19,38 |
| 20–24 | 22,5 | 16714 | 60412 | 8335 | 444 | 70 | 7 | 2 | 7 | 22,08 |
| 25–29 | 27,5 | 4577 | 41370 | 25630 | 2265 | 354 | 83 | 27 | 12 | 24,33 |
| 30–34 | 32,5 | 653 | 5813 | 8108 | 3529 | 804 | 265 | 75 | 15 | 27,29 |
| 35–39 | 37,5 | 142 | 1311 | 2452 | 2043 | 1273 | 572 | 211 | 57 | 31,12 |
| 40–49 | 45 | 73 | 470 | 1114 | 1412 | 1682 | 2028 | 1449 | 646 | 38,36 |
| 50–59 | 55 | 11 | 51 | 133 | 200 | 398 | 851 | 1684 | 5879 | 49,01 |
| Średni wiek mężczyzn | | 23,13 | 25,08 | 28,50 | 34,16 | 39,89 | 44,82 | 49,00 | 53,74 | × |

Źródło: „Rocznik Demograficzny", GUS, Warszawa 2001.

$$\operatorname{cov}(X,Y)=\frac{\sum_{i=1}^{n}(x_i-\bar{x})\cdot(y_i-\bar{y})}{n}, \tag{4.2}$$

$$\operatorname{var}(X)=\frac{\sum_{i=1}^{n}(x_i-\bar{x})^2}{n}; \qquad \operatorname{var}(Y)=\frac{\sum_{i=1}^{n}(y_i-\bar{y})^2}{n},$$

zatem:

$$r=\frac{\frac{\sum_{i=1}^{n}(x_i-\bar{x})\cdot(y_i-\bar{y})}{n}}{\sqrt{\frac{\sum_{i=1}^{n}(x_i-\bar{x})^2}{n}}\cdot\sqrt{\frac{\sum_{i=1}^{n}(y_i-\bar{y})^2}{n}}}. \tag{4.3}$$

W obliczeniach wygodniej jest posługiwać się wzorem, który uzyskujemy, upraszczając przez $n$ ułamek dany wzorem (4.3). Otrzymujemy w rezultacie[34]:

$$r=\frac{\sum_{i=1}^{n}(x_i-\bar{x})(y_i-\bar{y})}{\sqrt{\sum_{i=1}^{n}(x_i-\bar{x})^2\sum_{i=1}^{n}(y_i-\bar{y})^2}}. \tag{4.4}$$

**Współczynnik korelacji liniowej przyjmuje wartości z przedziału [–1; 1]** i określa zarówno siłę, jak i kierunek powiązań między zmiennymi.

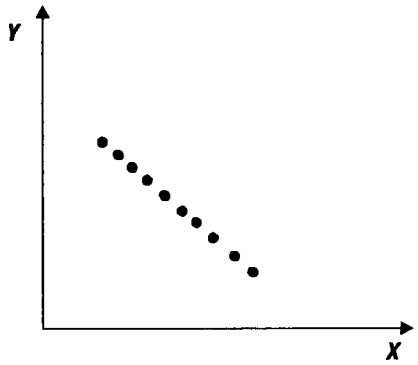
Korelacja liniowa może być:

- **dodatnia** – jeżeli wzrostowi wartości jednej zmiennej towarzyszy wzrost średnich wartości drugiej zmiennej,
- **ujemna** – jeżeli wzrostowi jednej zmiennej odpowiada spadek średnich wartości drugiej zmiennej.

Rozpatrzmy następujące szczególne przypadki.

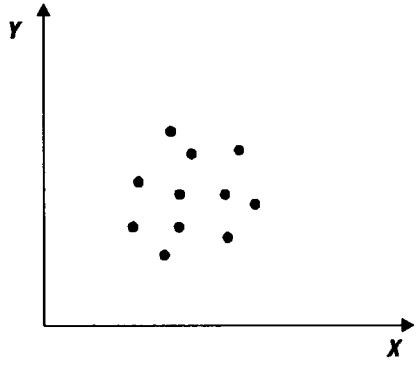
Jeśli **r = –1**, to zachodzi ścisła liniowa zależność; jest to korelacja doskonała ujemna. Przedstawia ją rysunek 4.4.

[34] Dla uproszczenia zapisu, we wzorze (4.4) pominięto symbole zmiennych ($XY$). Postępowanie takie jest uzasadnione, gdy nie będzie nieporozumienia co do jednoznacznej interpretacji wyników.



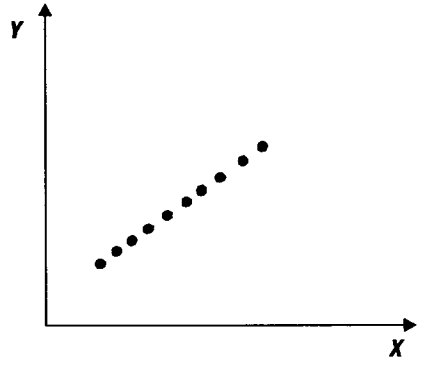

Rys. 4.4. Korelacja liniowa doskonała ujemna

Gdy $r = 0$, to wówczas brak jest korelacji liniowej, co zilustrowano na rysunku 4.5.




Rys. 4.5. Brak korelacji liniowej

$r = 1$ oznacza, że zachodzi ścisła liniowa korelacja dodatnia; jest to korelacja doskonała dodatnia, którą przedstawiono na rysunku 4.6.




Rys. 4.6. Doskonała korelacja liniowa dodatnia

**Kierunek związku** określa znak algebraiczny („+" lub „–"), zaś jego **siłę** wartość bezwzględna współczynnika korelacji. Określenie siły powiązań ułatwiają proponowane w literaturze przedziały bezwzględnych wartości | r |, którym przypisano odpowiednią interpretację. Podano je w tablicy 4.2.

**Tablica 4.2.** Interpretacja wartości bezwzględnych współczynnika korelacji liniowej Pearsona

| Wartość bezwzględna współczynnika \| r \| | Interpretacja |
|---|---|
| 0,0–0,2 | brak związku liniowego między zmiennymi |
| 0,2–0,4 | korelacja liniowa wyraźna, lecz słaba |
| 0,4–0,7 | korelacja liniowa umiarkowana |
| 0,7–0,9 | korelacja liniowa znacząca |
| 0,9–1,0 | korelacja liniowa bardzo silna |

Źródło: A. Zeliaś, *Metody statystyczne*, Warszawa 2000.

Na szczególną uwagę zasługuje równa lub bliska zeru wartość współczynnika korelacji. Musimy pamiętać, że współczynnik korelacji Pearsona jest miarą zależności liniowej. W związku z tym r = 0 może oznaczać jedynie brak korelacji liniowej między zmiennymi, a nie brak jakichkolwiek powiązań.

Współczynnik korelacji liniowej Pearsona charakteryzuje się **symetrią**. Oznacza to, że wartość tej miary jest taka sama zarówno przy badaniu zależności między zmienną $X$ i $Y$, jak i przy rozważaniu zależności między zmienną $Y$ i $X$, czyli r (Y, X) = r (Y, X). Na przykład zależność między wzrostem i masą ciała będzie taka sama, jak zależność między masą ciała i wzrostem.

### Przykład 4.1

Władze miasta opracowują plany dotyczące ochrony przeciwpożarowej. Jednym z punktów analizy jest sprawdzenie, czy występuje związek między rozmiarami zniszczeń spowodowanych pożarami a odległością najbliższej jednostki straży pożarnej od miejsca pożaru. Zgromadzono dane o dwunastu pożarach, które miały miejsce na terenie miasta. Informacje zostały podane w tablicy 4.3.

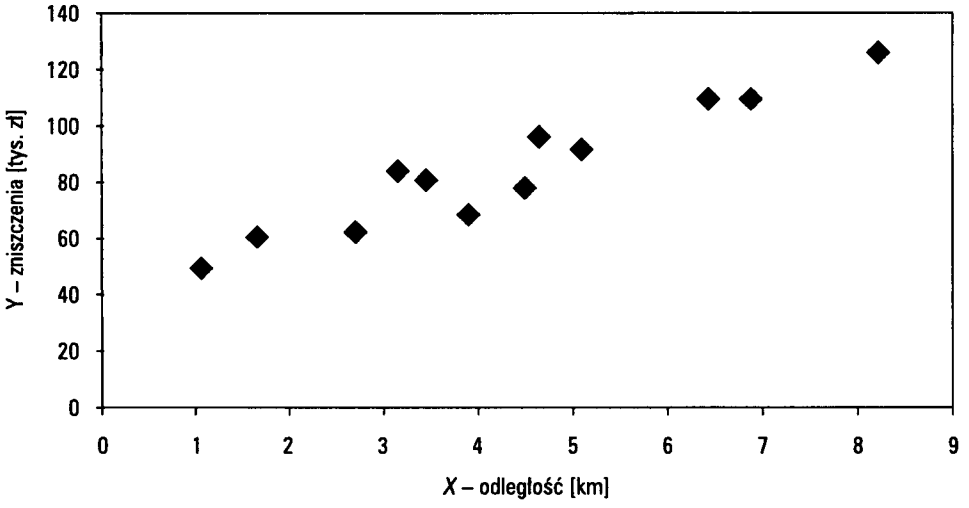
W celu sprawdzenia, czy występuje zależność między zmiennymi $X$ i $Y$, sporządzimy diagram korelacyjny (rysunek 4.7).

Na podstawie wykresu możemy przeprowadzić wstępną analizę. Widzimy, że zaznaczone punkty układają się w taki sposób, że możemy przypuszczać, iż związek między zmiennymi ma postać liniową. Ponadto wraz ze wzrostem wartości zmiennej

Tablica 4.3. Dane dotyczące pożarów w mieście

| L.p. | Odległość siedziby straży pożarnej od pożaru [km] $x_i$ | Zniszczenia [tys. zł] $y_i$ |
|---|---|---|
| 1 | 5,1 | 91,7 |
| 2 | 2,7 | 62,3 |
| 3 | 6,9 | 109,6 |
| 4 | 3,5 | 80,9 |
| 5 | 4,7 | 96,3 |
| 6 | 8,3 | 126,0 |
| 7 | 1,1 | 49,4 |
| 8 | 4,5 | 78,1 |
| 9 | 3,9 | 68,6 |
| 10 | 6,5 | 109,6 |
| 11 | 3,2 | 84,0 |
| 12 | 1,7 | 60,6 |

Źródło: dane umowne.




Rys. 4.7. Diagram korelacyjny wielkości zniszczeń względem odległości miejsca pożaru od siedziby straży pożarnej

$X$ rosną również wartości zmiennej $Y$. Możemy stwierdzić, że występuje liniowa korelacja dodatnia. Ponieważ punkty są niezbyt rozproszone, przeto można wnioskować, że związek ten jest raczej silny. Dla sformułowania bardziej precyzyjnych wniosków obliczymy wartość współczynnika korelacji liniowej Pearsona. W tablicy 4.4 przedstawiono przebieg obliczeń pomocniczych.

**Tablica 4.4.** Obliczenia pomocnicze do obliczenia współczynnika korelacji dla przykładu 4.1

| L.p. | Odległość [km] $x_i$ | Zniszczenia [tys. zł] $y_i$ | $x_i - \bar{x}$ | $(x_i - \bar{x})^2$ | $y_i - \bar{y}$ | $(y_i - \bar{y})^2$ | $(x_i - \bar{x})(y_i - \bar{y})$ |
|---|---|---|---|---|---|---|---|
| 1 | 5,1 | 91,7 | 0,79 | 0,62 | 6,97 | 48,58 | 5,51 |
| 2 | 2,7 | 62,3 | –1,61 | 2,59 | –22,43 | 503,10 | 36,11 |
| 3 | 6,9 | 109,6 | 2,59 | 6,71 | 24,82 | 616,03 | 64,28 |
| 4 | 3,5 | 80,9 | –0,86 | 0,74 | –3,88 | 15,05 | 3,34 |
| 5 | 4,7 | 96,3 | 0,34 | 0,12 | 11,52 | 132,71 | 3,92 |
| 6 | 8,3 | 126,0 | 3,94 | 15,52 | 41,27 | 1703,21 | 162,60 |
| 7 | 1,1 | 49,4 | –3,26 | 10,63 | –35,38 | 1251,74 | 115,34 |
| 8 | 4,5 | 78,1 | 0,19 | 0,04 | –6,68 | 44,62 | –1,27 |
| 9 | 3,9 | 68,6 | –0,41 | 0,17 | –16,13 | 260,18 | 6,61 |
| 10 | 6,5 | 109,6 | 2,14 | 4,58 | 24,82 | 616,03 | 53,11 |
| 11 | 3,2 | 84,0 | –1,16 | 1,35 | –0,73 | 0,53 | 0,85 |
| 12 | 1,7 | 60,6 | –2,66 | 7,08 | –24,18 | 584,67 | 64,32 |
| Razem | 51,8 | 1016,8 | × | 50,14 | × | 5776,48 | 514,72 |

Źródło: dane umowne.

- średnia arytmetyczna:

$$\bar{x} = \frac{51,8}{12} = 4,31\,[\text{kilometry}], \qquad \bar{y} = \frac{1016,8}{12} = 84,73\,[\text{tys. złotych}],$$

- odchylenia standardowe:

$$s_x = \sqrt{\frac{50,14}{12}} = 2,04\,[\text{kilometry}], \qquad s_y = \sqrt{\frac{5776,48}{12}} = 21,94\,[\text{tys. złotych}].$$

Następnym krokiem jest obliczenie kowariancji według wzoru (4.2):

$$\sum_{i=1}^{n}(x_i-\bar{x})(y_i-\bar{y})=514{,}72,$$

$$\operatorname{cov}(X,Y)=\frac{1}{12}\cdot 514{,}72=42{,}89.$$

Korzystając ze wzoru (4.3), obliczamy wartość współczynnika korelacji liniowej Pearsona:

$$r=\frac{42{,}89}{2{,}04\cdot 21{,}94}=0{,}956.$$

Na podstawie otrzymanego rezultatu możemy stwierdzić, że związek między odległością miejsca pożaru od jednostki straży pożarnej a wielkością zniszczeń spowodowanych pożarem jest bardzo silny i ma kierunek dodatni (por. rys. 4.7). Oznacza to, że jeśli wzrasta odległość siedziby straży od miejsca pożaru należy oczekiwać wzrostu zniszczeń spowodowanych wypadkiem. Zatem władze miasta powinny zoptymalizować liczbę jednostek pożarniczych w taki sposób, aby odległość od potencjalnego pożaru była jak najmniejsza. Powinno to doprowadzić do zmniejszenia wielkości strat.

W rzeczywistości rozważane zjawisko jest kształtowane przez wiele czynników. Uwarunkowania tego typu możemy badać za pomocą metod analizy korelacji cząstkowej i wielorakiej.

## 4.3. Współczynnik korelacji cząstkowej Kendalla

Analiza korelacji cząstkowej polega na badaniu związku między dwiema zmiennymi, przy wyłączeniu wpływu pozostałych zmiennych. W tym celu możemy posłużyć się **współczynnikiem korelacji cząstkowej**[35].

Rozpatrzymy najprostszą sytuację, w której rozważamy trzy zmienne oznaczone odpowiednio: $X$, $Y$, $Z$. Będziemy dążyć do ustalenia kierunku i siły powiązań między dwiema zmiennymi przy wyłączeniu oddziaływania trzeciej, czyli:

- między $Y$ a $X$ przy wyłączeniu wpływu $Z$,
- między $Y$ a $Z$ przy wyłączeniu wpływu $X$,
- między $X$ a $Z$ przy wyłączeniu wpływu $Y$.

Dla wymienionych sytuacji ustalimy odpowiednio wartości następujących współczynników korelacji cząstkowej:

$$r_{YZ.X} \qquad r_{YX.Z} \qquad r_{XZ.Y}.$$

[35] Por. G.U. Yule, M.G. Kendall, *Wstęp do teorii statystyki*, Warszawa 1966.

Poszczególne współczynniki będziemy obliczać według następujących wzorów:

$$r_{YX.Z} = \frac{r_{YX} - r_{YZ} \cdot r_{XZ}}{\sqrt{(1 - r_{YZ}^2) \cdot (1 - r_{XZ}^2)}}, \tag{4.5}$$

$$r_{YZ.X} = \frac{r_{YZ} - r_{YX} \cdot r_{XZ}}{\sqrt{(1 - r_{YX}^2) \cdot (1 - r_{XZ}^2)}}, \tag{4.6}$$

$$r_{XZ.Y} = \frac{r_{XZ} - r_{YX} \cdot r_{YZ}}{\sqrt{(1 - r_{YX}^2) \cdot (1 - r_{YZ}^2)}}. \tag{4.7}$$

Współczynniki korelacji liniowej między dwiema zmiennymi ($r_{YX}$, $r_{YZ}$, $r_{XZ}$) obliczamy jako współczynniki korelacji liniowej Pearsona zgodnie ze wzorem (4.4).

Współczynnik korelacji cząstkowej przyjmuje wartości z przedziału [–1; 1]. Interpretację siły związku, podobnie jak w przypadku współczynnika korelacji liniowej Persona, ułatwi zestawienie podane w tablicy 4.2. Za każdym razem należy zaznaczyć, że jest to związek przy ustalonym poziomie zmiennej kontrolowanej.

### Przykład 4.2

Firma doradcza otrzymała zlecenie od inwestora, który chce zainwestować kapitał w amerykański przemysł filmowy. Inwestor oczekuje odpowiedzi na pytanie, czy dochody generowane przez produkcje filmowe zależą od wysokości nakładów poniesionych na te przedsięwzięcia.

Uwzględniając specyfikę branży, specjaliści z firmy uznali, że ponoszone nakłady należy podzielić na dwie grupy kosztowe: koszty produkcji oraz koszty promocji. Określono też wielkość próby, którą stanowić będzie 20 losowo wybranych filmów. Analitycy otrzymali następujące dane w milionach dolarów (tablica 4.5).

Analiza będzie wymagała określenia współczynników korelacji cząstkowej. Na początku musimy obliczyć współczynniki korelacji liniowej Pearsona dla zmiennych $X$ (dochody), $Y$ (koszty produkcji) i $Z$ (koszty promocji) zgodnie ze wzorem (4.1). Otrzymaliśmy następujące wartości:

$$r_{YX} = 0{,}87,$$

$$r_{YZ} = 0{,}88,$$

$$r_{XZ} = 0{,}64.$$

Na podstawie uzyskanych wyników stwierdzamy, że jeśli dążymy do zwiększenia dochodów, to musimy liczyć się z poniesieniem większych kosztów produkcji oraz większych kosztów promocji. Pomiędzy rozpatrywanymi zmiennymi występuje bowiem znacząca korelacja o kierunku dodatnim. Wyróżnione grupy kosztów połączone są związkiem o kierunku dodatnim i umiarkowanej sile.

**Tablica 4.5.** Dochody i koszty z produkcji filmowych w milionach dolarów

| Film $i$ | Dochód brutto $y_i$ | Koszty produkcji $x_i$ | Koszty promocji $z_i$ |
|---|---|---|---|
| 1 | 28 | 4,2 | 1,0 |
| 2 | 35 | 6,0 | 3,0 |
| 3 | 50 | 5,5 | 6,0 |
| 4 | 20 | 3,3 | 1,0 |
| 5 | 75 | 12,5 | 11,0 |
| 6 | 60 | 9,6 | 8,0 |
| 7 | 15 | 2,5 | 0,5 |
| 8 | 45 | 10,8 | 5,0 |
| 9 | 50 | 8,4 | 3,0 |
| 10 | 34 | 6,6 | 2,0 |
| 11 | 48 | 10,7 | 1,0 |
| 12 | 82 | 11,0 | 15,0 |
| 13 | 24 | 3,5 | 4,0 |
| 14 | 50 | 6,9 | 10,0 |
| 15 | 58 | 7,8 | 9,0 |
| 16 | 63 | 10,1 | 10,0 |
| 17 | 30 | 5,0 | 1,0 |
| 18 | 37 | 7,5 | 5,0 |
| 19 | 45 | 6,4 | 8,0 |
| 20 | 72 | 10,0 | 12,0 |

Źródło: dane umowne.

Następnie, korzystając ze wzorów (4.5)–(4.7), obliczamy wartości współczynników korelacji cząstkowej:

$$r_{YX \cdot Z} = \frac{0{,}87 - 0{,}88 \cdot 0{,}64}{\sqrt{[1-(0{,}88)^2] \cdot [1-(0{,}64)^2]}} = 0{,}84.$$

Korelacja między dochodami a kosztami produkcji filmu przy ustalonych kosztach promocji jest znacząca, dodatnia.

$$r_{YZ.X} = \frac{0{,}88 - 0{,}87 \cdot 0{,}64}{\sqrt{[1-(0{,}87)^2] \cdot [1-(0{,}64)^2]}} = 0{,}86.$$

Korelacja między dochodami a kosztami promocji filmu z wyłączeniem oddziaływania kosztów produkcji jest również znacząca i dodatnia.

$$r_{XZ.Y} = \frac{0{,}64 - 0{,}87 \cdot 0{,}88}{\sqrt{[1-(0{,}87)^2] \cdot [1-(0{,}88)^2]}} = -0{,}54.$$

Korelacja między kosztami produkcji a kosztami promocji przy ustalonych dochodach z produkcji filmowej jest umiarkowana i ujemna.

W tym miejscu należy zwrócić uwagę, że współczynniki korelacji prostej i cząstkowej mogą różnić się zarówno co do znaku, jak i co do wartości bezwzględnej.

## 4.4. Współczynnik korelacji wielorakiej

**Współczynnik korelacji wielorakiej (wielokrotnej)** informuje o sile związku między wybraną zmienną, którą będziemy określać mianem objaśnianej, a całym zespołem zmiennych, nazywanych objaśniającymi[36].

W rozpatrywanym przypadku trzech zmiennych (jednej objaśnianej i dwóch objaśniających) współczynnik ten jest zdefiniowany wzorem:

$$R_{Y.XZ} = \sqrt{\frac{r_{YX}^2 + r_{YZ}^2 - 2 \cdot r_{YX} \cdot r_{YZ} \cdot r_{XZ}}{1 - r_{XZ}^2}}. \tag{4.8}$$

Współczynnik korelacji wielorakiej przyjmuje wartości z przedziału **[0; 1]**. Informuje on jedynie o sile związku, nie mówiąc nic o jej kierunku, albowiem w zbiorze zmiennych objaśniających nie wszystkie muszą mieć ten sam kierunek powiązań ze zmienną objaśnianą.

Należy zawsze brać pod uwagę, że na uzyskaną wartość współczynnika korelacji wielorakiej ma wpływ nie tylko siła zależności między zmienną objaśnianą a zmiennymi objaśniającymi, ale także siła zależności pomiędzy zmiennymi objaśniającymi. Silne skorelowanie zmiennych objaśniających może powodować sztuczne podwyższenie wartości współczynnika i utrudniać prawidłową interpretację. Z tego powodu

[36] Por.: A. Zeliaś, *Metody statystyczne*, *op. cit.*

do zbioru zmiennych objaśniających należy dobierać zmienne w taki sposób, by nie były one ze sobą skorelowane lub były skorelowane w jak najmniejszym stopniu[37].

Jeśli będziemy rozważać tylko dwie zmienne ($Y$ i $X$), to możemy przyjąć, że $r_{YZ} = 0$ oraz $r_{XZ} = 0$. Wtedy:

$$R_{Y.X} = \sqrt{r_{YX}^2} = r_{YX}.$$

Dochodzimy zatem do współczynnika korelacji liniowej Pearsona, ale w ten sposób możemy ustalić tylko siłę związku zmiennych $X$ i $Y$, a nie potrafimy określić kierunku powiązań.

Interesującą interpretację posiada kwadrat współczynnika korelacji wielorakiej ($R_{Y.XZ}^2$) noszący nazwę **współczynnika determinacji**. Wyrażony w procentach informuje, jaki procent zmienności zmiennej objaśnianej jest wyjaśniony przez zmienność zespołu zmiennych objaśniających. Innymi słowy współczynnik ten wskazuje, jaki procent zmienności zmiennej objaśnianej możemy wyrazić jako liniową funkcję zmiennych objaśniających. Dopełnienie do jedności współczynnika determinacji jest nazywane **współczynnikiem braku determinacji** lub **współczynnikiem zbieżności**, którego wartość obliczamy według wzoru:

$$\varphi_{Y.XZ}^2 = 1 - R_{Y.XZ}^2. \tag{4.9}$$

Współczynnik braku determinacji wyrażony w procentach wskazuje, jaki procent zmienności zmiennej objaśnianej nie jest wyjaśniony przez zespół przyjętych zmiennych objaśniających. Uważamy, że jest to rezultat zmienności czynników traktowanych jako przypadkowe.

### Przykład 4.3

Powróćmy do przykładu 4.2. dotyczącego produkcji filmowych. Spróbujmy określić siłę związku między dochodem i kosztami produkcji i promocji.

$$R_{y.xz} = \sqrt{\frac{(0{,}87)^2 + (0{,}88)^2 - 2 \cdot 0{,}87 \cdot 0{,}88 \cdot 0{,}64}{1 - (0{,}64)^2}} = 0{,}97.$$

Związek między dochodem a kosztami produkcji i promocji łącznie jest bardzo silny.

Współczynnik determinacji równy:

$$R_{y.xz}^2 = 0{,}9409,$$

wyrażony w procentach informuje, że zmienność dochodów generowanych przez produkcje filmowe w 94,09% jest wyjaśniona przez łączną zmienność kosztów produkcji i kosztów promocji.

---

[37] Por. też: A. Zeliaś, *Metody statystyczne*, *op. cit.*

Współczynnik braku determinacji równy:[38]

$$\varphi^2_{y.xz} = 1 - 0{,}9409 = 0{,}0591,$$

oznacza, że 5,91% zmienności dochodów z produkcji filmowej jest niewyjaśniona przez zmienność kosztów produkcji i promocji.

## 4.5. Współczynnik korelacji rang Spearmana

W badaniach związków między zjawiskami możemy mieć do czynienia z cechami o charakterze jakościowym, których nie da się bezpośrednio zmierzyć. Czasami brakuje odpowiednich danych, aby ustalić wartości współczynnika korelacji liniowej Pearsona.

Sytuacja taka może na przykład wystąpić w przypadku badania zależności między opiniami krytyków filmowych dotyczących wyświetlanych właśnie filmów albo przy analizowaniu związku między uzdolnieniami do matematyki i do historii. Podobnie nie będziemy mogli wykorzystać współczynnika korelacji Pearsona, jeśli będziemy mieć dostęp jedynie do informacji o uszeregowaniu firm pod względem zysku i wypłacanej dywidendy, bez dokładnych danych liczbowych o tych wielkościach.

W takich przypadkach możemy posłużyć się **współczynnikiem korelacji rang Spearmana**. Miara ta pozwala badać zależność w sytuacji, gdy poszczególnym zmiennym można nadać rangi. **Rangowanie** oznacza, że każdej obserwacji przyporządkowujemy numer w uporządkowanym zbiorze wartości zmiennej. W prosty sposób rangowanie możemy zilustrować na następującym przykładzie:

W szkole nauczyciele WF często zaczynają zajęcia od zbiórki polegającej na tym, że uczniowie ustawiają się według wzrostu (cechą jest tu wzrost), a następnie na komendę: „kolejno odlicz", uczniowie podają kolejne liczby odpowiadające ich miejscu w szeregu. To właśnie jest rangowanie. Na przykład wśród 10 uczniów najwyższy otrzymuje 1, a najniższy 10. W taki sam sposób będziemy porządkować zbiorowości pod względem innych rozważanych cech.

W przypadku współczynnika korelacji rang Spearmana daną zbiorowość będziemy porządkować za pomocą rangowania pod względem dwóch zmiennych, czyli każda obserwacja będzie miała przypisane równocześnie dwie rangi, odpowiednio ze względu na pierwszą ($Y$) i drugą cechę ($X$).

Współczynnik korelacji rang Spearmana obliczamy z następującego wzoru:

$$r_s = 1 - \frac{6 \cdot \sum_{i=1}^{n} (x_i - y_i)^2}{n \cdot (n^2 - 1)}. \tag{4.10}$$

[38] Podobnie jak w przypadku współczynnika korelacji Pearsona, w symbolu współczynników determinacji i braku determinacji można pominąć symbole zmiennych, jeśli nie będzie nieporozumienia przy interpretacji.

Miara ta przyjmuje wartości z przedziału [–1; 1].
Gdy:
$r_s = 1$ – korelacja jest doskonała dodatnia, występuje pełna zgodność uporządkowań pod względem badanych zmiennych,
$r_s = 0$ – brak korelacji, brak zgodności uporządkowań,
$r_s = -1$ – korelacja jest doskonała ujemna, występuje pełna przeciwstawność uporządkowań.

Pozostałe wartości $r_s$ (co do siły i kierunku powiązań) interpretujemy w taki sam sposób, jak współczynnik korelacji liniowej Pearsona (por. tablica 4.2).

### Przykład 4.4

Poproszono dwóch znawców malarstwa, aby każdy z nich wyraził opinie o 10 obrazach, uporządkowując je od najlepszego do najgorszego i nadając poszczególnym dziełom rangi od 1 (najlepszy) do 10 (najgorszy). Gdyby obaj znawcy byli zgodni w swych opiniach, to rangowanie wyglądałoby tak, jak to podano w kolumnie A tablicy 4.6. Jeżeli znawcy byliby dokładnie przeciwnego zdania, to ranking przedstawiałby się tak jak w kolumnie B.

Tablica 4.6. Wyniki rangowania obrazów przez dwóch znawców malarstwa oraz obliczenia pomocnicze

| Obraz | A. Idealna zgodność ocen | | | B. Idealna niezgodność ocen | | |
|---|---|---|---|---|---|---|
| | Znawca 1 $x_i$ | Znawca 2 $y_i$ | $(x_i - y_i)^2$ | Znawca 1 $x_i$ | Znawca 2 $y_i$ | $(x_i - y_i)^2$ |
| 1 | 5 | 5 | 0 | 5 | 6 | 1 |
| 2 | 10 | 10 | 0 | 10 | 1 | 81 |
| 3 | 1 | 1 | 0 | 1 | 10 | 81 |
| 4 | 2 | 2 | 0 | 2 | 9 | 49 |
| 5 | 6 | 6 | 0 | 6 | 5 | 1 |
| 6 | 8 | 8 | 0 | 8 | 3 | 25 |
| 7 | 3 | 3 | 0 | 3 | 8 | 25 |
| 8 | 4 | 4 | 0 | 4 | 7 | 9 |
| 9 | 7 | 7 | 0 | 7 | 4 | 9 |
| 10 | 9 | 9 | 0 | 9 | 2 | 49 |
| Razem | × | × | 0 | × | × | 330 |

Źródło: dane umowne.

Obliczmy współczynnik korelacji rang Spearmana dla dwóch sytuacji: A i B.

Sytuacja A

$$\sum_{i=1}^{n}(x_i - y_i)^2 = 0,$$

Sytuacja B

$$\sum_{i=1}^{n}(x_i - y_i)^2 = 330.$$

Podstawiając do wzoru (4.10), otrzymujemy:

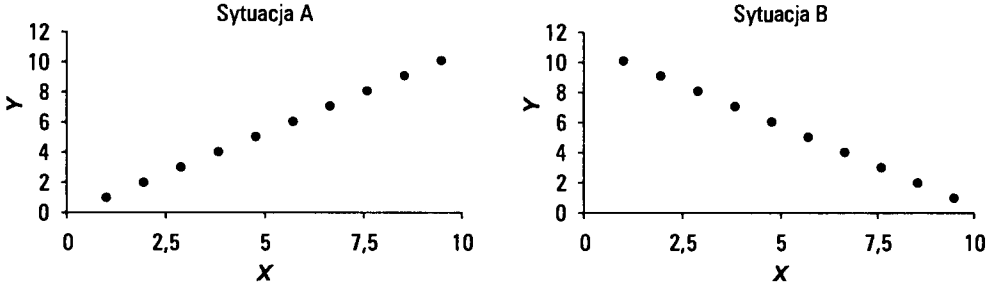
Sytuacja A

$$r_s = 1 - \frac{6 \cdot 0}{10 \cdot (10^2 - 1)} = 1 - 0 = 1,$$

Sytuacja B

$$r_s = 1 - \frac{6 \cdot 330}{10 \cdot (10^2 - 1)} = 1 - \frac{1980}{990} = 1 - 2 = -1.$$

Widzimy, że jeżeli znawcy byli zgodni co oceny obrazów, to korelacja była doskonała dodatnia. Natomiast jeżeli byli dokładnie przeciwnego zdania, to korelacja była doskonała ujemna. Spójrzmy na odpowiednie diagramy korelacyjne (rys. 4.8).




Rys. 4.8. Diagramy korelacyjne dla przykładu 4.4

W tym miejscu należy zwrócić uwagę na to, że jeżeli w badanej zbiorowości występują jednostki o takiej samej wartości cechy (np. znawca malarstwa oceni dwa obrazy jednakowo), to przypisuje się im rangę równą średniej arytmetycznej z kolejnych rang. Na przykład dwa obrazy powinny otrzymać kolejno rangę 5 i 6, ale ze względu na to, że są jednakowo oceniane przez krytyka, otrzymają rangę będącą średnią arytmetyczną z tych rang, czyli $\frac{5+6}{2} = 5{,}5$. Następny z kolei uzyska rangę 7. Rangi tego typu są określane mianem wiązanych.

### Przykład 4.5

Analityk banku inwestycyjnego otrzymał zadanie sprawdzenia stopnia powiązania zysku z wielkością aktywów przedsiębiorstw, których akcje posiada ten bank, mając do dyspozycji wyniki dwóch rankingów, które zostały przedstawione w tablicy 4.7. Rankingi są tak skonstruowane, że przedsiębiorstwo o największym zysku lub o największych aktywach otrzymało „10".

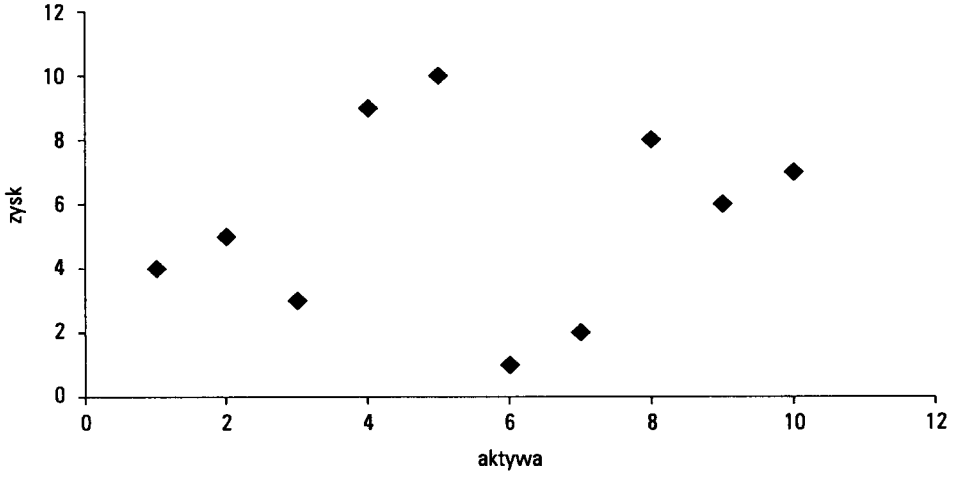
Tablica 4.7. Ranking przedsiębiorstw pod względem aktywów i zysku

| Ranking ze względu na wielkość aktywów w 2002 roku | |
|---|---|
| Nazwa | Aktywa (X) |
| Lexmark International | 1 |
| SAP | 2 |
| LM Ericsson | 3 |
| Nokia | 4 |
| Samsung | 5 |
| Alcatel | 6 |
| Motorola | 7 |
| Intel | 8 |
| Telefonica | 9 |
| Siemens | 10 |

| Ranking ze względu na poziom osiągniętego zysku w 2002 roku | |
|---|---|
| Nazwa | Zysk (Y) |
| Alcatel | 1 |
| Motorola | 2 |
| LM Ericsson | 3 |
| Lexmark International | 4 |
| SAP | 5 |
| Telefonica | 6 |
| Siemens | 7 |
| Intel | 8 |
| Nokia | 9 |
| Samsung | 10 |

Źródło: opracowanie własne na podstawie „BusinessWeek Global 1000" (2003).

Sporządzimy najpierw diagram korelacyjny, który przedstawia rysunek 4.9.




Rys. 4.9. Diagram korelacyjny aktywów i zysku

Rozrzut punktów prowadzi do przypuszczenia, że prawdopodobnie nie występuje związek między analizowanymi zmiennymi. Dla sprawdzenia obliczymy współczyn-

nik korelacji rang Spearmana. Niezbędne wyliczenia przedstawiono w tablicy 4.8, a następnie uzyskane wyniki podstawiono do wzoru (4.10).

**Tablica 4.8.** Obliczenia pomocnicze do przykładu 4.5

| Firma | $x_i$ | $y_i$ | $(x_i - y_i)$ | $(x_i - y_i)^2$ |
|---|---|---|---|---|
| Lexmark International | 1 | 4 | -3 | 9 |
| SAP | 2 | 5 | -3 | 9 |
| LM Ericsson | 3 | 3 | 0 | 0 |
| Nokia | 4 | 9 | -5 | 25 |
| Samsung | 5 | 10 | -5 | 25 |
| Alcatel | 6 | 1 | 5 | 25 |
| Motorola | 7 | 2 | 5 | 25 |
| Intel | 8 | 8 | 0 | 0 |
| Telefonica | 9 | 6 | 3 | 9 |
| Siemens | 10 | 7 | 3 | 9 |
| Razem | × | × | × | 136 |

Źródło: obliczenia własne.

$$r_s = 1 - \frac{6 \cdot 136}{10 \cdot (10^2 - 1)} = 1 - \frac{816}{990} = 1 - 0{,}824 = 0{,}175.$$

Otrzymany rezultat upoważnia do stwierdzenia, że nie występuje korelacja między aktywami a zyskiem firm, których akcje posiada bank.

## 4.6. Analiza regresji

**Analiza regresji** polega na badaniu związku między zmiennymi wyrażonymi za pomocą funkcji o określonej postaci analitycznej. Zajmiemy się związkiem między dwiema zmiennymi, który można przedstawić za pomocą funkcji liniowej. W tym celu musimy ściśle określić **zmienną objaśnianą** ($Y$) oraz **zmienną objaśniającą** ($X$). Rozróżnienie to nie było konieczne w przypadku analizy korelacji. Możemy również rozważać nieliniowe powiązania. Tym problemem nie będziemy się jednak tutaj zajmować.

Liniowy związek między zmiennymi przedstawiamy w postaci:

$$y = a + bx + \varepsilon, \tag{4.11}$$

gdzie:
a, b – parametry funkcji regresji,
y – zmienna objaśniana,
x – zmienna objaśniająca,
ε – składnik przypadkowy.

Możemy tutaj wyróżnić dwa składniki:

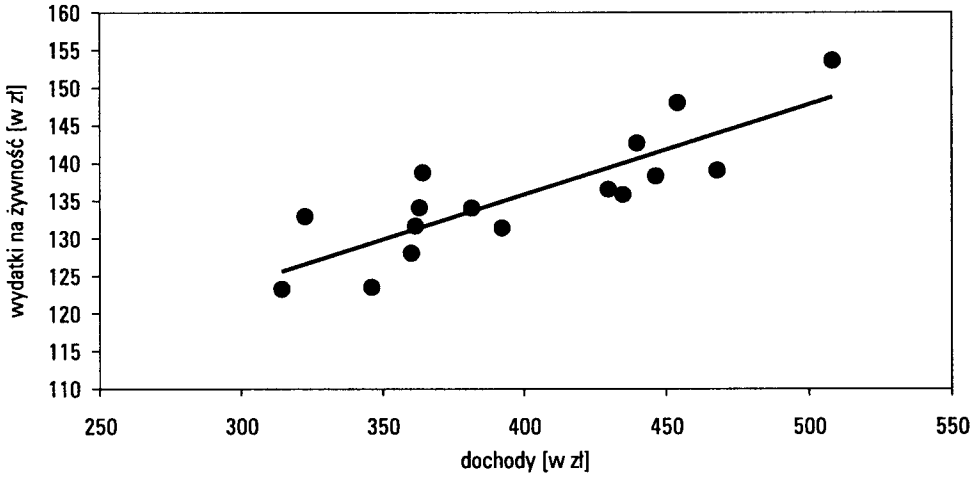
1) wyrażający związek między zmienną objaśnianą $Y$ i zmienną objaśniającą $X$, przy założeniu, że nie występują inne uwarunkowania zapisany jako:

$$\hat{y} = a + bx, \tag{4.12}$$

2) określony mianem składnika przypadkowego (ε), ujmujący oddziaływanie na zmienną $Y$ wszystkich innych uwarunkowań.

Podstawowym zadaniem analizy regresji jest znalezienie nieznanych wartości parametrów funkcji regresji[39] ($a$ i $b$). W tym celu musimy przeprowadzić obserwację interesujących nas zmiennych w badanej populacji. Jako przykład rozważymy związek między wydatkami na żywność ($Y$) oraz dochodami gospodarstw domowych ($X$).

Uzyskane rezultaty obserwacji przedstawiono na rysunku 4.10.




Rys. 4.10. Diagram korelacyjny wydatków na żywność względem wysokości dochodów w gospodarstwach domowych

[39] Por. też: S. Ostasiewicz, Z. Rusnak, U. Siedlecka, *op. cit.*, Wrocław 1999; M. Woźniak (red.), *op. cit.*; A. Zeliaś, *Metody statystyczne, op. cit.*

Zaobserwowane wartości zmiennych $X$ oraz $Y$ oznaczymy jako $(x_i, y_i)$. Każdemu gospodarstwu domowemu przypisujemy zatem zaobserwowaną wysokość wydatków na żywność $(y_i)$ oraz uzyskany dochód $(x_i)$, gdzie $i = 1, 2, ..., n$. Symbol $n$ oznacza liczebność populacji.

Układ punktów sugeruje, że badany związek może być opisany za pomocą funkcji o postaci liniowej. W związku z tym należy zastanowić się, w jaki sposób do danych empirycznych dopasować odpowiednią funkcję, której odpowiada prosta na rysunku 4.10.

Jedną z najczęściej stosowanych w tym celu procedur jest **metoda najmniejszych kwadratów**. Polega ona na wyznaczeniu takich wartości a i b, aby funkcja:

$$\psi = \sum_{i=1}^{n}(y_i - \hat{y}_t)^2 = \sum_{i=1}^{n}[(y_i - (a + bx_i)]^2 = \min. \tag{4.13}$$

Określenie minimum funkcji oznacza poszukiwanie ekstremum funkcji polegające w naszym przypadku na ustaleniu miejsc zerowych pierwszych pochodnych cząstkowych odpowiednio ze względu na a oraz na b. Otrzymujemy zatem:

$$\frac{\delta\psi}{\delta a} = -2 \cdot \sum_{i=1}^{n} y_i - na - b\sum_{i=1}^{n} x_i = 0$$

$$\frac{\delta\psi}{\delta b} = -2 \cdot \sum_{i=1}^{n} x_i y_i - a \cdot \sum_{i=1}^{n} x_i - b\sum_{i=1}^{n} x_i y_i = 0.$$

W wyniku odpowiednich przekształceń otrzymujemy **układ równań normalnych** o postaci:

$$\begin{cases} \sum_{i=1}^{n} y_i = a \cdot n + b \cdot \sum_{i=1}^{n} x_i \\ \sum_{i=1}^{n} x_i \cdot y_i = a \cdot \sum_{i=1}^{n} x_i + b \cdot \sum_{i=1}^{n} x_i^2 \end{cases} \tag{4.14}$$

Układ równań możemy rozwiązać, stosując wzory Cramera. Ustalamy w tym celu wartość następujących wyznaczników:

$$|A| = \begin{vmatrix} n & \sum_{i=1}^{n} x_i \\ \sum_{i=1}^{n} x_i & \sum_{i=1}^{n} x_i^2 \end{vmatrix} \tag{4.15}$$

$$|A_1| = \begin{vmatrix} \sum_{i=1}^{n} y_i & \sum_{i=1}^{n} x_i \\ \sum_{i=1}^{n} x_i \cdot y_i & \sum_{i=1}^{n} x_i^2 \end{vmatrix} \tag{4.16}$$

$$|A_2| = \begin{vmatrix} n & \sum_{i=1}^{n} y_i \\ \sum_{i=1}^{n} x_i & \sum_{i=1}^{n} x_i \cdot y_i \end{vmatrix} \tag{4.17}$$

$$a = \frac{|A_1|}{|A|} \tag{4.18}$$

$$b = \frac{|A_2|}{|A|} \tag{4.19}$$

Wartość parametru $a$ możemy zinterpretować jako oszacowanie wartości zmiennej $Y$ przy założeniu, że zmienna objaśniająca $X$ przyjmie wartość równą zeru. W tym miejscu należy zwrócić uwagę na to, że parametr ten nie zawsze musi mieć interpretację. $Z$ sytuacją taką możemy mieć do czynienia, gdy na przykład badamy związek między wysokością wydatków gospodarstw domowych i liczbą osób w rodzinie. Nie istnieje bowiem gospodarstwo, jeśli brak w nim jakiejkolwiek osoby.

Parametr $b$ funkcji regresji nazywamy **współczynnikiem regresji**. Przyjmuje on wartości z przedziału $(-\infty; +\infty)$. Znak współczynnika regresji wskazuje na kierunek związku między zmiennymi. Jeśli $b \in (-\infty; 0)$, to oznacza to, że jeśli wartość zmiennej objaśniającej $X$ wzrośnie o jednostkę, to wartość zmiennej objaśnianej zmniejszy się średnio o b jednostek. Jeśli natomiast $(0; +\infty)$, to jeśli wartość zmiennej objaśniającej $X$ wzrośnie o jednostkę, to wartość zmiennej objaśnianej zwiększy się średnio o $b$ jednostek. Jeżeli $b \in = 0$, to znaczy, że między zmiennymi $X$ i $Y$ związek nie występuje, bowiem zmienna $Y$ nie reaguje na zmiany zmiennej $X$.

### Przykład 4.6

Departament Kadr otrzymał od Zarządu Banku polecenie zbadania związku między generowanymi przychodami a wielkością zatrudnienia w placówkach detalicznych banku.

Dokonano losowego wyboru 10 oddziałów, otrzymując dane przedstawione w tablicy 4.9. Należy zbadać związek między tymi zmiennymi, wykorzystując metodę analizy regresji. W tym przypadku zmienną objaśnianą ($Y$) są przychody banków, a zmienną objaśniająca ($X$) wielkość zatrudnienia.

W pierwszej kolejności sporządzamy diagram korelacyjny (rysunek 4.11).

Z układu punktów możemy odczytać, że związek między zmiennymi ma charakter liniowy, dzięki czemu możemy go opisać przez liniową funkcję regresji.

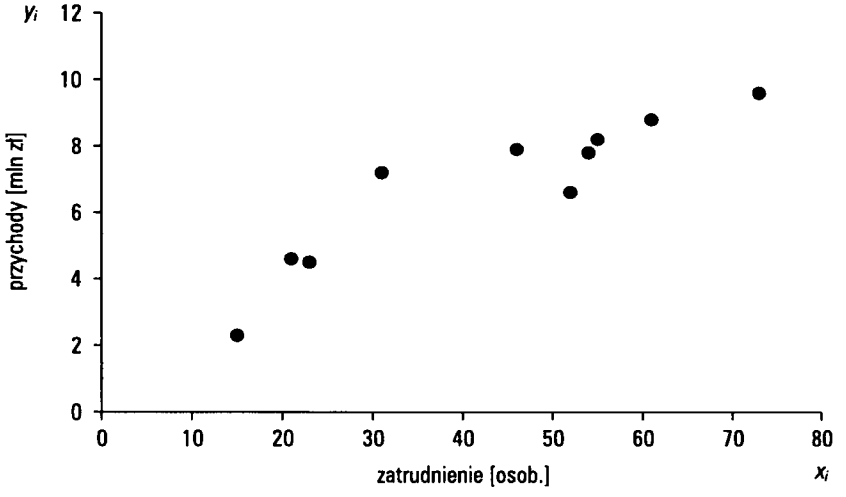
Po wykonaniu niezbędnych obliczeń otrzymujemy następujący układ równań normalnych:

$$\begin{cases} 67{,}5 = 10a + 431b \\ 3274{,}2 = 431a + 21967b \end{cases}$$

**Tablica 4.9.** Zatrudnienie i wielkość przychodów w placówkach banku oraz obliczenia pomocnicze

| L.p. | Zatrudnienie $x_i$ | Przychody $y_i$ | $x_i^2$ | $x_i \cdot y_i$ |
|---|---|---|---|---|
| 1 | 15 | 2,3 | 225 | 34,5 |
| 2 | 21 | 4,6 | 441 | 96,6 |
| 3 | 52 | 6,6 | 2704 | 343,2 |
| 4 | 31 | 7,2 | 961 | 223,2 |
| 5 | 46 | 7,9 | 2116 | 363,4 |
| 6 | 23 | 4,5 | 529 | 103,5 |
| 7 | 61 | 8,8 | 3721 | 536,8 |
| 8 | 55 | 8,2 | 3025 | 451,0 |
| 9 | 73 | 9,6 | 5329 | 700,8 |
| 10 | 54 | 7,8 | 2916 | 421,2 |
| Razem | 431 | 67,5 | 21967 | 3274,2 |

Źródło: dane umowne.




Rys. 4.11. Diagram korelacyjny przychodów względem zatrudnienia w placówkach banku

Ustalamy wartość odpowiednich wyznaczników, posługując się wzorami (4.15)–(4.17):

$$|A| = \begin{vmatrix} 10 & 431 \\ 431 & 21967 \end{vmatrix} = 219670 - 185761 = 33909,$$

$$|A_1| = \begin{vmatrix} 67{,}5 & 431 \\ 3274{,}2 & 21967 \end{vmatrix} = 1482773 - 1411180 = 71592{,}3,$$

$$|A_2| = \begin{vmatrix} 10 & 67{,}5 \\ 431 & 3274{,}2 \end{vmatrix} = 32742 - 29092{,}5 = 3649{,}5,$$

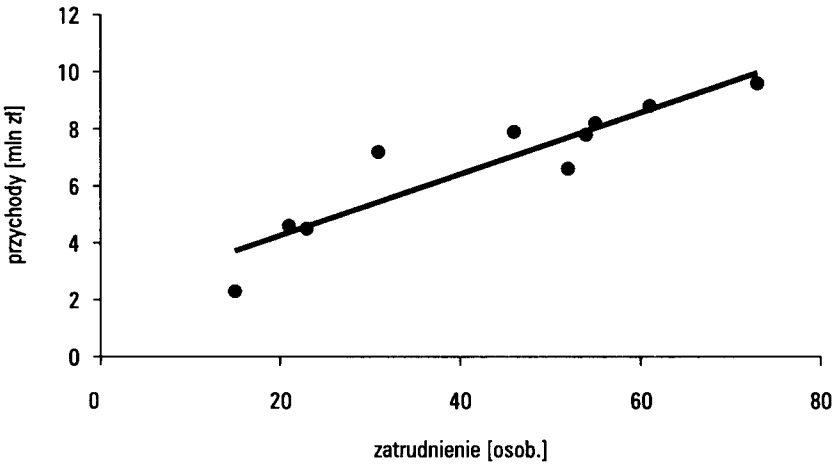
a następnie, zgodnie ze wzorami (4.18) i (4.19), ustalamy wartości parametrów:

$$a = \frac{|A_1|}{|A|} = \frac{71592{,}3}{33909} = 2{,}111, \qquad b = \frac{|A_2|}{|A|} = \frac{3649{,}5}{33909} = 0{,}108.$$

Funkcja regresji opisująca związek przychodów i zatrudnienia w placówkach banku ma postać:

$$\hat{y} = 2{,}111 + 0{,}108x.$$

Wykres tej funkcji umieszczamy na sporządzonym wcześniej diagramie korelacyjnym, który przedstawiono na rysunku 4.12. Na podstawie otrzymanych rezultatów można stwierdzić, że jeśli w placówce banku $Z$ zwiększymy zatrudnienie o jedną osobę, to należy oczekiwać, że przychody wzrosną średnio o 0,108 miliona złotych. Wyraz wolny „a" pozostaje bez interpretacji.




Rys. 4.12. Związek między wysokością przychodów i wielkością zatrudnienia w placówkach detalicznych banku

Po wyznaczeniu równania regresji przystępujemy do oceny dobroci jego dopasowania do danych empirycznych. Posługujemy się w tym celu odpowiednimi miarami.

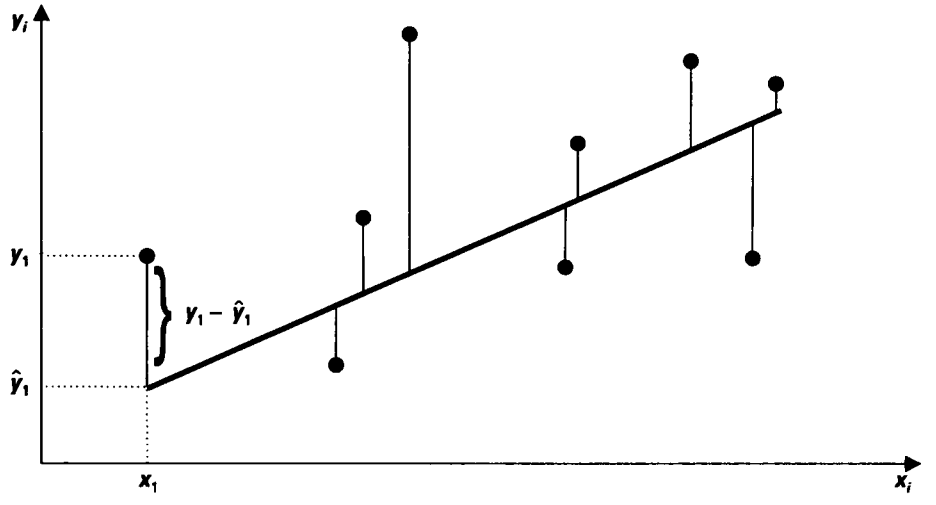
## 4.7. Miary dobroci dopasowania funkcji regresji do danych empirycznych

Dobroć dopasowania równania regresji do danych empirycznych oceniamy na podstawie następujących miar[40]:

a) wariancja składnika resztowego,
b) odchylenie standardowe składnika resztowego (średni błąd szacunku),
c) współczynnik zmienności przypadkowej (resztowej),
d) współczynnik braku determinacji (zbieżności),
e) współczynnik determinacji.

Przedstawimy je teraz kolejno.

Miary te są zdefiniowane na podstawie różnicy pomiędzy empirycznymi i teoretycznym wartościami zmiennej objaśnianej ($Y$). Wartości teoretyczne są obliczane na podstawie funkcji regresji. Oznaczają one wartości, jakie przyjęłaby zmienna objaśniana, gdyby była ona funkcją tylko od zmiennej objaśniającej. Odchylenia teoretyczne i empiryczne zostały przedstawione na rysunku 4.13. Linie pomiędzy punktami empirycznymi ($x_i$, $y_i$) a odpowiadającymi im punktami teoretycznymi ($x_i$, $\hat{y}_i$) znajdującymi się na linii regresji obrazują poszczególne odchylenia, które sumarycznie ujmujemy za pomocą wariancji i odchylenia standardowego składnika resztowego.




Rys. 4.13. Odchylenia wartości empirycznych i teoretycznych obliczonych na podstawie funkcji regresji

[40] Por. też: S. Ostasiewicz, Z. Rusnak, U. Siedlecka, *op. cit.*; M. Woźniak (red.), *op. cit.*, A. Zeliaś, *Metody statystyczne*, *op. cit.*; W. Starzyńska, *Statystyka praktyczna*, Warszawa 2000.

a) **Wariancja składnika resztowego** zdefiniowana jest jako:

$$s_\varepsilon^2 = \frac{1}{n-k}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2. \tag{4.20}$$

Mierzy ona rozproszenie empirycznych wartości zmiennej $Y$ wokół funkcji regresji. Im wariancja jest większa tym gorsze jest dopasowanie funkcji regresji. W mianowniku wyrażenia (4.20) występuje liczba stopni swobody oznaczona jako $n - k$. Ustalamy ją, pomniejszając liczebność populacji ($n$) o liczbę parametrów funkcji regresji ($k$). W przypadku liniowej regresji dwóch zmiennych $k = 2$. Wariancję składnika resztowego pozostawiamy bez interpretacji. Jest to miara mianowana, a jednostką jest kwadrat jednostki w jakiej wyrażona jest zmienna objaśniana.

b) **Odchylenie standardowe składnika resztowego** zdefiniowane wzorem:

$$s_\varepsilon = \sqrt{\frac{\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}{n-k}}, \tag{4.21}$$

informuje, o ile, średnio rzecz biorąc, wartości empiryczne ($y_i$) zmiennej $Y$ odchylają się od jej wartości teoretycznych ($\hat{y}_i$) obliczonych na podstawie funkcji regresji. Im większą wartość przyjmuje ta miara, tym gorsze jest dopasowanie funkcji regresji. Odchylenie standardowe składnika resztowego jest miarą mianowaną o mianie zgodnym z mianem zmiennej objaśnianej $Y$.

c) **Współczynnik zmienności przypadkowej (resztowej)** oblicza się na podstawie następującego wzoru:

$$V_\varepsilon = \frac{S_\varepsilon}{\bar{y}} \cdot 100\%, \quad \bar{y} \neq 0. \tag{4.22}$$

Informuje on, jaki procent średniej wartości zmiennej objaśnianej stanowi odchylenie standardowe składnika resztowego. Jest miarą niemianowaną, a jej wartość podajemy w procentach.

W przypadku idealnego dopasowania funkcji regresji do danych empirycznych współczynnik zmienności resztowej osiąga wartość 0. Dlatego im lepsze dopasowanie, tym miara ta przyjmuje mniejszą wartość. W praktyce ustala się pewną wartość progową współczynnika $V_\varepsilon^*$ (np. na poziomie $V_\varepsilon^* = 30\%$), której przekroczenie powoduje odrzucenie funkcji regresji, ze względu na zbyt wysoki udział czynników przypadkowych[41]. Dążymy zatem do uzyskania:

$$V_\varepsilon < V_\varepsilon^*. \tag{4.23}$$

[41] M. Woźniak (red.), *op. cit.*; A. Zeliaś, *Metody statystyczne*, *op. cit.*

d) **Współczynnik braku determinacji (zbieżności)** wskazuje, jaka część zmienności zmiennej objaśnianej $Y$ nie jest wyjaśniona przez zmienność zmiennej objaśniającej $X$. Dla ułatwienia interpretacji wyrażamy go w procentach i wówczas dostarcza informacji o tym, jaki procent zmienności zmiennej objaśnianej $Y$ nie jest wyjaśniony przez zmienność zmiennej objaśniającej $X$. Miara jest zdefiniowana wzorem:

$$\varphi^2 = \frac{\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}{\sum_{i=1}^{n}(y_i - \bar{y})^2}. \tag{4.24}$$

Współczynnik braku determinacji (zbieżności) jest miarą unormowaną, a mianowicie:

$$0 \leq \varphi^2 \leq 1.$$

Im $\varphi^2$ jest bliższe 0, tym dopasowanie funkcji regresji jest lepsze.

e) **Współczynnik determinacji i współczynnik korelacji wielorakiej**
Współczynnik determinacji jest dopełnieniem do jedności współczynnika braku determinacji, a zatem:

$$R^2 = 1 - \varphi^2, \tag{4.25}$$

$$0 \leq R^2 \leq 1.$$

Wyrażony w procentach wskazuje, jaka część zmienności zmiennej objaśnianej $Y$ jest wyjaśniona przez zmienność zmiennej objaśniającej $X$.
Wartości współczynnika determinacji zawierają się w przedziale [0; 1]. Przy czym wartość równa 1 oznacza, że zmienność zmiennej objaśnianej jest w pełni wyjaśniona (w 100%) przez zmienność zmiennej objaśniającej.
W przypadku liniowej funkcji regresji dwóch zmiennych współczynnik determinacji jest równy kwadratowi współczynnika korelacji liniowej Pearsona[42].

### Przykład 4.7

Ocenimy teraz dopasowanie do danych empirycznych funkcji regresji z przykładu 4.6. dotyczącego zależności przychodów od zatrudnienia w placówkach banku. Każdej empirycznej wartości ($y_i$) zmiennej $Y$ przyporządkowujemy jej wartość teoretyczną ($\hat{y}_i$) obliczoną na podstawie funkcji o postaci:

$$\hat{y}_i = 2{,}111 + 0{,}108 x_i.$$

[42] Por. punkt 4.3.

Odpowiednie obliczenia pomocnicze przedstawiamy w tablicy 4.10.

**Tablica 4.10.** Obliczenia pomocnicze od uzyskanych miar dopasowania funkcji regresji do danych empirycznych

| Nr $i$ | Zatrudnienie $x_i$ | Przychody $y_i$ | $\hat{y}_i$ | $y_i - \hat{y}_i$ | $(y_i - \hat{y}_i)^2$ | $y_i - \bar{y}$ | $(y_i - \bar{y})^2$ |
|---|---|---|---|---|---|---|---|
| 1 | 15 | 2,3 | 3,73 | –1,43 | 2,05 | –4,45 | 19,80 |
| 2 | 21 | 4,6 | 4,38 | 0,22 | 0,05 | –2,15 | 4,62 |
| 3 | 52 | 6,6 | 7,73 | –1,13 | 1,27 | –0,15 | 0,02 |
| 4 | 31 | 7,2 | 5,46 | 1,74 | 3,03 | 0,45 | 0,20 |
| 5 | 46 | 7,9 | 7,08 | 0,82 | 0,67 | 1,15 | 1,32 |
| 6 | 23 | 4,5 | 4,60 | –0,10 | 0,01 | –2,25 | 5,06 |
| 7 | 61 | 8,8 | 8,70 | 0,10 | 0,01 | 2,05 | 4,20 |
| 8 | 55 | 8,2 | 8,05 | 0,15 | 0,02 | 1,45 | 2,10 |
| 9 | 73 | 9,6 | 10,00 | –0,40 | 0,16 | 2,85 | 8,12 |
| 10 | 54 | 7,8 | 7,94 | –0,14 | 0,02 | 1,05 | 1,10 |
| Razem | 431 | 67,5 | 67,66 | × | 7,29 | × | 46,57 |

Źródło: dane umowne.

W pierwszej kolejności obliczamy wartości teoretyczne $\hat{y}_i$, podstawiając do funkcji regresji liniowej kolejne wartości $x_i$. Na przykład dla $x_1 = 15$ mamy:

$$\hat{y}_1 = 2{,}111 + 0{,}108 \cdot 15 = 3{,}73.$$

Wartości poszczególnych miar zgodnie z podanymi wzorami i wynikami obliczeń podanymi w tablicy 4.9 uzyskujemy jako:

- wariancje składnika resztowego

$$s_\varepsilon^2 = \frac{7{,}29}{10-2} = 0{,}911 \text{[mln zł}^2\text{]}.$$

  Wartość tę pozostawiamy bez interpretacji.
- odchylenie standardowe składnika resztowego

$$s_\varepsilon = \sqrt{\frac{7{,}29}{10-2}} = \sqrt{0{,}911} = 0{,}954 \text{[mln zł]}.$$

  Rzeczywiste przychody placówek banku różnią się od wartości teoretycznych obliczonych na podstawie równania regresji średnio o 954 tysiące złotych. Sza-

cując przychody placówek banku na podstawie funkcji regresji, musimy liczyć się ze średnim błędem oszacowania równym 0,954 miliona złotych.

- współczynnik zmienności przypadkowej (resztowej)

$$V_\varepsilon = \frac{0{,}954}{6{,}75} \cdot 100\% = 14{,}1\%.$$

Przyjmijmy wartość progową na poziomie $V_\varepsilon^*$ 30%.
Odchylenie standardowe składnika resztowego stanowi 14,1% przeciętnego przychodu oddziału banku, co świadczy o tym, że uzyskaną funkcję regresji możemy uznać za dobrze opisującą badany związek z punktu widzenia przyjętego kryterium.

- współczynnik braku determinacji

$$\varphi^2 = \frac{7{,}29}{46{,}57} = 0{,}156,$$

$$\varphi^2 \cdot 100\% = 15{,}6\%.$$

Zmienność przychodów oddziałów banku w 15,6% jest niewyjaśniona przez zmienność wielkości zatrudnienia.

- współczynnik determinacji

$$R^2 = 1 - 0{,}156 = 0{,}844,$$

$$R^2 \cdot 100\% = 84{,}4\%.$$

Zmienność przychodów oddziałów banku jest wyjaśniona w 84,4% zmiennością zatrudnienia.

Na podstawie uzyskanych wartości miar możemy uznać dopasowanie funkcji regresji do danych empirycznych za zadowalające.

## 4.8. Uproszczona metoda najmniejszych kwadratów

Przedstawimy teraz procedurę znajdowania wartości parametrów funkcji regresji tak zwaną metodą uproszczoną. Wartości zmiennych $X$ i $Y$ zastąpimy odchyleniami od ich średnich arytmetycznych. Oznacza to równoległe przesunięcie początku układu współrzędnych prostokątnych do punktu $P(\bar{x}, \bar{y})$.

Wtedy układ równań normalnych przyjmuje postać:

$$\begin{cases} \sum_{i=1}^{n}(y_i - \bar{y}) = n \cdot a + b \cdot \sum_{i=1}^{n}(x_i - \bar{x}) \\ \sum_{i=1}^{n}(y_i - \bar{y}) \cdot (x_i - \bar{x}) = a \cdot \sum_{i=1}^{n}(x_i - \bar{x}) + b \cdot \sum_{i=1}^{n}(x_i - \bar{x})^2 \end{cases}$$

Uwzględniając własności średniej arytmetycznej (por. punkt 3.1.1), otrzymujemy:

$$\begin{cases} n \cdot a = 0 \\ \sum_{i=1}^{n} (x_i - \bar{x}) \cdot (y_i - \bar{y}) = b \cdot \sum_{i=1}^{n} (x_i - \bar{x})^2. \end{cases}$$

Z pierwszego równania układu mamy: $n \neq 0$, a zatem, $a = 0$, a na podstawie drugiego otrzymujemy:

$$b = \frac{\sum_{i=1}^{n} (x_i - \bar{x}) \cdot (y_i - \bar{y})}{\sum_{i=1}^{n} (x_i - \bar{x})^2}. \tag{4.26}$$

Funkcja regresji ma postać:

$$y - \bar{y} = b \cdot (x_i - \bar{x}). \tag{4.27}$$

Wyraz wolny równy zeru oznacza, że linia będąca wykresem tej funkcji przechodzi przez początek układu współrzędnych $(\bar{x}, \bar{y})$. Współczynnik regresji $b$ przyjmuje w obydwu układach tę samą wartość. Dokonaliśmy bowiem przesunięcia równoległego. Wartość parametru a w układzie prostokątnym $(x, y)$ możemy znaleźć z równania (4.27) w następujący sposób:

$$y = \bar{y} + b \cdot x - b \cdot \bar{x} = \bar{y} - b \cdot \bar{x} + b \cdot x,$$

$$a = \bar{y} - b \cdot \bar{x}. \tag{4.28}$$

Niezbędne obliczenia zawiera tablica 4.11.

**Tablica 4.11.** Obliczenia pomocnicze

| $y_i$ | $x_i$ | $y_i - \bar{y}$ | $x_i - \bar{x}$ | $(x_i - \bar{x}) \cdot (y_i - \bar{y})$ | $(x_i - \bar{x})^2$ |
|---|---|---|---|---|---|
| 2,3 | 15 | –4,45 | –28,10 | 125,05 | 789,61 |
| 4,6 | 21 | –2,15 | –22,10 | 47,52 | 488,41 |
| 6,6 | 52 | –0,15 | 8,90 | –1,34 | 79,21 |
| 7,2 | 31 | 0,45 | –12,10 | –5,44 | 146,41 |
| 7,9 | 46 | 1,15 | 2,90 | 3,33 | 8,41 |
| 4,5 | 23 | –2,25 | –20,10 | 45,23 | 404,01 |
| 8,8 | 61 | 2,05 | 17,90 | 36,70 | 320,41 |
| 8,2 | 55 | 1,45 | 11,90 | 17,26 | 141,61 |
| 9,6 | 73 | 2,85 | 29,90 | 85,21 | 894,01 |
| 7,8 | 54 | 1,05 | 10,90 | 11,45 | 118,81 |
| 67,5 | 431 | × | × | 364,95 | 3390,90 |

Źródło: obliczenia własne.

Powrócimy do przykładu 4.6. W celu znalezienia wartości parametrów funkcji regresji przychodów placówek banku względem wielkości zatrudnienia zastosujemy metodę najmniejszych kwadratów w wersji uproszczonej.

$$b = \frac{364,95}{3390,90} = 0,108; \qquad a = 6,75 - 0,108 \cdot 4,31 = 2,111.$$

Uzyskaliśmy zatem identyczne rezultaty jak poprzednio. Metoda najmniejszych kwadratów w wersji uproszczonej jest przydatna wówczas, gdy współzależność między zjawiskami badamy równocześnie metodą analizy regresji i korelacji. Nie musimy wówczas wykonywać dodatkowych obliczeń.

## 4.9. Uwagi o regresji wielu zmiennych

Przedstawiając metodę analizy korelacji, zwróciliśmy uwagę na badanie związków między wieloma zmiennymi. Ten sam problem pojawia się w przypadku analizy regresji. Funkcję regresji dwu zmiennych daną wzorem (4.11) możemy rozwinąć, wprowadzając wiele zmiennych objaśniających. Będziemy je oznaczać jako: $X_1, X_2, \ldots X_k$. Funkcja regresji wielu zmiennych przyjmie postać:

$$y = a_0 + a_1 x_1 + a_2 x_2 + \ldots + a_k x_k + \varepsilon, \tag{4.29}$$

gdzie:

| | |
|---|---|
| $y$ | – zmienna objaśniana, |
| $x_1, x_2, \ldots x_k$ | – zmienne objaśniające, |
| $a_0, a_1, a_2, \ldots a_k$ | – parametry funkcji regresji, |
| $a_0$ | – wyraz wolny, |
| $a_1, a_2, \ldots a_k$ | – współczynniki regresji cząstkowej, |
| $\varepsilon$ | – składnik przypadkowy. |

Wartości parametrów funkcji regresji znajdujemy metodą najmniejszych kwadratów, której kryterium przyjmuje postać:

$$\psi = \sum_{i=1}^{n} (y_i - \hat{y}_i)^2 = \sum_{i=1}^{n} [(y_i - (a_0 + a_1 x_1 + a_2 x_2 + \ldots a_k x_k)]^2 = \min. \tag{4.30}$$

Znajdując minimum tej funkcji, otrzymujemy następujący układ równań normalnych:

$$\begin{cases} \sum_{i=1}^{n} y_i = na_0 + a_1 \sum_{i=1}^{n} x_{i1} + a_2 \sum_{i=1}^{n} x_{i2} + \ldots + a_k \sum_{i=1}^{n} x_{ik} \\ \sum_{i=1}^{n} y_i = a_0 \sum_{i=1}^{n} x_{i1} + a_1 \sum_{i=1}^{n} x_{i1}^2 + a_2 \sum_{i=1}^{n} x_{i1} \cdot x_{i2} + \ldots + a_k \sum_{i=1}^{n} x_{i1} \cdot x_{ik} \\ \sum_{i=1}^{n} x_{i2} y_i = a_0 \sum_{i=1}^{n} x_{i2} + a_1 \sum_{i=1}^{n} x_{i1} \cdot x_{i2} + a_2 \sum_{i=1}^{n} x_{i2}^2 + \ldots + a_k \sum_{i=1}^{n} x_{i2} \cdot x_{ik} \\ \cdot \\ \cdot \\ \sum_{i=1}^{n} x_{ik} y_i = a_0 \sum_{i=1}^{n} x_{ik} + a_1 \sum_{i=1}^{n} x_{i1} \cdot x_{ik} + a_2 \sum_{i=1}^{n} x_2 \cdot x_{ik} + \ldots + a_k \sum_{i=1}^{n} x_{ik}^2 \end{cases} \quad (4.31)$$

Układ równań (4.31) należy rozwiązać. Najczęściej posługujemy się w tym celu metodą macierzową[43]. Obliczenia są znacznie bardziej pracochłonne niż w przypadku regresji dwóch zmiennych. Posługujemy się odpowiednimi pakietami statystycznymi, na przykład pakietem Statistica firmy Statsoft.

Po przeprowadzeniu obliczeń należy zinterpretować uzyskane wyniki. Współczynniki regresji cząstkowej przyjmują wartości z przedziału $(-\infty; +\infty)$. Znak współczynnika informuje o kierunku powiązań między zmienną $y$ i odpowiednią zmienną objaśniającą $x_j$ ($j = 1, 2, \ldots k$). Kierunek może być dodatni lub ujemny.

Jeśli $a_j < 0$, to znaczy, że jeśli wartość zmiennej objaśniającej $x_j$ zwiększy się o jednostkę, to wartość zmiennej objaśnianej $y$ zmniejszy się o $a_j$ jednostek przy założeniu, że wartości pozostałych zmiennych objaśniających będą ustalone.

Jeśli $a_j > 0$, to znaczy, że jeśli wartość zmiennej objaśniającej $x_j$ zwiększy się o jednostkę, to wartość zmiennej objaśnianej $y$ wzrośnie się o $a_j$ jednostek przy założeniu, że wartości pozostałych zmiennych objaśniających będą ustalone.

### Przykład 4.8

Powrócimy do problemu przedstawionego w przykładzie 4.2, w którym zajmowaliśmy się związkiem pomiędzy dochodami brutto wytwórni filmowej (zmienna $Y$), kosztami produkcji (zmienna $X_1$) i kosztami promocji (zmienna $X_2$). Przeprowadzimy analizę regresji zmiennej $y$ względem wyróżnionych zmiennych objaśniających. W rozważanym przypadku funkcja regresji ma postać:

$$y = a_0 + a_1 x_1 + a_2 x_2 + \varepsilon. \quad (4.32)$$

Do znalezienia wartości parametrów funkcji (4.32) wykorzystano pakiet Statistica. Otrzymano następującą funkcję regresji:

$$\hat{y}_i = 8{,}15 + 3{,}26 x_{1i} + 2{,}34 x_{2i}.$$

[43] Por. np.: M. Woźniak, (red.), *op. cit.*, A. Zeliaś, *Metody statystyczne*, *op. cit.*

Uzyskane wyniki wskazują, że zwiększając nakłady na produkcję filmów (koszty produkcji) o 1 milion dolarów, możemy oczekiwać, że dochody brutto wzrosną średnio o 3,26 miliona dolarów, jeśli nakłady (koszty) na promocję będą ustalone. Zwiększając nakłady na promocję filmów o 1 milion dolarów, oczekujemy wzrostu dochodów brutto o 2,34 miliona dolarów, przy założeniu, że nakłady na produkcję filmów będą ustalone.

Dobroć dopasowania funkcji regresji do danych empirycznych oceniamy za pomocą tych samych miar, co dla regresji dwóch zmiennych. Podajemy je wraz z wynikami otrzymanymi jako rezultat obliczeń za pomocą pakietu Statistica:

- wariancja składnika resztowego zdefiniowana jest jako:

$$s_\varepsilon^2 = \frac{1}{n-k}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2 = 25{,}2536\,[\text{mln } \$^2],$$

- odchylenie standardowe składnika resztowego zdefiniowane wzorem:

$$s_\varepsilon = \sqrt{\frac{\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}{n-k}} = 5{,}02\,[\text{ mln } \$],$$

co oznacza, że zaobserwowane dochody brutto różnią się od dochodów teoretycznych średnio o 5,02 mln $.

- współczynnik zmienności przypadkowej (resztowej):

$$V_\varepsilon = \frac{S_\varepsilon}{\bar{y}} \cdot 100\% = \frac{5{,}02}{46{,}05} \cdot 100\% = 10{,}90\%,$$

co oznacza, że odchylenia przypadkowe stanowią 10,90% średnich dochodów brutto z produkcji filmowej.

- współczynnik braku determinacji (zbieżności):

$$\varphi^2 = \frac{\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}{\sum_{i=1}^{n}(y_i - \bar{y})^2} = 0{,}06,$$

zatem 6% zmienności dochodów brutto z produkcji filmowej pozostaje niewyjaśniona przez zmienność nakładów na produkcję i promocję filmów.

- współczynnik determinacji:

$$R^2 = 1 - \varphi^2 = 0{,}94,$$

czyli 94% zmienności dochodów brutto z produkcji filmowej jest wyjaśniona przez zmienność nakładów na produkcję i promocję filmów.

- współczynnik korelacji wielorakiej:

$$R = \sqrt{R^2} = \sqrt{0{,}94} = 0{,}97,$$

oznacza, że istnieje silny związek między dochodami brutto wytwórni filmowych i nakładami na produkcję i promocję filmów.

# Analiza szeregów czasowych

Rozdział 5

## 5.1. Szereg czasowy i jego składniki

Prezentowane dotychczas statystyczne metody badania zjawisk odnosiły się do obserwacji przeprowadzanych w ustalonym okresie. Jako przykład można podać strukturę ludności Polski według stanu cywilnego w dniu 31.12.2002 roku, rozkład liczebności gospodarstw domowych według wysokości dochodu przypadającego na jedną osobę w województwie Z w 2003 roku (zob. rozdział 2 punkty 2.1.1 i 2.1.2).

Analiza dynamiki wymaga obserwowania zjawiska w różnych okresach. Powstaje wówczas **szereg czasowy**, który jest zbiorem wartości zmiennej uporządkowanych według czasu[44]. Przykładem takiego szeregu może być produkt krajowy brutto (PKB) w Polsce w latach 1995–2005, konsumpcja artykułów spożywczych w gospodarstwach domowych w województwie małopolskim w latach 2000–2002, sprzedaż kosmetyków w Krakowie w latach 1998–2003, liczba absolwentów Akademii Ekonomicznej w Krakowie w latach 1990–2000, stopa bezrobocia w Polsce w latach 1990–2004.

Analiza dynamik z jednej strony pozwala poznać przebieg interesującego nas zjawiska w przeszłości (na podstawie danych historycznych). Rezultaty te mogą stanowić podstawę do wnioskowania o przyszłości, czyli do tworzenia prognoz[45].

W szeregu czasowym wyodrębniamy następujące składniki:

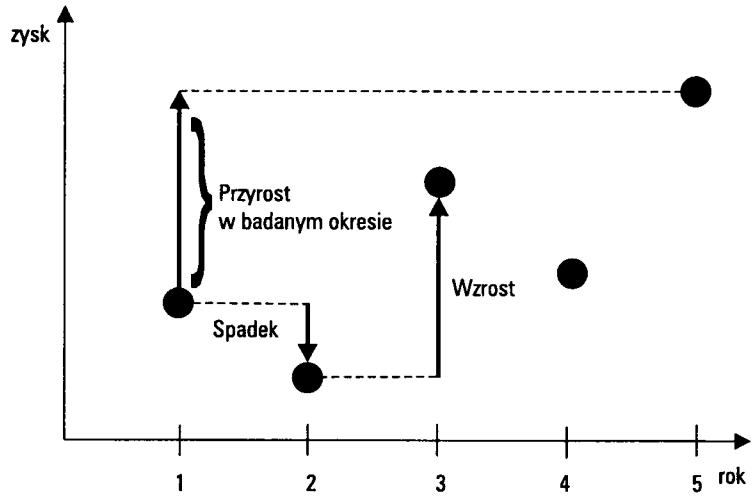
- tendencję rozwojową (trend),
- wahania okresowe,
- zmiany zachodzące pod wpływem czynników przypadkowych.

**Trend** (tendencja rozwojowa) reprezentuje trwałe zmiany zachodzące sposób powolny, ciągły i regularny w ogólnym poziomie zjawiska w badanym okresie[46]. Na rysunku 5.1 przedstawiono tendencję rozwojową zysku pewnego przedsiębiorstwa

---

[44] A. Aczel, *Statystyka w zarządzaniu. Pełny wykład*, Warszawa 2000.
[45] A. Zeliaś, *Teoria prognozy*, Warszawa 1997.
[46] A. Zeliaś, *Metody statystyczne*, *op. cit.*

w pięciu okresach (tutaj w latach). Porównując poszczególne wartości, stwierdzamy, że zysk ten na przemian wzrasta i maleje w porównaniu do poprzedniego roku.



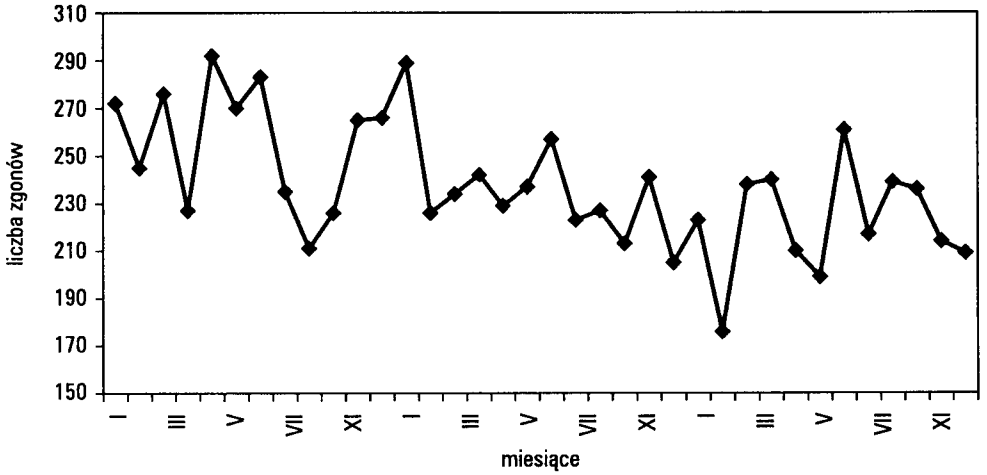

Rys. 5.1. Zysk przedsiębiorstwa w pięciu kolejnych latach działalności

Zmiany te są rezultatem oddziaływania różnorodnych czynników kształtujących warunki gospodarowania rozważanego przedsiębiorstwa. Stwierdzamy, że pomimo pojawiających się wzrostów i spadków zysk jednak wzrastał w badanym okresie. Powiemy, że charakteryzował się tendencją wzrostową lub rosnącym trendem.

Niektóre zjawiska podlegają **wahaniom okresowym**, które powstają na skutek czynników oddziaływujących w pewnych odstępach czasu. Powodują one regularne odchylenia od ogólnej tendencji rozwojowej. Takimi czynnikami są między innymi zmiany pogodowe związane z porą roku. Wpływają one na konsumpcję wielu dóbr i usług, np. węgla opalowego (największe zużycie w okresach zimowych), lodów, olejku do opalania, usług turystycznych itp. Tego rodzaju periodyczność nazywamy sezonowością. Wahania okresowe mogą być także rezultatem obyczajów. Przykładem takich zjawisk jest zawieranie małżeństw (por. rys. 2.12, s. 44). Na rysunku 5.2 przedstawiono jako przykład liczbę zgonów niemowląt w poszczególnych miesiącach okresu 2000–2002.

Wahania okresowe możemy podzielić na:

- **wahania krótkookresowe** powtarzające się w obrębie miesięcy, tygodni lub dni, np. stale powtarzający się wzrost sprzedaży w hipermarketach w czasie weekendu,
- **wahania sezonowe** występujące regularnie w odstępach rocznych, np. związane z czynnikami klimatycznymi lub zwyczajami (święta religijne),
- **wahania cykliczne** pojawiające się w okresach dłuższych niż rok, np. wahania koniunktury ekonomicznej.




Rys. 5.2. Liczba zgonów niemowląt według miesięcy w latach 2000–2003

**Wahania przypadkowe** zakłócają przebieg zjawiska w poszczególnych okresach. Są one spowodowane czynnikami, których nie jesteśmy w stanie przewidzieć. Wahania przypadkowe są rezultatem bardzo wielu czynników o różnokierunkowych oddziaływaniach. Zaliczamy je do tak zwanych przyczyn ubocznych (por. rozdział 1 punkt 1.1).

Celem statystycznej analizy dynamiki zjawisk jest wyodrębnienie i oszacowanie poszczególnych składników szeregu czasowego, a mianowicie trendu, wahań okresowych oraz odchyleń przypadkowych. W tym celu posługujemy się odpowiednimi metodami, które przedstawiamy w niniejszym rozdziale.

## 5.2. Indywidualne miary dynamiki

Dynamikę zjawisk możemy mierzyć za pomocą przyrostów (absolutnych i względnych), indywidualnych indeksów dynamiki oraz indeksów agregatowych.

### 5.2.1. Przyrosty

Najprostszym sposobem zbadania zmian poziomu rozpatrywanej zmiennej w czasie jest obliczenie różnicy między wartością tej zmiennej w okresie badanym i jej wartością w okresie przyjętym za podstawę porównań. Różnicę tę nazywamy przyrostem absolutnym. Przyrosty te mogą mieć podstawę stałą lub zmienną.

**Przyrosty o stałej podstawie (jednopodstawowe)** uzyskujemy, gdy wartości zmiennej zaobserwowane w kolejnych okresach porównujemy do wartości z ustalonego okresu bazowego (podstawowego) i wtedy mamy:

$$\Delta_{t/k} = y_t - y_k, \qquad t = 1,2,\ldots,n, \tag{5.1}$$

gdzie:
$k$ – okres bazowy,
$y_t$ – wartość zmiennej w okresie $t$,
$y_k$ – wartość zmiennej w okresie stanowiącym podstawę porównań.

**Przyrosty o podstawie zmiennej (łańcuchowe)** uzyskujemy, gdy jako bazowy przyjmiemy okres bezpośrednio poprzedzający okres badany. Przyrosty te definiujemy jako:

$$\Delta_{t/t-1} = y_t - y_{t-1}. \tag{5.2}$$

### Przykład 5.1

Obserwujemy wysokość miesięcznych rachunków telefonicznych pewnego gospodarstwa domowego w I kwartale 2003 roku. Odpowiednie informacje podano w tablicy 5.1.

**Tablica 5.1.** Wysokość miesięcznych rachunków telefonicznych w I kwartale 2003 roku

| Wyszczególnienie | Styczeń | Luty | Marzec |
|---|---|---|---|
| | w złotych | | |
| Wysokość rachunku | 150 | 122 | 164 |

Źródło: dane umowne.

Obliczmy przyrosty absolutne o stałej podstawie (wzór 5.1) dla lutego i marca. Za okres bazowy przyjmiemy styczeń 2003 roku:

$$\Delta_{\text{luty/styczeń}} = y_{\text{luty}} - y_{\text{styczeń}} = 122 - 150 = -28 \text{ złotych},$$

$$\Delta_{\text{marzec/styczeń}} = y_{\text{marzec}} - y_{\text{styczeń}} = 164 - 150 = 14 \text{ złotych}.$$

Wysokość rachunku w lutym była niższa od rachunku ze stycznia o 28 złotych. Rachunek z marca był o 14 złotych wyższy niż w styczniu.

Przyrost absolutny łańcuchowy uzyskujemy według wzoru (5.2):

$$\Delta_{\text{luty/styczeń}} = y_{\text{luty}} - y_{\text{styczeń}} = 122 - 150 = -28 \text{ złotych}.$$

W tym przypadku wartości obydwu miar są jednakowe, ponieważ styczeń jest okresem poprzedzającym luty, czyli okres bazowy pozostaje w tym przypadku bez zmian.

$$\Delta_{\text{marzec/luty}} = y_{\text{marzec}} - y_{\text{luty}} = 164 - 122 = 42 \text{ [zł]}.$$

Rachunek w marcu w porównaniu z poprzednim miesiącem był o 42 złote wyższy.

Przyrosty absolutne są miarami **mianowanymi** i dlatego nie pozwalają na porównania zjawisk scharakteryzowanych zmiennymi wyrażanymi w różnych jednostkach. Ponadto są uzależnione od skali. W celu wyeliminowania tych niedogodności możemy użyć **przyrostów względnych**, które będą wyrażać procentową zmianę wartości obserwowanej zmiennej. Miarę tę często określamy **tempem zmian**. Jest ona często wykorzystywana w analizie finansowej. Jako przykład zastosowania przyrostów względnych można podać badanie sprawozdań finansowych, kiedy to rozpatrujemy zmiany poszczególnych składników bilansu, rachunku zysków i strat oraz przepływów pieniężnych w wybranym okresie.

Podobnie jak w przypadku przyrostów absolutnych wyróżniamy przyrosty względne:

- o stałej podstawie (jednopodstawowe):

$$\delta_{t/k} = \frac{y_t - y_k}{y_k} \cdot 100\%, \tag{5.3}$$

- o zmiennej podstawie (łańcuchowe):

$$\delta_{t/t-1} = \frac{y_t - y_{t-1}}{y_{t-1}} \cdot 100\%, \tag{5.4}$$

$$\delta_{t/t-1} = \left( \frac{y_t}{y_{t-1}} - 1 \right) \cdot 100\%. \tag{5.5}$$

### Przykład 5.2

Powróćmy do przykładu 5.1. dotyczącego rachunków telefonicznych. Obliczymy tempo zmian wysokości opłat za korzystanie z telefonu. Styczeń przyjmiemy za stałą podstawę porównań (okres bazowy).

$$\delta_{\text{luty/styczeń}} = \frac{y_{\text{luty}} - y_{\text{styczeń}}}{y_{\text{styczeń}}} \cdot 100\% = \frac{122 - 150}{150} \cdot 100\% = -18{,}7\%.$$

Kwota rachunku telefonicznego w lutym obniżyła się w stosunku do stycznia o 18,7%.

$$\delta_{\text{marzec/styczeń}} = \frac{y_{\text{marzec}} - y_{\text{styczeń}}}{y_{\text{styczeń}}} \cdot 100\% = \frac{164 - 150}{150} \cdot 100\% = 9{,}3\%.$$

Wysokość opłat telefonicznych w marcu wzrosła o 9,3% w porównaniu do stycznia.

Scharakteryzujemy teraz dynamikę wydatków poniesionych w związku z rozmowami telefonicznymi za pomocą przyrostów względnych o podstawie zmiennej (łańcuchowych). Wartość dla lutego będzie taka sama jak w przypadku przyrostów o stałej podstawie. Dla marca otrzymujemy:

$$\delta_{\text{marzec/luty}} = \frac{y_{\text{marzec}} - y_{\text{luty}}}{y_{\text{luty}}} \cdot 100\% = \frac{164-122}{122} \cdot 100\% = 34{,}4\%.$$

Wysokość rachunku telefonicznego wzrosła w marcu w stosunku do poprzedniego miesiąca o 34,4%.

## 5.2.2. Indywidualne indeksy dynamiki

**Indywidualne indeksy dynamiki** określane mianem prostych są najczęściej wykorzystywanymi miarami przedstawiającymi zmiany badanego zjawisku w czasie. Indeks indywidualny (prosty) jest stosunkiem dwóch wartości zmiennej wyrażonym w procentach. Ze względu na przyjmowany okres bazowy możemy wyróżnić indeksy dynamiki:

- o stałej podstawie (jednopodstawowe), gdy przyjmujemy jeden, stały okres bazowy, do którego wartości odnosimy poziom badanej zmiennej w pozostałych okresach, liczony ze wzoru:

$$i_{t/k} = \frac{y_t}{y_k} \cdot 100\%, \tag{5.6}$$

- o zmiennej podstawie (łańcuchowe), jeśli podstawą do obliczeń jest poziom zjawiska z poprzedniego okresu.

$$i_{t/t-1} = \frac{y_t}{y_{t-1}} \cdot 100\%. \tag{5.7}$$

**Przykład 5.3**

Na podstawie danych z przykładu 5.1. obliczamy indeks dynamiki o stałej podstawie (wzór (5.6)). Za okres bazowy przyjmiemy styczeń:

$$i_{\text{luty/styczeń}} = \frac{y_{\text{luty}}}{y_{\text{styczeń}}} \cdot 100\% = \frac{122}{150} \cdot 100\% = 81{,}3\%,$$

$$i_{\text{marzec/styczeń}} = \frac{y_{\text{marzec}}}{y_{\text{styczeń}}} \cdot 100\% = \frac{164}{150} \cdot 100\% = 109{,}3\%.$$

Wysokość rachunku telefonicznego z lutego stanowiła 81,3%, a z marca 109,3%, jego wartości ze stycznia, a więc rachunek z lutego był 18,7% niższy, a z marca o 9,3% wyższy niż rachunek ze stycznia.

Otrzymane wyniki możemy interpretować *tylko w stosunku do roku bazowego*. Nie możemy zatem powiedzieć, że rachunek z marca w stosunku do lutego wzrósł o 109,3% minus 81,3%, czyli o 28%, gdyż jest to nieprawda, o czym przekonamy się, posługując się indeksem o zmiennej podstawie zdefiniowanym wzorem (5.7):

- dla lutego wartość indeksu będzie taka sama jak w przypadku indeksu o stałej stopie (ten sam okres bazowy).
- dla marca:

$$i_{\text{marzec/luty}} = \frac{y_{\text{marzec}}}{y_{\text{luty}}} \cdot 100\% = \frac{164}{122} \cdot 100\% = 134{,}4\%.$$

Rachunek z marca stanowił 134,4% wartości rachunku z okresu poprzedniego – wzrósł o 34,4% w porównaniu z lutym.

Indeksy o zmiennej podstawie możemy otrzymać z indeksów o stałej podstawie w rezultacie następującego przekształcenia[47]:

$$i_{t/t-1} = \frac{y_t}{y_k} : \frac{y_{t-1}}{y_k} = \frac{y_t}{y_{t-1}}. \tag{5.8}$$

Zauważamy, że z prostego indeksu dynamiki możemy w łatwy sposób otrzymać przyrost względny (tempo zmian), stosując formuły:

- dla stałej podstawy:

$$\delta_{t/k} = i_{t/k} - 100\%, \tag{5.9}$$

- dla zmiennej podstawy:

$$\delta_{t/t-1} = i_{t/t-1} - 100\%. \tag{5.10}$$

Na postawie indeksów łańcuchowych obliczamy średnioroczne tempo zmian jako ich średnią geometryczną[48].

$$\bar{i} = \sqrt[n-1]{\prod_{t=1}^{n-1} i_{t/t-1}} = \sqrt[n-1]{\frac{y_{n-1}}{y_0}}. \tag{5.11}$$

**Przykład 5.4**

Należy obliczyć średnioroczne tempo zmian wysokości opłat za telefon pewnego gospodarstwa domowego w I kwartale 2003 roku (por. tab. 5.1).

$$\bar{i} = \sqrt[2]{\frac{164}{150}} = \sqrt[2]{1{,}0933} = 1{,}046.$$

Oznacza to, że średni wzrost opłat za telefon w 2003 roku wynosił 4,6%.

Jeśli stopień pierwiastka jest wyższy, gdy analizujemy dynamikę w dłuższym okresie, to dla obliczenia średniego tempa logarytmujemy stronami formułę (5.11).

---

[47] M. Woźniak (red.), *Statystyka ogólna*, *op. cit.*

[48] Por. np.: J. Jóźwiak, J. Podgórski, *Statystyka od podstaw*, Warszawa 1992.

**Przykład 5.5**

Obliczamy średnie tempo zmian liczby ludności w Polsce w latach 1995–2003. W 1995 roku ludność Polski wynosiła 38609 tysięcy, a w 2002 roku 38232 tysiące[49]. Średnie tempo zmian wynosiło:

$$\bar{i} = \sqrt[7]{\frac{38232}{38609}} = \sqrt[7]{0{,}9902},$$

$$\log \bar{i} = \frac{1}{7} \cdot \log 0{,}9902 = \frac{1}{7} \cdot (-0{,}00426) = -0{,}00061,$$

$$\bar{i} = 0{,}9986,$$

$$(i-1) \cdot 100\% = (0{,}9986 - 1) \cdot 100\% = -0{,}14\%.$$

W latach 1995–2002 liczba ludności w Polsce zmniejszała się średnio o 0,14% na rok.

## 5.2.3. Indeksy agregatowe

**Indeksy agregatowe** mierzą względne zmiany zjawiska scharakteryzowanego za pomocą zmiennych, w których można wyróżnić składniki stanowiące pewne struktury. Na przykład badając dynamikę akcji obserwowaną na giełdzie, musimy wziąć pod uwagę, że dynamika ta może być wywołana zarówno zmianą cen akcji, jak i ich liczby. Mamy tu zatem do czynienia z agregatem, którego składniki musimy uwzględnić w badaniach dynamiki.

Klasycznym obszarem zastosowań indeksów agregatowych są badania dynamiki cen, masy towarowej, wydajności pracy. Znajdują one również zastosowanie w badaniach demograficznych w procedurze określanej mianem standaryzacji współczynników demograficznych[50].

Przedstawimy teraz metodę badania dynamiki cen i masy towarowej. W tym zakresie najczęściej stosuje się indeksy cen i masy towarowej Laspeyresa lub Paaschego.

- **Indeks cen Laspeyresa** jest zdefiniowany jako:

$$I_p^L = \frac{\sum_{j=1}^{m} q_{kj} \cdot p_{tj}}{\sum_{j=1}^{m} q_{kj} \cdot p_{kj}} \cdot 100\%, \tag{5.12}$$

[49] Źródło: www.stat.gov.pl

[50] Por. np.: J. Z. Holzer, *Demografia*, Warszawa 2004; J. Kurkiewicz, *Podstawowe metody analizy demograficznej*, Warszawa 1992.

gdzie:
$m$ – liczba składowych agregatu,
$q_{tj}$ – ilość $j$-tego składnika agregatu w okresie badanym $t$,
$q_{kj}$ – ilość $j$-tego składnika agregatu w okresie bazowym $k$,
$p_{tj}$ – cena $j$-tego składnika agregatu w okresie badanym $t$,
$p_{kj}$ – cena $j$-tego składnika agregatu w okresie bazowym $k$.

W symbolu $I_p^L$ indeks górny $L$ oznacza, że jest to indeks Laspeyresa, a dolny $p$ oznacza, że indeks ten odnosi się do zmian cen (ang. *price*).

W tym przypadku ceny dóbr w obydwu okresach (badanym i bazowym) są ważone ilością dóbr z okresu bazowego. Zatem nie musimy znać ilości dóbr w roku badanym. Posługujemy się jedynie ilościami z roku bazowego. Indeks ten ma zastosowanie w sytuacjach, gdy ilości dóbr nie zmieniają się gwałtownie z okresu na okres. Gdy występują duże różnice, indeks nie oddaje zbyt dobrze efektu zmian poziomu cen[51].

### Przykład 5.6

Producent jogurtów wytwarza swoje produkty z czterech podstawowych surowców: mleka, cukru oraz owoców (bananów i jabłek). Ceny tych dóbr (w złotych) oraz ilości (zapotrzebowanie miesięczne) w roku 1995, 1998 i 2000 podano w tablicy 5.2. Zarządzający przedsiębiorstwem interesują się dynamiką cen. Zbadamy ją za pomocą indeksu cen Laspeyresa dla surowców zużywanych w procesie produkcyjnym. Jako bazowy przyjmujemy 1995 rok.

**Tablica 5.2.** Dynamika ilości i cen surowców do produkcji jogurtów w latach 1998–2000

| Surowiec | Jednostka miary | 1995 | | 1998 | | 2000 | |
|---|---|---|---|---|---|---|---|
| | | liczba $q_{1995j}$ | cena $p_{1995j}$ | liczba $q_{1998j}$ | cena $p_{1998j}$ | liczba $q_{2000j}$ | cena $p_{2000j}$ |
| Mleko | tys. litrów | 500 | 1730 | 750 | 2080 | 1000 | 2200 |
| Cukier | tona | 50 | 1900 | 51 | 2120 | 70 | 2670 |
| Jabłko | tona | 17 | 1590 | 16,5 | 1870 | 25 | 1860 |
| Banan | tona | 15 | 2380 | 14 | 2370 | 17 | 3250 |

Źródło: *Ceny – Mały Rocznik Statystyczny*, GUS, Warszawa 2000 i 2003, ilości – dane umowne.

Zgodnie ze wzorem (5.12) obliczamy sumy iloczynów ilości poszczególnych surowców z roku bazowego (1995) i ich cen z badanego okresu:

[51] Por. np.: A. Aczel, *Statystyka w zarządzaniu. Pełny wykład*, *op. cit.*

Wartość mianownika będzie stała równa:

$$\sum_{j=1}^{m} q_{1995j} \cdot p_{1995j} = 500 \cdot 1730 + 50 \cdot 1900 + 17 \cdot 1590 + 15 \cdot 2380 = 102273.$$

Wartość licznika dla 1998 roku otrzymujemy jako:

$$\sum_{j=1}^{m} q_{1995j} \cdot p_{1998j} = 500 \cdot 2080 + 50 \cdot 2120 + 17 \cdot 1870 + 15 \cdot 2370 = 121334.$$

Licznik dla 2000 roku:

$$\sum_{j=1}^{m} q_{1995j} \cdot p_{2000j} = 500 \cdot 2200 + 50 \cdot 2670 + 17 \cdot 1860 + 15 \cdot 3250 = 131387.$$

Podstawiając do wzoru (5.12), otrzymujemy następujące wartości indeksu:
dla 1995: $I_p^L = 100\%$,
dla 1998: $I_p^L = 118{,}64\%$,
dla 2000: $I_p^L = 128{,}47\%$.

W rezultacie przeprowadzonych obliczeń stwierdzamy, że ceny agregatu składającego się z czterech surowców do produkcji jogurtów wzrosły w 1998 roku o 18,64%, a w 2000 roku o 28,47% w stosunku do roku bazowego, przy założeniu stałej wielkości zapotrzebowania przedsiębiorstwa na poziomie z 1995 roku.

Przykładem zastosowania indeksu cen Laspeyresa jest amerykański Indeks Cen Konsumpcyjnych (CPI – *consumer price index*) publikowany przez U.S. Bureau of Labor Statistics[52]. Jest on obliczany na podstawie koszyka dóbr składającego się z kilkuset artykułów używanych przez amerykańskie gospodarstwa domowe. Indeks ten pozwala ustalić siłę nabywczą dochodów gospodarstw domowych. Podobną miarą jest *Wskaźnik cen towarów i usług konsumpcyjnych* stosowany dla Polski przez Główny Urząd Statystyczny[53].

Jeśli w porównywanych okresach występują znaczne wahaniach ilości dóbr, to bardziej poprawne będzie zastosowanie **indeksu cen Paaschego**. Wagami w tym przypadku są bowiem ilości z okresu badanego. Miarę tę definiujemy za pomocą następującego wzoru:

$$I_p^P = \frac{\sum_{j=1}^{m} q_{tj} \cdot p_{tj}}{\sum_{j=1}^{m} q_{tj} \cdot p_{kj}} \cdot 100\%. \qquad (5.13)$$

[52] Por. www.bls.gov
[53] Por. www.stat.gov.pl

Indeks cen Paaschego ma również pewne ograniczenia. Pierwsze z związane jest z koniecznością każdorazowego pozyskiwania danych o ilości w badanych okresach, co często jest bardzo kosztowne, a czasem nawet niemożliwe. Wymaga na przykład przeprowadzania corocznych badań ilości dóbr konsumowanych przez gospodarstwa domowe. Drugi problem powstaje przy porównaniach indeksów cen Paaschego dla dwóch różnych okresów, z których żaden nie jest okresem bazowym. Wówczas nie wiemy, czy różnica między indeksami wynika ze zmian cen, czy też z różnych ilości dóbr w poszczególnych okresach.

**Przykład 5.7**

Obliczmy teraz wartość indeksu cen Paaschego w celu ustalenia dynamiki cen składników jogurtów (przykład 5.6). Dla roku bazowego wartość indeksu wynosi 100%. Obliczamy licznik i mianownik dla 1998 roku:

$$\sum_{j=1}^{m} q_{1998\,j} \cdot p_{1998\,j} = 750 \cdot 2080 + 51 \cdot 2120 + 16{,}5 \cdot 1870 + 14 \cdot 2370 = 1732155.$$

$$\sum_{j=1}^{m} q_{1998\,j} \cdot p_{1995\,j} = 750 \cdot 1730 + 51 \cdot 1900 + 16{,}5 \cdot 1590 + 14 \cdot 2380 = 1453955.$$

Dla 2000 roku:

$$\sum_{j=1}^{m} q_{2000\,j} \cdot p_{2000\,j} = 1000 \cdot 2200 + 70 \cdot 2670 + 25 \cdot 1860 + 17 \cdot 3250 = 2488650.$$

$$\sum_{j=1}^{m} q_{2000\,j} \cdot p_{1995\,j} = 1000 \cdot 1730 + 70 \cdot 1900 + 25 \cdot 1590 + 17 \cdot 2380 = 1943210.$$

Podstawiając do wzoru (5.13), otrzymamy:
- dla 1995: $I_p^P = 100\%$,
- dla 1998: $I_p^P = 119{,}13\%$,
- dla 2000: $I_p^P = 128{,}07\%$.

Ceny agregatu składającego się z czterech surowców do produkcji jogurtu w 1998 roku wzrosły o 19,13%, a w 2000 roku o 28,07% w porównaniu z rokiem bazowym, przy założeniu stałego zapotrzebowania z roku badanego.

Indeksy dla poszczególnych lat możemy przyrównywać do roku bazowego. Jednak nie możemy porównać między sobą wyników indeksu z roku 1998 i 2000, ponieważ ich mianowniki są różne. Problem ten nie występował w przypadku indeksu cen Laspeyresa. Posłużymy się więc średnią geometryczną indeksów cen Laspeyresa i Paaschego. Średnia ta jest nazywana **indeksem cen Fishera.** Obliczamy ją według wzoru:

$$I_p^F = \sqrt{I_p^L \cdot I_p^P}. \tag{5.14}$$

Indeks ten podaje średni wzrost cen agregatu dóbr w porównywanych okresach.

**Przykład 5.8**

Obliczamy wartość indeksu cen Fishera dla przykładu 5.6 dotyczącego agregatu surowców do produkcji jogurtów.

- dla 1995 roku: $I_p^F = 100\%$,
- dla 1998 roku: $I_p^F = \sqrt{118{,}64 \cdot 119{,}13} = 118{,}89\%$,
- dla 2000: $I_p^F = \sqrt{128.47 \cdot 128.07} = 128{,}27\%$.

Zatem średni wzrost cen agregatu składającego się z czterech surowców do produkcji jogurtów w 1998 roku wynosi 18,89%, a w 2000 roku jest równy 28,27% w relacji do 1995 roku.

W przypadku **indeksów agregatowych masy towarowej** badamy dynamikę ilości towaru i wówczas wagami stają się ceny. Zajmiemy się dwiema miarami tego typu: indeksem masy towarowej Laspeyresa oraz indeksem masy towarowej Paaschego. Obliczamy je na podstawie następujących wzorów:

**Indeks masy towarowej Laspeyresa:**

$$I_q^L = \frac{\sum_{j=1}^{m} q_{tj} \cdot p_{kj}}{\sum_{j=1}^{m} q_{kj} \cdot p_{kj}} \cdot 100\%. \tag{5.15}$$

**Indeks masy towarowej Paaschego:**

$$I_q^P = \frac{\sum_{j=1}^{m} q_{tj} \cdot p_{tj}}{\sum_{j=1}^{m} q_{kj} \cdot p_{tj}} \cdot 100\%. \tag{5.16}$$

Oznaczenie $q$ stojące przy symbolu I informuje, że jest to indeks ilości przy stałym poziomie cen. Zauważamy, że we wzorach (5.15) i (5.16) indeks Laspeyresa przyjmuje w liczniku i mianowniku ceny z roku bazowego, a indeks Paaschego ceny z roku badanego $t$.

W tym przypadku możemy również zdefiniować **indeks masy towarowej Fishera** dany jako:

$$I_q^F = \sqrt{I_q^L \cdot I_q^P}. \tag{5.17}$$

**Przykład 5.9**

Należy zbadać dynamikę ilości surowców zużywanych do produkcji jogurtów, przyjmując za rok bazowy 1995 (przykład 5.6).

- dla 1995 roku (rok bazowy):

$$I_q^L = 100\%, \quad I_q^P = 100\%, \quad I_q^F = 100\%.$$

- dla 1998 roku:

$$I_q^L = \frac{\sum_{j=1}^{m} q_{1998j} \cdot p_{1995j}}{\sum_{j=1}^{m} q_{1995j} \cdot p_{1995j}} \cdot 100\% = \frac{1453955}{102273} \cdot 100\% = 142{,}16\%,,$$

$$I_q^L = \frac{\sum_{j=1}^{m} q_{2000j} \cdot p_{1995j}}{\sum_{j=1}^{m} q_{1995j} \cdot p_{1995j}} \cdot 100\% = \frac{1943210}{102273} \cdot 100\% = 190{,}00\%.$$

Ilość nabywanych surowców wchodzących w skład analizowanego agregatu w 1998 roku wzrosła o 42,16%, a w 2000 roku o 90%, w porównaniu z 1995 rokiem, przy założeniu, że ceny tych surowców w 1998 i 2000 roku były na tym samym poziomie, co w 1995 roku.

$$I_q^P = \frac{\sum_{j=1}^{m} q_{1998tj} \cdot p_{1998j}}{\sum_{j=1}^{m} q_{1995j} \cdot p_{1998j}} \cdot 100\% = \frac{1732155}{121334} \cdot 100\% = 142{,}76\%,$$

$$I_q^P = \frac{\sum_{j=1}^{m} q_{2000tj} \cdot p_{2000j}}{\sum_{j=1}^{m} q_{1995j} \cdot p_{2000j}} \cdot 100\% = \frac{2488650}{131387} \cdot 100\% = 189{,}41\%.$$

Ilość nabywanych surowców wzrosła w 1998 roku o 42,76% w odniesieniu do 1995 roku, przy założeniu, że ceny w badanych okresach były na poziomie z 1998 roku. W 2000 roku ilość ta wzrosła o 89,41% w porównaniu z 1995 rokiem, przy założeniu, że ceny surowców były w obydwu okresach na poziomie z 2000 roku.

Obliczamy wartość indeksu Fishera:

- 1998 rok

$$I_q^F = \sqrt{142{,}16 \cdot 142{,}76} = 142{,}46\%,$$

- 2000 rok

$$I_q^F = \sqrt{190{,}00 \cdot 189{,}41} = 189{,}71\%.$$

Średni łączny wzrost ilości kupowanych surowców badanego agregatu w 1998 roku wynosi 42,46%, a w 2000 roku 89,71% w stosunku do 1995 roku.

## 5.3 Metody wyodrębniania trendu

Występowanie trendu oraz wahań przypadkowych w dynamice zjawisk jest związane z oddziaływaniem dwóch rodzajów przyczyn: głównych oraz ubocznych. Przyczyny główne wpływają stale i regularnie na badany proces, wywołując w nim systematyczne zmiany. Przyczyny uboczne mają natomiast charakter przypadkowy i zakłócają systematyczne zmiany w procesie (por. rozdział 1 punkt 1.1).

W celu zilustrowania oddziaływania tych przyczyn wyobraźmy sobie następujący przykład: słuchamy koncertu saksofonisty w źle wyciszonej sali koncertowej. Co jakiś czas, nieregularnie przed budynkiem przejeżdża tramwaj i do naszych uszu dochodzi jego stukot. W tym przypadku procesem, który obserwujemy jest dźwięk muzyki saksofonowej. Przyczyną główną jest wysiłek muzyka, który przy użyciu instrumentu tworzy muzykę, natomiast czynnikiem ubocznym, zakłócającym są dźwięki tramwaju.

W analizowanych zjawiskach chcielibyśmy wyodrębnić zmiany zachodzące w szeregu czasowym pod wpływem obydwu omawianych przyczyn, a w szczególności zależy nam na poznaniu tendencji rozwojowej (trendu), która jest wynikiem przyczyn głównych.

Aby określić trend, musimy wyeliminować z szeregu czasowego działanie wahań przypadkowych. W tym celu możemy wykorzystać dwa rodzaje metod:

- mechaniczne metody eliminacji wahań przypadkowych,
- analityczne metody eliminacji wahań przypadkowych.

### 5.3.1 Mechaniczne metody wyodrębniania trendu

Jedną z metod zaliczanych do procedur mechanicznego wyodrębniania trendu jest **metoda średnich ruchomych**. Jest to prosta i często stosowana metoda wygładzania szeregów czasowych. Polega na obliczaniu średnich arytmetycznych z kilku kolejnych wartości szeregu czasowego, zaczynając od początkowych wartości aż do wyczerpania wszystkich wyrazów szeregu. Otrzymujemy w ten sposób ciąg średnich ruchomych. Wykorzystując tę metodę, musimy określić $m$, czyli liczbę jednostek czasu, które obejmuje średnia.

W zależności od tego, czy $m$ jest parzyste, czy nieparzyste będziemy wykorzystywać różne formuły obliczeniowe. Dla $m$ nieparzystych średnią ruchomą obliczamy jako:

$$\overline{y}_{t+\frac{m-1}{2}} = \frac{y_t + y_{t+1} + \ldots + y_{t+m-1}}{m}, \tag{5.18}$$

gdzie:

$t = 1, 2, \ldots n - m + 1$,

$m$ – liczba nieparzystych okresów czasu, które obejmuje średnia (np. 3, 5, 7 itd.),

$y_t$ – wartość zmiennej w $t$-tym okresie czasu.

Przy takim sposobie postępowania tracimy obserwacje dla $\frac{m-1}{2}$ początkowych i $\frac{m-1}{2}$ końcowych okresów badanego zjawiska. Na przykład wzór dla średniej ruchomej 3-okresowej będzie miał postać:

$$\bar{y}_{t+1} = \frac{y_t + y_{t+1} + y_{t+2}}{3}. \tag{5.19}$$

Stwierdzamy, że w przypadku 3-okresowej średniej nie będziemy posiadać wygładzonej wartości dla pierwszego i ostatniego okresu.

### Przykład 5.10

Właściciele firmy fonograficznej chcą dowiedzieć się, jaki jest trend sprzedaży najnowszego albumu znanego artysty rockowego, który nagrywa płyty w tej właśnie firmie. Od premiery minęło 10 miesięcy. Dane przedstawiające liczbę sprzedanych płyt podano w tablicy 5.3. Wyznaczamy trend za pomocą 3-okresowej i 5-okresowej średniej ruchomej. Obliczamy kolejne średnie ruchome 3-okresowe według wzoru (5.19), rozpoczynając od wygładzonej wartości dla drugiego miesiąca.

$$\bar{y}_2 = \frac{175 + 250 + 147}{3} = 190{,}7 \text{ [tys. sztuk]}.$$

$$\bar{y}_3 = \frac{250 + 147 + 129}{3} = 150{,}3 \text{ [tys. sztuk]}.$$

**Tablica 5.3.** Liczba sprzedanych płyt w kolejnych miesiącach po premierze oraz obliczenia pomocnicze

| Miesiąc po premierze $t$ | Liczba sprzedanych egzemplarzy[tys. szt.] $y_i$ | Średnia ruchoma | |
|---|---|---|---|
| | | 3-miesięczna | 5-miesięczna |
| 1 | 175 | – | – |
| 2 | 250 | 190,7 | – |
| 3 | 147 | 150,3 | 171,0 |
| 4 | 129 | 143,3 | 148,0 |
| 5 | 154 | 139,3 | 140,0 |
| 6 | 135 | 141,3 | 140,0 |
| 7 | 135 | 139,0 | 139,4 |
| 8 | 147 | 136,0 | 131,4 |
| 9 | 126 | 129,0 | – |
| 10 | 114 | – | – |

Źródło: dane umowne.

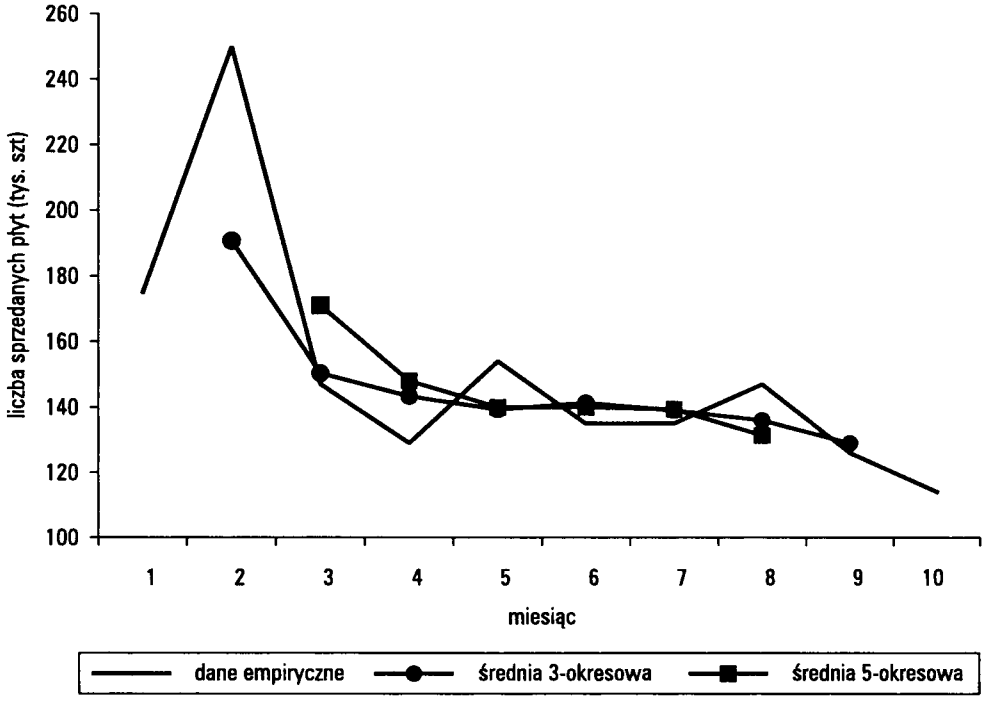
Kolejne obliczone wartości znajdują się w trzeciej kolumnie tablicy 5.3. Następnie obliczamy 5-okresowe średnie ruchome. W tym przypadku tracimy wartości trendu dla dwóch początkowych i dwóch końcowych okresów.

$$\bar{y}_3 = \frac{175+250+147+129+154}{5} = 171 \text{[tys. sztuk]},$$

$$\bar{y}_4 = \frac{250+147+129+154+135}{5} = 148 \text{[tys. sztuk]}.$$

Obliczone wartości dla kolejnych okresów znajdują się w czwartej kolumnie tablicy 5.3.

Graficznie otrzymane wyniki przedstawiamy na rysunku 5.3.




Rys. 5.3. Liczba sprzedanych płyt artysty rockowego w ciągu 10 miesięcy od premiery

Jeżeli liczba okresów, jakie obejmuje średnia jest wartością parzystą ($m$ – parzyste) przy wyznaczaniu trendu posługujemy się tzw. **średnią ruchomą scentrowaną**, obliczaną jako:

$$\bar{\bar{y}}_{t+\frac{m}{2}} = \frac{\frac{1}{2}y_t + y_{t+1} + \ldots + y_{t+m-1} + \frac{1}{2}y_{t+m}}{m}. \tag{5.20}$$

Jak zauważamy, w celu obliczenia średniej scentrowanej bierzemy połowy wartości skrajnych $y_t$ i $y_{t+m}$. W przypadku średnich scentrowanych utracimy wartości trendu dla $\frac{m}{2}$ początkowych i $\frac{m}{2}$ końcowych okresów obserwacji analizowanego zjawiska.

Sprzedaż płyt w ciągu 10 miesięcy od momentu premiery wykazywała tendencję malejącą. Dla $m = 4$ formuła obliczeniowa przyjmuje postać:

$$\bar{\bar{y}}_{t+2} = \frac{\frac{1}{2}y_t + y_{t+1} + y_{t+2} + y_{t+3} + \frac{1}{2}y_{t+4}}{4}. \tag{5.21}$$

Ciąg średnich ruchomych wyznacza trend badanego zjawiska. Istotnym problemem jest tutaj ustalenie długości okresów, z których będziemy obliczali średnie. Musimy pamiętać, że im dłuższe będą okresy, tym lepiej będzie wyrównany szereg czasowy, inaczej mówiąc, lepiej „oczyszczony" z wahań przypadkowych. Z drugiej jednak strony im dłuższe okresy, tym wygładzony szereg czasowy jest krótszy. Tracimy informacje dla skrajnych okresów szeregu.

Przyjmuje się, że ilość okresów, z których oblicza się średnią ruchomą $m$-okresową nie powinna przekraczać $0{,}5n$ okresów w przypadku szeregu o parzystej liczbie okresów oraz $0{,}5(n-1)$ okresów, dla szeregu o nieparzystej liczbie okresów[54].

Metoda średnich ruchomych ma wiele ograniczeń, między innymi nie daje w rezultacie czystego trendu. Średnie ruchome obarczone są błędami wynikającymi z prostoty obliczeń, która nie pozwala w pełni wyeliminować wahań przypadkowych. Ograniczeniem tej metody jest również brak możliwości ekstrapolacji trendu na okresy przyszłe, czyli prognozowania badanego zjawiska[55].

### Przykład 5.11

Dynamikę sprzedaży płyt artysty rockowego zbadamy, stosując scentrowaną średnią ruchomą 2 i 4-okresową (przykład 5.10). Dane i obliczenia pomocnicze podano w tablicy 5.4. Scentrowane średnie ruchome 2-okresowe obliczamy, korzystając ze wzoru (5.20) w następujący sposób:

$$\bar{\bar{y}}_2 = \frac{\frac{1}{2}\cdot 175 + 250 + \frac{1}{2}\cdot 147}{2} = 205{,}5 \text{ [tys. sztuk]}.$$

$$\bar{\bar{y}}_3 = \frac{\frac{1}{2}\cdot 250 + 147 + \frac{1}{2}\cdot 129}{2} = 168{,}3 \text{ [tys. sztuk] itd.}$$

Kolejne wygładzone wartości zapisano w trzeciej kolumnie tablicy 5.4.

---

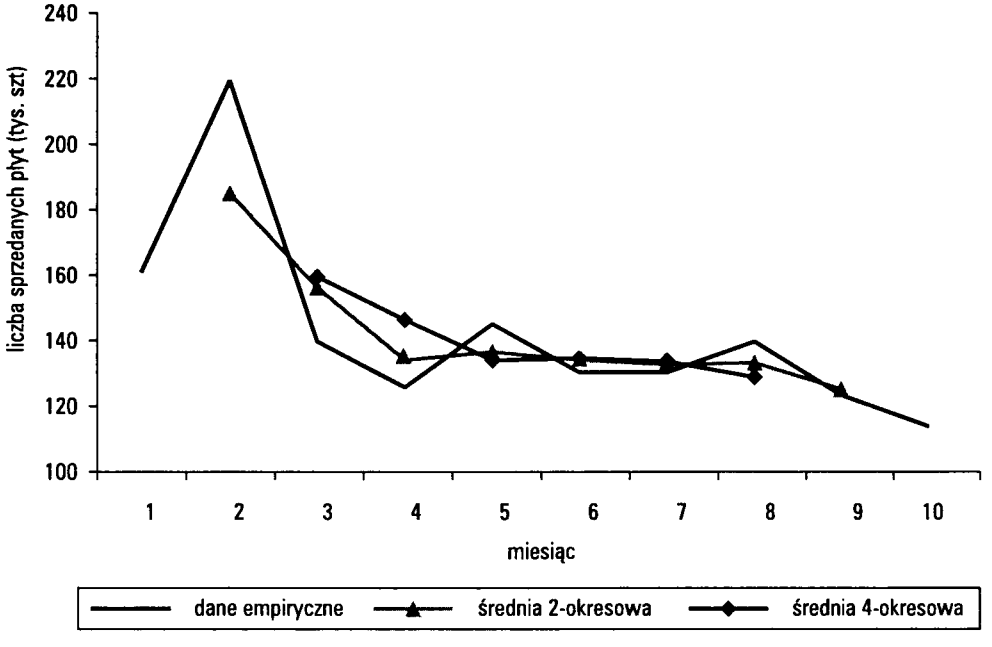
[54] Por. M. Woźniak (red.), *Statystyka ogólna*, *op. cit.*

[55] A. Zeliaś, *Metody statystyczne*, *op. cit.*

**Tablica 5.4.** Liczba sprzedanych płyt artysty oraz obliczenia pomocnicze

| Miesiąc po premierze $t$ | Liczba sprzedanych egzemplarzy[tys. szt.] $y_t$ | Średnia ruchoma | |
|---|---|---|---|
| | | 2-miesięczna | 4-miesięczna |
| 1 | 175 | – | – |
| 2 | 250 | 205,5 | – |
| 3 | 147 | 168,3 | 172,6 |
| 4 | 129 | 139,8 | 155,6 |
| 5 | 154 | 143,0 | 139,8 |
| 6 | 135 | 139,8 | 140,5 |
| 7 | 135 | 138,0 | 139,3 |
| 8 | 147 | 138,8 | 133,1 |
| 9 | 126 | 128,3 | – |
| 10 | 114 | – | – |

Źródło: dane umowne.




Rys. 5.4. Liczba sprzedanych płyt artysty rockowego

Wyznaczamy średnie scentrowane 4-okresowe (wzór (5.21)):

$$\bar{\bar{y}}_3 = \frac{\frac{1}{2}\cdot 175 + 250 + 147 + 129 + \frac{1}{2}\cdot 154}{4} = 172{,}6\,[\text{tys. sztuk}],$$

$$\bar{\bar{y}}_4 = \frac{\frac{1}{2}\cdot 250 + 147 + 129 + 154 + \frac{1}{2}\cdot 135}{4} = 155{,}6\,[\text{tys. sztuk}]\ \text{itd.}$$

Obliczone wartości średnich 4-okresowych dla rozważanego przykładu podano w czwartej kolumnie tablicy 5.4. Trendy wyznaczone mechaniczną metodą scentrowanych średnich ruchomych 2 i 4-okresowych dla przykładu 5.11 zostały przedstawione na rysunku 5.4.

## 5.3.2. Analityczne metody wyodrębniania trendu

**Metody analityczne wyodrębniania trendu** polegają na dopasowaniu do danych empirycznych funkcji, w której *zmienną objaśnianą* jest poziom zjawiska obserwowanego w określonych okresach czasu, a *zmienną objaśniającą* jest czas $t$.

Funkcja trendu będzie przedstawiała prawidłowości w kształtowaniu się badanego zjawiska w przeszłości. Postać funkcyjna umożliwia ekstrapolację trendu w okresy przyszłe. W takiej sytuacji musimy brać pod uwagę charakter badanego zjawiska, między innymi, czy zmienność okoliczności wpływających na badane zjawisko w przyszłości możliwa jest do przewidzenia.

Podstawowym problemem w przypadku metody analitycznej jest wybór odpowiedniej postaci funkcji trendu, czyli takiej, która będzie w najlepszy sposób opisywała kształtowanie się badanego zjawiska w czasie. Najczęściej wykorzystywane są funkcje: liniowa, wielomianowa, logarytmiczna, wykładnicza, potęgowa. Jednym ze sposobów pomagających w podjęciu decyzji jest *analiza graficzna*. W układzie współrzędnych prostokątnych zaznaczamy punkty empiryczne odpowiadające wartości przyjmowanych przez zmienną objaśnianą $Y$ w kolejnych okresach czasu. Sposób, w jaki układają się zaznaczone punkty, może wskazywać na rodzaj krzywej, która będzie opisywała zjawisko[56].

Jeśli w poszczególnych okresach bezwzględne przyrosty wartości zmiennej $Y_t$ są w przybliżeniu stale, to tendencję rozwojową (trend) możemy opisać za pomocą funkcji liniowej o postaci:

$$y_t = a + b \cdot t + \varepsilon_t, \tag{5.22}$$

[56] Por. M. Woźniak (red.), *Statystyka ogólna*, *op. cit.*; A. Zeliaś, *Metody statystyczne*, *op. cit.*; W. Starzyńska, *Statystyka praktyczna*, *op. cit.*

gdzie:

$y_t$ – zmienna objaśniana obrazująca dynamikę badanego zjawiska w czasie $t$,

$t$ – zmienna objaśniająca w postaci zmiennej czasowej,

$a$, $b$ – parametry modelu,

$\varepsilon_t$ – składnik przypadkowy charakteryzujący odchylenia przypadkowe zmiennej od linii trendu.

Funkcję trendu liniowego możemy przedstawiać w postaci:

$$\hat{y}_t = a + b \cdot t. \tag{5.23}$$

Przyrost zmiennej $t$ między dwoma okresami (na przykład między latami, miesiącami, kwartałami) jest równy 1. W obliczeniach zamiast lat kalendarzowych (np. 1991, 1992 itd.), możemy posłużyć się numerami zmiennej $t$, rozpoczynając od $t = 1$ oznaczającej pierwszy badany okres (czyli np. 1991 rok).

Wyraz wolny „a" funkcji (5.23) możemy interpretować jako oszacowanie poziomu rozważanej zmiennej w okresie $t = 0$, a więc bezpośrednio poprzedzającym okres obserwacji. Współczynnik kierunkowy „b" interpretujemy jako średnioroczny przyrost lub spadek wartości zmiennej w rozważanym okresie.

Najczęściej stosowaną metodą dopasowania funkcji trendu do danych empirycznych jest **metoda najmniejszych kwadratów**, która została przedstawiona podczas omawiania regresji liniowej (rozdział 4 punkt 4.6). Sposób postępowania dotyczący określania postaci funkcji liniowej trendu jest analogiczny do poszukiwania funkcji regresji liniowej i dlatego w tym miejscu pomijamy szczegółowe rozważania. Przypominamy jedynie, że podstawowym zadaniem tej metody jest ustalenie nieznanych wartości parametrów $a$ i $b$ na podstawie danych pochodzących z obserwacji. Dodajmy, że wartości zmiennej objaśnianej ($x_i$) w regresji liniowej zastępujemy zmienną czasową t, otrzymując układ równań normalnych o postaci:

$$\begin{cases} \sum_{t=1}^{n} y_t = n \cdot a + b \cdot \sum_{t=1}^{n} t \\ \sum_{t=1}^{n} t \cdot y_t = a \cdot \sum_{t=1}^{n} t + b \cdot \sum_{t=1}^{n} t^2 \end{cases}. \tag{5.24}$$

Układ ten możemy rozwiązać metodą wyznaczników:

$$|A| = \begin{vmatrix} n & \sum_{t=1}^{n} t \\ \sum_{t=1}^{n} t_t & \sum_{t=1}^{n} t^2 \end{vmatrix} \tag{5.25}$$

$$|A_1| = \begin{vmatrix} \sum_{t=1}^{n} y_t & \sum_{t=1}^{n} t \\ \sum_{t=1}^{n} t \cdot y_t & \sum_{t=1}^{n} t^2 \end{vmatrix} \tag{5.26}$$

$$|A_2| = \begin{vmatrix} n & \sum_{t=1}^{n} y_t \\ \sum_{t=1}^{n} t & \sum_{t=1}^{n} t \cdot y_t \end{vmatrix} \tag{5.27}$$

$$a = \frac{|A_1|}{|A|}, \qquad b = \frac{|A_2|}{|A|}. \tag{5.28}$$

Sposób dopasowywania funkcji trendu do danych empirycznych zilustrujemy w przykładzie 5.12.

### Przykład 5.12

Zarządzający przedsiębiorstwem produkującym sprzęt gospodarstwa domowego interesują się dynamiką zysku netto firmy w latach 1991–2001. Niezbędne dane zawiera tablica 5.5. Tendencję rozwojową chcemy opisać za pomocą funkcji trendu o postaci liniowej.

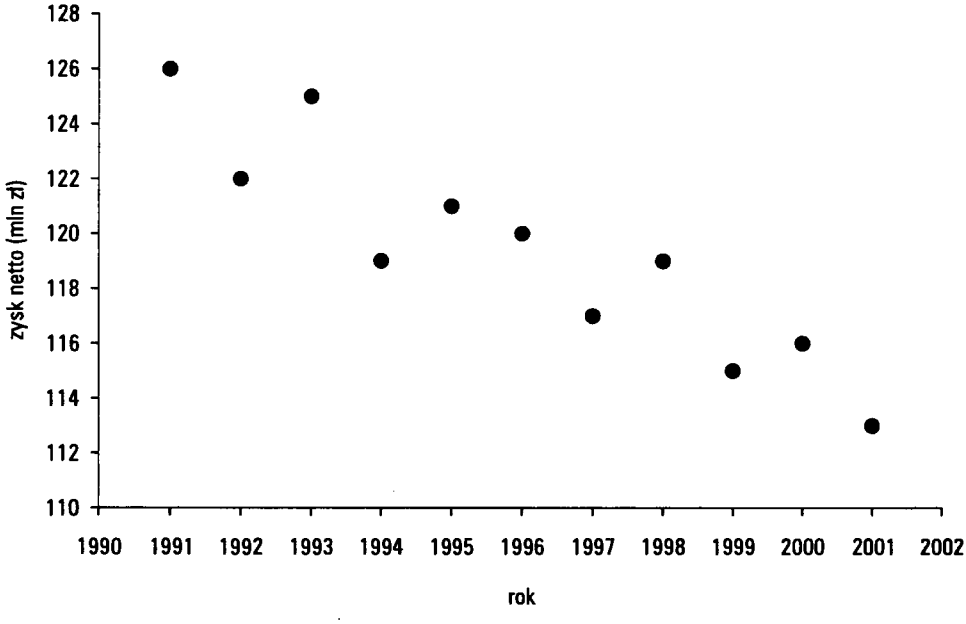
**Tablica 5.5.** Zysk netto przedsiębiorstwa produkującego sprzęt AGD w latach 1991–2001

| Rok | Zysk netto [w mln zł] |
|---|---|
| 1991 | 126 |
| 1992 | 122 |
| 1993 | 125 |
| 1994 | 119 |
| 1995 | 121 |
| 1996 | 120 |
| 1997 | 117 |
| 1998 | 119 |
| 1999 | 115 |
| 2000 | 116 |
| 2001 | 113 |

Źródło: dane umowne.

Analizę rozpoczynamy od sporządzenia rysunku 5.5. Na podstawie analizy wykresu możemy stwierdzić, że punkty empiryczne układają się wzdłuż pewnej linii prostej. Przystąpmy do wyznaczenia wartości nieznanych parametrów funkcji. Zastosu-

jemy w tym celu metodę najmniejszych kwadratów. Według wzorów (5.25, 5.26, 5.27, 5.28) obliczamy wartości a oraz b. Działania pomocnicze podano w tablicy 5.6.




Rys. 5.5. Zysk netto (w mln zł) przedsiębiorstwa produkującego sprzęt AGD w latach 1991–2001

Wyniki obliczeń są następujące:

$$|A| = \begin{vmatrix} 11 & 66 \\ 66 & 506 \end{vmatrix} = 5566 - 4356 = 1210,$$

$$|A_1| = \begin{vmatrix} 1313 & 66 \\ 7755 & 506 \end{vmatrix} = 664378 - 511830 = 152548,$$

$$|A_2| = \begin{vmatrix} 11 & 1313 \\ 66 & 7755 \end{vmatrix} = 85305 - 86558 = -1353,$$

$$a = \frac{152548}{1210} = 126{,}07, \qquad b = \frac{-1353}{1210} = -1{,}12.$$

Funkcja trendu przyjmuje postać:

$$\hat{y}_t = 126{,}07 - 1{,}12t.$$

Na podstawie uzyskanych rezultatów możemy stwierdzić, że kształtowanie się zysku netto przedsiębiorstwa produkującego sprzęt AGD w badanym okresie charakteryzowało się trendem malejącym. Średnioroczny spadek zysku wynosił 1,12 mln złotych. Możemy również oszacować wielkość zysku w 1990 roku. Jeśli podstawimy $t = 0$, to uzyskamy $y_0 = 126{,}07$ mln zł.

Kolejnym krokiem jest obliczenie wartości hipotetycznych ($\hat{y}_t$) dla poszczególnych okresów. Podano je w tablicy 5.6. Natomiast na rysunku 5.6 przedstawiono przebieg wyznaczonej linii trendu wraz z punktami empirycznymi.

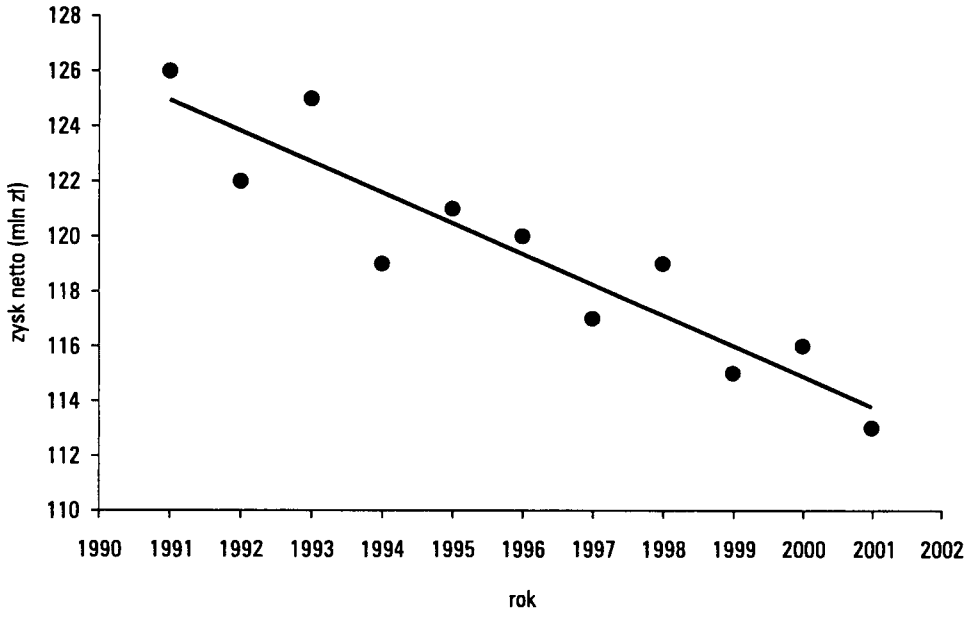
**Tablica 5.6.** Zysk netto przedsiębiorstwa AGD oraz obliczenia pomocnicze

| Rok | Okres $t$ | Zysk netto (w mln zł) $y_t$ | $t^2$ | $t \cdot y_t$ | $\hat{y}$ | $(y_t - \hat{y})^2$ | $(y_t - \bar{y}_t)^2$ |
|---|---|---|---|---|---|---|---|
| 1991 | 1 | 126 | 1 | 126 | 124,95 | 1,10 | 44,04 |
| 1992 | 2 | 122 | 4 | 244 | 123,83 | 3,35 | 6,95 |
| 1993 | 3 | 125 | 9 | 375 | 122,71 | 5,24 | 31,77 |
| 1994 | 4 | 119 | 16 | 476 | 121,59 | 6,71 | 0,13 |
| 1995 | 5 | 121 | 25 | 605 | 120,47 | 0,28 | 2,68 |
| 1996 | 6 | 120 | 36 | 720 | 119,35 | 0,42 | 0,40 |
| 1997 | 7 | 117 | 49 | 819 | 118,23 | 1,51 | 5,59 |
| 1998 | 8 | 119 | 64 | 952 | 117,11 | 3,57 | 0,13 |
| 1999 | 9 | 115 | 81 | 1035 | 115,99 | 0,98 | 19,04 |
| 2000 | 10 | 116 | 100 | 1160 | 114,87 | 1,28 | 11,31 |
| 2001 | 11 | 113 | 121 | 1243 | 113,75 | 0,56 | 40,50 |
| Razem | 66 | 1313 | 506 | 7755 | 1312,85 | 25,01 | 162,55 |

Źródło: dane umowne.

Podobnie jak w przypadku regresji liniowej, musimy ocenić dopasowanie funkcji trendu do danych empirycznych. W tym celu będziemy korzystać z przedstawionych w punkcie 4.7 miar, do których należą:

- odchylenie składnika resztowego (średni błąd szacunku) $s_\varepsilon$,
- współczynnik zmienności resztowej $V_\varepsilon$,
- współczynnik braku determinacji $\varphi^2$,
- współczynnik determinacji $R^2$.




Rys. 5.6. Dynamika zysku przedsiębiorstwa AGD w latach 1991–2001

### Przykład 5.13

Ocenimy dopasowanie funkcji trendu do danych empirycznych dla przykładu 5.12, gdzie:

$$\hat{y}_t = 126{,}07 - 1{,}12t.$$

Obliczenia pomocnicze do uzyskania wartości poszczególnych miar dobroci dopasowania zawiera tablica 5.6.

- odchylenie standardowe składnika resztowego:

$$s_\varepsilon = \sqrt{\frac{\sum_{i=1}^{n}(y_i - \hat{y}_t)^2}{n-k}} = \sqrt{\frac{25{,}01}{11-2}} = \sqrt{2{,}278} = 1{,}67 \text{ mln złotych}].$$

  Zaobserwowane w poszczególnych latach wielkości zysku netto w przedsiębiorstwie produkującym sprzęt AGD różnią się od wielkości teoretycznych obliczonych na podstawie funkcji trendu średnio o 1,67 mln złotych.

- współczynnik zmienności przypadkowej:

$$V_\varepsilon = \frac{s_\varepsilon}{\bar{y}} \cdot 100\%.$$

W celu obliczenia współczynnika zmienności przypadkowej, obliczamy wartość średniej arytmetycznej:

$$\bar{y} = \frac{1313}{11} = 119{,}36 \text{ [mln. złotych]}.$$

Średni zysk netto osiągany przez przedsiębiorstwo w badanym okresie wynosił 119,36 mln złotych.

$$V_\varepsilon = \frac{1{,}67}{119{,}36} \cdot 100\% = 1{,}4\%.$$

Odchylenie standardowe składnika resztowego stanowi 1,4% przeciętnego zysku netto przedsiębiorstwa, co świadczy o dobrym dopasowaniu wyznaczonej funkcji trendu.

- współczynnik braku determinacji (zbieżności):

$$\varphi^2 = \frac{\sum_{t=1}^{n}(y_i - \hat{y}_t)}{\sum_{t=1}^{n}(y_t - \bar{y})^2} = \frac{25{,}01}{162{,}55} = 0{,}154.$$

  Zmienność zysku netto przedsiębiorstwa w 15,4% nie jest wyjaśniona przez funkcję trendu.

- współczynnik determinacji:

$$R^2 = 1 - \varphi^2 = 1 - 0{,}154 = 0{,}846.$$

  Zmienność zysku netto jest wyjaśniona w 84,6% przez funkcję trendu.

Podsumowując uzyskane rezultaty, możemy stwierdzić, że dopasowanie funkcji trendu do danych empirycznych jest zadowalające.

## 5.4. Analiza wahań okresowych

Celem analizy wahań okresowych jest wyodrębnienie regularnych odchyleń od ogólnej tendencji rozwojowej, występujących w ustalonych odstępach czasu. W tym celu będziemy rozpatrywać szeregi czasowe, w których jednostką obserwacji jest na przykład: miesiąc, kwartał, półrocze, a więc okresy krótsze niż rok. Rozważania zilustrujemy przykładem.

**Przykład 5.14**

Dla zapewnienia ciągłości produkcji firma Z utrzymuje pewien poziom zapasów surowca. W tablicy 5.7 podano stan zapasów w poszczególnych półroczach w latach 2000–2004.

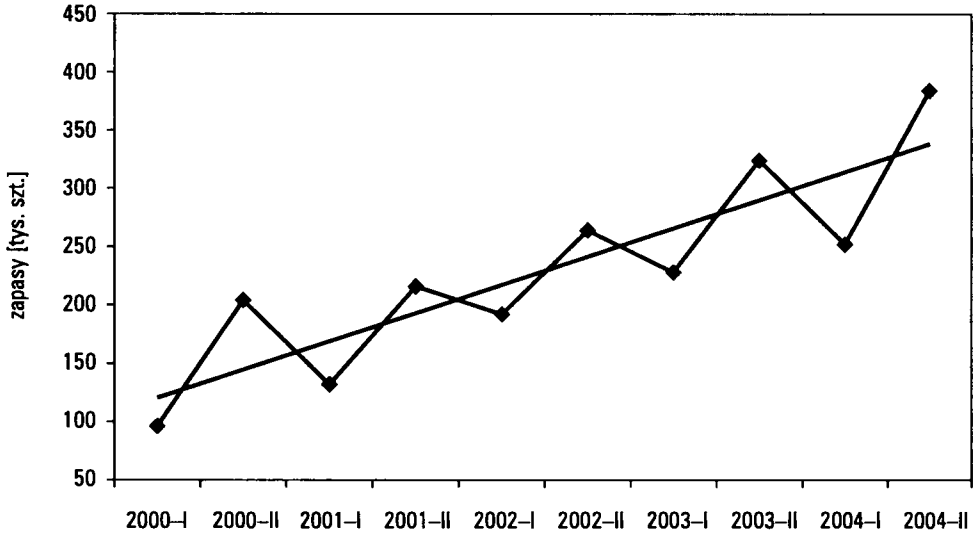
Na rysunku 5.7 przedstawiono kształtowanie się zapasów w rozważanym okresie. Przebieg krzywej pozwala przypuszczać, iż w tym szeregu czasowym możemy wyróżnić: trend, wahania okresowe i wahania przypadkowe. Zajmiemy się wyodrębnianiem poszczególnych składników. Analiza składa się z następujących etapów:

1) wygładzenie szeregu czasowego metodą mechaniczną lub analityczną,
2) wyeliminowanie trendu,
3) wyodrębnienie wahań przypadkowych – określenie surowych wskaźników sezonowości,
4) ustalenie czystych wskaźników sezonowości.

Tablica 5.7. Zapasy surowca w firmie Z w półroczach 2000–2004

| Rok | Półrocze | Zapas [tys. szt.] |
|---|---|---|
| 2000 | I | 96 |
| 2000 | II | 204 |
| 2001 | I | 132 |
| 2001 | II | 216 |
| 2002 | I | 192 |
| 2002 | II | 264 |
| 2003 | I | 228 |
| 2003 | II | 324 |
| 2004 | I | 252 |
| 2004 | II | 384 |

Źródło: dane umowne.




Rys. 5.7. Zapasy surowca w firmie Z w półroczach 2000–2004

Wygładzanie szeregu czasowego polega na przyporządkowaniu każdej wartości empirycznej odpowiedniej wartości teoretycznej, którą obliczamy albo jako średnią ruchomą (por. 5.3.1), albo ustalamy na podstawie funkcji trendu (por. 5.3.2).

W rozważanym przykładzie posłużymy się analityczną metodą wyodrębniania tendencji rozwojowej. Na postawie rysunku 5.7 można uznać, że zapasy przedsiębiorstwa charakteryzowały się trendem liniowym. W związku z tym metodą najmniejszych kwadratów znajdujemy wartości parametrów funkcji o postaci:

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

Po wykonaniu odpowiednich obliczeń otrzymujemy: $\hat{y}_t = 96{,}00 + 24{,}22 \cdot t$.

Obliczmy wartości teoretyczne zmiennej $Y$, które zostały zapisane w tablicy 5.8.

**Tablica 5.8.** Empiryczne i teoretyczne wartości zapasów

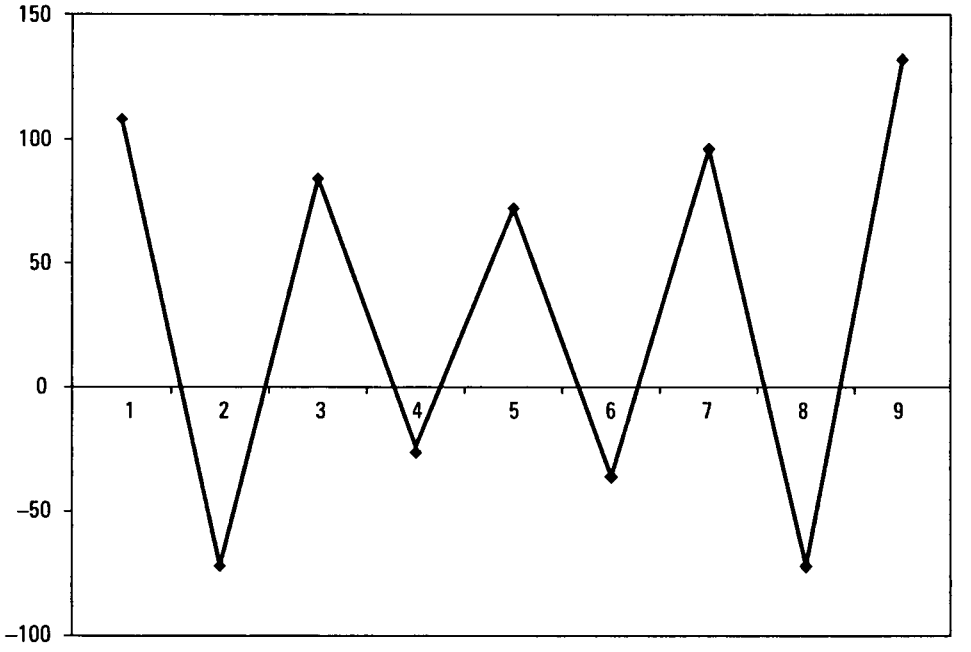
| $y_t$ | $\hat{y}_t$ | $w_t = \frac{y_t}{\hat{y}_t}$ |
|---|---|---|
| 96 | 120 | 0,799 |
| 204 | 144 | 1,412 |
| 132 | 169 | 0,783 |
| 216 | 193 | 1,120 |
| 192 | 217 | 0,884 |
| 264 | 241 | 1,094 |
| 228 | 266 | 0,859 |
| 324 | 290 | 1,118 |
| 252 | 314 | 0,803 |
| 384 | 338 | 1,135 |

Źródło: dane umowne.

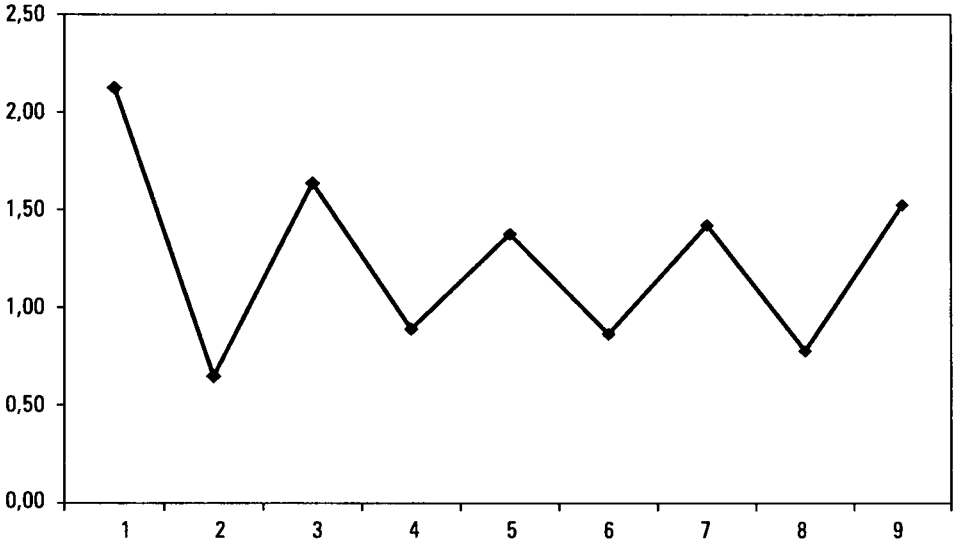
Otrzymaliśmy wygładzony szereg czasowy zapasów. W dalszej kolejności należy zastanowić się, w jaki sposób wyeliminować trend. Można to uczynić dwoma sposobami, a mianowicie:

1) od wartości empirycznych ($y_t$) odejmujemy odpowiadające im wartości teoretyczne ($\hat{y}_t$)
2) wartości empiryczne ($y_t$) dzielimy przez odpowiadające im wartości teoretyczne ($\hat{y}_t$).

W podjęciu decyzji pomoże prześledzenie, jak zmieniają się przyrosty bezwzględne (wzór 5.2) lub względne (wzór 5.7). Przebieg ich przedstawiono na rysunkach 5.8 (przyrosty bezwzględne) oraz 5.9 (przyrosty względne).




Rys. 5.8. Przyrosty bezwzględne zapasów w półroczach 2000–2004




Rys. 5.9. Stopy przyrostu zapasów w półroczach okresu 2000–2004

Stopy przyrostu charakteryzują się stałością, dlatego trend wyeliminujemy, dzieląc wartości empiryczne przez odpowiadające im wartości teoretyczne. Oznaczymy je jako $w_t$. Wyniki obliczeń zapisano w tablicy 5.8. Szereg $w_t$ zawiera wahania okresowe i przypadkowe. Wahania przypadkowe usuwamy, obliczając średnie arytmetyczne wartości $w_t$ dla jednoimiennych okresów (dla tego samego cyklu wahań). Średnie te nazywamy surowymi wskaźnikami sezonowości.

**Surowe wskaźniki sezonowości** informują, jaki byłby poziom obserwowanej zmiennej (poziom zapasów), gdyby nie występowały wahania, a tendencja rozwojowa kształtowałaby się zgodnie z funkcją trendu[57]. W celu łatwiejszego zaprezentowania omawianej procedury dane z tablicy 5.8 zgrupujemy według półroczy, które są okresami jednoimiennymi, czyli tymi samymi cyklami wahań. Zapiszmy je w tablicy 5.9.

**Tablica 5.9.** Ustalanie wskaźników sezonowości zapasów surowców

| Rok | Półrocze | |
|---|---|---|
| | I | II |
| | $w_{tk}$ | |
| 2000 | 0,800 | 1,412 |
| 2001 | 0,780 | 1,120 |
| 2002 | 0,880 | 1,094 |
| 2003 | 0,860 | 1,118 |
| 2004 | 0,800 | 1,135 |
| Suma | 4,120 | 5,879 |
| Średnie półroczne wskaźniki surowe $\overline{w}_k$ | 0,8240 | 1,1758 |
| $\overline{w}_I + \overline{w}_{II}$ | 1,9998 | |
| Wskaźniki czyste | 0,8241 | 1,1759 |

Źródło: dane umowne.

Surowe wskaźniki sezonowości ustalamy jako:

$$\overline{w}_k = \frac{\sum_{j=1}^{s} w_{jk}}{s}, \tag{5.29}$$

$s$ – liczba jednoimiennych okresów,
$k$ – liczba faz wahań.

[57] Por. np.: S. Ostasiewicz, Z. Rusnak, U. Siedlecka, *op. cit.*, Wrocław 1999.

Dla rozważanego przykładu $s = 5$ lat, ponieważ w okresie 10 lat występowało po pięć pierwszych i drugich półroczy, zaś $k = 2$, gdyż występowały dwie fazy (I i II półrocze).

W rezultacie obliczeń otrzymujemy:

$$\overline{w_I} = \frac{\sum_{j=1}^{s} w_{jI}}{s} = \frac{4{,}120}{5} = 0{,}8240,$$

$$(0{,}8240 - 1) \cdot 100\% = -17{,}49\%,$$

$$\overline{w_I} = \frac{\sum_{j=1}^{s} w_{jI}}{s} = \frac{5{,}879}{5} = 1{,}1758,$$

$$(1{,}1758 - 1) \cdot 100\% = 17{,}58\%.$$

**Surowe wskaźniki** sezonowości informują, że gdyby nie występowały wahania okresowe, to poziom zapasów w pierwszych półroczach byłby o 17,49% niższy, a w drugich półroczach o 17,58% niższy przy tendencji rozwojowej zgodnej z funkcją trendu.

Suma surowych wskaźników powinna być równa 2, ponieważ tyle jest faz cyklu. W naszym przykładzie mamy: 0,8240 + 1,1758 = 1,9998. Należy skorygować wskaźniki. W tym celu każdy z nich dzielimy przez ich średnią arytmetyczną.

Obliczamy:

1) średnią arytmetyczną surowych wskaźników:

$$\frac{0{,}8240 + 1{,}1758}{2} = \frac{1{,}9998}{2} = 0{,}9999,$$

2) czyste wskaźniki sezonowości:

dla I kwartałów: $w_I = \frac{0{,}8240}{0{,}9999} = 0{,}8241,$

dla I kwartałów: $w_{II} = \frac{1{,}1758}{0{,}9999} = 1{,}1759,$

0,8241 + 1,1759 = 2,0000.

Czyste wskaźniki informują, że gdyby nie było wahań okresowych, to stan zapasów w I półroczu byłby o 17,59% wyższy, a w II półroczu o 17,59% niższy od poziomu przeciętnego.

# Podstawy rachunku prawdopodobieństwa

Rozdział 6

## 6.1. Zmienna losowa

Wprowadzenie zmiennej losowej rozpoczniemy od następującego przykładu[58].

**Przykład 6.1**

Gra polega na rzucie trzema monetami i otrzymaniu wygranej w wysokości 100 złotych, jeśli wypadną trzy orły; przegranej 50 złotych, gdy wypadną trzy reszki. W pozostałych przypadkach nie ma wypłat. Należy obliczyć prawdopodobieństwo poszczególnych wygranych.

Najpierw rozważamy realizację zdarzeń losowych, jakimi są wyniki rzutu trzema monetami. Liczebność zbioru wszystkich możliwych zdarzeń (wyników rzutu trzema monetami) wyznaczamy zgodnie ze wzorem na wariacje z powtórzeniami jako $n^k = 2^3 = 8$ elementów. Każdy wynik ma jednakowe prawdopodobieństwo realizacji równe 1/8. W tablicy 6.1 podano wszystkie możliwe wyniki rzutu trzema monetami wraz z odpowiadającymi im prawdopodobieństwami i wysokością wygranych.

**Tablica 6.1.** Prawdopodobieństwo zdarzeń i odpowiednich wygranych w rzucie trzema monetami

| Realizacje zdarzenia ($A_i$) | OOO | RRR | ORR | ROR | RRO | RRO | RRO | ORR |
|---|---|---|---|---|---|---|---|---|
| Prawdopodobieństwo $P(A_i)$ | 1/8 | 1/8 | 1/8 | 1/8 | 1/8 | 1/8 | 1/8 | 1/8 |
| Wygrana ($x_i$) | 100 | –50 | 0 | 0 | 0 | 0 | 0 | 0 |

Legenda: O – orzeł; R – reszka.

[58] Por.: T. Gerstenkorn, T. Śródka, *Kombinatoryka i rachunek prawdopodobieństwa*, Warszawa 1973, s. 187.

Wygrana w grze jest funkcją określoną na zbiorze zdarzeń elementarnych podanych w tablicy 6.1. Jeśli zrealizuje się zdarzenie $A_1$ (uzyskanie trzech orłów), to funkcja przyjmie wartość 100, a prawdopodobieństwo jej uzyskania jest takie samo jak zajścia zdarzenia $A_1$. W kontekście ustalonych warunków gry, nie musimy rozróżniać wyników rzutu, w których nie uzyskujemy wypłat. Takich przypadków jest sześć. Funkcja przyjmie wartość 0 z prawdopodobieństwem 6/8. Mamy tu do czynienia z sumą sześciu wyłączających się zdarzeń. Jeśli zrealizuje się zdarzenie $A_8$, to przegramy 50 złotych, co oznacza, że funkcja przyjmie wartość –50 z prawdopodobieństwem 1/8. Wysokość wygranej w grze jest zatem zmienną losową. Rozkład prawdopodobieństwa wygranych podamy w tablicy 6.2.

**Tablica 6.2.** Prawdopodobieństwo wygranych

| $x_i$ | $P(X = x_i) = p_i$ |
|---|---|
| –50 | 1/8 |
| 0 | 6/8 |
| 100 | 1/8 |
| Suma | 1 |

**Zmienna losowa** jest to funkcja, której polem jest podstawowy zbiór zdarzeń elementarnych[59]. Zmienne losowe oznaczamy dużymi literami, najczęściej $X$, $Y$, Z.

Możemy interesować się, jakie jest prawdopodobieństwo, że w opisanej grze wygramy mniej niż 100 złotych, a więc szukamy $P(X < 100)$, które obliczamy jako:

$$P(X<100) = P(X=-50) + P(X=0) = \frac{7}{8}.$$

Możemy znaleźć wartości tej funkcji w każdym punkcie $x_i$.

$$P(X<-50) = 0.$$

$$P(X<0) = P(X=-50) = \frac{1}{8}.$$

$$P(X<100) = P(X=-50) + P(X=0) = \frac{7}{8}.$$

$$P(X<100+\varepsilon) = P(X=-50) + P(X=0) + P(X=100) = 1.$$

[59] Por.: M. Fisz, *Rachunek prawdopodobieństwa i statystyka matematyczna*, Warszawa 1967; T. Gerstenkorn, T. Śródka, *op. cit.*

Prawdopodobieństwo, że zmienna losowa $X$ przyjmie wartość mniejszą od liczby rzeczywistej x nazywamy **dystrybuantą** lub **funkcją rozkładu prawdopodobieństwa** zmiennej losowej $X$, co zapisujemy:

$$F(x) = P(X < x). \tag{6.1}$$

Funkcja F($x$) posiada następujące własności, które podamy bez dowodu[60]:

1) $0 \le F(x) \le 1$,
2) $F(x)$ jest przynajmniej lewostronnie ciągła,
3) $F(x)$ jest niemalejąca,
4) $\lim_{x \to -\infty} F(x) = F(-\infty) = 0, \quad \lim_{x \to \infty} F(x) = F(\infty) = 1.$

Własności te stanowią warunek konieczny i dostateczny, aby funkcja $F(x)$ była dystrybuantą.

Prawdopodobieństwo, że zmienna losowa przyjmie wartość z przedziału $x_1 \le x < x_2$ obliczamy jako:

$$P(x_1 \le X < x_2) = F(x_2) - F(x_1). \tag{6.2}$$

Wśród zmiennych losowych wyróżniamy zmienne typu skokowego i ciągłego.

**Zmienną losową skokową** nazywamy taką zmienną losową $X$, dla której istnieje funkcja $P(X = x_k) = p_k > 0$ (k = 1, 2, ...) taka, że dla każdego rzeczywistego $x$ zachodzi relacja:

$$F(x) = P(X < x) = \sum_{x_k < x} p(X = x_k). \tag{6.3}$$

Funkcję

$$F(x) = P(X < x) = \sum_{x_k < x} p(X = x_k). \tag{6.4}$$

nazywamy funkcją prawdopodobieństwa skokowej zmiennej losowej $X$. Wartości $x_k$ nazywamy punktami skokowymi, a prawdopodobieństwa $p_k$ – skokami[61].

$$\sum_k p_k = 1. \tag{6.5}$$

### Przykład 6.2

Liczbę dzieci w rodzinie możemy rozpatrywać jako zmienną losową. Załóżmy, że przyjmuje ona następujące wartości: 0, 1, 2, 3, 4, 5, 6. Rozkład i dystrybuantę zmiennej losowej można podać w postaci: tabeli, wykresu, wzoru analitycznego oraz ująć za pomocą odpowiednich miar. W tablicy 6.3 podano rozkład prawdopodobieństwa

---

[60] Dowody można znaleźć na przykład w pracach: M. Fisz, *op. cit.*, s. 43; T. Gerstenkorn, T. Śródka, *op. cit.*, s. 189.

[61] *Ibidem*.

(funkcję prawdopodobieństwa) i dystrybuantę (funkcję rozkładu prawdopodobieństwa) tej zmiennej losowej.

Na rysunkach 6.1 i 6.2 przedstawiono odpowiednio rozkład prawdopodobieństwa i dystrybuantę liczby dzieci w rodzinie rozpatrywanej jako zmienna losowa.

**Zmienną losową ciągłą** nazywamy zmienną losową $X$, dla której istnieje taka nieujemna funkcja f($x$), że dla każdego rzeczywistego $x$ zachodzi relacja:

$$F(x) = \int_{-\infty}^{x} f(t)dt. \tag{6.6}$$

Funkcję $f(x)$ spełniająca warunek (6.6) nazywamy gęstością prawdopodobieństwa lub gęstością zmiennej losowej ciągłej.

Funkcja $f(x)$ musi spełniać warunek:

$$\int_{-\infty}^{\infty} f(x) = 1. \tag{6.7}$$

Jeżeli gęstość $f(x)$ jest funkcją ciągłą w punkcie $x$, to zachodzi związek:

$$F'(x) = f(x). \tag{6.8}$$

**Przykład 6.3**

Autobus kursuje co 15 minut i dla tego okresu prawdopodobieństwo przyjazdu na przystanek jest stałe. Czas oczekiwania na autobus rozpatrujemy jako zmienną losową. Należy podać jej funkcję gęstości i dystrybuantę. Funkcję gęstości zapiszemy jako:

$$f(x) = \begin{cases} 0 & x < 0 \\ c & 0 \le x < 15. \\ 0 & x \ge 15 \end{cases}$$

Wartość $c$ wyznaczymy z warunku (6.7), a więc:

$\int_{-\infty}^{\infty} f(x) = \int_{-\infty}^{0} 0\,dx + \int_{0}^{15} c\,dx + \int_{15}^{\infty} 0\,dx = 1$, a zatem:

$$\int_{0}^{15} c\,dx = 1 \qquad cx\Big|_{0}^{15} = 1 \qquad 15c = 1 \qquad c = \frac{1}{15}.$$

Funkcję gęstości możemy zapisać w postaci:

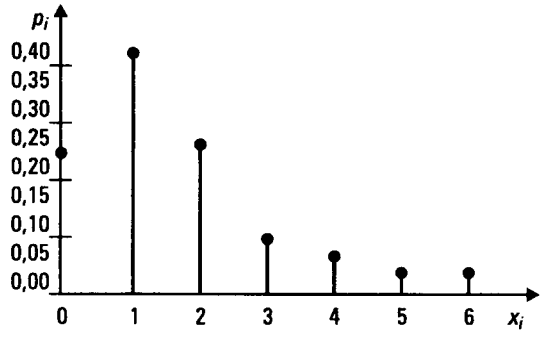
$$f(x) = \begin{cases} 0 & x < 0 \\ \dfrac{1}{15} & 0 \le x < 15. \\ 0 & x \ge 15 \end{cases}$$

Dla wyznaczenia dystrybuanty będziemy rozpatrywać kolejno poszczególne prze-

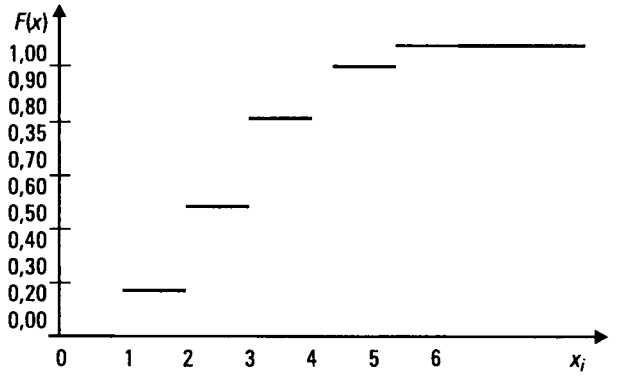
**Tablica 6.3.** Rozkład prawdopodobieństwa i dystrybuanta zmiennej losowej $X$ (liczba dzieci w rodzinie)

| Liczba dzieci w rodzinie ($x_i$) | $P(X = x_i)$ | Liczba dzieci w rodzinie ($X < x$) | $F(x)$ |
|---|---|---|---|
| 0 | 0,2097 | 0 | 0,0000 |
| 1 | 0,3670 | 1 | 0,2097 |
| 2 | 0,2753 | 2 | 0,5767 |
| 3 | 0,1147 | 3 | 0,8520 |
| 4 | 0,0287 | 4 | 0,9667 |
| 5 | 0,0043 | 5 | 0,9953 |
| 6 | 0,0004 | 6 | 0,9996 |
| Suma | 1,0000 | $6 + \varepsilon$ | 1,0000 |

Źródło: dane umowne.




Rys. 6.1. Rozkład prawdopodobieństwa zmiennej losowej typu skokowego




Rys. 6.2. Dystrybuanta zmiennej losowej typu skokowego

działy wartości, jakie może przyjąć zmienna $X$.

1) dla $x \in (-\infty; 0]$

$$F(x) = \int_{-\infty}^{x} f(t)dt = \int_{-\infty}^{x} 0\,dt = 0,$$

2) dla $x \in (0; 15]$

$$F(x) = \int_{-\infty}^{0} 0\,dt + \int_{0}^{x} \frac{1}{15}dt = 0 + \frac{1}{15}t\Big|_{0}^{x} = \frac{1}{15}x,$$

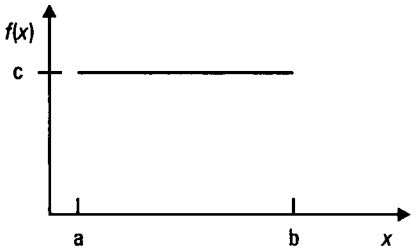
3) dla $x \in (15; \infty)$

$$F(x) = \int_{-\infty}^{0} 0\,dt + \int_{0}^{15} \frac{1}{15}dt + \int_{15}^{+\infty} 0\,dt = 0 + \frac{1}{15}t\Big|_{0}^{15} = \frac{1}{15}\cdot 15 = 1.$$

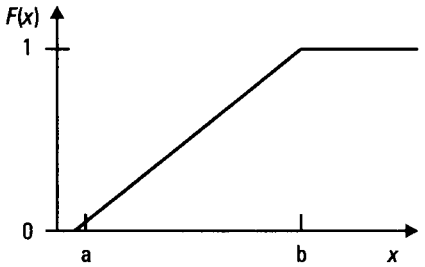
Zapisujemy zatem:

$$F(x) = \begin{cases} 0 & x \le 0 \\ \frac{1}{15}x & 0 < x \le 15 \\ 1 & x > 15 \end{cases}.$$

Funkcję gęstości i dystrybuantę zmiennej losowej, którą jest czas oczekiwania na autobus przedstawiono na rysunkach 6.3 i 6.4.




Rys. 6.3. Funkcja gęstości




Rys. 6.4. Dystrybuanta

## 6.2. Parametry rozkładu prawdopodobieństwa zmiennej losowej

Rozkład prawdopodobieństwa zmiennej losowej możemy ująć sumarycznie za pomocą pewnych wartości liczbowych, które nazywamy parametrami rozkładu. Omówimy tylko niektóre z nich, a mianowicie te, które są używane najczęściej. Należą do nich wartość oczekiwana (nadzieja matematyczna), wariancja i odchylenie standardowe.

**Wartość oczekiwana (nadzieja matematyczna)** jest to wartość przeciętna zmiennej losowej. Oznaczamy ją jako $E(X)$.

Jest ona zdefiniowana odpowiednio:

- dla zmiennej losowej typu skokowego:

$$E(X) = \sum_{i=1}^{n} x_i \cdot p_i, \tag{6.9}$$

- dla zmiennej losowej typu ciągłego:

$$E(X) = \int_{-\infty}^{\infty} x \cdot f(x). \tag{6.10}$$

Dla przykładu 6.2 obliczymy wartość oczekiwaną liczby dzieci w rodzinie. W tym celu wykonamy odpowiednie działania zapisane w tablicy 6.4:

$$E(X) = \sum_{i=1}^{n} x_i \cdot p_i = 1{,}39 \approx 1{,}4 \text{ [dzieci]}.$$

Średnia liczba dzieci w rodzinie wynosi w przybliżeniu 1,4 dzieci.

**Tablica 6.4.** Obliczanie wartości oczekiwanej zmiennej losowej $X$ (liczba dzieci w rodzinie)

| Liczba dzieci w rodzinie ($x_i$) | $P(X = x_i) = p_i$ | $x_i \cdot p_i$ |
|---|---|---|
| 0 | 0,2097 | 0,0000 |
| 1 | 0,3670 | 0,3670 |
| 2 | 0,2753 | 0,5505 |
| 3 | 0,1147 | 0,3441 |
| 4 | 0,0287 | 0,1147 |
| 5 | 0,0043 | 0,0215 |
| 6 | 0,0004 | 0,0022 |
| Suma | 1,0000 | 1,3999 |

Źródło: dane umowne.

Obliczymy średni czas oczekiwania na autobus w warunkach określonych w przykładzie 6.3.

$$E(X)=\int_{-\infty}^{0} x\cdot 0\,dx+\int_{0}^{15}\frac{1}{15}\cdot x\,dx+\int_{15}^{\infty}0\,dx=\frac{1}{15}\cdot\frac{1}{2}x^2\Big|_0^{15}=\frac{1}{30}\cdot(15)^2=\frac{225}{30}=7{,}5\,[\text{min}].$$

Średni czas oczekiwania na autobus jest równy 7,5 minuty.

**Wariancja** jest miarą rozproszenia wartości zmiennej losowej wokół wartości oczekiwanej. Jest to wartość oczekiwana zmiennej $[X - E(X)]^2$ zdefiniowana wzorem:

$$D^2(X)=E[X-E(X)]^2. \tag{6.11}$$

Można wykazać, że

$$D^2(X)=E(X^2)-[E(X)]^2. \tag{6.12}$$

Wariancję obliczamy według odpowiednich wzorów:

- dla zmiennej losowej typu skokowego:

$$D^2(X)=\sum_{i=1}^{n}[x_i-E(X)]^2\cdot p_i, \tag{6.13}$$

- dla zmiennej losowej typu ciągłego:

$$D^2(X)=\int_{-\infty}^{\infty}[x-E(X)]^2\cdot f(x)dx. \tag{6.14}$$

Odchylenie standardowe zmiennej losowej jest dane wzorem:

$$D(X)=\sqrt{D^2(X)}. \tag{6.15}$$

Odchylenie standardowe informuje, o ile średnio wartości zmiennej $X$ różnią się od wartości oczekiwanej E($X$).

Wracamy do przykładów 6.2 i 6.3. Obliczymy najpierw wariancję liczby dzieci w rodzinie (przykład 6.2). Odpowiednie obliczenia zapisano w tablicy 6.5.

Wariancja jest równa 1,1196 [dzieci$^2$]. Obliczamy odchylenie standardowe jako $D(X)=\sqrt{1{,}1196}=1{,}0581\approx 1{,}06$ dzieci. Liczba dzieci w poszczególnych rodzinach różni się od wartości oczekiwanej średnio o 1,06 [dzieci].

Jeśli posłużymy się wzorem (6.12), to otrzymujemy:

$$E(X^2)=\sum_{i=1}^{n}x_i\,p_i=3{,}0794 \quad [E(X)]^2=1{,}9597 \quad D^2(X)=3{,}0794-1{,}9597=1{,}1196.$$

Zgodnie z oczekiwaniami rezultat jest identyczny. Posługując się wzorem (6.12), unikamy uciążliwego obliczania odchyleń wartości zmiennej od wartości oczekiwanej.

Tablica 6.5. Obliczanie wariancji liczby dzieci w rodzinie

| Liczba dzieci w rodzinie $x_i$ | $P(X = x_i) = p_i$ | $X_i - E(X)$ | $[x_i - E(X)]^2$ | $[x_i - E(X)]^2 p_i$ | $x^2 p_i$ |
|---|---|---|---|---|---|
| 0 | 0,2097 | –1,3999 | 1,9597 | 0,4110 | 0,0000 |
| 1 | 0,3670 | –0,3999 | 0,1599 | 0,0587 | 0,3670 |
| 2 | 0,2753 | 0,6001 | 0,3601 | 0,0991 | 1,1010 |
| 3 | 0,1147 | 1,6001 | 2,5603 | 0,2936 | 1,0322 |
| 4 | 0,0287 | 2,6001 | 6,7605 | 0,1938 | 0,4588 |
| 5 | 0,0043 | 3,6001 | 12,9606 | 0,0557 | 0,1075 |
| 6 | 0,0004 | 4,6001 | 21,1608 | 0,0076 | 0,0129 |
| Suma: | 1,0000 | | | 1,1196 | 3,0794 |

Źródło: dane umowne.

Zbadamy teraz zróżnicowanie czasu oczekiwania na autobus w porównaniu z czasem średnim (przykład 6.3). Tym razem posłużymy się wyłącznie wzorem (6.12). Przypomnimy, że funkcja gęstości ma postać:

$$f(x) = \begin{cases} 0 & x < 0 \\ \dfrac{1}{15} & 0 \le x < 15. \\ 0 & x \ge 0 \end{cases}$$

Obliczamy:

$$E(X^2) = \int_{-\infty}^{\infty} x^2 \cdot f(x)\,dx,$$

$$E(X^2) = \int_{-\infty}^{0} 0 \cdot x^2\,dx + \int_{0}^{15} \frac{1}{15} \cdot x^2 + \int_{15}^{\infty} 0 \cdot x^2\,dx = \frac{1}{15} \cdot \frac{1}{3} \cdot x^3 \Big|_0^{15} = \frac{1}{45} \cdot (15)^3 = \frac{3375}{45} = 75,$$

$$D^2(X) = 75 - (7{,}5)^2 = 75 - 56{,}25 = 18{,}75\,[\text{min}^2],\ [\text{min}^2]$$

$$D(X) = \sqrt{18{,}75} = 4{,}33\,[\text{min}].$$

Przychodząc na przystanek autobusowy, musimy się liczyć z tym, że czas oczekiwania na autobus różni się od czasu średniego przeciętnie o 4,33 minuty.

# 6.3. Wybrane rozkłady prawdopodobieństwa zmiennej losowej

## 6.3.1. Podstawowe rozkłady prawdopodobieństwa zmiennej typu skokowego

Rozpoczniemy od prostego przypadku, w którym zmienna losowa przyjmuje tylko dwie wartości $x_1 = 0$ oraz $x_2 = 1$. Tak zdefiniowaną zmienną nazywamy zmienną zerojedynkową. Rozkład prawdopodobieństwa tej zmiennej podano w tablicy 6.6.

**Tablica 6.6.** Rozkład prawdopodobieństwa i parametry zmiennej losowej zerojedynkowej

| $x_k$ | $p_k$ | $x_k \cdot p_k$ | $x_k^2$ | $x_k^2 \cdot p_k$ |
|---|---|---|---|---|
| 0 | $q = 1-p$ | 0 | 0 | 0 |
| 1 | $p = 1-q$ | $p$ | 1 | $p$ |
| Suma | 1 | $p$ | | $p$ |

Wartość oczekiwana: $E(X) = \sum_{k=1}^{n} x_k \cdot p_k = p$ jest równa prawdopodobieństwu, że zmienna losowa przyjmie wartość 1.

Wariancja: $D^2(X) = E(X^2) - [E(X)]^2 = p - p^2 = p \cdot (1-p) = p \cdot q$.

### Rozkład dwumianowy

### Przykład 6.4

Dwie drużyny siatkówki rozgrywają serię spotkań złożoną z czterech meczów. Oznaczymy drużyny jako A i B. Dla drużyny A prawdopodobieństwo wygrania meczu (sukces) jest równe $p = 0{,}6$, a prawdopodobieństwo przegrania (porażka) wynosi $q = 0{,}4$. Traktując liczbę meczów wygranych przez drużynę A jako zmienna losową, podamy rozkład jej prawdopodobieństwa. Najpierw zanalizujemy rozważany przypadek:

1) Zmienna losowa przyjmuje wartości: 0, 1, 2, 3, 4. Są to liczby całkowite nieujemne.
2) Zmienne te możemy traktować jako niezależne, ponieważ wyniki meczu są niezależne.
3) Mecz może zakończyć się zwycięstwem (zmienna przyjmie wartość 1) lub porażką (zmienna przyjmie wartość 0), a zatem w pojedynczym przypadku mamy do czynienia z rozkładem zerojedynkowym.

Rozważana zmienna $X$ jest sumą $n$ – niezależnych zmiennych losowych o rozkładzie zerojedynkowym. Taka zmienna podlega rozkładowi dwumianowemu, który jest także nazywany **rozkładem Bernoulli'ego**. Funkcja rozkładu jest dana wzorem:

$$P(X=k)=\binom{n}{k}\cdot p^k\cdot q^{n-k}, \tag{6.16}$$

gdzie: $\binom{n}{k}=\frac{n!}{k!(n-k)!}$.

Rozkład zmiennej losowej dla przykładu 6.4 podano w tablicy 6.7.

**Tablica 6.7.** Rozkład prawdopodobieństwa liczby meczów wygranych przez drużynę A

| $x_k$ | $P(X = k) = p_i$ | $x_k \cdot p_k$ | $x_k^2 \cdot p_k$ |
|---|---|---|---|
| 0 | $P(X=0)=\binom{4}{0}\cdot(0{,}6)^0\cdot(0{,}4)^4=0{,}0256$ | 0,0000 | 0,0000 |
| 1 | $P(X=1)=\binom{4}{1}\cdot(0{,}6)^1\cdot(0{,}4)^3=0{,}1536$ | 0,1536 | 0,1536 |
| 2 | $P(X=2)=\binom{4}{2}\cdot(0{,}6)^2\cdot(0{,}4)^2=0{,}3456$ | 0,6912 | 1,3824 |
| 4 | $P(X=3)=\binom{4}{3}\cdot(0{,}6)^3\cdot(0{,}4)^1=0{,}3456$ | 1,0368 | 3,1104 |
| 5 | $P(X=4)=\binom{4}{4}\cdot(0{,}6)^4\cdot(0{,}4)^0=0{,}1296$ | 0,5184 | 2,0736 |
| Suma | | 2,4 | 6,7200 |

Źródło: obliczenia własne.

Obliczymy teraz średnią liczbę wygranych meczów, wariancję i odchylenie standardowe:

- $E(X)=\sum_{i=1}^{n} x_i\cdot p_i=2{,}4,$
- $D^2(X)=E(X^2)-[E(X)]^2=6{,}72-(2{,}4)^2=6{,}72-5.76=0{,}96,$
- $D(X)=\sqrt{0{,}96}\approx 0{,}98.$

Drużyna A może oczekiwać, że w seriach po 4 spotkania z drużyną B wygra średnio 2,4 meczów]. Liczba wygranych meczów w poszczególnych seriach różni się od średniej przeciętnie o 0,98 meczu, czyli o około 1 mecz.

W tym miejscu należy zwrócić uwagę na to, że jeśli zmienna ma rozkład dwumianowy, to dla ustalenia wartości oczekiwanej i wariancji nie musimy znać całego rozkładu. Wartości tych parametrów możemy obliczyć na podstawie następujących wzorów[62]:

$$E(X) = n \cdot p. \tag{6.17}$$

$$D^2(X) = n \cdot p \cdot q. \tag{6.18}$$

Dla przykładu 6.4 mamy:

$$E(X) = 4 \cdot 0{,}6 = 2{,}4, \qquad D^2(X) = 4 \cdot 0{,}6 \cdot 0{,}4 = 0{,}96.$$

Wyniki są więc identyczne.

## 6.3.2. Wybrane rozkłady zmiennej losowej typu ciągłego

Spośród rozkładów zmiennej losowej typu ciągłego omówimy tylko dwa, a mianowicie:

1) rozkład jednostajny (prostokątny),
2) rozkład normalny.

**Rozkład jednostajny (prostokątny)** został wprowadzony praktycznie w przykładzie 6.3, w którym jako zmienną losową rozważaliśmy czas oczekiwania na autobus. Tutaj podamy ogólną jego postać. Zmienna losowa podlega rozkładowi jednostajnemu, jeśli jej funkcja gęstości jest dana jako:

$$f(x) = \begin{cases} 0 & x < a \\ \dfrac{1}{b-a} & a \le x < b. \\ 0 & x \ge b \end{cases} \tag{6.19}$$

Dystrybuanta ma postać:

$$F(x) = \begin{cases} 0 & x < a \\ \dfrac{x-a}{b-a} & a \le x < b. \\ 1 & x \ge b \end{cases} \tag{6.20}$$

[62] Wyprowadzenie wzorów można znaleźć w podręcznikach do rachunku prawdopodobieństwa.

Wartość oczekiwana jest równa:

$$E(X)=\frac{a+b}{2}. \tag{6.21}$$

Wariancja jest dana wzorem:

$$D^2(X)=\frac{(b-a)^2}{12}. \tag{6.22}$$

Średni czas oczekiwania na autobus z przykładu 6.3 jest równy:

$$E(X)=\frac{0+15}{2}=7{,}5[\text{min}].$$

Wariancja jest równa:

$$D^2(X)=\frac{(15-0)^2}{12}=\frac{225}{12}=18{,}75[\text{min}^2].$$

Odchylenie standardowe $D(X)=\sqrt{18{,}75}=4{,}33[\text{min}]$.

Rezultaty obliczeń są identyczne jak te, które otrzymaliśmy według wzorów definicyjnych.

**Rozkład normalny**

Zmienna losowa podlega rozkładowi normalnemu, jeśli funkcja gęstości jest dana wzorem:

$$f(x)=\frac{1}{\sigma\sqrt{2\pi}}\cdot\exp\left[-\frac{1}{2}\cdot\frac{(x-\mu)^2}{\sigma^2}\right], \tag{6.23}$$

gdzie:
$\mu = E(X)$ – wartość oczekiwana zmiennej losowej $X$,
$\sigma^2 = D^2(X)$ – wariancja zmiennej losowej $X$,
$\sigma = D(X)$ – odchylenie standardowe zmiennej losowej $X$.

Dystrybuanta jest zdefiniowana jako:

$$F(x)=\frac{1}{\sigma\sqrt{2\pi}}\int_{-\infty}^{x}\exp\left[-\frac{1}{2}\cdot\frac{(x-\mu)^2}{\sigma^2}\right]dx. \tag{6.24}$$

Rozkład normalny jest w pełni określony przez dwa parametry: wartość oczekiwaną ($\sigma$) oraz przez odchylenie standardowe ($\mu$; $\sigma$). W dalszym ciągu używać będziemy zapisu $N(\mu;\sigma)$, który czytamy jako: zmienna $X$ podlega rozkładowi normalnemu z parametrami $E(X)=\mu$; $D(X)=\sigma$. Sformułujemy teraz podstawowe własności funkcji gęstości danej wzorem (6.23).

1. Jest to rozkład zmiennej losowej typu ciągłego, która przyjmuje wartości z przedziału $(-\infty; \infty)$.
2. Jest to rozkład symetryczny, a osią symetrii jest wartość przeciętna $E(X) = \mu$. Wynika stąd, że $P(X \leq \mu) = P(X \geq \mu) = 0{,}5$.
3. Posiada dwa punkty przegięcia o odciętych $x_1 = \mu - \sigma$ oraz $x_2 = \mu + \sigma$.
4. Wartość maksymalną przyjmuje w punkcie $\mu$ i jest ona równa $f(x) = \frac{1}{\sigma\sqrt{2\pi}}$.
5. Wartość $\mu$ określa położenie rozkładu w układzie współrzędnych, a odchylenie standardowe $\sigma$ jego spłaszczenie.

Funkcja gęstości rozkładu normalnego nie należy do funkcji elementarnych, a więc jej analityczna postać jest dość skomplikowana. Dlatego obliczanie na przykład prawdopodobieństwa $P(x_1 \leq X < x_2)$ wymaga obliczenia odpowiedniej całki z funkcji gęstości danej wzorem (6.24). Dla uniknięcia tej niedogodności wprowadzamy zmienną standaryzowaną U zdefiniowaną jako:

$$U = \frac{X - E(X)}{D(X)} = \frac{X - \mu}{\sigma}. \tag{6.25}$$

Zmienna ta ma stałe parametry, a mianowicie wartość przeciętna E(U) = 0, a wariancja $D^2(U) = 1$.

Biorąc pod uwagę, że:

1) wartość przeciętna stałej C jest równa E(C) = 0,
2) wartość oczekiwana różnicy zmiennych jest równa różnicy wartości oczekiwanych:
$$E(X-Y) = E(X) - E(Y),$$
3) $E[X - E(X)] = 0$,
4) wariancja stałej jest równa zeru; $D^2(C) = 0$,
5) wariancja iloczynu stałej i zmiennej jest równa: $D^2(CX) = C^2 \cdot D^2(X)$, otrzymujemy:

$$E(U) = E\left(\frac{X - E(X)}{D(X)}\right) = \frac{1}{D(X)} \cdot \{E[(X - E(X)]\} = \frac{1}{D(X)} \cdot [E(X) - E(X)] = 0,$$

$$D^2(U) = D^2\left(\frac{X - E(X)}{D(X)}\right) = \frac{1}{D^2(X)} \cdot D^2[X - E(X)] = \frac{1}{D^2(X)} \cdot D^2(X) = 1.$$

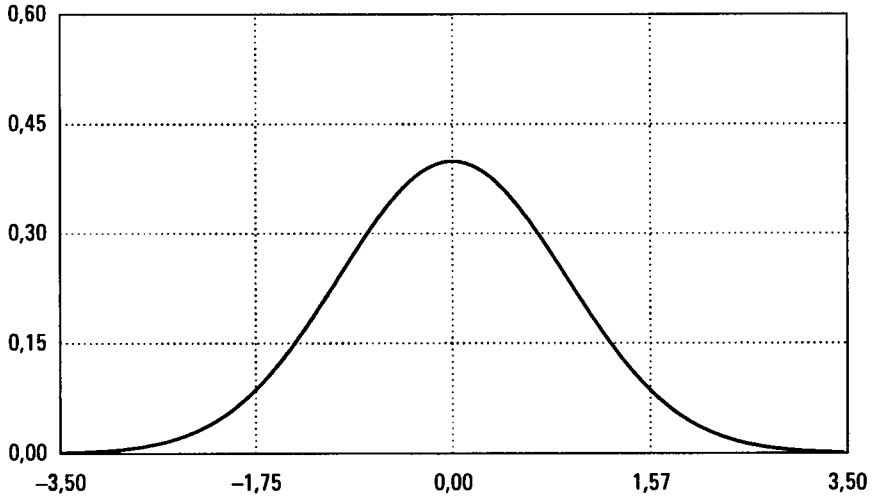
Funkcja gęstości standaryzowanej zmiennej U jest dana jako:

$$\varphi(u) = \frac{1}{\sqrt{2\pi}} \cdot \exp\left[\frac{u^2}{2}\right]. \tag{6.26}$$

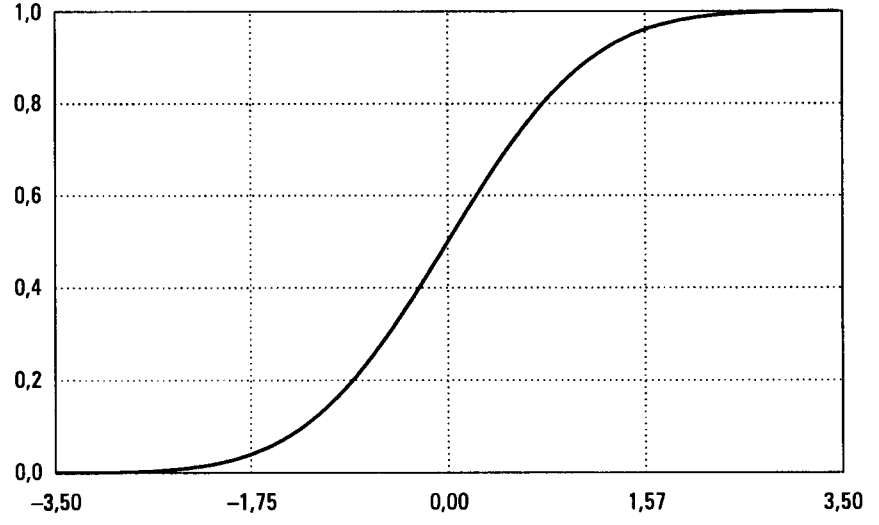
Dystrybuanta ma postać:

$$\Phi(u) = \frac{1}{\sqrt{2\pi}} \cdot \int_{-\infty}^{u} \varphi(u)\,du. \tag{6.27}$$

Na rysunku 6.5 przedstawiono funkcję gęstości, a na rysunku 6.6 dystrybuantę rozkładu normalnego dla zmiennej standaryzowanej U. Rozkład normalny zmiennej standaryzowanej ma te same własności, co rozkład zmiennej $X$.




Rys. 6.5. Funkcja gęstości rozkładu normalnego zmiennej standaryzowanej N(0; 1)




Rys. 6.6. Dystrybuanta zmiennej losowej standaryzowanej

Wartości funkcji gęstości i dystrybuanty zmiennej standaryzowanej zależą tylko od wartości zmiennej U. Opracowano więc tablice, z których korzystamy, gdy interesuje nas prawdopodobieństwo, że zmienna losowa przyjmie wartość z danego przedziału. Tablica 6.8 zawiera fragment tablicy dystrybuanty rozkładu normalnego zdefiniowanej wzorem (6.27).

### Przykład 6.5

Wzrost mężczyzn jest zmienną losową o rozkładzie normalnym z parametrami $\mu = 170$ cm i odchyleniem standardowym $\sigma = 3$ cm.

- Obliczyć prawdopodobieństwo, że losowo wybrany mężczyzna ma wzrost niższy niż 184 cm. Mamy zatem znaleźć: $P(X<184)$. Skorzystamy z tablic rozkładu normalnego. W tym celu zmienną $X$ przekształcimy w zmienną U według wzoru (6.24):

$$u = \frac{185-170}{3} = \frac{5}{3} = 1{,}67.$$

  Szukane prawdopodobieństwo znajdujemy na skrzyżowaniu wiersza 1,6 i kolumny 0,07, gdzie odczytujemy, że $P(U < 1{,}67) = P(X < 184) = 0{,}9525$.
- Jakie jest prawdopodobieństwo, że losowo wybrany mężczyzna jest niższego wzrostu niż 168 [cm]?

$$P(X<168) = P\left(U < \frac{168-170}{3}\right) = P(U < -0{,}67) = 0{,}2514.$$

  Prawdopodobieństwo to odczytujemy na przecięciu wiersza –0,60 z kolumną –0,07.
- Obliczyć prawdopodobieństwo, że losowo wybrany mężczyzna ma przynajmniej 176 cm wzrostu.

$$P(X \geq 176) = P\left(U \geq \frac{176-170}{3}\right) =$$

$$= P(U \geq 2) = 1 - P(U < 2) = 1 - 0{,}9767 = 0{,}0233.$$

- Znaleźć prawdopodobieństwo, że losowo wybrany mężczyzna ma przynajmniej 165 cm wzrostu.

$$P(X \geq 165) = P\left(U \geq \frac{165-170}{3}\right) = P(U \geq -1{,}67) =$$

$$= 1 - P(U < -1{,}67) = 1 - 0{,}9525 = 0{,}0475.$$

  Z tablicy dystrybuanty możemy odczytać $P(U < 2)$. Skorzystajmy więc z relacji: $P(U < 2) = 1 - P(U \leq 2)$.

- Jakie jest prawdopodobieństwo, że wzrost losowo wybranego mężczyzny wynosi przynajmniej 167 cm, ale nie przekracza 173 cm?

$$P(167 \le X < 173) = P\left(\frac{167-170}{3} \le U < \frac{173-170}{3}\right) = P(-1 \le U < 1),$$

$$P(167 \le X < 173) = P(-1 \le U < 1) = 0{,}6827.$$

Przykład 6.6

Dom Mody postanawia przygotować garnitury dla mężczyzn o nietypowych wymiarach. Do takich zaliczono osoby o wzroście co najwyżej 160 [cm] i przynajmniej 205 [cm]. Należy ustalić, jaki procent klientów mieści się w tych przedziałach wzrostu. Znajdujemy zatem:

- $P(X \le 160) = P\left(U \le \frac{160-170}{3}\right) = P(U \le -3) = 0{,}0013,$
- $P(X > 180) = P\left(U > \frac{180-170}{3}\right) = P(U > 3) = 0{,}0013.$

Wyrażając prawdopodobieństwa w procentach, otrzymujemy, że na obydwa przedziały wzrostu przypada 0,26% populacji.

Otrzymane wyniki możemy wykorzystać do oszacowania zapotrzebowania na garnitury dla mężczyzn uznanych za nietypowych.

Według danych Narodowego Spisu Ludności i Mieszkań w 2002 roku w Polsce zarejestrowano 14962106 mężczyzn w wieku przynajmniej 15 lat. Gdyby populacja ta podlegała rozkładowi normalnemu z parametrami $\mu = 170$ [cm] i odchyleniem standardowym $\sigma = 3$ [cm], to można oszacować, że wśród nich jest $14962106 \cdot 0{,}0026 = 38901{,}48 \approx 38901$ osób, które na podstawie przyjętego kryterium można zaliczyć do nietypowych. Jest to populacja potencjalnych nabywców przygotowywanej kolekcji.

**Tablica 6.8.** Dystrybuanta rozkładu normalnego

| u | –0,01 | –0,03 | –0,05 | –0,07 | –0,09 | –0,11 | 0 | 0,01 | 0,03 | 0,05 | 0,07 | 0,09 | 0,11 |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| **–3,0** | 0,0013 | 0,0012 | 0,0011 | 0,0011 | 0,0010 | 0,0009 | 0,0013 | 0,0014 | 0,0015 | 0,0016 | 0,0017 | 0,0018 | 0,0019 |
| **–2,8** | 0,0025 | 0,0023 | 0,0022 | 0,0021 | 0,0019 | 0,0018 | 0,0026 | 0,0026 | 0,0028 | 0,0030 | 0,0032 | 0,0034 | 0,0036 |
| **–2,6** | 0,0045 | 0,0043 | 0,0040 | 0,0038 | 0,0036 | 0,0034 | 0,0047 | 0,0048 | 0,0051 | 0,0054 | 0,0057 | 0,0060 | 0,0064 |
| **–2,2** | 0,0136 | 0,0129 | 0,0122 | 0,0116 | 0,0110 | 0,0104 | 0,0139 | 0,0143 | 0,0150 | 0,0158 | 0,0166 | 0,0174 | 0,0183 |
| **–2,0** | 0,0222 | 0,0212 | 0,0202 | 0,0192 | 0,0183 | 0,0174 | 0,0228 | 0,0233 | 0,0244 | 0,0256 | 0,0268 | 0,0281 | 0,0294 |
| **–1,8** | 0,0351 | 0,0336 | 0,0322 | 0,0307 | 0,0294 | 0,0281 | 0,0359 | 0,0367 | 0,0384 | 0,0401 | 0,0418 | 0,0436 | 0,0455 |
| **–1,6** | 0,0537 | 0,0516 | 0,0495 | 0,0475 | 0,0455 | 0,0436 | 0,0548 | 0,0559 | 0,0582 | 0,0606 | 0,0630 | 0,0655 | 0,0681 |
| **–1,4** | 0,0793 | 0,0764 | 0,0735 | 0,0708 | 0,0681 | 0,0655 | 0,0808 | 0,0823 | 0,0853 | 0,0885 | 0,0918 | 0,0951 | 0,0985 |
| **–1,2** | 0,1131 | 0,1093 | 0,1056 | 0,1020 | 0,0985 | 0,0951 | 0,1151 | 0,1170 | 0,1210 | 0,1251 | 0,1292 | 0,1335 | 0,1379 |
| **–1,0** | 0,1562 | 0,1515 | 0,1469 | 0,1423 | 0,1379 | 0,1335 | 0,1587 | 0,1611 | 0,1660 | 0,1711 | 0,1762 | 0,1814 | 0,1867 |
| **–0,6** | 0,2709 | 0,2643 | 0,2578 | 0,2514 | 0,2451 | 0,2389 | 0,2743 | 0,2776 | 0,2843 | 0,2912 | 0,2981 | 0,3050 | 0,3121 |
| **–0,4** | 0,3409 | 0,3336 | 0,3264 | 0,3192 | 0,3121 | 0,3050 | 0,3446 | 0,3483 | 0,3557 | 0,3632 | 0,3707 | 0,3783 | 0,3859 |
| **0,0** | 0,4960 | 0,4880 | 0,4801 | 0,4721 | 0,4641 | 0,4562 | 0,5000 | 0,5040 | 0,5120 | 0,5199 | 0,5279 | 0,5359 | 0,5438 |
| **0,2** | 0,5753 | 0,5675 | 0,5596 | 0,5517 | 0,5438 | 0,5359 | 0,5793 | 0,5832 | 0,5910 | 0,5987 | 0,6064 | 0,6141 | 0,6217 |
| **0,4** | 0,6517 | 0,6443 | 0,6368 | 0,6293 | 0,6217 | 0,6141 | 0,6554 | 0,6591 | 0,6664 | 0,6736 | 0,6808 | 0,6879 | 0,6950 |
| **0,6** | 0,7224 | 0,7157 | 0,7088 | 0,7019 | 0,6950 | 0,6879 | 0,7257 | 0,7291 | 0,7357 | 0,7422 | 0,7486 | 0,7549 | 0,7611 |
| **1,0** | 0,8389 | 0,8340 | 0,8289 | 0,8238 | 0,8186 | 0,8133 | 0,8413 | 0,8438 | 0,8485 | 0,8531 | 0,8577 | 0,8621 | 0,8665 |
| **1,2** | 0,8830 | 0,8790 | 0,8749 | 0,8708 | 0,8665 | 0,8621 | 0,8849 | 0,8869 | 0,8907 | 0,8944 | 0,8980 | 0,9015 | 0,9049 |

**Tablica 6.8.** ciąg dalszy

| u | –0,01 | –0,03 | –0,05 | –0,07 | –0,09 | –0,11 | 0 | 0,01 | 0,03 | 0,05 | 0,07 | 0,09 | 0,11 |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| **1,4** | 0,9177 | 0,9147 | 0,9115 | 0,9082 | 0,9049 | 0,9015 | 0,9192 | 0,9207 | 0,9236 | 0,9265 | 0,9292 | 0,9319 | 0,9345 |
| **1,6** | 0,9441 | 0,9418 | 0,9394 | 0,9370 | 0,9345 | 0,9319 | 0,9452 | 0,9463 | 0,9484 | 0,9505 | 0,9525 | 0,9545 | 0,9564 |
| **1,8** | 0,9633 | 0,9616 | 0,9599 | 0,9582 | 0,9564 | 0,9545 | 0,9641 | 0,9649 | 0,9664 | 0,9678 | 0,9693 | 0,9706 | 0,9719 |
| **2,0** | 0,9767 | 0,9756 | 0,9744 | 0,9732 | 0,9719 | 0,9706 | 0,9772 | 0,9778 | 0,9788 | 0,9798 | 0,9808 | 0,9817 | 0,9826 |
| **2,2** | 0,9857 | 0,9850 | 0,9842 | 0,9834 | 0,9826 | 0,9817 | 0,9861 | 0,9864 | 0,9871 | 0,9878 | 0,9884 | 0,9890 | 0,9896 |
| **2,8** | 0,9974 | 0,9972 | 0,9970 | 0,9968 | 0,9966 | 0,9964 | 0,9974 | 0,9975 | 0,9977 | 0,9978 | 0,9979 | 0,9981 | 0,9982 |
| **3,0** | 0,9986 | 0,9985 | 0,9984 | 0,9983 | 0,9982 | 0,9981 | 0,9987 | 0,9987 | 0,9988 | 0,9989 | 0,9989 | 0,9990 | 0,9991 |

# Podstawy wnioskowania statystycznego

Rozdział 7

## 7.1. Podstawowe statystyki z próby i ich rozkłady

W punkcie 1.2 wprowadzono pojęcie zbiorowości (populacji) generalnej w sposób intuicyjny. Pojęcie to można również definiować formalnie[63]. Według M. Fisza: „W statystyce przyjęto nazywać zbiorowość, której elementy obserwujemy, populacją generalną lub zbiorowością generalną. Elementy populacji możemy badać ze względu na różne cechy. Jeżeli mówimy, że populacja generalna ma rozkład $F(x)$, to chcemy przez to powiedzieć, że badamy cechę $X$ elementów tej populacji generalnej i że ta cecha $X$ jest zmienną losową o dystrybuancie $F(x)$. Zespół pewnej części elementów populacji generalnej przyjęto w statystyce nazywać próbą. My będziemy nazywali próbą ciąg wartości badanej cechy pewnej ilości elementów populacji generalnej"[64].

W dalszych rozważaniach będziemy wnioskować o cechach populacji generalnej na podstawie próby wybranej w drodze losowego doboru. „Metoda wyboru jest losowa, jeżeli cecha stanowiąca kryterium wyboru jest niezależna od cechy badanej"[65].

**Przykład 7.1**

Należy określić strukturę zatrudnionych według poziomu wykształcenia w Krakowie. Jako próbę traktujemy ogół pracowników zatrudnionych w wybranej uczelni wyższej. Taka próba nie może stanowić podstawy do wnioskowania o strukturze zatrudnionych według wykształcenia w całym Krakowie. W uczelniach wyższych przeważają bowiem osoby posiadające wykształcenie wyższe. Kryterium wyboru (uczelnia wyższa) jest zależne od badanej cechy (struktura według wykształcenia). Jeśli natomiast interesować nas będzie odległość miejsca zamieszkania od miejsca pracy,

---

[63] Por. np.: A. Iwasiewicz., Z. Paszek, *Statystyka z elementami statystycznych metod monitorowania procesów*, Kraków 2004.

[64] M. Fisz, *op. cit.*, s. 351–352.

[65] *Ibidem*, s. 522.

to taką próbę moglibyśmy potraktować jako wybraną w sposób losowy. „Stąd wynika, że pewna metoda wyboru może być losowa względem jednej cechy i nie być losowa względem innej"[66].

Do zrozumienia metod wnioskowania statystycznego konieczne jest uświadomienie sobie, że próba jest punktem w $n$-wymiarowej przestrzeni. Zilustrujemy to następującym przykładem.

**Przykład 7.2**

Populację generalną stanowią studenci wydziału $F$ pewnej uczelni w roku akademickim t. Interesującą nas cechą są oceny otrzymane przez studentów w roku akademickim t. Cechę tę oznaczamy jako $X$. Może ona przyjąć sześć następujących wartości: 2,0; 3,0; 3,5; 4,0; 4,5; 5,0. Takie bowiem oceny mogą uzyskiwać studenci na egzaminach. W uczelnianej bazie danych każdy student posiada swój dokument, w którym zapisane są wszystkie jego oceny. Zbiorowość generalną stanowi zbiór wszystkich dokumentów w bazie danych. Ich liczebność jest równa liczbie wszystkich studentów oznaczonej jako N.

Z tej zbiorowości pobieramy losowo ze zwracaniem $n = 10$ dokumentów. Zaobserwowane wartości zmiennej losowej $X$, czyli zespół liczb $(x_1, x_2, \ldots, x_{10})$ są wartościami 10-wymiarowej zmiennej losowej. Przestrzeń prób składa się ze wszystkich możliwych podzbiorów liczb $(x_1, x_2, \ldots, x_{10})$. Liczebność przestrzeni prób ustalamy jako wariacje z powtórzeniami. Jest ich $\bar{V}_k^n = k^n$. Rozważana w przykładzie przestrzeń składa się więc z $6^{10}$ punktów. Jest to ilość prób jaką możemy wylosować z rozważanej populacji.

Teraz określimy ważne w statystyce matematycznej pojęcie, jakim jest statystyka z próby. **Statystyka z próby** jest to zmienna losowa będącą funkcją obserwowanej łącznej zmiennej losowej $(X_1, X_2, \ldots X_n)$.

**Przykład 7.3**

Nawiązujemy do przykładu 7.2. Chcemy ustalić średnią ocen uzyskanych przez studentów Wydziału $F$ w roku akademickim $t$. Średnia zdefiniowana wzorem:

$$\bar{X} = \frac{X_1 + X_2 + \ldots + X_n}{n} = \frac{\sum_{i=1}^{n} X_i}{n} \tag{7.1}$$

jest statystyką z próby.

Jako zmienna losowa statystyka z próby posiada rozkład prawdopodobieństwa. Możemy zatem rozważać:

1) jaki jest dokładny rozkład statystyki $Z_n = Z(X_1, X_2, \ldots, X_n)$, co oznacza, że szukamy rozkładu tej statystyki dla każdego naturalnego $n$,

[66] *Ibidem*.

2) jaki jest rozkład graniczny statystyki $Z_n$, to znaczy, że interesuje nas rozkład tej statystyki gdy $n \to \infty$.

Rozkłady dokładne mają szczególne znaczenie w przypadku małych prób, a rozkłady graniczne rozważamy wówczas, gdy dysponujemy dużą próbą statystyczną. Kryterium zaliczenia próby do dużych lub małych zależy od rozważanych statystyk z próby. Nie istnieje natomiast kryterium ogólne wyróżniania prób małych i dużych[67].

Przedstawimy teraz rozkłady wybranych statystyk z próby, a mianowicie tych, które będą później wykorzystane do wnioskowania statystycznego.

### Rozkład średniej arytmetycznej niezależnych zmiennych losowych o rozkładach normalnych

Zajmiemy się najpierw sytuacją, w której zmienna $X$ w populacji generalnej ma rozkład normalny określony wzorem:

$$f(x) = \frac{1}{\sigma\sqrt{2\pi}} \exp\left[-\frac{1}{2} \cdot \frac{(x-\mu)^2}{\sigma^2}\right], \tag{7.2}$$

gdzie:

$\mu$ – wartość przeciętna,

$\sigma$ – odchylenie standardowe zmiennej losowej $X$.

Rozpatrujemy średnią z próby określoną wzorem:

$$\bar{X} = \frac{1}{n} \sum_{k=1}^{n} X_k, \tag{7.3}$$

gdzie zmienne $X_k$ są niezależne i mają jednakowy rozkład określony wzorem (7.2). Metodą funkcji charakterystycznych[68] znaleziono rozkład tej zmiennej dany wzorem:

$$f_1(\bar{x}) = \frac{1}{\frac{\sigma}{\sqrt{n}}\sqrt{2\pi}} \exp\left[-\frac{1}{2} \cdot \frac{(x-\mu)^2}{\frac{\sigma^2}{n}}\right]. \tag{7.4}$$

Zmienna $\bar{X}$ ma tę samą wartość przeciętną ($\mu$), co zmienna $X$. Wariancja zmiennej $\bar{X}$ jest $n$ razy mniejsza od wariancji zmiennej $X$.

Jest ona równa:

$$\sigma_{\bar{x}}^2 = \frac{\sigma^2}{n}. \tag{7.5}$$

---

[67] Por.: A. Iwasiewicz, Z. Paszek, *op. cit.*

[68] Por. np.: M. Fisz, *op. cit.*, s. 354.

Odchylenie standardowe jest równe zmiennej $\bar{X}$ i jest równe:

$$\sigma_{\bar{x}} = \frac{\sigma}{\sqrt{n}}. \tag{7.6}$$

Oznacza to, że rozproszenie wartości zmiennej $\bar{X}$ wokół wartości przeciętnej $\mu$ jest mniejsze niż zmiennej $X$ wokół tej samej wartości.

**Przykład 7.4**

Zmienną $X$ jest ciężar czekoladek produkowanych przez automat. Zmienna ta podlega rozkładowi normalnemu z parametrami $\mu = 10$ [g] i $\sigma = 2$ [g], co zapiszemy krótko N(10; 2).

Rozpatrujemy średni ciężar czekoladek w próbach pobranych w sposób losowy i liczących po $n = 16$ czekoladek.:

a) obliczymy prawdopodobieństwo, że ciężar losowo wybranej czekoladki różni się od ciężaru średniego nie więcej niż o 2 [g]:

$$P\{8 \le X < 12\} = P\left\{\frac{8-10}{2} \le \frac{X-10}{2} < \frac{12-10}{2}\right\} = P\left\{-1 \le \frac{X-10}{2} < 1\right\} = 0{,}68266.$$

Z tablicy 6.6 odczytujemy w wierszu „–1,0" i kolumnie „0,00" oraz w wierszu „1,0" i w kolumnie 0,00, odpowiednio prawdopodobieństwa: 0,83891 i 0,15625. Prawdopodobieństwa te odejmujemy i otrzymujemy podany wyżej wynik. Prawdopodobieństwo wylosowania czekoladki, której ciężar różni się od średniego ciężaru czekoladek nie więcej niż o 2 [g] jest równe 0,68266.

b) znajdziemy rozkład zmiennej losowej $\bar{X}$:

$$f_1(\bar{x}) = \frac{1}{\frac{2}{\sqrt{16}}\sqrt{2\pi}} \exp\left[-\frac{(x-10)^2}{\frac{2\cdot 4}{16}}\right] = \frac{1}{0{,}5\sqrt{2\pi}} \exp\left[-\frac{(x-10)^2}{0{,}5}\right].$$

Wariancja zmiennej $\bar{X}$ jest równa $\sigma^2 = 0{,}25$ [g$^2$], a odchylenie standardowe $\sigma = 0{,}5$ [g].

c) dla wylosowanego ciągu 16-elementowych prób niezależnych obliczymy prawdopodobieństwo, że uzyskany na ich podstawie średni ciężar różni się od średniego ciężaru czekoladek w populacji generalnej nie więcej niż o 2 [g]:

$$P\{8 \le \bar{X} < 12\} = P\left\{\frac{8-10}{0{,}5} \le \frac{\bar{X}-10}{0{,}5} < \frac{12-10}{0{,}5}\right\} = P\left\{-4 \le \frac{\bar{X}-10}{0{,}5} < 4\right\} = 0{,}99997.$$

W tym przypadku prawdopodobieństwo, że średni ciężar czekoladek w wybranej losowo próbie różni się od średniego ich ciężaru w populacji generalnej nie więcej niż o 2 [g] jest równe 0,99997.

Uzyskane rezultaty wskazują, że w dużej serii 16-elementowych prób prostych 99997 razy na 100000 otrzymamy takie wartości zmiennej $\bar{X}$, które będą różnić się od wartości $\mu$ nie więcej niż o 2 [g]. Gdybyśmy do szacowania średniego ciężaru czekoladek użyli prób jednoelementowych, a więc wielokrotnie losowalibyśmy po jednym elemencie, to tylko 68266 razy na 10000 otrzymamy wartości $X$ różniące się od wartości przeciętnej w populacji generalnej nie więcej niż o 2 [g].

### Rozkład $\chi^2$ (Chi$^2$)

Interesuje nas rozkład $n$ niezależnych zmiennych losowych $X_k$ (k = 1, 2, ..., $n$) o jednakowym rozkładzie normalnym z wartością przeciętną $\mu = 0$ i odchyleniem standardowym $\sigma$, a więc o gęstości prawdopodobieństwa danej wzorem:

$$f(x) = \frac{1}{\sigma\sqrt{2\pi}} \exp\left[\frac{x^2}{2 \cdot \sigma^2}\right]. \tag{7.7}$$

Rozpatrujemy statystykę, która jest sumą kwadratów $n$ niezależnych zmiennych losowych $X$, z których każda ma rozkład normalny o gęstości danej wzorem (7.7). Jest to statystyka $\chi^2$ zdefiniowana jako:

$$\chi^2 = \sum_{k=1}^{n} X_k^2. \tag{7.8}$$

Wartość przeciętna tej zmiennej jest równa:

$$m_1 = n \cdot \sigma^2, \tag{7.9}$$

wariancja jest równa:

$$\mu_2 = 2 \cdot n \cdot \sigma^4. \tag{7.10}$$

Parametr $n$ nosi nazwę liczby stopni swobody.

### Przykład 7.5

Zmienne losowe $X_k$ ($k$ = 1, 2, ..., 16) są niezależne i każda z nich ma jednakowy rozkład N(0; 2). Rozpatrujemy statystykę:

$$\chi^2 = \sum_{k=1}^{16} X_k^2.$$

Zmienna ta ma 16 stopni swobody oraz wartość przeciętną:

$$m_1 = 16 \cdot 2^2 = 64,$$

wariancję równą:

$$\mu_2 = 2 \cdot 16 \cdot 2^4 = 32 \cdot 16 = 512,$$

odchylenie standardowe równe:

$$\sqrt{\mu_2} = \sqrt{512} = 22{,}63.$$

**Rozkład Studenta**

Rozważamy zmienną losową $\bar{X}$ zdefiniowaną jako:

$$\bar{X} = \frac{1}{n}\sum_{k=1}^{n} X_k,$$

gdzie:
zmienne losowe $X_k$ są niezależne i mają jednakowy rozkład normalny $N(\mu; \sigma)$. Zmienna losowa $\bar{X}$ ma rozkład normalny dany wzorem (7.4). Rozkład ten jest określony, jeśli znamy $\mu$ i $\sigma$.

Jeśli znamy tylko $\mu$, a nie znamy $\sigma$, to rozkład zmiennej losowej $\bar{X}$ jest nieznany[69]. Dlatego musimy rozważyć taką statystykę, która, będąc funkcją $\mu$, nie zależy od $\sigma$. Statystykę taką wprowadził Gosset, który bardziej znany jest pod pseudonimem Student. Wykazał on, że jeśli zmienne $X_k$ są niezależnymi zmiennymi losowymi mającymi ten sam rozkład normalny $N(\mu; \sigma)$, a zmienna $\bar{X}$ jest określona jako:

$$\bar{X} = \frac{1}{n}\sum_{k=1}^{n} X_k,$$

zaś S jest dane jako:

$$S = \frac{1}{n}\sqrt{\sum_{k=1}^{n}(X_k - \bar{X})^2}, \tag{7.12}$$

to statystyka dana wzorem:

$$t = \frac{\bar{X} - \mu}{S}\sqrt{n-1}, \tag{7.13}$$

nie zależy od $\sigma$.

Zmienna określoną wzorem (7.13) jest zmienną losową o rozkładzie $t$ – Studenta o $n$-1 stopniach swobody.

## 7.2. Estymacja jako metoda indukcyjnego wnioskowania statystycznego

### 7.2.1. Estymatory i ich własności

Estymacja jest to metoda postępowania prowadząca do oszacowania nieznanych wartości parametrów na podstawie wyników zaobserwowanych w próbie. Uzyskanie szacunków nieznanych wartości parametrów przebiega w następujących etapach:

1) ustalenie parametru, którego wartość należy oszacować;

[69] W miejsce nie możemy wstawić wartości s obliczonej na podstawie próby, ponieważ jest to realizacja zmiennej losowej $S$.

2) określenie modelu populacji generalnej;
3) scharakteryzowanie próby statystycznej;
4) wybór odpowiedniej statystyki z próby, której rozkład zależy od interesującego nas parametru;
5) obliczenie wartości tej statystyki dla wylosowanej próby statystycznej.

W tym miejscu możliwe są dwie metody postępowania. Jedną z nich jest to metoda **estymacji punktowej**, która polega na tym, że obliczoną dla próby wartość statystyki przyjmujemy za oszacowanie parametru. Drugą procedurą jest metoda **estymacji przedziałowej**, polegajaca na skonstruowaniu przedziału, który z zadanym z góry prawdopodobieństwem pokrywa szacowaną wartość parametru.

W obydwu przypadkach posługujemy się pewnymi statystykami, które nazywamy estymatorami. Wprowadzamy następujące oznaczenia:

$Q$ – parametr, którego wartość chcemy oszacować,
$\hat{Q}_n$ – estymator parametru $Q$,
$Q^*$ – oszacowana wartość (ocena parametru $Q$).

Estymatorem $\hat{Q}_n$ parametru $Q$ nazywamy statystykę, której rozkład zależy od parametru $Q$ i która spełnia następujący warunek:

$$\lim_{n\to\infty}\left(P\left|\hat{Q}_n - Q\right| \geq c\right) = 0. \tag{7.14}$$

Wynika stąd, że jeśli $\hat{Q}_n$ jest estymatorem parametru $Q$, to, zwiększając liczebności próby, ($n \to \infty$) oczekujemy, że prawdopodobieństwo, iż wartość estymatora w $n$-elementowej próbie nie odchyli się od wartości szacowanego parametru więcej niż o dowolną stałą $c$ zmierza do zera.

W niektórych podręcznikach warunek ten jest podawany jako:

$$\lim_{n\to\infty}\left(P\left|\hat{Q}_n - Q\right| < c\right) = 1, \tag{7.15}$$

co oznacza że jeśli jest estymatorem parametru $Q$, to, zwiększając liczebności próby, ($n \to \infty$) oczekujemy, że prawdopodobieństwo, iż wartość estymatora w $n$-elementowej próbie odchyli się od wartości szacowanego parametru mniej niż o dowolną stałą c zmierza do jedności.

Spośród statystyk, które mogą być estymatorami danego parametru będziemy chcieli wybierać takie, które dadzą jak najlepsze oszacowania. Kryterium wyboru są własności estymatorów[70]. Wyróżniamy bowiem:

[70] Własności te podamy bez dowodów, które można znaleźć w pracach ze statystyki matematycznej (por.: M. Fisz, *op. cit.*; A. Iwasiewicz, Z. Paszek, *op. cit.*).

1) **estymatory zgodne**
Ciąg $\{\hat{Q}_n\}$ (n = 1, 2, ....) estymatorów parametru $Q$ jest zgodny, jeśli spełnia następujący warunek:

$$\lim_{n\to\infty}\left(P\left|\hat{Q}_n - Q\right| \geq \varepsilon\right) = 0 \tag{7.16}$$

lub

$$\lim_{n\to\infty}\left(P\left|\hat{Q}_n - Q\right| < \varepsilon\right) = 1. \tag{7.17}$$

Warunek (7.16) oznacza, że jeśli $\hat{Q}_n$ jest zgodnym estymatorem parametru $Q$, to, zwiększając liczebności próby, ($n \to \infty$), oczekujemy, że prawdopodobieństwo, iż wartość estymatora w próbie $n$-elementowej odchyli się od wartości szacowanego parametru więcej niż o dowolnie małą wartość $\varepsilon$ zmierza do zera. Biorąc pod uwagę warunek (7.17), możemy powiedzieć, że jeśli $\hat{Q}_n$ jest zgodnym estymatorem parametru $Q$, to zwiększając liczebności próby ($n \to \infty$), oczekujemy, że prawdopodobieństwo, iż wartość estymatora w $n$-elementowej próbie odchyli się od wartości $\varepsilon$ szacowanego parametru mniej niż o dowolnie małą wartość zmierza do jedności. O takich estymatorach mówimy, że są one stochastycznie zbieżne do wartości parametru. Zgodność estymatora rozważamy w odniesieniu do dużych prób statystycznych. Następna własność dotyczy estymatorów o dowolnej liczebności;

2) **estymatory nieobciążone**
Wartość estymatora uzyskana dla danej próby statystycznej różni się od wartości parametru w populacji generalnej. Rozważamy wartości tych estymatorów w serii prób. Gauss wykazał, że jeśli estymator jest nieobciążony, to jego wartość oczekiwana (w serii prób) jest równa wartości szacowanego parametru. Zachodzi zatem związek:

$$E(\hat{Q}_n) = Q. \tag{7.18}$$

Różnicę $B_n = (\hat{Q}_n) - Q$ nazywamy obciążeniem estymatora;

3) **estymatory najefektywniejsze**
Estymatorem najefektywniejszym jest ten spośród estymatorów nieobciążonych, który ma najmniejszą wariancję. Taki estymator daje najmniejszy rozrzut wyników uzyskanych w serii prób, a zatem większe jest wówczas prawdopodobieństwo, że oszacowana wartość parametru będzie bliska jego prawdziwej wartości. Efektywność estymatora została wprowadzona przez Fishera.

## 7.2.2. Estymacja punktowa

Zasady estymacji punktowej przedstawimy na przykładzie wnioskowania o średnim czasie dojazdu studentów do uczelni.

### Przykład 7.6

Studenci studiów dziennych pewnej krakowskiej uczelni zostali poddani obserwacji ze względu na czas poświęcany na wykonywanie najważniejszych czynności w ciągu tygodnia, do których zaliczono również czas wymagany na dojazd do uczelni.

Zbiorowość studentów tej uczelni stanowi populację generalną. Cechą statystyczną jest czas dojazdu do uczelni. Jest to zmienna losowa typu ciągłego mierzona w skali interwałowej. Podjęto decyzję o przeprowadzeniu badań metodą reprezentacyjną. Próba będzie wybierana w drodze losowania indywidualnych dokumentów z uczelnianej bazy danych o studentach.

Czas dojazdu do uczelni oznaczamy jako $X$. Przyjmujemy założenie, że jest to zmienna losowa o rozkładzie normalnym, którego wartość przeciętna $\mu$ jest nieznana. Znane jest natomiast odchylenie standardowe $\sigma = 8$ [min]. Rozkład tej zmiennej ujmiemy sumarycznie przez przeciętny czas dojazdu do uczelni. Wyrazimy go za pomocą średniej arytmetycznej. Interesuje nas średni czas, jaki poświęcają na dojazd wszyscy studenci uczelni. Wnioskować będziemy na podstawie próby liczącej $n$ elementów. Do wnioskowania użyjemy statystyki z próby, którą jest średnia arytmetyczna dana wzorem (7.3). Jest to zgodny, nieobciążony i najefektywniejszy estymator wartości przeciętnej w populacji generalnej[71]. W rozważanym przypadku ma ona rozkład normalny (por. wzory (7.4)–(7.6)) o wartości przeciętnej $\mu$ oraz odchyleniu standardowym $\frac{\sigma}{\sqrt{n}}$, co zapisujemy jako: $N(\mu;\frac{\sigma}{\sqrt{n}})$.

Dla oszacowania średniego czasu dojazdu studentów do uczelni wylosowano próbę o liczebności $n = 9$ studentów. Otrzymano następujące wyniki:

| $x_i$ [min] | 18 | 20 | 24 | 28 | 28 | 35 | 40 | 45 | 50 |
|---|---|---|---|---|---|---|---|---|---|

Po wykonaniu obliczeń uzyskano średni czas dojazdu do uczelni równy $\bar{x} = 32$ [min].

Jest to wartość estymatora $\bar{X}$. Postępując zgodnie z zasadami estymacji punktowej, wartość tę przyjmujemy za oszacowanie średniego czasu dojazdu w populacji generalnej. W rozważanym przypadku znamy wartość odchylenia standardowego w populacji generalnej. Dzięki temu możemy podać średni błąd oszacowania tego parametru. Jest on równy:

---

[71] Formalne dowody na to, że średnia arytmetyczna jest estymatorem zgodnym, nieobciążonym i najefektywniejszym, znajdujemy np. w pracy M. Fisza, *op. cit.*; A. Iwasiewicza, Z. Paszka, *op. cit.*

$$\sigma_{\bar{x}} = \frac{\sigma}{\sqrt{n}} = \frac{8}{\sqrt{9}} = 2{,}67\,[\text{min}].$$

Oznacza to, że przyjmując jako średni czas dojazdu do uczelni wartość średniej arytmetycznej uzyskaną na podstawie 9-elementowej próby statystycznej, musimy liczyć się ze średnim odchyleniem od wartości parametru (z błędem oszacowania) równym 2,67 [min].

Jeśli odchylenie standardowe w populacji generalnej jest nieznane, to nie potrafimy podać dokładności oszacowania parametru.

**Przykład 7.7**

Przeprowadzamy analizę zachowań matrymonialnych. Obserwujemy wiek kobiet w chwili zawarcia pierwszego małżeństwa. Wiek ten rozpatrujemy jako zmienną losową $X$ o rozkładzie normalnym, którego parametry są nieznane. Interesuje nas średni wiek i wariancja wieku kobiet w chwili zawierania pierwszego małżeństwa. Wylosowano 10 kobiet, które zawierały pierwszy związek małżeński w podanym niżej wieku:

| $x_i$ [lata] | 21 | 19 | 18 | 24 | 22 | 23 | 26 | 21 | 22 | 24 |
|---|---|---|---|---|---|---|---|---|---|---|

W tej próbie średni wiek kobiet jest równy $\bar{x} = \frac{220}{10} = 22\,[\text{lata}]$.

Do oszacowania wariancji możemy wybrać[72]:

1) estymator obciążony, który jest dany wzorem:

$$S^2 = \frac{\sum_{i=1}^{n}(X_i - \bar{x})^2}{n}. \tag{7.19}$$

2) estymator nieobciążony zdefiniowano jako:

$$\hat{S}^2 = \frac{\sum_{i=1}^{n}(X_i - \bar{x})^2}{n-1}. \tag{7.20}$$

Posłużymy się estymatorem nieobciążonym. Dla wylosowanej próby przyjmuje on wartość:

$$\hat{s}^2 = \frac{5{,}2}{9} = 5{,}78\,[\text{lat}^2].$$

Odchylenie standardowe jest równe:

[72] Własności estymatorów wariancji i odchylenia standardowego są szczegółowo przedstawione np. w pracach: M. Fisz, *op. cit.*; A. Iwasiewicz, Z. Paszek, *op. cit.*

$$\hat{s} = \sqrt{\frac{52}{9}} = \sqrt{5{,}78} = 2{,}40\,[\text{lat}].$$

Zgodnie z zasadami estymacji punktowej wartości estymatorów $\overline{X}$, $\hat{S}^2$, $\hat{S}$ przyjmujemy za oszacowanie odpowiednio średniego wieku, wariancji i odchylenia standardowego. W tym przypadku nie możemy podać błędów oszacowania, ponieważ nie są znane wartości wariancji w populacji generalnej (wartości parametrów).

### 7.2.3. Estymacja przedziałowa

Estymacja przedziałowa polega na skonstruowaniu przedziału nazywanego przedziałem ufności, który z zadanym z góry prawdopodobieństwem bliskim jedności równym $1 - \alpha$ pokrywa wartość parametru $Q$. Prawdopodobieństwo to nazywamy **współczynnikiem (poziomem) ufności**. Przedział ufności jest przedziałem losowym. Musimy się liczyć z istnieniem przedziałów, które nie pokrywają wartości parametru. Przyjmując współczynnik ufności, określamy, jaka będzie częstość przedziałów pokrywających wartość parametru w długiej serii niezależnych prób. Ze wszystkich przedziałów wybieramy ten, który przy danym $1 - \alpha$ ma możliwie najmniejszą długość. Przedział ufności konstruujemy na podstawie rozkładu odpowiedniej statystyki z próby (estymatora). Mogą to być rozkłady dokładne lub graniczne (por. punkt 7.1).

Zasady estymacji przedziałowej przedstawimy na przykładach 7.8 do 7.14.

**Przykład 7.8**

Podjęto badania mające na celu określenie średnich wydatków na opiekę zdrowotną w gospodarstwach domowych emerytów i rencistów. Wiadomo, że wydatki te są zmienną losową o rozkładzie normalnym, którego wartość średnia $\mu$ jest nieznana. Znane jest natomiast odchylenie standardowe $\sigma = 20$ [zł]. Jako estymatora użyjemy średniej arytmetycznej danej wzorem (7.3), który w tym miejscu podajemy dla przypomnienia:

$$\overline{X} = \frac{1}{n}\sum_{k=1}^{n} X_k.$$

Estymator ten ma rozkład $N(\mu; \frac{\sigma}{\sqrt{n}})$. Aby skonstruować przedział ufności, przyjmujemy współczynnik ufności $1 - \alpha = 0{,}95$. Oznacza to, że w serii złożonej ze 100 niezależnych prób uzyskamy 95 przedziałów pokrywających średnie wydatki na opiekę zdrowotną w populacji generalnej w gospodarstwach domowych emerytów i rencistów. Poszukujemy granic przedziału:

$$P\left\{-u_\alpha < \frac{\overline{X} - \mu}{\frac{\sigma}{\sqrt{n}}} < u_\alpha\right\} = 1 - \alpha. \tag{7.21}$$

Po prostych przekształceniach otrzymujemy:

$$P\left\{\overline{X} - u_\alpha \cdot \frac{\sigma}{\sqrt{n}} < \mu < \overline{X} + u_\alpha \cdot \frac{\sigma}{\sqrt{n}}\right\} = 1 - \alpha. \tag{7.22}$$

Wartość $u_\alpha$ odczytujemy z tablic rozkładu normalnego dla zadanego współczynnika ufności $1 - \alpha$ (por. tabl. 7.1), w taki sposób, aby $P(|U| < u_\alpha) = 1 - \alpha$.

Dla przedziałowego oszacowania średnich wydatków na ochronę zdrowia wylosowano 16 gospodarstw domowych emerytów i rencistów. Zaobserwowano następujące wydatki ($x_i$ w zł):

| | | | | | | | |
|---|---|---|---|---|---|---|---|
| 100 | 180 | 300 | 500 | 120 | 200 | 380 | 600 |
| 160 | 200 | 400 | 650 | 180 | 240 | 430 | 320 |

Po wykonaniu obliczeń dla tej zbiorowości otrzymano średnie wydatki równe:

$$\overline{x} = \frac{4960}{16} = 310\,[\text{zł}].$$

Wartości $u_\alpha$ odczytujemy z tablic dystrybuanty rozkładu normalnego, której fragment podano w tablicy 7.1.

**Tablica 7.1.** Fragment dystrybuanty rozkładu normalnego

| $u_\alpha$ | **–0,01** | **–0,03** | **–0,04** | **–0,05** | **–0,06** |
|---|---|---|---|---|---|
| –1,9 | 0,0281 | 0,0268 | 0,0262 | 0,0256 | 0,0250 |
| –1,6 | 0,0537 | 0,0516 | 0,0505 | 0,0495 | 0,0485 |
| | | | | | |
| $u_\alpha$ | **0,01** | **0,03** | **0,04** | **0,05** | **0,06** |
| 1,6 | 0,9463 | 0,9484 | 0,9406 | 0,9505 | 0,9515 |
| 1,9 | 0,9719 | 0,9732 | 0,9686 | 0,9744 | ***0,9750*** |

Źródło: opracowanie własne.

Po podstawieniu odpowiednich wartości do wzoru (7.21) mamy:

$$\left\{310 - 1{,}96 \cdot \frac{20}{\sqrt{16}} < \mu < 310 + 1{,}96 \cdot \frac{20}{\sqrt{16}}\right\} = \{310 - 1{,}96 \cdot 5 < \mu < 310 + 1{,}96 \cdot 5\}$$

$$\{310 - 9{,}8 < \mu < 310 + 9{,}8\}.$$

Ostatecznie przedział ufności ma postać: $\{300{,}2 < \mu < 319{,}8\}$.

Jest to jeden z losowych przedziałów, które z prawdopodobieństwem 0,95 pokrywają średnie wydatki na opiekę zdrowotną w gospodarstwach domowych emerytów i rencistów.

Przykład 7.9

W celu podjęcia decyzji w zakresie zarządzania personelem przeprowadzono analizę fluktuacji zatrudnienia w firmie G. Z wykazu pracowników wylosowano w sposób niezależny 26 osób i zebrano dane dotyczące stażu pracy. Cecha ta jest zmienną losową o rozkładzie normalnym, którego parametry są nieznane. Należy oszacować średni staż pracy wszystkich zatrudnionych w firmie G.

Wartość odchylenia standardowego w populacji generalnej jest nieznana i dlatego musimy wprowadzić jego punktowe oszacowanie uzyskane według wzorów (7.18) lub (7.19). W przypadku estymacji przedziałowej nie ma znaczenia, czy wybierzemy estymator obciążony, czy nieobciążony. W obydwu przypadkach otrzymamy identyczne wyniki.

W celu skonstruowania przedziału ufności musimy wykorzystać funkcję $t$-Studenta zdefiniowaną wzorem (7.15). Poszukujemy granic przedziału ufności danego odpowiednio:

- dla estymatora obciążonego:

$$P\left\{-t_\alpha < \frac{\overline{X}-\mu}{\frac{\hat{s}}{\sqrt{n}}} < t_\alpha\right\} = 1-\alpha. \tag{7.23}$$

- dla estymatora nieobciążonego:

$$P\left\{-t_\alpha < \frac{\overline{X}-\mu}{\frac{s}{\sqrt{n-1}}} < t_\alpha\right\} = 1-\alpha. \tag{7.24}$$

W rezultacie odpowiednich przekształceń otrzymujemy następujące przedziały ufności:

$$P\left\{\overline{X} - t_\alpha \cdot \frac{\hat{s}}{\sqrt{n}} < \mu < \overline{X} + t_\alpha \cdot \frac{\hat{s}}{\sqrt{n}}\right\} = 1-\alpha. \tag{7.25}$$

$$P\left\{\overline{X} - t_\alpha \cdot \frac{s}{\sqrt{n-1}} < \mu < \overline{X} + t_\alpha \cdot \frac{s}{\sqrt{n-1}}\right\} = 1-\alpha. \tag{7.26}$$

Wartość $t$ odczytujemy z tablic rozkładu Studenta dla $n - 1$ stopni swobody i współczynnika ufności $1 - \alpha$ w taki sposób, aby $P(|T| < t_\alpha) = 1 - \alpha$. Tablice rozkładu Studenta są tak skonstruowane, że wartość statystyki $t_\alpha$ odczytujemy dla poziomu. Wybrane wartości statystyki $t$ – Studenta podano w tablicy 7.2.

Tablica 7.2. Fragment tablicy rozkładu Studenta

| $n-1$ | $\alpha$ | | | |
|---|---|---|---|---|
| | 0,01 | 0,05 | 0,09 | 0,1 |
| 5 | 4,032 | 2,571 | 2,098 | 2,015 |
| 8 | 3,355 | 2,306 | 1,928 | 1,860 |
| 9 | 3,250 | 2,262 | 1,899 | 1,833 |
| 10 | 3,169 | 2,228 | 1,877 | 1,812 |
| 15 | 2,947 | 2,131 | 1,812 | 1,753 |
| 20 | 2,845 | 2,086 | 1,782 | 1,725 |
| 25 | 2,787 | 2,060 | 1,764 | 1,708 |
| 30 | 2,750 | 2,042 | 1,752 | 1,697 |

Źródło: opracowanie własne.

Do oszacowania średniego stażu pracy zebrano dane dotyczące wybranych do próby 26 pracowników ($x_i$ staż pracy w latach):

| | | | | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| 1 | 7 | 14 | 18 | 28 | 2 | 7 | 14 | 20 | 30 | 3 | 7 | 15 |
| 22 | 33 | 5 | 10 | 17 | 25 | 35 | 5 | 12 | 18 | 27 | 35 | 25 |

Po wykonaniu obliczeń otrzymano:

- średni staż pracy równy: $\bar{x} = \frac{435}{26} = 17{,}4\,[\text{lata}]$,
- wariancja stażu pracy jest równa: $s^2 = \frac{2828{,}76}{26} = 113{,}1504\,[\text{lata}^2]$,
- odchylenie standardowe równe: $s = \sqrt{113{,}1504} = 10{,}64\,[\text{lata}]$.

Przyjmujemy współczynnik ufności $1 - \alpha = 0{,}9$, któremu przy 25 stopniach swobody odpowiada wartość $t_\alpha = 1{,}708$ (por. tab. 7.2). Podstawiamy do wzoru (7.25) i otrzymujemy:

$$\left\{18 - 1{,}708 \cdot \frac{3{,}59}{\sqrt{25}} < \mu < 18 + 1{,}708 \cdot \frac{3{,}59}{\sqrt{25}}\right\}.$$

Po wykonaniu obliczeń mamy: $\{16{,}77 < \mu < 19{,}22\}$.

Jest to jeden z losowych przedziałów, które z prawdopodobieństwem 0,90 pokrywają średnie zatrudnienie w firmie G. W tym przypadku wielkościami losowymi są zarówno końce przedziału, jak i jego długość wynosząca $S \cdot t_\alpha$.

### Przykład 7.10

W budżecie czasu studentów powinien się znaleźć czas poświęcony na lekturę literatury pięknej. Przeprowadzając badania metodą reprezentacyjną, zebrano dane dotyczące czytelnictwa wśród wybranych losowo 100 studentów pewnej uczelni. Czas poświęcony na lekturę jest zmienną losową o nieznanym rozkładzie. Należy oszacować metodą przedziałową średni czas poświęcony na lekturę literatury pięknej przez wszystkich studentów uczelni, przyjmując współczynnik ufności $1 - \alpha = 0{,}90$.

Gdy postać rozkładu zmiennej w populacji generalnej nie jest znana, a dysponujemy dużą próbą statystyczną, to dla skonstruowania przedziału ufności wykorzystamy statystykę daną wzorem (7.16):

$$U = \frac{\overline{X} - \mu}{s} \cdot \sqrt{n}.$$

Poszukujemy zatem końców przedziału:

$$P\left\{-u_\alpha < \frac{\overline{X} - \mu}{\frac{s}{\sqrt{n}}} < u_\alpha\right\} = 1 - \alpha. \tag{7.26}$$

Po prostych przekształceniach otrzymujemy:

$$P\left\{\overline{X} - u_\alpha \cdot \frac{s}{\sqrt{n}} < \mu < \overline{X} + u_\alpha \cdot \frac{s}{\sqrt{n}}\right\} = 1 - \alpha. \tag{7.27}$$

**Tablica 7.3.** Rozkład liczebności studentów pewnej uczelni według czasu poświęcanego tygodniowo na czytanie literatury pięknej

| $x_i$ | $f_i$ | $x_i'$ | $f_i \cdot x_i'$ | $f_i \cdot (x' - \overline{x})^2$ |
|---|---|---|---|---|
| 0–2 | 17 | 1 | 17 | 242,90 |
| 2–4 | 24 | 3 | 72 | 76,04 |
| 4–6 | 27 | 5 | 135 | 1,31 |
| 6–8 | 17 | 7 | 119 | 83,78 |
| 8–10 | 15 | 9 | 135 | 267,13 |
| Suma | 100 | × | 478 | 671,16 |

Źródło: dane umowne.

Dla wylosowanej próby skonstruowano szereg rozdzielczy podany w tablicy 7.3. Po wykonaniu odpowiednich obliczeń otrzymano:

- średni tygodniowy czas czytania literatury pięknej:

$$\bar{x} = \frac{478}{100} = 4{,}78\,[\text{godz.}],$$

- wariancję tygodniowego czasu czytania literatury pięknej:

$$s^2 = \frac{671{,}16}{100} = 6{,}71\,[\text{godz.}^2].\ [\text{godz}^2.],$$

- odchylenie standardowe tygodniowego czasu czytania:

$$s = \sqrt{6{,}71} = 2{,}59\,[\text{godz.}].$$

Odczytana z tablic wartość $u_\alpha = 1{,}65$. Podstawiając do wzoru, otrzymujemy:

$$\left\{4{,}78 - 1{,}65 \cdot \frac{2{,}59}{\sqrt{100}} < \mu < 4{,}78 + 1{,}65 \cdot \frac{2{,}59}{\sqrt{100}}\right\}$$

$$\{4{,}78 - 0{,}43 < \mu < 4{,}78 + 0{,}43\}.$$

Ostateczna postać przedziału ufności jest następująca: $\{4{,}25 < \mu < 5{,}21\}$.

Jest to jeden z losowych przedziałów, które z prawdopodobieństwem 0,90 pokrywają średni czas tygodniowo poświęcany przez studentów pewnej uczelni na czytanie literatury pięknej.

Zastanowimy się teraz, jak liczna powinna być próba, aby z zadanym z góry współczynnikiem ufności zapewnić dokładność oszacowania wartości przeciętnej z błędem nie przekraczającym ustalonej wielkości $d$. Jako maksymalny dopuszczalny błąd (dokładność oszacowania) przyjmiemy połowę długości przedziału ufności.

### Przykład 7.11

Dla sytuacji rozważanej w przykładzie 7.7 ustalimy liczebność próby pozwalającą oszacować średnie wydatki na opiekę zdrowotną w gospodarstwach domowych emerytów i rencistów z błędem nie większym niż 4,50 zł. Połowa długości przedziału ufności przyjmowana jako błąd oszacowania jest dana wzorem:

$$d = \frac{u_\alpha \cdot \sigma}{\sqrt{n}}. \tag{7.28}$$

Ponieważ poszukujemy liczebności próby $n$, to przekształcamy odpowiednio wzór (7.28).

W rezultacie otrzymujemy:

$$n = \frac{\sigma^2 \cdot u_\alpha^2}{d^2}. \tag{7.29}$$

Podstawiając będące do dyspozycji dane, otrzymujemy:

$$n = \frac{400 \cdot (1{,}96)^2}{25{,}25} = 75{,}88 \approx 76.$$

Aby zapewnić zadaną dokładność oszacowania, należy dysponować próbą o liczebności $n = 76$ obserwacji. Wielkość próby ustalamy, zaokrąglając do najmniejszej liczby całkowitej przekraczającej uzyskany wynik. Należałoby więc dolosować 76 – 16 = 50 gospodarstw domowych.

W rozważanym przykładzie znana była *a priori* wartość odchylenia standardowego w populacji generalnej. W praktyce trudno jest spotkać taki przypadek. Najczęściej wartość $\sigma$ jest nieznana. Postępujemy wówczas tak, jak w podanym niżej przykładzie.

**Przykład 7.12**

Przygotowywane są reprezentacyjne badania czasu pozostawania bezrobotnym przez absolwentów wyższych uczelni. Jednym z szacowanych parametrów jest średni czas pozostawania bezrobotnym. Należy ustalić wielkość próby, która przy współczynniku ufności $1 - \alpha = 0{,}99$, da oszacowanie tego parametru z błędem maksymalnym $d = 0{,}5$ [miesiąca]. Zakładamy, że czas przebywania absolwentów w stanie bezrobocia jest zmienną losową o rozkładzie normalnym, którego parametry są nieznane. Najpierw znajdziemy wzór podający zadaną dokładność oszacowania, czyli ustalimy połowę długości przedziału ufności danego wzorem (7.25). Jest ona równa:

$$d = \frac{t_\alpha \cdot \hat{s}}{\sqrt{n}}. \tag{7.30}$$

Po przekształceniu otrzymujemy:

$$n = \frac{t_\alpha^2 \cdot \hat{s}^2}{d^2},$$

$$n = \frac{t_\alpha^2 \cdot \hat{s}^2}{d^2}. \tag{7.31}$$

Dla ustalenia wielkości $n$ musimy znać wartość estymatora $\hat{S}^2$. W tym celu przeprowadzamy badanie pilotażowe. Losujemy małą próbę statystyczną o liczebności $n_0 < 30$. Obliczamy wartość estymatora $\hat{S}^2$ (wzór (7.22)). Ustalamy niezbędną wielkość próby $n$ według wzoru (7.31).

Jeśli okaże się, że $n > n_0$, to musimy dolosować brakujące $n - n_0$ elementów. Jeśli natomiast $n < n_0$, to dalszą analizę przeprowadzamy na podstawie próby pilotażowej, ponieważ jej liczebność zapewnia większą od założonej dokładność oszacowania.

Do wstępnego oszacowania wariancji czasu pozostawania bezrobotnych spośród absolwentów wyższej uczelni wylosowano 9 osób. Otrzymano następujące realizacje badanej zmiennej ($x_i$ w miesiącach).

**Tablica 7.4.** Czas pozostawania bezrobotnym po ukończeniu uczelni wyższej

| $x_i$ | 5 | 6 | 6 | 7 | 9 | 12 | 13 | 14 | 18 |
|---|---|---|---|---|---|---|---|---|---|
| $(x_i - \bar{x})^2$ | 16 | 25 | 16 | 9 | 1 | 4 | 9 | 16 | 64 |

Źródło: dane umowne.

Na podstawie tych danych ustalamy:

$$\hat{s}^2 = \frac{160}{9} = 20 \text{[miesięcy}^2\text{]}.$$

Dla $n - 1 = 8$ stopni swobody i dla $1 - \alpha = 0{,}99$, to znaczy dla $\alpha = 0{,}01$ z tablic rozkładu $t$ – Studenta (por. tabl. 7.2) odczytujemy $t_\alpha = 3{,}355$. Podstawiając do wzoru, mamy:

$$n = \frac{t_\alpha^2 \cdot \hat{s}^2}{d^2} = \frac{(3{,}335)^2 \cdot 20}{(0{,}25)^2} = \frac{66{,}7}{0{,}25} = 889{,}778.$$

Zgodnie z zasadą zaokrąglania do najmniejszej liczby całkowitej przekraczającej uzyskany wynik, stwierdzamy, że dla zapewnienia oszacowania, z zadaną dokładnością średniego czasu, jaki absolwenci uczelni wyższych pozostają bezrobotnymi należy dysponować próbą licząca $n = 890$ [osoby]. Wobec tego należy dolosować $n - n_0 = 790 - 9 = 781$ [osoby].

Zajmiemy się teraz konstruowaniem przedziału ufności dla wariancji i odchylenia standardowego.

### Przykład 7.13

Przeprowadzamy analizę zróżnicowania plac zasadniczych pracowników firmy M. Zakładamy, że cecha ta jest zmienną losową o rozkładzie normalnym, którego parametry są nieznane. Jako miarę zróżnicowania przyjęto wariancję. Dla jej oszacowania z listy płac wylosowano 8 pracowników.

W rozważanym przypadku wiemy, że cecha w populacji generalnej ma rozkład normalny o nieznanych parametrach. Z populacji tej wylosowano małą próbę statystyczną. Do skonstruowania przedziału ufności użyjemy więc statystyki z próby zdefiniowanej jako:

$$\chi^2 = \frac{n \cdot S^2}{\sigma^2}. \tag{7.32}$$

Poszukujemy zatem końców przedziału:

$$P\left\{c_1 < \frac{n \cdot S^2}{\sigma^2} < c_2\right\} = 1 - \alpha, \tag{7.33}$$

a po przekształceniach otrzymujemy przedział ufności:

$$P\left\{\frac{n \cdot S^2}{c_2} < \sigma^2 < \frac{n \cdot S^2}{c_1}\right\} = 1 - \alpha, \tag{7.34}$$

gdzie:

$S^2$ jest estymatorem wariancji w populacji generalnej zdefiniowany wzorem (7.19), $c_1$ oraz $c_2$ oznaczają wartości zmiennej losowej odczytane z tablic rozkładu $\chi^2$ dla $n - 1$ stopni swobody i poziomu istotności $1 - \alpha$ odpowiednio jako:

$$c_1 = \chi^2_{n-1;1-\frac{\alpha}{2}}, \qquad c_2 = \chi^2_{n-1;\frac{\alpha}{2}}.$$

Jeśli posłużymy się estymatorem nieobciążonym, który jest zdefiniowane wzorem (7.20), to przedział ufności przyjmie postać:

$$P\left\{\frac{(n-1 \cdot \hat{S}^2}{c_2} < \sigma^2 < \frac{(n-1) \cdot \hat{S}^2}{c_1}\right\} = 1 - \alpha. \tag{7.35}$$

W sposób najbardziej ogólny przedział ufności dla wariancji przedstawiamy jako:

$$P\left\{\frac{\sum_{i=1}^{n}(x_i - \bar{x})^2}{c_2} < \sigma^2 < \frac{\sum_{i=1}^{n}(x_i - \bar{x})^2}{c_1}\right\} = 1 - \alpha. \tag{7.36}$$

Na podstawie wylosowanej próby należy skonstruować przedział ufności, przyjmując współczynnik ufności $1 - \alpha = 0{,}95$. Zebrane w tym celu informacje podano w tablicy 7.5.

**Tablica 7.5.** Płace zasadnicze pracowników firmy M

| $x_i$ | 1000 | 1200 | 1280 | 1300 | 1400 | 1570 | 1700 | 1700 |
|---|---|---|---|---|---|---|---|---|
| $(x_i - \bar{x})^2$ | 155039,1 | 37539,1 | 12939,1 | 8789,1 | 39,1 | 31064,1 | 93789,1 | 93789,1 |

Źródło: dane umowne.

Po wykonaniu odpowiednich obliczeń otrzymano:

- średnią płacę zasadniczą: $\bar{x} = \frac{11150}{8} = 1393{,}75\,[\text{zł}]$,
- wariancję płac zasadniczych: $s^2 = \frac{432987{,}5}{8} = 54123{,}48\,[\text{zł}^2]$.

Teraz należy znaleźć wartości $c_1$ oraz $c_2$. W tym celu ustalamy: $\alpha = 0{,}05; \frac{\alpha}{2} = 0{,}025$; $n - 1 = 7$.

W tablicy 7.6 podano fragment rozkładu $\chi^2$.

Tablica 7.6. Fragment tablicy rozkładu $\chi^2$

| Stopnie swobody $n-1$ | Poziom istotności $\alpha$ | | | | | |
|---|---|---|---|---|---|---|
| | 0,025 | 0,05 | 0,01 | 0,975 | 0,95 | 0,99 |
| 5 | 12,832 | 11,070 | 15,086 | 0,831 | 1,145 | 0,554 |
| 6 | 14,449 | 12,592 | 16,812 | 1,237 | 1,635 | 0,872 |
| 7 | **16,013** | 14,067 | 18,475 | **1,690** | 2,167 | 1,239 |
| 8 | 17,535 | 15,507 | 20,090 | 2,180 | 2,733 | 1,647 |
| 9 | 19,023 | 16,919 | 21,666 | 2,700 | 3,325 | 2,088 |
| 10 | 20,483 | 18,307 | 23,209 | 3,247 | 3,940 | 2,558 |

Źródło: opracowanie własne.

Odczytane wartości są następujące:

$$c_1 = \chi^2_{7;0,975} = 1{,}690. \qquad c_2 = \chi^2_{7;0,025} = 16{,}013.$$

Po podstawieniu do wzoru (7.35) otrzymujemy:

$\left\{\frac{8 \cdot 54123{,}44}{16{,}013} < \sigma^2 < \frac{8 \cdot 54123{,}44}{1{,}690}\right\}$, a po wykonaniu obliczeń przedział ufności ma postać:

$$\{27039{,}75 < \sigma^2 < 160128{,}50\}.$$

Jest to jeden z losowych przedziałów, które z prawdopodobieństwem 0,95 pokrywają wariancję płac zasadniczych pracowników firmy M.

Jeśli interesuje nas przedziałowe oszacowanie odchylenia standardowego płac zasadniczych pracowników firmy M, to wystarczy obliczyć pierwiastki kwadratowe z końców przedziału dla wariancji. Otrzymujemy w rezultacie:

$$\{\sqrt{27039{,}75} < \sigma < \sqrt{160128{,}50}\},$$

$$\{164{,}44 < \sigma < 400{,}16\}.$$

Jest to jeden z losowych przedziałów, które z prawdopodobieństwem 0,95 pokrywają odchylenie standardowe płac zasadniczych pracowników firmy M.

**Przykład 7.14**

W celu usprawnienia obsługi klientów pewnego banku przeprowadzono analizę czasu ich oczekiwania na obsługę przy okienku kasowym. Cecha ta jest zmienną losową, której rozkład jest nieznany. Należy oszacować odchylenie standardowe czasu oczekiwania. W tym celu wybrano w sposób losowy próbę o liczebności $n = 100$ klientów. Rozważana tutaj sytuacja różni się od poprzedniej, a mianowicie: nie jest znany rozkład i dysponujemy dużą próbą statystyczną.

Przedziałowe oszacowanie odchylenia standardowego otrzymujemy jako:

$$P\left\{\frac{S}{1+\frac{u_\alpha}{\sqrt{2\cdot n}}} < \sigma < \frac{S}{1-\frac{u_\alpha}{\sqrt{2\cdot n}}}\right\} = 1-\alpha. \tag{7.37}$$

Po wykonaniu odpowiednich przekształceń otrzymujemy równoważną postać wzoru (7.37):

$$P\left\{S - u_\alpha \cdot \frac{S}{\sqrt{2\cdot n}} < \sigma < S + u_\alpha \cdot \frac{S}{\sqrt{2\cdot n}}\right\} = 1-\alpha. \tag{7.38}$$

Dla wybranej losowo próby otrzymano rozkład liczebności według czasu oczekiwania na obsługę podany w tablicy 7.7.

**Tablica 7.7.** Rozkład liczebności klientów według czasu oczekiwania na obsługę w banku

| $x_i$ | $f_i$ | $x_i'$ | $f_i \cdot x_i'$ | $f_i \cdot (x_i' - \bar{x})^2$ |
|---|---|---|---|---|
| 4–8 | 15 | 6 | 90 | 693,6 |
| 8–12 | 25 | 10 | 250 | 196,0 |
| 12–16 | 40 | 14 | 560 | 57,6 |
| 16–20 | 15 | 18 | 270 | 405,6 |
| 20–24 | 5 | 22 | 110 | 423,2 |
| | 100 | × | 1280 | 1776 |

Źródło: dane umowne.

Dla tej próby obliczono:

- średni czas oczekiwania na obsługę: $\bar{x} = \frac{1280}{100} = 12{,}80\,[\text{min}]$,
- wariancję czasu oczekiwania na obsługę: $s^2 = \frac{1776}{100} = 17{,}76\,[\text{min}^2]$,
- odchylenie standardowe czasu oczekiwania na obsługę: $s = \sqrt{17{,}76} = 4{,}21\,[\text{min}]$.

Skonstruowano przedział ufności dla odchylenia standardowego czasu oczekiwania na obsługę, przyjmując współczynnik ufności $1 - \alpha = 0{,}95$. Ma on postać:

$$\left\{ \frac{4{,}21}{1+\frac{1{,}96}{\sqrt{2\cdot 100}}} < \sigma < \frac{4{,}21}{1-\frac{1{,}96}{\sqrt{2\cdot 100}}} \right\},$$

i ostatecznie $\{3{,}69 < \sigma < 4{,}89\}$.

Jest to jeden z losowych przedziałów, które z prawdopodobieństwem 0,95 pokrywają odchylenie standardowe czasu oczekiwania na obsługę w banku.

Jeśli konieczne jest oszacowanie wariancji rozpatrywanej zmiennej, to wystarczy podnieść do kwadratu końce przedziału ufności dla odchylenia standardowego. Otrzymujemy:

$$\{13{,}67 < \sigma^2 < 23{,}89\}.$$

Jest to jeden z losowych przedziałów, które z prawdopodobieństwem 0,95 pokrywają wariancję czasu oczekiwania na obsługę w banku.

Po wprowadzeniu zarysu zasad estymacji wartości wybranych parametrów w populacji generalnej, zajmiemy się teraz drugą metodą wnioskowania statystycznego, którą jest weryfikacja hipotez.

## 7.3. Weryfikacja hipotez jako metoda indukcyjnego wnioskowania statystycznego

Poniżej omówimy metodę wnioskowania polegającą na weryfikacji hipotez statystycznych. **Hipotezy statystyczne** są przypuszczeniami, w których sformułowano sądy o nieznanym rozkładzie zmiennej losowej w populacji generalnej. Przypuszczenia te mogą odnosić się do wartości parametru lub do analitycznej postaci rozkładu rozważanej zmiennej. Słuszność wysuniętych przypuszczeń jest weryfikowana na podstawie wyników zaobserwowanych w próbie statystycznej.

Najczęściej zanim hipoteza zostanie sformułowana posiadamy już pewne informacje *a priori*. Informacje te wyznaczają **zbiór hipotez dopuszczalnych**. Oznacza to, że ze zbioru wszystkich możliwych przypuszczeń potrafimy wyeliminować te, których rozważanie nie byłoby uzasadnione.

Zbiór dopuszczalnych hipotez oznaczamy przez $\Omega$. Jeżeli elementy tego zbioru różnią się tylko wartościami parametrów rozkładu zmiennej w populacji generalnej, to hipotezy te nazywamy **parametrycznymi**. Jeśli natomiast elementy zbioru $\Omega$ mogą różnić się zarówno wartościami parametrów, jak i postacią rozkładu zmiennej w populacji generalnej, to hipotezy te nazywamy **nieparametrycznymi**.

Hipotezy są weryfikowane za pomocą testów statystycznych. **Testem statystycznym** nazywamy regułę postępowania rozstrzygającą, jakie wyniki uzyskane na podstawie próby wskazują, że postawioną hipotezę należy odrzucić, a jakie nie dają podstaw do podjęcia takiej decyzji. Decyzja o przyjęciu hipotezy może być podejmowana tylko w ściśle określonych warunkach. Podziałowi hipotez na parametryczne i nieparametryczne odpowiada takie samo rozróżnienie testów parametrycznych i nieparametrycznych.

Jeżeli wysuwamy tylko jedną hipotezę i celem testu jest zweryfikowanie, czy jest ona prawdziwa, to taki test nazywamy **testem istotności**. Rezultatem jego zastosowania jest decyzja o odrzuceniu hipotezy lub uznanie, że nie ma podstaw do jej odrzucenia. W tym przypadku możliwe jest jedynie wyeliminowanie hipotez fałszywych, a niemożliwe jest ustalenie, czy sformułowana hipoteza jest prawdziwa. Na podstawie testu istotności nie można podjąć decyzji o przyjęciu hipotezy.

Hipoteza jest weryfikowana na podstawie wyników zaobserwowanych w próbie statystycznej. Decyzję o jej odrzuceniu podejmujemy wówczas, gdy uzyskane tą drogą wyniki są mało prawdopodobne, jeśli sprawdzana hipoteza jest prawdziwa[73]. Musimy przyjąć kryterium, na podstawie którego uznamy, że zrealizowało się takie mało prawdopodobne zdarzenie. W tym celu przyjmujemy **poziom istotności** $\alpha$. Jest to prawdopodobieństwo odrzucenia weryfikowanej hipotezy wówczas, gdy jest ona prawdziwa. Jeżeli prawdopodobieństwo otrzymania wyników zaobserwowanych w próbie jest nie większe niż założony poziom istotności, to odrzucamy sprawdzaną hipotezę. Jeżeli prawdopodobieństwo uzyskania wyniku jest większe niż założony poziom istotności $\alpha$, to należy rozumieć, że przeprowadzone badanie nie przeczy postawionej hipotezie i nie ma podstaw do jej odrzucenia.

Przyjmowanie hipotezy na podstawie jednego doświadczenia, w wyniku którego nastąpiło zdarzenie mogące przy prawdziwości postawionej hipotezy pojawić się częściej niż $\alpha$ na 100, nie jest uzasadnione.

Najczęściej bezpośrednie porównywanie wartości $\alpha$ z odpowiednim prawdopodobieństwem uzyskania wyniku, który zrealizował się w próbie statystycznej, zastępuje się porównaniem wartości odpowiednich statystyk. Statystyki te określane są mianem testów statystycznych.

**Test statystyczny**, czyli sprawdzian hipotezy, jest pewną statystyką z próby. Sprawdzian hipotezy jest miernikiem rozbieżności między wynikami uzyskanymi na podstawie próby (wartość empiryczna) a przypuszczeniami sformułowanymi w postaci hipotezy przy założeniu, że jest ona prawdziwa (wartość teoretyczna). Podjęcie decyzji jest poprzedzone porównaniem wartości teoretycznej z empiryczną, co jest równoznaczne z porównywaniem prawdopodobieństw. Rezygnujemy z porównywania prawdopodobieństw ze względu na większą pracochłonność ich obliczania w porównaniu z obliczaniem wartości testu statystycznego. Jeśli obliczenia są przeprowadza-

[73] Zob. też: H. M. Blalock, *Statystyka dla socjologów*, Warszawa 1975; M. Fisz, *op. cit.*; A. Iwasiewicz, Z. Paszek, *op. cit.*

ne za pomocą statystycznych pakietów komputerowych, to obydwie wartości otrzymujemy równocześnie.

W dalszym ciągu zajmować się będziemy testami istotności, czyli będziemy sprawdzać, czy postawiona hipoteza jest prawdziwa. W tym celu formułujemy:

1) **hipotezę zerową**, to znaczy weryfikowaną, oznaczaną jako $H_0$,
2) **hipotezy alternatywne**, czyli inne dopuszczalne hipotezy.

Do sprawdzenia słuszności sformułowanego przypuszczenia wybieramy odpowiedni test. Jest to statystyka (funkcja) przyjmująca w różnych próbach różne wartości. Ze zbioru tych wartości wyodrębniamy **obszar krytyczny**.

Zbiór wszystkich wartości testu oznaczymy jako $U$, a obszar krytyczny jako $v$. Granice jego ustalamy tak, aby spełniony był warunek, że jeśli hipoteza $H_0$ jest prawdziwa, to prawdopodobieństwo, że wartość testu znajdzie się w obszarze krytycznym jest równe przyjętemu poziomowi istotności, co zapisujemy, że $P\{U \in v | H_0\} = \alpha$. Oznacza to, że za każdym razem, gdy wartość testu znajdzie się w obszarze krytycznym, to zerową hipotezę odrzucamy, na korzyść hipotezy alternatywnej $H_1$. Obszar przyjęć jest dopełnieniem obszaru krytycznego. Gdy wartość testu znajdzie się poza obszarem krytycznym ($U \in v$), to uznajemy, że nie ma podstaw do odrzucenia hipotezy $H_0$.

Odrzucenie prawdziwej hipotezy jest decyzją błędną. Błąd polegający na odrzuceniu hipotezy $H_0$, gdy jest ona prawdziwa nazywamy **błędem pierwszego rodzaju**. Prawdopodobieństwo jego popełnienia jest równe $\alpha$.

Można popełnić błąd w przeciwnym kierunku, to znaczy przyjąć hipotezę $H_0$, gdy prawdziwa jest hipoteza alternatywna. Jest to **błąd drugiego rodzaju**. Prawdopodobieństwo takiego zdarzenia jest: $P\{U \notin v | H_1\} = \beta$. Spośród prawdopodobieństw $\alpha$ i $\beta$ znacznie trudniej jest ocenić prawdopodobieństwo $\beta$. Musimy bowiem znać rozkład cechy w populacji generalnej oraz posiadać informację o stopniu rozbieżności między hipotetyczną i prawdziwą wartością parametru[74].

## 7.3.2. Weryfikacja hipotez o wartości przeciętnej w populacji generalnej

Rozważana zmienna losowa $X$ ma rozkład normalny o nieznanej wartości przeciętnej $\mu$ oraz o znanym odchyleniu standardowym $\sigma$. Formułujemy hipotezę zerową, głoszącą, że wartość przeciętna zmiennej $X$ w populacji generalnej jest równa hipotetycznej wartości $\mu_0$, co zapisujemy jako:

$H_0$: $\mu = \mu_0$.

Równocześnie stawiamy hipotezę alternatywną, która może przyjąć jedną z trzech podanych niżej postaci:

---

[74] Por.: M. Fisz, *op. cit.*; A. Iwasiewicz, Z. Paszek, *op. cit.*; S. Ostasiewicz, Z. Rusnak, U. Siedlecka, *op. cit.*; A. Zeliaś, *op. cit.*

a) $H_1: \mu \neq \mu_0$,
b) $H_1: \mu < \mu_0$,
c) $H_1: \mu > \mu_0$.

Do zweryfikowania hipotezy wylosowano małą próbę statystyczną o liczebności $n < 30$. W określonych warunkach test statystyczny (sprawdzian hipotezy) ma postać:

$$U = \frac{\overline{X} - \mu_0}{\sigma} \cdot \sqrt{n}, \tag{7.39}$$

gdzie:
U jest zmienną losową standaryzowaną o rozkładzie normalnym.

Po obliczeniu wartości U (empiryczna wartość testu) musimy podjąć decyzję o losach hipotezy zerowej. W tym celu wyznaczamy obszar krytyczny. Zmienna U przyjmuje wartości z przedziału $(-\infty; +\infty)$. Obszar krytyczny wyodrębniamy przy założeniu, że hipoteza zerowa jest prawdziwa. W tym celu niezbędne są:

- statystyka testowa (U),
- hipoteza alternatywna ($H_1$),
- poziom istotności $\alpha$.

Obszar krytyczny będzie konstruowany na podstawie rozkładu normalnego, bo jest to rozkład testu danego wzorem (7.39). Hipoteza alternatywna typu a) wyznacza dwustronny obszar krytyczny. Hipotezy typu b) i c) wyznaczają jednostronne obszary krytyczne odpowiednio lewo- (typ b) i prawostronny (typ c). Poziom istotności określa granice obszaru krytycznego.

Postępowanie prowadzące do zweryfikowania przypuszczenia o wartości przeciętnej w populacji generalnej przedstawimy na przykładzie.

### Przykład 7.15

Rozpatrujemy sytuację przedstawioną w przykładzie 7.6, w którym przeprowadzano analizę czasu dojazdu studentów na zajęcia w uczelni. Przypominamy założenia ważne nie tylko w przypadku estymacji przedziałowej, ale również dla zweryfikowania postawionej hipotezy, a mianowicie:

- czas dojazdu do uczelni jest zmienną losową $X$, która ma rozkład normalny o nieznanej wartości przeciętnej oraz o znanym odchyleniu standardowym $\sigma$,
- estymator tej średniej dany wzorem (7.3) ma rozkład normalny z parametrami oraz $\frac{5}{\sqrt{n}}$, co zapisujemy $N\left(\mu; \frac{\sigma}{\sqrt{n}}\right)$.

Przypuszczamy, że średni czas dojazdu do uczelni w całej populacji studentów może być równy $\mu_0 = 35$ [min]. Jest to hipotetyczny średni czas dojazdu do uczelni. Przypuszczenie to jest hipotezą zerową poddawaną weryfikacji. Zapisujemy to jako:

$H_0$: $\mu = 35$ [min], co oznacza, że hipoteza zerowa głosi, że średni czas dojazdu do uczelni jest równy 35 [min].

Równocześnie stawiamy konkurującą z nią hipotezę alternatywną $H_1$, która przyjmie postać: $H_1$: $\mu \neq 35$ [min], a więc przypuszczamy, że czas ten jest różny od 35 [min].

Będziemy sprawdzać, czy różnica między wartością hipotetyczną i wartością zaobserwowaną w próbie jest istotna. Wobec przyjętych założeń jako sprawdzian tej hipotezy wybieramy test dany wzorem 7.39. Licząc się z możliwością podjęcia błędnej decyzji, polegającej na odrzuceniu hipotezy prawdziwej, przyjmiemy prawdopodobieństwo popełnienia tego błędu równe $\alpha = 0{,}05$.

Do zweryfikowania sformułowanej hipotezy z populacji generalnej studentów wybrano w sposób losowy $n = 9$ osób. Dla tej próby otrzymano następujące wyniki:

| $x_i$ [min] | 18 | 20 | 24 | 28 | 28 | 35 | 40 | 45 | 50 |
|---|---|---|---|---|---|---|---|---|---|

Obliczony średni czas dojazdu do uczelni jest równy $\bar{x} = 23$ [min].

W celu sprawdzenia, czy różnica między wartością hipotetyczną i empiryczną jest statystycznie istotna, obliczamy wartość testu.

$$u = \frac{32-35}{8} \cdot \sqrt{9} = \frac{3}{8} \cdot 3 = \frac{9}{8} = 1{,}125.$$

W rozważanym przykładzie zmienna losowa służąca do testowania hipotezy zerowej wskazuje, że obszar krytyczny będzie konstruowany na podstawie rozkładu normalnego. Hipoteza alternatywna oznacza, że będzie to dwustronny obszar krytyczny. Jeśli bowiem hipoteza alternatywna jest zapisana jako $H_1$: $\mu \neq 35$[min], to oznacza ona, że dopuszczamy odchylenia zarówno *in plus*, jak i *in minus*. Obszar krytyczny jest wyznaczony tak, aby:

$$P(|U| > u_\alpha) = \alpha. \tag{7.40}$$

Jest to prawdopodobieństwo, że wartość testu znajdzie się w obszarze krytycznym, jeśli postawiona hipoteza zerowa jest prawdziwa. Wartość $u_\alpha$ odczytujemy z tablic rozkładu normalnego dla zadanego poziomu istotności $\alpha$.

Dla przyjętego na początku poziomu istotności $\mu = 0{,}05$ znajdujemy:

$$P(-u_\alpha < -U) \cup (U > u_\alpha) = 0{,}05.$$

Wartości $u_\alpha$ odczytujemy z tablic rozkładu normalnego, których fragment zamieszczono w tablicy 7.8. W tym przypadku znane jest prawdopodobieństwo. Należy podać wartości zmiennej wyznaczające granice obszaru krytycznego. W przypadku dwustronnego obszaru krytycznego prawdopodobieństwo $\alpha = 0{,}05$ zostało podzielone na połowy, tak aby: $P(u_\alpha < -U) = 0{,}025$ oraz $P(U > u_\alpha) = 0{,}025$.

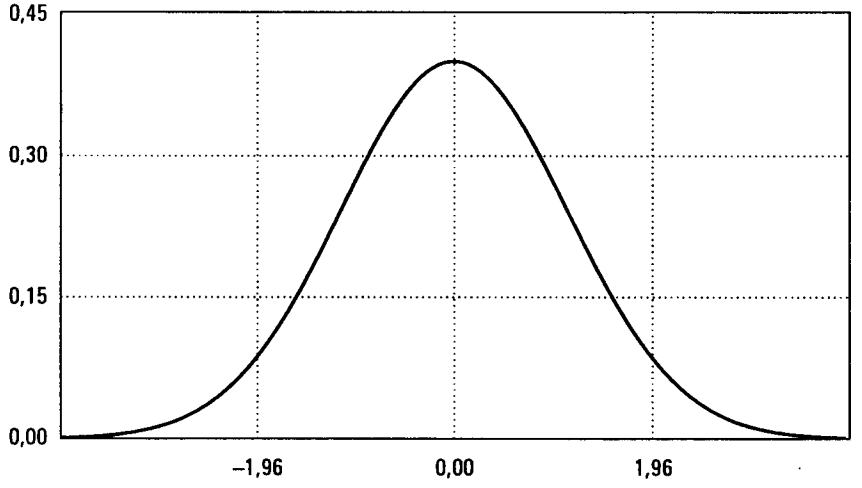
W tablicy dystrybuanty rozkładu normalnego znajdujemy więc 0,025, któremu odpowiada wartość $-u_\alpha = -1{,}96$ oraz 0,975, któremu odpowiada $u_\alpha = 1{,}96$.

Tablica 7.8. Fragment dystrybuanty rozkładu normalnego

| $u_\alpha$ | –0,01 | –0,03 | –0,04 | –0,05 | –0,06 |
|---|---|---|---|---|---|
| –1,9 | 0,0281 | 0,0268 | 0,0262 | 0,0256 | **0,0250** |
| –1,6 | 0,0537 | 0,0516 | *0,0505* | 0,0495 | 0,0485 |
| | | | | | |
| $u_\alpha$ | **0,01** | **0,03** | **0,04** | **0,05** | **0,06** |
| 1,6 | 0,9463 | 0,9484 | 0,9406 | 0,9505 | 0,9515 |
| 1,9 | 0,9719 | 0,9732 | 0,9686 | 0,9744 | **0,9750** |

Źródło: opracowanie własne.

Do obszaru krytycznego należą zatem wartości testu mniejsze niż –1,96 ($-1{,}96 < U$) albo większe niż 1,96 ($1{,}96 > U$). Obszar ten przedstawiono na rysunku 7.1




Rys. 7.1. Dwustronny obszar krytyczny dla poziomu istotności 0,05 – rozkład normalny

W omawianym przykładzie wartość empiryczna testu wynosi $u = 1{,}125$. Porównujemy ją z wartością krytyczną $u_\alpha = 1{,}96$ i stwierdzamy że $1{,}125 < 1{,}96$. Zatem nie ma podstaw do odrzucenia hipotezy zerowej ($H_0$), że średni czas dojazdu do uczelni jest równy 35 minut.

Przedstawiony w tym miejscu przykład należy traktować jako teoretyczny. Trudno bowiem znaleźć sytuację, w której znane jest odchylenie standardowe w populacji

generalnej. Zajmiemy się teraz przykładami, które częściej możemy napotykać w praktyce.

### Przykład 7.16

Zarządzający pewnym bankiem interesują się czasem, po jakim klienci indywidualni spłacają kredyt. Czas zadłużenia traktujemy jako zmienną losową. Przyjmujemy założenie, że posiada ona rozkład normalny. Parametry tego rozkładu są nieznane. Przypuszcza się, że średni czas zadłużenia klientów jest równy $\mu_0 = 24$ [miesiące]. Dla zweryfikowania słuszności tej hipotezy przeprowadzono badania metodą reprezentacyjną. Wylosowano w tym celu próbę o liczebności $n = 26$ wybraną z populacji klientów, którzy w ciągu roku $t$ spłacili kredyty zaciągnięte w analizowanym banku. Populacja ta została poddana bezpośredniemu badaniu.

Formułujemy następujące hipotezy:

$H_0$: $\mu = 24$ [miesiące],

$H_1$: $\mu \neq 24$ [miesiące].

W określonej wyżej sytuacji jako test statystyczny wykorzystamy statystykę z próby zdefiniowaną jako:

$$T = \frac{\overline{X} - \mu_0}{s}\sqrt{n-1}, \tag{7.41}$$

która posiada rozkład Studenta o $n - 1$ stopniach swobody.

Do podjęcia decyzji niezbędny jest poziom istotności. Przyjmujemy, że będzie równy $\alpha = 0{,}1$.

**Tablica 7.9.** Fragment rozkładu Studenta

| $n-1$ | $\alpha$ | | | |
|---|---|---|---|---|
| | 0,01 | 0,05 | 0,09 | 0,1 |
| 5 | 4,032 | 2,571 | 2,098 | 2,015 |
| 8 | 3,355 | 2,306 | 1,928 | 1,860 |
| 9 | 3,250 | 2,262 | 1,899 | 1,833 |
| 10 | 3,169 | 2,228 | 1,877 | 1,812 |
| 15 | 2,947 | 2,131 | 1,812 | 1,753 |
| 20 | 2,845 | 2,086 | 1,782 | 1,725 |
| 25 | 2,787 | 2,060 | 1,764 | 1,708 |
| 30 | 2,750 | 2,042 | 1,752 | 1,697 |

Źródło: opracowanie własne.

Konstruujemy obszar krytyczny, biorąc pod uwagę: rozkład statystyki służącej jako test, hipotezę alternatywną oraz poziom istotności. Obszar ten wyznaczamy na podstawie rozkładu Studenta.

Podobnie jak poprzednio jest to obszar dwustronny. Jego granice odczytujemy z tablic dla zadanej wartości $\alpha$ oraz dla $n - 1$ stopni swobody. Odpowiednie wartości zapisano w tablicy 7.9.

Obszar krytyczny jest wyznaczony tak, aby: $P(|T| > t_\alpha) = \alpha$.

Dla przyjętego na początku poziomu istotności $\alpha = 0{,}1$ znajdujemy $t_\alpha$ tak aby:

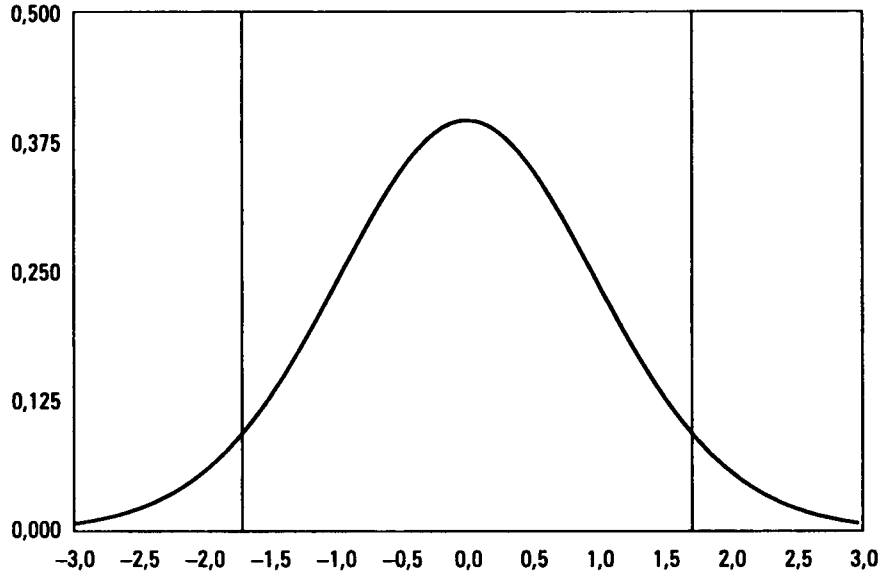
$$P(-t_\alpha < -T) \cup (T > t_\alpha) = 0{,}1.$$

Dla $n - 1 = 25$ stopni swobody i $\alpha = 0{,}1$ odczytujemy: $t_\alpha = 1{,}708$.

Do obszaru krytycznego tworzą zatem wartości $T$ należące do przedziałów:

$$(-1{,}708 < -U) \cup (U > 1{,}708).$$

Obszar ten przedstawiono na rysunku 7.2.




Rys. 7.2. Rozkład $t$ – Studenta; dwustronny obszar krytyczny dla poziomu istotności $\alpha = 0{,}1$ oraz dla 25 stopni swobody

Na podstawie przeprowadzonych badań otrzymano następujące wyniki:

$\bar{x} = 26$ [miesięcy]; $s = 5{,}06$ [miesiąca].

Na tej podstawie ustalamy empiryczną wartość testu:

$$t = \frac{24-26}{5{,}06} \cdot \sqrt{25} = -\frac{2}{5{,}06} \cdot 5 = -\frac{10}{5{,}06} = -1{,}97.$$

Porównujemy empiryczną wartość (–1,97) testu z wartością krytyczną ($t_\alpha = -1{,}708$) i stwierdzamy, że $-t_\alpha < -t$. Oznacza to, że empiryczna wartość testu znalazła się w obszarze krytycznym. Na tej podstawie odrzucamy hipotezę zerową na korzyść hipotezy alternatywnej głoszącej, że średni czas zadłużenia jest różny od 24 miesięcy. Znak minus pozwala przypuszczać, że prawdopodobnie jest on dłuży od wartości hipotetycznej.

**Przykład 7.17**

Poddamy weryfikacji hipotezę dotyczącą średniego czasu, jaki studenci poświęcają tygodniowo na czytanie literatury pięknej. Weryfikację przeprowadzamy w takich samych warunkach jak w przykładzie 7.10. Czas poświęcany na lekturę jest zmienną losową $X$ o nieznanym rozkładzie. Przypuszczamy, że czytanie literatury pięknej zajmuje studentom średnio 5 godzin na tydzień. Formułujemy zatem następujące hipotezy:

$H_0$: $\mu = 5{,}0$ [godziny],

$H_1$: $\mu \neq 5{,}0$ [godziny].

Z populacji generalnej wybrano w sposób losowy 100 studentów. Wprawdzie postać rozkładu zmiennej w populacji generalnej nie jest znana, ale dysponujemy dużą próbą statystyczną ($n > 30$). Jako test wykorzystamy statystykę o rozkładzie normalnym daną jako:

$$U = \frac{\overline{X} - \mu}{s} \cdot \sqrt{n}. \tag{7.42}$$

Na podstawie odpowiednich obliczeń wykonanych w przykładzie 7.10 otrzymaliśmy:

- średni tygodniowy czas czytania literatury pięknej: $\bar{x} = \frac{478}{100} = 4{,}78$ [godz.],
- odchylenie standardowe tygodniowego czasu czytania: $s = \sqrt{6{,}71} = 2{,}59$ [godz.].

Podstawiając do wzoru (7.16), otrzymujemy:

$$u = \frac{4{,}78 - 5{,}00}{2{,}59} \cdot \sqrt{100} = \frac{0{,}22}{2{,}59} \cdot 10 = 0{,}85.$$

Odczytana z tablic wartość $u_\alpha = 1{,}65$.

Porównując wartość $u_\alpha = 0{,}85$ oraz $u = 1{,}96$, stwierdzamy, że $u_\alpha < u$. Zatem nie ma podstaw do odrzucenia hipotezy, że średni czas przeznaczany przez studentów na czytanie literatury pięknej jest równy 5 godzin tygodniowo.

### 7.3.2. Weryfikacja hipotezy o wariancji w populacji generalnej

Istnieją zagadnienia, w których oprócz wartości przeciętnej interesujemy się także rozproszeniem wyników wokół niej.

Przykład 7.18

Automat napełnia cukrem torebki, w których nominalnie powinien mieścić się 1 kg. Dopuszcza się tolerancję równą 3 [g] (*in plus, in minus*). Podjęto badanie mające na celu ocenę prawidłowości funkcjonowania automatu. Sprawdzono, czy odchylenia ciężaru torebek cukru od ciężaru nominalnego nie przekraczają założonej tolerancji. W rozważanym przypadku populację generalną stanowią pakowane przez automat torebki cukru, których ciężar podlega obserwacji. Jest to zmienna losowa $X$ o rozkładzie normalnym, którego parametry są nieznane. Z bieżącej produkcji pobrano $n = 7$ torebek cukru. Dopuszczamy prawdopodobieństwo popełnienia błędu pierwszego rodzaju $\alpha = 0{,}1$.

Weryfikacji podlegać będzie hipoteza:
$H_0$: $\sigma^2 = \sigma_0^2$,
$H_1$: $\sigma^2 > \sigma_0^2$.

W rozważanej sytuacji mamy $n < 30$ obserwacji. W zdefiniowanych wyżej warunkach jako test wybierzemy statystykę daną wzorem:

$$\chi^2 = \frac{n \cdot S^2}{\sigma_0^2} \tag{7.43}$$

lub

$$\chi^2 = \frac{(n-1) \cdot \hat{S}^2}{\sigma_0^2}, \tag{7.44}$$

gdzie:
$\hat{S}^2$ oznacza wariancję zmiennej $X$, której wartość obliczymy na podstawie próby;
$\sigma_0^2$ jest hipotetyczną wartością wariancji w populacji generalnej.

Statystyka ta ma rozkład $\chi^2$ o $n - 1$ stopniach swobody. Hipoteza alternatywna wyznacza jednostronny (prawostronny) obszar krytyczny, którego granicę ustala wartość krytyczna $\chi_\alpha^2$ odczytana z tablic rozkładu $\chi^2$ dla zadanego poziomu istotności $\alpha$ oraz $n - 1$ stopni swobody, tak aby: $P\{\chi^2 \geq \chi_\alpha^2\} = \alpha$.

Dla rozważanego przykładu weryfikacji poddajemy hipotezę o postaci:
$H_0$: $\sigma^2 = 9$ [g$^2$],
$H_1$: $\sigma^2 > 9$ [g$^2$].

Dopuszczalne prawdopodobieństwo popełnienia błędu pierwszego rodzaju przyjmujemy $\alpha = 0{,}1$. Na tej podstawie wyników zaobserwowanych w próbie rozważanej w przykładzie 7.7 otrzymano:

- średni ciężar torebek cukru:

$$\bar{x} = 998{,}77\ [g],$$

- wariancję ciężaru torebek cukru:

$$s^2 = 13{,}56\ [g^2].$$

Po podstawieniu do wzoru (7.44) otrzymujemy:

$$\chi^2 = \frac{94{,}95}{9} = 10{,}5498.$$

Wartość krytyczną odczytujemy z tablicy 7.10 zawierającej fragment rozkładu $\chi^2$.

**Tablica 7.10.** Fragment tablicy rozkładu $\chi^2$

| | 0,01 | 0,05 | 0,10 |
|---|---|---|---|
| 5 | 15,086 | 11,070 | 9,236 |
| 6 | 16,812 | 12,592 | **10,645** |
| 7 | 18,475 | 14,067 | 12,017 |
| 8 | 20,090 | 15,507 | 13,362 |
| 9 | 21,666 | 16,919 | 14,684 |

Źródło: opracowanie własne.

Dla 6 stopni swobody i przyjętego poziomu istotności wartość krytyczna jest $\chi^2 = 10{,}645$. Wyznaczony obszar krytyczny przedstawiono na rysunku 7.3.

Wartość obliczona $\chi^2 = 10{,}645 < \chi^2_\alpha = 10{,}645$. Znajduje się ona poza obszarem krytycznym, dlatego nie ma podstaw do odrzucenia hipotezy zerowej, że wariancja ciężaru torebek cukru jest zgodna z normą, czyli równa $\sigma^2 = 3\ [g^2]$.

Zweryfikujemy dodatkowo hipotezę, że średni ciężar pakowanych torebek jest zgodny z normą, a więc równy 1000 [g]. Formułujemy hipotezy:

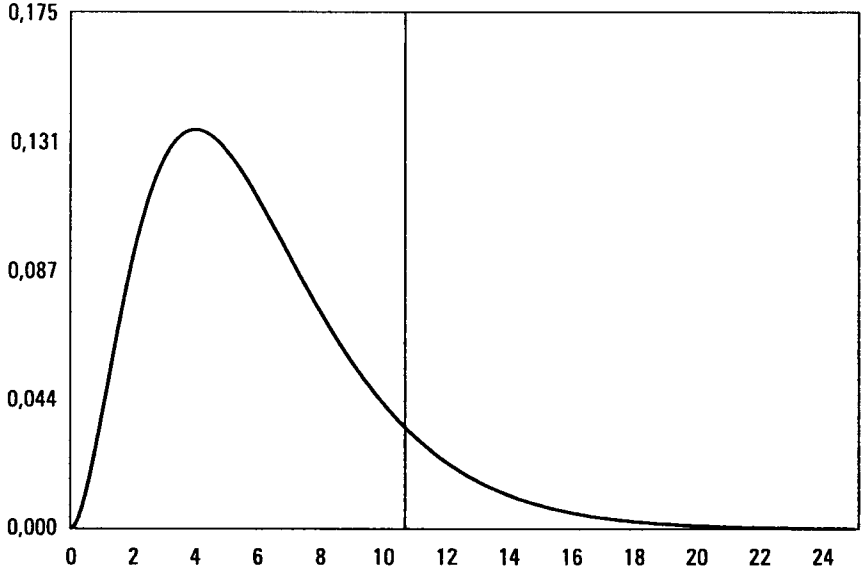
$H_0$: $\mu = 1000$ [g],

$H_1$: $\mu \neq 1000$ [g].

Przyjmujemy poziom istotności $\alpha = 0{,}1$.

Użyjemy statystyki danej wzorem (7.41). Dla rozważanej sytuacji otrzymujemy:

$$t = \frac{\bar{x} - \mu_0}{s}\sqrt{n-1} = \frac{998{,}31 - 1000}{3{,}68} \cdot \sqrt{6} = \frac{-1{,}69}{3{,}68} \cdot 2{,}45 = \frac{-4{,}13}{3{,}68} = -1{,}12.$$




Rys. 7.3. Rozkład $\chi^2$ – jednostronny obszar krytyczny dla poziomu istotności $\alpha = 0,1$ oraz dla 6 stopni swobody

Z tablicy rozkładu $t$ – Studenta odczytujemy wartość krytyczną $t_\alpha = 2,447$ (zob. tabl. 7.2). Okazuje się, że $-t_\alpha < -t$. Obliczona wartość testu znajduje się poza obszarem krytycznym. Wobec powyższego nie ma podstaw do odrzucenia hipotezy zerowej, że ciężar torebek napełnianych cukrem przez automat jest równy 1000 [g], czyli zgodnie z normą.

### Przykład 7.19

Należy zbadać zróżnicowanie dochodów gospodarstw domowych województwa M. Przypuszcza się, że odchylenie standardowe tych dochodów jest równe $\sigma = 200$ [zł]. Wariancja jest więc równa $\sigma^2 = 40000$ [zł$^2$]. Rozkład dochodów w badanej populacji jest nieznany. Do zweryfikowania hipotezy o wartości wariancji wylosowano 225 gospodarstw domowych. Otrzymano wyniki podane zostały w tablicy 7.11.

W rozważanym przypadku dochody gospodarstw domowych w województwie M są zmienną losową, której rozkład jest nieznany. Badania będą przeprowadzone metodą reprezentacyjną na podstawie dużej próby wybranej drogą losowania niezależnego. Dopuszczamy prawdopodobieństwo popełnienia błędu pierwszego rodzaju $\alpha$.

Formułujemy odpowiednie hipotezy:

$H_0$: $\sigma^2 = \sigma_0^2$,

$H_1$: $\sigma^2 > \sigma_0^2$.

W rozważnym przypadku jako test statystyczny wykorzystujemy statystykę daną wzorem:

$$U = \sqrt{\frac{2 \cdot n \cdot S^2}{\sigma_0^2}} - \sqrt{2 \cdot n - 3}. \tag{7.45}$$

Statystyka ta ma graniczny rozkład normalny[75]. Wartość krytyczną u odczytujemy z tablic tego rozkładu (por. tablica 6.8 i 7.1).

W rozważanym przykładzie mamy:

$H_0$: $\sigma^2 = 40000$ [zł$^2$],

$H_1$: $\sigma^2 > 40000$ [zł$^2$].

Przyjmujemy poziom istotności $\alpha = 0{,}05$.

Dla wylosowanej próby otrzymano wyniki podane w tablicy 7.11.

Tablica 7.11. Rozkład liczebności dochodów wylosowanych gospodarstw domowych według wysokości dochodu

| $x_i$ | $f_i$ | $x_i'$ | $f_i \cdot x_i'$ | $f_i \cdot (x_i' - \bar{x})^2$ |
|---|---|---|---|---|
| 1000–1200 | 10 | 1100 | 11000 | 1541289 |
| 1200–1400 | 21 | 1300 | 27300 | 778930 |
| 1400–1600 | 24 | 1500 | 36000 | 1317 |
| 1600–1800 | 14 | 1700 | 23800 | 602250 |
| 1800–2000 | 12 | 1900 | 22800 | 1991770 |
| Suma | 81 | × | 120900 | 4915556 |

Źródło: dane umowne.

W celu uzyskania wartości testu musimy znaleźć wartość wariancji badanej próby. Obliczamy:

- średni dochód badanej zbiorowości gospodarstw:

$$\bar{x} = \frac{120900}{81} = 1492{,}59\,[\text{zł}],$$

- wariancję dochodów badanej zbiorowości gospodarstw:

$$s^2 = \frac{4915556}{81} = 246{,}3450\,[\text{zł}^2].$$

Następnie ustalamy wartość testu:

$$U = \sqrt{\frac{2 \cdot 81 \cdot 60685{,}87}{40000}} - \sqrt{2 \cdot 81 - 3} = \sqrt{\frac{9831111{,}11}{40000}} - \sqrt{162 - 3}$$

[75] Gdy $n \to \infty$, to rozkład zmiennej losowej $U$ zdąża do rozkładu N(0; 1).

$$U = \sqrt{245,77} - \sqrt{159} = 15,68 - 12,61 = 3,07.$$

Z tablicy rozkładu normalnego (zob. tabl. 7.1) wyznaczamy granicę prawostronnego obszaru krytycznego (porównaj hipotezę alternatywną), tak aby $P(U > u_\alpha) = \alpha$. Znajdujemy wartość $u_\alpha$, której odpowiada wartość dystrybuanty dla $1 - \alpha = 1 - 0,05 = 0,95$. Jest to $u_\alpha = 1,65$. Obliczona wartość testu $u = 3,07 > u_\alpha = 1,65$. Znalazła się ona zatem w obszarze krytycznym, a więc odrzucamy hipotezę, że wariancja w populacji generalnej jest równa $\sigma^2 = 40000$ [zł$^2$], na korzyść hipotezy alternatywnej głoszącej, że $\sigma^2 > 40000$ [zł$^2$].

### 7.3.4. Weryfikacja hipotez w zakresie badania związków między zjawiskami

Rozważamy populację generalną, w której obserwujemy dwuwymiarową zmienną losową $XY$ o rozkładzie normalnym o nieznanych parametrach, którymi są: wartości przeciętne ($\mu_x$, $\mu_y$), odchylenia standardowe ($\sigma_x$, $\sigma_y$) oraz współczynnik korelacji ($\rho$). Chcemy sprawdzić, czy współczynnik korelacji $\rho$ istotnie różni się od zera. Interesujemy się zatem, czy korelacja między badanymi zmiennymi jest statystycznie istotna. Formułujemy następujące hipotezy:

$H_0$: $\rho = 0$,

$H_1$: $\rho \neq 0$.

Słuszność tych przypuszczeń zweryfikujemy za pomocą testu o postaci:

$$T = \frac{r}{\sqrt{1-r^2}}\sqrt{n-2}. \tag{7.46}$$

Statystyka ta ma rozkład $t$ – Studenta o $n - 2$ stopniach swobody. Do podjęcia decyzji o losach hipotezy zerowej przyjmujemy poziom istotności $\alpha$. Biorąc pod uwagę hipotezę alternatywną, konstruujemy dwustronny obszar krytyczny. Należą do niego wartości testu $t$ spełniające warunek: $P(t_\alpha < -T) \cup (T > t_\alpha) = \alpha$. Wartość $t_\alpha$ odczytujemy z tablic rozkładu $t$ – Studenta dla zadanego poziomu istotności $\alpha$ oraz $n - 2$ stopni swobody.

**Przykład 7.20**

Przeprowadzono badania współzależności między ilością zakładów przemysłowych i emisją zanieczyszczeń pyłowych. Oznaczamy odpowiednio:

- zmienna $Y$ – emisja zanieczyszczeń w tysiącach ton,
- zmienna $X$ – liczba zakładów przemysłowych.

Dla wybranych losowo $n = 6$ zakładów otrzymano następujące wyniki, na podstawie których obliczono wartość współczynnika korelacji $r$ Pearsona. Dane wraz z obliczeniami zawiera tablica 7.12.

**Tablica 7.12.** Obliczenia pomocnicze w celu zbadania współzależności między liczbą zakładów przemysłowych i emisją zanieczyszczeń pyłowych

| $y_i$ | $x_i$ | $y_i - \bar{y}$ | $x_i - \bar{x}$ | $(y_i - \bar{y}) \cdot (x_i - \bar{x})$ | $(y_i - \bar{y})^2$ | $(x_i - \bar{x})^2$ |
|---|---|---|---|---|---|---|
| 2 | 3 | -5 | -4 | 20 | 25 | 16 |
| 4 | 6 | -3 | -1 | 3 | 9 | 1 |
| 8 | 7 | 1 | 0 | 0 | 1 | 0 |
| 7 | 5 | 0 | -2 | 0 | 0 | 4 |
| 10 | 11 | 3 | 4 | 12 | 9 | 16 |
| 11 | 10 | 4 | 3 | 12 | 16 | 9 |
| 42 | 42 | | | 47 | 60 | 46 |

Źródło: dane umowne.

Dla wylosowanej próby wartość współczynnika korelacji jest równa:

$$r = \frac{47}{\sqrt{60 \cdot 46}} = 0{,}8946.$$

Przyjmujemy poziom istotności $\alpha = 0{,}05$.
Obliczamy wartość testu:

$$t = \frac{0{,}8946}{\sqrt{1-(8946)^2}} \cdot \sqrt{6-2} = \frac{0{,}8946}{0{,}4469} \cdot 2 = 4{,}004.$$

Wartość krytyczną $t_\alpha = 2{,}776$ odczytujemy z tablic rozkładu $t$ – Studenta dla poziomu istotności $\alpha = 0{,}05$ i dla $6 - 2 = 4$ stopni swobody. Empiryczna wartość testu znalazła się w obszarze krytycznym ($t > t_\alpha$). Odrzucamy zatem hipotezę zerową na korzyść hipotezy alternatywnej głoszącej, że wartość współczynnika korelacji $\rho$ istotnie różni się od zera. Istnieje więc związek między emisją pyłów i liczbą zakładów przemysłowych.

Związek między zjawiskami możemy również badać za pomocą analizy regresji (por. punkt 4.6). Wówczas do danych empirycznych dopasowujemy funkcję regresji, która dla populacji generalnej przybiera postać:

$$\hat{Y} = \alpha + \beta X. \tag{7.47}$$

Symbolami $\alpha$ oraz $\beta$ oznaczono parametry funkcji regresji, których wartości należy oszacować na podstawie danych zaobserwowanych w próbie, a następnie sprawdzić istotność współczynnika regresji $\beta$. Postępowanie badawcze prowadzimy w populacji generalnej, której charakterystyką poprzedzono przykład 7.19. Jeśli badamy

związek między zjawiskami metodą analizy regresji, to weryfikacji poddajemy następujące hipotezy:

$H_0$: $\beta = 0$,

$H_1$: $\beta \neq 0$.

Słuszność wysuniętych przypuszczeń sprawdzimy za pomocą testu o postaci:

$$T = \frac{b}{s_\varepsilon} \cdot \sqrt{\sum_{i=1}^{n} (x_i - \bar{x})^2}. \tag{7.48}$$

Aby podjąć decyzję o losach hipotezy zerowej, musimy przyjąć poziom istotności. Konstruujemy dwustronny obszar krytyczny, biorąc pod uwagę postać hipotezy alternatywnej. Na podstawie wyników zaobserwowanych w próbie obliczamy wartość testu danego wzorem (7.48). Porównując empiryczną wartość testu $t$ z wartością krytyczną ($t_\alpha$), decydujemy: „odrzucamy hipotezę zerową" albo „nie ma podstaw do odrzucenia hipotezy zerowej". Odrzucenie hipotezy oznacza istotność współczynnika regresji $\beta$. Jeśli nie ma podstaw do podjęcia takiej decyzji, to wartość ta różni się od zera nieistotnie, co oznacza brak związku między zmiennymi $X$ oraz $Y$.

**Przykład 7.21**

Badamy związek między ilością zakładów przemysłowych i emisją zanieczyszczeń pyłowych. Tym razem posłużymy się metodą analizy regresji. Na podstawie danych i wyników obliczeń z tablicy 7.12 ustalamy:

- współczynnik regresji: $b = \frac{47}{46} = 1{,}0218$,
- wyraz wolny funkcji regresji: $a = \bar{y} - b \cdot \bar{x} = 7 - 1{,}0218 \cdot 7 = -0{,}1522$.

Funkcja regresji ma więc postać:

$$\hat{y} = -0{,}15 + 1{,}02x.$$

Obliczamy wartość odchylenia standardowego składnika resztowego:

$$s_\varepsilon = \sqrt{\frac{\sum_{i=1}^{n} (y_i - \hat{y}_i)^2}{n-k}} = \sqrt{\frac{11{,}9783}{4}} = \sqrt{2{,}9946} = 1{,}73.$$

Obliczamy wartość testu danego wzorem (7.48):

$$t = \frac{1{,}022}{1{,}730} \cdot \sqrt{46} = 0{,}59 \cdot 6{,}78 = 4{,}006.$$

Wartość $t_\alpha = 2{,}776$. Okazuje się, że $t = 4{,}006 > t_\alpha = 2{,}776$. Odrzucamy hipotezę zerową na korzyść hipotezy alternatywnej. Oznacza to, że współczynnik regresji $\beta$

istotnie różni się od zera. Istnieje zatem związek między ilością zakładów przemysłowych i emisją zanieczyszczeń pyłowych.

W niniejszym rozdziale przedstawiliśmy tylko wybrane przykłady wnioskowania statystycznego zarówno metodą estymacji parametrów, jak i weryfikacji hipotez o ich wartościach w populacji generalnej. Rozważania te mają jedynie ilustrować zasady postępowania i skierować zainteresowania Czytelnika na postępowanie w przypadku innych parametrów oraz na procedury pozwalające zweryfikować hipotezy dotyczące postaci rozkładu zmiennej losowej w populacji generalnej. Zagadnienia te przedstawiają na przykład w pracach A. Iwasiewicz i Z. Paszek[76], S. Ostasiewicz, Z. Rusnak, U. Siedlecka[77], M. Woźniak[78] (i in.), A. Zeliaś[79].

[76] M. Fisz, *Rachunek prawdopodobieństwa i statystyka matematyczna*, Warszawa 1967, A. Iwasiewicz, Z. Paszek, *Statystyka z elementami statystycznych metod monitorowania procesów*, Kraków 2004.

[77] S. Ostasiewicz, Z. Rusnak, U. Siedlecka, *Statystyka. Elementy teorii i zadania*, Wrocław 1999.

[78] M. Woźniak (red.), *Statystyka ogólna*, Kraków 1997.

[79] A. Zeliaś, *Metody statystyczne*, Warszawa 2000.

# Literatura


Aczel A., *Statystyka w zarządzaniu. Pełny wykład*, PWN, Warszawa, 2000.

Blalock H.M., *Statystyka dla socjologów*, PWN, Warszawa 1975.

Bracha C., *Teoretyczne podstawy metody reprezentacyjnej*, PWN, Warszawa 1996.

Fisz M., *Rachunek prawdopodobieństwa i statystyka matematyczna*, PWN, Warszawa 1967.

Gerstenkorn T., Śródka T., *Kombinatoryka i rachunek prawdopodobieństwa*, PWN, Warszawa 1973.

Góralski A., *Metody opisu i wnioskowania statystycznego w psychologii*, PWN, Warszawa 1974.

Holzer J. Z., *Demografia*, PWE, Warszawa 2004.

Iwasiewicz A., Paszek Z., *Statystyka z elementami statystycznych metod monitorowania procesów*, Wyd. AE w Krakowie, Kraków 2004.

Jóźwiak J., Podgórski J., *Statystyka od podstaw*, PWE, Warszawa 1992.

Kurkiewicz J., *Podstawowe metody analizy demograficznej*, PWN, Warszawa 1992.

Ostasiewicz S., Rusnak Z., Siedlecka U., *Statystyka. Elementy teorii i zadania*, Wyd. AE we Wrocławiu, Wrocław 1999.

Starzyńska W., *Statystyka praktyczna*, PWN, Warszawa 2000.

Steczkowski J., *Metoda reprezentacyjna w poglądach jej twórców*, PN, AE we Wrocławiu", nr 513 (1990).

Woźniak M., (red.), *Statystyka ogólna*, Wyd. AE w Krakowie, Kraków 1997.

Yule G. U., Kendall M. G., *Wstęp do teorii statystyki*, PWN, Warszawa 1966.

Zając K., *Zarys metod statystycznych*, PWE, Warszawa 1988.

Zasępa R., *Zarys metody reprezentacyjnej*, „Biblioteka Wiadomości Statystycznych", t. 30, GUS-PTS, Warszawa 1991.

Zeliaś A., *Metody statystyczne*, PWE, Warszawa 2000.

Zeliaś A., *Teoria prognozy*, PWE, Warszawa 1997.

# Spis tablic

# Spis rysunków