Iwona Jażdżewska

STATYSTYKA

Podręcznik dla studentów turystyki i rekreacji

STATYSTYKA

Podręcznik dla studentów turystyki i rekreacji

WYDAWNICTWO
UNIWERSYTETU
ŁÓDZKIEGO

Iwona Jażdżewska

STATYSTYKA

Podręcznik dla studentów turystyki i rekreacji

WYDAWNICTWO
UNIWERSYTETU
ŁÓDZKIEGO
Łódź 2019

Iwona Jażdżewska – Uniwersytet Łódzki, Wydział Nauk Geograficznych
Instytut Geografii Miast i Turyzmu, Zakład Geoinformacji, 90-142 Łódź, ul. Kopcińskiego 31

RECENZENT
Jerzy Runge

REDAKTOR INICJUJĄCY
Beata Koźniewska

REDAKTOR WYDAWNICTWA UŁ
Katarzyna Gorzkowska

SKŁAD I ŁAMANIE
Munda – Maciej Torz

OPRACOWANIE TECHNICZNE RYSUNKÓW
Anna Wosiak

PROJEKT OKŁADKI
Katarzyna Turkowska

Zdjęcie wykorzystane na okładce autorstwa Iwa Hencza

Wydane przez Wydawnictwo Uniwersytetu Łódzkiego
Wydanie I. W.07409.16.0.S

Ark. wyd. 11,0; ark. druk. 15,75

ISBN 978-83-8142-535-3
e-ISBN 978-83-8142-536-0
https://doi.org/10.18778/8142-536-0
Wydawnictwo Uniwersytetu Łódzkiego
90-131 Łódź, ul. Lindleya 8
www.wydawnictwo.uni.lodz.pl
e-mail: ksiegarnia@uni.lodz.pl
tel. (42) 665 58 63

SPIS TREŚCI

WPROWADZENIE

W badaniach obejmujących szeroko rozumianą turystykę występuje wiele informacji, które zazwyczaj, choć nie zawsze, przyjmują formę liczb. Niekiedy jest ich kilka i można je natychmiast poddać analizie, jednak z czasem ich przybywa – np. liczba informacji wzrasta do kilkuset tysięcy i wtedy, aby je zinterpretować czy wykryć pewne prawidłowości, trzeba posłużyć się procedurami badawczymi oferowanymi przez statystykę.

Celem autorki niniejszego podręcznika jest przedstawienie podstawowej wiedzy na temat metod statystycznych oraz umiejętności ich wykorzystania do analizy i oceny wyników prowadzonych badań statystycznych przez studentów różnych kierunków studiów związanych z turystyką. Wiedza ta może być przydatna podczas przygotowywania pracy licencjackiej lub magisterskiej.

Zanim rozpocznie się własne badania, bardzo ważne jest postawienie wielu pytań. Na początek należy określić, po co je wykonujemy. Czy chcemy opisać jedno zjawisko, czy kilka? Jaki problem chcemy rozwiązać? Czy badania będą statyczne (jednokrotne), czy dynamiczne (prowadzone w ciągu kilku lat)? Czy chcemy wskazać związki między badanymi zmiennymi, czy różnice między nimi? Jakimi danymi dysponujemy? Czy przeprowadzimy badania całej populacji, czy określonej próby (badania cząstkowe)? Jakie metody

statystyczne możemy wykorzystać? Jakie narzędzia komputerowe są w zasięgu naszych zdolności i możliwości wykorzystania? Jeśli analizujemy wyniki innych autorów (gdy ktoś już podobne badania wykonał), powinniśmy mieć możliwość oceny ich rzetelności, prawidłowo sformułowanych hipotez, zebranych danych źródłowych oraz metod statystycznych, które pozwoliły autorowi na uogólnione wnioski.

Przetwarzanie danych statystycznych wymaga znajomości metod, a także odpowiedniego oprogramowania komputerowego, które eliminuje czasochłonne obliczenia. Można sobie jednak poradzić w sytuacjach, gdy nie mamy go pod ręką i trzeba będzie wykonać obliczenia proste. Ich wykonanie z użyciem kalkulatora lub bez niego pozwala na lepsze zrozumienie stosowanej procedury. Metody tych obliczeń będą prezentowane na przykładach.

W tym miejscu chciałabym podziękować Studentkom i Studentom kierunku turystyka i rekreacja, kształcącym się na Wydziale Nauk Geograficznych Uniwersytetu Łódzkiego, z którymi miałam przez wiele lat przyjemność prowadzić zajęcia ze statystyki i którzy byli pierwszymi czytelnikami tekstu, a także jako pierwsi rozwiązywali zadania oraz stosowali wybrane metody w pracach magisterskich.

Autorka

ROZDZIAŁ 1

ZAGADNIENIA WSTĘPNE

Do rozwiazywania zadań ze statystyki potrzebna jest umiejętność posługiwania się podstawowymi operacjami, symbolami i oznaczeniami matematycznymi, takimi jak np.: +, –, ×, /, $\sqrt{}$, !, %, ‰, <, >, ≤, ≥, ≠, ∈, ∉, $\sum$ (suma), $\prod$ (iloczyn), pojęciami: macierz, zbiór liczb naturalnych, całkowitych, rzeczywistych, a także znajomość kilku liter greckich wykorzystywanych w matematyce: α, β, γ, ν, μ, δ, Σ, σ, π, Π, ε, ξ, χ, ϕ (warto sprawdzić, jak się je czyta).

1.1. PODSTAWOWE POJĘCIA STATYSTYCZNE

Jednym z początkowych etapów badania statystycznego powinno być precyzyjne określenie **podmiotu badań** pod względem **rzeczowym** (co lub kogo badamy?), **czasowym** (w jakim okresie lub kiedy odbywają się badania?) oraz **przestrzennym** (gdzie one się odbywają?). Po spełnieniu tych założeń można przystąpić do określenia **zbiorowości statystycznej** (populacji statystycznej[1]), którą jest **ogół elementów poddanych badaniu statystycznemu**. Z kolei **jednostkami statystycznymi** są elementy zbiorowości statystycznej

---

1 Obydwa określenia są równorzędne, lecz niekiedy przyjmuje się, że **populacja statystyczna** (od łac. *populatio* – ludność) dotyczy zbioru ludności.

powiązane ze sobą logicznie, tak aby można je było przyporządkować danej populacji. Każdą z jednostek charakteryzują określone cechy i ich wartości (przykład 1.1.1).

Przykład 1.1.1

Można mówić np. o zbiorowości turystów w Ciechocinku w maju 2013 r. lub hoteli w Londynie w grudniu 2015 r. W każdej z tych zbiorowości występują jednostki statystyczne (turysta, hotel), które wyróżniają się określonymi cechami o określonych wartościach (matematycznych).

Jeśli jednostką statystyczną jest turysta, to cechą wspólną łączącą ją z innymi osobami przebywającymi w Ciechocinku jest fakt, że przyjechał tu w celach turystycznych, a cechami różniącymi są np. takie, jak: wiek, płeć, miejsce zamieszkania, liczba dzieci, z którymi przyjechał, wydatki poniesione podczas pobytu w uzdrowisku, stopień zadowolenia z usług, z jakich korzystał.

Jeśli jednostką statystyczną jest hotel, to cechą wspólną łączącą ją z innymi obiektami noclegowymi jest funkcja, jaką ma do spełnienia, czyli udzielenie noclegu, a cechami różniącymi są m.in.: liczba miejsc noclegowych, liczba zatrudnionych, kategoria (liczba gwiazdek), położenie geograficzne, koszty noclegów, wyposażenie w dodatkowe usługi, np. bar, fryzjer, Wi-Fi.

Z jakimi zbiorowościami możemy spotkać się w badaniach turystycznych? Wbrew pozorom odpowiedź na to pytanie nie jest prosta, gdyż turystyka to zjawisko złożone i może być badana w wielu aspektach.

Nie byłoby turystyki, gdyby nie było osób chętnych do jej uprawiania, dlatego najczęściej wymienianą i badaną populacją są właśnie turyści. Wraz z innymi **osobami związanymi z turystyką** stanowią oni dużą grupę, którą można podzielić m.in. na następujące kategorie:

a) turyści,

b) pracownicy zatrudnieni w przedsiębiorstwach turystycznych,

c) urzędnicy – wydziały promocji i turystyki w administracji publicznej,

d) pracownicy organizacji turystycznych (np. Polskiej Organizacji Turystycznej – POT, Polskiego Towarzystwa Turystyczno-Krajoznawczego – PTTK),

e) duchowni zajmujący się pielgrzymami,

f) mieszkańcy obszarów, na które przyjeżdżają turyści,

g) pracownicy innych instytucji związanych pośrednio lub bezpośrednio z turystyką,

h) inne osoby.

Działalnością turystyczną zajmuje się wiele podmiotów – **przedsiębiorstw**, które mogą być również podmiotami zainteresowania i badania statystycznego. Część z nich związana jest bezpośrednio lub pośrednio z turystyką. Można je podzielić na następujące branże[2]:

- hotelarstwo (cała baza noclegowa),
- gastronomia (wszystkie typy),
- transport kolejowy, lądowy, wodny, lotniczy (obsługa pasażerów i bagażu),
- organizatorzy turystyki, agenci i pośrednicy turystyczni,
- pośrednictwo finansowe (ubezpieczenie podróżnych, wymiana walut),
- administracja publiczna (informacja i administracja turystyczna, wyspecjalizowane oddziały policji, straży granicznej, ochrona lotnisk, wydawanie wiz, pozwoleń itd.),
- organizacje i instytucje turystyczne,
- handel (sprzedaż pamiątek, literatury i map turystycznych),
- kultura (obiekty muzealne, galerie, teatry, parki rozrywki),
- sport i rekreacja (obiekty sportowo-rekreacyjne, ogrody botaniczne i zoologiczne),
- zdrowie i uroda (gabinety odnowy, kosmetyczne, SPA),
- edukacja (studia w zakresie turystyki, hotelarstwa, rekreacji oraz kursy pilotażu, żeglarskie i kursy dla turystów, np. jazdy na nartach),
- inne.

Wyjazdy turystyczne mogą być związane z różnymi **wydarzeniami**, które są elementami zbiorowości statystycznej. Można badać wydarzenia:

1) kulturalne – muzealne, folklorystyczne, muzyczne, teatralne, filmowe itd.,

---

2 Światowa Organizacja Turystyki (UNWTO) w zakres gospodarki turystycznej włącza 10 sektorów gospodarki, które są zgodne z Międzynarodową Klasyfikacją Turystyczną (SICTA – Standard International Classification of Tourism Activities).

2) sportowe – jednej z dyscyplin (np. tenisa) lub kilku (np. lekkoatletyczne, olimpijskie),

3) religijne – zależą one od religii; obejmują święta religijne cykliczne, jubileusze, święta patronów (odpusty) i inne,

4) ekonomiczne – kongresy, szkolenia, targi itd.,

5) inne.

Turysta opuszcza czasowo swoje miejsce zamieszkania i udaje się do innej miejscowości, regionu, kraju. Przedmiotami badań w geografii turyzmu są zagadnienia związane z kilkoma aspektami, do których wykorzystuje się odpowiednie metody badań. Są wśród nich badania dotyczące walorów turystycznych (przyrodniczych i antropogenicznych), zagospodarowania turystycznego, ruchu turystycznego.

Przestrzeń geograficzna regionów turystycznych może być przedmiotem badań geograficznych, a badaniem mogą być objęte struktury i procesy obserwowane w jednostkach przestrzennych. Ich położenie da się określić w różny sposób, np.:

a) administracyjny, uwzględniający m.in. podział na:
- jednostki administracyjne (w Polsce: województwa, powiaty gminy),
- jednostki osadnicze (miasto, wieś),
- obszary chronione (np. parki narodowe, rezerwaty);
- państwa (europejskie, należące do UE, z innych kontynentów),

b) przyrodniczy, z punktu widzenia np.:
- strefy klimatycznej (międzyzwrotnikowej, umiarkowanej itd.),
- ukształtowania terenu (górski, wyżynny, nizinny),
- położenia względem akwenów i sieci rzecznej,
- typów krajobrazów przyrodniczych;

c) z uwzględnieniem pełnionych funkcji:
- regiony turystyczne,
- regiony rolnicze,
- regiony ekonomiczne.

Przedstawione przykłady zbiorowości statystycznych, które mogą być objęte badaniami z zakresu przestrzeni geograficznej, wskazują na szerokie spektrum dziedzin naukowych i związanych z nimi metod badawczych. Podstawowe metody statystyczne są identyczne w każdej z dyscyplin, jednak warto sięgnąć po podręczniki dedykowane naukom społecznym (Frankfort-Nachmias, Nachmias 2001; Babbie 2004, 2013; Francuz, Mackiewicz 2005; Bedyńska, Brzezicka-Rotkiewicz, red. 2007), ekonomicznym (Luszniewicz, Słaby 1996; Pociecha 2002), geograficznym (Jażdżewska 2013; Runge 2007), politologicznym (Mider, Marcinkowska 2013).

Kolejnym etapem w części wstępnej badań jest określenie **obszaru badań**, na którym się one odbywają. Najczęściej bywa on wskazywany jako jednostka administracyjna, np. województwo lubelskie, miasto Kraków, państwo Słowacja. Nie jest to jedyna możliwość zdefiniowania miejsca, w którym przeprowadzono badania statystyczne. W zależności od potrzeb, skali badań i umiejętności można posługiwać się różnymi procedurami **określania położenia geograficznego**, np. takimi, jak:

- nominalną (podaje się nazwę kraju, miasta lub obiektu – np. Polska, Kraków, Wawel),
- współrzędnych lokalnych (określa się obiekt i odległość – np. w promieniu 500 m od szlaku turystycznego, 50 m od rzeki Warty),
- porządkową (numeracja domów – np. wzdłuż ulicy Piotrkowskiej w Łodzi, od numeru 1 do 63),
- współrzędnych geograficznych (długość i szerokość geograficzna),
- topologiczną (czyli sąsiedztwa obiektów – np. wszystkie państwa graniczące z Morzem Bałtyckim).

Kiedy już zostanie określone, co lub kto będą poddane badaniom i gdzie będą się one odbywały, należy wskazać kiedy się one odbędą (przykład 1.2). W statystyce można wykorzystać dwa sposoby **określenia czasu badań**. Pierwszy polega na podaniu okresu badań, np. tydzień, miesiąc, kwartał, cały rok. Jest to badanie ciągłe zjawiska, m.in. przychodów w hotelu w ciągu roku. Drugi dotyczy badania wykonywanego najczęściej w konkretnym dniu, np. 31 grudnia, lub w dniu występowania wydarzenia turystycznego, np. mecz finałowy w Mistrzostwach Świata w siatkówce kobiet.

Przykład 1.1.2

Jeśli zadaniem badawczym jest wskazanie wielkości ruchu turystycznego w Sopocie w ciągu całego roku, to mamy do czynienia z pierwszym sposobem. Jeśli zadaniem jest podanie liczby pielgrzymów w dniu święta maryjnego (15 sierpnia) w jednym z sanktuariów maryjnych – wykorzystuje się drugi sposób.

W niektórych przypadkach badanie całej zbiorowości statystycznej – nazywane **badaniem pełnym lub wyczerpującym** – jest bardzo kosztowne (np. Narodowy Spis Powszechny – NSP[3]) lub niemożliwe do zrealizowania (np. badanie pielgrzymów podczas mszy, kontrola jakości produktów spożywczych). Wówczas wykonuje się **badania częściowe lub cząstkowe** na próbie losowej lub nielosowej (szerzej w podrozdziale 1.3).

W zależności od tego, jaki charakter miały badania, wyróżnia się dwa podejścia w statystyce: **statystykę opisową**, gdy mamy do czynienia z całą zbiorowością, lub **wnioskowanie statystyczne**, gdy wnioskujemy na temat całej populacji na podstawie pomiarów z próby[4].

W statystyce występują liczby **bezwzględne (absolutne)** i **względne**. Liczby bezwzględne są to wielkości, które otrzymujemy w wyniku mierzenia lub sumowania jednostek zbiorowości. Każda z nich, wyrażająca rozmiar badanego zjawiska, jest mianowana, np. koszty noclegu w złotych, dolarach amerykańskich, rublach lub euro. Liczby względne powstają przez porównanie ze sobą dwóch liczb. Odgrywają one ważną rolę przy porównywaniu zjawisk. Są to wielkości procentowe lub wskaźniki natężenia opisujące relacje między różnymi cechami, np. gęstość zaludnienia, przeciętne wydatki turysty na noclegi.

Przykład 1.1.3

Dla jednostek statystycznych z przykładu 1.1.1 wiek turysty będzie wartością bezwzględną, a średnie wydatki dzienne poniesione podczas pobytu

---

3 Nie każdy NSP jest badaniem pełnym, niekiedy były to również badania częściowe, np. NSP w 2011 r. w Polsce.

4 W podręczniku zaprezentowana będzie głównie statystyka opisowa, z elementami wnioskowania statystycznego.

w uzdrowisku – wartością względną. Liczba miejsc noclegowych w hotelu będzie wartością bezwzględną, a procent wykorzystania miejsc noclegowych w ciągu roku – wartością względną.

Własności, jakimi charakteryzują się jednostki statystyczne nazywamy **cechami statystycznymi**. Przystępując do badań statystycznych, określamy, ile cech będziemy analizować. Zbiorowość może być jednocechowa (jednowymiarowa) i wielocechowa. Zbiór cech dzielimy na **mierzalne (ilościowe)** i **niemierzalne (jakościowe)**.

Cechy mierzalne można przedstawić za pomocą liczb i jednostek miary, np. koszt podróży (zł), długość podróży (km), czas podróży (dni), waga bagażu (kg), liczba osób towarzyszących (osoby), wiek turysty (lata). Dzieli się je na **ciągłe** i **skokowe**. Zmienna skokowa może przyjmować wartości z określonego przedziału liczb zbioru liczb naturalnych lub całkowitych nieujemnych, np. liczba osób towarzyszących, liczba dzieci, liczba bagażu. Zmienna ciągła może przyjmować wszystkie wartości z określonego przedziału liczb zbioru liczb rzeczywistych, np. koszt biletu, odległość od miejsca zamieszkania.

Przykład 1.1.4

Dla jednostek statystycznych z przykładu 1.1.1

Wydatki dzienne turysty poniesione podczas pobytu w uzdrowisku są cechą mierzalną ciągłą. Liczba dzieci, z którymi przyjechał, są cechą mierzalną skokową. Płeć jest cechą niemierzalną. Stopień zadowolenia z usług, z jakich korzystał, jest **cechą niemierzalną stopniowalną**.

Liczba miejsc noclegowych w hotelu jest cechą mierzalną skokową, a dochody miesięczne są cechą mierzalną ciągłą. Położenie geograficzne, np. ulica, jest cechą niemierzalną, a kategoria hotelu (liczba gwiazdek) jest cechą niemierzalną stopniowalną, wyposażenie w dodatkowe usługi (np. fryzjer, bar, Wi-Fi) jest **cechą niemierzalną dwudzielną**.

1.2. SKALE POMIARU

W statystyce operuje się pojęciem **skali pomiarowej**. Jej zrozumienie jest niezwykle ważne, gdyż część metod statystycznych można wykonywać jedynie dla danych w określonej skali. Wykorzystanie – często bezmyślne – programów komputerowych dla danych w złej skali pomiarowej prowadzi do błędnych wniosków i nierzetelnych analiz statystycznych. Dlatego niezbędna jest znajomość skali, w jakiej prezentowane są dane. Wyróżnia się następujące skale pomiarowe: nominalną, porządkową, interwałową i ilorazową. Cechy jakościowe są mierzone w skalach nominalnej i porządkowej, które nazywane są skalami słabymi. Cechy ilościowe są mierzone w skalach interwałowej i ilorazowej, które nazywane są skalami mocnymi. Im silniejsza jest skala pomiaru, tym więcej metod statystycznych ma do dyspozycji badacz.

Skalę nominalną wykorzystuje się w badaniach jakościowych; określa ona np.:

a) narodowość: Francuz, Niemiec, Polak, Portugalczyk, Rosjanin, Węgier,

b) wyznanie: chrześcijanie, muzułmanie, żydzi,

c) użytkowanie ziemi: lasy, tereny zabudowane, tereny komunikacyjne, wody,

d) własność: państwowa, prywatna, kościelna, gminna,

e) branże w gospodarce: przemysł, rolnictwo, turystyka.

Należy pamiętać, że każda jednostka statystyczna badanej zbiorowości może należeć tylko do jednej klasy, czyli podział na nie jest rozłączny, oraz wszystkie jednostki statystyczne są skategoryzowane, zatem klasyfikacja na nie jest zupełna.

Dane przedstawione w skali nominalnej możemy zliczać według kategorii i podać liczebność każdej z nich lub wskazać kategorię najliczniejszą albo najmniej liczną. Nie można ich porównywać, odejmować, dzielić, porządkować (np. rosnąco), ustalać ich rangi („rangować"), a jedyną relacją, jaką można im przyporządkować, jest relacja równości. Są im dedykowane specjalne metody statystyczne, takie jak np. dominanta.

Skalę porządkową również wykorzystuje się w badaniach jakościowych, ale ma ona inne podstawy, gdyż określa cechy, które można uporządkować, takie jak:

a) wykształcenie: brak, podstawowe, zawodowe, średnie, wyższe,

b) zasięg przestrzenny wydarzenia turystycznego: regionalny, krajowy, europejski, światowy,

c) stopień trudności szlaku turystycznego: mały, średni, duży,

d) jakość usług hotelowych: liczba gwiazdek hotelu,

e) zadowolenie z usług przewodnika: dobry, trudno powiedzieć, zły.

Na danych w skali porządkowej dozwolone są wszystkie działania, które mogły być wykonane dla danych w skali nominalnej. Ponadto można porządkować je według rangi („rangować"), wskazywać relacje między nimi (równości, większości, mniejszości, np. *x* jest lepiej wykształcony niż *y*). Danych w tej skali nie odejmuje się i nie dzieli, gdyż różnica między oceną „dobry" i „zły" nie jest jednoznaczna, nie możemy precyzyjnie wskazać „o ile lepsza" jest ocena. Statystyki stosowane dla danych w skali porządkowej to m.in. dominanta, mediana, współczynnik korelacji Spearmana. Skala porządkowa jest silniejsza niż nominalna, ale słabsza niż interwałowa.

W **skali interwałowej (przedziałowej)** przedstawiane są cechy mierzalne, których wartości można odejmować i wskazać precyzyjnie różnicę pomiędzy nimi, np. temperaturę powietrza w stolicach europejskich o określonej godzinie. Nie mają one absolutnego zera, jest ono ustalone w sposób arbitralny. Nowy Rok w kalendarzu gregoriańskim przypada na inny dzień niż w kalendarzu chińskim. Wartości podawane są zawsze z jednostką miary. W przypadku np. temperatury –459,67 °F (Farenheita) jest równe 273,15 °C (Celsjusza) i 0 °K (Kelvina). Temperatura 0 °C oznacza konkretną temperaturę, a nie brak zjawiska, jak to ma miejsce np. w przypadku wagi (czy bagaż może ważyć 0 g?). Dlatego też nie ma sensu przedstawianie ich jako ilorazu. Dane w tej postaci możemy klasyfikować, „rangować", a także obliczyć różnicę (interwał) między wartościami, np. między temperaturą w nocy i w południe. Możliwe jest wskazanie relacji między obiektami (równości, większości, mniejszości). Należy pamiętać, że wzrost podany precyzyjnie w cm jest w skali ilorazowej, ale zdefiniowany jako „niski (140–160)", „średni (161–180)" i „wysoki (181 i więcej)" jest już w skali porządkowej, gdyż nie możemy podać dokładnej wartości w cm różnicy pomiędzy „średni" i „niski" itp.

W celu poznania (pomiaru) opinii, postaw, poglądów badanych turystów można w kwestionariuszu tak zadać pytanie, aby wykorzystać **skalę Likerta**. Przyjmuje się, że jest to skala przedziałowa, która nie posiada naturalnego punktu zerowego (błędem jest uznawanie za punkt 0 odpowiedzi „ani tak, ani nie"). W skali tej zakładamy, że odległości między poszczególnymi wartościami na skali są jednakowe (w przeciwieństwie do skali porządkowej).

Przykładowo, zadajmy pytanie: „Czy uważa Pan/i, że sam/a powinien/powinna ubezpieczyć się na wypadek upadłości biura podróży?".

Liczba odpowiedzi powinna być nieparzysta; możliwe odpowiedzi to:

1) zdecydowanie nie,

2) raczej nie,

3) ani tak, ani nie,

4) raczej tak,

5) zdecydowanie tak.

Należy pamiętać, że odpowiedzi „trudno powiedzieć", „nie mam zdania", „ani tak, ani nie" powinny znajdować się dokładnie pomiędzy odpowiedziami pozytywnymi i negatywnymi. W przypadku, gdy będą one umieszczone na końcu zestawu odpowiedzi (1 – nie, 2 – raczej nie, 3 – raczej tak, 4 – tak, 5 – nie mam zdania), nie można mówić o skali porządkowej. Odpowiedzi w skali Likerta są uporządkowane od najsłabszej do najsilniejszej, co oznacza, że skala ta jest jedną z odmian skali porządkowej.

Skala ilorazowa obejmuje dane o cechach mierzalnych, które podawane są z jednostką miary, np.: waga, wzrost, powierzchnia, liczba osób. Mają one absolutne 0, zatem jeśli dana jednostka ma cechę o wartości 0, to oznacza jej brak (np. jeśli liczba dzieci towarzyszących w podróży wynosi 0, to oznacza, że turyści podróżowali bez dzieci; jeśli liczba hoteli w danej miejscowości wynosi 0, to nie ma tam ani jednego hotelu). Przykładami danych w skali ilorazowej mogą być:

1. Liczba dni spędzonych poza domem (dni).

2. Wydatki na podróże jednego turysty (zł).

3. Liczba osób na wycieczce (osoby).

4. Waga bagażu turysty (kg).

5. Spożycie alkoholu przez turystę na dzień (ml/dzień).
6. Powierzchnia działki letniskowej (m^2).
7. Przeciętna cena za nocleg (USD).
8. Liczba imprez w mieście w ciągu roku (liczba/rok).
9. Zużycie paliwa w czasie podróży (l).
10. Prędkość samochodu (km/h).
11. Gęstość obiektów turystycznych (obiekty/km^2).
12. Różnica wysokości względnych (m).

Dane w tej postaci możemy klasyfikować, „rangować", a także wykonywać wszystkie obliczenia matematyczne, w tym różnicę, iloraz wartości. Można je przedstawiać w postaci ułamka:

1. Liczba osób na km^2.
2. Liczba punktów widokowych na 10 km^2.
3. Spożycie lodów w szt. na 1 turystę.
4. Liczba stacji benzynowych na 1000 mieszkańców.
5. Przeciętna liczba miejsc noclegowych na 1 obiekt.
6. Liczba imprez na dzień w ciągu roku w % (liczba imprez/365 × 100).
7. Powierzchnia obszarów chronionych w stosunku do powierzchni gminy (w %).

Możliwe jest wskazanie relacji między obiektami (równości, większości, mniejszości).

Przyjęto, że skala ilorazowa jest skalą najsilniejszą, a nominalna – najsłabszą. Można przejść z jednej skali w inną, jednak nie w każdą stronę. Na przykład dane w skali interwałowej i ilorazowej można przedstawić w skali porządkowej (np. wzrost człowieka: wysoki, średni, niski, lub wiek: przedziały 20–29, 30–39, 40–49). W drugą stronę jest to niemożliwe.

Przykład 1.2.1

Na podstawie danych zawartych w tab. 1.2.1 określ zbiorowość statystyczną i jednostkę statystyczną. Jakie cechy statystyczne można zbadać w tej zbiorowości? Jaka będzie ich skala pomiarowa?

Tabela 1.2.1. Tereny rekreacyjne w Tomaszowie Mazowieckim w 2011 r.

Tereny rekreacyjne	Powierzchnia (ha)
Lasy	522,0
Ogródki działkowe	73,0
Cmentarze	50,0
Rezerwaty/skanseny	30,0
Parki	15,0
MOSiR	12,0
Zieleń osiedlowa	8,2
Zieleńce	8,1

Źródło: opracowanie J. Brockhusen na podstawie danych z Urzędu Miasta w Tomaszowie Mazowieckim.

Zbiorowością statystyczną będą wszystkie tereny rekreacyjne, jakie występowały w 2011 r. w Tomaszowie Mazowieckim. Na podstawie tab. 1.2.1 nie możemy określić ich liczebności (np. liczby parków), a jedynie sumę powierzchni. Jednostką statystyczną będzie jeden rodzaj terenu (rekreacyjny), gdzie mieszkańcy i goście mogą uprawiać rekreację.

W zależności od celu przeprowadzanego badania można brać pod uwagę następujące cechy statystyczne opisujące tę zbiorowość:

a) liczba ławek – skala ilorazowa,

b) liczba koszy na śmieci – skala ilorazowa,

c) powierzchnia terenu – skala ilorazowa,

d) rodzaj terenu (lasy, parki itd.) – skala nominalna,

e) własność terenu (gminna, kościelna, państwowa, spółdzielcza itd.) – skala nominalna,

f) dostępność terenu dla wszystkich (T/N) – skala nominalna,

g) ocena zagospodarowania rekreacyjnego terenu (bardzo dobra, dobra, przeciętna, zła, bardzo zła) – skala porządkowa lub Likerta,

h) temperatura powietrza w południe – skala interwałowa.

1.3. METODA REPREZENTACYJNA

Nie zawsze można zbadać całą zbiorowość statystyczną, wobec tego badania prowadzone są w części zbiorowości – na próbie losowej lub nielosowej. **Próbą losową** nazywamy część populacji statystycznej wybraną za pomocą określonego sposobu losowania w celu zbadania własności całej populacji[5]. Aby informacje pochodzące z próby były obiektywne i wiarygodne, musi być ona losowana w specjalny sposób – nie może być pobierana w sposób tendencyjny. Oznacza to, że fakt zaliczenia obiektu do próby nie może zależeć od wielkości cechy przypisanej obiektowi. W przypadku prób nielosowych wnioskowanie statystyczne jest obarczone bardzo dużym ryzykiem popełnienia błędu przy wnioskowaniu.

Metoda reprezentacyjna (lub reprezentatywna) polega na tym, że na podstawie losowo wybranej próby wnioskujemy o całości populacji. Aby wnioskowanie było poprawne, badana część zbiorowości musi być wybrana w sposób reprezentatywny. Wybór jednostek statystycznych do próby powinien uwzględniać strukturę badanej zbiorowości, odpowiedni sposób losowania oraz liczebność próby. Używa się różnych technik losowania: 1) ze zwracaniem elementów, tzw. próba z powtórzeniami, lub 2) bez ich zwracania, tzw. próba bez powtórzeń. Pierwszą nazywa się losowaniem niezależnym, drugą – zależnym.

Wyróżnia się następujące sposoby **losowania próby**:

- bezpośrednie,
- systematyczne,
- z wykorzystaniem liczb losowych,
- warstwowe.

Jeśli populacja wybrana do analizy jest nieduża, to można każdej badanej jednostce nadać numer, zapisać go na kartce, a następnie po wymieszaniu kartek dokonać losowania bez zwracania. Jest to **bezpośredni rodzaj losowania**.

---

5 Wykorzystuje się często **estymację**, która jest procesem wnioskowania o numerycznych wartościach nieznanych wielkości charakteryzujących populację generalną na podstawie niekompletnych danych, takich jak próba (Kendall, Buckland 1986), lub **predykcję**, która jest procesem określania przyszłych wielkości zmiennych losowych.

Losowanie systematyczne polega na wybieraniu elementów próbki co pewien z góry ustalony krok. Długość kroku (interwału) jest dobierana każdorazowo w zależności od liczebności próby.

Przykład 1.3.1

W biurze podróży na liście klientów, którzy korzystali z organizowanych przez biuro wycieczek do Paryża znajduje się 2000 nazwisk. Aby przeprowadzić badania wśród tej grupy, należy określić liczebność próby. Jeśli zamierzamy wylosować 100 osób, należy wybrać z listy co 20. osobę (2000/100 = 20), zaczynając w dowolnym miejscu na liście.

Jeśli wielkość próby określamy w procentach (np. 15% zbiorowości), to zaczynamy od obliczenia odsetka (15% × 2000 = 300 osób), a następnie obliczamy interwał (2000 / 300 = 6,7). Jest on ułamkiem – nie możemy losować z listy co 6,7 osób – dlatego zaokrąglamy go w górę i losujemy od dowolnego miejsca na liście co 7. osobę, aż uzyskamy 300-osobową listę nazwisk uczestników wycieczek do Paryża wytypowanych do badania.

Aby skorzystać z **liczb losowych**, można posłużyć się programem komputerowym do generowania liczb losowych bądź tablicami liczb losowych (załącznik 5). Liczby losowe mogą być jednocyfrowe, np. 2, 4, 5, 8, 3, dwucyfrowe, np. 23, 02, 90, 01, trzycyfrowe, np. 234, 567, 012, 453, 003 987. Każdemu elementowi zbiorowości statystycznej przyporządkowujemy numer od 1 do *n*, a następnie odczytujemy z tablic odpowiednią liczbę jednostek w zależności od wielkości próby.

Przykład 1.3.2

Ze zbiorowości liczącej 300 elementów należy wybrać próbę 30-elementową, korzystając z liczb losowych.

Algorytm[6] postępowania jest następujący:

Numerujemy elementy zbioru od 1 do 300.

Ze zbioru liczb losowych wybieramy liczby trzycyfrowe.

---

6 Algorytm to opis wykonania w określonym porządku skończonej liczby operacji prowadzących do otrzymania rozwiązania zadania.

Jeśli pierwsza wybrana liczba jest mniejsza bądź równa 300, to element o tym numerze będzie wylosowany, a jeśli wybrana liczba jest większa od 300, to ją odrzucamy.

Odczytujemy kolejną liczbę losową z tablicy (załącznik 5).

Jeśli wybrana liczba jest mniejsza bądź równa 300, to element o tym numerze będzie wylosowany, a jeśli wybrana liczba jest większa od 300, to ją odrzucamy.

Kroki 4. i 5. powtarzamy tak długo, aż otrzymamy 30 liczb, czyli tyle, ile chcemy, aby zawierała próba.

Losowanie warstwowe (kwotowe) próby stosuje się wtedy, gdy zbiorowość składa się z podgrup o różnej liczbie elementów, np. jeśli badaniu poddano miejsca noclegowe w Warszawie (np. motele, hotele, hostele, kwatery prywatne, kempingi), których liczebność jest mocno zróżnicowana, to z każdej grupy obiektów można pobrać liczbę elementów proporcjonalną (np. 10%) do liczebności warstwy. Losowanie warstwy wymaga znajomości rozkładów zmiennych całej populacji (np. przy badaniach opinii mieszkańców miasta na temat zagospodarowania terenów rekreacyjnych należy na podstawie danych Głównego Urzędu Statystycznego (GUS) poznać strukturę wieku i płci mieszkańców tego miasta, a następnie wybrać do badań odpowiednią kwotę według wieku i płci).

Przykład 1.3.3

Na Wydziale Nauk Geograficznych Uniwersytetu Łódzkiego w roku akademickim 2015/2016 było 460 studentów na I roku na wszystkich kierunkach prowadzanych przez ten wydział (tab. 1.3.1). Wybierz 25-procentową próbę studentów do badań na temat ich wyjazdów turystycznych.

Tabela 1.3.1. Struktura studentów I roku na Wydziale Nauk Geograficznych Uniwersytetu Łódzkiego w roku akademickim 2015/2016

Lp.	Kierunek studiów	Liczba studentów	Liczba studentów w warstwie próby
1.	Geografia	60	25% × 60 = 15
2.	Gospodarka przestrzenna	100	25% × 100 = 25

Tabela 1.3.1. cd.

Lp.	Kierunek studiów	Liczba studentów	Liczba studentów w warstwie próby
3.	Geoinformacja	60	25% × 60 = 15
4.	Geomonitoring	60	25% × 60 = 15
5.	Studia regionalne	40	25% × 40 = 10
6.	Turystyka i rekreacja	140	25% × 150 = 35
Razem		460	115

Źródło: dane umowne.

Algorytm. Zbiorowość liczy 460 osób, czyli badania będą przeprowadzone na grupie 115 studentów (25% z 460). Z każdego kierunku należy wybrać 25% liczby studentów, czyli następującą liczbę osób: z geografii 15, z gospodarki przestrzennej 25, z geoinformacji 15, z geomonitoringu 15, ze studiów regionalnych 10, z turystyki i rekreacji 35. Losując odpowiednią liczbę studentów z każdego kierunku, można wykorzystać ich listę z dziekanatu i skorzystać z metody losowania systematycznego.

Ostateczny wybór metody należy zawsze do przeprowadzającego badanie. Musi on się zastanowić nad tym, czy wszystkie elementy zbiorowości miały **jednakowe szanse** bycia wybranymi.

Jedną z często wykorzystywanych w badaniach turystycznych metod pozyskiwania uczestników jest **dobór celowy**. Jest to **metoda nielosowego doboru próby**, w której badacz powinien mieć dobrą wiedzę na temat populacji, ale nie może określić jej pełnego wykazu. Stosowany jest w małych zbiorowościach statystycznych lub w badaniach pilotażowych. Należy mieć na uwadze, że wnioskowanie statystyczne na temat całej populacji jest obarczone ryzykiem, choć uznaje się, że dobrze dobrana próba celowa jest zbliżona do reprezentatywnej (więcej: Frankfort-Nachmias, Nachmias 2001). Innym nielosowym sposobem doboru uczestników do badań jest **metoda kuli śnieżnej**, która polega na tym, że ankietowani polecają kolejne osoby do badań (czyli zazwyczaj swoich znajomych). Wyniki analiz trudno więc uogólniać na całą populację, np. turystów, mieszkańców czy inne zbiorowości. Jest uznawana za niereprezentatywną.

1.4. TECHNIKI ZBIERANIA INFORMACJI

Wybór technik zbierania informacji zależy od celu badania, możliwości dotarcia do podmiotu badań, kosztochłonności i wielu innych aspektów, które pojawiają się w trakcie tej procedury. Na wstępie należy założyć, czy będą to badania pełne czy częściowe, a jeśli te drugie, to czy jest możliwość pobrania próby losowej lub nielosowej. Można samodzielnie zbierać informacje o jednostkach statystycznych albo skorzystać ze źródeł wtórnych – zebranych przez inne osoby lub instytucje, np. Główny Urząd Statystyczny (GUS), Światową Organizację Turystyki (WTO) lub instytuty turystyki.

Spośród wielu metod zbierania informacji przydatnych w badaniach turystycznych można wyróżnić:

1) inwentaryzację za pomocą kart inwentaryzacyjnych,

2) inwentaryzację fotograficzną,

3) bieżącą rejestrację (pomiar),

4) bieżącą samorejestrację,

5) spisy,

6) badania społeczne, np.
- badania ankietowe,
- obserwację uczestniczącą,
- wywiady,
- sondaże,

7) źródła wtórne.

Inwentaryzacja w terenie może dotyczyć wielu aspektów związanych z turystyką, m.in. walorów i zagospodarowania turystycznego. Powinna być ona przemyślana i przygotowana wcześniej. Warto znaleźć w literaturze podobne badania, aby mieć możliwość odniesienia się do nich w analizie, np. inwentaryzację obiektów krajoznawczych PTTK[7]. Na wstępie należy przygotować **karty inwentaryzacyjne**, które w miarę możliwości kompleksowo opiszą badany obiekt (załącznik 3). Jednym z elementów inwentaryzacji w terenie albo samodzielnym zadaniem może być **inwentaryzacja fotograficzna**

---

7 https://www.pttk.pl – zakładka „Krajoznawstwo", „Inwentaryzacja krajoznawcza".

(załącznik 4). Daje ona wiele cennych informacji wizualnych o obiektach nieożywionych, a oprócz tego może być dokumentem obrazującym zachowania turystyczne, ruch turystyczny, a także nastrój osób, pogodę, otoczenie i inne.

Dobrym źródłem informacji jest **bieżąca rejestracja (pomiar)** wykonywana przez pewien okres. Można ją przeprowadzić w różnych instytucjach związanych z turystyką, np. w hotelach, muzeach, przedsiębiorstwach środków transportu (por. podrozdział 1.2). Dzięki bieżącej rejestracji możliwe jest uwzględnienie w badaniach czasu z dokładnością np. do 1 godziny.

Przykład 1.4.1

Zamierzamy zbadać ruch turystyczny w Muzeum Włókiennictwa w Łodzi w 2014 r. Interesuje nas, ile osób korzysta z muzeum w określonych porach dnia, w poszczególnych dniach tygodnia oraz miesiącach. Można to zrobić w różny sposób. Jeśli mamy możliwość skorzystania z raportów kasowych (zakładamy, że jest to muzeum płatne), to wystarczy zliczyć liczbę osób w poszczególnych godzinach otwarcia placówki. Jeśli wydruki kasowe są niedostępne, to nie pozostaje nic innego, jak zliczanie odwiedzających przez cały okres badań. Jeśli badania wykonywane są osobiście, można zebrać dodatkowe informacje na temat ruchu turystycznego, np. podzielić odwiedzających ze względu na płeć, ewentualnie wiek, bez konieczności prowadzenia badań kwestionariuszowych czy wywiadów z badanymi.

Bieżąca samorejestracja jest jedną z technik zbierania informacji, która jest przygotowana pod kątem zebrania jak największej ilości informacji w trakcie uprawiania turystyki. Dzięki niej można poznać zachowania turystyczne: sposób spędzania czasu, motywacje, wydatki itd. W zależności od celu badania należy osobie, która zgodziła się prowadzić samorejestrację, przygotować tabelę, w której będzie wpisywać swój rozkład zajęć w ciągu dnia (lub doby), poniesione wydatki, towarzystwo i inne. Można ją też wyposażyć w odbiornik GPS do zbierania informacji o jej aktywności w terenie.

Przykład 1.4.2

Zamierzamy zbadać zachowania turystyczne oraz wydatki emerytów przebywających w Ciechocinku. Interesuje nas, jaką kwotę seniorzy zostawiają w mieście podczas pobytu, jaka jest struktura ich wydatków, jak spędzają

czas. Osobom, które zgodziły się na samorejestrację przygotowujemy tabelę wydatków z podziałem na grupy, np. noclegi, wyżywienie podstawowe (śniadanie, obiad, kolacja), wyżywienie dodatkowe (kawiarnia, lody, owoce, alkohol itp.), leki, ubrania, wydatki na kulturę (książki, kino, teatr), wycieczki, usługi (fryzjer, kosmetyczka), pamiątki i inne. Ponadto dzielimy każdy dzień na godziny i prosimy o wpisanie aktywności w ciągu dnia.

Przykładami **spisów** są Narodowy Spis Powszechny czy Narodowy Spis Rolny. Są to trudne, czasochłonne i kosztochłonne badania, zazwyczaj wyczerpujące zasób danych[8]. W badaniach turystycznych spisy mogą obejmować wspomnianą powyżej inwentaryzację, np. zagospodarowania turystycznego lub spis turystów wchodzących do parku narodowego.

Badania społeczne stosuje się w socjologii, psychologii i geografii społecznej, są one szczegółowo opisane w podręcznikach (Frankfort-Nachmias, Nachmias 2001; Francuz, Mackiewicz 2005; Bedyńska, Brzezicka-Rotkiewicz, red. 2007). Mają zazwyczaj postać badań ankietowych, obserwacji uczestniczącej, wywiadów, sondaży.

Bardzo dobrym źródłem informacji turystycznej są źródła wtórne. Mogą to być np. dane zbierane przez GUS w Polsce, który umieszcza swoje wyniki na stronie internetowej https://stat.gov.pl lub publikuje w postaci roczników statystycznych. Ta pierwsza forma jest dużo lepsza, gdyż dane można bez problemu importować na swój komputer w postaci plików, które od razu można poddać analizie statystycznej w arkuszach kalkulacyjnych lub programach statystycznych, bez konieczności ich przepisywania. Nowością GUS jest internetowy portal geostatystyczny[9], prezentujący wynikowe informacje statystyczne w postaci kartogramów lub kartodiagramów. W zakładce „Turystyka" znajdują się dane z tzw. Banku Danych Lokalnych, które można zaprezentować na mapie, m.in. placówki gastronomiczne w obiektach turystycznych, hotele, motele i pensjonaty według kategorii, noclegi udzielone turystom według wybranych krajów, a także wybrane wskaźniki.

Oprócz danych liczbowych dobrym źródłem są opracowania kartograficzne (takie jak mapy topograficzne w skali 1 : 25 000 lub większej), na

8 W 2011 r. po raz pierwszy w Polsce NSP miał charakter spisu częściowego.

9 https://geo.stat.gov.pl.

których można odczytać wiele cennych informacji na temat walorów turystycznych obszaru. Korzystając z mapy topograficznej w postaci papierowej, należy zwrócić uwagę na datę jej wykonania i zweryfikować w terenie jej aktualność. Innym współczesnym źródłem informacji o terenie są wektorowe mapy topograficzne oferowane przez ośrodki geodezyjne (Centralny Ośrodek Dokumentacji Geodezyjnej i Kartograficznej – CODGiK), które można kupić w postaci papierowej lub wektorowej. Ta druga forma daje możliwość ich przestrzennej analizy z wykorzystaniem oprogramowania GIS (Geographic Information Systems). Kolejnymi źródłami informacji wtórnej są przewodniki turystyczne, plany miast z obiektami turystycznymi. Do badań porównawczych warto zaopatrzyć się w archiwalne materiały kartograficzne, takie jak *Topograficzna mapa Królestwa Polskiego z 1863 r.*, mapy topograficzne Polski[10] czy opracowania książkowe: *Słownik geograficzny Królestwa Polskiego i innych krajów słowiańskich* (1880–1914)[11], *Słownik geograficzno-krajoznawczy Polski* z 1983 r., *Leksykon miast polskich* z 1998 r.

Internet jest dobrym źródłem informacji turystycznej, o ile można sprawdzić rzetelność podawanych danych. Do takich źródeł zalicza się też strony internetowe:

- instytucji rządowych (na całym świecie mają one zazwyczaj rozszerzenie gov.), np. portal amerykański: www.bea.gov, japoński: www.stat.go.jp, polski: stat.gov.pl, francuski: www.insee.fr, hiszpański: www.ine.es, włoski: www.istat.it, brytyjski: www.ons.gov.uk, izraelski: cbs.gov.il,
- Unii Europejskiej: ec.europa.eu/eurostat/web/tourism/data/database;
- organizacji powołanych do badania ruchu turystycznego i zjawisk turystycznych, np. WTO (Światowej Organizacji Turystycznej): stat.wto.org, POT (Polskiej Organizacji Turystycznej): www.pot.gov.pl, WTTC (World Travel and Tourism Council): www.wttc.org/research/economic-research,
- Organizacji Współpracy i Rozwoju (OECD): www.oecd.org/cfe/tourism,

a także:

- portale prezentujące informacje o ruchu pasażerskim i lotniskach: www.world-airport-codes.com,

10 http://polski.mapywig.org/news.php.

11 http://dir.icm.edu.pl/pl/Slownik_geograficzny.

- portale prezentujące ofertę noclegową, np. www.booking.com,
- portale prezentujące ofertę gastronomiczną, np. www.gastonauci.pl,
- geoportale, np. geoportal.gov.pl, geoportaltatry.pl,
- blogi turystyczne, np. www.travelblog.org,
- portale prezentujące diagnozę społeczną, np. www.diagnoza.com.

1.5. ZADANIA

ZADANIE 1.5.1

Odpowiedz na pytania: Co to jest zbiorowość statystyczna? Jak powinna być określona? Jakie skale pomiarowe stosujemy w statystyce?

ZADANIE 1.5.2

Na podstawie danych zawartych w tab. 1.5.1 nazwij: zbiorowość statystyczną, jednostkę statystyczną, liczebność zbiorowości. W jakiej skali pomiarowej jest podana cecha dotycząca parków w Piotrkowie Trybunalskim? Jakie inne informacje o każdym parku można pozyskać? W jakiej skali pomiarowej będą określone?

Tabela 1.5.1. Parki miejskie w Piotrkowie Trybunalskim w 2011 r.

Park	Powierzchnia (ha)
im. ks. J. Poniatowskiego	7,2
Belzacki	7,5
im. Jana Pawła II	6,0
im. kard. S. Wyszyńskiego	1,5

Źródło: opracowanie Lena Florczyk na podstawie danych z Urzędu Miasta w Piotrkowie Trybunalskim.

ZADANIE 1.5.3

Podane niżej obiekty są elementami określonych zbiorowości statystycznych. Wskaż zbiorowość, do której mogą przynależeć.

Nazwa jednostki statystycznej	Zbiorowość statystyczna
Hotel „Qubus" w Łodzi	
Kuracjusz w Nałęczowie	
Góra Śnieżka	
Gniezno	
Cmentarz Łyczakowski	
Muzeum na Wawelu	
Szlak św. Jakuba	
Lotnisko Heathrow	
Festiwal Filmowy w Cannes	
Bar mleczny w Gdańsku	
Stadion Real Madryt	
Zegar Big Ben w Londynie	
Kamping w Mielnie	
Pol'and'Rock Festival w Kostrzynie nad Odrą	
Lazurowe Wybrzeże	
Ojcowski Park Narodowy	
Wyspy Kanaryjskie	

ZADANIE 1.5.4

Na podstawie danych GUS (https://stat.gov.pl) wybierz najnowsze informacje o hotelach w Polsce w województwie mazowieckim. Określ zbiorowość statystyczną, jednostkę statystyczną i liczebność. Jakie cechy mogą charakteryzować tę zbiorowość? Określ ich skalę pomiarową.

ZADANIE 1.5.5

Jaką skalę pomiarową reprezentują wymienione w poniższej tabeli cechy, z którymi można spotkać się podczas badań statystycznych? Jeśli to możliwe, określ jednostkę miary.

Cecha statystyczna	Skala	Jednostka miary
Liczba turystów na Helu		
Waga bagażu oddanego do odprawy na lotnisku Okęcie		

Cecha statystyczna	Skala	Jednostka miary
Kategoria hoteli w Lizbonie		
Wykształcenie pracowników biur podróży		
Liczba dzieci towarzyszących rodzicom na wycieczce		
Liczba hoteli w Poznaniu		
Płeć pielgrzymów		
Narodowość podróżnych w pociągu do Madrytu		
Zadowolenie z usług hotelu		
Wydatki na żywność pielgrzymów do Rzymu		
Temperatura powietrza w Suwałkach o godz. 12^{00}		
Grubość pokrywy śnieżnej na Giewoncie		

ZADANIE 1.5.6

Przedsiębiorstwo komunikacyjne zatrudnia 3000 kierowców. Każdy z nich posiada numer identyfikacyjny. W celu zbadania sposobów wykorzystania urlopu przez pracowników postanowiono wybrać losowo 100 z nich. Wyjaśnij, jak pobrać próbę w sposób systematyczny lub warstwowy.

ZADANIE 1.5.7

W województwie X w 2016 r. było 1100 obiektów noclegowych, w tym: hotele, motele, pensjonaty, schroniska i kwatery prywatne (tab. 1.5.2). Wyjaśnij, jak pobrać warstwową próbę losową, jeśli ma ona liczyć 165 obiektów.

Tabela 1.5.2. Obiekty noclegowe w województwie X w 2016 r.

Rodzaj obiektu	Liczba
Hotele	120
Motele	180
Pensjonaty	300
Schroniska	25
Kwatery prywatne	475
Razem	1100

Źródło: dane umowne.

ZADANIE 1.5.8

Korzystając z podanych liczb losowych, wybierz 10-osobową próbę z grupy 50 studentów prawa I roku Uniwersytetu Jagiellońskiego: 4 2 1 5 5 4 3 7 8 7 0 7 0 5 2 7 0 9 1 4 0 4 4 5 0 1 2 6 1 4 8 6 4 7 1 6 4 7 5 8 7 2 1 0 7 6 1 0 3 5 5 0 3 7 1 7 1 7 1 9 8 6 3 2 6 4 5 4 5 1 6 3 0 7 6 8 4 4 0 3 0 7 0 1 9 3 4 1 6 2 7 8 6 2 9 7 1 3 2 8 9 2 2 0 8 9.

ZADANIE 1.5.9

Wylosuj 5-procentową próbę gmin w Polsce w 2015 r., wykorzystując metodę warstwową (w zależności od liczby gmin w województwie). Skorzystaj ze strony: https://stat.gov.pl.

ZADANIE 1.5.10

Wylosuj 10-procentową próbę obiektów noclegowych w województwach koszalińskim, łódzkim i nowosądeckim (tab. 1.5.3). Jaka metoda losowania będzie najlepsza?

Tabela 1.5.3. Obiekty noclegowe turystyki według rodzaju w województwach: koszalińskim, łódzkim i nowosądeckim w 1994 r.

Województwo	Ośrodki		Domy pracy twórczej	Kempingi	Pola biwakowe	Pokoje gościnne (kwatery prywatne)
	kolonijne	szkoleniowo-wypoczynkowe				
Koszalińskie	38	18	1	5	17	92
Łódzkie	2	4	0	9	2	11
Nowosądeckie	3	16	14	14	27	265

Źródło: GUS (1995c).

ZADANIE 1.5.11

Do biura podróży w ciągu dnia weszło 110 osób. Na podstawie danych wybierz losowo trzy próby o różnej liczebności (10 osób, 20 osób i 30 osób). Oblicz średni wiek klienta w każdej próbie. Porównaj wyniki ze sobą i średnim wiekiem całej zbiorowości.

Wiek osób: 18, 25, 26, 45, 58, 69, 78, 25, 25, 46, 36, 38, 49, 58, 51, 62, 65, 18, 70, 81, 36, 46, 52, 69, 57, 58, 25, 47, 36, 58, 56, 41, 43, 29, 27, 40, 36, 39, 19, 54, 58, 68, 62, 47, 43, 25, 29, 53, 19, 20, 36, 88, 65, 63, 47, 42, 55, 59, 33, 30, 18, 18, 20, 23, 28, 59, 58, 47, 41, 36, 39, 23, 90, 20, 26, 69, 36, 36, 24, 50, 45, 47, 18, 70, 41, 75, 36, 38, 24, 25, 66, 56, 80, 47, 65, 45, 19, 40, 47, 36, 38, 66, 56, 47, 58, 80, 45, 23, 38, 36, 63.

ZADANIE 1.5.12

Ania miała przeprowadzić badania ankietowe na temat rekreacji wśród 500 mieszkańców Kalisza. Postanowiła wykonać je wśród znajomych, którzy do dalszych badań polecali swoich znajomych. Czy jej próba będzie reprezentatywna i wnioski wysnute na podstawie pobranej próby można uogólnić na wszystkich mieszkańców Kalisza?

ZADANIE 1.5.13

Studia licencjackie trwają trzy lata. Na każdym roku jest od 1 do 4 grup studentów liczących po około 30 osób. Masz przeprowadzić ankietę wśród kolegów, ale nie ma potrzeby przeprowadzania jej ze wszystkimi. Odpowiedz na pytania: Jak wybierzesz próbę? Czy wystarczy, jeśli przeprowadzisz ją tylko w twojej grupie, a wyniki uogólnisz? Uzasadnij odpowiedź.

ZADANIE 1.5.14

Masz przeprowadzić badania reprezentatywne z 500 pełnoletnimi mieszkańcami twojego miasta. Jak to zrobisz?

1) wejdziesz do dużego kina (na film dla widzów powyżej 18 lat) i przeprowadzisz ankietę?

2) poprosisz znajomych i ich znajomych o kontakty telefoniczne do mieszkańców?

3) pójdziesz do przychodni lekarskiej (tam mają nazwiska i adresy mieszkańców) i poprosisz o pozwolenie na losowe wybranie 500 osób, a następnie pójdziesz do nich przeprowadzić ankietę?

4) będziesz przesiadywał w kawiarni przez dwa tygodnie i przepytasz 500 osób?

5) opublikujesz ankietę w mediach społecznościowych?

6) poprosisz w Urzędzie Wojewódzkim o listę wyborczą mieszkańców miasta i wylosujesz 500 osób, a potem pójdziesz do nich przeprowadzić ankietę?

Która metoda będzie reprezentatywna i dlaczego?

ZADANIE 1.5.15

Podziel polskie miasta na małe (do 20 tys. ludności), średnie (20–100 tys. ludności) oraz duże (powyżej 100 tys.) i wylosuj 20-procentową próbę warstwową do badań ruchu turystycznego. Skorzystaj z danych o liczbie mieszkańców miast na stronie: https://stat.gov.pl.

ZADANIE 1.5.16

Masz przeprowadzić badania bazy gastronomicznej wzdłuż ulicy Piotrkowskiej w Łodzi, znanej z licznej oferty kawiarni, pubów i restauracji. Możesz posłużyć się metodą reprezentatywną. Jak wybierzesz obiekty do badań?

ZADANIE 1.5.17

Skorzystaj z Internetu, wejdź na stronę Unii Europejskiej: https://ec.europa.eu/eurostat/web/tourism/data/main-tables, poszukaj informacji *Number of nights spent by country/world region of destination*. Wybierz wartości dla ostatnich trzech lat dla wszystkich podanych krajów. Zobacz ich prezentację w formie wykresu i mapy.

ZADANIE 1.5.18

Zaproponuj badania działek letniskowych. Przy każdym pytaniu podaj skalę pomiaru.

1.6. ODPOWIEDZI DO WYBRANYCH ZADAŃ

ZADANIE 1.5.2

Zbiorowością statystyczną będą parki w Piotrkowie Trybunalskim w 2011 r., a jednostką statystyczną jeden park. Liczebność tej zbiorowości wynosi 4, gdyż tyle było parków w tym mieście w 2011 r. Powierzchnia parku jest cechą mierzalną ciągłą w skali pomiarowej ilorazowej.

W zależności od celu przeprowadzanego badania można brać pod uwagę następujące cechy statystyczne:

1) liczbę osób odwiedzających park, liczbę ławek – skala ilorazowa,

2) powierzchnię parku, długość ogrodzenia, długość ścieżek – skala ilorazowa,

3) ocenę stanu czystości, bezpieczeństwa parku – skala porządkowa,

4) własność parku – skala nominalna,

5) temperatura w południe – skala interwałowa,

6) czy park jest zabytkowy (tak lub nie) – skala nominalna.

ZADANIE 1.5.3 (WYBRANE ODPOWIEDZI)

Hotel „Qubus" w Łodzi – hotele w Łodzi w 2011 r.

Kuracjusz w Nałęczowie – turyści przebywający w Nałęczowie w 2016 r.

Festiwal Filmowy w Cannes – festiwale filmowe na świecie w XXI w.

Zegar Big Ben w Londynie – atrakcje turystyczne Londynu w 2016 r.

Lazurowe Wybrzeże – regiony turystyczne Francji w 2017 r.

Ojcowski Park Narodowy – parki narodowe w Polsce w 2015 r.

ZADANIE 1.5.5

Cecha statystyczna	Skala
Liczba turystów na Helu	ilorazowa
Waga bagażu oddanego do odprawy na lotnisku Okęcie	ilorazowa
Kategoria hoteli w Lizbonie	porządkowa
Wykształcenie pracowników biur podróży	porządkowa

Cecha statystyczna	Skala
Liczba dzieci towarzyszących rodzicom na wycieczce	ilorazowa
Liczba hoteli w Poznaniu	ilorazowa
Płeć pielgrzymów	nominalna
Narodowość podróżnych w pociągu do Madrytu	nominalna
Zadowolenie z usług hotelu	porządkowa
Wydatki na żywność pielgrzymów do Rzymu	ilorazowa
Temperatura powietrza w Suwałkach o godz. 12^{00}	interwałowa
Grubość pokrywy śnieżnej na Giewoncie	ilorazowa

Źródło: opracowanie własne.

ZADANIE 1.5.6

Aby wybrać próbę w sposób systematyczny, należy sporządzić listę pracowników lub ich numerów identyfikacyjnych i wybrać co 30. osobę, poczynając od dowolnego miejsca na liście. Jeśli chcemy pobrać próbę warstwową, to należy podzielić kierowców na grupy (warstwy), np. kobiety i mężczyzn lub ze względu na staż pracy (do 10 lat, 10–20, powyżej 20 lat), i z każdej warstwy wylosować 3,33-procentową próbę, ponieważ próba licząca 100 to 3,33% populacji.

ZADANIE 1.5.7

Próba licząca 165 obiektów to 15% zbiorowości. Należy wziąć do badania po 15% z każdej warstwy: 15% ze 120 hoteli, czyli 18; 15% ze 180 moteli, czyli 27; 15% z 300 pensjonatów, czyli 45; 15% z 25 schronisk, czyli 3,75 ≈ 4; 15% z 475 kwater prywatnych, czyli 71,25 ≈ 71.

ZADANIE 1.5.8

Należy wybierać liczby dwucyfrowe. Ich wskazywanie można zacząć w dowolnym miejscu tablicy liczb losowych (załącznik 5). Dlatego przez różnych badaczy mogą być wybrani inni studenci do badań. Jeśli zacznie się wybór studentów do próby od początku przykładowej listy liczb losowych, to będą to osoby z numerami: 42, 15, 37, 07, 05, 27, 09, 14, 04, 45.

ZADANIE 1.5.12

Nie, gdyż Ania nie miała kontroli nad wyborem kolejnych respondentów. Ankietowani mogą zaliczać się do podgrupy mieszkańców Kalisza; nie wszyscy mieli szansę bycia wylosowanymi. Lepiej byłoby, gdyby ograniczyła swoje badania do mniejszej zbiorowości statystycznej, np. rówieśników, i przeprowadziła je poprawnie.

ZADANIE 1.5.14

Pamiętając o tym, że wszystkie elementy zbiorowości muszą mieć jednakowe szanse bycia wybranymi, najodpowiedniejsza będzie ostatnia metoda.

ZADANIE 1.5.17

Dane Eurostat w języku angielskim dotyczące liczby udzielonych noclegów według krajów w latach 2015–2017 uporządkowano alfabetycznie.

Tabela 1.5.4. Number of nights spent by country/world region of destination 2015–2017

Country	2015	2016	2017
	number		
European Union (28 countries)	:	:	:
Austria	82 754 133	88 691 566	89 058 720
Belgium	76 889 873	89 334 394	77 748 107[b]
Bulgaria	12 928 014	14 885 649	21 658 233
Croatia	33 814 257	22 598 566	25 839 243
Cyprus	13 404 592	14 092 285	15 171 290
Czechia	116 883 483	129 748 871	136 403 342
Denmark	101 047 950	83 270 005	80 189 098[b]
Estonia	10 109 006	10 716 235	11 441 839
Finland	105 210 291	105 468 301	109 586 331
France	999 268 210	978 136 095	980 829 271
Germany	1 031 784 641	1 060 510 060	1 002 224 318
Greece	54 157 096	51 648 324	58 405 097

Country	2015	2016	2017
	number		
Hungary	59 812 837	59 661 406	60 772 356
Iceland	:	:	:
Ireland	48 369 739	51 219 852	53 457 933
Italy	234 286 402	262 776 247	277 641 451
Latvia	11 808 952	12 957 612	12 066 489
Liechtenstein	:	:	:
Lithuania	16 025 099	15 541 811	15 669 617
Luxembourg	8 698 423	8 631 469	10 459 661[b]
Malta	2 380 585	2 824 363	3 103 764
Netherlands	223 399 732	231 864 728	240 924 015
Norway	83 281 900	83 854 018	89 174 572
Poland	261 414 880	283 893 181	304 811 969
Portugal	56 062 622	55 238 291	61 805 741
Romania	60 536 665	62 292 509	63 778 070
Slovakia	33 682 111	36 925 292	38 860 418
Slovenia	17 983 668	18 433 682	18 039 776
Spain	524 915 698[b]	550 963 282	564 892 253
Sweden	135 846 149	143 073 805	216 902 819[b]
Switzerland	100 574 472	103 006 552	116 368 273
United Kingdom	:	:	:

: – Not available.

[b] – Break in time series.

Źródło: https://ec.europa.eu/eurostat/tgm/table.do?tab=table&init=1&language=en&p-code=tin00193&plugin=1 (dostęp: 13.06.2019).

ZADANIE 1.5.18

Skorzystaj z badań R. Szkupa (2003) – załącznik 3.

ROZDZIAŁ 2

PREZENTACJA DANYCH STATYSTYCZNYCH

Zebrany podczas badań materiał statystyczny jest zazwyczaj dostępny w formie nieuporządkowanej – w postaci zeszytów, kart inwetaryzacyjnych, fotografii, ankiet, nagrań lub komputerowych baz danych (arkusze kalkulacyjne). Materiał ten należy uporządkować i zaprezentować w formie czytelnej i oczywiście poprawnej. Można do tego użyć odpowiedniego, bezpłatnego lub komercyjnego, oprogramowania komputerowego. Są to np.:

a) arkusze kalkulacyjne,

b) programy statystyczne,

c) programy graficzne,

d) programy do analiz przestrzennych (GIS).

Warto wybrać takie oprogramowanie, które pozwala na bezproblemowe przenoszenie raz wpisanych danych. Dobrze jest poradzić się kompetentnej osoby, aby nie było problemów z konwersją danych, a w konsekwencji – konieczności ich przepisywania. Wiele bezpłatnych programów ma duże możliwości przetwarzania informacji i można je bez wątpienia polecić do tego celu. Programy komercyjne oferują wiele możliwości, ale są drogie. Należy jednak sprawdzić, czy uczniowie i studenci nie mogą

skorzystać z ich wersji bezpłatnych. Wielu producentów oprogramowania dysponuje wersjami dla studentów (z możliwością korzystania z nich od 60 dni do roku). Ostateczny wybór oprogramowania należy zawsze do osoby przeprowadzającej badania.

Zebrany materiał statystyczny jest zazwyczaj obszerny, więc należy go uporządkować, pogrupować i zaprezentować w formie syntetycznej, np. w postaci:

- szeregów statystycznych,
- tablic statystycznych,
- wykresów statystycznych,
- map,
- opisowej włączonej do tekstu.

Przed rozpoczęciem prezentacji i analizy danych statystycznych należy określić, w jakiej skali pomiarowej są one przedstawione: nominalnej, porządkowej, interwałowej czy ilorazowej (por. rozdział 1).

2.1. SZEREGI STATYSTYCZNE

Szeregiem statystycznym nazywamy ciąg wielkości statystycznych uporządkowanych według określonej cechy. Klasyfikacja musi być przeprowadzona w sposób rozłączny oraz zupełny. Oznacza to, że poszczególne jednostki o określonych cechach są jednoznacznie przyporządkowane odpowiedniej klasie, a klasy są tak skonstruowane, że obejmują wszystkie cechy występujące w danej zbiorowości.

Rozróżniamy szeregi szczegółowe i rozdzielcze, które z kolei dzielimy na:

- punktowe,
- przedziałowe,
- strukturalne z cechą jakościową,
- geograficzne,
- dynamiczne,
- kumulacyjne.

Szereg szczegółowy obejmuje wszystkie pojedyncze wartości zmiennej, uporządkowane rosnąco lub malejąco. Na przykład dla szeregu składającego się z $N = 11$ elementów będzie miał postać:

$$x_1 \leq x_2 \leq x_3 \leq x_4 \leq x_5 \leq x_6 \leq x_7 \leq x_8 \leq x_9 \leq x_{10} \leq x_{11}$$

lub

$$x_1 \geq x_2 \geq x_3 \geq x_4 \geq x_5 \geq x_6 \geq x_7 \geq x_8 \geq x_9 \geq x_{10} \geq x_{11}$$

Z tego wynika, że szeregu szczegółowego nie można skonstruować dla danych w skali nominalnej, a jedynie dla porządkowej, interwałowej lub ilorazowej.

Przykład 2.1.1

Zanotowano wiek osób kolejno wchodzących w godz. od 16^{00} do 16^{10} na koncert Erica Claptona w czerwcu 2013 r. w Atlas Arenie w Łodzi[1]: 23, 45, 56, 44, 34, 12, 45, 55, 61, 41, 32, 71, 66, 48, 28, 51, 37, 24, 61, 65, 49, 20, 20, 44, 38. Należy przedstawić tę informację w postaci szeregu szczegółowego.

W tym celu należy uporządkować zmienną rosnąco lub malejąco. Można skorzystać z funkcji „sortuj", jeśli dane te mamy wprowadzone do arkusza kalkulacyjnego lub programu statystycznego, albo zrobić to ręcznie, co jest bardziej pracochłonne przy dużej liczbie danych. W wyniku sortowania uzyskujemy szereg uporządkowanych wartości tej zmiennej, np. rosnąco: 12, 20, 20, 23, 24, 28, 32, 34, 37, 38, 41, 44, 44, 45, 45, 48, 49, 51, 55, 56, 61, 61, 65, 66, 71.

Szereg rozdzielczy otrzymujemy w przypadku, gdy rozdzielamy zbiorowość na klasy według określonej cechy (x_i) i podajemy liczebność w każdej z tych klas (f_i). Stosowany jest on zazwyczaj dla danych w skali porządkowej lub wyższej.

Uzyskane liczby mogą przyjąć postać szeregu **rozdzielczego punktowego** (tab. 2.1.1). Jest on stosowany do cech w skali porządkowej lub wyższej do cech mierzalnych skokowych, np. dla zmiennej „liczba dzieci w rodzinie", „ocena zadowolenia z usługi".

1 Na koncercie było ok. 15 tys. osób.

Tabela 2.1.1. Struktura cechy x_i

Wielkość cechy x_i	**Liczebność** f_i
x_1	f_1
x_2	f_2
...	...
x_n	f_n
Ogółem	f

Źródło: opracowanie własne.

Przykład 2.1.2

Na lotnisku Okęcie w Warszawie 15 lipca 2013 r. podróżni czekali na wylot do Londynu na wakacje. Wielu z nich podróżowało z dziećmi. Postanowiono przeprowadzić badanie wśród rodzin pod kątem liczby dzieci, które towarzyszyły im w podróży. Rodziny zabierały w podróż następującą liczbę potomstwa: 2, 3, 1, 4, 1, 1, 1, 4, 0, 0, 3, 3, 2, 2, 2, 2, 1, 1, 1, 1. Przedstaw tę informację w postaci szeregu szczegółowego oraz rozdzielczego punktowego.

Wybrany został szereg punktowy, gdyż liczba dzieci jest wartością podaną w skali ilorazowej. Dane mają jedynie 5 klas i są liczbami całkowitymi nieujemnymi (czyli reprezentują cechy mierzalne skokowe).

Aby przedstawić dane w postaci szeregu szczegółowego, należy uporządkować je rosnąco lub malejąco. W prezentowanym przykładzie zmienna x_i przyjmuje wielkości: x_1, x_2, ..., x_{20}. Szereg szczegółowy będzie miał następującą postać:

x_1	x_2	x_3	x_4	x_5	x_6	x_7	x_8	x_9	x_{10}	x_{11}	x_{12}	x_{13}	x_{14}	x_{15}	x_{16}	x_{17}	x_{18}	x_{19}	x_{20}
0	0	1	1	1	1	1	1	1	1	2	2	2	2	2	3	3	3	4	4

Tabela 2.1.2. Rodziny na lotnisku Okęcie w Warszawie wybierające się do Londynu 15 lipca 2013 r. z uwzględnieniem liczby towarzyszących im dzieci

Liczba dzieci x_i	**Liczba rodzin** f_i
0	2
1	8
2	5

Liczba dzieci x_i	Liczba rodzin f_i
3	3
4	2
Ogółem	20

Źródło: opracowanie własne (dane umowne).

Do prezentacji jednej cechy zbiorowości wykorzystuje się też **szereg rozdzielczy przedziałowy** (tab. 2.1.3). Stosuje się go wówczas, gdy dane są w skali interwałowej lub ilorazowej i do cech mierzalnych ciągłych, takich jak np. wydatki na turystykę, wiek podróżnych, waga bagażu. Dobór klas, ich wartości minimalnej oraz ich rozpiętości należy do decyzji badacza (przykład 2.1.3).

Tabela 2.1.3. Struktura cechy x_i

Rozpiętość przedziału klasowego $<x_{id} - x_{ig})$	Liczebność f_i
$<x_{1d} - x_{1g})$	f_1
$<x_{2d} - x_{2g})$	f_2
...	...
$<x_{nd} - x_{ng})$	f_n
Razem	f

Objaśnienia: x_{id} – dolna granica i-tego przedziału, x_{ig} – górna granica i-tego przedziału, n – liczba przedziałów.

Źródło: opracowanie własne.

Przy konstrukcji szeregu statystycznego rozdzielczego warto sprawdzić wyniki innych podobnych badań statystycznych. Dzięki temu możliwe jest zapoznanie się z rodzajami klasyfikacji lub porównanie wyników.

Przykład 2.1.3

Badaniu poddano wydatki (zł) poniesione podczas pobytu w sanatorium w Sopocie w styczniu 2016 r. przez kuracjuszy. Informację przedstawiono w postaci szeregu szczegółowego: 200, 250, 270, 340, 410,

490, 500, 700, 710, 800, 850, 900, 1000, 1200, 1450. Przedstawmy wydatki w postaci szeregu rozdzielczego przedziałowego. Wybrano ten rodzaj szeregu, gdyż wydatki są cechą mierzalną ciągłą podaną w skali ilorazowej.

Zanim przystąpimy do tworzenia szeregu, należy zastanowić się, jaka rozpiętość przedziału będzie najodpowiedniejsza. Należy zwrócić uwagę na to, aby nie było „pustych" przedziałów oraz aby ich liczba była optymalna. W przykładzie wybrano rozpiętość 250 zł. Dolną granicę pierwszego przedziału wybieramy tak, aby można go było porównywać z podobnymi wynikami badań. Dlatego rozpoczęto od wartości 0, pomimo że najmniejszą wartością, jaka występowała w szeregu, było 200 zł. Rozpiętość przedziałów w szeregu można zaprezentować w dwojaki sposób (tab. 2.1.4).

Tabela 2.1.4. Wydatki poniesione przez kuracjuszy podczas pobytu w sanatorium (w zł) w Sopocie w styczniu 2017 r. Dwa sposoby przedstawiania szeregu

Wielkość wydatków (zł) $\langle x_{id} ; x_{ig})$	**Liczba osób f_i**
0–250	1
250–500	5
500–750	3
750–1000	3
1000–1250	2
1250–1500	1
Ogółem	15

Wielkość wydatków (zł) $\langle x_{id} ; x_{ig} \rangle$	**Liczba osób f_i**
0–249,99	1
250–499,99	5
500–749,99	3
750–999,99	3
1000–1249,99	2
1250–1499,99	1
Ogółem	15

Źródło: opracowanie własne.

W niektórych opracowaniach statystycznych można wykorzystać szeregi rozdzielcze o **otwartych przedziałach klasowych**. Są to szeregi, w których minimum jeden z przedziałów (pierwszy lub ostatni) nie ma wskazanych wartości granicznych i zastąpiony jest wyrażeniem „poniżej", „powyżej" „mniej", „i więcej" (przykład 2.1.4).

Przykład 2.1.4

Zebrano informacje o wydatkach w czasie podróży urlopowych wśród pracowników jednego z banków w 2014 r. Wynosiły one w zł: 300, 450, 580, 600, 700, 880, 1200, 1300, 1900, 100 000. Jeden z pracowników kupił na pamiątkę obraz w antykwariacie za znaczącą kwotę, odbiegającą od wydatków poniesionych przez innych pracowników. Aby przedstawić te dane w postaci szeregu rozdzielczego, można skorzystać z możliwości konstrukcji przedziału otwartego. W tekście należy skomentować tę wartość – jest ona nazywana **nietypową wartością skrajną**.

Tabela 2.1.5. Wydatki poniesione przez pracowników jednego z banków w 2014 r. podczas podróży urlopowych

Wydatki (zł) $<x_{id}; x_{ig})$	**Liczba osób** f_i
0–500	2
500–1000	4
1000–1500	2
1500–2000	1
2000 i więcej	1
Ogółem	10

Źródło: opracowanie własne.

Wśród szeregów rozdzielczych wyróżniamy **szeregi proste** i **skumulowane**. W szeregu prostym podane są liczebności poszczególnych klas f_i (tabele powyżej). Budowę szeregu kumulacyjnego rozpoczynamy, dodając liczebności kolejnych przedziałów. W pierwszym wierszu przepisuje się liczebność pierwszego przedziału, a sumę dwóch kolejnych wpisuje się w drugim wierszu f_{ic}. Następnie dodajemy liczebność trzech pierwszych przedziałów i wpisuje się ją na trzeciej pozycji itd. Można tworzyć szereg kumulacyjny zarówno z wartości bezwzględnych, jak i procentowych (tab. 2.1.6).

Szereg skumulowany daje możliwość dodatkowej analizy zjawiska, np. można z niego odczytać, że 40% kuracjuszy wydało podczas pobytu w uzdrowisku mniej niż 500 zł, a 80% mniej niż 1000 zł.

Przykład 2.1.5

Oblicz udział procentowy wydatków kuracjuszy z tab. 2.1.4 oraz przedstaw go w postaci szeregu skumulowanego.

Tabela 2.1.6. Wydatki (w zł) poniesione podczas pobytu w sanatorium w Sopocie w styczniu 2017 r. przez kuracjuszy

Wydatki (w zł)	Liczba osób	Udział (%)	Wartości skumulowane	
	f_i		f_{ic}	% f_{ic}
0–250	1	6,7	1	6,7
250–500	5	33,3	6	40,0
500–750	3	20,0	9	60,0
750–1000	3	20,0	12	80,0
1000–1250	2	13,3	14	93,3
1250–1500	1	6,7	15	100,0
Razem	15	100,0	×	×

Źródło: opracowanie własne na podstawie tab. 2.1.4.

Szeregi strukturalne dotyczące cech jakościowych występują w skali nominalnej lub porządkowej. Mogą nimi być: płeć, rodzaj, cel i kierunek wyjazdu.

Przykład 2.1.6

W wyniku badań wykonanych w Instytucie Turystyki na zlecenie Ministerstwa Sportu i Turystyki (w ramach Programu badań statystycznych statystyki publicznej na rok 2008 – temat nr 1.30.06 (087): *Aktywność turystyczna Polaków*) uzyskano następującą strukturę wyjazdów Polaków według kierunku i długości wyjazdu turystycznego (tab. 2.1.7) oraz celu wyjazdu (tab. 2.1.8).

Tabela 2.1.7. Krajowe i zagraniczne podróże Polaków w wieku 15 i więcej lat (w mln) w latach 2006–2008

Rodzaj wyjazdu	2006	2007	2008
Krajowy długookresowy	16,90	15,80	14,30
Krajowy krótkookresowy	21,60	19,10	20,60
Krajowy (ogółem)	38,50	34,90	34,90
Zagraniczny długookresowy	b.d.	4,75	5,35
Zagraniczny krótkookresowy	b.d.	1,45	1,65

Rodzaj wyjazdu	2006	2007	2008
Zagraniczny (ogółem)	b.d.	6,20	7,00
Krajowe i zagraniczne (ogółem)	b.d.	41,10	41,90

b.d. – brak danych.

Uwaga: uczestnictwo w wyjazdach (podróżach) turystycznych odnosi się do osób, które co najmniej raz wzięły udział w danego rodzaju wyjeździe (podróży); część osób uczestniczyła w więcej niż jednym rodzaju podróży.

Źródło: badania ankietowe Instytutu Turystyki na ogólnopolskiej reprezentatywnej próbie OBOP obejmującej mieszkańców Polski w wieku 15 i więcej lat (*N* = 4059).

Tabela 2.1.8. Cele zagranicznych podróży Polaków w latach 2007–2008

Cel	Liczba osób	
	2007	2008
Turystyczno-wypoczynkowy	42	52
Odwiedziny u krewnych lub znajomych	25	25
Służbowy	26	18
Inne długookresowe	7	5

Źródło: badania Instytutu Turystyki.

Jeśli szereg przedstawia dane w skali nominalnej, to cechy porządkujemy według alfabetu albo według wielkości zjawiska (tab. 2.1.8). W przypadku gdy konstruujemy szereg dla cechy w skali porządkowej (np. wykształcenie), nie powinno się porządkować danych w tabeli według wielkości zjawiska, lecz zaprezentować je według określonego porządku cechy (tab. 2.1.9).

Tabela 2.1.9. Uczestnictwo (w %) Polaków w wyjazdach krótkookresowych w latach 2006–2008 według wykształcenia

Wykształcenie	Udział osób (%)		
	2006	2007	2008
Podstawowe	18,2	17,6	15,9
Zasadnicze zawodowe	16,6	16,1	16,0
Średnie	27,3	26,5	23,4
Wyższe	39,7	38,1	29,9

Źródło: badania Instytutu Turystyki.

Szczególnym rodzajem szeregu ukazującego strukturę według występowania zjawiska w przestrzeni (np. kierunków wyjazdów) jest **szereg geograficzny**, który ukazuje rozmieszczenie zjawiska w przestrzeni (np. według kontynentów, państw, jednostek administracyjnych różnego szczebla). W szeregu geograficznym wiersze mogą być uporządkowane według alfabetu (gdy wymieniamy wszystkie jednostki administracyjne) lub według wielkości zjawiska (tab. 2.1.10).

Tabela 2.1.10. Przyjazdy turystów do Japonii w latach 2007–2012

Kraj	Liczba turystów					
	2007	2008	2009	2010	2011	2012
Ogółem przyjazdy	8 346 969	8 350 835	6 789 658	8 611 175	6 218 752	8 358 105
Europa razem	877 531	886 723	800 085	853 166	569 279	775 840
Austria	13 217	13 453	13 684	14 440	8 539	11 633
Belgia	14 828	15 773	13 899	15 981	10 708	14 608
Dania	14 305	14 486	13 116	14 606	10 821	13 594
Finlandia	18 870	20 025	17 797	16 960	10 943	15 529
Francja	137 787	147 580	141 251	151 011	95 438	130 412
Hiszpania	33 478	40 852	42 484	44 076	20 814	35 207
Holandia	33 290	34 487	31 186	32 837	23 450	30 266
Irlandia	13 681	12 513	10 450	10 738	8 294	10 358
Niemcy	125 193	126 207	110 692	124 360	80 772	108 898
Norwegia	10 668	10 848	9 855	10 302	7 905	11 447
Portugalia	13 351	10 280	8 463	10 313	6 227	8 408
Rosja	64 244	66 270	46 952	51 457	33 793	50 176
Szwajcaria	23 996	24 364	23 091	26 005	16 410	24 329
Szwecja	29 792	30 129	26 384	29 188	21 806	30 458
Wlk. Brytania	221 945	206 564	181 460	184 045	140 099	173 994
Włochy	54 022	56 243	59 607	62 394	34 035	51 801
Pozostałe europejskie	54 864	56 649	49 714	54 453	39 225	54 722

Źródło: http://www.stat.go.jp (dostęp: 26.09.2015).

Szeregi dynamiczne przedstawiają rozmiary zjawiska w określonym czasie, wykorzystuje się je do danych w skali interwałowej lub ilorazowej. Szeregi dynamiczne dzielimy na szeregi okresów i szeregi momentów. **Szereg okresów** określa zmiany zjawiska w przeciągu pewnego okresu, np. miesiąca, kwartału, roku (tab. 2.1.11).

Tabela 2.1.11. Ruch pasażerów w morskich portach handlowych w Polsce w latach: 1960, 1965, 1970, 1975

Wyszczególnienie	Liczba pasażerów			
	1960	1965	1970	1975
Przyjazdy pasażerów do kraju[a] przez port:	6 005	16 859	50 745	117 296
Gdańsk	663	1 023	1 212	31 147
Gdynia	4 994	6 100	7 889	7 670
Szczecin	348	9 736	41 644	78 479
Wyjazdy pasażerów z kraju[b] przez port:	7 904	19 902	47 376	111 050
Gdańsk	221	718	1 032	30 164
Gdynia	7 157	10 028	6 206	5 694
Szczecin	526	9 156	41 138	75 192

[a] Pasażerowie, którzy przyjechali z portów zagranicznych do portów polskich jako portów docelowych w ich podróży morskiej na statkach pasażerskich, towarowych i promach.

[b] Pasażerowie, którzy wyjechali w podróż morską z portów polskich do portów zagranicznych na statkach pasażerskich, towarowych i promach.

Źródło: GUS (1978), s. 277.

Szereg momentów określa rozmiar zjawiska w ściśle określonym momencie, np. ostatniego dnia w roku (tab. 2.1.12).

Tabela 2.1.12. Ludność Polski na podstawie Narodowych Spisów Powszechnych z lat: 1946, 1950, 1960 i 1970

Data spisu	Ogółem	Mężczyźni	Kobiety	Miasto	Wieś
	w tys.				
14 lutego 1946 r.	23 930[a]	10 954	12 976	7 517	16 109
3 grudnia 1950 r.	25 008[b]	11 928	13 080	9 605	15 009
6 grudnia 1960 r.	29 776[c]	14 404	15 372	14 219	15 187
8 grudnia 1970 r.	32 642[d]	15 854	16 788	17 064	15 578

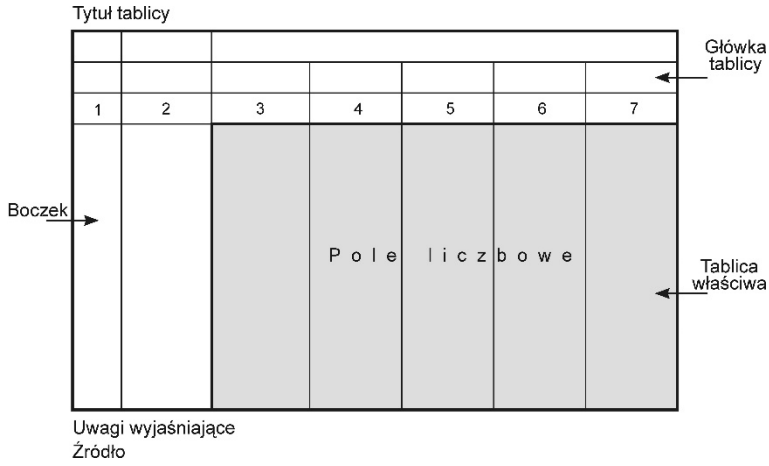
Objaśnienia: w danych dotyczących miast i wsi nie uwzględniono: [a] 304 tys., [b] 394 tys., [c] 370 tys. w podziale administracyjnym z 1 stycznia 1971 r.

Źródło: GUS (1978), s. 25.

2.2. TABLICE STATYSTYCZNE

Jak łatwo zauważyć, szeregi rozdzielcze zaprezentowane w podrozdziale 2.1 miały postać tabeli nazywanej **tablicą statystyczną**. Zgrupowane i opracowane materiały statystyczne przedstawiane są zazwyczaj w formie tablic statystycznych, których budowa jest ściśle określona. Poprawnie zbudowana tablica (tabela) statystyczna składa się z następujących elementów:

- tytułu,
- tablicy właściwej,
- uwag wyjaśniających lub objaśnień (opcjonalnie),
- informacji o źródle danych.



Jeśli tekst prezentujący wyniki badań zawiera kilka tablic statystycznych, to każdą z nich numeruje się w kolejności występowania. W tekście powinny znaleźć się powołania na nie, tak aby czytelnik mógł porównać analizę statystyczną autora z danymi zawartymi w tabeli, którą interpretuje.

Tytuł tablicy – powinien być sformułowany krótko i precyzyjnie, określać badaną zbiorowość, badaną cechę oraz czas i miejsce badania. Niekiedy w tytule występuje jednostka miary wspólna dla wszystkich jednostek statystycznych, np. zł (tab. 2.1.4) lub liczebność próby badawczej *N* (tab. 2.3.1).

Tablica właściwa – składa się z główki (zawierającej tytuły kolumn), boczku (zapisuje się w nim warianty cech przyporządkowane jednostkom) i pola liczbowego. Jeżeli w tablicy występują różne jednostki miary, to są one umieszczane w pierwszej kolumnie po boczku. W pierwszym wierszu główki oraz w pierwszej kolumnie boczku wszystkie tytuły kolumn oraz warianty cech w kolumnie pierwszej zapisujemy, zawsze zaczynając nazwy wielką literą (nawet jeśli jest to nazwa województwa).

W przypadku rozbudowanej tablicy numeruje się również kolumny, tak aby kontynuując tabelę na następnych stronach, wpisać jedynie numery kolumn (bez ich opisów).

Uwagi wyjaśniające – zamieszcza się je bezpośrednio pod tablicą właściwą. W tablicy umieszcza się odnośniki do poszczególnych uwag (małymi literami lub symbolami z indeksem górnym: ^{a}, *, Uwaga). Uwagi dotyczą kwestii spornych, sposobu grupowania danych; przykładowe uwagi znajdują się pod tab. 2.1.7, 2.1.11 i 2.1.12.

Źródło (danych) – zawiera informacje na temat pochodzenia danych zamieszczonych w tablicy. Podanie źródła jest szczególnie ważne w przypadku wykorzystania danych z innych publikacji ze względu na ochronę praw autorskich. Poznanie źródła danych pozwala często ocenić wiarygodność informacji. Sposób zapisu zależy od tego, czy informacje pochodzą z badań własnych czy publikowanych. Źródło powinno być tak podane, aby można było do niego dotrzeć (m.in. tab. 2.1.10), np. wpisując adres internetowy źródła danych. Przykładowe źródła danych:

- opracowanie własne (gdy wszystkie dane pochodzą z pomiarów przeprowadzonych przez autora),
- opracowanie własne na podstawie danych GUS (dane pochodzą z innych publikacji, np. GUS, ale były przez autora opracowane w innej formie),
- *Rocznik statystyczny 1977*, GUS, Warszawa, s. 277 (dane są identyczne jak w cytowanej publikacji),
- *Rocznik statystyczny 2003*, GUS, Warszawa, s. 58, zmodyfikowane (dane pochodzą z cytowanej publikacji, ale zostały przez autora zgeneralizowane, np. zamiast danych o ruchu turystycznym dla poszczególnych lat 2000, 2001, 2002 przedstawiono ich sumę za lata 2000–2002),

- strony internetowe, np. http://www.stat.go.jp, https://ec.europa.eu/eurostat/web/tourism/data/database, z datą dostępu,
- badania Instytutu Turystyki (dane uzyskane z Instytutu Turystyki),
- dane udostępnione przez Muzeum Narodowe w Warszawie.

Treść tablicy powinna być zgodna z tytułem, a umieszczone informacje czytelne i zrozumiałe oraz porównywalne z innymi opracowaniami. Wszystkie komórki w polu liczbowym muszą być wypełnione liczbami lub znakami umownymi (np. brak danych – b.d.). Pole liczbowe powinno być wypełnione czytelnie, z zachowaniem odpowiednich odstępów.

Niektóre stosowane znaki umowne w tablicach statystycznych:

(–) kreska – zjawisko nie występuje,

(·) kropka lub b.d. – brak danych lub brak wiarygodnych informacji,

(0) zero – zjawisko występuje w zbyt małych ilościach, w rozmiarze mniejszym niż jednostki miary przyjęte w tablicy, np. jeśli jednostką miary byłyby miliony turystów, a wartość, jaką należałoby wpisać, wynosiłaby 100, to oznaczona byłaby w tablicy jako 0,

(0,0) zero, zero – zjawisko występuje w wielkości mniejszej niż 0,05,

(×) – dana pozycja nie może być wypełniona ze względów formalnych, np. w tablicy przedstawiającej czas przejazdu pociągów między Warszawą a Warszawą, Lublinem i Lublinem,

(♦) – występuje w „Przeglądzie Międzynarodowym"; oznacza, że dane dla Polski w części międzynarodowej różnią się od danych w części krajowej,

w tym – oznacza, że nie podaje się wszystkich składników sumy (por. tab. 2.1.10).

Wszystkie tabele z podrozdziału 2.1 spełniają warunki, jakie stawia się tablicom statystycznym. Jeśli wykorzystujemy publikacje, w których są tablice statystyczne ze znakami umownymi, warto sprawdzić, czy autorzy stosują te same znaki umowne. Są one zazwyczaj opisane na początku publikacji. Jeśli korzystamy ze źródeł zagranicznych, należy sprawdzić, jakie znaki obowiązują w danym kraju, np. w zbiorach danych Eurostat[2] dwukropek (:) oznacza brak danych (*not available*) (por. tab. 1.5.4).

---

2 http://appsso.eurostat.ec.europa.eu/nui/show.do.

2.3. GRAFICZNA PREZENTACJA DANYCH STATYSTYCZNYCH

Tablica statystyczna jest dobrym narzędziem do studiowania badanego zjawiska, lecz często duża liczba informacji w niej zawartych nie pozwala na zaprezentowanie istoty badanego problemu. Wówczas lepszym narzędziem okazuje się wykres lub mapa, które nie zastępują tabeli, lecz są środkami pomocniczymi.

Wykres statystyczny składa się z **pola wykresu** i **części opisowej** (nad lub pod wykresem). **Tytuł** wykresu umieszczany jest zazwyczaj – w przeciwieństwie do tabeli – pod rysunkiem. Powinien być sformułowany krótko oraz informować o przedmiocie, czasie i miejscu przedstawianego zagadnienia. **Podtytuł** zawiera informacje uzupełniające. **Legenda** umieszczana jest w polu wykresu lub pod nim. **Źródło** podajemy zawsze pod wykresem. Gdy korzystamy z tabeli występującej w tym samym tekście, wystarczy podać numer tabeli. Jeśli w opracowaniu jest kilka wykresów, to numeruje się je w kolejności, w jakiej są prezentowane.

Pierwszy etap wizualizacji danych za pomocą wykresu stanowi poszukiwanie jak najlepszej metody z uwzględnieniem skali, w jakiej są prezentowane dane. W tym celu należy zadać pytania:

- w jakiej skali pomiarowej są dane?
- wartości ilu zmiennych chcemy pokazać na wykresie?
- jaki typ szeregu rozdzielczego wykorzystano do prezentacji danych w tabeli statystycznej?

W celu pokazania struktury jednego zjawiska w skali nominalnej warto sięgnąć po wykresy kołowe lub słupkowe (rys. 2.3.1, 2.3.2) i umieścić w opisie informację o wielkości zjawiska w procentach. Jeśli do analizy dołączona jest tabela, to powinna zawierać wartości bezwzględne, a nie procentowe, aby nie powielać informacji (przykład 2.3.1).

Przykład 2.3.1

W wyniku przeprowadzenia badania ankietowego wśród 100 mieszkańców Pabianic na temat form spędzania czasu wolnego w parku Wolności uzyskano odpowiedzi, które zawarto w tab. 2.3.1. Są to informacje w skali nominalnej uporządkowane w szeregu strukturalnym. Przedstawmy je graficznie.

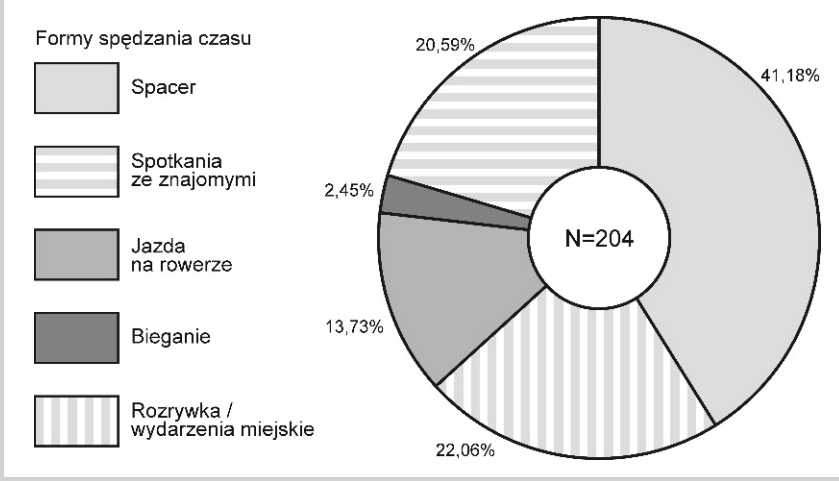
Tabela 2.3.1. Formy spędzania czasu wolnego przez mieszkańców Pabianic w parku Wolności w 2011 r. (*N* = 100)

Lp.	Formy spędzania czasu	Liczba odpowiedzi*
1	Spacer	84
2	Spotkania ze znajomymi	45
3	Jazda na rowerze	28
4	Bieganie	5
5	Rozrywka/wydarzenia miejskie	42
Razem		204

* Możliwość wskazania kilku form przez jedną osobę.

Źródło: opracowanie J. Szkobel na podstawie badań terenowych.

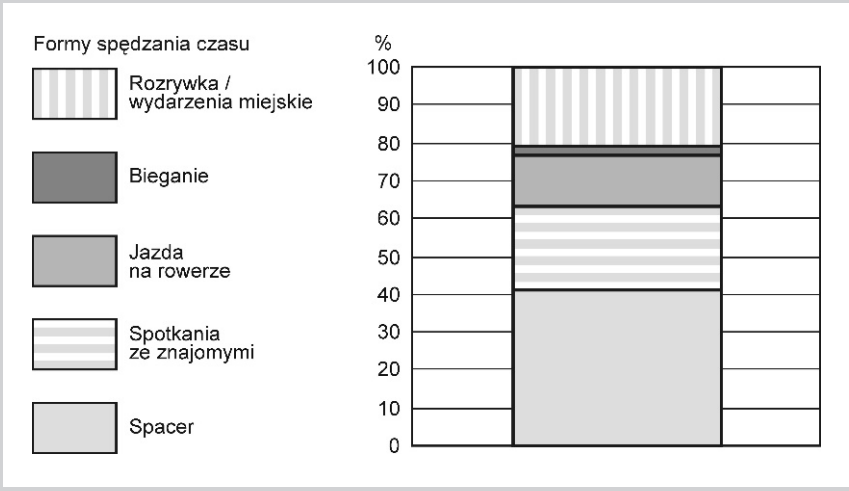
Przedstawione w tab. 2.3.1 dane można zaprezentować w postaci diagramu strukturalnego kołowego.



Rysunek 2.3.1. Formy spędzania czasu wolnego przez mieszkańców Pabianic w parku Wolności w 2011 r. (*N* = 204)

Źródło: opracowanie własne na podstawie tab. 2.3.1.

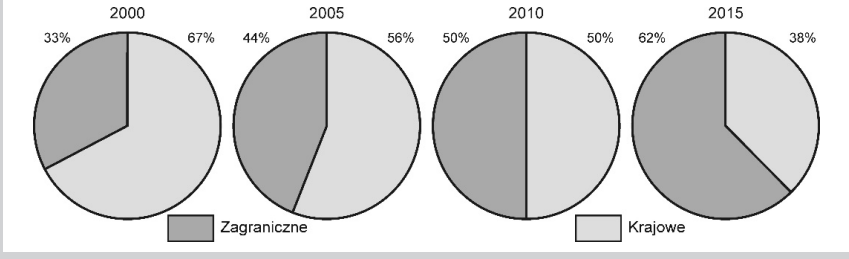
Dane z tab. 2.3.1. można zaprezentować również jako diagram strukturalny słupkowy (rys. 2.3.2).



Rysunek 2.3.2. Formy spędzania czasu wolnego przez mieszkańców Pabianic w parku Wolności w 2011 r.

Źródło: opracowanie własne na podstawie tab. 2.3.1.

Przyjęto, że za pomocą tego typu wykresów nie ilustruje się zjawisk, w których występują tylko dwie możliwości (np. podział na płeć, wyjazdy krajowe i zagraniczne itd.). Można odstąpić od tej zasady, gdy na jednym rysunku umieszcza się kilka wykresów w celach porównawczych (rys. 2.3.3).



Rysunek 2.3.3. Kierunki wyjazdów urlopowych pracowników Lasów Państwowych w latach 2000–2015

Źródło: opracowanie własne na podstawie danych umownych.

Aby jak najlepiej zilustrować graficznie zjawisko prezentowane w skali interwałowej lub ilorazowej, należy zastanowić się nad wyborem odpowiedniego wykresu oraz właściwej skali. Najczęściej używany jest prostokątny układ współrzędnych. Jeśli prezentowane dane mają wartości dodanie, to wykorzystujemy pierwszą ćwiartkę tego układu.

W statystyce rozróżnia się następujące skale stosowane przy konstrukcji wykresów (Zając 1988):

a) prostoliniowe,

b) krzywoliniowe (rzadziej używane),

c) punktowe,

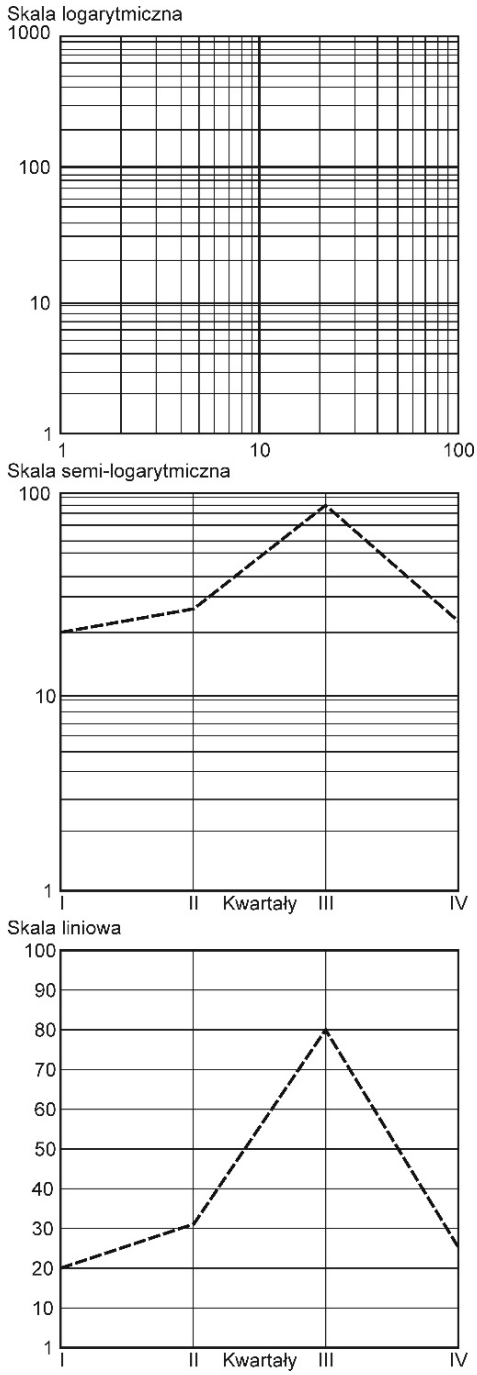
d) punktowo-obrazkowe.

Skala wykresu to krzywa (w większości przypadków prosta), której punktom przyporządkowuje się wartości pewnej zmiennej. Wyróżnione punkty na skali (najczęściej w postaci prostopadłych do skali kresek) tworzą jej podziałkę i umożliwiają wyznaczenie wartości zmiennej przyporządkowanej dowolnemu punktowi skali. Niektóre kreski opatrzone są wartościami zmiennej. Przedział między dwiema sąsiednimi kreskami nazywa się **działką elementarną**. Podziałka kreskowa nazywana jest **jednostajną**, jeżeli wszystkie jej działki elementarne mają jednakową długość, a **równomierną** – jeżeli mają jednakową wartość.

Podziałki jednostajna i równomierna są podziałkami **regularnymi**, które stanowią przykłady podziałki liniowej. Do podziałek **nieliniowych** zalicza się natomiast podziałkę logarytmiczną, kwadratową i inne.

Podziałki skali wykresu należy dobrać w ten sposób, aby odczytanie dowolnego punktu nie sprawiło trudności oraz aby dotyczyło wyłącznie wartości zmiennej. Wyznaczając podziałkę skali, należy pamiętać o proporcjach wykresu. Osie *X* i *Y* nie muszą przecinać się w punkcie (0,0). Osie współrzędnych muszą być dokładnie opisane, bowiem bez tego wykres jest bezwartościowy.

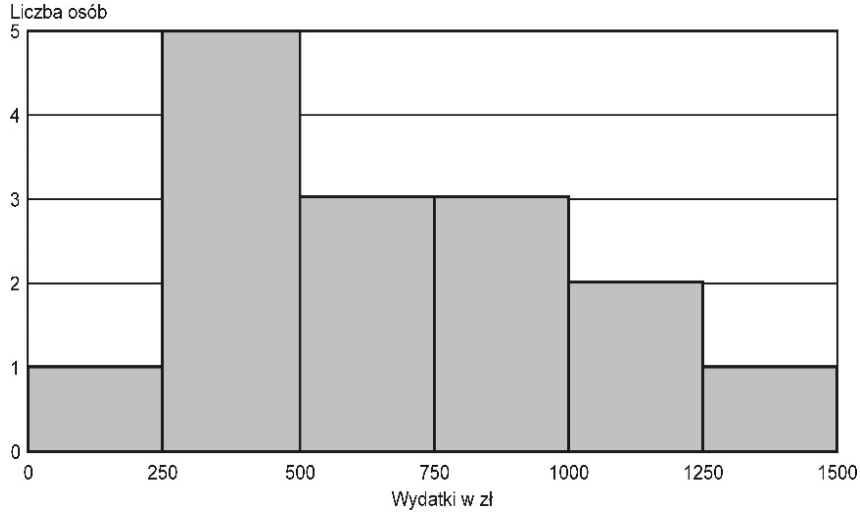
W celu porównania względnych różnic na ogół korzysta się ze skali nierównomiernej, np. logarytmicznej. Wówczas na osi *Y* odkładamy skalę logarytmiczną, a oś *X* ma skalę równomierną. Wykres taki nosi nazwę półlogarytmicznego lub semilogarytmicznego.



Rysunek 2.3.4. Typy skal wykorzystywanych do komponowania wykresu

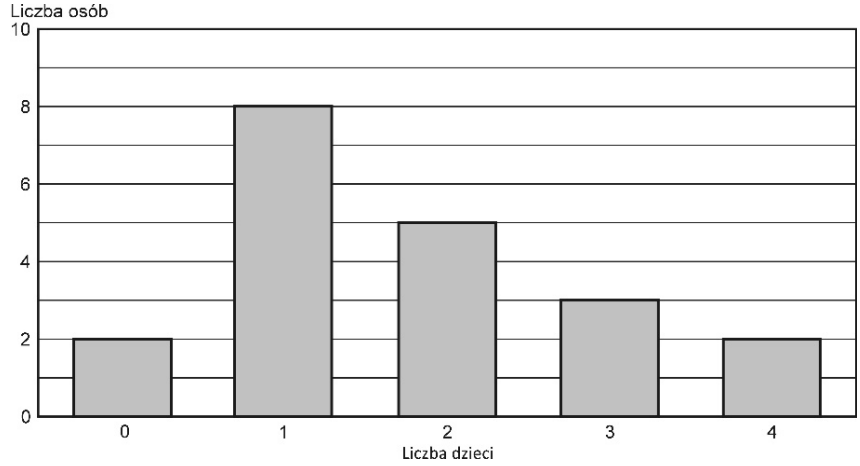
Źródło: Jażdżewska (2013).

Dane można przedstawić w formie **histogramu**, czyli wykresu w prostokątnym układzie współrzędnych, w którym na osi *X* zapisuje się cechy, a na osi *Y* liczbę wskazań lub ich udział procentowy (rys. 2.3.5–2.3.6).



Rysunek 2.3.5. Histogram rozkładu wydatków kuracjuszy podczas pobytu w sanatorium w Sopocie w styczniu 2017 r.

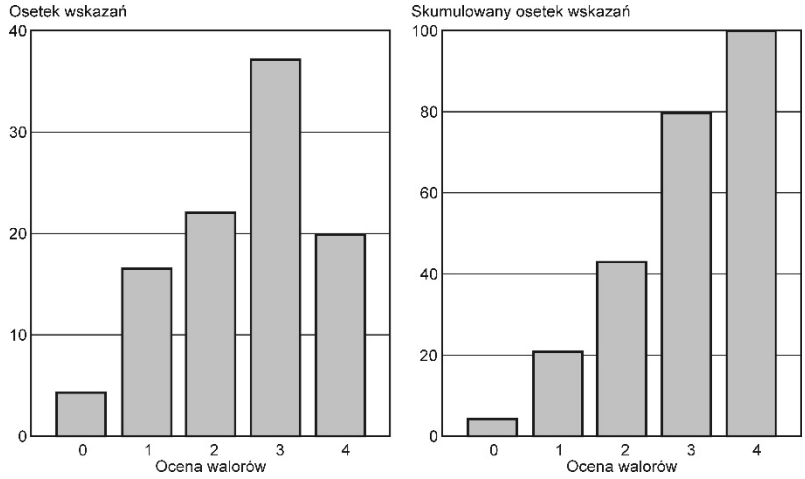
Źródło: opracowanie własne na podstawie tab. 2.1.4.



Rysunek 2.3.6. Histogram rozkładu rodzin podróżujących do Londynu z lotniska Okęcie w Warszawie 15 lipca 2013 r. według liczby towarzyszących im dzieci

Źródło: opracowanie własne na podstawie tab. 2.1.2.

Dane w skali porządkowej można zaprezentować w formie diagramu (rys. 2.3.7), na którym na osi *X* zapisuje się cechy (według ustalonego porządku), a na osi *Y* liczbę wskazań lub ich udział procentowy. W formie diagramu można przedstawić tzw. szereg kumulacyjny. Na osi *X* znajdują się cechy, zaś na osi *Y* – liczebności skumulowane.



Rysunek 2.3.7. Ocena walorów przyrodniczych ścieżek rowerowych w Sieradzu w 2011 r. (*N* = 91), szereg prosty i skumulowany

Źródło: opracowanie J. Szkobel na podstawie badań terenowych.

Przykład 2.3.2

Przedstaw w postaci graficznej zagraniczne podróże studentów kierunku turystyka i rekreacja w 2010 r. według krajów.

Tabela 2.3.2. Zagraniczne podróże studentów kierunku turystyka i rekreacja w 2010 r. według krajów

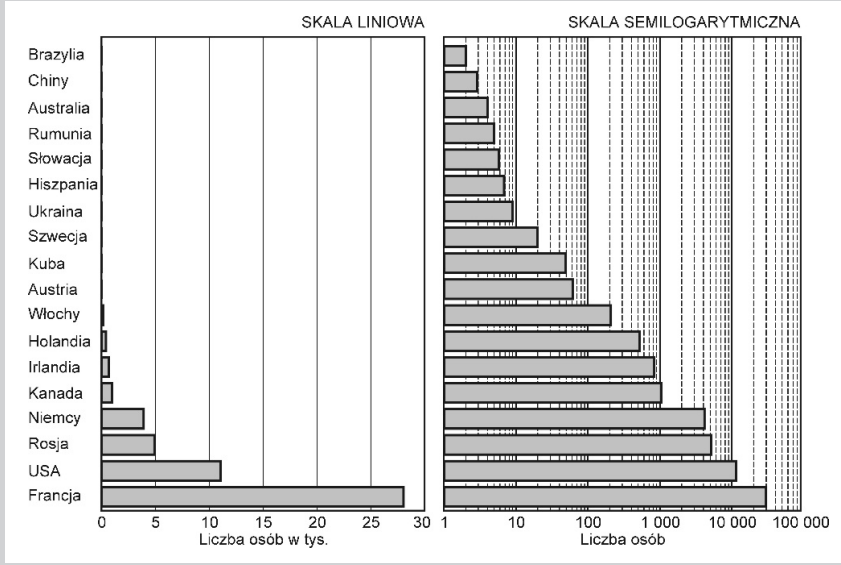
Kraj	Liczba osób	Kraj	Liczba osób
Francja	28 000	Kuba	50
USA	11 000	Szwecja	20
Rosja	5 000	Ukraina	9
Niemcy	4 000	Hiszpania	7

Tabela 2.3.2 cd.

Kraj	Liczba osób	Kraj	Liczba osób
Kanada	1 000	Słowacja	6
Irlandia	800	Rumunia	5
Holandia	500	Australia	4
Włochy	200	Chiny	3
Austria	60	Brazylia	2

Źródło: opracowanie własne (dane umowne).

Najpierw zazwyczaj tworzymy wykres w skali liniowej (rys. 2.3.8A). Bardzo małe wartości (poniżej 100 osób) są jednak zupełnie niewidoczne na wykresie, dlatego warto skorzystać z możliwości zastąpienia skali liniowej skalą logarytmiczną. W efekcie otrzymuje się wykres, na którym rozkład danych jest bardzo czytelny (rys. 2.3.8B).

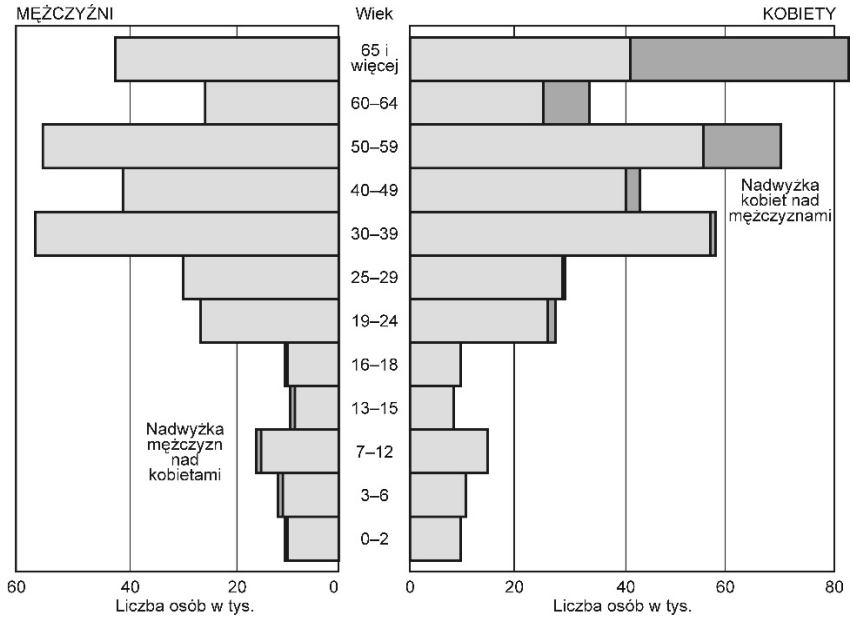


A B

Rysunek 2.3.8. Zagraniczne podróże studentów kierunku turystyka i rekreacja według krajów w 2010 r. w skali liniowej i w skali semilogarytmicznej

Źródło: opracowanie własne na podstawie tab. 2.3.2.

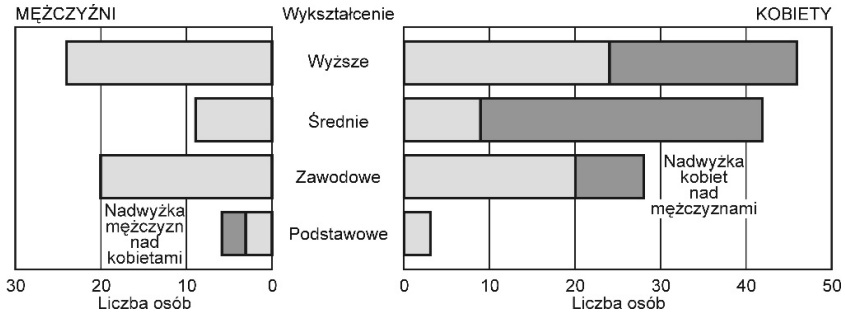
Szczególnym przypadkiem szeregu rozdzielczego jest piramida płci i wieku (rys. 2.3.9). Prezentuje ona dane w skali nominalnej (płeć) oraz ilorazowej (wiek).



Rysunek 2.3.9. Struktura płci i wieku mieszkańców Łodzi w 2010 r.

Źródło: opracowanie własne na podstawie danych z https://stat.gov.pl (dostęp: 13.06.2019).

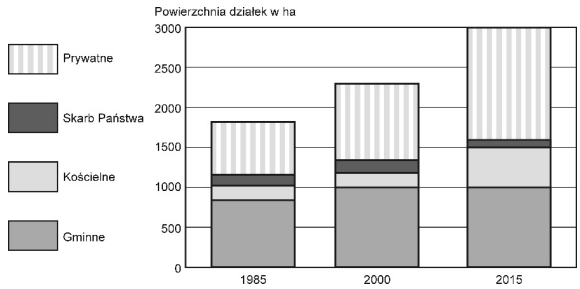
Zilustrowanie struktury płci (skala nominalna) i wykształcenia (skala porządkowa) badanej grupy jest możliwe za pomocą wykresu podobnego do piramidy płci i wieku (rys. 2.3.10).



Rysunek 2.3.10. Piramida płci i wykształcenia pielgrzymów (respondentów) w sanktuarium w Kodniu w lipcu 2009 r.

Źródło: opracowanie K. Wasiluk na podstawie badań terenowych.

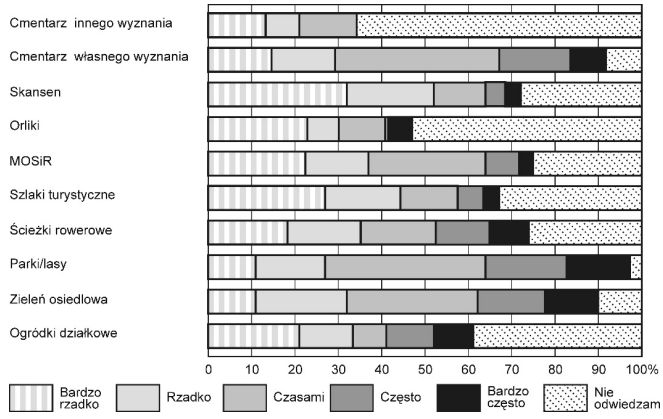
Jeśli mamy za zadanie prezentację struktury cech jakościowych w skali nominalnej w pewnym okresie, wówczas możemy posłużyć się wykresem słupkowym lub kolumnowym złożonym. Poszczególne prostokąty odpowiadają jednemu rokowi i podzielone są na części odpowiadające wielkości każdej cechy (rys. 2.3.11).



Rysunek 2.3.11. Powierzchnia (w ha) działek przeznaczonych na rekreację według ich własności w latach 1985–2015

Źródło: opracowanie własne (dane umowne).

Struktura odpowiedzi wyrażonych w skali porządkowej może być przedstawiona w wartościach bezwzględnych albo procentowych w postaci wykresu słupkowego złożonego (rys. 2.3.12). Należy tak ułożyć odpowiedzi, aby opinie pozytywne były po jednej stronie osi, a negatywne po drugiej. Zjawisko dobrze zilustruje odpowiedni dobór kolorów, np. zimnych dla odpowiedzi negatywnych i ciepłych dla pozytywnych.



Rysunek 2.3.12. Częstotliwość odwiedzin wybranych terenów rekreacyjnych Sieradza w 2011 r. przez respondentów (*N* = 120)

Źródło: opracowanie J. Ziemniak na podstawie badań terenowych.

Szeregi dynamiczne, które – przypomnijmy – ilustrują rozmiary zjawiska w pewnym czasie, można zaprezentować za pomocą histogramów, ale częściej wykorzystuje się do tego celu wykresy liniowe lub słupkowe (rys. 2.3.13). Jeśli zachodzi potrzeba, można na jednym wykresie umieścić dwie zmienne. W przypadku gdy mają one różne jednostki miary albo gdy dane występują w postaci liczb bezwzględnych i względnych, albo gdy prezentuje się zjawiska o różnych rzędach wielkości (rys. 2.3.14), tworzy się dwie osie *OY* po obu stronach wykresu.

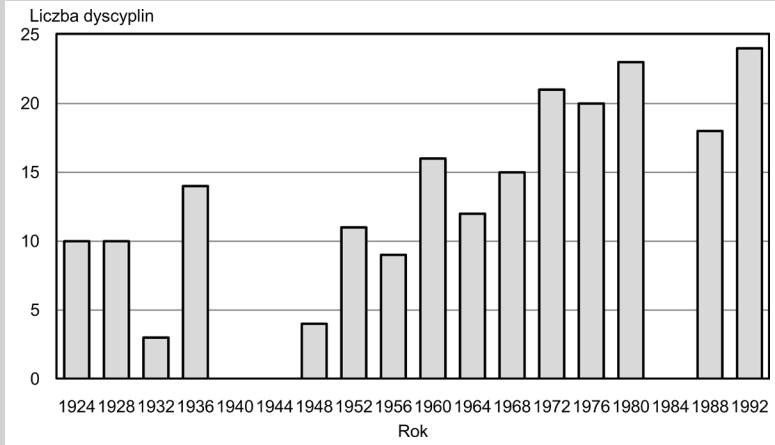
Przykład 2.3.3

Przedstawimy za pomocą wykresów udział Polski w letnich igrzyskach olimpijskich: liczbę zawodników i zdobyte medale (tab. 2.3.3). Spróbuj wyjaśnić, dlaczego nie ma dla Polski danych w latach: 1940, 1944 i 1984.

Tabela 2.3.3. Udział Polski w igrzyskach olimpijskich w latach 1924–1992

Rok, miasto	Liczba		
	zawodników	reprezentowanych dyscyplin	zdobytych medali
1924 – Paryż	66	10	2
1928 – Amsterdam	64	10	5
1932 – Los Angeles	20	3	7
1936 – Berlin	112	14	6
1948 – Londyn	24	4	1
1952 – Helsinki	128	11	4
1956 – Melbourne	64	9	9
1960 – Rzym	186	16	21
1964 – Tokio	140	12	23
1968 – Meksyk	177	15	18
1972 – Monachium	290	21	21
1976 – Montreal	223	20	24
1980 – Moskwa	306	23	32
1984 – Los Angeles	–	–	–
1988 – Seul	143	18	16
1992 – Barcelona	207	24	19

Źródło: GUS (1994).

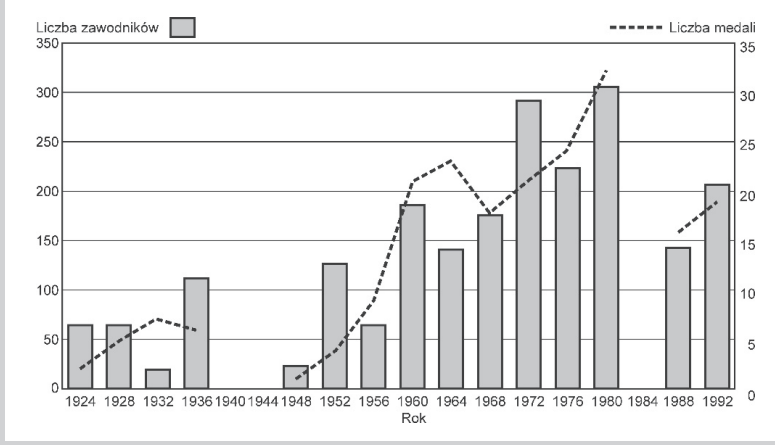


Rysunek 2.3.13. Liczba dyscyplin reprezentowanych przez polskich sportowców na letnich igrzyskach olimpijskich w latach 1924–1992

Źródło: opracowanie własne na podstawie tab. 2.3.3.

Do przedstawienia jednej zmiennej wybrano wykres słupkowy. Na osi *Y* zaznaczono czas (*t*), a na osi *X* liczebność dyscyplin (f_i) (rys. 2.3.13).

Jeśli chcemy zilustrować równocześnie wartość dwóch zmiennych w tym samym czasie, to odpowiedni będzie wykres liniowy z dwiema osiami: jedna prezentuje przykładowo liczbę polskich sportowców biorących udział w igrzyskach, a druga – liczbę zdobytych medali (rys. 2.3.14).



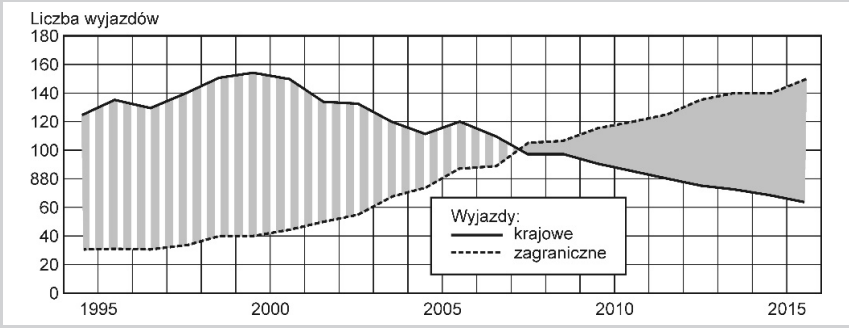
Rysunek 2.3.14. Zawodnicy polscy i zdobyte medale na letnich igrzyskach olimpijskich w latach 1924–1992

Źródło: opracowanie własne na podstawie tab. 2.3.3.

Nałożenie dwóch wykresów liniowych z jedną osią Y może prowadzić do przedstawienia trzeciej zmiennej, a mianowicie różnicy między dwiema badanymi cechami (rys. 2.3.15). W ten sposób można pokazać różnicę między wyjazdami turystycznymi krajowymi a zagranicznymi. W przykładzie 2.3.4 widać nie tylko liczbę wyjazdów, lecz także moment, w którym liczba wyjazdów zagranicznych przekroczyła liczbę wyjazdów krajowych oraz tendencję.

Przykład 2.3.4

Na podstawie liczby podróży krajowych i zagranicznych w latach 1995–2016 pracowników banku wskaż saldo i kierunki tych wyjazdów.



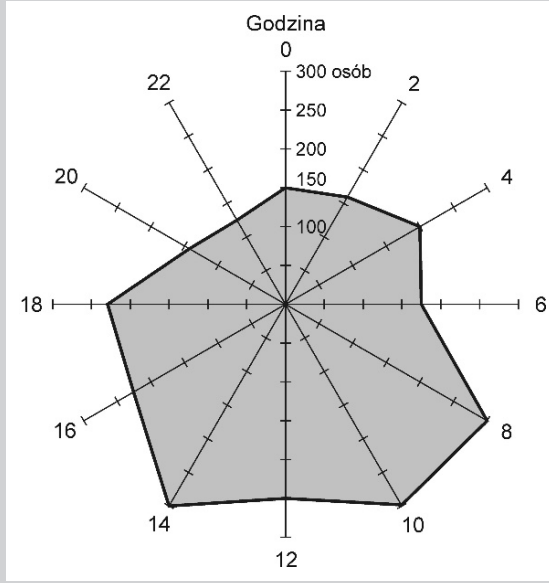
Rysunek 2.3.15. Wyjazdy krajowe i zagraniczne pracowników banku w latach 1995–2016

Źródło: opracowanie własne (dane umowne).

Do zmian cyklicznych jednej cechy, np. co dwie godziny w ciągu doby, można zastosować **radiogram**. Jest to wykres o układzie współrzędnych biegunowych, w których jedną zmienną wyraża się w postaci kąta, a drugą – odległością od środka (początku) układu współrzędnych (rys. 2.3.16). Za jego pomocą można zilustrować wielkość cechy w skali interwałowej. Można go wykorzystywać do badania aktywności turystyczno-rekreacyjnej w ciągu doby, np. na szlaku, lotnisku, w kawiarni.

Przykład 2.3.5

Zbadano liczbę osób przebywających na dworcu w Lublinie 31 sierpnia 2015 r. Co dwie godziny liczono podróżnych.



Rysunek 2.3.16. Dobowa zmiana liczby osób na dworcu w Lublinie 31 sierpnia 2015 r.

Źródło: opracowanie własne (dane umowne).

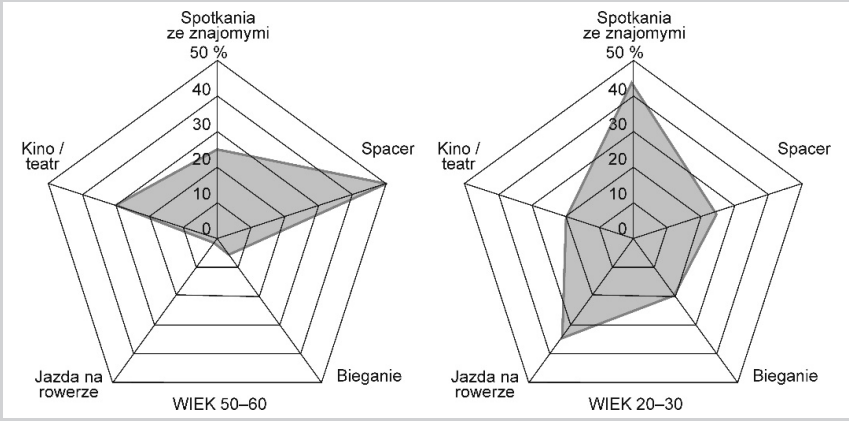
Interpretacja. Rysunek 2.3.16 prezentuje dobową zmianę liczby osób na dworcu w Lublinie i wskazuje najmniejszy udział pasażerów (ok. 150 osób) w porze nocnej – od godz. 20^{00} do 6^{00}. Grupa podróżnych rośnie gwałtownie od godz. 8^{00} i stan ten utrzymuje się do godz. 10^{00}. Następnie liczba pasażerów nieco spada w południe, ale już o godz. 14^{00} wzrasta do maksymalnej wartości 300 osób. Później systematycznie spada i ok. godz. 18^{00} osiąga stan 230 osób.

Badana doba to ostatni dzień wakacji, wobec czego należałoby jeszcze przygotować wykres prezentujący to zagadnienie w dzień powszedni roku szkolnego i porównać z omówionym.

Innym rodzajem wykresu jest **typogram** (Uhorczak, Ostrowski 1972), w którym liczba osi może przedstawiać liczbę wariantów cech, a na poszczególnych osiach odkładane są wartości bezwzględne lub względne (%) różnych cech (rys. 2.3.17).

Przykład 2.3.6

Chcemy dowiedzieć się, jakie są różnice w zachowaniach rekreacyjnych dwóch ankietowanych grup osób w wieku 50–60 i 20–30 lat. Przedstawimy strukturę procentową zachowań rekreacyjnych ankietowanych osób za pomocą typogramu.



Rysunek 2.3.17. Struktura procentowa zachowań rekreacyjnych respondentów w zależności od ich wieku

Źródło: opracowanie własne (dane umowne).

Interpretacja. Zaprezentowane w postaci typogramów zachowania rekreacyjne badanych osób pozwalają na stwierdzenie, że obydwie grupy różnią się. Wśród respondentów w wieku 50–60 lat największym zainteresowaniem cieszą się spacery (50%), podczas gdy dla ankietowanych w wieku 20–30 są one na trzecim miejscu (25%). Młodsi respondenci preferowali spotkania ze znajomymi (45%), jazdę na rowerze (25%) i bieganie (20%). Z kolei starsi nie uczestniczyli tak często w aktywnych formach rekreacji (jazda na rowerze 2%, bieganie 6%), ale częściej bywali w kinie i w teatrze (30%) niż młodsi (20%).

Chcąc skonstruować szeregi geograficzne, stanowiące jedną z odmian szeregów strukturalnych, należy koniecznie zapoznać się z literaturą kartograficzną (Kraak, Ormeling 1998; Ratajski 1989; Pasławski 2003, red. 2006; Żyszkowska, Spallek, Borowicz 2012). Można je wprawdzie przedstawić w postaci wykresu, ale dużo więcej informacji uzyska się z mapy. Jest to oddzielny obszerny temat do zaprezentowania, wymagający znajomości kartografii; w tym podręczniku nie będzie on poruszany.

2.4. ZADANIA

ZADANIE 2.4.1

W celu porównania standardu schronisk młodzieżowych w Polsce w 1999 r. wylosowano 25 schronisk i w każdym z nich zbadano następujące cechy: liczba łóżek, wyposażenie w centralne ogrzewanie, cena jednego noclegu w pokoju dwuosobowym, możliwość dojazdu PKP, PKS lub innym środkiem komunikacji. Na podstawie uzyskanych informacji zaprojektuj szeregi rozdzielcze.

ZADANIE 2.4.2

Na podstawie wpisu do księgi meldunkowej w pensjonacie „Rybitwa" w Szczecinie w czerwcu 2015 r. uzyskano następujące dane liczbowe dotyczące wieku gości: 60, 50, 20, 20, 30, 30, 12, 12, 13, 35, 15, 15, 13, 44, 56, 47, 38, 59, 70, 24, 23, 1, 2, 9, 8, 33, 5, 6, 66, 34, 22, 8, 6, 41, 8, 31, 34, 24, 56, 78, 2, 12, 13. Przedstaw szeregi rozdzielcze o wielkości przedziałów: 2 lata, 5 lat, 10 lat, 20 lat.

ZADANIE 2.4.3

Przedstaw graficznie przyjazdy cudzoziemców do Polski w 1994 r. według celów: odwiedziny – 16 675 100 osób, handlowy – 5 904 000, turystyczny – 39 923 300 i tranzyt – 6 747 100.

Źródło: GUS (1995c).

ZADANIE 2.4.4

W jakiej skali pomiarowej przedstawiane są następujące dane statystyczne?

a) typy obiektów noclegowych (hotel, motel, schronisko, kwatera prywatna),

b) wykonywany zawód (nauczyciel, księgowy, stolarz, rolnik, ślusarz, pielęgniarka),

c) miejsce urodzenia (miasto lub województwo),

d) religia (rzymskokatolicka, greckokatolicka, islam),

e) standard hoteli (trzygwiazdkowe, czterogwiazdkowe itd.),

f) typy szkół (podstawowa, gimnazjum, liceum, pomaturalna, wyższa),

g) temperatura wody w morzu,

h) uzdrowiska według liczby kuracjuszy,

i) wydatki turysty na przejazd, wyżywienie, pamiątki,

j) długość szlaków turystycznych w Pieninach.

ZADANIE 2.4.5

Studenci I roku geoinformacji na Uniwersytecie Łódzkim w 2018 r. otrzymali następujące oceny z przedmiotów „statystyka" i „wstęp do geoinformacji":

a) statystyka: 3, 3, 4, 5, 3, 2, 2, 2, 3, 5, 5, 4, 2, 4, 2, 3, 5, 4, 4, 2, 4, 3, 3, 3, 3, 3, 3, 2, 5, 4, 5, 3, 4, 5, 5, 2, 3, 4, 2, 4, 5, 5, 3, 4, 4, 3, 3, 2, 3,

b) wstęp do geoinformacji: 5, 5, 5, 5, 5, 2, 5, 5, 3, 3, 4, 4, 2, 4, 4, 3, 3, 5, 2, 2, 4, 4, 4, 4, 4, 3, 4, 3, 4, 3, 2, 4, 2, 4, 5, 2, 5, 3, 4, 4, 4, 3, 4, 5, 4, 3, 3, 4.

Dokonaj podziału studentów według otrzymanych ocen i przedstaw wyniki w postaci szeregów statystycznych rozdzielczych.

ZADANIE 2.4.6

Pogrupuj przedsiębiorstwa branży gastronomicznej prowadzące działalność w Lublinie według zysków, jakie osiągnęły w 2013 r. Przyjmij rozpiętość przedziałów 25 tys. zł. Zyski badanych przedsiębiorstw były następujące (w tys. zł): 10, 23, 56, 23, 25, 26, 56, 45, 63, 69, 59, 88, 104, 120, 11, 12, 36, 38, 56, 49, 48, 78, 88, 48, 98, 66, 35, 75, 76, 45, 110, 45, 56, 15, 48, 25, 46, 18, 45, 47, 15, 48, 15, 18, 48, 96, 78, 15, 100.

ZADANIE 2.4.7

Zebrano informacje o liczbie rodzeństwa studentów grupy A kierunku turystyka i rekreacja na Uniwersytecie im. Adama Mickiewicza w Poznaniu w 2010 r. Na podstawie poniższych danych sporządź szereg rozdzielczy i odpowiedz na pytania.

Liczba rodzeństwa: 1, 2, 1, 1, 1, 1, 3, 0, 0, 0, 7, 4, 3, 2, 2, 1, 0, 0, 0, 0, 0, 2, 3, 4, 1, 2, 2, 1, 3, 4, 1, 2, 0, 0, 1;

- czy jest to cecha mierzalna czy niemierzalna?
- jaka liczba rodzeństwa powtarza się najczęściej?
- jaka jest liczebność zbiorowości?
- ile razem rodzeństwa mają wszyscy studenci tej grupy? Utwórz szereg skumulowany;
- ile osób miało więcej niż dwoje rodzeństwa?
- jaki odsetek osób to jedynacy?

ZADANIE 2.4.8

Na podstawie zebranych wśród przyjaciół informacji dotyczących ich wzrostu sporządź szeregi rozdzielcze co 5 cm i 10 cm. Odpowiedz na pytania:

- jaki odsetek osób ma wzrost powyżej 180 cm?
- ile osób ma wzrost poniżej 165 cm?
- jakiego wzrostu jest najwyższa i najniższa osoba w grupie?

ZADANIE 2.4.9

Na podstawie zebranych wśród przyjaciół informacji na temat powierzchni mieszkania, jakie zajmują, rodzaju własności i dzielnicy, w jakiej mieszkają, sporządź szeregi rozdzielcze: co 10 m^2, szereg geograficzny, szereg strukturalny. Odpowiedz na pytania:

- ile osób mieszkało w lokalach o powierzchni poniżej 60 m^2?
- jaka powierzchnia mieszkania występowała najczęściej?
- w której dzielnicy mieszka najmniej przyjaciół?
- jaki odsetek osób mieszka w lokalach własnościowych, komunalnych i spółdzielczych?

ZADANIE 2.4.10

Na podstawie najnowszego rocznika statystycznego przedstaw strukturę ludności Polski w postaci piramidy płci i wieku.

ZADANIE 2.4.11

Na podstawie najnowszego rocznika statystycznego przedstaw strukturę ludności Polski według płci i wieku dla miasta oraz wsi w postaci graficznej. Porównaj otrzymane piramidy.

ZADANIE 2.4.12

Dla danych zawartych w tab. 2.4.1 utwórz szereg rozdzielczy przedziałowy, dobierając odpowiednią liczbę klas oraz ich rozstęp.

Tabela 2.4.1. Pomoc udzielona w ramach Planu Marshalla w latach 1948–1951 według krajów

Kraj	1948/1949	1949/1950	1950/1951	Razem
	[mln USD]			
Austria	232	166	70	468
Belgia i Luksemburg	195	222	360	777
Dania	103	87	195	385
Francja	1085	691	520	2296
Grecja	175	156	45	376
Holandia	471	302	355	1128
Irlandia	88	45	0	133
Islandia	6	22	15	43
Niemcy	510	438	500	1448
Norwegia	82	90	200	372
Portugalia	0	0	70	70
Szwajcaria	0	0	250	250
Szwecja	39	48	260	347
Turcja	28	59	50	137
Wielka Brytania	1316	921	1060	3297
Włochy i Triest	594	405	205	1204

Źródło: Gardner (2001), s. 120.

ZADANIE 2.4.13

Sprawdź co 4 godziny liczbę osób w wybranym miejscu turystycznym w twoim mieście. Do prezentacji wyników wykorzystaj radiogram.

ZADANIE 2.4.14

Temperatura powietrza w °C w lipcu i styczniu w stacji meteorologicznej na Śnieżce w 1999 r. była następująca (wartości umowne):

- w styczniu: 2, 3, –2, –5, –5, –7, –6, –8, –9, –10, –7, –8, –4, –5, –3, –6, –10, –4, –3, –2, –1, 0, 0, 0, 0, –1, 1, 1, 1, 2,
- w lipcu: 19, 19, 20, 18, 19, 25, 24, 23, 25, 24, 25, 22, 21, 23, 18, 18, 18, 19, 18, 20, 20, 21, 23, 26, 28, 29, 27, 25, 26, 21, 18.

Przedstaw dane w postaci dwóch szeregów. Na jednym układzie współrzędnych przedstaw wyniki w postaci histogramu. Porównaj oba rozkłady.

ZADANIE 2.4.15

Narysuj diagram prezentujący strukturę wieku pielgrzymów do Santiago de Compostela 31 sierpnia 2016 r. na podstawie tab. 2.4.2.

Tabela 2.4.2. Pielgrzymi do Santiago de Compostela 31 sierpnia 2016 r. według wieku

Wiek	Liczba osób
0–9	196
10–19	284
20–29	265
30–39	268
40–49	333
50–59	170
65 i więcej	226

Źródło: opracowanie własne na podstawie danych umownych.

ZADANIE 2.4.16

Tabela 2.4.3 odzwierciedla strukturę wieku i płci kuracjuszy w sanatorium „Gryf" w Połczynie Zdroju w 2011 r. Przedstaw graficznie (w postaci piramidy) zebrane dane. Przeprowadź ich analizę.

Tabela 2.4.3. Kuracjusze sanatorium „Gryf" w Połczynie Zdroju w 1998 r. według płci i wieku

Wiek	Płeć		Wiek	Płeć	
	Mężczyźni	Kobiety		Mężczyźni	Kobiety
22	4	6	47	47	35
23	1	9	48	47	29
24	1	9	49	50	44
25	4	5	50	56	50
26	5	6	51	56	56
27	10	20	52	45	56
28	11	23	53	40	59
29	17	20	54	44	58
30	11	25	55	45	46
31	16	30	56	50	50
32	20	31	57	56	54
33	25	30	58	59	54
34	25	36	59	58	60
35	23	25	60	45	60
36	26	29	61	58	60
37	40	36	62	56	63
38	45	44	63	58	50
39	40	45	64	70	40
40	39	40	65	65	40
41	36	40	66	69	20
42	40	42	67	68	20
43	42	46	68	50	12
44	41	40	69	58	10
45	47	40	70	30	5
46	45	36			

Źródło: opracowanie własne na podstawie danych umownych.

ZADANIE 2.4.17

Narysuj diagram prezentujący strukturę dyscyplin sportowych, w których Polacy zdobyli złote medale na igrzyskach olimpijskich w Atlancie w 1996 r.

Tabela 2.4.4. Złote medale zdobyte przez reprezentantów Polski na igrzyskach olimpijskich w Atlancie w 1996 r.

Dyscyplina sportowa	Liczba medali
Ogółem	7
Zapasy	3
Judo	1
Lekkoatletyka	1
Strzelectwo	1
Żeglarstwo	1

Źródło: GUS (1997).

ZADANIE 2.4.18

Porównaj zachowania rekreacyjne rówieśników z wynikami zaprezentowanymi na rys. 2.3.1. Przeprowadź odpowiednie badania na grupie minimum 25 osób.

ZADANIE 2.4.19

Przedstaw w postaci graficznej dane dotyczące ludności według wieku w Polsce w miastach i na wsi (tab. 2.4.5). Jaki typ wykresu będzie najodpowiedniejszy?

Tabela 2.4.5. Ludność w miastach i na wsi w Polsce w 1994 r. według wieku (w tys. osób)

Wiek	Liczba ludności w tys. osób	
	w miastach	na wsi
Ogółem	23 858	14 686
0–2	811	690
3–6	1 264	969
7–14	3 195	2 070
15–17	1 216	704
18–19	787	455
20–24	1 670	1 072
25–29	1 478	984
30–34	1 717	1 043

Wiek	Liczba ludności w tys. osób	
	w miastach	na wsi
35–39	2 194	1 102
40–44	2 146	984
45–49	1 622	749
50–54	1 145	587
55–59	1 169	690
60–64	1 121	734
65–69	906	676
70–74	644	514
75–79	310	279
80 i więcej	453	382

Źródło: GUS (1995a).

ZADANIE 2.4.20

Uaktualnij tabelę z przykładu 2.3.3 i przedstaw dane w niej zamieszczone na podobnym wykresie.

ZADANIE 2.4.21

Przedstaw graficznie na wykresie liniowym z dwiema osiami zmiany w liczbie miejsc noclegowych i liczbie pokoi w hotelach pięciogwiazdkowych w Polsce w latach 2012–2018.

Tabela 2.4.6. Liczba miejsc noclegowych i pokoi w hotelach pięciogwiazdkowych w Polsce w latach 2012–2018

Rok	Liczba miejsc noclegowych	Liczba pokoi
2012	11907	6575
2013	11311	6287
2014	12999	6990
2015	14027	7429
2016	14532	7617
2017	15461	8116
2018	16649	8804

Źródło: opracowanie własne na podstawie danych z https://stat.gov.pl (dostęp: 13.06.2019).

ZADANIE 2.4.22

Dane z tab. 2.4.7 przedstaw w postaci wykresów liniowych dla poszczególnych krajów.

Tabela 2.4.7. Przyjazdy cudzoziemców do Polski (w tys. osób) według krajów w latach 1985–1994

Kraj	Rok				
	1985	1986	1987	1988	1989[a]
Ogółem, w tym:	3 436,2	3 851,2	4 776,4	6 195,6	8 232,6
Austria	29,6	29,8	37,8	53,1	75,5
Holandia	30,1	26,9	38,3	45,8	56,1
Niemcy	1 037,8	1 086,8	1 305,5	1 533,3	1 844,4
Węgry	215,9	351,1	469,6	567,0	698,5
Włochy	24,4	23,7	34,6	40,4	59,7

Kraj	Rok			
	1991[a]	1992	1993	1994
Ogółem, w tym:	36 845,8	49 015,0	60 951,2	74 252,8
Austria	133,2	192,2	231,9	292,2
Holandia	159,4	178,9	189,1	340,5
Niemcy	20 885,4	30 687,7	42 574,0	47 488,5
Węgry	179,9	187,3	164,5	178,6
Włochy	122,7	110,6	123,5	174,2

[a] Dane częściowo szacunkowe.

Źródło: GUS (1995c).

ZADANIE 2.4.23

Dane z tab. 2.4.8 przedstaw w postaci graficznej.

Tabela 2.4.8. Sposób organizacji podróży ankietowanych turystów (dane w %) według krajów

Organizacja wyjazdu	Kanada	USA	Szwecja	Ukraina
Zakup pakietu	15	8	10	5
Zakup części usług	25	14	20	10
Tylko rezerwacja	20	22	25	10
Przyjazd samodzielnie zorganizowany	40	55	45	70
Brak danych	0	1	0	5

Źródło: dane umowne.

ZADANIE 2.4.24

Dane z tab. 2.4.9 przedstaw w postaci graficznej.

Tabela 2.4.9. Struktura wydatków poniesionych przez turystów na terenie Polski (w %) w latach 1990–2015

Wydatek	Odsetek turystów		
	1990	2000	2015
Noclegi	30	32	25
Wyżywienie	20	19	21
Transport	11	10	18
Usługi rekreacyjne	5	8	10
Zakupy na własne potrzeby	20	22	25
Zakupy w celach odsprzedaży	10	5	0
Inne	3	3	1
Brak danych	1	1	0

Źródło: dane umowne.

ZADANIE 2.4.25

Dane z tab. 2.4.10 przedstaw w postaci graficznej – w formie wykresu liniowego.

Tabela 2.4.10. Liczba miejsc noclegowych (w tys.) w Monachium według typu obiektu noclegowego w latach 1990–2010

Rok	Miejsca noclegowe			
	hotel	hostel	motel	kemping
1990	160	140	8	–
2000	200	160	10	8
2010	340	120	8	44
Razem	700	420	26	52

Źródło: dane umowne.

ZADANIE 2.4.26

Przedstaw w postaci wykresu słupkowego dane z tab. 2.4.11. Wybierz odpowiednią skalę.

Tabela 2.4.11. Zagraniczne podróże lekarzy według krajów w 2010 r.

Lp.	Kraj	Liczba osób
1.	Hiszpania	30 000
2.	Grecja	21 000
3.	Francja	11 000
4.	Włochy	4000
5.	Brazylia	2000
6.	Ekwador	800
7.	Chiny	400
8.	Włochy	350
9.	Kuba	60
10.	Egipt	50
11.	Norwegia	20
12.	Ukraina	10
13.	Kongo	8
14.	Australia	7
15.	Japonia	6
16.	Rosja	4
17.	Maroko	1

Źródło: dane umowne.

ZADANIE 2.4.27

W Łodzi w 1996 r. powierzchnie poszczególnych rodzajów terenów zieleni były następujące:

1) parki: 503,8 ha

2) zieleńce: 276,9 ha

3) zieleń przyuliczna: 78,1 ha

4) zieleń osiedlowa: 844 ha

5) zieleń towarzysząca zabudowie: 615 ha

6) ogrody działkowe: 711,4 ha
7) cmentarze: 201 ha
8) Ogród Botaniczny: 64,5 ha
9) Miejski Ogród Zoologiczny: 17 ha
10) ośrodki sportu i rekreacji: 98 ha

Źródło: *Założenia*... (1997).

Przedstaw strukturę zieleni miejskiej w postaci wykresu kołowego. Odpowiedz na pytania:

- jaki odsetek zajmują w Łodzi parki?
- jakiego typu zieleni jest najwięcej?

ZADANIE 2.4.28

Przedstaw graficznie wyniki badań czeskiego rynku turystycznego, korzystając z danych z tab. 2.4.12.

Tabela 2.4.12. Zakwaterowanie Czechów podczas ich wyjazdów zagranicznych w 2013 r. (w %)

Rodzaj wyjazdu	Zakwaterowanie				Razem
	w hotelu	u przyjaciół lub krewnych	w innych obiektach zbiorowych noclegów	w innych miejscach	
Długi	64	12	6	18	100%
Krótki	43	30	5	22	100%

Źródło: *Badanie*... (2014), http://www.pot.gov.pl (dostęp: 26.09.2015).

ZADANIE 2.4.29

W trakcie badań czeskiego rynku turystycznego poproszono Czechów o ocenę kilku krajów.

Poniżej znajdziesz kilka cech. Oceń każdą z nich, zwracając uwagę na to, czy według Ciebie pasuje do Polski, czy nie.

Wykonaj typogramy oceny Czechów dla Litwy, Słowacji, Polski, Węgier i Estonii oraz porównaj je, korzystając z danych zawartych w tab. 2.4.13.

Tabela 2.4.13. Ocena wybranych krajów przez Czechów w 2014 r.

Cecha	Litwa	Słowacja	Polska	Węgry	Estonia
	liczba odpowiedzi przy				
	N = 288	*N* = 227	*N* = 1008	*N* = 278	*N* = 244
Nowoczesny	20	46	30	16	24
Podobny do krajów Europy Zachodniej	17	37	25	16	21
Posiada bogate dziedzictwo, historię	46	74	73	35	44
Zdrowy, ekologiczny	26	48	22	20	31
Nieznany	57	10	20	11	54
Warty poznania	60	87	71	40	60
Pełen energii życiowej	24	62	34	26	30

Źródło: *Badanie*... (2014), http://www.pot.gov.pl (dostęp: 26.09.2015).

2.5. ODPOWIEDZI DO WYBRANYCH ZADAŃ

ZADANIE 2.4.4

a–d – nominalna; e–f – porządkowa; g – interwałowa, h–j – ilorazowa.

ZADANIE 2.4.10

Skorzystaj z rys. 2.3.9: *Struktura płci i wieku mieszkańców Łodzi w 2010 r.*

ZADANIE 2.4.11

Skorzystaj z rys. 2.3.9: *Struktura płci i wieku mieszkańców Łodzi w 2010 r.*

ZADANIE 2.4.12

Skorzystaj z przykładu 2.1.3; rozstęp może wynosić 1000 mln USD.

ZADANIE 2.4.13

Skorzystaj z przykładu 2.3.6.

ZADANIE 2.4.17
Skorzystaj z przykładu 2.3.1.

ZADANIE 2.4.23
Skorzystaj z rys. 2.3.12.

ZADANIE 2.4.25
Skorzystaj z przykładu 2.3.1, skala logarytmiczna.

ROZDZIAŁ 3

ROZKŁADY ZMIENNYCH LOSOWYCH I ICH WŁASNOŚCI

W badaniach statystycznych dość często jesteśmy zmuszeni (np. z powodu zbyt wysokich kosztów, trudności z dostępnością informacji) do wnioskowania na podstawie pobranej próby. Wówczas niezbędna jest podstawowa znajomość rachunku prawdopodobieństwa i teorii zmiennych losowych. Bez wnikliwego wdawania się w matematyczne szczegóły tej teorii, poniżej zostaną zaprezentowane podstawowe definicje i najczęściej spotykane rozkłady zmiennej losowej.

Zbiór wszystkich możliwych wyników jakiegoś pomiaru, losowania, nazywa się **przestrzenią zdarzeń elementarnych**[1]. Oznacza się ją zazwyczaj wielką grecką literą Ω, a jej elementy małą ω. Podzbiory zbioru Ω nazywa się **zdarzeniami losowymi**. W turystyce przestrzenią zdarzeń elementarnych mogą być obiekty noclegowe, a podzbiorami trzy typy obiektów (hotele, motele, hostele). Zdarzeniem pewnym nazywamy zdarzenie, które musi zajść, a zdarzeniem niemożliwym jest pusty podzbiór zbioru Ω, co oznacza, że takie zdarzenie nie wystąpi (np. nie można wylosować do badań kwater prywatnych).

---

1 Zdarzenie elementarne jest pojęciem pierwotnym, czyli takim, którego się nie definiuje. W wyniku rzutu monetą mamy dwa zdarzenia elementarne, a rzutu kostką – sześć zdarzeń elementarnych.

Każdą funkcję rzeczywistą określoną na zbiorze zdarzeń elementarnych nazywamy **zmienną losową** i oznaczamy dużymi literami od końca alfabetu. Zmienna losowa może przyjmować postać dyskretną lub ciągłą.

Przykład 3.1

Zmienna losowa może przyjmować wartości liczbowe w zależności od opisywanego przypadku, np. liczba zaćmień księżyca w ciągu roku, liczba dni słonecznych (od 0 do 365 dni), temperatura powierza (od –20 °C do 40 °C), wzrost (wysokość w cm) turystów (od 56 cm do 210 cm), wydatki na noclegi (od 0 zł do 500 zł za noc) itp.

Zmienne losowe mogą przyjmować **postać dyskretną** (skokową), jeżeli jej możliwymi wartościami są izolowane liczby (x_1, x_2, x_3, ..., x_n) przyjmowane przez te zmienne z określonym prawdopodobieństwem (p_1, p_2, p_3, ..., p_n).

Najczęściej rozpatrywanymi rozkładami zmiennej losowej dyskretnej są:

- rozkład zero-jedynkowy,
- rozkład dwumianowy (Bernoulliego),
- rozkład Poissona.

Rozkład zero-jedynkowy jest rezultatem takiego doświadczenia, w wyniku którego określone zdarzenie A wystąpi lub nie wystąpi. Zdarzeniom elementarnym realizującym zdarzenie A przyporządkowana jest liczba 1, a zdarzeniom elementarnym nierealizującym zdarzenia – liczba 0. Zmienna losowa X ma rozkład zero-jedynkowy, jeżeli może przyjmować dwie wartości, tj. 0 i 1, z następującymi prawdopodobieństwami:

$$x = \begin{cases} 1 - \text{sukces z } p \\ 0 - \text{porażka z } q \end{cases}$$

gdzie: $q = 1 - p$.

Przykład 3.2

Załóżmy, że turysta zarezerwował nocleg w hotelu i jednakowo prawdopodobne jest, że będzie on miał lub nie własny bezpłatny parking. Wówczas mamy do czynienia z rozkładem zero-jedynkowym, który jest rezultatem

doświadczenia, w wyniku którego określone zdarzenie A wystąpi (hotel ma bezpłatny parking dla gości) lub nie wystąpi (hotel nie ma bezpłatnego parkingu).

Jeżeli wykonujemy serię doświadczeń, to jest ona zgodna ze schematem **Bernoulliego**, gdy doświadczenia są wykonywane niezależnie, a realizacja każdego zdarzenia może być sukcesem (oznaczenie A) z prawdopodobieństwem p lub porażką (oznaczenie B) z prawdopodobieństwem $q = 1 - p$, a prawdopodobieństwo zajścia zdarzeń A i B jest stałe.

Zmienną losową X, oznaczającą liczbę sukcesów w n niezależnych próbach, w których prawdopodobieństwo jest stałe, nazywamy zmienną dwumianowaną (Bernoulliego). Niech $b(k; n, p)$ oznacza prawdopodobieństwo $P(X = k)$, że w n próbach Bernoulliego o prawdopodobieństwach p dla sukcesu, a q dla porażki daje w wyniku k sukcesów ($k - 1, \ldots, n$) i $n - k$ porażek. Wówczas:

$$b(k;n,p) = \binom{n}{k} p^k (1-p)^{n-k},$$

gdzie: $k = 1, \ldots, n$

$$\binom{n}{k} = \frac{n!}{n!(n-k)!}.$$

Przy samych sukcesach prawdopodobieństwo to jest równe $b(n; n, p) = p^n$, a przy n porażkach mamy: $b(0; n, p) = q^n$.

Przykład 3.3

W pewnym mieście turystycznym prawdopodobieństwo znalezienia obiektu gastronomicznego, w którym serwowane są posiłki dla wegan jest równe 0,4. Oblicz prawdopodobieństwo, że spośród wybranych do badania sześciu lokali posiłki dla wegan serwowane były w dwóch lokalach.

Stosujemy wzór: $b(k; n, p)$, gdzie $n = 6$, $p = 0{,}4$, $k = 2$, a $q = 0{,}6$. Zatem prawdopodobieństwo, że w 6 doświadczeniach 2 dadzą pomyślny wynik wyniesie:

$$b(2;\ 6,\ 0{,}4) = \binom{6}{2} 0{,}4^2 (1-0{,}4)^{6-2} = 0{,}311.$$

Rozkład Poissona jest szczególnym rodzajem rozkładu dwumianowanego o parametrach n i p, których iloczyn jest wielkością stałą. Wzór Poissona oznacza tzw. prawo rzadkich zdarzeń, gdy liczba prób jest duża, a prawdopodobieństwo sukcesu – małe:

$$P(X = k) = \frac{\lambda^k}{k!} e^{-\lambda}$$

gdzie:

$\lambda > 0$, $\lambda = n \cdot p$,

e – podstawa logarytmu naturalnego,

k – liczba zrealizowanych „sukcesów".

W celu uniknięcia żmudnych rachunków opracowano tablice pozwalające uzyskać $P(X = k) = pk$ dla różnych λ. Wykres zależy od jego parametrów.

Rozkład Poissona ma szerokie zastosowanie praktyczne z tego względu, że wiele zjawisk i procesów można opisać za pomocą zmiennej losowej o tym rozkładzie.

Jednym z prostszych sposobów określenia uporządkowania przestrzennego, tj. wzajemnego usytuowania punktów, jest wykorzystanie własności rozkładu Poissona. Możemy dzięki niemu sprawdzić, czy położenie tych punktów ma charakter losowy, czy nie. Punktami tymi mogą być: lokalizacja budynków hotelowych w przestrzeni miasta, położenie osad w przestrzeni województwa albo kraju, lub też rozmieszczenie turystów na plaży.

Dzielimy wówczas badany obszar na jednakowej wielkości kwadraty, zliczamy w każdym z nich liczbę „sukcesów", a następnie wstawiamy do wzoru. Jeśli rozmieszczenie punktów miało charakter losowy, to liczebności empiryczne powinny odpowiadać liczebnościom uzyskanym za pomocą rozkładu Poissona[2].

Zmienną losową X nazywamy **ciągłą**, jeżeli może ona przyjmować każdą wartość rzeczywistą z pewnego skończonego lub nieskończonego przedziału liczbowego. Dla zmiennej tej istnieje rzeczywista funkcja $f(x) \geq 0$, całkowalna w tym przedziale i spełniająca warunek:

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

2 Przykład zastosowania rozkładu Poissona w ujęciu przestrzennym można znaleźć w: Jaždżewska (2013), s. 144.

Funkcję $f(x)$ nazywamy **gęstością** zmiennej losowej X.

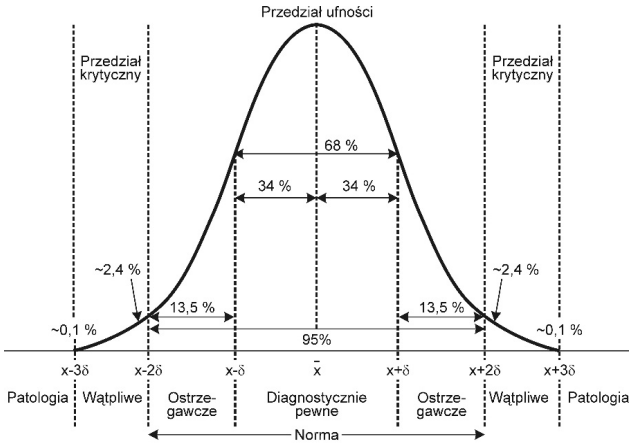
Jednym z najczęściej występujących rozkładów zmiennych losowych ciągłych **jest rozkład normalny** (zwany również rozkładem **Gaussa**). Wiele zjawisk i procesów opisuje się za pomocą zmiennych losowych o takim rozkładzie. Rozkład normalny mają następujące zjawiska: ciepło promieniowania, wartości ciśnienia atmosferycznego, temperatura powietrza i wody, rozmiary organizmów, koncentracja pierwiastków w skałach i glebach.

Zmienna losowa ma taki rozkład, jeżeli jej funkcja gęstości prawdopodobieństwa wyraża się wzorem:

$$f(x) = \frac{1}{\sigma\sqrt{2\pi}} e \frac{(x-m)^2}{2\sigma^2}$$

gdzie: m i σ są stałe. Parametr m to wartość średnia populacji, względem której rozkład jest symetryczny, a parametr σ to odchylenie standardowe[3].

Najczęściej nie jest znany parametr m i szacuje się go na podstawie średniej obliczonej z próby. Wykres tej funkcji ma kształt dzwonu, stąd **krzywa Gaussa** nazywana jest często krzywą dzwonową (rys. 3.1). Jej charakterystyczne cechy będą wykorzystywane w analizie statystycznej.



Rysunek 3.1. Krzywa rozkładu normalnego (Gaussa)

Źródło: Jażdżewska (2013), zmodyfikowane.

3 Definicje obydwu parametrów można znaleźć w rozdziale 4.

Rozkład normalny ma szerokie zastosowanie w statystyce, gdyż jego własności i kształt służą do porównywania rozkładów empirycznych (czyli danych uzyskanych w trakcie badań jakiegoś zjawiska) z rozkładem Gaussa. Niezmiernie rzadko zdarza się, że dane zebrane podczas eksperymentu są zbliżone do rozkładu normalnego, w którym około 68% populacji znajduje się w odległości jednego odchylenia standardowego od średniej, około 95,5% w odległości dwóch odchyleń standardowych i około 99,7% w odległości trzech odchyleń standardowych (reguła trzech sigm). Mogą występować różne odchylenia, które są opisywane i stanowią przedmiot analizy statystycznej opisowej (rozdział 4).

ROZDZIAŁ 4

ANALIZA JEDNEJ ZMIENNEJ

Analizy statystyczne rozpoczyna się często od metod statystyki opisowej, które m.in. zawierają analizę jednej zmiennej. Obejmuje ona obliczenie miar: średnich, rozproszenia, asymetrii i koncentracji. Zmienna, która będzie analizowana nie może być w skali nominalnej, lecz wyższej: porządkowej, interwałowej lub ilorazowej. Dla zmiennej w skali nominalnej można jedynie wskazać (a nie obliczyć) dominantę.

W zależności od techniki obliczania wymienione miary dzielimy na **klasyczne i pozycyjne**. W przypadku miar klasycznych bierze się pod uwagę wszystkie elementy szeregu, natomiast w przypadku miar pozycyjnych uwzględnia się tylko niektóre wartości zmiennej, stojące na określonej pozycji.

4.1. MIARY ŚREDNIE

Do grupy miar średnich zalicza się:

a) miary klasyczne:
- średnią arytmetyczną,
- średnią harmoniczną,
- średnią geometryczną;

b) miary pozycyjne:
- kwantyle, w tym:
  - kwartyle – podział na cztery części danego szeregu,
  - kwintyle – podział na pięć części,
  - decyle – podział na dziesięć części,
  - percentyle – podział na sto części,
- dominantę.

ŚREDNIA ARYTMETYCZNA

Średnia arytmetyczna jest miarą bardzo często stosowaną w analizie statystycznej. Bardzo często jest ona elementem innych miar statystycznych. Należy do grupy miar klasycznych. Otrzymujemy ją w wyniku podzielenia sumy wartości wszystkich jednostek zbiorowości przez jej liczebność. Oznaczamy ją $\overline{x}$. Jest ona wielkością mianowaną, tzn. że interpretując otrzymany wynik, nie można zapominać o jednostce miary danej cechy.

Średnia arytmetyczna prosta liczona dla szeregów szczegółowych ma postać:

$$\overline{x} = \frac{\sum_{i=1}^{n} x_i}{n},$$

gdzie:

n – liczebność próbki lub zbiorowości,

x_i – wartości cechy statystycznej.

Przykład 4.1.1

Długość poszczególnych szlaków pieszych (km) w jednej z miejscowości turystycznych wynosiła: 10, 12, 15, 18, 20 km. Zjawisko przedstawione jest w postaci szeregu szczegółowego, wobec czego należy zastosować średnią arytmetyczną prostą. Średnia długość szlaków pieszych wynosi więc:

$$(10 + 12 + 15 + 17 + 20) / 5 = 14{,}8 \text{ km}.$$

Warto po obliczeniu tej miary sprawdzić, czy jej wartość zawiera się w przedziale: $x_{min} < \overline{x} < x_{max}$, tzn. 14 km < 14,8 km < 20 km.

Średnia arytmetyczna ważona liczona dla szeregów rozdzielczych ma postać:

$$\overline{X} = \frac{\sum_{i=1}^{n} x_i f_i}{\sum_{i=1}^{n} f_i},$$

gdzie:

f_i – liczebność w i-tym przedziale,
x_i – wartości cechy statystycznej,
n – liczebność próbki lub zbiorowości.

Przy obliczaniu średniej arytmetycznej bierze się pod uwagę wszystkie elementy szeregu, dlatego jest ona bardzo wrażliwa na wartości skrajne. Z tego względu należy się nią posługiwać ostrożnie. Warto przed rozpoczęciem obliczeń wykonać histogram rozkładu zmiennych, aby sprawdzić, czy jej wartość ma sens.

Przykład 4.1.2

Nauczyciele Szkoły Podstawowej nr 1 w Toruniu w roku szkolnym 2013/2014 często zabierali dzieci na wycieczki do muzeów. Oblicz średnią liczbę wycieczek, jaką odbył nauczyciel na podstawie danych z tab. 4.1.1.

Tabela 4.1.1. Wycieczki do muzeów dzieci ze Szkoły Podstawowej nr 1 w Toruniu w roku szkolnym 2013/2014

Lp.	Liczba		Iloczyn $x_i f_i$
	wycieczek x_i	nauczycieli f_i	
1.	0	3	0
2.	1	3	3
3.	2	8	16
4.	3	8	24
5.	4	2	8
6.	5	1	5
Suma		28	56

Źródło: dane umowne.

Gdy dane są przedstawione w postaci szeregu rozdzielczego, wówczas korzysta się ze wzoru na średnią arytmetyczną ważoną. W kolumnie 3 obliczono iloczyn wartości cechy x_i i liczby nauczycieli, którzy odbyli określoną liczbę wycieczek f_i.

Stąd:

$$\overline{x} = \frac{\sum_{i=1}^{n} x_i f_i}{\sum_{i=1}^{n} f_i} = \frac{56}{28} = 2,$$

Interpretacja. W roku szkolnym 2013/2014 przeciętnie jeden nauczyciel w Szkole Podstawowej nr 1 w Toruniu wybrał się z uczniami na dwie wycieczki do muzeum.

Jeśli informacje przedstawione są w postaci szeregu rozdzielczego przedziałowego, wówczas w miejsce x_i obliczamy x_i', które oznaczają środek przedziału klasowego x_i (przykład 4.1.3).

Przykład 4.1.3

W pierwszej dekadzie maja 2000 r. do Ojcowskiego Parku Narodowego przybyło 1000 turystów. Oblicz średni wiek turystów na podstawie danych przedstawionych za pomocą szeregu rozdzielczego (tab. 4.1.2).

Tabela 4.1.2. Wiek turystów odwiedzających Ojcowski Park Narodowy w pierwszej dekadzie maja 2000 r.

Wiek turystów $\langle x_{id} - x_{ig})^*$	**Liczba turystów** f_i	**Środek przedziału** x_i'	**Iloczyn** $x_i'f_i$
10–15	300	12,5	3750
15–20	350	17,5	6125
20–25	100	22,5	2250
25–30	150	27,5	4125
30–35	50	32,5	1625
35–40	30	37,5	1125
40–45	20	42,5	850
Suma	1000	×	19 850

* Przedział lewostronnie domknięty, a prawostronnie otwarty.

Źródło: dane umowne.

Algorytm. Do wyznaczenia średniej arytmetycznej potrzebne będą środki przedziałów (x_i'). Należy je wpisać do kolumny 3, a następnie obliczyć iloczyn środków przedziałów (x_i') i liczebności w odpowiednich przedziałach (f_i). Wynik mnożenia należy umieścić w kolumnie 4. Wynik dzielenia sumy kolumny 4 przez sumę kolumny 2 jest szukaną średnią arytmetyczną ważoną.

Wartość średniej arytmetycznej ważonej obliczamy, korzystając ze wzoru:

$$\overline{x} = \frac{\sum_{i=1}^{n} x_i f_i}{\sum_{i=1}^{n} f_i} = \frac{19850}{1000} = 19{,}85,$$

gdzie:

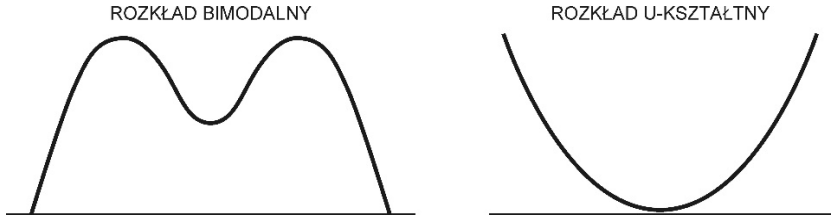
f_i – liczebność w *i*-tej klasie,

x_i' – wartość środka przedziału cechy dla *i*-tej jednostki.

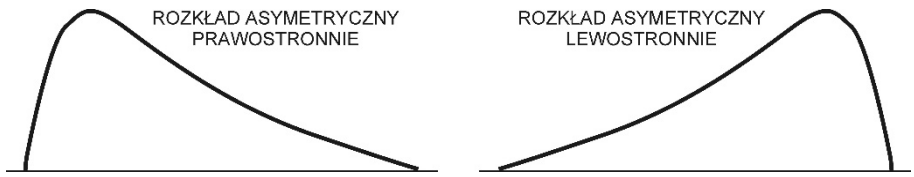
Interpretacja. Przeciętny wiek turystów w Ojcowskim Parku Narodowym w pierwszej dekadzie maja 2000 r. wynosił prawie 20 lat.

Jak wspomniano, średnia arytmetyczna jest jedną z powszechniej stosowanych miar, lecz nie zawsze jej wynik można interpretować. Dlatego nie stosujemy jej do:

- szeregów rozdzielczych o otwartych przedziałach klasowych (por. tab. 2.1.5),
- szeregów, w których występują nietypowe wartości skrajne (por. przykłady 2.1.5, 4.1.5),
- rozkładów bimodalnych (rys. 4.1.1),
- rozkładów typu U-kształtnego (rys. 4.1.1),
- rozkładów skrajnie asymetrycznych (rys. 4.1.2),
- obliczeń, które nie mają sensu (por. przykład 4.1.4).



Rysunek 4.1.1. Rozkład bimodalny i U-kształtny



Rysunek 4.1.2. Rozkład asymetryczny prawostronnie i lewostronnie

Przykład 4.1.4

Na spacerze w parku jest 10 osób i 8 psów. Ile średnio nóg mają właściciele i ich podopieczni? Czy obliczanie tej średniej ma sens? Czy może ona posłużyć do dalszej analizy?

Przykład 4.1.5

W rodzinie składającej się z 4 osób analizowano ich wiek. Matka i ojciec mieli po 40 lat, a ich potomstwo 2 i 6 lat. Jaki jest średni wiek tej rodziny? Czy wartość średnia, tzn. 22 lata, wiarygodnie przedstawia średni wiek zbiorowości? Czy mając jedynie informację o średnim wieku osób można im zaproponować wyjazd wakacyjny z grupą studentów?

ŚREDNIA HARMONICZNA

Średnia harmoniczna jest odwrotnością średniej arytmetycznej. Stosujemy ją wówczas, gdy wartości zbiorowości statystycznej są podane w formie odwrotności, tj. gdy wartości jednej zmiennej są podane w przeliczeniu na stałą jednostkę innej zmiennej (np. km/h), czyli w postaci wielkości względnych.

Używa się jej przy obliczeniach:

a) przeciętnej szybkości pojazdów (km/h),

b) przeciętnego czasu potrzebnego do wykonania pewnej czynności (szt./h),

c) wskaźnika natężenia gęstości zaludnienia (os./km^2).

Stosowana jest w postaci prostej i ważonej.

Średnia harmoniczna prosta liczona dla szeregów szczegółowych ma postać:

$$\overline{x}_h = \frac{n}{\sum_{i=1}^{n} \frac{1}{x_i}},$$

gdzie:

n – liczebność próbki lub zbiorowości,

x_i – wartości cechy statystycznej.

W celu wyjaśnienia zasadności użycia średniej harmonicznej można posłużyć się przykładem.

Przykład 4.1.6

Droga podróży na wakacje miała długość 400 km. Podzielona została na cztery stukilometrowe odcinki, które przejechano z różną prędkością. Pierwsze 100 km przebyto z prędkością 10 km/h, drugi odcinek – 40 km/h, trzeci – 100 km/h, a ostatnie 100 km z prędkością 50 km/h. Jaka była średnia prędkość tej podróży?

Wykorzystamy postać prostą średniej, gdyż odcinki były równej długości.

Gdybyśmy do obliczenia średniej szybkości wykorzystali średnią arytmetyczną, wówczas:

$$\overline{x} = \frac{10 + 40 + 100 + 50}{4} = 40 \text{ km/h}$$

Korzystając ze wzoru na średnią harmoniczną:

$$\overline{x}_h = \frac{4}{\frac{1}{10} + \frac{1}{40} + \frac{1}{100} + \frac{1}{50}} = 25{,}8 \text{ km/h}$$

Różnica w obliczeniach jest znaczna, dlatego zastanówmy się, z czego ona wynika. Na początek obliczmy, jaki był czas przejazdu poszczególnych odcinków przy takich prędkościach?

Czas przejazdu pierwszego 100-kilometrowego odcinka z prędkością 10 km/h to 10 h, czas przejazdu drugich 100 km z prędkością 40 km/h to 2,5 h, czas przejazdu trzecich 100 km z prędkością 100 km/h to 1 h, czas przejazdu czwartych 100 km z prędkością 50 km/h to 2 h. Czas podróży wynosił: 10 + 2,5 + 1 + 2 = 15,5 h.

Przy takich prędkościach podróż 400 km trwała 15,5 godziny. Dlatego średni czas przejazdu 400 km wynosił 25,8 km/h. Korzystając z wzoru na średnią harmoniczną uzyskujemy wynik poprawny.

Interpretacja. Średnia prędkość podróży na wakacje wynosiła 25,8 km/h.

Postać **ważona średniej harmonicznej** to:

$$\overline{x}_h = \frac{\sum_{i=1}^{n} f_i}{\sum_{i=1}^{n} \frac{f_i}{x_i}},$$

gdzie:

f_i – liczebność w i-tym przedziale,
x_i – wartości cechy statystycznej,
n – liczebność próbki lub zbiorowości.

Przykład 4.1.7

Pewien nadmorski region turystyczny składa się z pięciu podregionów, w których wypoczywała różna liczba turystów (f_i).

Obliczmy, jaka była przeciętna gęstość turystów na 1 km^2 w całym nadmorskim regionie, wiedząc że:

do I regionu przyjechało 6000 osób, gęstość wynosiła 30 osób/km^2,
do II regionu przyjechało 5000 osób, gęstość wynosiła 20 osób/km^2,
do III regionu przyjechało 9000 osób, gęstość wynosiła 30 osób/km^2,
do IV regionu przyjechało 8000 osób, gęstość wynosiła 40 osób/km^2,
do V regionu przyjechało 5000 osób, gęstość wynosiła 10 osób/km^2.

Oblicz średnią gęstość turystów w całym regionie.

Cechą statystyczną x_i, którą należy uśrednić jest gęstość. Dlaczego korzystamy ze średniej harmonicznej ważonej? Policzmy, ilu turystów przyjechało do całego regionu i jaka jest jego całkowita powierzchnia. W sumie przyjechało do regionu 33 tys. turystów. Nie mamy podanej powierzchni całego regionu nadmorskiego. Skąd wziąć jego powierzchnię? Skoro mamy gęstość i liczbę osób w podregionach, to możemy obliczyć powierzchnię każdego z podregionów i je zsumować:

I region odwiedziło 6000 osób, gęstość wynosiła 30 osób/km^2, dlatego jego powierzchnia to: 6000/30 = 200 km^2,
II region odwiedziło 5000 osób, gęstość wynosiła 20 osób/km^2, dlatego jego powierzchnia to: 5000/20 = 250 km^2,
III region odwiedziło 9000 osób, gęstość wynosiła 30 osób/km^2, dlatego jego powierzchnia to: 9000/30 = 300 km^2,
IV region odwiedziło 8000 osób, gęstość wynosiła 40 osób/km^2, dlatego jego powierzchnia to: 8000/40 = 200 km^2,

V region odwiedziło 5000 osób, gęstość wynosiła10 osób/km^2, dlatego jego powierzchnia to: 5000/10 = 500 km^2.

Powierzchnia całego regionu to: 200 + 250 + 300 + 200 + 500 = 1450 km^2.

Teraz już można przekonać się, jak była obliczana średnia harmoniczna ważona:

$$\overline{x}_h = \frac{\sum_{i=1}^{n} f_i}{\sum_{i=1}^{n} \frac{f_i}{x_i}}.$$

Interpretacja. Przeciętna gęstość turystów na tym obszarze wynosi 22,76 osób/km^2.

ŚREDNIA GEOMETRYCZNA

Średnia geometryczna to pierwiastek *n*-tego stopnia z iloczynu *n* wartości:

$$\overline{x}_g = \sqrt[n]{x_1 \cdot x_2 \cdot \ldots \cdot x_n} = \sqrt[n]{\prod_{i=1}^{n} x_i},$$

gdzie: $x_i \geq 0$.

Średnia geometryczna jest stosowana do obliczania średniego tempa wzrostu określonych zjawisk przedstawianych za pomocą szeregów dynamicznych (czasowych), np. przyrost produkcji, zatrudnienia, ludności (więcej w rozdziale 6).

MEDIANA

Ważną średnią pozycyjną jest **mediana**, zwana wartością środkową. Dzieli ona szereg na dwie równe części. W jednej części znajduje się 50% jednostek o wartościach wyższych od mediany, a w drugiej 50% jednostek o wartościach poniżej jej wartości. Z tego wynika, że powyżej i poniżej mediany znajduje się jednakowa liczba jednostek.

Dla szeregu rozdzielczego szczegółowego wyznaczenie mediany zaczynamy od ustalenia, czy liczba jednostek jest parzysta, czy nie. Dla nieparzystej liczby jednostek medianę liczymy ze wzoru:

$$m_x = x_{\frac{n+1}{2}},$$

gdzie: *n* – liczba obserwacji.

Przykład 4.1.8

Na kąpielisko termalne w Uniejowie przyszło 35 osób w różnym wieku (dane umowne): 14, 12, 12, 35, 14, 15, 14, 13, 12, 8, 8, 9, 8, 7, 10, 11, 11, 13, 14, 18, 18, 18, 40, 19, 17, 18, 17, 16, 15, 14, 13, 13, 13, 8, 8 lat.

Aby odszukać medianę, trzeba przedstawić dane w postaci szeregu szczegółowego, tj. 7, 8, 8, 8, 8, 8, 9, 10, 11, 11, 12, 12, 12, 13, 13, 13, 13, 13, 14, 14, 14, 14, 14, 15, 15, 16, 17, 17, 18, 18, 18, 18, 19, 35, 40.

Szereg ten ma nieparzystą liczbę wyrazów $n = 35$, należy więc zastosować wzór $x_{(n+1)/2}$. Obliczając $(35 + 1)/2 = 18$, stwierdzamy, że medianą będzie wartość x_{18} szeregu szczegółowego, która wynosi 13 lat.

Interpretacja. Połowa osób w kąpielisku w Uniejowie miała poniżej 13 lat, a połowa była starsza.

Dla parzystej liczby jednostek szeregu statystycznego medianę liczymy ze wzoru:

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

Jeżeli dwa środkowe elementy mają równe wartości, to mediana jest równa tej wartości.

Przykład 4.1.9

Autobusem jedzie 20 osób w różnym wieku: 2, 3, 50, 60, 12, 14, 16, 20, 21, 23, 24, 30, 40, 7, 10, 12, 40, 45, 50, 61 lat. Dane trzeba przedstawić w postaci szeregu szczegółowego, czyli: 2, 3, 7, 10, 12, 12, 14, 16, 20, 21, 23, 24, 30, 40, 40, 45, 50, 50, 60, 61 lat.

Aby znaleźć medianę dla szeregu o parzystej liczbie elementów, szukamy $x_{n/2}$ i $x_{(n/2)+1}$ oraz średniej arytmetycznej tych liczb. Dla $n = 20$ będą to wartości $x_{n/2} = x_{10} = 21$ oraz $x_{(n/2)+1} = x_{10+1} = x_{11} = 23$. Dlatego mediana będzie średnią arytmetyczną $m_x = (21+23)/2 = 22$.

Interpretacja. Połowa pasażerów autobusu była w wieku poniżej 22 lat, a połowa starsza.

Mediana jest szczególnie przydatna, gdy mamy do czynienia z szeregiem o wartościach skrajnych. Na przykład wydatki podczas urlopu w grupie pracowników banku wynosiły w zł: 450, 550, 650, 800, 1000, 1200, 1250, 1300, 1500, 50 000. W tym przykładzie jedna z osób wydała o wiele więcej pieniędzy podczas podróży, gdyż kupiła na pamiątkę wartościowy obraz w antykwariacie za znaczącą kwotę, odbiegającą od wydatków poniesionych przez innych pracowników.

Gdy obliczono średnią arytmetyczną i medianę okazało się, że średnia arytmetyczna wydatków wynosi 5870 zł, a mediana 1100 zł.

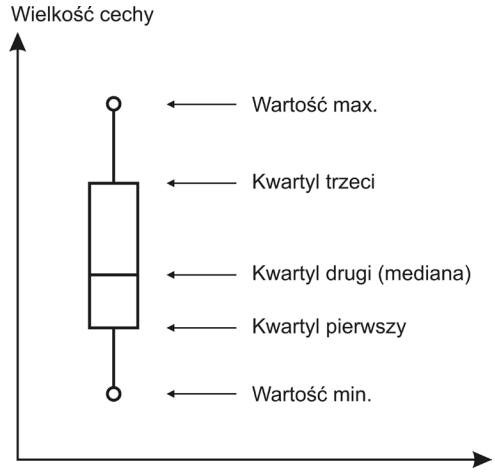
Spróbuj odpowiedzieć na pytanie: Która miara lepiej oddaje wielkość przeciętnych wydatków pracowników banku podczas podróży urlopowych? Uzasadnij odpowiedź.

Podobna sytuacja ma miejsce, gdy mamy do czynienia z szeregiem mocno asymetrycznym, dlatego przeciętne wynagrodzenia lepiej przedstawia wartość mediany niż średniej arytmetycznej (więcej na stronie internetowej: http://wynagrodzenia.pl).

W badaniach statystycznych stosowane są inne miary pozycyjne, które dzielą zbiorowość na równe części: kwartyle, kwintyle, decyle, percentyle.

Kwartyle dzielą zbiorowość na cztery równe części. **Kwartyl pierwszy** to wartość, poniżej której znajduje się ¼ wyrazów szeregu, **kwartyl drugi** jest równy medianie, **kwartyl trzeci** to wartość, poniżej której znajduje się ¾ wyrazów szeregu. Elementy podziału na pięć części nazywamy **kwintylami**, podziału na 10 części – **decylami**, podziału na 100 części – **percentylami**. Aby wyznaczyć te miary, modyfikujemy wzór na medianę w zależności od podziału na cztery, pięć, 10 lub 100 części.

Średnie pozycyjne, takie jak mediana i kwartyle, można również interpretować graficznie. Służy do tego **diagram pudełkowy**, nazywany też wykresem skrzynkowym lub „pudełkiem z wąsami" (rys. 4.1.3). Wartości między kwartylem pierwszym i trzecim zawierają 50% obserwacji i można przypuszczać, że są one typowe dla tej zbiorowości. Powyżej i poniżej pudełka znajduje się kolejne 25% obserwacji. Diagram jest szczególnie przydatny, gdy trzeba porównać kilka zbiorowości.



Rysunek 4.1.3. Diagram pudełkowy, tzw. pudełko z wąsami

Przykład 4.1.10

Do biura podróży zgłosiły się trzy instytucje z prośbą o zorganizowanie wyjazdu na wypoczynek dla pracowników i podopiecznych. Organizator na wstępie zapytał o wiek uczestników. Pierwsza grupa (grupa A) liczyła 26 osób w wieku: 30, 50, 55, 58, 59, 59, 60, 61, 62, 65, 65, 65, 65, 66, 67, 67, 68, 68, 69, 69, 69, 70, 70, 70, 70, 71, druga (grupa B) – 12 osób w wieku: 30, 31, 32, 34, 35, 36, 37, 37, 38, 39, 39, 40, trzecia (grupa C) to młodzież szkolna, w której było: dwoje opiekunów w wieku 30 lat oraz jeden 10-latek, sześciu 11-latków, siedmiu 12-latków, trzech 13-latków, dwóch 14-latków i trzech 15-latków. Obliczymy medianę i kwartyle dla każdej zbiorowości. Wykreślimy potrójny diagram pudełkowy i ocenimy te trzy zbiorowości. Z którą z nich moglibyśmy wysłać naszą babcię (60 lat), wujka (41 lat), koleżankę (20 lat) lub młodszą siostrę (14 lat) na wypoczynek.

Pierwszą zbiorowość, liczącą 26 osób, charakteryzuje parzysta liczba jednostek. Po ich uporządkowaniu obliczamy medianę:

$$m_x = \frac{x_{n/2} + x_{(n/2)+1}}{2} = \frac{x_{13} + x_{14}}{2} = \frac{65 + 66}{2} = 65{,}5$$

Kwartyle to miary dzielące zbiorowość na cztery części (26/4 = 6,5 ≈ 7), dlatego kwartyl pierwszy to element znajdujący się na siódmym miejscu w szeregu szczegółowym.

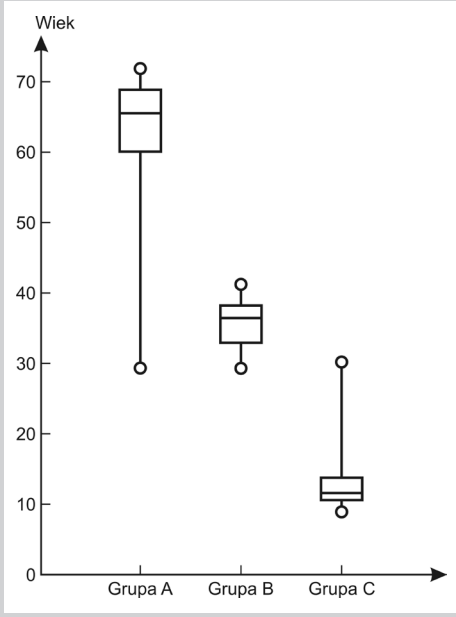
$$Q_1 = \frac{x_{n/4}}{2} = \frac{x_7}{2} = 60.$$

Kwartyl trzeci to element znajdujący się na 20. miejscu w omawianym szeregu (26/4 × 3 = 19,5 ≈ 20).

$$Q_3 = \frac{x_{(n/4)\cdot 3}}{2} = \frac{x_{20}}{2} = 69.$$

Pozostałe miary dla dwóch kolejnych zbiorowości obliczamy analogicznie. Poniższe zestawienie przedstawia wyniki dla wszystkich trzech grup turystycznych.

Grupa	Min.	Max.	Q1	Q2	Q3
A	30	71	60	65,5	69,0
B	30	40	33	36,5	38,5
C	10	30	11	12,0	14,0



Rysunek 4.1.4. Diagram pudełkowy – wiek uczestników wycieczek

Interpretacja. Wiek uczestników wypoczynku w trzech grupach bardzo się różni. W pierwszej waha się od 30 do 71 lat, lecz „trzon" grupy, czyli 50%, to osoby w wieku 60–69 lat. Drugą grupę stanowią ludzie młodsi i ich wiek mieści się w przedziale od 30 do 40 lat. Są więc grupą bardziej jednorodną pod względem wieku. Połowa uczestników jest w wieku od 33 do

38,5 lat, czyli są oni około 30 lat młodsi od członków grupy A. Należy się zastanowić, czy mogą oni pojechać w to samo miejsce, czy raczej trzeba by ich umieścić w innych obiektach noclegowych. Trzecią grupę stanowi młodzież, której wiek waha się od 10 do 30 lat, lecz dwoje z nich to wychowawcy w wieku 30 lat. Najmłodszy uczestnik ma 10 lat, a najstarszy 15 lat. Połowa wszystkich członków tej grupy jest wieku 11–14 lat. Trzeba w tym miejscu postawić pytanie: czy uczestnicy grupy szkolnej wypoczną w towarzystwie dwóch pozostałych i odwrotnie? Sześćdziesięcioletniej babci można zaproponować wyjazd z grupą A, która najlepiej odpowiada jej wiekowi. Z kolei 41-letni wujek chętnie pojedzie z grupą B – będzie, co prawda, najstarszy, ale powinien dobrze się czuć w tym towarzystwie. Koleżanka (20 lat) może nie znaleźć dla siebie towarzystwa w żadnej z grup i zapewne zrezygnuje z wyjazdu lub będzie się musiała zastanowić nad wyborem jednej z nich. Trudno jej cokolwiek zasugerować. Natomiast siostrę, uczennicę gimnazjum, można z pewnością wysłać na wakacje z grupą C – osób w podobnym wieku pod czujnym okiem wychowawców.

DOMINANTA

Dominanta (zwana też modą, wartością modalną) należy do średnich pozycyjnych. Jest to wartość występująca z największą częstotliwością. Jej zaletą jest to, że nie mają na nią wpływu skrajne wartości szeregu. Może być wyznaczona liczbowo tylko wtedy, gdy spełnione są następujące warunki (Luszniewicz, Słaby 1996):

- indywidualny materiał statystyczny jest pogrupowany do postaci szeregu rozdzielczego,
- rozkład empiryczny jest jednomodalny, tzn. ma jedno wyraźnie zaznaczone maksimum,
- rozkład nie jest skrajnie asymetryczny,
- rozpiętości klasowe przedziałów są równe.

Korzystając z programów komputerowych, należy przy obliczaniu dominanty wykazać dużą ostrożność i sprawdzić, w jaki sposób jest ona uzyskiwana. Zdarza się, że jest ona liczona z surowego materiału statystycznego (niepogrupowanego do postaci szeregów) i nie może być prawidłowo interpretowana.

Wyliczenie, a właściwie wskazanie dominanty z szeregu punktowego lub strukturalnego jest proste (por. podrozdział 2.1), gdyż dominantą jest ta wartość cechy, którą przyjmuje największa liczba jednostek.

Przykład 4.1.11

Jaka ocena z wychowania fizycznego dominowała wśród studentów?

Tabela 4.1.3. Oceny z wychowania fizycznego studentów turystyki i rekreacji UŁ w 2017 r.

Ocena	Liczba studentów
2	0
3	15
3,5	25
4	40
4,5	30
5	20
Razem	130

Źródło: dane umowne.

Interpretacja. Dominującą oceną z wychowania fizycznego, jaką uzyskali studenci turystyki i rekreacji UŁ w 2017 r., była ocena dobra.

Aby obliczyć dominantę szeregu rozdzielczego przedziałowego (metodą interpolacji), korzysta się ze wzoru:

$$D_x = x_0 + h_d \frac{f_d - f_{d-1}}{(f_d - f_{d-1}) + (f_d - f_{d+1})},$$

gdzie:

x_0 – dolna granica przedziału dominanty,
f_d – liczebność przedziału dominanty,
f_{d-1} – liczebność przedziału poprzedzającego przedział dominanty,
f_{d+1} – liczebność przedziału następującego po przedziale dominanty,
h_d – rozpiętość przedziału dominanty.

Przykład 4.1.12

Na podstawie danych z tab. 4.1.8 wyznacz dominujące wydatki na obiady studentów wędrujących po Tatrach w 2017 r.

Tabela 4.1.4. Wydatki na obiady studentów wędrujących po Tatrach w 2017 r.

Wydatki w zł x_i	**Liczba studentów f_i**	
5–10	5	
10–15	9	
15–20	10	– przedział poprzedzający przedział dominanty
20–25	16	– przedział dominanty
25–30	13	– przedział następujący po przedziale dominanty
30–35	10	
35–40	7	
Razem	70	

Źródło: dane umowne.

Na wstępie należy prawidłowo wyznaczyć przedział dominanty. Będzie nim czwarty przedział, w którym 16 studentów wydawało na obiady od 20 do 25 zł:

$f_d = 16$ $x_0 = 20$ $f_{d-1} = 10$ $f_{d+1} = 13$ $h_d = 5$

Po podstawieniu do wzoru otrzymuje się wartość modalną:

$$D_x = 20 + 5\frac{16-10}{(16-10)+(16-13)} = 20 + 3,(33) = 23,33.$$

Interpretacja. Badani studenci wędrujący po Tatrach w 2017 r. najczęściej wydawali na obiady około 23,33 zł.

Dla zmiennej w skali nominalnej, np. cele zagranicznych podróży Polaków (tab. 2.1.8), można wskazać dominantę. Wśród celów zagranicznych wyjazdów Polaków dominował cel turystyczno-wypoczynkowy – 42% odpowiedzi w 2007 r. i 52% w 2008 r.

4.2. MIARY ROZPROSZENIA

Analizując strukturę badanej zbiorowości, nie można zapomnieć o miarach rozproszenia, gdyż miary średnie nie charakteryzują w pełni zbiorowości statystycznej. Należy jeszcze poznać strukturę tej zbiorowości. Zadaniem miar rozproszenia jest ukazanie, w jaki sposób wartości jednostek statystycznych koncentrują się wokół wartości centralnej. Znaczenie średniej wzrasta wraz ze zmniejszaniem się stopnia zmienności wokół niej. Zdarza się, że dla dwóch różnych szeregów średnia arytmetyczna jest taka sama, ale szeregi te różnią się znacząco między sobą skupieniem i zmiennością poszczególnych wartości wokół tej średniej (przykład 4.2.1).

Przykład 4.2.1

Obliczyć średnią arytmetyczną i medianę dla następujących szeregów przedstawiających liczbę godzin spędzonych na uprawianiu rekreacji w ciągu miesiąca przez dwie 7-osobowe grupy uczniów:

szereg A: 1, 5, 20, 50, 80, 95, 99,

szereg B: 49, 50, 50, 50, 50, 50, 51.

Dla szeregu A: $\overline{x} = \frac{350}{7} = 50; \;\; m_x = 50.$

Dla szeregu B: $\overline{x} = \frac{350}{7} = 50; \;\; m_x = 50.$

Interpretacja. W każdej grupie uczniowie spędzali przeciętnie po 50 godzin w miesiącu na uprawianiu rekreacji. W obydwu grupach połowa badanych uprawiała rekreację poniżej 50 godzin w miesiącu, a połowa więcej niż 50 godzin w miesiącu. Pomimo że obydwie statystyki są równe, wyraźnie widać, że ich struktury są mocno zróżnicowane. Dlatego warto je zbadać.

Do zbadania zróżnicowania cechy służą miary rozproszenia, zwane również miarami dyspersji, miarami odchyleń lub miarami zmienności. Do ich obliczenia wykorzystuje się wcześniej obliczone miary średnie klasyczne lub pozycyjne.

Do miar rozproszenia zalicza się:

a) miary klasyczne:
- odchylenie przeciętne,
- odchylenie standardowe,
- współczynniki zmienności;

b) miary pozycyjne:
- obszar zmienności (rozstęp),
- rozstęp kwartylny,
- odchylenie ćwiartkowe.

OBSZAR ZMIENNOŚCI (ROZSTĘP)

Najprostszą miarą rozproszenia jest **obszar zmienności**. Miara ta używana jest zazwyczaj w początkowej fazie analizy, kiedy ustala się granice zmienności, zwłaszcza wówczas gdy jesteśmy zainteresowani ekstremalnymi wartościami cechy. Rozstęp jest wielkością mianowaną i obrazuje różnicę między wartością największą i najmniejszą cechy w badanej zbiorowości.

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

Przykład 4.2.2

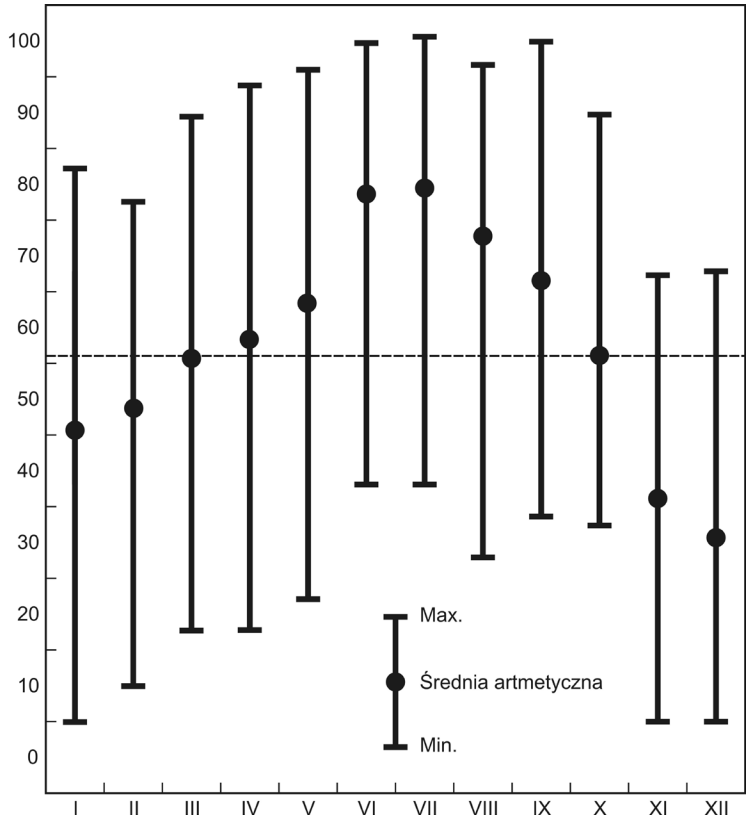
Obszar zmienności dla danych z przykładu 4.2.1 wynosi:

dla grupy A: 99 – 1 = 98,

dla grupy B: 51 – 49 = 2.

Interpretacja. Grupa A była bardziej zróżnicowana co do czasu uprawiania rekreacji, gdyż różnica między minimalnym i maksymalnym czasem spędzonym na rekreację wynosiła 98 godzin, zaś w grupie B jedynie 2 godziny.

Rozstęp można obliczyć (przykład 4.2.2) lub przedstawić graficznie. Jest on jednym z elementów wykresu pudełkowego (podrozdział 4.1). Można na nim pokazać wahania, wartości minimalne, maksymalne i przeciętne w pewnym okresie (rys. 4.2.1), np. liczbę osób przybywających do sanktuarium/hotelu/muzeum w ciągu roku, z podziałem na miesiące, z uwzględnieniem dni o najniższych wartościach, przeciętnych i najwyższych w miesiącu.



Rysunek 4.2.1. Zróżnicowanie liczby gości w hotelu Kapitol w Poznaniu w 2015 r. (w miesiącach).

Źródło: opracowanie własne na podstawie danych umownych.

Dzięki określeniu obszaru zmienności wiemy, jaka jest różnica między krańcowymi wartościami cechy, nie mamy jednak żadnych informacji o zróżnicowaniu pozostałych jej wartości. Nie charakteryzuje on bliżej wewnętrznej struktury badanej zbiorowości. Aby ją określić, należy obliczyć kolejne miary rozproszenia.

ODCHYLENIE PRZECIĘTNE

Odchylenie przeciętne to średnia arytmetyczna bezwzględnych wartości odchyleń poszczególnych wartości od średniej arytmetycznej. W zależności od rodzaju szeregu statystycznego ma postać prostą i ważoną. Dla szeregu szczegółowego otrzymuje się ją ze wzoru:

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

gdzie:

x_i – wartość zmiennej,

$\overline{x}$ – średnia arytmetyczna wartości zmiennej,

n – liczba obserwacji.

Przykład 4.2.3

Obliczmy odchylenie przeciętne dla danych z przykładu 4.2.1.

$$d_A = \frac{|1-50|+|5-50|+|20-50|+|50-50|}{7} +$$
$$+\frac{|80-50|+|95-50|+|99-50|}{7} = \frac{248}{7} = 35,43.$$

$$d_B = \frac{|49-50|+|50-50|+|50-50|+|50-50|}{7} +$$
$$+\frac{|50-50|+|50-50|+|51-50|}{7} = \frac{2}{7} = 0,29.$$

Interpretacja. Rozproszenie liczby godzin spędzanych na rekreacji w grupie A i grupie B różniło się znacznie. W grupie A liczba godzin przeznaczonych na rekreację odchylała się od średniej dla całej grupy (50 h) przeciętnie o 35,43 h, podczas gdy w grupie B jedynie o 0,29 h.

Dla szeregu rozdzielczego odchylenie przeciętne otrzymuje się ze wzoru:

$$d_x = \frac{\sum_{i=1}^{n} |x_i' - \overline{x}| \cdot f_i}{\sum_{i=1}^{n} f_i},$$

gdzie:

x_i' – środek i-tego przedziału klasowego,

$\overline{x}$ – średnia arytmetyczna ważona wartości zmiennej,

n – liczba obserwacji,

f_i – liczebność i-tego przedziału klasowego.

Przykład 4.2.4

Zbadajmy, jakie było zróżnicowanie wieku pracowników hotelu Star w Krakowie w 2013 r. Informacje są przedstawione w postaci szeregu rozdzielczego (tab. 4.2.1, kolumny 1 i 2). Aby obliczyć wartość przeciętną, wypełniamy kolejno cztery ostatnie kolumny. Suma kolumny 4 i 2 pozwala na obliczenie średniej arytmetycznej, a suma kolumny 6 i 3 jest potrzebna do obliczenia wartości przeciętnej.

Tabela 4.2.1. Struktura wieku pracowników hotelu Star w Krakowie w 2013 r.

Wiek pracowników	Liczba pracowników	x_i'	$x_i'f_i$	$\lvert x_i' - \bar{x} \rvert$	$\lvert x_i' - \bar{x} \rvert \cdot f_i$
[1]	[2]	[3]	[4]	[5]	[6]
20–25	5	22,5	112,5	16,45	82,25
25–30	10	27,5	275,0	11,45	114,50
30–35	12	32,5	390,0	6,45	77,40
35–40	20	37,5	750,0	1,45	29,00
40–45	30	42,5	1275,0	3,55	106,50
45–50	23	47,5	1092,5	8,55	196,65
Razem	100	×	3895,0	×	606,30

Źródło: dane umowne.

$$\bar{x} = \frac{3895}{100} = 38{,}95 \approx 39 \text{ lat} \qquad d_x = \frac{606{,}3}{100} \approx 6 \text{ lat}$$

Interpretacja. Średni wiek pracowników hotelu wynosi 39 lat, a wiek poszczególnych pracowników przeciętnie odchyla się od średniej arytmetycznej o około 6 lat.

ODCHYLENIE STANDARDOWE

Najczęściej stosowaną miarą rozproszenia jest jednak **odchylenie standardowe**, obliczane jako pierwiastek kwadratowy z **wariancji** (znana jest również jako **moment centralny rzędu drugiego**). Oznacza się go grecką literą sigma σ , natomiast wariancję σ^2.

Wariancję dla szeregu szczegółowego obliczamy, korzystając ze wzoru:

$$\sigma^2 = \frac{\sum_{i=1}^{n}(x_i - \overline{x})^2}{n},$$

gdzie:

x_i – wartość zmiennej,

$\overline{x}$ – średnia arytmetyczna wartości zmiennej,

n – liczba obserwacji.

Odchylenie standardowe obliczamy ze wzoru:

$$\sigma = \sqrt{\sigma^2}$$

Przykład 4.2.5

Oblicz odchylenie standardowe dla danych z przykładu 4.2.1, przyjmując:

szereg A: 1, 5, 20, 50, 80, 95, 99,

szereg B: 49, 50, 50, 50, 50, 50, 51.

Przypomnijmy, że średnie arytmetyczne oraz mediany dla obydwu szeregów są równe i wynoszą 50 godzin (przykład 4.2.1), lecz wartości szeregu są mocno zróżnicowane.

Tabela 4.2.2. Liczba godzin spędzonych na uprawianiu rekreacji w ciągu miesiąca w grupie A

Liczba godzin A	Odchylenie od średniej $x - \overline{x}$	Kwadrat odchylenia $(x - \overline{x})^2$
1	1 – 50 = –49	2 401
5	5 – 50 = –45	2 025
20	20 – 50 = –30	900
50	50 – 50 = 0	0
80	80 – 50 = 30	900
95	95 – 50 = 45	2 025
99	99 – 50 = 49	2 401
Suma		10 652

Źródło: opracowanie własne na podstawie przykładu 4.2.1.

$$\sigma^2 = \frac{\sum_{i=1}^{n}(x_i - \overline{x})^2}{n} = \frac{10652}{7} = 1521{,}7, \qquad \sigma = \sqrt{1521{,}7} = 39 \text{ godzin}$$

Tabela 4.2.3. Liczba godzin spędzonych na uprawianiu rekreacji w ciągu miesiąca w grupie B

Liczba godzin B	Odchylenie od średniej $x - \bar{x}$	Kwadrat odchylenia $(x - \bar{x})^2$
49	49 – 50 = –1	1
50	50 – 50 = 0	0
50	50 – 50 = 0	0
50	50 – 50 = 0	0
50	50 – 50 = 0	0
50	50 – 50 = 0	0
51	51 – 50 = 1	1
Suma		2

Źródło: opracowanie własne. na podstawie przykładu 4.2.1.

$$\sigma^2 = \frac{\sum_{i=1}^{n}(x_i - \bar{x})^2}{n} = \frac{2}{7} = 0{,}3; \qquad \sigma = \sqrt{\sigma^2} = 0{,}53 \text{ godziny}$$

Interpretacja. Zbiorowości A i B, mające równe wartości średniej arytmetycznej i mediany, różnią się bardzo swoją strukturą. W zbiorowości A wartości oznaczające liczbę godzin spędzonych na rekreacji przez badane osoby przeciętnie odchylają się od średniej arytmetycznej (która wynosi 50 h) o 39 godzin, natomiast w zbiorowości B tylko o 0,53 godziny. Stąd wniosek, że zbiorowość B jest bardziej jednorodna co do zachowań rekreacyjnych niż zbiorowość A, której wartości są bardziej rozproszone.

Wariancję dla szeregu rozdzielczego obliczamy ze wzoru:

$$\sigma^2 = \frac{\sum_{i=1}^{n} f_i(x_i' - \bar{x})^2}{\sum_{i=1}^{n} f_i},$$

gdzie:

x_i' – środek *i*-tego przedziału klasowego,

$\bar{x}$ – średnia arytmetyczna wartości zmiennej,

n – liczba obserwacji,

f_i – liczebność *i*-tego przedziału klasowego.

Odchylenie standardowe obliczamy ze wzoru:

$$\sigma = \sqrt{\sigma^2}$$

Przykład 4.2.6

Przeprowadzono badania dotyczące kosztów noclegów w 49 hotelach pięciogwiazdkowych w Polsce w 2017 r. Informacje przedstawiono w postaci szeregu rozdzielczego (tab. 4.2.4). Zbiorowością statystyczną będą hotele pięciogwiazdkowe, a cechą przeciętny koszt noclegu. Zbadajmy, jakie było zróżnicowanie kosztów noclegów w 49 hotelach pięciogwiazdkowych. Obliczenia należy zacząć od wyznaczenia średniej arytmetycznej, a następnie odchylenia standardowego. Skorzystajmy z pomocniczych kolumn w tab. 4.2.4.

Tabela 4.2.4. Przeciętne koszty noclegów w 49 hotelach pięciogwiazdkowych w Polsce w 2017 r.

Koszty noclegów w zł	Liczba hoteli	x_i'	$x_i' f_i$	$(x_i' - x)^2$	$(x_i' - \bar{x})^2 \cdot f_i$
400–450	8	425	3 400	4 529,3	36 234,4
450–500	27	475	12 825	299,3	8 081,1
500–550	8	525	4 200	1 069,3	8 554,4
550–600	3	575	1 725	6 839,3	20 517,9
600–650	1	625	625	17 609,3	17 609,3
650–700	2	675	1 350	33 379,3	66 758,6
Razem	49	×	24 125	×	157 755,7

Źródło: opracowanie własne na podstawie danych umownych.

$$\bar{x} = \frac{24125}{49} \approx 492{,}3 \text{ zł} \qquad \sigma^2 = \frac{157755{,}7}{49} = 3219{,}5$$

Stąd:

$$\sigma = \sqrt{3219{,}5} = 56{,}7 \text{ zł}$$

Interpretacja. Średnie koszty noclegów w hotelach pięciogwiazdkowych w Polsce w 2017 r. wynosiły 492,3 zł, lecz w poszczególnych hotelach ceny przeciętnie różniły się od średniej dla 49 hoteli o 56,7 zł.

WSPÓŁCZYNNIK ZMIENNOŚCI

Aby porównać zmienność w dwóch próbach o różnych średnich arytmetycznych lub o różnych mianach, można ją wyrazić w procentach, obliczając **współczynniki zmienności**. Współczynniki te są wyrażone stosunkiem dyspersji (odchylenie przeciętne lub standardowe) do modułu średniej arytmetycznej:

$$V_x = \frac{d_x}{|\overline{x}|} \cdot 100\%$$

$$V_x = \frac{\sigma}{|\overline{x}|} \cdot 100\%$$

gdzie:

d_x – odchylenie przeciętne,
$\overline{x}$ – średnia arytmetyczna wartości zmiennej,
σ – odchylenie standardowe.

Współczynnik jest wielkością niemianowaną wyrażoną w procentach.

Przykład 4.2.7

Współczynnik zmienności dla danych z przykładu 4.2.4 wynosi:

$$V_x = \frac{6{,}06}{39} \cdot 100\% = 15{,}5\%$$

Interpretacja. Dyspersja wieku pracowników hotelu Star w Krakowie w 2013 r. nie jest wysoka – wynosi 15,5%.

Współczynniki zmienności mogą mieć zastosowanie w badaniach zróżnicowania takich wielkości, które są w skali interwałowej lub wyższej, jak np. wydatki na podróże, obroty w hotelach, wiek podróżnych, temperatury powietrza, liczby turystów i inne. Można je również wykorzystać do prezentacji zjawisk na mapie (oczywiście należy przyswoić wiedzę z kartografii na temat wizualizacji tych wartości).

ROZSTĘP KWARTYLNY

Jeśli jako podstawę analizy statystycznej przyjęto miary pozycyjne, to do badania rozproszenia rozkładu stosuje się **rozstęp kwartylny**, który jest różnicą

pomiędzy kwartylem trzecim i pierwszym oraz **odchylenie ćwiartkowe** określone wzorem:

$$Q = \frac{Q_3 - Q_1}{2},$$

gdzie:

Q_3 – kwartyl trzeci,

Q_1 – kwartyl pierwszy.

Dzięki rozstępowi kwartalnemu, w granicach którego mieści się dokładnie 50% badanych obiektów, możemy określić stopień rozproszenia zbiorowości (por. przykład 4.1.10).

STANDARYZACJA DANYCH

Załóżmy, że chcemy porównać dwie zbiorowości statystyczne o różnej liczebności i jednostkach miary, np. wydatki na rekreację 100 uczniów szkół w Polsce i 77 uczniów szkół francuskich. Pierwsi ponoszą koszty w zł, a drudzy w euro, obydwie zbiorowości mają inną liczebność: $n_P = 100$ i $n_F = 77$. Aby wyniki obliczonych statystyk były porównywalne, surowe dane o wydatkach $i = 1, \ldots, n$ poddaje się **procedurze standaryzacji** według wzoru:

$$z_i = \frac{x_i - \overline{x}}{\sigma},$$

gdzie:

x_i – dane empiryczne (surowe),

z_i – dane standaryzowane,

$\overline{x}$ – średnia arytmetyczna wartości zmiennej,

σ – odchylenie standardowe.

4.3. MIARY ASYMETRII I KONCENTRACJI

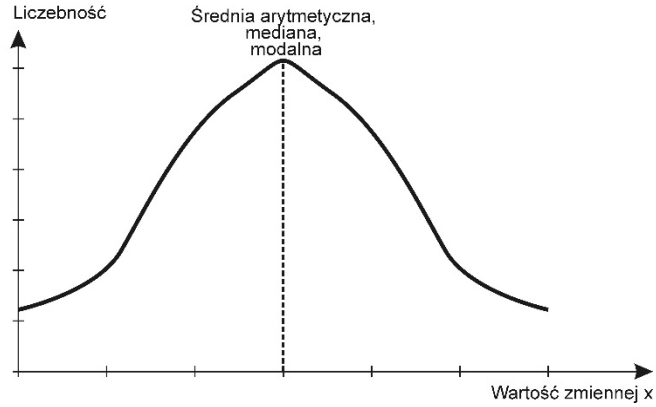
Kolejnym etapem analizy jednej zmiennej jest sprawdzenie, czy jej rozkład jest symetryczny i jaka jest jego koncentracja. Analiza tego zagadnienia pozwala na poszerzenie wiedzy na temat struktury zbiorowości. W grupie miar

asymetrii wyróżnia się: porównanie średnich, wskaźniki asymetrii, moment centralny rzędu trzeciego, współczynnik skośności. Do miar koncentracji zalicza się: moment centralny rzędu czwartego, współczynnik Giniego.

MIARY ASYMETRII

Jeśli rozkład nie jest symetryczny, to istotną kwestią jest zbadanie, czy odchylenia od wartości średniej w jedną stronę są mniej lub więcej liczne od odchyleń w drugą stronę. Można to zbadać za pomocą miar asymetrii, inaczej nazywanych miarami skośności. W szeregu idealnie symetrycznym średnia arytmetyczna, mediana i modalna są równe:

$$\overline{x} = m_x = d_x.$$

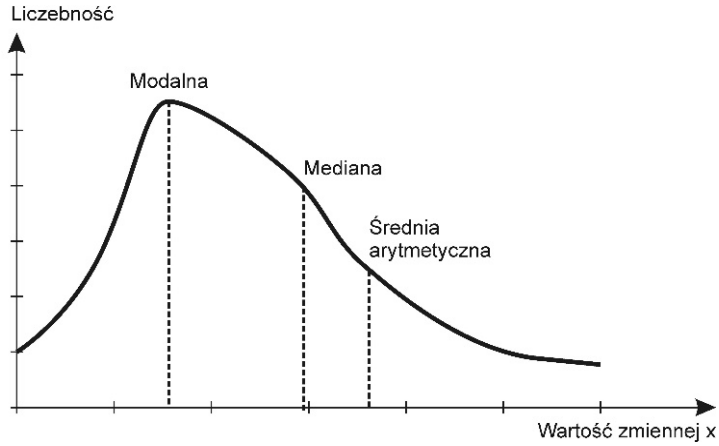


Rysunek 4.3.1. Rozkład symetryczny szeregu statystycznego

Źródło: opracowanie własne.

W szeregach asymetrycznych wartości dominanty, mediany i średniej arytmetycznej różnią się. Im większa jest skośność szeregu, tym większe są różnice między nimi. Po stwierdzeniu, że mamy do czynienia z asymetrią, należy określić jej kierunek i natężenie.

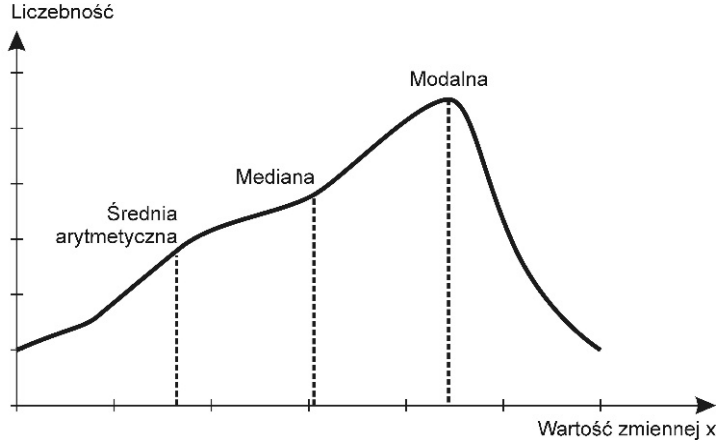
W szeregu o skośności prawostronnej (dodatniej) wartości skrajne położone są z prawej strony średniej. Powoduje to przesunięcie średniej arytmetycznej w prawo w stosunku do wartości najczęstszej (tj. dominanty) i mediany.



Rysunek 4.3.2. Rozkład asymetryczny szeregu statystycznego, asymetria prawostronna, nachylenie dodatnie

Źródło: opracowanie własne.

W szeregu o skośności lewostronnej (ujemnej) wartości skrajne położone są z lewej strony średniej. Powoduje to przesunięcie średniej arytmetycznej w lewo w stosunku do wartości najczęstszej (tj. dominanty) i mediany.



Rysunek 4.3.3. Rozkład asymetryczny szeregu statystycznego, asymetria lewostronna, nachylenie ujemne

Źródło: opracowanie własne.

Najprostszą do obliczeń miarą skośności jest różnica między średnią arytmetyczną i dominantą, wskazującą jednocześnie kierunek asymetrii. Jeżeli $\overline{x} - d_x > 0$, wówczas jest to szereg o asymetrii prawostronnej; jeśli $\overline{x} - d_x < 0$, wówczas jest to szereg o asymetrii lewostronnej. Różnica ta jest bezwzględną miarą skośności, tzw. **miernikiem skośności**.

Ponieważ bezwzględne miary skośności są mało przydatne, zwłaszcza przy porównywaniu cech zbiorowości mierzonych za pomocą odmiennych jednostek, różnicę między średnią arytmetyczną i dominantą dzielimy przez odchylenie standardowe. Otrzymujemy wówczas tzw. **współczynnik skośności**:

$$A_S = \frac{\overline{x} - d_x}{\sigma},$$

gdzie:

d_x – dominanta,
$\overline{x}$ – średnia arytmetyczna wartości zmiennej,
σ – odchylenie standardowe.

Tak obliczony współczynnik jest równy zeru, gdy zbiorowość jest symetryczna, dodatni w przypadku asymetrii prawostronnej i ujemny w przypadku asymetrii lewostronnej.

Siłę i kierunek asymetrii można również zmierzyć, posługując się **momentem centralnym rzędu trzeciego**. Dla szeregów rozdzielczych określony jest wzorem:

$$M_3 = \frac{\sum_{i=1}^{n} f_i (x_i - \overline{x})^3}{\sum_{i=1}^{n} f_i},$$

gdzie:

x_i – wartość zmiennej,
$\overline{x}$ – średnia arytmetyczna wartości zmiennej,
f_i – liczebność i-tego przedziału klasowego,
n – liczba obserwacji.

Dla szeregów symetrycznych $M_3 = 0$, dla szeregów o asymetrii lewostronnej jest ujemny, a prawostronnej – dodatni. Moment centralny rzędu trze-

ciego jest miarą skośności wyrażoną w tych samych jednostkach, co zmienna i podniesiony do potęgi trzeciej. Fakt ten utrudnia interpretację i porównywanie asymetrii, dlatego w celu uzyskania miary asymetrii porównywalnej należy obliczyć stosunek wartości momentu centralnego rzędu trzeciego do sześcianu odchylenia standardowego, czyli **współczynnik asymetrii**:

$$A = \frac{M_3}{\sigma^3}.$$

Znak określa kierunek asymetrii. Przy $A > 0$ mamy do czynienia z asymetrią prawostronną, przy $A < 0$ asymetria jest lewostronna, zaś dla $A = 0$ szereg jest symetryczny. Wartość bezwzględna z A wskazuje na siłę asymetrii. Z reguły współczynnik ten mieści się w granicach od –2 do +2.

Współczynnik asymetrii wykorzystujący moment rzędu trzeciego może być przydatny do oceny wielu zjawisk, np. zarobków w turystyce. Wówczas symetria zerowa odpowiada symetrycznemu rozkładowi zarobków, symetria dodatnia odpowiada rozkładowi z przewagą zarobków mniejszych niż przeciętna, a symetria ujemna – rozkładowi z przewagą większych niż przeciętna.

Przykład 4.3.1

GUS przeprowadził badania dotyczące średnich kosztów noclegów w Polsce w 1995 r. w poszczególnych województwach. Informacje przedstawiono w postaci szeregu rozdzielczego (tab. 4.3.1).

Z przykładu 4.2.6 wiemy, że:

$$\bar{x} = \frac{24125}{49} \approx 492{,}3 \text{ i } \sigma = \sqrt{3219{,}5} = 56{,}7.$$

Zbadajmy kierunek i siłę asymetrii, korzystając z momentu centralnego rzędu trzeciego. Dodajmy pomocnicze kolumny w tab. 4.2.4.

Tabela 4.3.1. Przeciętne koszty noclegów w 49 hotelach pięciogwiazdkowych w Polsce w 2017 r.

Koszty noclegów w zł	Liczba hoteli	x_i'	$(x_i' - \bar{x})^3$	$f_i \cdot (x_i' - \bar{x})^3$
400–450	8	425	–304 821,2	–2 438 569,6
450–500	27	475	–5 545,2	–149 720,4
500–550	8	525	34 965,8	279 726,4

Koszty noclegów w zł	Liczba hoteli	x_i'	$(x_i' - \bar{x})^3$	$f_i \cdot (x_i' - \bar{x})^3$
550–600	3	575	565 609,3	1 696 827,8
600–650	1	625	2 336 752,8	2 336 752,8
650–700	2	675	6 098 396,3	12 196 792,6
Suma	49	×	×	11 585 056,5

Źródło: opracowanie własne na podstawie danych umownych.

$$M_3 = \frac{\sum_{i=1}^{n} f_i \cdot (x_i - \bar{x})^3}{\sum_{i=1}^{n} f_i} = \frac{11585056,5}{49} = 236429,7$$

$$A_S = \frac{M_3}{\sigma^3} = \frac{236429,7}{182284,263} = 1,3$$

Interpretacja. Dla szeregów symetrycznych współczynnik jest równy 0, dlatego mamy w tym przykładzie silną asymetrię prawostronną dodatnią. Oznacza to, że średnia arytmetyczna jest położona na prawo od dominujących wartości. Przeciętny koszt noclegu w badanych hotelach wynosił 492,3 zł, a współczynnik asymetrii równy 1,3 wskazuje, że w większości hoteli koszty noclegu były dużo niższe od przeciętnego.

Jeśli jako podstawę analizy statystycznej przyjęto miary pozycyjne, to do badania asymetrii rozkładu stosuje się **współczynnik skośności** określony wzorem:

$$A_Q = \frac{Q_1 + Q_3 - 2M_x}{2Q},$$

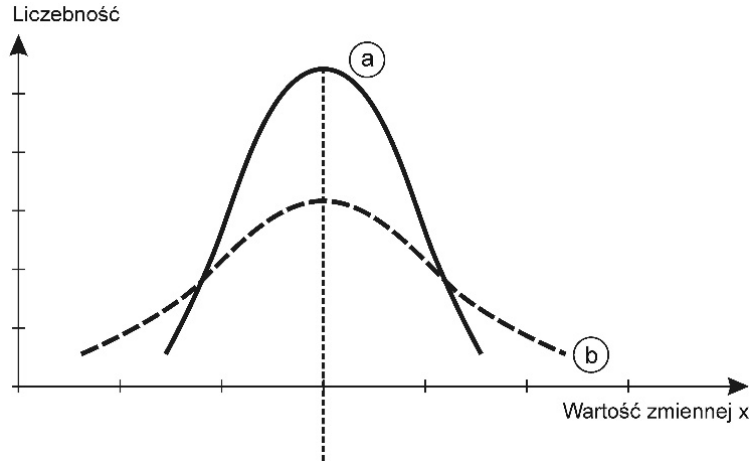
gdzie:

Q_1 – kwartyl pierwszy,
Q_3 – kwartyl trzeci,
M_x – mediana,
Q – odchylenie ćwiartkowe.

MIARY KONCENTRACJI

W statystyce terminu **koncentracja** używa się w celu określenia ścisłości skupienia pojedynczych obserwacji zmiennej wokół pewnej wartości (np. średniej arytmetycznej) oraz do określenia stopnia rozproszenia lub skupienia elementów pewnego zbioru w przestrzeni. Znaczenie tego terminu wynika często z punktu widzenia badającego, np. geografa lub ekonomisty.

Oprócz omawianych dotychczas problemów dotyczących tendencji centralnej, rozproszenia i skośności, zbiorowość statystyczna może być badana pod kątem koncentracji (skupienia) poszczególnych wartości zmiennej wokół średniej arytmetycznej. Skupienie wartości wokół średniej zależy oczywiście od rozproszenia. Im większe jest rozproszenie, tym mniejsza koncentracja i odwrotnie. Jednak dwa szeregi, charakteryzujące się takim samym lub bardzo podobnym odchyleniem przeciętnym czy standardowym (a więc szeregi o tym samym lub podobnym rozproszeniu), mogą różnić się pod względem koncentracji, jeżeli obszar zmienności tych szeregów jest różny. Różnice w skupieniu zbiorowości wokół średniej łatwiej można zaobserwować na wykresie (rys. 4.3.4).



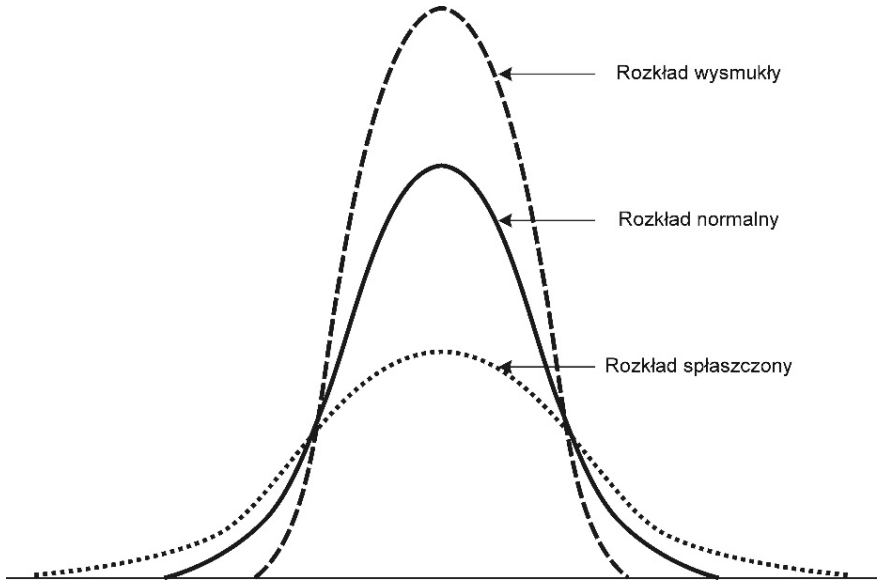
Rysunek 4.3.4. Krzywe rozkładów statystycznych

Źródło: opracowanie własne.

Krzywa oznaczona na wykresie symbolem *a* przedstawia rozkład o większym skupieniu poszczególnych jednostek zbiorowości wokół

średniej. Wysmukły kształt dowodzi, że większa część wartości zmiennej pozostaje w bezpośrednim sąsiedztwie średniej, a tylko niewielka część wartości różni się znacznie od średniej. Krzywa oznaczona na wykresie symbolem *b* ma kształt spłaszczony w porównaniu z kształtem krzywej *a*, co jest równoznaczne z mniejszą koncentracją poszczególnych jednostek szeregu wokół średniej. Aby określić koncentrację zbiorowości wokół średniej, trzeba porównać badany rozkład z innymi rozkładami, np. rozkład przedstawiony za pomocą krzywej *b* z rozkładem zaprezentowanym za pomocą krzywej *a*.

W celu uniknięcia dowolności w wyborze rozkładu, który ma stanowić podstawę porównania, za punkt odniesienia przyjęto rozkład normalny (por. rozdział 3). Tak więc szereg, którego wykres ma postać bardziej wysmukłej krzywej niż krzywa rozkładu normalnego jest szeregiem o większym skupieniu poszczególnych wartości wokół średniej, zaś szereg, którego krzywa jest mniej wysmukła w porównaniu z krzywą rozkładu normalnego (czyli bardziej spłaszczona) jest szeregiem o mniejszej koncentracji poszczególnych wartości wokół średniej.



Rysunek 4.3.5. Krzywe rozkładów statystycznych

Źródło: opracowanie własne.

Miarą natężenia koncentracji zbiorowości wokół średniej jest tzw. **moment centralny rzędu czwartego**:

$$M_4 = \frac{\sum_{i=1}^{n} f_i (x_i - \overline{x})^4}{\sum_{i=1}^{n} f_i},$$

gdzie:

x_i – wartość zmiennej,

$\overline{x}$ – średnia arytmetyczna wartości zmiennej,

f_i – liczebność i-tego przedziału klasowego,

n – liczba obserwacji.

Moment centralny rzędu czwartego jest miarą koncentracji wyrażoną w tych samych jednostkach, co zmienna i podniesiony do potęgi czwartej. Fakt ten utrudnia porównywanie koncentracji różnych szeregów. Dlatego w celu uzyskania miary koncentracji porównywalnej należy obliczyć stosunek wartości momentu centralnego czwartego rzędu do odchylenia standardowego podniesionego do potęgi czwartej. Miara ta nazywana jest **kurtozą**.

$$K = \frac{M_4}{\sigma^4}$$

W przypadku rozkładu normalnego przyjmuje on wartość 3. Wartości większe od 3 mówią o większej koncentracji niż w rozkładzie normalnym, a wartości mniejsze od 3 – o spłaszczeniu rozkładu w porównaniu z rozkładem normalnym. Bardzo wysoka wartość współczynnika dowodzi, że istnieje tendencja do skupiania się wartości wokół średniej.

Przykład 4.3.2

Dane z przykładu 4.3.1:

$$\overline{x} = \frac{24125}{49} \approx 492{,}3 \quad \text{i} \quad \sigma = \sqrt{3219{,}5} = 56{,}7.$$

Rozbudujmy tabelę, aby było łatwiej obliczyć kurtozę.

Tabela 4.3.2. Przeciętne koszty noclegów w 49 hotelach pięciogwiazdkowych w Polsce w 2017 r.

Koszty noclegów w zł	Liczba hoteli	x_i'	$(x_i' - \bar{x})^4$	$(x_i' - \bar{x})^4 \cdot f_i$
400–450	8	425	20 514 467,9	164 115 743,2
450–500	27	475	89 574,5	2 418 511,5
500–550	8	525	1 143 381,1	9 147 048,8
550–600	3	575	46 775 887,7	140 327 663,1
600–650	1	625	310 087 094,3	310 087 094,3
650–700	2	675	114 177 000,9	2 228 354 001,8
Suma	49	×	×	2 854 450 062,7

Źródło: opracowanie własne (dane umowne).

$$M_4 = \frac{\sum_{i=1}^{n} f_i (x_i - \bar{x})^4}{\sum_{i=1}^{n} f_i} = \frac{2854450062,7}{49} = 58254082,9$$

$$K = \frac{M_4}{\sigma^4} = \frac{58254082,9}{10335517,7121} = 5,64$$

Interpretacja. Tak wysoka wartość kurtozy świadczy o dużej smukłości rozkładu i skupianiu się kosztów noclegu w poszczególnych hotelach wokół kosztów przeciętnych. Należy jednak pamiętać o asymetrii (1,3), mówiącej o niższych niż przeciętna kosztach noclegu w większości badanych hoteli.

WSPÓŁCZYNNIK GINIEGO I KONCENTRACJI LORENZA

Współczynnik Giniego (wskaźnik nierówności społecznej) jest znany w literaturze statystycznej jako miara mówiąca o nierównościach dochodów mieszkańców poszczególnych państw. Do jego obliczenia potrzebne są dane o zarobkach, które ustawia się w szeregu szczegółowym i przelicza na procenty. W Polsce według GUS współczynnik ten dla dochodów Polaków osiągał wartość około 30, co plasowało Polskę w połowie wszystkich krajów europejskich.

W turystyce **współczynnik Giniego** może być stosowany jako współczynnik koncentracji dochodów z turystyki, wykorzystania obiektów hotelowych lub w formie mapy koncentracji przestrzennej turystów w określonych jednostkach administracyjnych (szerzej: Jażdżewska 2013). Załóżmy, że badamy zarobki 200 pracowników branży turystycznej. Jeśli współczynnik byłby równy 0, to oznaczałoby, że wszyscy zarabiają tyle samo. Jeśli współczynnik wynosiłby 100, to oznaczałoby, że tylko jeden pracownik osiąga dochody, a reszta pracuje społecznie. Interpretacja współczynnika Giniego zależy od uzyskanych wartości. Jeśli osiąga on wartości bliskie 100, wtedy mówimy o wysokiej koncentracji (nierówności), ale gdy są bliskie 0, wówczas mówimy o małej koncentracji. Rozwinięcie współczynnika Giniego zaproponował **Lorenz**, którego **współczynnik koncentracji** ma szerokie zastosowanie w ekonomii, demografii i innych naukach.

Wskaźnik oblicza się na podstawie tabeli i krzywej Lorenza (przykład 4.3.3). Wartość współczynnika określa się wzorem:

$$K = 1 - \frac{1}{5000} \sum_{i=1}^{k} \frac{s_i'(M_{i-1} - M_i)}{2},$$

gdzie:

s' – procent obiektów,

m' – procent miejsc noclegowych.

$$M_n = m_1' + m_2' + \ldots + m_n'$$

Przykład 4.3.3

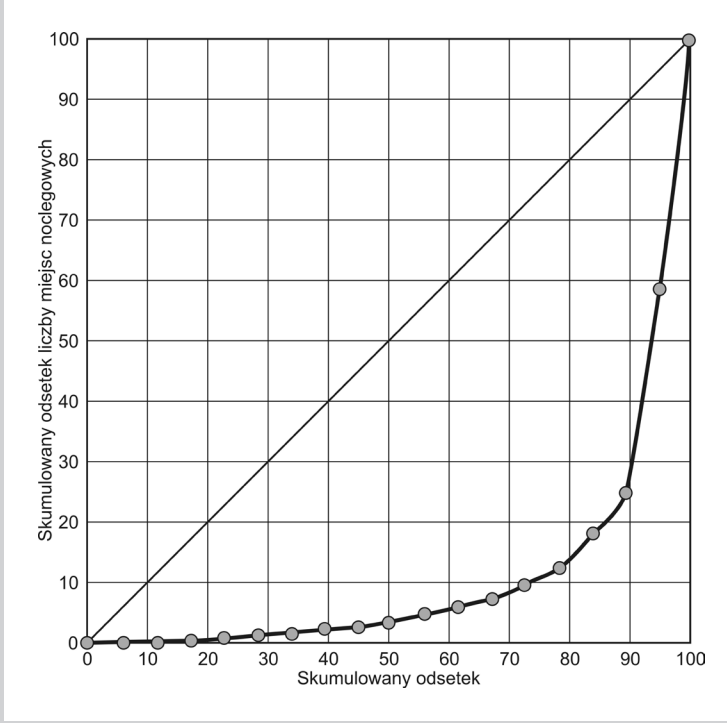
Chcemy poznać koncentrację liczby miejsc w 18 obiektach noclegowych Bratysławy w 2017 r. Przystępując do obliczenia współczynnika koncentracji Lorenza, wstawiamy dane do tablicy (najlepiej w arkuszu kalkulacyjnym), tworzymy z nich szereg statystyczny (sortujemy wartości rosnąco), a następnie dodajemy kolumnę z liczbą obiektów. Obliczamy udział procentowy liczby miejsc noclegowych i liczby obiektów (tab. 4.3.3).

Tabela 4.3.3. Obliczenia współczynnika koncentracji Lorenza

Liczba		Odsetek			
miejsc noc-legowych	**obiektów**	**liczby miejsc noc-legowych**	**obiektów**	**obiektów kumulo-wany**	**liczby miejsc noc-legowych kumulo-wany**
m	***s***	***m'***	***s'***	***s' cum***	***m' cum***
25	1	0,06	5,56	5,56	0,06
132	1	0,30	5,56	11,11	0,35
143	1	0,32	5,56	16,67	0,67
167	1	0,37	5,56	22,22	1,04
187	1	0,42	5,56	27,78	1,46
246	1	0,55	5,56	33,33	2,01
250	1	0,56	5,56	38,89	2,57
253	1	0,57	5,56	44,44	3,14
412	1	0,92	5,56	50,00	4,06
423	1	0,95	5,56	55,56	5,01
530	1	1,19	5,56	61,11	6,19
620	1	1,39	5,56	66,67	7,58
995	1	2,23	5,56	72,22	9,80
1 227	1	2,74	5,56	77,78	12,55
2 679	1	5,99	5,56	83,33	18,54
2 965	1	6,63	5,56	88,89	25,17
15 016	1	33,59	5,56	94,44	58,76

Źródło: opracowanie własne (dane umowne).

Dwie ostatnie kolumny służą do wykreślenia tzw. **krzywej koncentracji** Lorenza (rys. 4.3.6), do której dodajemy przekątną nazywaną linią równomiernego rozkładu. Gdyby w każdym obiekcie było tyle samo miejsc, to krzywa pokrywałaby się z przekątną. Im większe pole między krzywą i przekątną, tym większa koncentracja.



Rysunek 4.3.6. Krzywa koncentracji miejsc noclegowych w obiektach noclegowych w Bratysławie w 2017 r.

Źródło: opracowanie własne na podstawie tab. 4.3.3.

Interpretacja. Koncentracja miejsc noclegowych w obiektach noclegowych w Bratysławie w 2017 r. była bardzo wysoka (K = 0,97). Ponad 80% obiektów posiadało niespełna 20% miejsc noclegowych, a pozostałe 20% obiektów skupiało ponad 80% miejsc noclegowych.

Podsumowując omówienie metod opisu statystycznego, należy stwierdzić, że pełny opis statystyczny jednej zmiennej powinien zawierać podstawowe miary empiryczne wybierane w zależności od rodzaju cechy, ale także od rozkładu (tab. 4.3.4). Nie ma potrzeby liczenia wszystkich miar z każdej omawianej grupy, lecz aby mieć pełną charakterystykę rozkładu cechy, należy obliczyć zarówno miary średnie, jak i rozproszenia, asymetrii oraz koncentracji. Zakres zastosowań statystyk zależy przede wszystkim od regularności badanego rozkładu (tab. 4.3.4).

Tabela 4.3.4. Podstawowe metody opisu statystycznego i zakres ich zastosowań

Opisowe miary	Klasyczne	Pozycyjne
Zakres zastosowań	rozkłady regularne (miarkowane zróżnicowanie, niewielka asymetria, nieznaczna kurtoza)	rozkłady nieregularne (silnie zróżnicowane, znaczna asymetria, wyraźna kurtoza)
Miary tendencji centralnej	średnia arytmetyczna, średnia harmoniczna	mediana, kwantyle
Miary rozproszenia	wariancja (moment centralny rzędu drugiego), odchylenie standardowe, współczynniki zmienności	obszar zmienności (rozstęp), rozstęp kwartylny, odchylenie ćwiartkowe
Miary asymetrii	moment centralny rzędu trzeciego, współczynnik asymetrii	współczynnik skośności
Miary koncentracji	moment centralny rzędu czwartego, kurtoza, współczynnik Giniego, współczynnik koncentracji Lorenza	×

Źródło: Luszniewicz, Słaby (1996), zmodyfikowane.

Przykład 4.3.4

Wśród uczniów I i III klasy II Liceum Ogólnokształcącego w Łodzi w 2002 r. przeprowadzono badania ankietowe (klasy liczyły po 38 osób). Pytano m.in. o liczbę dni przeznaczonych na podróże podczas wakacji letnich. (Uwaga: jeden uczeń mógł wyjeżdżać kilka razy). Odpowiedzi po obliczeniu 6 statystyk były następujące.

Klasa	Średnia arytmetyczna	Mediana	Dominanta	Minimum	Maksimum	Skośność
I	12,7	14	15	2	21	–0,47
III	21,8	19	14	10	48	0,92

Źródło: obliczenia własne.

Interpretacja. Po przyjrzeniu się obliczonym statystykom widzimy, że uczniowie I klasy jeździli w krótsze podróże podczas wakacji letnich, natomiast klasy III preferowali wyjazdy dłuższe. Przeciętnie młodsi uczniowie wyjeżdżali na 13 dni, a starsi na 22 dni. Zauważono dużą asymetrię, dlatego

posłużono się medianą. Okazało się, że połowa uczniów I klasy spędziła w podróży ponad 2 tygodnie, a połowa mniej niż 2 tygodnie. Natomiast połowa uczniów klasy III przebywała w podróży ponad 19 dni, a połowa mniej niż 19 dni. Rozstęp między minimalną liczbą dni podróży dla uczniów klasy I wynosił 19 dni, a dla III aż 38 dni, co świadczy o dużym zróżnicowaniu długości trwania wyjazdów. Porównanie średniej arytmetycznej i dominanty w obu przypadkach wskazuje na asymetrię: dla klasy I lewostronną, a dla klasy III prawostronną. Siłę i kierunek asymetrii podaje również współczynnik skośności. Dla klasy I wynosi on –0,47 i świadczy o tym, że większość uczniów spędzała w podróży więcej niż 14 dni. Dla klasy III asymetria wynosi 0,92 i mówi o tym, że więcej uczniów tej klasy podróżowało w wakacje letnie 2002 r. krócej niż 3 tygodnie.

4.4. ZADANIA

ZADANIE 4.4.1

Oblicz średnią arytmetyczną i medianę następujących zbiorów liczb:

a) 96, 89, 88, 85, 93, 87, 79, 100, 102,

b) 46, 41, 23, 26, 0, 2, 20, 48, 63, 55,

c) 50, 50, 50, 50, 100, 100, 100, 100,

d) 30, 30, 30, 30, 30, 30, 1000,

e) 16, 14, 18, 12, 10, 17, 20, 18, 19, 14,

f) 10, 12, 13, 14, 15, 15, 16, 16, 16, 17, 17, 19, 20.

Przypomnij własności średniej arytmetycznej i mediany. Czy każda obliczona średnia arytmetyczna dobrze opisuje podane zbiorowości? Którą z miar należy zastosować i interpretować? Jaka jest liczebność zbiorowości?

ZADANIE 4.4.2

Waga plecaków (w kg) na lotnisku Okęcie uczestników wycieczki do Izraela 22 lipca 2000 r. przedstawiała się następująco: 9, 11, 12, 13, 14, 10, 15, 22, 23, 25, 18, 17, 16, 15, 17, 20, 21, 21, 10, 9, 11, 10, 19, 15, 14, 13, 11, 10, 10, 13, 12, 16, 18, 19, 11. Utwórz szereg rozdzielczy o rozpiętości przedzia-

łów co 5 kg. Oblicz przeciętną wagę plecaka. Jaki odsetek osób ma plecaki lżejsze niż 15 kg, a ilu podróżnych musi zapłacić dodatkowo za bagaż, jeśli bez opłaty można wziąć do 20 kg?

ZADANIE 4.4.3

Wydatki na rekreację uczniów pierwszej klasy III Liceum Ogólnokształcącego w Łodzi we wrześniu 2016 r. kształtowały się następująco (w zł): 100, 200, 140, 180, 190, 200, 240, 280, 120, 210, 220, 170, 150, 160, 110, 300, 330, 350, 360, 340, 140, 170, 120, 310, 320, 370, 390, 250, 305, 130. Oblicz, ile przeciętnie uczniowie wydali na rekreację w tym czasie.

ZADANIE 4.4.4

Na podstawie tab. 4.4.1 oblicz, jaka była przeciętna powierzchnia województw w Polsce, ile województw miało powierzchnię większą, a ile mniejszą bądź równą średniej.

Tabela 4.4.1. Powierzchnia województw w Polsce według stanu na 31 grudnia 1999 r.

Województwo	Powierzchnia w km^2
Dolnośląskie	19 948
Kujawsko-pomorskie	17 970
Lubelskie	25 114
Lubuskie	13 984
Łódzkie	18 219
Małopolskie	15 144
Mazowieckie	35 598
Opolskie	9 412
Podkarpackie	17 926
Podlaskie	20 180
Pomorskie	18 293
Śląskie	12 294
Świętokrzyskie	11 672
Warmińsko-mazurskie	24 203
Wielkopolskie	29 826
Zachodniopomorskie	22 902

Źródło: GUS (2000).

ZADANIE 4.4.5

Jaka była przeciętna temperatura w °C w lipcu o 7^{00} rano mierzona w stacji meteorologicznej na lotnisku na Lublinku w Łodzi w 2014 r. (dane umowne): 15, 16, 10, 10, 10, 14, 12, 17, 17, 18, 19, 19, 11, 11, 12, 12, 13, 14, 14, 15, 15, 16, 15, 14, 17, 18, 18, 12, 12, 13, 15.

ZADANIE 4.4.6

Zapytano o wiek osoby, które 25 listopada 1999 r. o godz. 15^{00} w kinie „Polonia" w Łodzi obejrzały film pt. *Pan Tadeusz* w reżyserii Andrzeja Wajdy. Wyniki (tab. 4.4.2) przedstawiono w postaci szeregu rozdzielczego. Jaki był przeciętny wiek widzów? Jaki dominował?

Tabela 4.4.2. Wiek widzów w kinie „Polonia" w Łodzi na seansie o godz. 15^{00} filmu pt. *Pan Tadeusz* w reżyserii Andrzeja Wajdy 25 listopada 1999 r.

Wiek widzów	Liczba osób
10–19	5
20–29	10
30–39	15
40–49	8
50–59	7
60–69	6

Źródło: dane umowne.

ZADANIE 4.4.7

Na stronie internetowej http://wynagrodzenia.pl sprawdź, jak obecnie są wynagradzani pracownicy w turystyce.

- recepcjonista,
- specjalista ds. turystyki,
- pilot wycieczek turystycznych,

Na podstawie mediany i kwartyli porównaj ich zarobki.

ZADANIE 4.4.8

Oblicz średnią gęstość zaludnienia w trzech krajach: Bułgarii, Rumunii i na Węgrzech w 1994 r. (tab. 4.4.3).

Tabela 4.4.3. Ludność i gęstość zaludnienia w Bułgarii, Rumunii i na Węgrzech w 1994 r.

Kraj	Liczba ludności	Gęstość zaludnienia (os./km^2)
Węgry	10 712 000	115
Rumunia	22 201 000	93
Bułgaria	8 862 000	80

Źródło: GUS (1995b).

ZADANIE 4.4.9

Trasę wycieczki autobusowej podzielono na trzy etapy. Pierwszy, o długości 100 km, przejechano z prędkością 80 km/h, drugi, o długości 100 km, z prędkością 45 km/h, a trzeci, o długości 100 km, z prędkością 25 km/h. Oblicz przeciętną prędkość, z jaką jechał autobus. Którą średnią wybierzesz i dlaczego?

ZADANIE 4.4.10

Długość urlopu spędzonego poza miejscem zamieszkania w 2016 r. przez pracowników banku PKO SA wynosiła (w dniach): 10, 12, 15, 10, 26, 23, 21, 15, 3, 23, 8, 9, 10, 5, 21, 16, 15, 13, 12, 17, 19, 5, 3, 24, 29, 12, 18, 27, 6. Oblicz średnią arytmetyczną oraz medianę i kwartyle. Podaj interpretację obliczonych wartości.

ZADANIE 4.4.11

Przeprowadzono badania w 100 obiektach noclegowych w Borach Tucholskich w 2015 r. pod względem liczby miejsc noclegowych (tab. 4.4.4). Jaka liczba miejsc noclegowych powtarzała się najczęściej?

Tabela 4.4.4. Miejsca noclegowe w obiektach wypoczynkowych w Borach Tucholskich w 2000 r.

Liczba miejsc $<x_{id} - x_{ig})$	Liczba obiektów f_i
20–40	3
40–60	16
60–80	13
80–100	7
100–120	5
120–140	3

Tabela 4.4.4 cd.

Liczba miejsc $<x_{id} - x_{ig})$	Liczba obiektów f_i
140–160	1
160–180	1

Źródło: dane umowne.

ZADANIE 4.4.12

Przed egzaminem maturalnym z matematyki przedstawiono poniżej czas nauki w dniach 50 uczniów liceum ogólnokształcącego: 57, 65, 61, 55, 42, 54, 36, 51, 32, 60, 57, 47, 58, 47, 61, 60, 54, 61, 28, 60, 42, 43, 61, 35, 73, 46, 32, 47, 51, 53, 27, 61, 49, 36, 29, 28, 55, 26, 49, 83, 65, 29, 74, 61, 36, 52, 42, 32, 57, 67. Zbuduj szereg rozdzielczy o rozpiętości przedziałów co 5 dni, rozpoczynając od 25 dni. Wyznacz medianę oraz dominantę szeregu. Jaki odsetek uczniów uczył się mniej niż 30 dni?

ZADANIE 4.4.13

Krzyś uzyskał następujące oceny z angielskiego: 1, 4, 6, 2, 3, 4, 5, 5, 3, 1, 5. Oblicz średnią arytmetyczną, medianę oraz dominantę dla tych danych. Którą z tych miar wybierze Krzyś, opowiadając rodzicom o swoich wynikach w nauce?

ZADANIE 4.4.14

Nie znając szczegółowych danych rozkładu, a jedynie statystyki opisowe (średnia arytmetyczna, mediana i dominanta), przedstaw graficznie kształt rozkładów prezentujących długość podróży wakacyjnych (w km) czterech badanych grup.

	$\bar{x}$	M_x	D_x
Rozkład I	300	300	300
Rozkład II	600	400	200
Rozkład III	400	400	200 i 700
Rozkład IV	100	500	800

Źródło: opracowanie własne.

ZADANIE 4.4.15

Oblicz następujące statystyki: minimum, maksimum, medianę oraz kwartyl pierwszy i trzeci dla danych z lat 2004, 2009 i 2014, a następnie narysuj trzy wykresy „pudełko z wąsami" i przeprowadź analizę statystyczną.

Tabela 4.4.5. Obiekty hotelowe w Polsce według województw w latach 2004, 2009 i 2014

Województwo	Rok		
	2004	2009	2014
Łódzkie	119	139	206
Mazowieckie	185	217	297
Małopolskie	257	363	449
Śląskie	171	232	304
Lubelskie	93	97	149
Podkarpackie	102	158	197
Podlaskie	33	51	87
Świętokrzyskie	54	84	128
Lubuskie	112	137	127
Wielkopolskie	199	282	344
Zachodniopomorskie	133	189	246
Dolnośląskie	239	309	381
Opolskie	45	49	71
Kujawsko-pomorskie	108	131	149
Pomorskie	154	248	314
Warmińsko-mazurskie	135	150	197

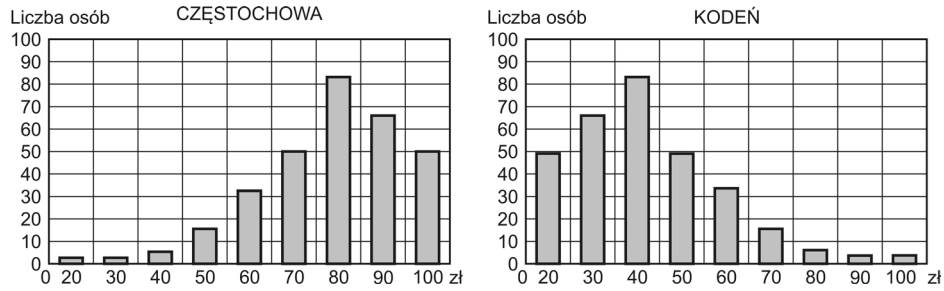
Źródło: https://stat.gov.pl (dostęp: 24.09.2015).

ZADANIE 4.4.16

Skorzystaj z danych na temat turystyki publikowanych przez GUS na stronie internetowej https://stat.gov.pl. Wybierz dane dotyczące turystycznych obiektów noclegowych według województw (jednostki terytorialne), a w nich: obiekty ogółem, obiekty zbiorowego zakwaterowania – kempingi i pola namiotowe dla trzech przedziałów czasowych (rok 2004, 2010 i najnowsze). Oblicz następujące statystyki: minimum, maksimum, medianę oraz kwartyl pierwszy i trzeci, a następnie narysuj trzy wykresy „pudełko z wąsami" i przeprowadź analizę statystyczną.

ZADANIE 4.4.17

Na podstawie poniższych histogramów oceń strukturę dziennych wydatków pielgrzymów do Częstochowy i Kodenia w 2016 r.



Rysunek 4.4.1. Rozkład dziennych wydatków pielgrzymów do Częstochowy i Kodenia w 2016 r.

Źródło: opracowanie własne na podstawie danych umownych.

ZADANIE 4.4.18

Na podstawie analizy statystycznej (tab. 4.4.6) oceń zyski spółek „Atos" i „Portos" działających w Kaliszu w 2013 r.

Tabela 4.4.6. Zyski spółek „Atos" i „Portos" w Kaliszu w 2013 r. według miesięcy

Zyski w tys. zł	„Atos"	„Portos"
0–9,99	1	1
10–19,99	2	1
20–29,99	3	2
30–39,99	3	2
40–49,99	2	3
50–59,99	1	3

Źródło: dane umowne.

ZADANIE 4.4.19

W czasie zawodów krajoznawczo-sportowych w Ojcowskim Parku Narodowym w 2000 r. zawodnicy mieli do pokonania trasę 500 km w ciągu 7 dni. W zawodach brało udział 26 osób. Po upływie trzech dni Sylwia zapisała rezultaty zawodników. Wyniki były następujące (w km): 85, 90, 100, 110, 120, 165, 210, 180, 140, 130, 168, 150, 150, 147, 163, 170, 180, 130, 174, 200, 110, 160,

120, 142, 150, 190. Przedstaw dane w postaci szeregu rozdzielczego o rozpiętości 10 km, rozpoczynając od 80 km. Przeprowadź obliczenia i analizę statystyczną dla danych szeregu szczegółowego i rozdzielczego. Porównaj wyniki.

ZADANIE 4.4.20

W hotelu „Ibis" w Grenoble w 2012 r. przebywała następująca liczba turystów zagranicznych:

- w lutym: 100, 120, 115, 116, 117, 115, 118, 121, 121, 117, 102, 105, 119, 110, 116, 124, 108, 112, 113, 106, 116, 111,
- w maju: 40, 32, 30, 31, 43, 42, 40, 44, 43, 40, 39, 38, 44, 45, 35, 44, 46, 38, 41, 47, 35, 49.

Przedstaw dane w postaci szeregu i histogramu. Przeprowadź wszechstronną analizę statystyczną.

ZADANIE 4.4.21

Zbadano ruch pasażerski PKP na trasie Łódź–Poznań w drugi weekend października 2010 r. Szczególną uwagę zwracano na wiek pasażerów, którzy korzystali z tej linii. Uzyskane dane przedstawia szereg w tab. 4.4.7.

Tabela 4.4.7. Pasażerowie PKP na trasie Łódź–Poznań w drugi weekend października 2000 r.

Wiek pasażerów	Liczba osób
15–19	18
20–24	137
25–29	58
30–34	39
35–39	18
40–44	20
45–49	8
50–54	42
55–59	12
60–64	37
65–69	11

Źródło: dane umowne.

Oblicz statystyki opisowe i przeprowadź analizę statystyczną.

ZADANIE 4.4.22

Szeregi przedstawiają długość pobytu turystów z Niemiec i Rosji w Polsce w 1993 r. (tab. 4.4.8). Porównaj przeciętny czas spędzony w Polsce dla obu zbiorowości, wykreśl histogramy, zbadaj średnią, zróżnicowanie, asymetrię i koncentrację.

Tabela 4.4.8. Turyści z Niemiec i Rosji w Polsce w 1993 r. według długości pobytu

Liczba dni	Liczba turystów w tys	
	z Niemiec	z Rosji
1–4	1,20	1,30
5–8	4,20	2,50
9–12	1,78	6,70
13–16	0,49	0,94
17–20	0,33	0,56

Źródło: dane umowne.

ZADANIE 4.4.23

Na podstawie danych dotyczących liczby turystów według województw w Polsce (GUS) oblicz współczynnik koncentracji Lorenza i sporządź wykres. Przeprowadź analizę.

ZADANIE 4.4.24

Oceń rozproszenie, asymetrię i koncentrację danych z zadania 4.4.15.

ZADANIE 4.4.25

Oceń rozproszenie, asymetrię i koncentrację danych z zadania 4.4.16.

4.5. ODPOWIEDZI DO WYBRANYCH ZADAŃ

ZADANIE 4.4.1

Zbiór	N	$\bar{x}$	M_x	Uwagi
a)	9	91	89	
b)	10	32,4	33,5	
c)	8	75	75	W tym przypadku trzeba wskazać, że zbiorowość składa się z dwóch podgrup.
d)	7	188,3	30	Należy zwrócić uwagę na występowanie wartości skrajnej, która zniekształca wynik średniej arytmetycznej.
e)	10	15,8	16,5	
f)	13	14,3	16	

ZADANIE 4.4.2

Tabela 4.4.9. Waga plecaków (w kg) uczestników wycieczki do Izraela 22 lipca 2000 r. na lotnisku Okęcie

Waga (kg)	Liczba osób	Udział (%)	Wartości skumulowane	
	f_i		f_{ic}	$\% f_{ic}$
5–9,99	8	22,9	8	22,9
10–14,99	13	37,1	21	60,0
15–19,99	9	25,7	30	85,7
20–24,99	5	14,3	35	100,0
×	35	100,0	×	×

Źródło: dane z zadania 4.4.2.

Przeciętna waga plecaka wynosiła 14,86 kg; 60% osób ma bagaż ważący 15 kg lub mniej, a 5 osób, tj. 14,3% uczestników wycieczki, musi zapłacić dodatkową opłatę za nadbagaż.

ZADANIE 4.4.3

Uczniowie wydali przeciętnie 228,5 zł.

ZADANIE 4.4.4

Średnia powierzchnia województw w Polsce w 1999 r. wynosiła 19 542,81 km^2. Większą powierzchnię od przeciętnej miało 7 województw, a mniejszą 9 województw.

ZADANIE 4.4.8

Liczymy średnią harmoniczną. Przeciętna gęstość zaludnienia wynosi 94,4 os./km^2.

ZADANIE 4.4.18

Tabela 4.4.10. Statystyki opisowe dla zysków spółek „Atos" i „Portos" w Kaliszu w 2013 r. według miesięcy

Spółka	$\bar{x}$	σ	V_x	A	K
„Atos"	29,58	13,14	44,4	–0,18	2,08
„Portos"	35,42	14,35	40,53	–0,76	2,38

Źródło: opracowanie własne na podstawie tab. 4.4.6.

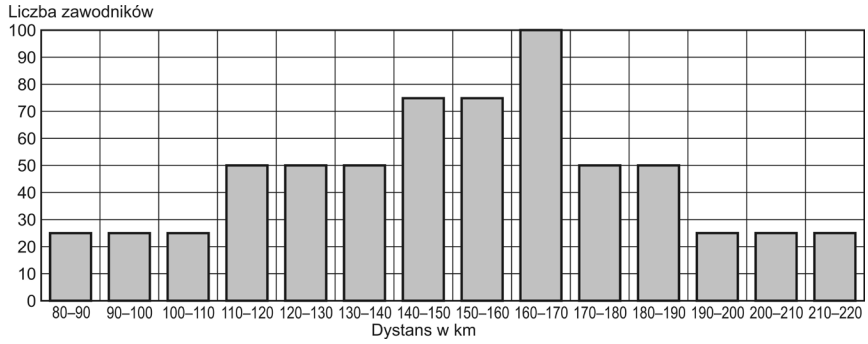
ZADANIE 4.4.19

Tabela 4.4.11. Wyniki zawodników uczestniczących w zawodach krajoznawczo-sportowych w Ojcowskim Parku Narodowym w 2000 r. po upływie trzech dni

Pokonany dystans w km	Liczba zawodników
80–89,99	1
90–99,99	1
100–109,99	1
110–119,99	2
120–129,99	2
130–139,99	2
140–149,99	3
150–159,99	3
160–169,99	4

Pokonany dystans w km	Liczba zawodników
170–179,99	2
180–189,99	2
190–199,99	1
200–209,99	1
210–219,99	1

Źródło: opracowanie własne.



Rysunek 4.4.2. Rozkład wyników w zawodach krajoznawczo-sportowych w Ojcowskim Parku Narodowym w 2000 r. po upływie trzech dni

Źródło: opracowanie własne na podstawie tab. 4.4.11.

Tabela 4.4.12. Statystyki opisowe dla wyników w zawodach krajoznawczo-sportowych w Ojcowskim Parku Narodowym w 2000 r. dla szeregu rozdzielczego i szczegółowego

Typ szeregu	Statystyki						
	rozstęp	$\bar{x}$	σ	M_3	A_s	M_4	K
Rozdzielczy	140	151,20	32,47	–3391,8	–0,099	2 702 470,5	2,43
Szczegółowy	105	147,46	32,48	–4333,2	–0,130	2 538 669,0	2,28

Źródło: opracowanie własne na podstawie zadania 4.4.19 i tab. 4.4.11.

ZADANIE 4.4.20

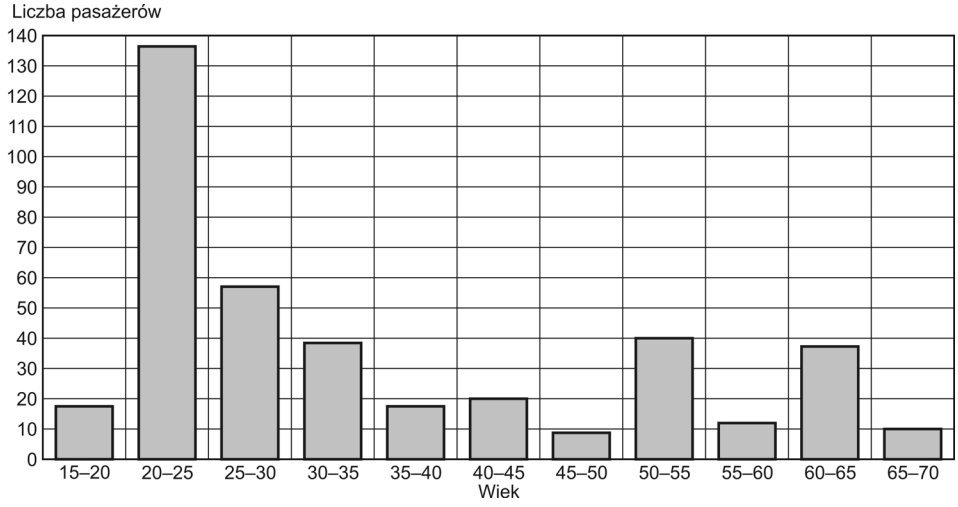
- luty: *śr* = 113,7 os., *m* = 115,5 os., *Q*1 = 110,25 os., *Q*3 = 117,75 os., σ = 6,4 os., *V* = 5,6%, *A* = –0,55; *K* = 2,31;
- maj: *śr* = 40,1 os., *m* = 40,5 os., *Q*1= 38 os., *Q*3 = 44 os., σ = 5,2 os., *V* = 12,9%, *A* = –0,08, *K* = 2,1.

ZADANIE 4.4.21

Tabela 4.4.13. Statystyki opisowe dla wieku pasażerów PKP na trasie Łódź–Poznań w drugi weekend października 2010 r.

Statystyka	Wartość statystyki	Jednostka miary
N – liczebność	400	osoby
Średnia arytmetyczna	35,3	lata
Dominanta	23,0	lata
Rozstęp	55	lata
Odchylenie przeciętne	13,2	lata
Wariancja – moment centralny rzędu drugiego	228,54	–
Odchylenie standardowe	15,12	lata
Współczynniki zmienności	37,0 i 42,8	%
Moment centralny rzędu trzeciego	2663,09	–
Współczynnik asymetrii	0,77	lata
Moment centralny rzędu czwartego	111 636,59	–
Kurtoza	2,14	lata

Źródło: opracowanie własne na podstawie tab. 4.4.7.



Rysunek 4.4.3. Rozkład wieku pasażerów PKP na trasie Łódź–Poznań w drugi weekend października 2000 r.

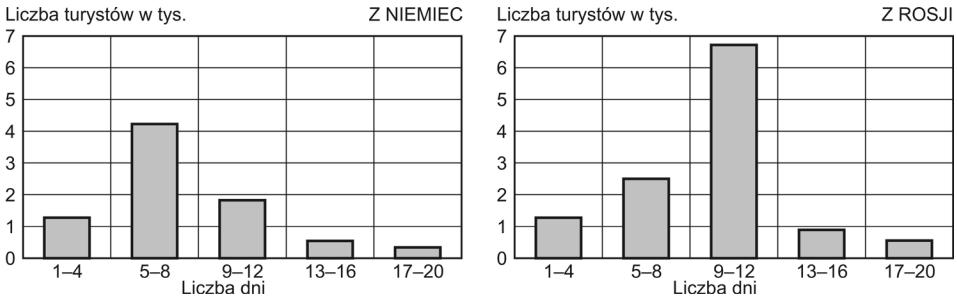
Źródło: opracowanie własne na podstawie tab. 4.4.7.

ZADANIE 4.4.22

Tabela 4.4.14. Turyści z Niemiec w Polsce w 1993 r. według długości pobytu

Liczba dni	Liczba turystów w tys.	x_i'	$x_i'f_i$	$(x_i' - \bar{x})^2 \cdot f_i$	$(x_i' - \bar{x})^3 \cdot f_i$	$(x_i' - \bar{x})^4 \cdot f_i$
1–4	1,20	2,5	3,00	33,39	–176,14	929,12
5–8	4,20	6,5	27,30	6,83	–8,71	11,10
9–12	1,78	10,5	18,69	13,22	36,02	98,15
13–16	0,49	14,5	7,11	22,16	149,03	1 002,23
17–20	0,33	18,5	6,11	37,96	407,10	4 366,20
Suma	8,00	×	62,20	113,56	407,31	6406,79

Źródło: opracowanie własne.



A B

Rysunek 4.4.4. Rozkład długości pobytu turystów z Niemiec (A) i Rosji (B) w Polsce w 1993 r.

Źródło: opracowanie własne na podstawie tab. 4.4.8.

Tabela 4.4.15. Statystyki opisowe dla długości pobytu turystów z Niemiec (A) i Rosji (B) w Polsce w 1993 r.

Narodowość turystów	Statystyki					
	$\bar{x}$	σ	M_3	A_s	M_4	K
Niemcy	7,08	3,77	50,91	0,95	800,85	3,97
Rosjanie	9,49	3,67	2,13	0,05	632,78	3,48

Źródło: opracowanie własne na podstawie tab. 4.4.8.

ROZDZIAŁ 5

ANALIZA KORELACJI I REGRESJI

Często można usłyszeć, że np. wysokość stypendium naukowego zależy od średniej ocen uzyskanych w roku poprzednim, waga człowieka zależy od liczby spożywanych kalorii, zużycie paliwa zależy od pojemności silnika, opinia o usługach zależy od płci lub narodowości respondentów. Do poszukiwania związków w zakresie współzależności między dwiema lub większą liczbą zmiennych wykorzystuje się analizę korelacji, a do przewidywania wartości jednej zmiennej na podstawie wartości drugiej zmiennej stosuje się metody regresji.

Jeśli zostanie wskazany związek statystyczny między dwiema zmiennymi *x* i *y*, czyli że wraz ze zmianą wielkości zmiennej *x* zmienia się wielkość zmiennej *y*, to można mówić o **współzależności** *x* i *y*. Kolejnym etapem badań jest określenie siły i kierunku związku między *x* i *y*, tzn. wskazanie współzależności zmiennych. Miary zależności nazywa się **współczynnikami korelacji**. Za pomocą **metod regresji** możliwe jest na podstawie wartości jednej zmiennej przewidywanie (predykcja) wartości drugiej zmiennej. W badaniach obejmujących zjawiska turystyczne nieczęsto możemy spotkać się z zależnością funkcyjną, gdyż zazwyczaj na wielkość zjawiska wpływa wiele czynników (np. pogoda, moda, ekonomia, polityka). Wówczas na podstawie poczynionych obserwacji dla jednej cechy poszukujemy przybliżonej wartości drugiej

cechy. Jedną z cech, której wartości zależą od drugiej, nazywamy **zmienną zależną** *y* (objaśnianą), a drugą **zmienną niezależną** *x* (objaśniającą). Zdarza się, że obie zmienne wpływają na siebie i wtedy obojętne jest, którą z nich oznaczymy jako zależną. Na przykład, jeśli planowane są badania zależności spożycia lodów od temperatury powietrza, to zmienną niezależną *x* będzie temperatura, a zależną *y* – wielkość sprzedaży. W rozdziale zaprezentowane zostaną wybrane metody korelacji i regresji dla dwóch cech.

5.1. ANALIZA KORELACJI

Przystępując do poszukiwania współzależności między cechami statystycznymi, należy na wstępie określić, w jakiej skali (nominalnej, porządkowej czy ilorazowej) są one przedstawione. Jest to niezwykle ważny moment, gdyż wybór odpowiedniej metody statystycznej do poszukiwania korelacji jest uzależniony od skali pomiarowej. Nie można swobodnie korzystać z zaprezentowanych metod, a w szczególności z procedur, które oferują pakiety statystyczne, zanim nie upewnimy się co do tego, z danymi w jakiej skali pomiarowej mamy do czynienia. Z tego powodu przedstawione zostaną wybrane metody korelacyjne dla każdego typu skali pomiarowej.

Podczas analizy korelacji należy zawsze pamiętać o tym, że znalezienie pozytywnej odpowiedzi na postawione pytania badawcze nie powinno być uważane za pewnik. Wyniku analizy korelacyjnej nie można traktować w kategoriach przyczynowo-skutkowych, gdyż możemy mieć do czynienia z **korelacją pozorną**[1].

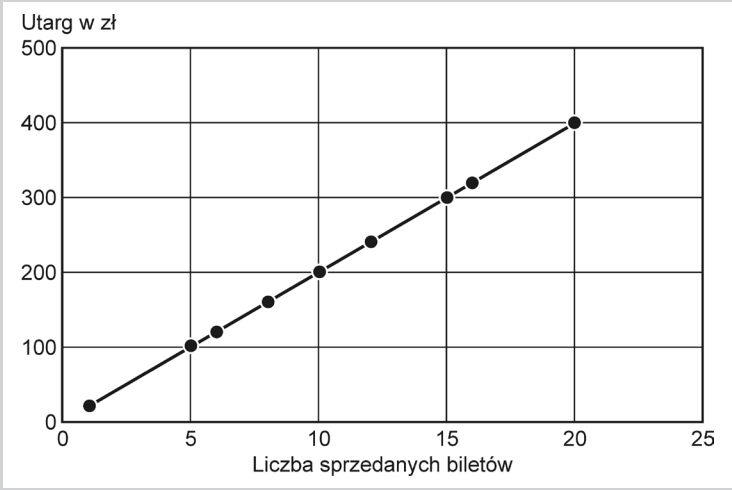
Zależność między zmiennymi reprezentującymi zjawiska bardzo rzadko ma charakter funkcyjny (przykład 5.1.1). Wówczas wraz ze zmianą jednej zmiennej (argumentów funkcji) następuje zmiana drugiej zmiennej (wartości funkcji) i zależność ta jest opisana za pomocą wzoru matematycznego: $y = f(x)$.

---

1 Może się zdarzyć, że dostaniemy niezwykle wysoką istotność wyników między dwoma zbiorami, które nie mają ze sobą nic wspólnego, np. między wielkością ruchu turystycznego w Egipcie a liczbą pingwinów na Antarktydzie w tych samych latach.

Przykład 5.1.1

Zależność między liczbą sprzedanych biletów w sztukach (*x*) i kwotą, jaka powinna być w kasie, czyli utargiem (*y*), ma charakter funkcyjny wyrażony wzorem: $y = 20x$ (gdzie 20 jest ceną biletu w zł) lub za pomocą wykresu (rys. 5.1.1).



Rysunek 5.1.1. Zależność między liczbą sprzedanych biletów w sztukach (*x*) i utargiem (*y*)

Źródło: dane umowne.

Poszukiwania współzależności między zmiennymi można przeprowadzić za pomocą:

a) obserwacji szeregu,

b) tablic korelacyjnych,

c) metody graficznej – wykresy i mapy,

d) wyliczenia współczynnika korelacji[2].

2 Obliczenie współczynnika korelacji, determinacji i zbadanie istotności będzie prowadziło do określenia współzależności. Wcześniejsze metody będą mówiły jedynie o współzmienności.

Każde z tych narzędzi daje mniej lub bardziej precyzyjne określenie związku między zmiennymi, ale należy pamiętać, że istnienie związku nie oznacza przyczynowości.

Najprostszą procedurą poszukiwania **współzmienności** między zmiennymi jest **obserwacja szeregu statystycznego**, polegająca na zestawieniu obok siebie wartości szeregów statystycznych szczegółowych w postaci macierzy A_{nx2}, gdzie *n* oznacza liczbę obserwacji, a 2 jest liczbą cech (przykład 5.1.2). Jeden z nich porządkujemy rosnąco lub malejąco i sprawdzamy, jak zachowują się wartości w drugim szeregu. Gdy szereg nie jest zbyt liczny, można zauważyć, że wraz ze wzrostem wartości *x* rosną wartości *y* (jest to **korelacja dodatnia**) lub wraz ze wzrostem wartości *x* maleją wartości *y* (jest to **korelacja ujemna**). Nie widząc żadnej zależności między zmiennymi, uznajemy, że korelacja nie występuje. Zdarza się, że początkowo między zmiennymi istnieje korelacja dodania, która następnie zmienia się w korelację ujemną. Można wtedy przypuszczać, że mamy do czynienia z **korelacją krzywoliniową**[3]. Metoda obserwacji szeregu nie jest warta polecenia, w przypadku gdy dysponujemy dużym zbiorem jednostek statystycznych – wówczas bowiem nie jesteśmy w stanie poprawnie wnioskować o współzmienności.

Przykład 5.1.2

Przeprowadzono badania sondażowe wśród polskich turystów, którzy byli na wycieczce w Barcelonie (weekend 12–14 lipca 2013 r.) i oczekiwali na lotnisku na lot do Polski. Zadano im kilka krótkich pytań – odpowiedzi znajdują się w tab. 5.1.1. Zmienne prezentują różne typy skali pomiarowej: wiek i wydatki – ilorazową, zadowolenie z wycieczki – porządkową, a płeć – nominalną.

Na podstawie dwóch zmiennych: wieku i wydatków (euro) poniesionych na noclegi poszukiwano odpowiedzi na pytanie: Czy istnieje zależność między wiekiem i kwotą przeznaczoną na noclegi przez turystów?

3 Nie będzie omawiana w podręczniku.

Tabela 5.1.1. Odpowiedzi udzielone przez turystów wracających z wycieczki do Barcelony (12–14 lipca 2013 r.)

Imię respondenta	Wiek	Wydatki (w euro)	Płeć	Zadowolenie z wycieczki*
	x_i	y_i	z_i	w_i
Hanna	22	48	k	8
Jan	25	48	m	9
Milena	26	58	k	7
Kamil	26	58	m	10
Jakub	26	51	m	6,5
Iwo	29	66	m	9,5
Wioleta	30	66	k	8,5
Stanisław	35	50	m	7,5
Jerzy	38	77	m	5
Krystyna	40	100	k	4
Adam	44	78	m	3,5
Ewa	49	130	k	5,5
Iwona	51	150	k	1
Ewa	56	145	k	3
Leon	59	145	m	2
Maryla	59	150	k	3

* W skali od 1 do 10 (1 – ocena najniższa, 10 – ocena najwyższa).

Źródło: dane umowne.

Interpretacja. Pierwsza prosta obserwacja szeregu (tab. 5.1.1) pozwala zauważyć, że im starsi respondenci, tym więcej wydawali na noclegi. Jednak nic nie możemy powiedzieć o sile oraz istotności statystycznej związku. Nasze spostrzeżenia są jednak na tyle wartościowe, że powinno się przejść do dalszych etapów analizy korelacyjnej.

Kolejnym sposobem analizy współzależności jest konstrukcja tablicy korelacyjnej i jej interpretacja. **Tablica korelacyjna** składa się z dwóch szeregów rozdzielczych, z których jeden znajduje się w kolumnach, a drugi w wierszach. W nagłówkach wierszy i kolumn wpisujemy granice przedziałów klasowych

lub wartość cechy (przykład 5.1.3). W poszczególnych komórkach tablicy zamieszczamy liczebność klas jednej zmiennej odpowiadającą drugiej zmiennej. Im więcej jest wypełnionych komórek wzdłuż przekątnej, tym większa jest zależność między zmiennymi. Prawidłowo wypełniona tablica korelacyjna może posłużyć do obliczenia współczynnika korelacji (Zając 1988; Runge 2007).

Przykład 5.1.3

Tablica korelacji dla danych z przykładu 5.1.2 powstała przez umieszczenie w pierwszej kolumnie szeregu rozdzielczego zmiennej niezależnej wieku turystów (x_i), a w pierwszym wierszu szeregu rozdzielczego zmiennej zależnej prezentującej wydatki (y_i). W polu tablicy znajduje się liczba osób.

Tabela 5.1.2. Tablica korelacji wieku oraz wydatków na noclegi turystów wracających z wycieczki do Barcelony (12–14 lipca 2013 r.)

Wiek turystów (x_i)	Wydatki (y_i) (w euro)		
	<0–50)	<50–100)	<100–150)
<20–30)	2	4	–
<30–40)	–	3	–
<40–50)	–	1	2
<50–60)	–	–	4

Źródło: opracowanie własne na podstawie tab. 5.1.1

Interpretacja. Cechy układają się wzdłuż przekątnej i wskazują, że wraz ze wzrostem wieku respondentów wzrasta kwota wydana przez turystów na noclegi. Starsi turyści wybierali prawdopodobnie obiekty o wyższym standardzie. Przypuszczamy, że zależność ta jest znaczna, ale nie możemy jeszcze określić jej siły. Posłużą do tego współczynniki korelacji.

Dla danych w skali nominalnej i porządkowej można stosować tablice kontyngencji. Zazwyczaj umieszczamy w lewym boku zmienną niezależną *x*, a w główce tablicy zmienną zależną *y*. Jej interpretacja polega na obserwacji rozkładu zmiennych i poszukiwaniu prawidłowości w ich ułożeniu, np. wzdłuż przekątnej (przykład 5.1.4). Jeśli zmienna *x* ma *p* odmian, a zmienna *y* ma *q* odmian, to tablica ma *p* wierszy i *q* kolumn.

Przykład 5.1.4

W trzech nadmorskich miejscowościach turystycznych: Sopocie (tab. 5.1.3), Władysławowie (tab. 5.1.4) i Helu (tab. 5.1.5), zapytano 150 turystów o cel przyjazdu (skala nominalna) oraz o wykształcenie (skala porządkowa). Za pomocą tablicy kontyngencji można sprawdzić, czy istnieje współzmienność między tymi cechami.

Tabela 5.1.3. Wykształcenie a cel podróży turystów w Sopocie w 2015 r. (idealna współzmienność)

Wykształcenie	Cel podróży			Suma
	odwiedziny u krewnych	służbowy	turystyczno-wypoczynkowy	
Wyższe	0	0	50	50
Średnie	0	50	0	50
Zawodowe	50	0	0	50
Suma	50	50	50	150

Źródło: dane umowne.

Tabela 5.1.4. Wykształcenie a cel podróży turystów we Władysławowie w 2015 r. (umiarkowana współzmienność)

Wykształcenie	Cel podróży			Suma
	odwiedziny u krewnych	służbowy	turystyczno-wypoczynkowy	
Wyższe	0	10	40	50
Średnie	10	30	10	50
Zawodowe	40	10	0	50
Suma	50	50	50	150

Źródło: dane umowne.

Tabela 5.1.5. Wykształcenie a cel podróży turystów w Helu w 2015 r. (minimalna współzmienność lub jej brak)

Wykształcenie	Cel podróży			Suma
	odwiedziny u krewnych	służbowy	turystyczno-wypoczynkowy	
Wyższe	15	15	20	50
Średnie	15	20	15	50
Zawodowe	20	15	15	50
Suma	50	50	50	150

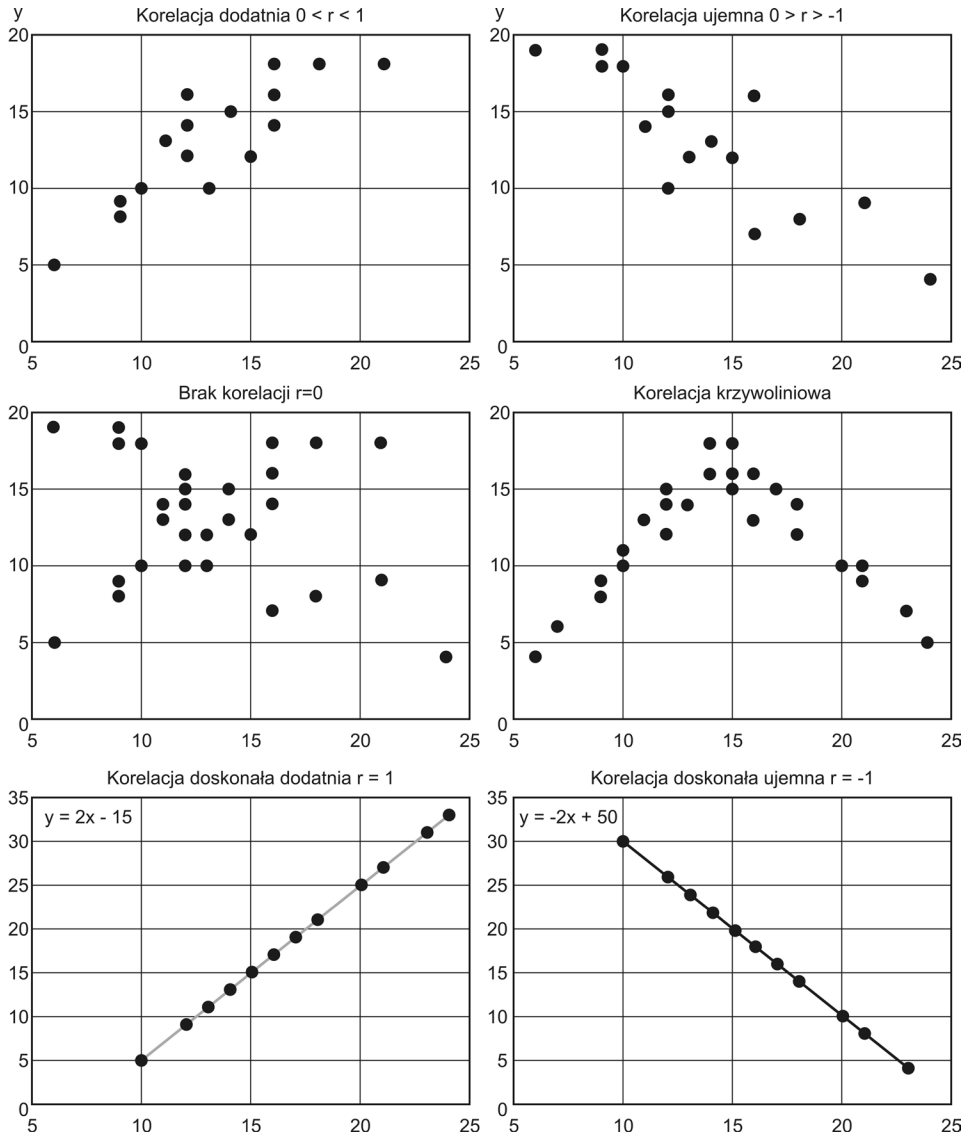
Źródło: dane umowne.

Interpretacja. Wyniki poszukiwań współzależności między wykształceniem i celem przyjazdu do Sopotu, Władysławowa i Helu dały odmienne rezultaty dla każdego z miast. Analiza cech w przypadku badań turystów w Sopocie (tab. 5.1.3) ukazała idealną współzmienność między cechami – każdy rodzaj wykształcenia turysty odnosił się do innego celu przyjazdu do Sopotu. Wyniki są tak idealne, że aż nieprawdopodobne (można zaproponować powtórzenie badań). Analiza cech w przypadku badań turystów we Władysławowie (tab. 5.1.4) sugeruje umiarkowaną współzmienność między cechami. Można powiedzieć, że turyści z wykształceniem wyższym w większości wybrali się do tego miasta w celach turystyczno-wypoczynkowych, z kolei turyści ze średnim wykształceniem byli tam służbowo, a z wykształceniem zawodowym przyjechali na weekend w odwiedziny do krewnych. Badania turystów w mieście Hel (tab. 5.1.5) wykazały brak współzależności między cechami.

METODA GRAFICZNA

W zależności od skali pomiarowej zmiennych wykorzystywana jest też odpowiednia metoda graficzna. Metoda graficzna dla danych w skali porządkowej i wyższej polega na nanoszeniu na układ współrzędnych prostokątnych każdej obserwacji. Wykres ten nazywany jest **wykresem rozrzutu** (rys. 5.1.2). Wartość x odpowiada zmiennej niezależnej, a y – zmiennej zależnej. Otrzymujemy chmurę punktów o współrzędnych x_i, y_i, która objaśnia zależność między zmiennymi. Na podstawie układu punktów można wnioskować o kie-

runku zależności i szacować, jak silny będzie współczynnik korelacji (nie zastępuje to konieczności jego obliczenia).

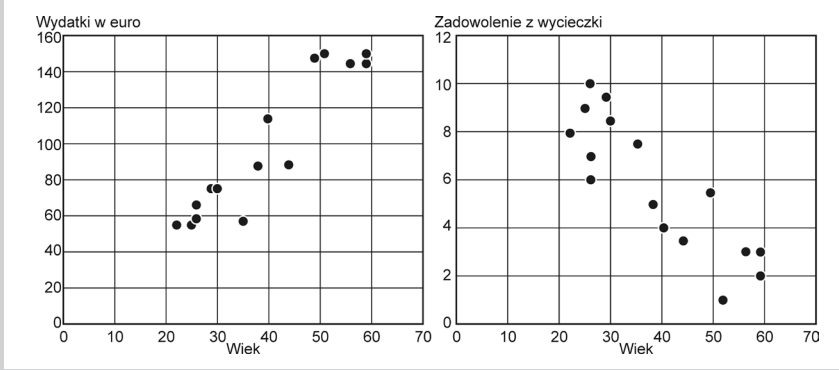


Rysunek 5.1.2. Graficzne przedstawienie zależności (dane w skali porządkowej i wyższej)

Źródło: opracowanie własne.

Przykład 5.1.5

Graficzna prezentacja danych (rys. 5.1.3) z przykładu 5.1.2 powstała przez umieszczenie na osi *X* wartości wieku turystów (x_i), a na osi *Y* wartości zmiennej reprezentującej wydatki (y_i) (A) oraz zadowolenie z wycieczki (B).



A B

Rysunek 5.1.3. Współzależność między wiekiem i wielkością wydatków (A) oraz zadowolenia z wycieczki (B) turystów wracających z Barcelony (12–14 lipca 2013 r.)

Źródło: opracowanie własne.

Interpretacja. Na podstawie wykresów można wnioskować, że zachodzi współzależność między wiekiem badanych i ich wydatkami podczas podróży oraz zadowoleniem z wycieczki. W pierwszym przypadku mamy do czynienia z korelacją dodatnią – możemy przypuszczać, że uczestnik był starszy, tym więcej wydawał. Punkty na drugim wykresie układają się inaczej – występuje korelacja ujemna, wobec tego możemy przypuszczać, że im uczestnik był starszy, tym mniej był zadowolony z wycieczki.

Jeśli przynajmniej jedna cecha jest w skali nominalnej, np. płeć, narodowość, to do prezentacji graficznej współzależności warto skorzystać z wykresu słupkowego skumulowanego procentowego.

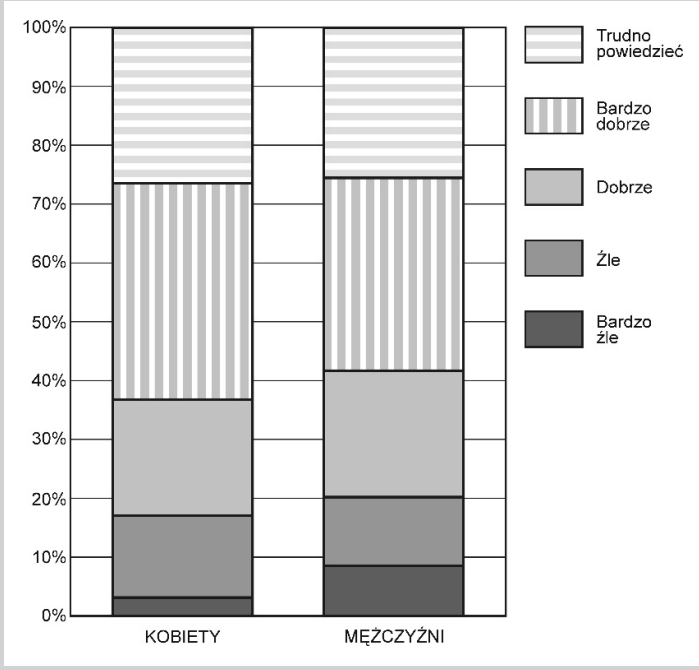
Przykład 5.1.6

Graficzna prezentacja zależności z tab. 5.1.6. Jedna z cech (płeć) jest w skali nominalnej.

Tabela 5.1.6. Opinia turystów korzystających z biura podróży „Odys" w 2015 r. na temat pracy przewodnika według płci respondentów

Płeć	Pracę przewodnika oceniono					
	bardzo źle	źle	dobrze	bardzo dobrze	trudno powiedzieć	Suma
Kobiety	5	20	30	55	40	150
Mężczyźni	14	19	35	54	42	164
Suma	19	39	65	109	82	314

Źródło: dane umowne.



Rysunek 5.1.4. Opinia turystów korzystających z biura podróży „Odys" w 2015 r. na temat pracy przewodnika według płci respondentów

Źródło: opracowanie własne na podstawie tab. 5.1.6.

Interpretacja. Cechy w skali nominalnej zostały przedstawione na osi *X*. Zależność między oceną pracy przewodnika a płcią respondentów korzystających z biura podróży „Odys" w 2015 r. występuje, ale jest nieduża. Kobiety nieco lepiej oceniały pracę przewodnika niż mężczyźni. Warto zwrócić uwagę na to, że co czwarta ankietowana osoba nie potrafiła dokonać oceny i wybrała odpowiedź: „trudno powiedzieć".

Wykres może być sporządzony dla cech w skali porządkowej. Należy pamiętać, aby cechy w skali porządkowej miały odpowiednią kolejność, zgodną z porządkiem cech (przykład 5.1.7).

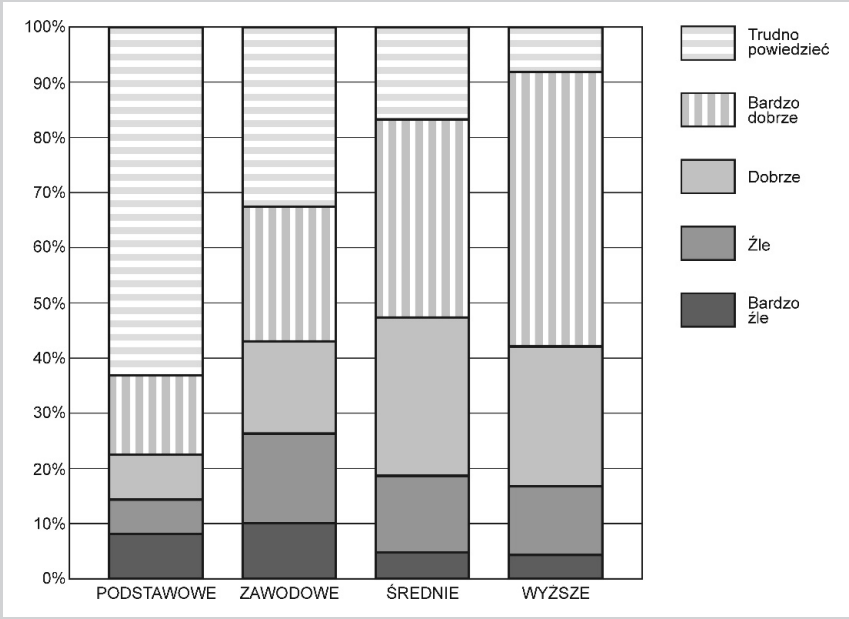
Przykład 5.1.7

Przedstawimy graficznie współzmienność między wykształceniem turystów i zadowoleniem z usług przewodnika.

Tabela 5.1.7. Opinia turystów korzystających z biura podróży „Odys" w 2015 r. na temat pracy przewodnika według wykształcenia respondentów

Wykształcenie	Pracę przewodnika oceniono					
	bardzo źle	źle	dobrze	bardzo dobrze	trudno powiedzieć	suma
Podstawowe	5	4	5	9	40	64
Zawodowe	6	10	10	15	20	61
Średnie	3	10	20	25	12	70
Wyższe	5	15	30	60	10	120
Suma	19	39	65	109	82	314

Źródło: dane umowne.



Rysunek 5.1.5. Opinia turystów korzystających z biura podróży „Odys" w 2015 r. na temat pracy przewodnika według wykształcenia respondentów

Źródło: opracowanie własne na podstawie tab. 5.1.7.

Interpretacja. Cechy w skali porządkowej zostały przedstawione na osi *X* według ich rangi od najniższego do najwyższego wykształcenia. Zależność między oceną pracy przewodnika i wykształceniem respondentów korzystających z biura podróży „Odys" w 2015 r. występuje i jest znaczna. Lepiej wykształceni oceniali wyżej pracę przewodnika niż osoby z niższym wykształceniem. 60% osób z najniższym wykształceniem (podstawowym) nie potrafiło dokonać oceny i wybrało odpowiedź: „trudno powiedzieć".

WSPÓŁCZYNNIKI KORELACJI

W badaniach statystycznych stosuje się różne sposoby pomiaru współzależności dwóch cech. Najczęściej stosowaną miarą, która pozwala określić kierunek i siłę związku między zjawiskami (cechami) są współczynniki korelacji (r_{xy}). W zależności od skali pomiarowej badanych zmiennych możemy wybrać: współczynnik korelacji liniowej Pearsona dla cech w skali ilorazowej lub interwałowej, współczynnik korelacji rang Spearmana dla cech w skali ilorazo-

wej oraz porządkowej, współczynnik korelacji Younga dla cech jakościowych dychotomicznych (możliwe tylko dwie odpowiedzi). Współczynnik Goodmana i Kruskala obliczany jest dla zmiennych w skali porządkowej, w których występują więcej niż dwie możliwe odpowiedzi. Do poszukiwania związku między zmiennymi w skali nominalnej (z wyjątkiem dychotomicznych), które mają więcej niż dwie kategorie, można wykorzystać współczynnik *lambda* Goodmana i Kruskala. W celu zbadania zależności danych w skali nominalnej warto poznać test χ^2 i współczynnik *V* Cramera.

WSPÓŁCZYNNIK KORELACJI DLA CECH W SKALI INTERWAŁOWEJ I ILORAZOWEJ

Współczynnik korelacji liniowej Pearsona można stosować dla cech w skali interwałowej i ilorazowej. Wykorzystuje on kowariancję, czyli średnią arytmetyczną iloczynu odchyleń poszczególnych wartości badanych cech x_i, y_i od wartości ich średnich. Należy wcześniej sprawdzić (np. graficznie), czy zależność między zmiennymi jest liniowa, bo tylko dla takiej wyliczony współczynnik ma sens. Kowariancję zapisujemy wzorem:

$$C_{xy} = \frac{\sum_{i=1}^{n}(x_i - \overline{x})(y_i - \overline{y})}{N},$$

gdzie:

N – liczba rozpatrywanych cech (x_i, y_i).

Współczynnik korelacji (czyli stosunek kowariancji do iloczynu odchyleń standardowych badanych zmiennych) określamy wzorem:

$$r_{xy} = \frac{C_{xy}}{\sigma_x \cdot \sigma_y},$$

gdzie:

σ_x – odchylenie standardowe.

Współczynnik korelacji przyjmuje wartości od –1 do 1. Aby go prawidłowo zinterpretować, należy zwrócić uwagę na dwie jego cechy: znak i wartość bezwzględną. Znak współczynnika wskazuje na kierunek korelacji. Jeśli jest on

dodatni, oznacza to, że wraz ze wzrostem jednej zmiennej (x_i) rośnie wartość drugiej zmiennej (y_i). Jeśli znak jest ujemny, to interpretacja jest następująca: wraz ze wzrostem jednej zmiennej (x_i) maleje wartość drugiej zmiennej (y_i).

Wartość bezwzględna współczynnika korelacji jest zawsze dodatnia i określa siłę korelacji. Dla $r_{xy} = 0$ współzależność nie występuje, dla $|r_{xy}| = 1$ współzależność jest całkowita (idealna). Im wyższa jest wartość bezwzględna współczynnika korelacji, tym wyższe jest jego znaczenie.

Obliczony w ten sposób współczynnik korelacji pozwala na opisanie współzależności cech w zbiorze danych, które były podstawą jego wyliczenia. Jeśli zebrane informacje były próbą losową z większej zbiorowości, to w tym momencie analizy nie można jeszcze formułować wniosków na temat całej populacji. Konieczne będą dalsze obliczenia, o których będzie mowa dalej.

Przykład 5.1.8

Sprawdźmy, ile będzie wynosił liniowy współczynnik korelacji dla cech: wiek (x_i) i wydatki (y_i) z przykładu 5.1.2. W tym celu obliczamy średnią arytmetyczną dla każdej cechy i budujemy tabelę (tab. 5.1.8). Można posłużyć się kalkulatorem, arkuszem kalkulacyjnym lub statystycznym programem komputerowym.

$$\overline{x} = \frac{\sum_{i=1}^{n} x_i}{n} \qquad \overline{y} = \frac{1420}{16} = 88{,}75 \text{ euro} \qquad \overline{x} = \frac{615}{16} = 38{,}4 \text{ lat.}$$

Tabela 5.1.8. Zależność między wiekiem turystów i ich wydatkami na noclegi

Imię turysty	x_i wiek	y_i wydatki	$x_i - \overline{x}$	$y_i - \overline{y}$	$(x_i - \overline{x}) \cdot (y_i - \overline{y})$	$(x_i - \overline{x})^2$	$(y_i - \overline{y})^2$
Hanna	22	48	–14,18	–35,53	503,68	200,97	1 262,34
Jan	25	48	–11,18	–35,53	397,09	124,91	1 262,34
Milena	26	58	–10,18	–25,53	259,80	103,56	651,75
Kamil	26	58	–10,18	–25,53	259,80	103,56	651,75
Jakub	26	51	–10,18	–32,53	331,03	103,56	1 058,16
Iwo	29	66	–7,18	–17,53	125,80	51,50	307,28

Tabela 5.1.8 cd.

Imię turysty	x_i wiek	y_i wydatki	$x_i - \bar{x}$	$y_i - \bar{y}$	$(x_i - \bar{x}) \cdot (y_i - \bar{y})$	$(x_i - \bar{x})^2$	$(y_i - \bar{y})^2$
Wioleta	30	66	–6,18	–17,53	108,27	38,15	307,28
Stanisław	35	50	–1,18	–33,53	39,45	1,38	1 124,22
Jerzy	38	77	1,82	–6,53	–11,91	3,33	42,63
Krystyna	40	100	3,82	16,47	62,98	14,62	271,28
Adam	44	78	7,82	–5,53	–43,26	61,21	30,57
Ewa	49	49	12,82	46,47	595,92	164,44	2 159,52
Iwona	51	150	14,82	66,47	985,33	219,74	4 418,34
Ewa	56	145	19,82	61,47	1 218,56	392,97	3 778,63
Leon	59	145	22,82	61,47	1 402,98	520,91	3 778,63
Maryla	59	150	22,82	66,47	1 517,09	520,91	4 418,34
Suma	×	×	×	×	7 752,61	2 625,73	25 523,07

Źródło: opracowanie własne na podstawie tab. 5.1.1

Tabela 5.1.9. Statystki badanych cech

Cechy	Średnia arytmetyczna	Odchylenie standardowe
Wiek (x_i)	38,4 lat	12,6 lat
Koszt noclegów (y_i)	88,75 euro	39,6 euro

Źródło: opracowanie własne na podstawie tab. 5.1.1.

$$C_{xy} = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(x_i - \bar{y})}{N} = \frac{7563,75}{16} = 472,73$$

$$\sigma_x = \sqrt{\frac{\sum_{i=1}^{n}(x_i - \bar{x})^2}{n}} = \sqrt{\frac{2543,94}{16}} = \sqrt{159} = 12,6$$

$$\sigma_y = \sqrt{\frac{\sum_{i=1}^{n}(y_i - \bar{y})^2}{n}} = \sqrt{\frac{25083}{16}} = \sqrt{51,96} = 39,6$$

$$r_{xy} = \frac{C_{xy}}{\sigma_x \cdot \sigma_y} = \frac{472,73}{12,6 \cdot 39,6} = 0,9$$

Interpretacja. Wielkość współczynnika wskazuje na wysoki, dodatni i bardzo silny związek korelacyjny między badanymi cechami (r = 0,9). Im starsi byli turyści, tym wydawali więcej pieniędzy niż młodsi. Potwierdziły się przypuszczenia z przykładu 5.1.5. Należy jeszcze upewnić się, że otrzymany wynik jest istotny statystycznie (por. przykład 4.2.1).

WSPÓŁCZYNNIK DETERMINACJI (WYJAŚNIENIA)

Następnym etapem analizy współzależności jest obliczenie współczynnika determinacji oraz sprawdzenie jego istotności (jeśli mamy do czynienia z próbą i chcemy wnioskować na całą populację). **Współczynnik determinacji** to kwadrat współczynnika korelacji (r^2). Określa on, jaką część zmienności zmiennej objaśnianej można wytłumaczyć przez zmienność wartości objaśniających. Sam współczynnik determinacji – na równi z inną miarą statystyczną – nie ustala jeszcze związku przyczynowego. Określa jedynie stopień współzależności.

Przykład 5.1.9

Obliczyliśmy współczynnik korelacji dla dwóch zmiennych – wieku turystów i ich wydatków (przykład 5.1.5); wynosi on r = 0,9. Kwadrat współczynnika korelacji pomnożony przez 100% wynosi 81%. Oznacza to, że zmienność wydatków na noclegi wśród badanych turystów w 81% była zdeterminowana zmiennością ich wieku, a 19% przypada na zmienność przypadkową zależną od innych czynników, np. płci, wykształcenia i innych.

Opisana procedura korelacyjna pozwoliła na wskazanie współzależności dwóch zmiennych, jej siły (wartość bezwzględna współczynnika korelacji) oraz kierunku (znak stojący przed wartością współczynnika). Możliwe jest jej wykorzystanie do przewidywania nieznanych wartości y przy pewnych założeniach x. Przykładowo, możemy oszacować, jakie wydatki na noclegi ponieśliby turyści w wieku od 15 do 80 lat, zakładając, że nasze wcześniejsze obliczenia są wiarygodne i istotne. Wykorzystujemy do tego metodę regresji, o której będzie szerzej w podrozdziale 5.2.

WSPÓŁCZYNNIK KORELACJI DLA CECH W SKALI PORZĄDKOWEJ

Do badania współzależności dwóch cech w skali porządkowej lub jednej w skali porządkowej, a drugiej wyższej stosujemy **współczynnik korelacji rangowej Spearmana**. Może być on stosowany do małych zbiorowości. Zależność między dwiema cechami bada się przez arbitralne uporządkowanie poszczególnych stanów każdej z cech, np. rosnąco, a następnie przyporządkowanie tym stanom kolejnych numerów odpowiadających miejscom zajmowanym przez te stany w uporządkowanych poprzednio ciągach. Przykładowo, dla pięciu osób dla cechy „wykształcenie" możemy nadać rangi: 1. – podstawowe, 2. – zawodowe (bez matury), 3. – średnie, 4. – licencjackie, 5. – magisterskie. Nadane numery w przykładzie od 1 do 5 nazywamy **rangami**.

Współczynnik korelacji rang określony jest wzorem:

$$r_s = 1 - \frac{6 \cdot \sum_{i-1}^{n} d_i^2}{n^3 - n},$$

gdzie:

d_i – różnica rang,

n – liczba obserwacji.

Współczynnik korelacji rang Spearmana, podobnie jak współczynnik korelacji Pearsona, przyjmuje wartości liczbowe z przedziału <–1; +1>, a jego interpretacja jest taka sama.

Przykład 5.1.10

Chcemy sprawdzić, czy istnieje zależność między wiekiem turystów i ich zadowoleniem z podróży. Wiek turystów jest w skali ilorazowej (x_i), a zadowolenie – w skali porządkowej (w_i), dlatego wybieramy współczynnik Spearmana. Zbudujmy tablicę korelacyjną dla danych z przykładu 4.1.2 (tab. 4.1.1). W tym celu w dodatkowych kolumnach wpisujemy rangi poszczególnych cech, zaczynając od 1 dla wartości najmniejszej. Jeśli zdarzy się, że dwoje turystów będzie miało przypisaną tę samą wartość (w przykładzie wiek 59 lat występuje dwukrotnie), to wpisujemy w kolumnę rangę będącą średnią arytmetyczną rang, które powinny wystąpić w tym miejscu (średnia arytmetyczna liczb 15 i 16 równa się 15,5). W następnym kroku odejmujemy rangi i ich różnicę podnosimy do kwadratu. Sumę ostatniej kolumny wpisujemy do licznika wzoru współczynnika korelacji Spearmana.

Tabela 5.1.10. Tablica korelacji rang. Zależność między zadowoleniem z wycieczki i wiekiem turystów

Imię turysty	Wiek x_i	Zadowolenie y_i	Ranga dx_i	Ranga dy_i	Różnica rang $d_i = dx_i - dy_i$	Kwadrat różnicy rang
Hanna	22	8	1	12	–11	121
Jan	25	9	2	14	–12	144
Milena	26	7	4	9	–5	25
Kamil	26	10	4	10	–6	36
Jakub	26	6	4	16	–12	144
Iwo	29	9,5	6	15	–9	81
Wioleta	30	8,5	7	13	–6	36
Stanisław	35	7,5	8	11	–3	9
Jerzy	38	5	9	7	2	4
Krystyna	40	4	10	6	4	16
Adam	44	3,5	11	5	6	36
Ewa	49	5,5	12	8	4	16
Iwona	51	1	13	1	12	144
Ewa	56	3	14	3	11	121
Marek	59	2	15,5	2	13,5	182,25
Krystyna	59	3	15,5	4	11,5	132,25
Suma	×	×	×	×	×	1247,5

Źródło: opracowanie własne na podstawie tab. 5.1.1.

$$r_s = 1 - \frac{6 \cdot \sum_{i=1}^{n} d_i^2}{n^3 - n} = 1 - \frac{6 \cdot 1247{,}5}{16^3 - 16} = 1 - \frac{7485}{4080} = 1 - 1{,}83 = -0{,}83.$$

Interpretacja. Mamy tu do czynienia z ujemną, wysoką zależnością, wskazującą na to, że im turyści byli starsi, tym mniej zadowoleni z podróży do Barcelony. Współczynnik determinacji informuje, że w około 70% odpowiedzi stopień zadowolenia z podróży mógł być wyjaśniony wiekiem turystów, a w 30% zależał od innych czynników, np. pogody, towarzystwa itd. Potwierdziły się przypuszczenia z przykładu 5.1.5. Należy jeszcze upewnić się, że otrzymany wynik jest istotny statystycznie (por. przykład 5.1.9).

KORELACJA CECH JAKOŚCIOWYCH DYCHOTOMICZNYCH, METODA Φ (*PHI*) YULE'A

Jest to miara nieparametryczna dla dwóch zmiennych nominalnych tabel 2 × 2. Niekiedy zachodzi potrzeba określenia zależności dwóch cech jakościowych. Badaniem tych związków zajmuje się teoria asocjacji (skojarzeń). Zagadnienie to przedstawiano na przykładzie zależności między dwiema dychotomicznymi cechami. Cechy mogą przyjmować po dwie wartości: *tak* lub *nie*, *duży* lub *mały*. Ogólnie tablicę czteropolową można przedstawić, posługując się następującym schematem (tab. 5.1.11):

Tabela 5.1.11. Tablica asocjacji, metoda Φ (*phi*) Yule'a

		Cecha pierwsza		
		tak	nie	
Cecha druga	tak	*a*	*b*	*a* + *b*
	nie	*c*	*d*	*c* + *d*
		a + *c*	*b* + *d*	*n*

a, *b*, *c*, *d* – liczby obserwacji każdego wariantu.

Źródło: opracowanie własne.

Korzystając z tab. 5.1.11, możemy posłużyć się współczynnikiem Φ (*phi*) Yule'a, określającym współzależność cech jakościowych.

$$\Phi = \frac{ad - bc}{\sqrt{(a+b)(c+d)(a+c)(b+d)}}$$

Współczynnik ten waha się w granicach $-1 \leq \Phi \leq +1$. Gdy jest równy zero, wówczas mówimy o braku współzależności. Uznajemy ją za przeciętną, gdy $0{,}3 \leq |\Phi| < 0{,}5$; za wysoką, gdy $0{,}5 \leq |\Phi| < 0{,}7$; za bardzo wysoką, gdy $|\Phi| \geq 0{,}7$. By wynik został uznany za istotny statystycznie, należy jeszcze zbadać jego istotność.

Przykład 5.1.11

Rafał postanowił sprawdzić, czy istnieje zależność między odpowiedzią respondentów na dwa pytania: „Czy biuro turystyczne, w którym kupujemy

wycieczkę jest wiarygodne?" oraz „Czy należy dodatkowo ubezpieczyć bagaż na czas podróży?". Wyniki umieścił w tab. 5.1.12.

Tabela 5.1.12. Tablica asocjacji (metoda Yule'a)

		Czy biuro turystyczne, w którym kupujemy wycieczkę, jest wiarygodne?		
		tak	nie	
Czy należy dodatkowo ubezpieczyć bagaż na czas podróży?	tak	5	20	25
	nie	15	10	25
		20	30	50

Źródło: opracowanie własne.

$$\Phi = \frac{ad - bc}{\sqrt{(a+b)(c+d)(a+c)(b+d)}} = \frac{-250}{612{,}4} = -0{,}41.$$

Interpretacja. Mamy tu do czynienia z korelacją ujemną. Można przypuszczać, że osoby, które ubezpieczały dodatkowo bagaż, nie miały potrzeby sprawdzania wiarygodności biura. Z kolei druga część ankietowanych bardziej ceniła sobie bezpieczeństwo pobytu bez przykrych niespodzianek, np. upadłości biura, niż ubezpieczenie bagażu. Należy jeszcze upewnić się, że otrzymany wynik jest istotny statystycznie (por. przykład 5.1.9).

Współczynnik Goodmana i Kruskala obliczany jest dla zmiennych w skali porządkowej, w których możliwe są więcej niż dwie odpowiedzi, czyli dla tablic większych niż 2 × 2. Jest to współczynnik symetryczny, co znaczy, że nie zakładamy, która zmienna jest zmienną zależną, a która niezależną. Oblicza się go ze wzoru:

$$\gamma = \frac{P_s - P_d}{P_s + P_d},$$

gdzie:

P_s – liczba par zgodności,

P_d – liczba par niezgodności odpowiedzi.

Współczynnik ten przyjmuje wartości od –1 do +1. Jeśli liczba par zgodnych jest większa niż niezgodnych, to wartość współczynnika będzie dodatnia, a zależność pozytywna. Jeśli w tablicy wszystkie pola będą miały tę samą wartość, to oznacza brak związku. Jeśli liczba par zgodnych jest mniejsza niż niezgodnych, to wartość współczynnika będzie ujemna, a zależność negatywna.

ISTOTNOŚĆ WSPÓŁCZYNNIKA KORELACJI

Gdy informacje pochodzą z próby i chcemy naszymi wnioskami objąć całą populację, należy wykazać, że otrzymany współczynnik korelacji nie jest równy zeru i że jego wartość nie jest spowodowana błędem próby. Jak wiemy, badania statystyczne mogą być wyczerpujące lub częściowe. Dokonując wyboru między badaniem wyczerpującym i reprezentacyjnym, należy zdać sobie sprawę z różnic zachodzących między nimi, tj. zasad opracowywania wyników i reguł ich interpretacji. Wyniki badań reprezentacyjnych są zawsze wynikami hipotetycznymi o określonych granicach niepewności, w przeciwieństwie do badań wyczerpujących, które są stwierdzeniami kategorycznymi. W zakresie opracowań wyników różnica polega na tym, że w przypadku badań wyczerpujących procedura ogranicza się do opisu statystycznego. W badaniach reprezentacyjnych opis statystyczny wyników z próby musi być uzupełniony postępowaniem umożliwiającym dokonanie uogólnień tych wyników na całą populację. Postępowanie to nazywamy **wnioskowaniem statystycznym**.

W postępowaniu zwanym wnioskowaniem statystycznym przyjmujemy zawsze dwie hipotezy: **hipotezę zerową** (H_0) i **alternatywną** (H_1). Hipoteza H_0 jest tak sformułowana, aby jej odrzucenie było jednoznaczne z przyjęciem hipotezy roboczej (alternatywnej). Weryfikacja hipotezy badawczej ma zatem charakter pośredni, gdyż bezpośrednio testowana jest hipoteza zerowa. Dopiero fakt jej przyjęcia lub odrzucenia na określonym poziomie istotności α ma konsekwencje dla hipotezy roboczej. Decydując o przyjęciu lub odrzuceniu hipotezy zerowej, ustalamy pewien poziom naszych wymagań.

Poziom istotności, zwany też poziomem α, to prawdopodobieństwo uzyskania z testu statystycznego takiej wartości, która nakazuje odrzucenie hipotezy zerowej na rzecz hipotezy roboczej, pomimo że w rzeczywistości H_0 może być prawdziwa. Poziomem istotności jest zatem ułamek wyrażający ryzyko popełnienia błędów. I tak, przy wybranym α = 0,01 odrzucimy ją 1% razy, mimo

że H_0 może być prawdziwa. Gdy $\alpha = 0{,}05$, odrzucimy H_0 5% razy, czyli ryzyko popełnienia błędu zachodzi średnio w pięciu przypadkach na 100 (5%).

Zapis informuje o popełnieniu błędu odrzucenia prawdziwej hipotezy H_0 jako nieprawdziwej. Przyjmując ryzyko popełnienia błędu w jednym przypadku na 100 (1%), decydujemy się na ryzyko popełnienia błędu $p = 0{,}01$. W praktyce najczęściej stosowane są poziomy istotności równe 0,05 i 0,01.

Wnioski statystyczne, u podstaw których leży pewność wynosząca co najmniej 95% ($p < 0{,}05$), nazywamy **istotnymi**. Kiedy podstawą odrzucenia hipotezy zerowej jest prawdopodobieństwo błędu mniejsze niż 1% (np. $\alpha = 0{,}001$), to wnioski takie określamy **wysoce istotnymi**.

Test służący do zweryfikowania hipotezy H_0, polegający wyłącznie na jej odrzuceniu lub stwierdzeniu braku podstaw do jej odrzucenia, nazywamy **testem istotności**.

Formułowanie i weryfikacja hipotez statystycznych obejmuje cztery etapy:

1. Sformułowanie hipotezy.
2. Wybór testu lub testów określających reguły postępowania przy weryfikacji hipotezy zerowej.
3. Określenie poziomu istotności (prawdopodobieństwa błędu), a tym samym wyznaczenie obszaru krytycznego hipotezy.
4. Sformułowanie (na podstawie wyniku z próby, testu i przyjętych założeń) wniosku końcowego: „sformułowaną hipotezę zerową odrzucić" albo „nie ma statystycznych podstaw do odrzucenia sformułowanej hipotezy zerowej".

W celu zbadania istotności współczynnika korelacji wykonujemy następujące czynności:

a) stawiamy hipotezę H_0 mówiącą, że współczynnik korelacji jest równy 0,

b) obliczamy wartość, korzystając z testu *t*-Studenta:

$$t = \frac{|r|\sqrt{n-2}}{\sqrt{1-r^2}},$$

gdzie: r oznacza współczynnik korelacji z n-elementowej próby pobieranej z populacji na drodze niezależnego losowania o $n - 2$ stopniach swobody,

c) dla przyjętego α (oznaczającego istotność, np. $\alpha = 0{,}1$, $\alpha = 0{,}05$, $\alpha = 0{,}01$) i $n - 2$ stopnia swobody odczytujemy t_α z tablicy (załącznik 1),

d) jeżeli $t > t_\alpha$, to współczynnik korelacji istotnie różni się od zera, jeżeli $t < t_\alpha$, przyjmujemy H_0, co oznacza, że korelacja jest nieistotna.

Przykład 5.1.12

W próbie o liczebności 12 obliczono współczynnik korelacji i otrzymano $r = 0{,}86$. Czy wynik ten przeczy, na poziomie istotności 0,05, hipotezie H_0, że współczynnik korelacji w badanej populacji jest równy zeru? Z tablicy odczytujemy wartość t (0,05; 10) = 2,2281.

Obliczona wartość t wynosi:

$$t = \frac{|r|\sqrt{n-2}}{\sqrt{1-r^2}} = \frac{|0{,}86|\sqrt{10-2}}{\sqrt{1-0{,}86^2}} = 5{,}329.$$

Interpretacja. Ponieważ 5,329 > 2,2281, uznajemy uzyskany współczynnik korelacji za istotny na poziomie 0,05. Sformułowaną hipotezę zerową odrzucamy.

TEST NIEZALEŻNOŚCI *CHI*-KWADRAT (X^2) PEARSONA

Test niezależności *chi*-kwadrat (χ^2) Pearsona jest stosowany do cech nominalnych. Jego wynik wykorzystywany jest do stwierdzenia, czy istnieje bądź nie związek pomiędzy dwiema zmiennymi, ale nie służy do oceny siły i kierunku tego związku – jak to miało miejsce w przypadku współczynnika korelacji liniowej Pearsona, stosowanego do danych w skali ilorazowej.

Formułowanie i weryfikacja hipotez statystycznych obejmuje kilka etapów:

1. Stawiamy dwie hipotezy: zerową i alternatywną:

a) H_0 – nie istnieje istotny statystycznie związek pomiędzy badanymi zmiennymi,

b) H_1 – istnieje statystycznie istotny związek pomiędzy zmiennymi.

2. Obliczamy E_{ij}, korzystając ze wzoru:

$$E_{ij} = \frac{\sum_{i=1}^{n} O_i \cdot \sum_{j=1}^{m} O_j}{\sum O},$$

gdzie:

E_{ij} – wartości oczekiwane dla komórki na przecięciu rzędu *i* oraz kolumny *j*,

O_i – wartości obserwowane dla rzędu *i*,

O_j – wartości obserwowane dla kolumny *j*,

O – wszystkie wartości obserwowane.

3. Obliczamy test, porównując wartości oczekiwane (tzn. bez żadnego związku między nimi) z wartościami obserwowanymi (liczebnościami empirycznymi).

4. Obliczamy liczbę swobody:

v = (liczba wartości pierwszej zmiennej – 1) · (liczba wartości drugiej zmiennej – 1).

5. Wartość obliczonego testu *chi*-kwadrat porównujemy z tablicami rozkładu *chi*-kwadrat (załącznik 2).

a) jeśli $t > t_{\alpha}$, to możemy odrzucić hipotezę zerową i stwierdzić, że test jest istotny statystycznie oraz możemy odrzucić hipotezę zerową i przyjąć hipotezę alternatywną o istnieniu związku między badanymi zmiennymi. Przyjmuje się, że wartość dla α powinna być mniejsza bądź równa 0,05,

b) Jeśli $t < t_{\alpha}$, to nie możemy odrzucić hipotezy zerowej i stwierdzamy, że test nie jest istotny statystycznie oraz nie możemy mówić o istnieniu związku między badanymi zmiennymi.

6. Jeśli tablica była rozmiarów 2 × 2, należy jeszcze obliczyć poprawkę ciągłości Yatesa.

Przykład 5.1.13

Przypuśćmy, że chcemy się dowiedzieć, czy istnieje związek pomiędzy płcią turysty i decyzją dotyczącą wzięcia udziału w spływie kajakowym nieznaną nikomu rzeką. W tym celu przeprowadziliśmy badania wśród 100 osób: 50 mężczyzn i 50 kobiet. Wyniki zestawiono w tab. 5.1.13.

Tabela 5.1.13. Rozkłady wartości empirycznych dla zmiennych „płeć" i „chęć uczestniczenia w spływie kajakowym"

		Czy wybrał(a)by się Pan(i) na spływ kajakowy rzeką Żejmeną na Litwie?		
		tak	nie	Razem
Płeć	kobieta	15	35	50
	mężczyzna	30	20	50
	Razem	45	55	100

Źródło: opracowanie własne.

$$E_{11} = \frac{\sum_{i=1}^{n} O_i \cdot \sum_{j=1}^{m} O_j}{\sum O} = \frac{50 \cdot 45}{100} = \frac{2250}{100} = 22{,}5,$$

$$E_{12} = \frac{\sum_{i=1}^{n} O_i \cdot \sum_{j=1}^{m} O_j}{\sum O} = \frac{50 \cdot 55}{100} = \frac{2750}{100} = 27{,}5,$$

$$E_{21} = \frac{\sum_{i=1}^{n} O_i \cdot \sum_{j=1}^{m} O_j}{\sum O} = \frac{50 \cdot 45}{100} = \frac{2250}{100} = 22{,}5,$$

$$E_{22} = \frac{\sum_{i=1}^{n} O_i \cdot \sum_{j=1}^{m} O_j}{\sum O} = \frac{50 \cdot 55}{100} = \frac{2750}{100} = 27{,}5,$$

Wyniki wstawiamy do tab. 5.1.14.

Tabela 5.1.14. Rozkłady wartości oczekiwanych dla zmiennych „płeć" i „chęć uczestnictwa w spływie kajakowym" (liczebności teoretyczne)

		Czy wybrał(a)by się Pan(i) na spływ kajakowy rzeką Żejmeną na Litwie?		
		tak	nie	Razem
Płeć	kobieta	22,5	27,5	50
	mężczyzna	27,5	22,5	50
	Razem	50,0	50,0	100

Źródło: opracowanie własne.

Różnice pomiędzy wartościami obserwowanymi i wartościami oczekiwanymi nazywamy resztami. Dzięki wartościom reszt możemy przypuszczać istnienie związku korelacyjnego. Jednak dopiero obliczenie statystyki χ^2 pozwala na właściwe wyciągnięcie wniosków.

$$\chi^2 = \sum = \frac{(O-E)^2}{E} = \frac{(15-22{,}5)^2}{22{,}5} + \frac{(35-27{,}5)^2}{27{,}5} + \frac{(30-27{,}5)^2}{27{,}5} + \frac{(20-22{,}5)^2}{22{,}5} = 5{,}05$$

Obliczamy tzw. liczbę **stopni swobody**:

v = (liczba wartości pierwszej zmiennej − 1) · (liczba wartości drugiej zmiennej − 1)

$v = (2 - 1) \cdot (2 - 1) = 1$

dla przyjętego α (oznaczającego istotność, np. $\alpha = 0{,}1$, $\alpha = 0{,}05$, $\alpha = 0{,}01$) t_α odczytujemy z tablicy. Jeśli wartość testu *chi*-kwadrat jest większa od wartości podanej w tabeli, to odrzucamy hipotezę zerową. W przykładzie dla $v = 1$ odczytujemy wartości $t_{0,1} = 2{,}706$, $t_{0,05} = 3{,}841$, a $t_{0,01} = 6{,}635$.

Przypomnijmy, że α oznacza ryzyko popełnienia błędu. Stąd wniosek, że powinno mieć najniższą wartość, ale niestety dla $\alpha = 0{,}01$ obliczona statystyka χ^2 ma wartość zbyt niską, gdyż $t_{0,01} = 6{,}635$, czyli $5{,}050 < 6{,}635$ i na tym poziomie błędu musielibyśmy wnioskować, że test nie jest statystycznie istotny.

Dla $\alpha = 0{,}05$ obliczona statystyka χ^2 ma wartość odpowiednią $t_{0,05} = 3{,}841$, gdyż $5{,}050 > 3{,}841$ i na poziomie błędu ($\alpha = 0{,}05$) możemy wnioskować, że test jest statystycznie istotny.

Wynika z tego, że możemy odrzucić hipotezę zerową i przyjąć hipotezę alternatywną – o istnieniu związku między badanymi zmiennymi.

W przykładzie tablica ma rozmiar 2 × 2, dlatego do prawidłowego zakończenia wnioskowania niezbędne jest jeszcze obliczenie poprawki na ciągłość (tzw. test *chi*-kwadrat z poprawką Yatesa). Poprawka ta polega na odjęciu liczby 0,5 od modułu różnicy między liczebnościami obserwowanymi a liczebnościami oczekiwanymi:

$$\chi^2 = \sum = \frac{(|O-E|-0{,}5)^2}{E} = \frac{(|15-22{,}5|-0{,}5)^2}{22{,}5} + \frac{(|35-27{,}5|-0{,}5)^2}{27{,}5} +$$
$$+ \frac{(|30-27{,}5|-0{,}5)^2}{27{,}5} + \frac{(|20-22{,}5|-0{,}5)^2}{22{,}5} = 4{,}283$$

Wartość testu *chi*-kwadrat z poprawką na ciągłość Yatesa pozwala na stwierdzenie, że pomiędzy zmiennymi istnieje współzależność na poziomie istotności $\alpha = 0{,}05$.

Mając χ^2, możemy obliczyć współczynnik *V* Craméra (współczynnik kontyngencji *V* Craméra lub *phi* Craméra), który służy do pomiaru siły zależności pomiędzy dwiema zmiennymi jakościowymi mierzonymi w skali nominalnej. Jest on zawsze dodatni. Aby go obliczyć, posługujemy się wzorem:

$$V = \sqrt{\frac{\chi^2}{N(m-1)}},$$

gdzie:

- m – liczba kolumn lub liczba wierszy w tabeli (bierzemy pod uwagę mniejszą liczebność),
- N – liczebność wszystkich jednostek analizy,
- χ^2 – wartość współczynnika.

Im wartość współczynnika jest bliższa zeru, tym siła zależności pomiędzy badanymi cechami jest mniejsza, a gdy zbliża się ona do jedności, tym siła zależności jest większa.

Przykład 5.1.14

Dla danych z przykładu 5.1.13 obliczamy współczynnik *V* Cramera.

$$V = \sqrt{\frac{4{,}283}{100(2-1)}} = \sqrt{0{,}04283} = 0{,}207.$$

Interpretacja. Możemy wnioskować o słabej zależności pomiędzy płcią i chęcią wyjazdu na spływ kajakowy rzeką Żejmeną na Litwie.

5.2. ANALIZA REGRESJI

Jeżeli oprócz ustalenia korelacji między dwiema cechami chcemy przewidzieć, jaką wartość będzie miała **zmienna zależna** *y* (wyjaśniana), przy ustalonej wartości **zmiennej niezależnej** *x* (wyjaśniającej), należy posłużyć się **funkcją**

regresji. Wyraża ona zależność między zmiennymi za pomocą funkcji matematycznej. Powinna być jak najlepiej dopasowana do danych liczbowych. Równanie regresji można przedstawić wzorem (Zając 1988):

$$y = f(x),$$

gdzie:

y – nieznana warunkowa średnia wartość zmiennej zależnej,
x – określona wartość zmiennej niezależnej,
f – określona postać funkcji.

Poszczególne wartości zmiennej zależnej y_i odchylają się od funkcji $f(x)$ o pewną wartość losową ξ. Wówczas model regresji prostej, podanej z dokładnością do składnika losowego ξ, wyraża się następująco:

$$y_i = f(x_i) + \xi,$$

gdzie:

y_i – zaobserwowana wartość zmiennej zależnej y,
ξ – składnik losowy (przypadkowy) o założeniach właściwych rozkładowi normalnemu.

Zakłada się, że składniki losowe mają rozkłady ze średnią równą zeru i stałą wariancję oraz że są nieskorelowane.

Rozpoczynając modelowanie regresyjne, należy określić, która zmienna jest zmienną objaśnianą, a która objaśniającą. Następnie trzeba właściwie dopasować funkcję regresji. Można wówczas wykorzystać matematyczną metodę najmniejszych kwadratów. Pozwala ona tak dopasować funkcję $f(x)$ do danych empirycznych, aby suma kwadratów odchyleń poszczególnych wartości empirycznych y_i od wartości funkcji regresji $f(x_i)$ równała się minimum (stąd pochodzi nazwa metody):

$$\psi = \sum_{i=1}^{n} [y_i - f(x_i)]^2 = \min$$

W zależności od kształtu krzywej rozróżniamy różne modele regresji: funkcję prostoliniową (tylko ta będzie omawiana), kwadratową, kubiczną, potęgową, wykładniczą i inne.

Najczęściej spotykaną funkcją regresji jest funkcja prostoliniowa: $y = \alpha + \beta x + \xi$. W tym przypadku mówimy o regresji zmiennej *y* względem zmiennej *x*. Jeżeli zaobserwujemy, że na skutek zwiększania się zmiennej niezależnej *x* zwiększa się wartość zmiennej zależnej *y*, mówimy, że jest ona **stymulantą**. W przeciwnym razie, gdy wzrost wartości zmiennej niezależnej *x* powoduje zmniejszanie się wartości zmiennej zależnej *y*, mówimy, że jest ona **destymulantą**. Jak widać na rys. 5.1.3, wiek może być zarówno stymulantą (A), jak i destymulantą (B).

Obydwie zmienne powinny mieć rozkład normalny, który możemy sprawdzić testem **Kołmogorowa-Smirnowa**[4]. Gdy mamy do czynienia z próbą i nie znamy średniej lub odchylenia standardowego całej populacji, test ten powinien być przeprowadzony z uwzględnieniem tzw. poprawki Lilleforsa. Test najlepiej wykonać, korzystając z programów statystycznych. Warto jeszcze sprawdzić kształt rozkładu za pomocą wykresu pudełkowego (rys. 3.1.4), a jeśli nadal nie mamy pewności, czy rozkład jest normalny lub o ile odbiega od normalnego, obliczamy moment centralny rzędu trzeciego i czwartego (współczynnik skośności i kurtozę z rozdziału 3). Wynik testu obejmuje, podobnie jak w teście niezależności *chi*-kwadrat (χ^2) Pearsona, określenie hipotezy zerowej H_0 (rozkład badanej cechy w populacji jest rozkładem normalnym) oraz alternatywnej H_1 (rozkład badanej cechy w populacji jest różny od rozkładu normalnego). Zakładamy, że jeśli istotność testu jest niższa niż $\alpha = 0{,}05$, to przyjmujemy H_1. Jeśli jest wyższa lub równa, wówczas nie ma podstaw do odrzucenia H_0, a więc przyjmujemy, że rozkład jest normalny.

Analiza regresji liniowej przebiega na kilku etapach, na których przyjmuje się kolejne ustalenia i ograniczenia oraz uzyskuje określone rezultaty numeryczne (Luszniewicz, Słaby 1996; zmodyfikowane).

Etap pierwszy – specyfikacja, czyli:

a) wybór zmiennych objaśnianych i objaśniających (muszą być w skali przedziałowej lub ilorazowej),

b) sprawdzenie liczebności populacji – minimum 15 obserwacji,

c) sprawdzenie, czy zmienne zależna i niezależna pochodzą z populacji o rozkładzie normalnym (test Kołmogorowa-Smirnowa lub Shapiro-Wilka),

---

4 Zamiennie do testu Kołmogorowa-Smirnowa, w przypadku analizy podobieństwa rozkładu do rozkładu normalnego, stosowany jest test Shapiro-Wilka.

d) wskazanie postaci funkcji regresji na podstawie rozrzutu punktów w układzie współrzędnych (należy ocenić, czy jest ona prostoliniowa).

Etap drugi – estymacja[5], tj. estymacja strukturalnych i stochastycznych parametrów modelu regresji na podstawie indywidualnych informacji statystycznych. Należy wykryć i wyeliminować informacje izolowane (nietypowe), które mogą zakłócić postać funkcji regresji i zafałszować późniejsze przewidywania; można skorzystać z graficznej analizy reszt statystycznego modelu regresji.

Etap trzeci – weryfikacja, czyli weryfikacja modelu regresji polegająca na badaniu jego zgodności z danymi empirycznymi.

Etap czwarty – predykcja[6] (inaczej prognoza), tzn. przewidywanie poziomów realizacji zmiennej objaśnianej przy określonych poziomach zmiennej objaśniającej.

Poszczególne etapy modelowania regresyjnego (specyfikacja, estymacja, weryfikacja i predykcja) tworzą ich naturalną sekwencję i powinny być traktowane w sposób integralny.

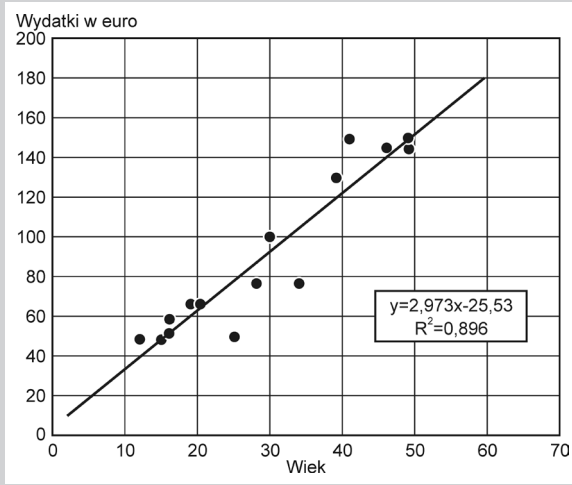
Przykład 5.2.1

Na potrzeby obecnego przykładu, wykorzystując wcześniejszy przykład 5.1.2, w którym poszukiwano współzależności między wiekiem turystów i ich wydatkami na noclegi (r = 0,9), można wykreślić linię regresji (rys. 5.2.1). Nie zawsze przechodzi ona przez punkty zmiennej niezależnej x_i (wiek) oraz zależnej y_i (wydatki), ale zawsze przechodzi przez ich wartości średnie. Linię tę można opisać za pomocą wzoru $y = f(x)$. Odchylenie punktów od linii regresji (w prezentowanym przykładzie będzie to $y = -25{,}53 + 2{,}973x$) wskazuje na różnicę między wartością przewidywaną i rzeczywistą. Jeśli spojrzymy na wykres rozrzutu z linią regresji, to możemy sprawdzić, które z naszych obserwacji leżą wzdłuż linii, a które mają najwyższe odchylenia od tej linii (np. 35-letni Stanisław, który wydał tylko 50 euro, a z funkcji regresji wynika, że dla osób w tym wieku przewidywano 78 euro).

---

5 Estymacja jest procesem wnioskowania o numerycznych wartościach nieznanych wielkości charakteryzujących populację generalną na podstawie niekompletnych danych, takich jak próba.

6 Predykcja jest procesem określania przyszłych wielkości zmiennych losowych.



Rysunek 5.2.1. Wykres rozrzutu z zaznaczoną linią regresji (*x* – zmienna niezależna, *y* – zmienna zależna)

Źródło: opracowanie własne.

Zakładamy hipotezę zerową H_0 o braku wpływu zmiennej niezależnej na zależną. Obliczamy wartość statystyki *t*-Studenta dla współczynnika korelacji $r = 0{,}9$ i $n - 2$ stopni swobody.

$$t = \frac{|r|\sqrt{n-2}}{\sqrt{1-r^2}} = \frac{|0{,}9|\sqrt{16-2}}{\sqrt{1-0{,}9^2}} = 8{,}26,$$

gdzie: $n = 16$, $r = 0{,}9$.

Wartość *t* wynosi 8,26 i należy ją porównać z t_α. Z tablicy statystycznej w załączniku 1 (rozkład *t*-Studenta) odczytujemy wartość t_α w zależności od α:

$$t_\alpha = (0{,}05;\ n - 2 = 14) = 2{,}1448;\ t_\alpha = (0{,}01;\ n - 2 = 14) = 2{,}9768;$$
$$t_\alpha = (0{,}001;\ n - 2 = 14) = 4{,}1403$$

Którą wartość t_α wybrać? Oczywiście tę, w której prawdopodobieństwo popełnienia błędu jest mniejsze, czyli dla $\alpha = 0{,}001$. Ponieważ $8{,}26 > 4{,}14$, uznajemy uzyskany współczynnik korelacji za wysoce istotny ($\alpha = 0{,}001$), a sformułowaną hipotezę zerową odrzucamy. Teraz możemy oszacować wydatki turystów w zależności od ich wieku.

Rzeczywiste i przewidywane wydatki na noclegi turystów w Barcelonie w 2013 r. na podstawie informacji o wieku i wydatkach na noclegi w badanej grupie turystów, korelacji między nimi (r = 0, 9) i reszty z regresji przedstawiono w tab. 5.2.1.

Tabela 5.2.1. Rzeczywiste i przewidywane wydatki na noclegi turystów w Barcelonie w 2013 r. oraz reszty z regresji

Wyszczególnienie	Wiek y	Wydatki x	Przewidywane wydatki $\hat{y} = -25{,}53 + 2{,}973x$	Reszty z regresji $Y - \hat{y}$
Anna	22	48	39,9	–8,1
Jan	25	48	48,8	0,8
Milena	26	58	51,8	–6,2
Kamil	26	58	51,8	–6,2
Jakub	26	51	51,8	0,8
Iwo	29	66	60,7	–5,3
Wioleta	30	66	63,7	–2,3
Stefan	35	50	78,5	28,5
Jerzy	38	77	87,4	10,4
Zofia	40	100	93,4	–6,6
Adam	44	78	105,3	27,3
Maria	49	49	120,1	–9,9
Iwona	51	150	126,1	–23,9
Ewa	56	145	141,0	–4,0
Leon	59	145	149,9	4,9
Maryla	59	150	149,9	–0,1

Źródło: opracowanie własne na podstawie tab. 5.1.1.

Interpretacja. Analiza korelacji pozwoliła na wskazanie statystycznie istotnej zależności (α = 0,001, r = 0,9), że wiek turystów miał wpływ na wysokość wydatków w czasie ich podróży do Barcelony. Im turyści byli starsi, tym więcej pieniędzy wydawali. Analiza regresji pozwala na oszacowanie potencjalnych wydatków turystów w zależności od ich wieku: $\hat{y} = -25{,}53 + 2{,}973x$. Wstawiając w miejsce x dowolny wiek turysty, możemy przypuszczać, ile wyda on pieniędzy. Na przykład, gdyby w grupie była osoba 55-letnia, to można przypuszczać, że wydałaby około 189 zł. Reszty z regresji przypominają, że nie jest to wartość precyzyjna.

Dodatkową informacją wykorzystywaną do analizy statystycznej są reszty z regresji (kolumna 5 w tab. 5.2.1), które są wynikiem odejmowania wartości empirycznej zmiennej objaśnianej *y* i wartości teoretycznej zmiennej otrzymanej z obliczeń funkcji regresji na podstawie modelu (kolumna 4 w tab. 5.2.1). Pozwalają one na ocenę stopnia dopasowania funkcji regresji do rozkładu punktów (gdyby była idealnie dopasowana, to wartości reszt byłyby równe zeru).

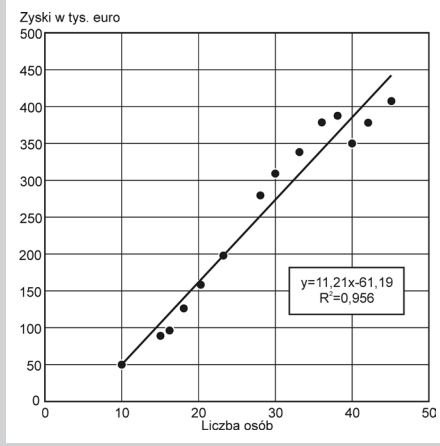
Przykład 5.2.2

Weźmy pod uwagę następujące dane: *x* – liczba osób przyjeżdżających do sieci hoteli „Ibis" (w tys. osób), *y* – jego zyski (w tys. euro) w 2003 r. (dane umowne).

x	10	15	16	18	20	23	28	30	33	36	38	40	42	45
y	50	90	96	126	160	200	280	310	340	380	390	350	380	410

Przyjmujemy, że zmienną niezależną jest liczba gości hotelowych *x*, a zmienną zależną zysk *y*. Przedstawmy zebrane dane w postaci wykresu punktowego, w którym na osi *X* umieścimy dane dotyczące liczby turystów, a na osi *Y* zyski.

Oszacowana na podstawie danych funkcja regresji ma postać: $y = -61{,}19 + 11{,}21x$, a współczynnik korelacji liniowej Pearsona: $r = 0{,}98$.



Rysunek 5.2.2. Wykres rozrzutu z zaznaczoną linią regresji (*x* – zmienna niezależna, *y* – zmienna zależna)

Źródło: opracowanie własne.

Zakładamy hipotezę zerową H_0 o braku wpływu zmiennej niezależnej na zależną. Obliczamy wartość statystyki *t*-Studenta dla współczynnika korelacji $r = 0{,}96$ i $n - 2$ stopni swobody. Wynosi ona 17,06. Z tablicy statystycznej rozkładu *t*-Studenta odczytujemy wartość:

$t_\alpha = (0{,}05;\ n - 2 = 12) = 2{,}201;$

$t_\alpha = (0{,}01;\ n - 2 = 14) = 3{,}106;$

$t_\alpha = (0{,}001;\ n - 2 = 14) = 4{,}437$

$$t = \frac{|r|\sqrt{n-2}}{\sqrt{1-r^2}} = \frac{|0{,}98|\sqrt{14-2}}{\sqrt{1-0{,}98^2}} = 17{,}06.$$

Ponieważ 17,06 > 4,437, uznajemy uzyskany współczynnik korelacji za wysoce istotny: $\alpha = 0{,}001$, a sformułowaną hipotezę zerową odrzucamy.

Współczynnik determinacji wynosi aż 95,67%, co oznacza że 95,67% zmian zmiennej zależnej zostało wyjaśnionych zmiennością liczby turystów. Korzystając z wykresu, można obliczyć wartości funkcji dla dowolnej zmiennej niezależnej, a w szczególności dla danych empirycznych.

Tabela 5.2.2. Rzeczywiste i przewidywane zyski sieci hoteli „Ibis" w 2003 r. oraz reszty z regresji

Liczba osób ***x***	**Zyski (w tys. euro)**	**Przewidywane zyski** ***y* = 11,21 *x* – 61,19**	**Reszty z regresji**
10	50	50,9	0,9
15	90	107,0	17,0
16	96	118,2	22,2
18	126	140,6	14,6
20	160	163,0	3,0
23	200	196,6	–3,4
28	280	252,7	–27,3
30	310	275,1	–34,9
33	340	308,7	–31,3
36	380	342,4	–37,6
38	390	364,8	–25,2
40	350	387,2	37,2
42	380	409,6	29,6
45	410	443,3	33,3

Źródło: opracowanie własne na podstawie danych umownych.

Ostatni etap analizy regresji to predykcja na podstawie określonego modelu regresji. Możemy zatem prognozować, jaki będzie zysk dla dowolnej liczby turystów. I tak, przykładowo, dla wartości zmiennej niezależnej $x = 60$ teoretyczna wartość zysku wynosi 611,71 euro.

PRZYKRE NIESPODZIANKI, CZYLI JAK UNIKNĄĆ BŁĘDÓW W INTERPRETACJI?

Zanim przystąpi się do obliczania współczynnika korelacji lub analizy regresji, warto przyjrzeć się wartościom zmiennych, gdyż możemy mieć do czynienia z sytuacją, gdy:

1) w zbiorze danych są obserwacje nietypowe,

2) zbiór danych może składać się z kilku podzbiorów,

3) zależność może być krzywoliniowa.

Wartości nietypowe mogą wynikać albo z błędu pomiaru, albo z błędu wpisywania danych do bazy danych (wystarczy, że dodamy jedno zero na końcu wartości), albo ze specyfiki zbiorowości (na wycieczkę ze studentami zabrał się dziadek milioner). Może też się zdarzyć, że zbiór danych obejmuje dwa różne podzbiory, np. dzieci na wycieczce szkolnej i dorosłych (bez dzieci) na urlopie.

Przedstawienie danych w postaci graficznej pozwala przed przystąpieniem do obliczeń upewnić się, czy nie mamy do czynienia z którąś z podanych sytuacji, a także ocenić przewidywany kształt krzywej rozkładu.

5.3. ZADANIA

ZADANIE 5.3.1

Przedstaw w postaci graficznej zależności między podanymi poniżej parami cech, oszacuj korelację, a następnie oblicz ją za pomocą współczynnika Spearmana lub Pearsona. Pamiętaj, że można korelować ze sobą szeregi równej długości. Czy dla wszystkich par potrafisz obliczyć współczynnik? Zbadaj istotność współczynników, korzystając z testu *t*-Studenta.

A	B	C	D	E	F	G	H	I	J	K	L
13	4	20	30	2	20	6	10	30	9	4	4
6	5	21	28	4	18	5	12	33	12	6	12
6	7	23	26	6	16	9	5	36	15	3	9
9	8	26	24	8	14	4	5	39	16	9	18
10	10	28	22	12	12	2	2	40	19	5	10
12	12	30	20	14	10	1	10	44	25	12	24
14	15	35	16	16	8	6	2	48	50	25	50
15	16	45	15	18	6	5	7	49	55	14	28
18	18	50	10	20	4	3	3	56	58	13	26
9	9			30	3	10	8	58	20	29	58
10	12					5	5	65	19	30	60
11	13							69	18	33	66
12	14							75	10		
12	16							78			
15	16										
21	18										

ZADANIE 5.3.2

Ustal, jaka jest zależność między sprzedażą lodów w kawiarni „Hortex" w Warszawie i średnią temperaturą dnia (°C) w ciągu dwóch tygodni lipca 2016 r. (tab. 5.3.1). Oblicz współczynnik korelacji Pearsona, gdyż cechy mają skalę ilorazową i interwałową.

Tabela 5.3.1. Sprzedaż lodów w kawiarni „Hortex" w Warszawie w lipcu 2000 r.

Dzień	Wartość sprzedanych lodów w zł	Średnia temperatura dnia (°C)
1.	6000	20
2.	6500	22
3.	6000	23
4.	7000	25
5.	7500	26
6.	6000	22
7.	5000	15
8.	5500	16
9.	4500	15

Tabela 5.3.1 cd.

Dzień	Wartość sprzedanych lodów w zł	Średnia temperatura dnia (°C)
10.	4000	15
11.	4000	17
12.	6000	17
13.	6200	18
14.	6800	19

Źródło: dane umowne.

ZADANIE 5.3.3

Jaka jest współzależność między wykształceniem i długością wypoczynku poza domem w grupie 10 pracowników spółki „Kormoran" w Zamościu w 2000 r. (tab. 5.3.2). Jaką metodę zastosujesz i dlaczego?

Tabela 5.3.2. Pracownicy spółki „Kormoran" w Zamościu w 2000 r. według wykształcenia i długości wypoczynku poza domem

Pracownik	Wykształcenie	Liczba dni
Anna	podstawowe	7
Bogdan	średnie	10
Celina	zawodowe	9
Damian	wyższe magisterskie	25
Ewa	policealne	20
Franciszek	wyższe licencjackie	22
Grażyna	podstawowe	9
Henryk	podstawowe	5
Irena	średnie	12
Joachim	policealne	18

Źródło: dane umowne.

ZADANIE 5.3.4

Jak sądzisz, czy istnieje korelacja pomiędzy przedstawionymi poniżej cechami, a jeżeli tak, to jaki jest jej kierunek?

a) zależność pomiędzy znajomością francuskiego i wyborem Francji na wycieczkę,

b) zależność pomiędzy wysokością zarobków a wysokością wydatków na prezenty,

c) zależność pomiędzy liczbą dzieci a odległością do miejsc wypoczynku,

d) zależność pomiędzy narodowością a wyborem środka transportu na wakacje,

e) zależność pomiędzy liczbą osób na dworcach a liczbą pociągów,

f) zależność pomiędzy wykształceniem a częstością odwiedzania muzeów,

g) zależność pomiędzy ilorazem inteligencji a wyborem kraju na urlop,

h) zależność pomiędzy głębokością rzeki a jej temperaturą,

i) zależność pomiędzy wysokością n.p.m. a ciśnieniem atmosferycznym,

j) zależność pomiędzy szerokością geograficzną a średnią roczną temperaturą,

k) zależność pomiędzy wiekiem a korzystaniem z aplikacji booking.com.

ZADANIE 5.3.5

Jakie współczynniki korelacji zastosujesz do obliczenia zależności z zadania 5.3.4? Zanim odpowiesz, zastanów się, w jakiej skali pomiarowej są dane.

ZADANIE 5.3.6

Zbadaj zależność między liczbą restauracji w turystycznych obiektach hotelowych (tab. 5.3.3) a liczbą tych obiektów (tab. 5.3.4) w Polsce w 2014 r. według województw.

Tabela 5.3.3. Placówki gastronomiczne w turystycznych obiektach noclegowych w Polsce w 2014 r. według województw

Województwo	Placówki gastronomiczne			
	restauracje	bary i kawiarnie	stołówki	punkty gastronomiczne
	liczba obiektów			
Łódzkie	171	116	36	27
Mazowieckie	270	162	53	21
Małopolskie	379	269	386	44
Śląskie	248	190	93	35

Tabela 5.3.3 cd.

Województwo	Placówki gastronomiczne			
	restauracje	bary i kawiarnie	stołówki	punkty gastronomiczne
	liczba obiektów			
Lubelskie	140	71	39	17
Podkarpackie	194	100	49	17
Podlaskie	76	46	24	1
Świętokrzyskie	114	60	24	4
Lubuskie	101	59	48	18
Wielkopolskie	314	186	63	47
Zachodniopomorskie	189	252	278	40
Dolnośląskie	295	224	205	58
Opolskie	63	39	12	9
Kujawsko-pomorskie	137	102	71	19
Pomorskie	273	240	229	28
Warmińsko-mazurskie	182	164	73	39

Źródło: opracowanie własne na podstawie danych GUS, https://stat.gov.pl (dostęp: 26.05.2015).

Tabela 5.3.4. Turystyczne obiekty noclegowe w Polsce w 2014 r. według województw

Województwo	Obiekty noclegowe		Miejsca noclegowe	
	hotelowe	pozostałe razem	w obiektach hotelowych	w pozostałych obiektach razem
Łódzkie	206	156	14 863	8 368
Mazowieckie	297	179	36 208	11 713
Małopolskie	449	969	40 529	46 679
Śląskie	304	332	26 040	18 956
Lubelskie	149	214	8 420	12 387
Podkarpackie	197	316	11 074	16 524
Podlaskie	87	161	5 971	6 877
Świętokrzyskie	128	103	7 977	6 921
Lubuskie	127	156	7 383	10 828
Wielkopolskie	344	367	22 591	20 006
Zachodniopomorskie	246	1 076	22 893	98 724

Województwo	Obiekty noclegowe		Miejsca noclegowe	
	hotelowe	pozostałe razem	w obiektach hotelowych	w pozostałych obiektach razem
Dolnośląskie	381	522	33 359	26 961
Opolskie	71	71	3 278	4 635
Kujawsko-pomorskie	149	182	10 165	16 742
Pomorskie	314	1 136	24 968	72 163
Warmińsko-mazurskie	197	299	16 802	23 018

Źródło: opracowanie własne na podstawie danych GUS, https://stat.gov.pl (dostęp: 26.05.2015).

ZADANIE 5.3.7

Zbadaj zależność między liczbą obiektów hotelowych i liczbą innych obiektów noclegowych (tab. 5.3.4) w Polsce w 2014 r. według województw.

ZADANIE 5.3.8

Zbadaj zależność między liczbą obiektów hotelowych i liczbą miejsc noclegowych w hotelach (tab. 5.3.4) w Polsce w 2014 r. według województw.

ZADANIE 5.3.9

Zbadaj, czy istniała zależność między całoroczną bazą noclegową w Polsce w latach 1980–1992 i liczbą przyjazdów do Polski cudzoziemców (tab. 5.3.5). Poszukaj na stronie https://stat.gov.pl aktualnych danych i porównaj korelacje.

Tabela 5.3.5. Baza noclegowa całoroczna w Polsce i liczba przyjazdów cudzoziemców do Polski w latach 1980–1992

Rok	Miejsca noclegowe w tys.	Liczba osób w tys.
1980	305,2	7030
1985	254,5	3410
1990	244,7	18 211
1991	216,9	36 846
1992	230,5	49 015
1993	263,2	60 951

Źródło: GUS (1993).

ZADANIE 5.3.10

Zbadaj, czy istniała zależność między liczbą hoteli w Polsce w latach 1980–1992 i liczbą przyjazdów cudzoziemców do Polski. Poszukaj na stronie https://stat.gov.pl aktualnych danych i porównaj korelacje.

Tabela 5.3.6. Liczba hoteli w Polsce i liczba przyjazdów cudzoziemców do Polski w latach 1980–1992

Rok	Liczba hoteli w tys.	Liczba osób w tys.
1980	49,9	7030
1985	51,3	3410
1990	57,4	18 211
1991	58,3	36 846
1992	64,7	49 015
1993	74,3	60 951

Źródło: GUS (1993).

ZADANIE 5.3.11

Rafał postanowił sprawdzić, czy istnieje zależność między wyborem pory roku na urlop i destynacją wśród znajomych babci i rodziców według danych z tab. 5.3.7. Oblicz współczynnik korelacji i podaj jego interpretację.

Tabela 5.3.7. Wybór kierunku wyjazdu urlopowego latem i zimą

Destynacja	Czy woli Pan/i spędzać urlop	
	latem	zimą
Morze	25	5
Góry	10	30

Źródło: dane umowne.

ZADANIE 5.3.12

Sprawdź graficznie, czy istnieje współzależność między płcią i ochotą na wędkowanie w grupie badanych osób.

Tabela 5.3.8. Płeć a zainteresowanie respondentów (*N* = 320) wędkowaniem

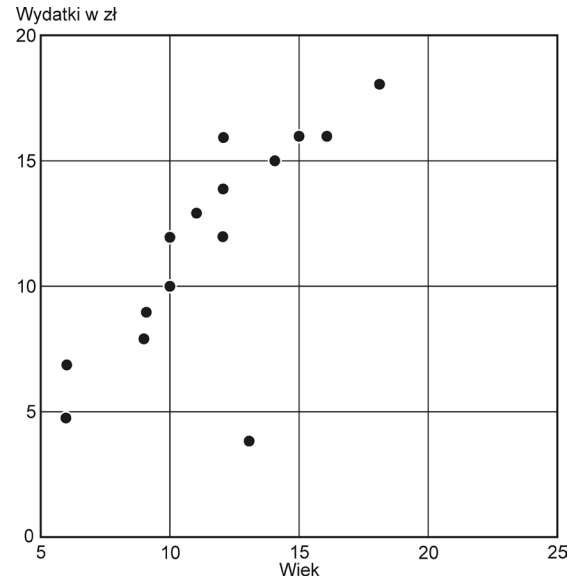
Płeć	Czy lubi Pan/i wędkować w wolnym czasie?					Suma
	bardzo lubię	lubię	trudno powiedzieć	nie lubię	bardzo nie lubię	
Kobiety	1	5	10	55	79	150
Mężczyźni	39	65	40	15	11	170
Razem	40	70	50	70	90	320

Źródło: dane umowne.

5.4. ODPOWIEDZI DO WYBRANYCH ZADAŃ

ZADANIE 5.3.1

Wyobraźmy sobie, że dwie pierwsze kolumny odpowiadają takim cechom, jak wiek i wydatki na słodycze. Czy istnieje statystyczna zależność między nimi? Narysujmy wykres, zbudujmy tablicę i obliczmy współczynnik korelacji Pearsona, gdyż są to dwie cechy w skali ilorazowej.



Rysunek 5.3.1. Ankietowani według wieku i wydatków na słodycze

Źródło: zadanie 5.3.1, dane umowne.

Tabela 5.3.9. Tablica korelacji wieku (x_i) oraz wydatków na słodycze (y_i) ankietowanych

Wiek (x_i)	Wydatki (y_i) w zł			
	<0–5)	**<5–10)**	**<10–15)**	**<15–20)**
<5–10)	1	3	–	–
<10–15)	1	–	5	2
<15–20)	–	–	–	3
<20–25)	–	–	–	1

Źródło: opracowanie własne na podstawie zadania 5.3.1.

Na podstawie rys. 5.3.1 i tab. 5.3.9 możemy przypuszczać, że istnieje współzmienność między badanymi cechami oraz że jest ona dodatnia i liniowa. Należy policzyć współczynnik korelacji, aby wskazać jej siłę.

Tabela 5.3.10. Zależność między wiekiem turystów a ich wydatkami na słodycze (N = 16)

Lp.	x_i (wiek)	y_i (wydatki)	$x_i - \bar{x}$	$y_i - \bar{y}$	$(x_i - \bar{x}) \cdot (y_i - \bar{y})$	$(x_i - \bar{x})^2$	$(y_i - \bar{y})^2$
1.	13	4	1,3	–8,1	–10,53	1,69	65,61
2.	6	5	–5,7	–7,1	40,47	32,49	50,41
3.	6	7	–5,7	–5,1	29,07	32,49	26,01
4.	9	8	–2,7	–4,1	11,07	7,29	16,81
5.	10	10	–1,7	–2,1	3,57	2,89	4,41
6.	12	12	0,3	–0,1	–0,03	0,09	0,01
7.	14	15	2,3	2,9	6,67	5,29	8,41
8.	15	16	4,3	3,9	16,77	8,49	15,21
9.	8	18	6,3	5,9	37,17	39,69	34,81
10.	9	9	–2,7	–3,1	8,37	7,29	9,61
11.	10	12	–1,7	–0,1	0,17	2,89	0,01
12.	11	13	–0,7	0,9	–0,63	0,49	0,81
13.	12	14	0,3	1,9	0,57	0,09	3,61
14.	12	16	0,3	3,9	1,17	0,09	15,21
15.	15	16	3,3	3,9	12,87	10,89	15,21
16.	25	18	9,3	5,9	54,87	86,49	34,81
Suma	187	193	×	×	211,62	248,64	300,96

Źródło: opracowanie własne.

Tabela 5.3.11. Statystyki badanych cech

Cechy	Średnia arytmetyczna	Odchylenie standardowe
Wiek: x_i	11,7 lat	3,94 lat
Wydatki na słodycze: y_i	12,1 zł	4,34 zł

Źródło: opracowanie własne na podstawie tab. 5.3.10.

$$C_{xy} = \frac{\sum_{i=1}^{n}(x_i - \overline{x})(y_i - \overline{y})}{N} = \frac{211{,}64}{16} = 13{,}23,$$

$$\sigma_x = \sqrt{\frac{\sum_{i=1}^{n}(x_i - \overline{x})^2}{n}} = \sqrt{\frac{248{,}64}{16}} = 3{,}94,$$

$$\sigma_y = \sqrt{\frac{\sum_{i=1}^{n}(y_i - \overline{y})^2}{n}} = \sqrt{\frac{300{,}96}{16}} = 4{,}34,$$

$$r_{xy} = \frac{C_{xy}}{\sigma_x \cdot \sigma_y} = \frac{12{,}23}{3{,}94 \cdot 4{,}34} = 0{,}77.$$

Interpretacja. Badana grupa liczyła 16 osób w wieku od 6 do 25 lat. Wielkość współczynnika wskazuje na dodatni i silny związek korelacyjny między badanymi cechami (r = 0,77). Im badani byli starsi, tym wydawali więcej na słodycze niż młodsi. Potwierdziły się przypuszczenia z przykładu 5.1.5. Należy jeszcze upewnić się, że otrzymany wynik jest istotny statystycznie (por. przykład 4.2.1).

ZADANIE 5.3.2

Obliczamy współczynnik korelacji Pearsona, gdyż cechy mają skalę ilorazową i interwałową. R = 0,8.

ZADANIE 5.3.3

Obliczamy współczynnik korelacji Spearmana, gdyż cechy mają skalę ilorazową i porządkową.

Tabela 5.3.12. Pracownicy spółki „Kormoran" w Zamościu w 2000 r. według wykształcenia i długości wypoczynku poza domem

Imię turysty	Wiek x_i	Długość wypoczynku y_i	Ranga dx_i	Ranga dy_i	Różnica rang $d_i = dx_i - dy_i$	Kwadrat różnicy rang
Anna	podstawowe	7	1,5	2,0	–0,50	0,25
Bogdan	średnie	10	5,5	5,0	0,50	0,25
Celina	zawodowe	9	3,0	3,5	–0,50	0,25
Damian	wyższe magisterskie	25	10,0	10,0	0,00	0,00
Ewa	policealne	20	7,5	8,0	–0,50	0,25
Franciszek	wyższe licencjackie	22	9,0	9,0	0,00	0,00
Grażyna	podstawowe	9	4,0	3,5	0,50	0,25
Henryk	podstawowe	5	1,5	1,0	0,50	0,25
Irena	średnie	12	5,5	6,0	–0,50	0,25
Joachim	policealne	18	7,5	7,0	0,55	0,25
					Suma	2

Źródło: opracowanie własne na podstawie tab. 5.3.2.

$$r_s = 1 - \frac{6 \cdot \sum_{i=1}^{n} d_i^2}{n^3 - n} = 1 - \frac{6 \cdot 2}{10^3 - 10} = 1 - \frac{12}{990} = 1 - 0{,}0(12) = -0{,}99.$$

Interpretacja. Mamy tu do czynienia z dodatnią, bardzo wysoką zależnością, świadczącą, że im lepiej byli wykształceni pracownicy, tym na dłużej wyjeżdżali na urlop poza miejsce zamieszkania. Współczynnik determinacji wskazuje, że w około 98% odpowiedzi długość wyjazdu mogła być wyjaśniona wykształceniem pracowników, a w 2% zależała od innych czynników, np. pogody, towarzystwa. Jeśli dane pochodziły z próby, należy jeszcze upewnić się, że otrzymany wynik jest istotny statystycznie (por. przykład 5.1.9).

ZADANIE 5.3.11

Obliczamy współczynnik korelacji Yule'a, gdyż cechy mają skalę nominalną dychotomiczną.

Tabela 5.3.13. Tablica korelacji (metoda Yule'a)

		Czy woli Pan/i spędzać urlop?		
		latem	zimą	Razem
Typ krajobrazu	morze	25	5	30
	góry	10	30	40
	Razem	35	35	70

Źródło: opracowanie własne.

$$\Phi = \frac{ad - bc}{\sqrt{(a+b)(c+d)(a+c)(b+d)}} = \frac{750 - 50}{\sqrt{(25+5)(10+30)(25+10)(5+30)}} =$$

$$= \frac{700}{1212} = 0{,}578$$

Interpretacja. Zachodzi tu silna korelacja dodatnia. Można przypuszczać, że osoby, które wolały spędzać urlop latem, przeważnie wybierały morze. Z kolei druga część ankietowanych wolała spędzać urlop zimą i wyjeżdżała w większości w góry.

ROZDZIAŁ 6

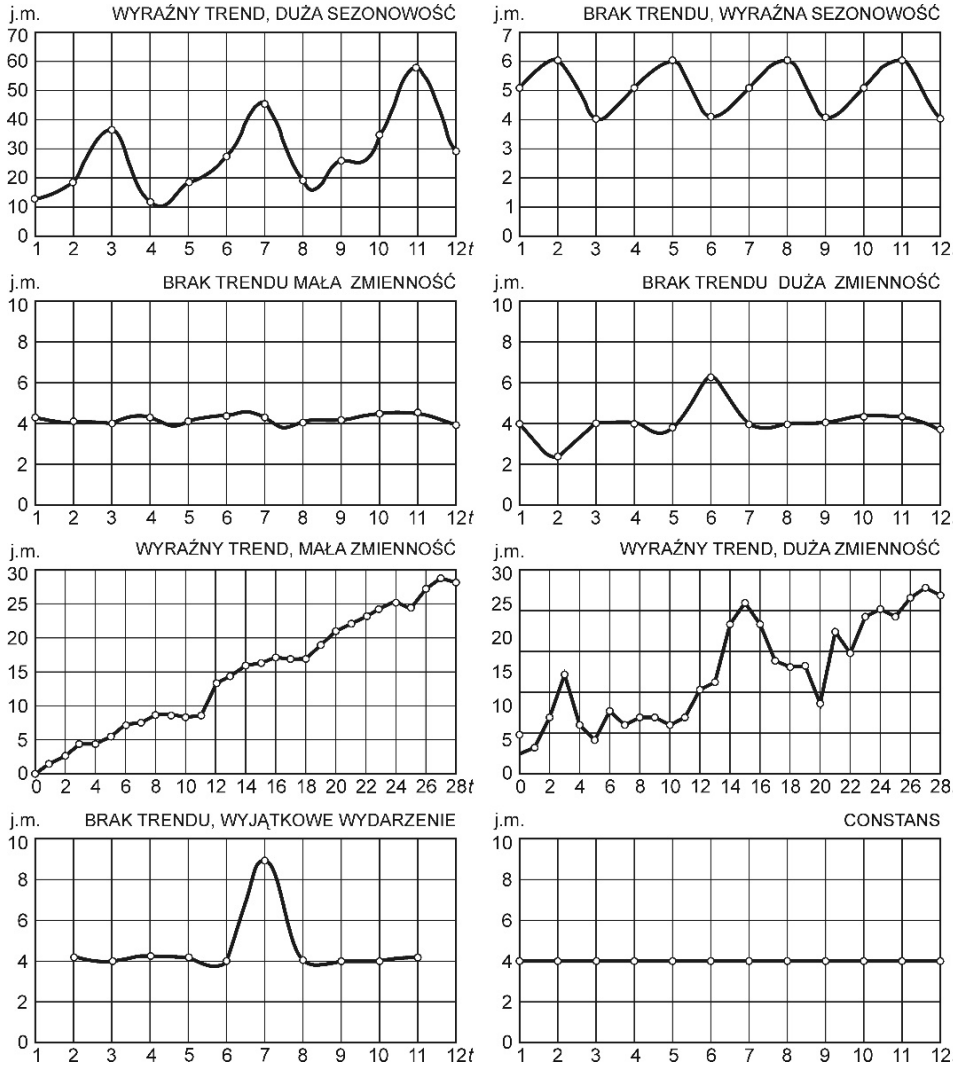
ANALIZA DYNAMIKI

Zarówno pojedyncze zjawiska, jak i zbiorowości statystyczne charakteryzują się tym, że podlegają zmianom w czasie. Dlatego w statystyce ujmuje się zjawiska nie tylko w sposób statyczny, lecz także uwzględnia się analizę dynamiki zjawisk. Dynamikę zjawisk prezentuje się w postaci szeregów chronologicznych, w których podane są momenty lub okresy i odpowiadające im wielkości badanego zjawiska (por. rozdział 2). W szeregach tych zmienna wyrażona jest w takich jednostkach, jak lata, kwartały, miesiące itp. Zmienną tę będziemy oznaczać przez T, a momenty, w których rejestrujemy zjawiska jako t_1, t_2, ..., t_n. Wartości odnotowane w tym czasie natomiast odpowiednio: z_1, z_2, ..., z_n. W takich szeregach ważny jest nie tylko zbiór wartości, ale też kolejność, w jakiej one występują. Metody analizy dynamiki są stosowane głównie w aspektach ekonomicznych i finansowych turystyki, a także ruchu turystycznego.

6.1. WYKRESY DYNAMIKI

Analiza statystyczna, w której jedną z cech jest czas, polega na określeniu zmian wielkości badanego zjawiska oraz kierunku, w jakim te zmiany przebiegają, tzn. na ustaleniu stopnia poziomu wzrostu lub spadku w obrębie

badanego zjawiska. Kiedy zjawisko jest zmienne w czasie, ustala się, czy podlega ono dużej czy małej zmienności. Na następnym etapie analiz poszukiwany jest trend, czyli wskazanie na trwałą zwyżkę lub spadek dotyczący danego zjawiska. W analizach turystycznych prezentuje się często sezonowość, np. ruchu turystycznego.



Rysunek 6.1.1. Wykresy liniowe prezentujące dynamikę zmian

Źródło: opracowanie własne.

Na wstępie analizy dynamiki można zmienność w czasie przedstawić na wykresach. Konstruuje się je w prostokątnym układzie współrzędnych, gdzie na osi *X* odmierzany jest czas t_i, a na osi *Y* wielkość zjawiska z_i w danym czasie. Zazwyczaj jest to wykres liniowy, słupkowy, liniowo-słupkowy, powierzchniowy (rys. 6.1.1, 6.1.2). Wahania zmiennej, które można wskazać na wykresie, mogą być: krótkookresowe – np. jednorazowa organizacja wydarzenia, takiego jak zawody sportowe w piłce nożnej o zasięgu europejskim; sezonowe – tj. zależne od pory roku; koniunkturalne – m.in. związane z koniunkturą finansową przedsiębiorstwa; przypadkowe – wynikające z wydarzenia nieprzewidzianego przez organizatorów turystyki (powódź, wydarzenie polityczne, zamach itp.).

Przykład 6.1.1

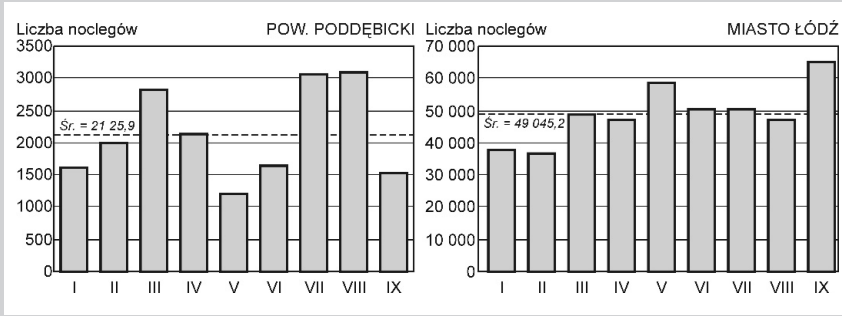
Według danych Urzędu Statystycznego w Łodzi do województwa łódzkiego w trzech pierwszych kwartałach 2011 r. przyjechało blisko 1,5 mln turystów. Ich liczba w poszczególnych miesiącach nie była równa. Nasuwa się pytanie: czy istnieje sezonowość w przyjazdach do Łodzi, powiatu poddębickiego i województwa łódzkiego? Można ją odczytać na rys. 6.1.2.

Tabela 6.1.1. Liczba noclegów udzielonych w obiektach zakwaterowania zbiorowego według miesięcy w trzech pierwszych kwartałach 2011 r. (dane Urzędu Statystycznego w Łodzi)

Obszar	Miesiąc				
	I	II	III	IV	V
Powiat poddębicki	1 606	2 006	2 823	2 145	1 207
Miasto Łódź	38 006	36 870	48 737	46 984	58 270
Województwo łódzkie	124 464	111 038	146 301	146 272	174 132

Obszar	Miesiąc			
	VI	VII	VIII	IX
Powiat poddębicki	1 650	3 075	3 098	1 523
Miasto Łódź	50 343	50 063	47 255	64 879
Województwo łódzkie	174 083	207 302	204 091	195 228

Źródło: *Ruch turystyczny…* (2012).



Rysunek 6.1.2. Liczba noclegów udzielonych w obiektach zakwaterowania zbiorowego według miesięcy w trzech pierwszych kwartałach 2011 r. w powiatach poddębickim i miejskim Łódź

Źródło: opracowanie własne na podstawie tab. 6.1.1.

Interpretacja. Zjawisko sezonowości ruchu turystycznego w badanych powiatach w trzech pierwszych kwartałach 2011 r. jest odmienne. W powiecie poddębickim sezon przypada na okres od lipca do sierpnia oraz marzec. Z kolei w Łodzi najwięcej noclegów – ponadprzeciętnie – udzielono we wrześniu i maju. Liczba udzielonych noclegów w poszczególnych miesiącach jest bardziej zmienna w powiecie poddębickim niż w Łodzi. Aby wyjaśnić sezonowość, należy lepiej poznać cel podróży turystów, ale można przypuszczać, że w powiecie poddębickim mamy do czynienia z turystyką o charakterze wypoczynkowym, a w Łodzi – z turystyką biznesową.

Jeśli chcemy przedstawić dwa zjawiska, które były obserwowane w tym samym czasie, a które mają różne jednostki miary lub różny zakres, można na wykresie dodać jeszcze jedną oś i zaprezentować je wspólnie (rys. 6.1.3).

Przykład 6.1.2

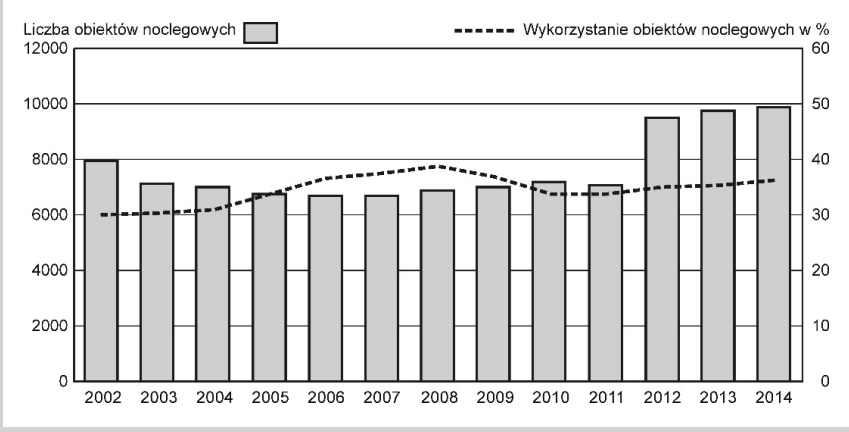
Główny Urząd Statystyczny zbiera i udostępnia dane na temat turystyki w Polsce. Przedstawimy na jednym wykresie liczbę obiektów noclegowych i ich wykorzystanie (w %) w Polsce w latach 2002–2014 (tab. 6.1.2).

Wartości w wierszach tab. 6.1.2 mają inne jednostki miary, z tego powodu należy skorzystać z **wykresu dwuosiowego**. Powinien być tak skonstruowany, aby nie było wątpliwości, które dane są odmierzone na poszczególnych osiach.

Tabela 6.1.2. Liczba obiektów noclegowych i ich wykorzystanie (w %) w Polsce w latach 2002–2014

Rok	Liczba obiektów noclegowych	Wykorzystanie obiektów (w %)
2002	7948	30,3
2003	7116	31,0
2004	6972	33,8
2005	6723	36,5
2006	6694	37,4
2007	6718	38,9
2008	6857	37,0
2009	6992	33,6
2010	7206	33,8
2011	7039	35,0
2012	9483	35,2
2013	9775	36,1
2014	9885	37,0

Źródło: https://stat.gov.pl (dostęp: 26.05.2015).



Rysunek 6.1.3. Obiekty noclegowe w Polsce i ich wykorzystanie (w %) w latach 2002–2014

Źródło: opracowanie własne na podstawie tab. 6.1.2.

Interpretacja. W pierwszych latach badanego okresu (2002–2014) liczba obiektów noclegowych malała z blisko 8 tys. w 2002 r. do 6,7 tys. w 2006 r. Następnie każdego roku sukcesywnie wzrastała o ok. kilkaset tysięcy. Gwałtowny wzrost liczby obiektów noclegowych miał miejsce w 2012 r., kiedy przekroczył 9 tys. Interesująco przedstawia się odsetek wykorzystania obiektów na tle ich liczby. W badanym okresie wahał się od 30% do 39%. Najwyższe wartości (37% i więcej) odnotowano w latach 2006–2008, gdy liczba obiektów była najniższa (rys. 6.1.3). Następnie w 2009 r. spadła do wartości 33,6%, ale później systematycznie wzrastała i w 2014 r. osiągnęła 37%.

6.2. WSKAŹNIKI DYNAMIKI

Po narysowaniu wykresów następny etap analizy dynamiki obejmuje proste miary statystyczne, takie jak: przyrost absolutny (bezwzględny), przyrost względny, wskaźniki łańcuchowe, agregatowe indeksy dynamiki, średnia geometryczna.

Przyrost absolutny, zwany też bezwzględnym, to różnica między wielkością zjawiska w okresie badanym i wielkością zjawiska w okresie poprzednim:

$$\Delta_1 = z_2 - z_1, \quad \Delta_2 = z_3 - z_2, \ldots, \quad \Delta_{n-1} = z_n - z_{n-1}.$$

Może on przyjmować wartości dodatnie, ujemne lub zero. Miara ta jest najprostszą miarą dynamiki. Jest ona wielkością mianowaną, wyrażaną w tych samych jednostkach, co badane zjawisko i obrazuje, o ile zmienił się poziom badanego zjawiska w okresie badanym w porównaniu do okresu podstawowego. Najczęściej wyznacza się przyrosty absolutne w stosunku do okresu poprzedniego, gdyż jesteśmy zainteresowani zróżnicowaniem zjawisk w następujących po sobie przedziałach czasowych. Można jednak założyć, że okresem podstawowym będzie pierwszy rok badawczy lub inny rok szczególnie istotny dla badań.

Przyrost względny jest stosunkiem przyrostu absolutnego do wartości z poprzedniego okresu:

$$(z_2 - z_1)/z_1, \quad (z_3 - z_2)/ z_2, \ldots, \quad (z_n - z_{n-1})/z_{n-1}.$$

Jest to wielkość niemianowana – dzięki temu umożliwia przeprowadzenie porównania i analizę dynamiki zjawisk o różnych miarach.

Tempo wzrostu to przyrost względny pomnożony przez 100%. Tempo wzrostu określa wyrażoną w procentach wielkość przyrostu (spadku) badanego zjawiska w okresie badanym w stosunku do okresu poprzedniego.

Przyrost absolutny, względny i tempo wzrostu mogą przyjmować wartości dodatnie, ujemne i zero:

- jeśli przyjmują wartości dodatnie, to znaczy, że poziom zjawiska wzrastał w stosunku do okresu poprzedniego,
- jeśli przyjmują wartości ujemne, to znaczy, że poziom zjawiska malał w stosunku do okresu poprzedniego,
- jeśli przyjmują wartość 0, to znaczy, że poziom zjawiska nie zmienił się w stosunku do okresu poprzedniego.

Wartość bezwzględna poszczególnych wskaźników mówi o sile wzrostu lub spadku.

Przykład 6.2.1

Na podstawie informacji z *Rocznika statystycznego* GUS: „Noclegi udzielone turystom z Włoch w obiektach zbiorowego zakwaterowania w Polsce w 2008 r." wyznaczamy przyrost absolutny, przyrosty względne i tempo wzrostu (dane w tab. 6.2.1).

Tabela 6.2.1. Noclegi udzielone turystom z Włoch w obiektach zbiorowego zakwaterowania w Polsce w 2008 r.

Miesiąc	Liczba osób	Przyrost absolutny (osoby) $z_n - z_{n-1}$	Przyrost względny $(z_n - z_{n-1})/z_{n-1}$	Tempo wzrostu (%)
Styczeń	33 640	–	–	–
Luty	28 359	–5 281	–0,15699	–15,70
Marzec	32 329	3 970	0,13999	14,00
Kwiecień	41 362	9 033	0,27941	27,94
Maj	36 613	–4 749	–0,11482	–11,48
Czerwiec	35 401	–1 212	–0,03310	–3,31
Lipiec	39 547	4 146	0,11712	11,71

Tabela 6.2.1 cd.

Miesiąc	Liczba osób	Przyrost absolutny (osoby) $z_n - z_{n-1}$	Przyrost względny $(z_n - z_{n-1})/z_{n-1}$	Tempo wzrostu (%)
Sierpień	64 967	25 420	0,64278	64,28
Wrzesień	37 188	–27 779	–0,42759	–42,76
Październik	29 464	–7 724	–0,20770	–20,77
Listopad	25 950	–3 514	–0,11926	–11,93
Grudzień	22 294	–3 656	–0,14089	–14,09
Ogółem	427 114	×	×	×

Źródło: opracowanie własne na postawie danych GUS, https://stat.gov.pl (dostęp: 26.05.2015).

Interpretacja. W 2008 r. udzielono 427 114 noclegów turystom z Włoch. Średnio były to 35 593 noclegi w miesiącu, a najwięcej przypadało na sierpień – blisko 65 tys. Przyrost absolutny (tab. 6.2.1) oznacza, że tylko w ciągu czterech miesięcy (marzec–kwiecień i lipiec–sierpień) liczba noclegów wrosła w stosunku do miesiąca poprzedniego od około 4 tys. w marcu, 9 tys. w kwietniu, 4 tys. w lipcu do 25,4 tys. w sierpniu. W pozostałych miesiącach przyrost naturalny był ujemny, a największy spadek udzielonych noclegów zanotowano we wrześniu – na poziomie 27,8 tys. Tempo wzrostu ukazuje to zjawisko w procentach, dzięki czemu można je porównywać z innymi danymi, np. noclegami udzielonymi turystom z innych krajów. Tempo wzrostu w sierpniu wynosiło aż 64% w stosunku do lipca, a największy spadek, aż 43%, odnotowano we wrześniu. Można to tłumaczyć okresem wakacyjnym i urlopowym we Włoszech, ale należałoby ten fakt jeszcze potwierdzić, sprawdzając rozkład roku szkolnego w tym kraju.

W analizie dynamiki stosuje się dwa typy **wskaźników łańcuchowych** (zwanych również **indeksami**) o stałej lub zmiennej podstawie.

Wskaźniki łańcuchowe jednopodstawowe o **stałej podstawie** definiuje się następująco:

$$I_2 = z_2/z_1 \cdot 100\%, \quad I_3 = z_3/z_1 \cdot 100\%, \ldots, \quad I_n = z_n/z_1 \cdot 100\%.$$

Wybór okresu podstawowego (w tym wypadku z_1) zależy od celu badania. Wybieramy np. okres charakterystyczny lub zwrotny w przebiegu zjawiska. Niewłaściwy wybór podstawy może spowodować zniekształcenie obrazu i wywołać błędne wrażenie. Podstawę może też stanowić średnia arytmetyczna szeregu.

Wskaźniki łańcuchowe o **zmiennej podstawie** są to ilorazy wyrazu następnego do poprzedniego:

$$I_2 = z_2/z_1 \cdot 100\%, \quad I_3 = z_3/ z_2 \cdot 100\%, \ldots, \quad I_n = z_n/z_{n-1} \cdot 100\%.$$

Wskaźniki łańcuchowe przyjmują zawsze wartości dodatnie, a ich interpretacja zależy od tego, czy ich wartość odchyla się od 100:

- jeżeli wartości są większe od 100%, oznacza to wzrost zjawiska,
- jeżeli wartości są mniejsze od 100%, oznacza to spadek zjawiska,
- jeżeli wartości są równe 100%, oznacza to brak zmian w obrębie zjawiska.

Przykład 6.2.2

Obliczymy indeksy o stałej podstawie (2006 r.) oraz o zmiennej podstawie przeciętnej ceny śniadania w hotelu „Inez" w Ustce w styczniu w ciągu sześciu lat 2006–2012.

Tabela 6.2.2. Ceny śniadania w hotelu „Inez" w Ustce w latach 2006–2012

Rok	Przeciętna cena w zł (w styczniu)	Indeksy o stałej podstawie $I_n = z_n/z_1$	Indeksy o zmiennej podstawie $I_n = z_n/z_{n-1}$
2006	10	×	×
2007	20	200	200,0
2008	40	400	200,0
2009	35	350	87,5
2010	74	740	211,0
2011	64	640	86,5
2012	76	760	118,8

Źródło: opracowanie własne.

Interpretacja. Za rok podstawowy przyjęto pierwszy rok funkcjonowania hotelu. Wszystkie wartości w kolumnie 3 są większe od 100, co znaczy, że cena śniadania ciągle rosła w stosunku do roku 2006, różna była jednak jej dynamika wzrostu – największa w latach 2010 i 2012. Dużo lepiej siłę dynamiki oddają indeksy o podstawie zmiennej, w której podstawą jest zawsze rok poprzedni. Widać, że cena śniadania w pierwszych dwóch latach podwoiła się w stosunku do roku poprzedniego. Następnie w 2009 r. cena spadła o 12,5% i stanowiła 87,5% ceny z roku 2008. Kolejne lata wskazują na wahania ceny – w 2010 r. znów wzrosła ponaddwukrotnie, a następnie spadła o 13,5%. Ponowny wzrost nie był już tak duży – w 2012 r. cena zwiększyła się w stosunku do 2011 r. o 18,8%.

Głębsza analiza dynamiki zjawiska wymaga też wyznaczenia **średniego tempa dynamiki** tych zmian przypadających na jednostkę czasu. Jego wartość oblicza się ze wzoru na średnią geometryczną, która należy do średnich klasycznych (por. rozdział 3); jest to pierwiastek stopnia $n - 1$ z iloczynu wskaźników łańcuchowych.

$$\overline{Z}_g = \sqrt[n-1]{\prod_{i=1}^{n} \frac{Z_i}{Z_{i-1}}}$$

Przykład 6.2.3

Obliczamy średnie tempo dynamiki zmian ceny śniadania w hotelu „Inez" w latach 2006–2012 (dane w tab. 6.2.2).

Aby je obliczyć, korzystamy ze wzoru na średnią geometryczną, obliczając pierwiastek stopnia 6 z iloczynu indeksu o podstawie zmiennej (kolumna 4 w tab. 6.2.2).

$$\overline{Z}_g = \sqrt[n-1]{\prod_{i=1}^{n} \frac{Z_i}{Z_{i-1}}} = \sqrt[6]{200 \cdot 200 \cdot 87{,}5 \cdot 211 \cdot 86{,}5 \cdot 118{,}8} = 140{,}2.$$

Interpretacja. Przeciętna cena śniadania w hotelu „Inez" w latach 2006–2012 wahała się między 10 i 76 zł i wynosiła 37 zł. Średnie tempo wzrostu ceny śniadania w hotelu „Inez" w latach 2006–2012 wynosiło 40,2%, co znaczy, że z roku na rok cena śniadania w hotelu „Inez" w latach 2006–2012 rosła przeciętnie o 40,2%.

Agregatowe indeksy dynamiki można wykorzystać do analizy pewnego zespołu (agregatu) cech, np. cen, kosztów utrzymania i innych. Jeśli założymy, że turysta podczas podróży wydaje określoną sumę pieniędzy na transport, wyżywienie, noclegi, pamiątki, bilety do muzeów, bilety do obiektów sportowych, to można poddać analizie dynamicznej poszczególne wydatki lub zespół określonych wydatków. Zespół wydatków można porównywać z wydatkami na grupy produktów w okresie podstawowym lub w roku poprzednim.

Agregatowy indeks wartości zespołu wydatków (produktów) jest ilorazem sumy iloczynów liczby produktów i ich cen do sumy iloczynów liczby produktów i ich cen w okresie podstawowym oraz wyrażany jest w procentach:

$$I_w = \frac{\sum_{j=1}^{m} p_{ij} \cdot q_{ij}}{\sum_{j=1}^{m} p_{0j} \cdot q_{0j}} \cdot 100$$

gdzie:

m – liczba produktów, n – liczba okresów,
p_{ij} – cena j-tego produktu (j = 1, ..., m) w i-tym okresie badań (i = 1, ..., n),
p_{0j} – cena j-tego produktu (j = 1, ..., m) w okresie podstawowym,
q_{ij} – liczba j-tego produktu (j = 1, ..., m) w i-tym okresie badań (i = 1, ..., n),
q_{0j} – liczba j-tego produktu (j = 1, ..., m) w okresie podstawowym.

Na wzrost agregatowego indeksu wartości zespołu wydatków mogły mieć wpływ zmiany cen zagregowanych produktów. Dlatego kolejnym krokiem analizy może być określenie zmian cen agregatu za pomocą indeksu cen Laspeyresa lub Paaschego. Pierwszy jest stosowany, gdy liczba produktów jest różna, ale różnice te nie są znaczne:

$$I_C^L = \frac{\sum_{j=1}^{m} p_{ij} \cdot q_{ij}}{\sum_{j=1}^{m} p_{0j} \cdot q_{0j}} \cdot 100,$$

gdzie:

m – liczba produktów, n – liczba okresów,
p_{ij} – cena j-tego produktu (j = 1, ..., m) w i-tym okresie badań (i = 1, ..., n),
p_{0j} – cena j-tego produktu (j = 1, ..., m) w okresie podstawowym,
q_{0j} – liczba j-tego produktu (j = 1, ..., m) w okresie podstawowym.

Indeks cen Paaschego:

$$I_C^P = \frac{\sum_{j=1}^{m} p_{ij} \cdot q_{ij}}{\sum_{j=1}^{m} p_{0j} \cdot q_{0j}} \cdot 100,$$

gdzie:

m – liczba produktów, n – liczba okresów,

p_{ij} – cena j-tego produktu (j = 1, ..., m) w i-tym okresie badań (i = 1, ..., n),

p_{0j} – cena j-tego produktu (j = 1, ..., m) w okresie podstawowym,

q_{ij} – liczba j-tego produktu (j = 1, ..., m) w i-tym okresie badań (i = 1, ..., n).

Należy pamiętać, że wartości indeksu cen Laspeyresa i Paaschego są różne i nie powinny być porównywane, czyli $I_C^L \neq I_C^P$.

Przykład 6.2.4

Przeprowadzimy analizę kosztów poniesionych na pamiątki rodziny, która od kilku lat spędza urlop w lipcu w Helu (dane w tab. 6.2.3).

Tabela 6.2.3. Wykaz wydatków (w zł) poniesionych na pamiątki przez rodzinę Nowakowskich w latach 2011–2013 podczas pobytu w Helu

Lp.	Produkty j = 1, ..., 10	Jednostka miary	2011		2012		2013	
			liczba q_{0j}	cena p_{0j}	liczba q_{1j}	cena p_{1j}	liczba q_{2j}	cena p_{2j}
1	Widokówki	szt.	4	0,65	7	0,65	8	0,8
2	Znaczki	szt.	4	1	7	1,2	8	1,7
3	Pluszowe foki	szt.	3	25	4	20	5	35
4	Bransoletki z bursztynów	szt.	3	30	2	30	2	35
5	Magnesy na lodówkę	szt.	3	5	4	5	8	5
6	Dzwonki z napisem Hel	szt.	1	15	2	20	3	20
7	Czapki	szt.	2	20	2	25	4	25
8	Podkoszulki	szt.	4	25	4	25	4	40
9	Breloki	szt.	2	5	1	10	4	15
10	Słodycze	kg	1,5	25	1	20	3	25

Produkty $j = 1, \ldots, 10$	2011	2012	2013	2011 /2013	2011 /2013	2011 /2012	2012 /2013
	$q_{0j} \cdot p_{0j}$	$q_{1j} \cdot p_{1j}$	$q_{2j} \cdot p_{2j}$	$q_{0j} \cdot p_{2j}$	$q_{2j} \cdot p_{0j}$	$q_{0j} \cdot p_{1j}$	$q_{1j} \cdot p_{2j}$
1	2,6	4,55	6,4	3,2	5,2	2,6	5,6
2	4,0	8,40	13,6	6,8	8,0	4,8	11,9
3	75,0	80,00	175,0	105,0	125,0	60,0	140,0
4	90,0	60,00	70,0	105,0	60,0	90,0	70,0
5	15,0	20,00	40,0	15,0	40,0	15,0	20,0
6	15,0	40,00	60,0	20,0	45,0	20,0	40,0
7	40,0	50,00	100,0	50,0	80,0	50,0	50,0
8	100,0	100,00	160,0	160,0	100,0	100,0	160,0
9	10,0	10,00	60,0	30,0	20,0	20,0	15,0
10	37,5	20,00	75,0	37,5	75,0	30,0	25,0
Suma	389,1	392,95	760,0	532,5	558,2	392,4	537,5

Źródło: opracowanie własne.

$$I_{(2012/2011)} = \frac{\sum_{j=1}^{10} p_{1j} \cdot q_{1j}}{\sum_{j=1}^{10} p_{0j} \cdot q_{0j}} \cdot 100 = 100{,}99\%, \quad I_{(2013/2012)} = \frac{\sum_{j=1}^{10} p_{2j} \cdot q_{2j}}{\sum_{j=1}^{10} p_{1j} \cdot q_{1j}} \cdot 100 = 193{,}41\%,$$

$$I_{(2013/2011)} = \frac{\sum_{j=1}^{10} p_{2j} \cdot q_{2j}}{\sum_{j=1}^{10} p_{0j} \cdot q_{0j}} \cdot 100 = 195{,}32\%, \quad I^{L}_{C(2013/20111)} = \frac{\sum_{j=1}^{10} p_{ij} \cdot q_{0j}}{\sum_{j=1}^{10} p_{0j} \cdot q_{0j}} \cdot 100 = 136{,}85\%,$$

$$I^{P}_{C(2013/2011)} = \frac{\sum_{j=1}^{10} p_{ij} \cdot q_{ij}}{\sum_{j=1}^{10} p_{0j} \cdot q_{0j}} \cdot 100 = 136{,}15\%,$$

Interpretacja. Wydatki poniesione na pamiątki przez rodzinę Nowakowskich w 2012 r. wzrosły w porównaniu z rokiem 2011 jedynie o 100,99%[1], w 2013 r. zwiększyły się w stosunku do wydatków z 2012 r. o 193,41%, a w 2013 r. w porównaniu z 2011 r. o 195,32%. Indeksy cen produktów kupowanych przez tę rodzinę wzrosły w ciągu dwóch lat: $I^{L}_{C} = 136{,}85\%$, a $I^{P}_{C} = 136{,}15\%$.

1 W interpretacji pamiętamy, że wartości powyżej 100% oznaczają wzrost, a poniżej 100% – spadek wydatków.

6.3. WYZNACZANIE TENDENCJI ROZWOJOWYCH

Wyznaczone w poprzednim podrozdziale wskaźniki i indeksy pozwoliły na szczegółową charakterystykę badanego okresu, obserwację kierunku i siły zmian w poszczególnych podokresach. Nie dały jednak podstaw do uogólnień (poza średnim tempem dynamiki) oraz wyznaczenia tendencji. Analiza szeregów chronologicznych obejmuje też wyznaczenie kierunku rozwoju zjawiska, tzw. trendu, i prognozowanie. W przypadku trendu liniowego prognozę można zapisać jako funkcję liniową, której dziedzinę rozszerza się o okres prognozowania. W przykładzie 6.3.1 dla funkcji liniowej oraz wielomianowej prognozowano na dwa lata do przodu, czyli dla roku 2016 (n = 15). Linie trendu na wykresach są wydłużone o dwie jednostki czasu (rys. 6.3.1). Można z nich odczytać prognozowane wartości lub obliczyć je, wstawiając do wzorów odpowiednią wartość t (w tym przypadku będzie to 15).

O ile dokładność wartości prognozowanych na podstawie funkcji $\hat{y} = f(t)$ dla lat 2002–2014 można sprawdzić z wartościami rzeczywistymi (tab. 6.3.1), to trafność przewidywań na podstawie linii trendu możemy stwierdzić dopiero na koniec 2016 r. – taką metodę nazywamy ***ex post***. Wartości przewidywane dla 2016 r. w zależności od wyboru linii trendu różnią się znacznie i wynoszą: $\hat{y}_{15}$ = 9282 lub $\hat{y}_{15}$ = 11 300. W rzeczywistości w 2016 r. według danych GUS było ich 10 509.

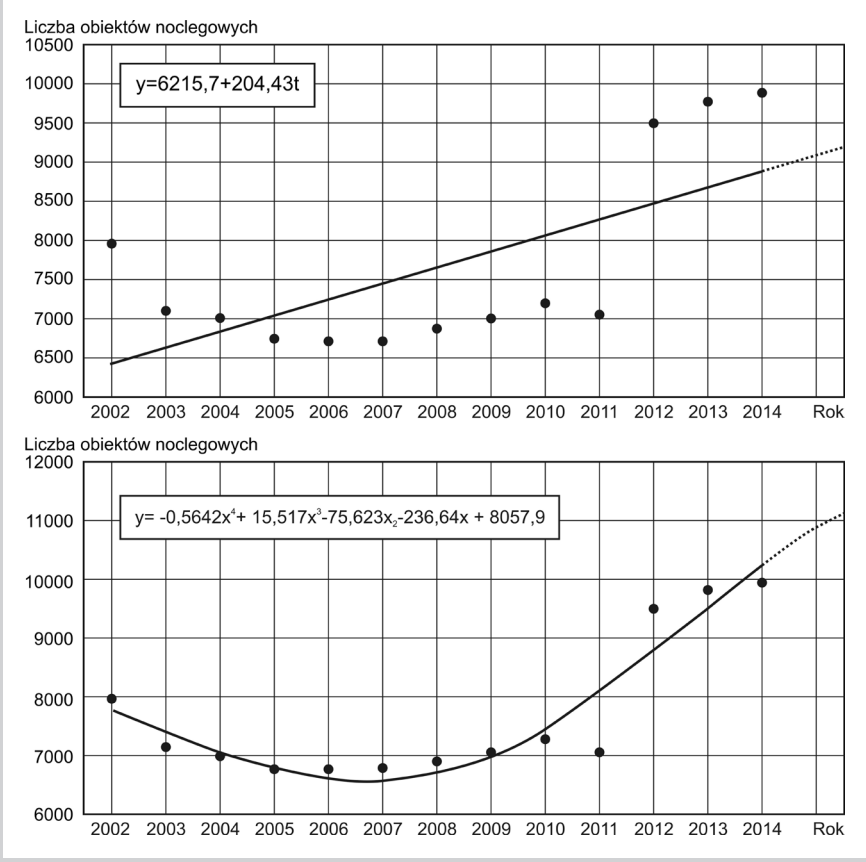
Tabela 6.3.1. Wartości prognozowane liczby obiektów noclegowych w Polsce (2005 i 2016 r.)

Wartość y_4	**2005**	**2016**
	Wartość $\hat{y} = 204{,}43\,x + 6215{,}7$	
	estymowana	
6723	$\hat{y}_4$ = 7033	$\hat{y}_{15}$ = 9282
	Wartość $y = -0{,}5642\,x^4 + 15{,}517x^3 - 75{,}623\,x^2 - 236{,}64\,x + 8057{,}9$	
6723	$\hat{y}_4$ = 6750	$\hat{y}_{15}$ = 11 300

Źródło: opracowanie własne na podstawie tab. 6.1.2.

Przykład 6.3.1

Narysujmy wykres przedstawiający liczbę obiektów noclegowych w Polsce w latach 2002–2014 wraz z **liniami trendu** liniową i wielomianową oraz prognozą na dwa lata.



Rysunek 6.3.1. Liczba obiektów noclegowych w Polsce w latach 2002–2014 wraz z liniami trendu i prognozą na dwa lata

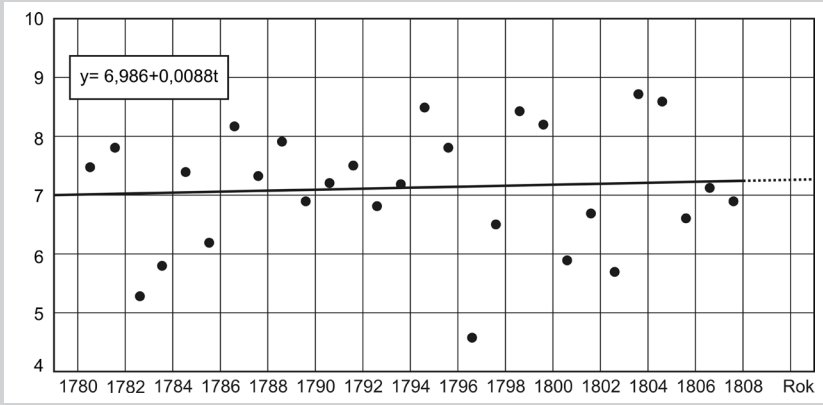
Źródło: opracowanie własne na podstawie tab. 6.1.2.

Przykład 6.3.2

Narysujemy wykres przedstawiający średnioroczną temperaturę (°C) w Warszawie w latach 1780–1810 oraz linię trendu z prognozą na 5 lat.

Średnie roczne temperatury (°C) w Warszawie w latach 1780–1800 były następujące: 7,5; 8,6; 7,3; 8,6; 7,8; 5,3; 5,8; 7,4; 6,2; 8,2; 7,3; 7,9; 6,9; 7,2; 7,5; 6,8; 7,2; 8,5; 7,8; 4,6; 6,5; 8,4; 8,2; 5,9; 6,7; 5,7; 8,7; 8,6; 6,6; 7,1; 6,9.

W tym przypadku lepiej prezentuje zjawisko wykres punktowy zamiast liniowego.



Rysunek 6.3.2. Trend zmian średniorocznej temperatury w Warszawie w latach 1780–1815

Źródło: opracowanie własne.

Interpretacja. Średnioroczne temperatury w Warszawie na przełomie XIX i XX w. były bardzo zróżnicowane. Linia trendu opisana jako funkcja liniowa $\hat{y} = 6{,}986 + 0{,}0088t$ wskazuje na przyrost średniej temperatury o 0,0088 °C. W roku 1815 przewiduje się średnioroczną temperaturę w Warszawie na podstawie linii trendu $\hat{y} = 6{,}986 + 0{,}0088 \cdot 36 = 7{,}3$ °C.

Średnia ruchoma jest jedną z metod wyznaczania trendu. Polega ona na zastąpieniu wartości z_i wartościami średnimi. W zależności od długości podstawowego okresu badań można obliczyć średnią trzyokresową, pięcio-okresową itp. Zazwyczaj przyjmuje się nieparzystą liczbę okresów.

Trzyokresową średnią ruchomą oblicza się w następujący sposób:

$$\overline{z}_{r_2} = (z_1 + z_2 + z_3)/3$$
$$\overline{z}_{r_3} = (z_2 + z_3 + z_4)/3$$
$$\dots$$
$$\overline{z}_{r_{n-1}} = (z_{n-2} + z_{n-1} + z_n)/3$$

Pięciookresową średnią ruchomą oblicza się podobnie, biorąc do obliczenia średniej pięć kolejnych okresów. W tym przypadku nie możemy mieć danych dla dwóch pierwszych i ostatnich okresów (por. tab. 6.3.1).

$$\overline{z}_{r_3} = (z_1 + z_2 + z_3 + z_4 + z_5)/5$$
$$\overline{z}_{r_4} = (z_2 + z_3 + z_4 + z_5 + z_6)/5$$
$$\dots$$
$$\overline{z}_{r_{n-2}} = (z_{n-4} + z_{n-3} + z_{n-2} + z_{n-1} + z_n)/5$$

Następnie rysuje się wykres, na którym umieszcza się zarówno dane empiryczne, jak i średnie ruchome. Pozwalają one na określenie tendencji rozwojowych zjawiska.

Przykład 6.3.3

Metodą średnich ruchomych wyznaczymy tendencję rozwojową liczby zwiedzających w Muzeum Kinematografii w Łodzi w latach 1999–2014.

Tabela 6.3.2. Liczba zwiedzających Muzeum Kinematografii w Łodzi w latach 1999–2014

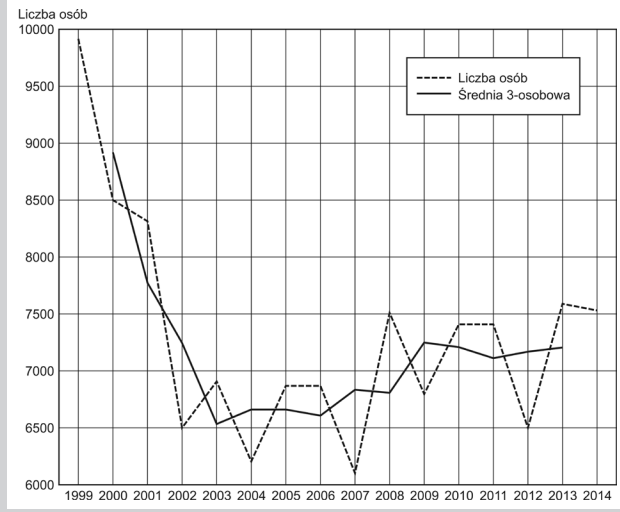
Rok	Liczba osób	Średnia	
		3-okresowa	5-okresowa
1999	9 920	×	×
2000	8 500	8 913,0	×
2001	8 319	7 773,0	8 027,8
2002	6 500	7 239,7	7 283,8
2003	6 900	6 533,3	6 958,4
2004	6 200	6 657,7	6 669,8
2005	6 873	6 649,7	6 589,8

Tabela 6.3.2 cd.

Rok	Liczba osób	Średnia	
		3-okresowa	5-okresowa
2006	6 876	6 616,3	6 713,2
2007	6 100	6 831,0	6 833,2
2008	7 517	6 805,7	6 941,6
2009	6 800	7 244,0	7 048,0
2010	7 415	7 207,7	7 128,0
2011	7 408	7 107,7	7 143,4
2012	6 500	7 167,3	7 290,0
2013	7 594	7 209,0	×
2014	7 533	×	×

Źródło: dane umowne.

Na podstawie danych z tab. 6.3.2 można narysować wykres (rys. 6.3.3), na którym będzie widać, jak średnia trzyokresowa wygładza linię danych empirycznych, wskazując na stabilizację linii odwiedzin w ostatnich latach na poziomie około 7,2 tys. zwiedzających rocznie.



Rysunek 6.3.3. Zwiedzający Muzeum Kinematografii w Łodzi w latach 1999–2014

Źródło: opracowanie własne na podstawie tab. 6.3.1.

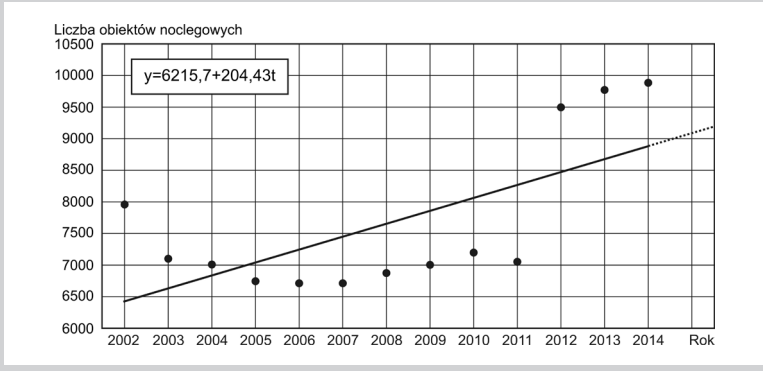
Funkcja trendu jest opisaną matematycznie linią ukazującą zmiany zachodzące w określonym czasie; wyraża się ją wzorem: $\hat{y} = f(t)$. Jest wyznaczana podobnie jak linia regresji, ale w tym przypadku zmienną niezależną jest czas (t). Linia może mieć różne kształty, które są związane z funkcją, jaką reprezentuje: liniową, potęgową, logarytmiczną, wielomianu n-tego stopnia i inne.

Jedną z metod wyznaczania linii trendu jest **metoda najmniejszych kwadratów** (stosowana również w regresji – por. rozdział 4). Polega ona na wykreśleniu funkcji $f(t_i)$ (linii) tak dopasowanej do danych empirycznych, aby suma kwadratów odchyleń poszczególnych wartości empirycznych z_i od wartości funkcji $f(t_i)$ równała się minimum. Algorytm obliczeń nie będzie podany, można posłużyć się odpowiednim oprogramowaniem komputerowym. Wyznaczając linię trendu, należy zwrócić uwagę na jej kształt. Nie zawsze najodpowiedniejsza będzie linia prosta. W niektórych przypadkach bardziej przydatne mogą być inne funkcje, np. wykładnicza, potęgowa, logarytmiczna lub wielomian wyższego rzędu.

Wykreślona linia trendu pozwala nie tylko na analizę zjawiska w badanym okresie, ale również na przewidywanie rozwoju zjawiska (oczywiście należy być bardzo ostrożnym przy przewidywaniu rozmiaru zjawiska w latach następnych, gdyż na jego wartość zazwyczaj ma wpływ wiele czynników).

Przykład 6.3.4

Narysujemy wykres liniowy przedstawiający liczbę obiektów noclegowych w Polsce wraz z linią trendu (dane z tab. 6.1.2).



Rysunek 6.3.4. Liczba obiektów noclegowych w Polsce w latach 2004–2014 wraz z linią trendu (funkcja liniowa)

Źródło: opracowanie własne na podstawie tab. 6.1.2.

Interpretacja. Linia trendu pozwala na ocenę przyrostu wartości w poszczególnych latach i prognozę na dwa lata. Wartość współczynnika kierunkowego prostej wskazuje na to, że każdego roku przybywały około 204,43 obiekty noclegowe w Polsce. Są to wartości przybliżone, a o ich dokładności mogą świadczyć wartości reszt, czyli różnicy $y_i - \hat{y}_i$.

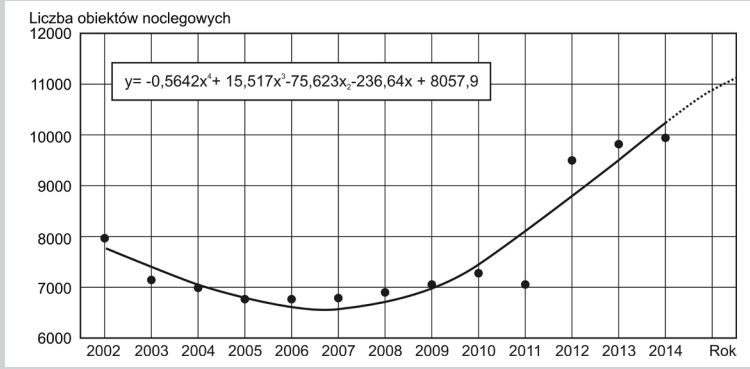
Tabela 6.3.3. Liczba obiektów noclegowych latach 2002–2014 w Polsce, funkcja liniowa $\hat{y} = 6215{,}7 + 204{,}43\,t$

Wartość	Rok						
	2002	2003	2004	2005	2006	2007	2008
y_i	7948	7116	6972	6723	6694	6718	6857
$\hat{y}_i$	6420,1	6624,6	6829,0	7033,4	7237,9	7442,3	7646,7
$y_i - \hat{y}_i$	1527,9	491,4	143,0	–310,4	–543,9	–724,3	–789,7

Wartość	Rok					
	2009	2010	2011	2012	2013	2014
y_i	6992	7206	7039	9483	9775	9885
$\hat{y}_i$	7851,1	8055,6	8260,0	8464,4	8668,9	8873,3
$y_i - \hat{y}_i$	–859,1	–849,6	–1221,0	1018,6	1106,1	1011,7

Źródło: opracowanie własne na podstawie https://stat.gov.pl (dostęp: 26.06.2015).

Wartości szacunkowe zależą od tego, jaką linię najlepiej dopasujemy do danych. Nie jest to proste i wymaga doświadczenia w przeprowadzeniu podobnych badań, zwłaszcza jeśli za wynikami mają iść decyzje, np. inwestycyjne. Porównajmy wyniki dla linii trendu przedstawionej jako linia prosta i wielomian czwartego stopnia.



Rysunek 6.3.5. Liczba obiektów noclegowych w Polsce w latach 2002–2014 wraz z linią trendu (wielomian czwartego stopnia)

Źródło: opracowanie własne na podstawie tab. 6.1.2.

Tabela 6.3.4. Liczba obiektów noclegowych latach 2002–2014 w Polsce, linia trendu: wielomian czwartego stopnia $y = -0{,}5642\,t^4 + 15{,}517\,t^3 - 75{,}623\,t^2 - 236{,}64\,t + 8057{,}9$

Wartość	Rok						
	2002	2003	2004	2005	2006	2007	2008
y_i	7948	7116	6972	6723	6694	6718	6857
$\hat{y}_i$	7760,6	7397,2	7040,6	6750,0	6571,1	6536,1	6663,6
$y_i - \hat{y}_i$	187,4	–281,2	–68,6	–27,0	122,9	181,9	193,4

Wartość	Rok					
	2009	2010	2011	2012	2013	2014
y_i	6992	7206	7039	9483	9775	9885
$\hat{y}_i$	6958,6	7412,9	8004,2	8697,2	9442,6	10178,0
$y_i - \hat{y}_i$	33,4	–206,9	–965,2	785,8	332,4	–293,0

Źródło: opracowanie własne na podstawie https://stat.gov.pl (dostęp:26.06.2015).

Warto porównać, która linia jest lepiej dopasowana, np. dla roku 2005. Wartość przewidywana różni się od rzeczywistej o 110 obiektów dla linii prostej i o 27 dla wielomianu.

6.4. ZADANIA

ZADANIE 6.4.1

Na podstawie danych z tab. 6.4.1. oblicz przyrosty absolutne, względne tempo wzrostu zysków w hotelu „Posejdon" w Gdyni w 2010 r. Przedstaw dane w postaci tabeli.

Tabela 6.4.1. Zyski hotelu „Posejdon" w Gdyni w 2010 r. w poszczególnych miesiącach (w zł)

Lp.	Miesiąc	Wartość zysku (w tys. zł)
1	Styczeń	300
2	Luty	320
3	Marzec	200
4	Kwiecień	250
5	Maj	250

Tabela 6.4.1 cd.

Lp.	Miesiąc	Wartość zysku (w tys. zł)
6	Czerwiec	400
7	Lipiec	490
8	Sierpień	500
9	Wrzesień	450
10	Październik	300
11	Listopad	300
12	Grudzień	350

Źródło: dane umowne.

ZADANIE 6.4.2

Zapytaj rodziców, ile zarabiali brutto w ciągu ostatnich 10 lat (np. według PIT) i oblicz ich średnie zarobki. Wylicz tempo wzrostu lub spadku ich wynagrodzenia.

ZADANIE 6.4.3

Sporządź wykresy, opierając się na danych z tab. 6.4.2 i oceń dynamikę zmian wykorzystania obiektów hotelowych (w %) w Polsce w latach 2002–2014 według kategorii.

Tabela 6.4.2. Wykorzystanie obiektów hotelowych (w %) w Polsce w latach 2002–2014 według kategorii

Kategoria	2002	2003	2004	2005	2006	2007	2008
Pięciogwiazdkowe	38,60	37,40	43,33	53,11	53,02	52,52	48,52
Czterogwiazdkowe	35,10	35,10	38,06	41,65	42,53	44,26	39,73
Trzygwiazdkowe	32,00	32,50	34,65	36,85	37,13	37,96	35,93
Dwugwiazdkowe	28,60	29,30	31,17	33,35	34,19	36,43	35,39
Jednogwiazdkowe	27,90	30,90	35,88	36,93	39,46	39,28	39,38

Kategoria	2009	2010	2011	2012	2013	2014
Pięciogwiazdkowe	45,15	48,63	49,3	52,9	55,8	54,3
Czterogwiazdkowe	36,12	38,53	40,3	40,9	41,4	42,2
Trzygwiazdkowe	32,33	31,46	32,8	32,7	33,2	34,2
Dwugwiazdkowe	32,01	31,44	31,5	30,3	31,1	31,3
Jednogwiazdkowe	35,19	35,55	36,5	33,5	34,6	36,6

Źródło: https://stat.gov.pl (dostęp: 26.06.2015).

ZADANIE 6.4.4

Oblicz przyrost absolutny, względny, tempo wzrostu liczebności kin oraz miejsc na widowni w kinach w Polsce w latach 2003–2014. Narysuj wykres dwuosiowy z danymi. Przeprowadź analizę zmian dynamiki.

Tabela 6.4.3. Kina oraz miejsca na widowni w kinach w Polsce w latach 2003–2014

Rok	Liczba kin	Liczba miejsc na widowni
2003	581	230 817
2004	545	225 454
2005	536	235 248
2006	505	232 471
2007	496	244 174
2008	483	249 533
2009	448	248 181
2010	438	248 029
2011	448	249 390
2012	447	257 849
2013	469	271 781
2014	463	266 479

Źródło: opracowanie własne na podstawie https://stat.gov.pl (dostęp: 26.06.2015).

ZADANIE 6.4.5

Oceń tempo zmian liczebności kin i sal w kinach w Polsce w latach 2003–2014 na podstawie indeksów oraz średniego tempa dynamiki. Narysuj wykres dwuosiowy z danymi.

Tabela 6.4.4. Kina oraz sale w kinach w Polsce w latach 2003–2014

Rok	Liczba kin	Liczba sal
2003	581	880
2004	545	870
2005	536	937
2006	505	931
2007	496	1 008
2008	483	1 043
2009	448	1 061

Tabela 6.4.4 cd.

Rok	Liczba kin	Liczba sal
2010	438	1 076
2011	448	1 122
2012	447	1 162
2013	469	1 243
2014	463	1 243

Źródło: opracowanie własne na podstawie https://stat.gov.pl (dostęp: 26.06.2015).

ZADANIE 6.4.6

Oblicz, jaki procent stanowiły seanse produkcji polskiej w stosunku do seansów ogółem wyświetlanych w polskich kinach w latach 2003–2014. Narysuj wykres dwuosiowy z danymi: liczba seansów produkcji polskiej i ich udział procentowy (na podstawie tab. 6.4.5). Oblicz tempo wzrostu i przyrost absolutny. Przeprowadź analizę.

Tabela 6.4.5. Seanse produkcji polskiej w stosunku do seansów ogółem wyświetlanych w polskich kinach w latach 2003–2014

Rok	Seanse ogółem	Seanse produkcji polskiej
2003	816 705	78 383
2004	889 868	75 151
2005	947 878	76 553
2006	1 037 138	150 059
2007	1 190 879	169 121
2008	1 326 008	252 144
2009	1 416 677	267 127
2010	1 476 905	198 069
2011	1 569 062	323 068
2012	1 565 688	224 748
2013	1 645 637	294 732
2014	1 756 954	296 560

Źródło: opracowanie własne na podstawie https://stat.gov.pl (dostęp: 26.06.2015).

ZADANIE 6.4.7

Na podstawie danych z tab. 6.1.1 sprawdź, czy występuje sezonowość udzielonych noclegów w województwie łódzkim w trzech pierwszych kwartałach 2011 r. i porównaj ją z przykładem 6.1.1.

ZADANIE 6.4.8

Przedstaw analizę dynamiki liczby seansów wyświetlanych w kinach w Polsce w latach 2003–2014 (tab. 6.4.5). Wykreśl linię trendu na trzy lata naprzód.

ZADANIE 6.4.9

Oblicz indeksy o stałej i zmiennej podstawie dla liczby kin i multipleksów w Polsce w latach 2003–2014 (tab. 6.4.6). Podaj interpretację wyników.

Tabela 6.4.6. Liczba kin i multipleksów w Polsce w latach 2003–2014

Rok	Liczba multipleksów	Liczba kin
2003	25	581
2004	28	545
2005	33	536
2006	34	505
2007	41	496
2008	44	483
2009	48	448
2010	48	438
2011	51	448
2012	51	447
2013	55	469
2014	54	463

Źródło: opracowanie własne na podstawie https://stat.gov.pl (dostęp: 26.06.2015).

ZADANIE 6.4.10

W Muzeum Archeologicznym w Atenach zestawiono liczbę sprzedanych biletów w ciągu pięciu lat. Oceń dynamikę odwiedzin tego muzeum. Sprawdź sezonowość w poszczególnych latach. Oblicz sumę sprzedanych biletów

w poszczególnych latach i na jej podstawie wylicz odpowiednie wskaźniki. Narysuj krzywą trendu i oceń, czy występują jakieś prawidłowości, podaj prognozę na pięć lat (na podstawie tab. 6.4.7).

Tabela 6.4.7. Sprzedaż biletów w Muzeum Archeologicznym w Atenach w latach 2010–2015

Miesiąc	Liczba sprzedanych biletów w tys. w roku					
	2010	2011	2012	2013	2014	2015
Styczeń	5	7	7	8	9	10
Luty	7	9	10	11	12	13
Marzec	8	10	11	12	16	17
Kwiecień	12	14	15	16	17	18
Maj	20	23	25	26	27	28
Czerwiec	30	35	36	37	37	37
Lipiec	50	52	53	54	58	59
Sierpień	49	51	52	54	57	60
Wrzesień	36	38	40	41	42	43
Październik	20	21	21	23	24	25
Listopad	10	10	10	11	11	12
Grudzień	9	9	8	9	10	10

Źródło: dane umowne.

ZADANIE 6.4.11

Oblicz przyrost bezwzględny i tempo wzrostu liczby noclegów udzielonych turystom zagranicznym w 2008 r. według krajów (tab. 6.4.8) i porównaj wyniki z obliczeniami z przykładu 6.2.1.

Tabela 6.4.8. Noclegi udzielone turystom zagranicznym w obiektach zbiorowego zakwaterowania w Polsce w 2008 r.

Miesiąc	Austria	Republika Czeska	Rumunia	Słowacja	Słowenia	Szwajcaria	Wielka Brytania
Styczeń	1 714	8 680	4 007	3 876	936	2 680	48 103
Luty	1 515	9 449	4 130	4 451	1 014	3 460	78 786
Marzec	1 695	11 907	6 503	4 611	910	4 110	79 644
Kwiecień	2 389	13 797	4 889	5 965	1 781	5 121	78 302

Miesiąc	Austria	Republika Czeska	Rumunia	Słowacja	Słowenia	Szwajcaria	Wielka Brytania
Maj	4 556	17 487	5 498	7 377	1 662	5 684	92 034
Czerwiec	4 521	14 699	4 787	6 949	1 399	7 047	85 819
Lipiec	6 659	17 652	6 761	6 298	1 138	10 758	88 924
Sierpień	5 080	16 116	5 587	6 138	1 795	8 780	86 332
Wrzesień	5 166	18 236	5 902	7 623	1 219	7 355	87 941
Październik	3 143	16 601	5 305	8 795	1 626	5 282	77 043
Listopad	1 358	15 176	4 093	7 185	1 171	3 291	47 198
Grudzień	2 515	10 884	3 110	5 248	1 239	2 987	40 037
Ogółem	40 311	170 684	60 572	74 516	15 890	66 555	890 163

Źródło: opracowanie własne na postawie danych GUS, https://stat.gov.pl (dostęp: 26.06.2015).

ZADANIE 6.4.12

Oblicz średnią trzyokresową średnich rocznych temperatur w Krakowie (tab. 6.4.9) i przedstaw ją na wykresie.

Tabela 6.4.9. Średnie roczne temperatury powietrza w Krakowie w °C w latach 1988–2001

Rok	1988	1989	1990	1991	1992	1993	1994	1995	1996	1997	1998	1999	2000	2001
Temperatura w °C	8,1	8,8	8,2	8,6	7,5	7,9	8,0	7,4	9,2	7,8	8,6	8,1	8,3	8,9

Źródło: dane umowne.

ZADANIE 6.4.13

Oblicz średnią pięciookresową wykorzystania obiektów noclegowych (w %) w Polsce w latach 2002–2014 (tab. 6.1.2) i przedstaw ją na wykresie.

ZADANIE 6.4.14

Przedstaw jak najbardziej wszechstronną analizę przyjazdów turystów do Japonii w latach 2007–2012 według kontynentów. Więcej danych na stronie japońskiej: http://www.stat.go.jp.

Tabela 6.4.10. Turyści przyjeżdżający do Japonii w latach 2007–2012 według kontynentów

Kontynent	2007	2008	2009	2010	2011	2012
Afryka	23 408	24 498	20 621	22 665	19 361	24 725
Ameryka Płn.	1 017 018	967 125	874 617	905 896	685 046	876 401
Ameryka Płd.	37 001	38 567	33 481	39 481	31 762	51 151
Azja	6 130 283	6 153 827	4 814 001	6 528 432	4 723 661	6 387 977
Europa	877 531	886 723	800 085	853 166	569 279	775 840
Oceania	260 788	278 988	246 213	260 872	189 150	241 513
Ogółem	8 346 969	8 350 835	6 789 658	8 611 175	6 218 752	8 358 105

Źródło: Japan National Tourism Organization.

ZADANIE 6.4.15

Na podstawie tab. 6.4.11 odpowiedz na pytania:

1. Z których trzech krajów azjatyckich przyjechało w latach 2010–2012 najwięcej turystów do Japonii? Ile procent wszystkich turystów stanowili?
2. Jaką część przyjazdów turystycznych stanowi turystyka biznesowa w badanych krajach? (wyniki przedstaw w postaci tabeli).
3. Czy tempo wzrostu liczby turystów jest zawsze dodatnie w poszczególnych krajach? (wyniki przedstaw w postaci tabeli).
4. W nazwach państw jest pisownia oryginalna, przedstaw wyniki w wersji polskiej.

Tabela 6.4.11. Turyści napływający do Japonii z Azji w latach 2010–2012 według celu podróży

Narodowość/kraj	Cel podróży								
	turystyczny			biznesowy			inny		
	2010	2011	2012	2010	2011	2012	2010	2011	2012
Israel	8 970	2 581	5 380	4 548	3 792	4 456	671	558	577
India	20 929	12 211	19 096	28 917	27 094	32 297	16 973	20 049	17 521
Indonesia	53 195	33 954	68 211	12 943	13 262	17 445	14 494	14 695	15 804
South Korea	1 963 002	1 199 020	1 569 278	334 592	306 757	332 132	142 222	152 296	141 365
Singapore	151 580	86 034	112 842	26 590	22 227	26 486	2 790	3 093	2 873
Thailand	165 901	95 185	201 623	30 661	27 886	36 459	18 319	21 898	22 558

Narodowość/kraj	Cel podróży								
	turystyczny			biznesowy			inny		
	2010	2011	2012	2010	2011	2012	2010	2011	2012
China	831 652	453 182	829 206	230 597	195 209	236 268	350 626	394 855	359 626
China (Taiwan)	1 139 339	868 010	1 329 331	95 159	90 712	102 158	33 780	35 252	34 264
China (Hong Kong)	473 031	333 773	447 486	30 949	26 708	30 112	4 711	4 384	4 067
Philippines	43 298	29 832	48 735	14 165	12 867	15 470	19 914	20 400	20 832
Vietnam	13 224	8 741	15 523	9 031	8757	11 321	19 607	23 550	28 312
Malaysia	80 308	50 312	95 030	24 974	20 824	24 975	9 237	10 380	10 178
Others	51 989	34 496	53 923	22 069	20 032	24 017	33 475	38 793	40 740

Źródło: http://www.stat.go.jp (dostęp: 26.09.2015).

ZADANIE 6.4.16

Wybierz trzy kraje europejskie (z tab. 6.4.12) i porównaj dynamikę zmian liczby obiektów hotelowych w latach 2005–2014. W tabeli pisownia nazw krajów jest oryginalna, przedstaw wyniki w wersji polskiej.

Tabela 6.4.12. Liczba obiektów hotelowych i podobnych w Unii Europejskiej w latach 2005–2014

Kraj	2005	2006	2007	2008	2009	2010	2011	2012	2013	2014
Belgium	1 899	1 955	2 013	2 009	2 036	2 088	2 091	1 771	1 713	1 653
Bulgaria	1 230	1 348	1 526	1 646	1 784	1 823	1 862	1 936	2 055	2 166
Czech Rep.	4 278	4 314	4 559	4 483	4 469	4 300	4 612	6 350	6 301	5 833
Denmark	482	473	477	470	471	482	519	515	514	533
Germany	36 575	36 201	35 941	35 891	35 814	35 867	35 579	35 215	34 692	33 997
Estonia	317	341	346	368	387	375	374	390	404	410
Ireland	4 407	4 296	4 087	3 947	3 624	3 451	3 071	2 945	2 462	2 438
Greece	9 036	9 111	9 207	9 385	9 559	9 732	9 648	9 665	9 675	10 123
Spain	17 607	18 304	17 827	18 026	18 387	18 635	19 262	19 532	19 610	19 563
France	18 689	18 361	18 135	17 970	17 723	17 290	17 189	17 189	17 171	17 336
Croatia	1 015	762	800	835	819	841	857	878	897	909

Tabela 6.4.12 cd.

Kraj	2005	2006	2007	2008	2009	2010	2011	2012	2013	2014
Italy	33 527	33 768	34 058	34 155	33 967	33 999	33 918	33 728	33 316	33 290
Cyprus	785	753	735	708	699	690	683	799	792	799
Latvia	337	321	318	387	451	495	496	247	255	258
Lithuania	331	338	348	365	380	381	379	397	414	421
Luxembourg	293	277	273	267	261	260	259	251	243	236
Hungary	2 061	2 032	1 999	2 001	2 042	2 033	1 927	2 094	2 064	2 123
Malta	173	173	160	155	158	153	149	150	153	149
Netherlands	3 135	3 099	3 196	3 180	3 151	3 172	3 194	3 155	3 510	3 561
Austria	14 267	14 051	14 204	13 756	13 645	13 461	13 134	13 203	13 073	12 839
Poland	2 200	2 301	2 443	2 642	2 836	3 223	3 285	3 414	3 485	3 646
Portugal	2 012	2 028	2 031	2 041	1 988	2 011	2 019	2 028	2 331	2 331
Romania	3 608	4 125	4 163	4 362	4 566	4 724	4 612	2 216	2 439	2 500
Slovenia	344	358	396	654	667	647	648	642	639	647
Slovakia	885	922	1 249	1 313	1324	1 322	1297	1473	1439	:
Finland	938	923	909	901	867	842	830	839	828	785
Sweden	1 857	1 888	1 893	1 940	1 982	1 985	1 998	2 003	2 045	2 033
United Kingdom	32 926	39 107	39 860	39 024	40 415	40 184	38 940	38 996	40 272	:
Iceland	319	308	294	301	296	343	:	332	344	:
Liechtenstein	46	46	47	45	41	40	40	38	36	40
Norway	1 136	1 119	1 112	1 108	1 122	1 128	1 115	1 102	1 201	1 145
Switzerland	5 836	5 693	5 635	5 582	5 533	5 477	5 396	5 257	5 191	5 129
Montenegro	:	:	:	:	:	:	:	351	:	:
Former Yugoslav Republic of Macedonia	:	:	:	128	149	172	186	209	225	233
Serbia	:	:	:	:	:	:	:	716	657	676
Turkey	:	:	:	:	:	:	:	:	:	:

: Not available/brak danych.

Źródło: http://ec.europa.eu/eurostat/web/tourism/data/database (dostęp: 26.09.2015).

6.5. ODPOWIEDZI DO WYBRANYCH ZADAŃ

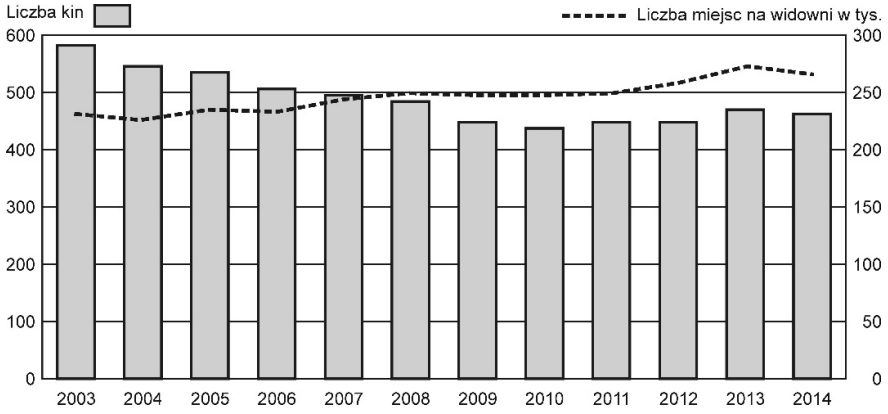
ZADANIE 6.4.1

Tabela 6.5.1. Zyski hotelu „Posejdon" w Gdyni w 2010 r. w poszczególnych miesiącach w zł

Miesiąc	Wartość zysku (tys. zł)	Przyrost absolutny (tys. zł) $z_n - z_{n-1}$	Przyrost względny $(z_n - z_{n-1})/z_{n-1}$	Tempo wzrostu (%)
Styczeń	300	×	×	×
Luty	320	20	0,067	6,67
Marzec	200	–120	–0,375	–37,50
Kwiecień	250	50	0,250	25,00
Maj	250	0	0,000	0,00
Czerwiec	400	150	0,600	60,00
Lipiec	490	90	0,225	22,50
Sierpień	500	10	0,020	2,04
Wrzesień	450	–50	–0,100	–10,00
Październik	300	–150	–0,333	–33,33
Listopad	300	0	0,000	0,00
Grudzień	350	50	0,167	16,67

Źródło: opracowanie własne na podstawie tab. 6.4.3.

ZADANIE 6.4.4



Rysunek 6.5.1. Kina oraz miejsca na widowni w kinach w Polsce w latach 2003–2014

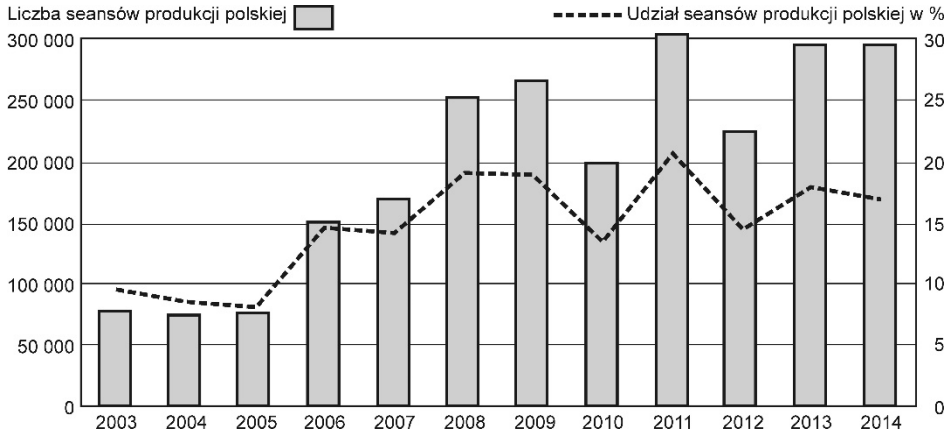
Źródło: opracowanie własne na podstawie tab. 6.4.3.

Tabela 6.5.2. Kina oraz miejsca na widowni w kinach w Polsce w latach 2003–2014

Rok	Liczba kin	Przyrost absolutny	Tempo wzrostu (%)
2003	581	×	×
2004	545	–36	–6,2
2005	536	–9	–1,7
2006	505	–31	–5,8
2007	496	–9	–1,8
2008	483	–13	–2,6
2009	448	–35	–7,2
2010	438	–10	–2,2
2011	448	10	2,3
2012	447	–1	–0,2
2013	469	22	4,9
2014	463	–6	–1,3

Źródło: opracowanie własne na podstawie tab. 6.4.3.

ZADANIE 6.4.6



Rysunek 6.5.2. Seanse produkcji polskiej i ich udział w seansach ogółem wyświetlanych w polskich kinach w latach 2003–2014

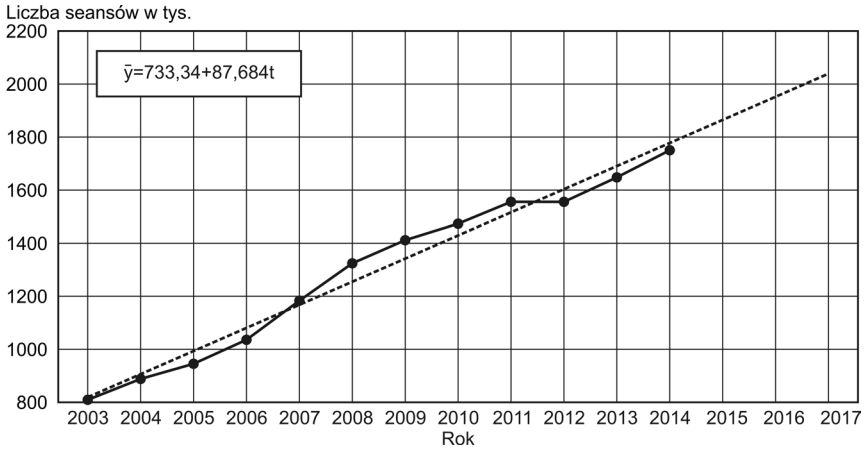
Źródło: opracowanie własne na podstawie tab. 6.4.5.

Tabela 6.5.3. Seanse produkcji polskiej oraz wskaźniki dynamiki

Rok	Liczba seansów produkcji polskiej	Przyrost absolutny	Tempo wzrostu (%)
2003	816 705	×	×
2004	889 868	73 163	9,0
2005	947 878	58 010	6,5
2006	1 037 138	89 260	9,4
2007	1 190 879	153 741	14,8
2008	1 326 008	135 129	11,3
2009	1 416 677	90 669	6,8
2010	1 476 905	60 228	4,3
2011	1 569 062	92 157	6,2
2012	1 565 688	–3 374	–0,2
2013	1 645 637	79 949	5,1
2014	1 756 954	111 317	6,8

Źródło: opracowanie własne na podstawie https://stat.gov.pl (dostęp: 26.05.2015).

ZADANIE 6.4.8



Rysunek 6.5.3. Seanse produkcji polskiej w latach 2003–2014 wraz z linią trendu i prognozą na trzy lata

Źródło: opracowanie własne na podstawie tab. 6.4.5.

Na podstawie 12 obserwacji (lata 2003–2014) przewidywana liczba seansów trzy lata naprzód, czyli w 2017 r., na podstawie linii trendu wynosi 2049 tys. seansów:

$$\hat{y} = 733{,}34 + 87{,}68\, t = 733{,}34 + 87{,}68 \cdot 15 = 2049 \text{ tys.}$$

ZADANIE 6.4.9

Obliczamy indeksy o podstawie stałej i zmiennej dla liczby multipleksów (z_i) i kin (w_i) w Polsce. Za podstawę stałą wybieramy pierwszy rok analizy danych, czyli 2003.

Tabela 6.5.4. Liczba kin i multipleksów w Polsce w latach 2003–2014 oraz indeksy o podstawie stałej i zmiennej

Rok	Liczba		Indeksy o podstawie			
	multipleksów z_i	kin w_i	stałej $I_n = z_n/z_{2003}$	zmiennej $I_n = z_n/z_{n-1}$	stałej $I_n = w_n/w_{2003}$	zmiennej $I_n = w_n/_{n-1}$
2003	25	581	×	×	×	×
2004	28	545	112,0	112,0	93,8	93,8
2005	33	536	132,0	117,9	92,3	98,3
2006	34	505	136,0	103,0	86,9	94,2
2007	41	496	164,0	120,6	85,4	98,2
2008	44	483	176,0	107,3	83,1	97,4
2009	48	448	192,0	109,1	77,1	92,8
2010	48	438	192,0	100,0	75,4	97,8
2011	51	448	204,0	106,3	77,1	102,3
2012	51	447	204,0	100,0	76,9	99,8
2013	55	469	220,0	107,8	80,7	104,9
2014	54	463	216,0	98,2	79,7	98,7

Źródło: opracowanie własne na podstawie tab. 6.4.6.

Interpretacja. Jeśli przyjąć za podstawę stałą liczbę multipleksów i kin w roku 2003, to wartości indeksów (kolumny 4 i 6 w tab. 6.4.16) wskazują na systematyczny wzrost liczby multipleksów, ale spadek liczby kin w Polsce w latach 2003–2014. Wszystkie wartości indeksów o podstawie stałej dla multiplek-

sów są wyższe od 100%, co wskazuje, że we wszystkich latach liczba multipleksów wzrastała w stosunku do liczby z 2003 r. (było ich wówczas 25).

Od roku 2011 do końca okresu badawczego było ponaddwukrotnie więcej multipleksów niż w 2003 r. Z kolei wszystkie wartości indeksów o podstawie stałej dla kin były niższe od 100%, co znaczy, że we wszystkich latach liczba kin była niższa niż w 2003 r. (było ich wówczas 581). Warto zwrócić uwagę, że w 2010 r. nie funkcjonowało już co czwarte kino w Polsce.

Od roku 2011 do końca okresu badawczego zwiększyła się nieznacznie liczba kin, ale nie osiągnęła wartości z 2003 r.

Jeśli uwzględnimy wartości indeksów o podstawie zmiennej (kolumny 5 i 7 w tab. 5.4.16), to również można zauważyć systematyczny wzrost liczby multipleksów i spadek liczby kin w Polsce w latach 2003–2014, jednak nie był on identyczny każdego roku. Największy przyrost liczby multipleksów w stosunku do roku poprzedniego nastąpił w 2010 r. (120,6%). W ciągu pierwszych 10 lat badawczych liczba multipleksów rosła, ale w 2014 r. indeks miał wartość niższą niż 100%, co oznacza, że zmalała (98,2%). Indeksy o zmiennej podstawie dla kin były w większości niższe od 100%, a najniższą wartość osiągnęły w 2008 r. (92,8%), co jednak nie świadczy o tym, że w całym okresie nie przybywało kin w Polsce, gdyż w 2011 r. i 2013 r. wartość indeksów wskazuje na ich przyrost. Nie był on jednak na tyle duży, aby liczba kin osiągnęła stan z 2003 r.

PODSUMOWANIE

PROPOZYCJA ETAPÓW BADANIA STATYSTYCZNEGO

Ze statystyką można spotkać się przy wielu okazjach w życiu codziennym, nie tylko podczas studiów. Z mojego doświadczenia nauczyciela akademickiego wynika, że studenci mają różny stosunek do statystyki. Przedmiot obligatoryjny traktują często jako „zło konieczne" do zaliczenia roku. W trakcie studiów zaczynają korzystać z metod i źródeł statystycznych do opracowania prac zaliczeniowych. Dopiero przygotowując się do pracy licencjackiej lub magisterskiej, kiedy dysponują zebranym przez siebie materiałem do opracowania i danymi wtórnymi, wracają do metod statystycznych oraz przychodzą na konsultacje do specjalisty. Często zbierają informacje chaotycznie, ze słabo sformułowanym problemem badawczym i nie wiedzą, co z nimi począć. Mają za dużo („Zebrałem je tak na wszelki wypadek") lub zbyt mało danych („Szkoda, że nie zapytałem w ankiecie o pewną kwestię"). Dlatego na zakończenie przedstawionych zostanie kilka etapów badania statystycznego, których kolejność wykonywania jest bardzo ważna.

Metody, jakie zostały przedstawione w poprzednich rozdziałach mogą stanowić część badań przeprowadzanych w ramach pracy licencjackiej, magisterskiej lub opracowania raportu na zadany temat z szeroko rozumianej

turystyki (por. rozdział 1). Aby tak się stało, proces badań powinien być podzielony na etapy. Można je uporządkować następująco:

1. Sformułowanie problemu badawczego (podmiot, przedmiot, cele).
2. Przegląd literatury naukowej.
3. Określenie zbiorowości statystycznej, ustalenie zasięgu przestrzennego, czasu badań.
4. Eksplikacja badania, postawienie hipotezy badawczej i określenie celów szczegółowych.
5. Wybór technik zbierania informacji, narzędzi, źródeł danych i metod badawczych.
6. Badania pilotażowe.
7. Dobór próby badawczej.
8. Realizacja badania (badania społeczne, inwentaryzacja terenowa, pozyskiwanie danych z różnych źródeł).
9. Wybór metod statystycznych.
10. Wybór oprogramowania.
11. Przetwarzanie danych.
12. Wstępna analiza danych.
13. Zaawansowana analiza statystyczna.
14. Raport z badań.

Sformułowanie problemu badawczego, który chcemy podjąć to bardzo ważny moment. Od niego zależą pozostałe działania. Jego konstrukcja powinna wyraźnie wskazywać, czym chcemy się zająć. Z jednej strony powinna być uniwersalna (ale nie oczywista), z drugiej – dobrze, aby opisywała i tłumaczyła interesujące dla badacza i czytelników zagadnienia turystyczne. Problem badawczy można przedstawić jako zestaw pytań, na które w trakcie postępowania badacz będzie się starał uzyskać odpowiedzi.

Kiedy mamy już wstępnie zarysowany problem badawczy, przechodzimy do drugiego punktu, czyli przeglądu literatury przedmiotu badań. Literatura naukowa jest bardzo obszerna, nie ma dość czasu, aby wszystko przestudiować, często trudno wybrać z niej ważne pozycje. Dlatego warto zastanowić się nad słowami kluczowymi, jakie można przypisać problematyce, którą za-

mierzamy się zająć, np. ruch turystyczny, opinie turystów, szlak turystyczny, przestrzeń turystyczna, rekreacja. Te określenia w językach polskim i angielskim mogą posłużyć do poszukiwań w bazach artykułów o podobnej tematyce na takich portalach, jak https://scholar.google.pl, http://www.ibuk.pl, http://www.sciencedirect.com, http://www.elsevier.pl, http://repozytorium.uni.lodz.pl. Nie polecam stron, które naruszają prawa autorskie. Warto zapoznać się z tą literaturą z kilku powodów. Po pierwsze, aby mieć szerszy pogląd na zjawisko, które chcemy badać, po drugie – aby poznać różną metodologię badań, po trzecie – by móc porównać własne wyniki z wynikami innych autorów. Literatura naukowa jest bardzo inspirująca dla początkujących badaczy. Często po jej przeglądzie modyfikujemy nasz problem badawczy, aby stał się bardziej interesujący.

W trakcie dwóch pierwszych etapów zastanawiamy się, co lub kogo będziemy badać oraz gdzie i kiedy będą się odbywały badania (podrozdział 1.2). Ale zanim do nich przystąpimy, należy bardzo precyzyjnie to określić. Ich wybór zależy od kilku czynników, m.in. od:

1. Czasu, jaki mamy do dyspozycji, np. do przygotowania pracy zaliczeniowej, licencjackiej, magisterskiej.
2. Kosztów, jakie trzeba ponieść, m.in. dojazdów, noclegów, zakupu źródeł danych.
3. Znajomości problematyki – należy uczciwie sobie odpowiedzieć, ile czasu potrzebujemy na zgłębienie tematu.

Jedną z ważnych kwestii jest odpowiedź na pytanie, czy przeprowadzamy badania całej zbiorowości czy jej fragmentu (podrozdział 1.3). Niedoświadczeni badacze powinni raczej wybierać mniej liczne zbiorowości i badać je kompleksowo, lub liczniejsze, ale takie, z których będzie można pobrać reprezentatywną próbę.

W tym momencie można przystąpić do **eksplikacji** (łac. *explicatio* – tłumaczenie, wyjaśnienie), czyli logicznego uściślenia treści pojęcia bez zmiany jego zakresu. Eksplikacja powinna charakteryzować się ścisłością, naukową użytecznością i prostotą; w logice tradycyjnej funkcję eksplikacji pełnią definicje; potocznie „wyjaśnianie”[1]. W tym momencie badania należy sformułować hipotezy badawcze i cele szczegółowe.

1 https://encyklopedia.pwn.pl/szukaj/eksplikacja.html (dostęp: 13.06.2019).

Kolejny etap stanowi przygotowanie narzędzi badawczych, źródeł danych i metod badawczych bezpośrednio związanych z postawioną hipotezą, sposobem wyboru zbiorowości.

Jeśli to możliwe, należy przeprowadzić badania pilotażowe. Pozwalają one na ocenę narzędzi badawczych, czasu potrzebnego do przeprowadzenia badań i ewentualną korektę technik oraz narzędzi badawczych przed badaniami właściwymi.

Nie wszystkie metody badawcze będą wykorzystywały statystykę, ale w większości przypadków badań jest ona konieczna. Już w trakcie konstruowania pytań szczegółowych, np. w ankietach, warto zastanowić się dłużej nad wyborem cech (zmiennych) do badań. To od ich skali pomiarowej będą zależały przyszłe metody statystyczne. Na przykład pytając o wydatki i dochody, można poprosić o przybliżoną wartość w zł (skala ilorazowa) lub dać do wyboru możliwość kilku przedziałów (skala porządkowa). Pamiętajmy, że skala ilorazowa daje bardziej precyzyjne wyniki, można ją w razie potrzeby przekształcić w porządkową, ale nigdy odwrotnie (podrozdział 1.2).

Dobór próby badawczej jest jednym z najważniejszych etapów badania, od którego zależy, czy wyniki będzie można uogólniać na całą zbiorowość. Należy określić liczbę i zasady losowania jednostek ze zbiorowości statystycznej. Źle dobrana próba prowadzi do błędnych wniosków, niepotrzebnych kosztów (jeśli jest zbyt duża).

Wszystkie dotychczas omówione czynności obejmowały fazę przygotowania badania, zatem teraz można przystąpić do jego realizacji. W zależności od typu badań, obejmujących szeroko rozumianą turystykę i rekreację, mogą to być m.in. badania społeczne, inwentaryzacja i pomiary terenowe, pozyskiwanie danych z różnych źródeł i inne (podrozdział 1.4). Mogą one wykorzystywać różne techniki zbierania informacji: papierowe ankiety lub karty inwetaryzacyjne, które wymagają przetworzenia (np. przepisania) do formy elektronicznej; urządzenia elektroniczne (smartfony, laptopy) z wbudowaną bazą danych do bezpośredniego wpisywania danych.

Po zebraniu danych pierwotnych i wtórnych nadszedł czas na wybór metod statystycznych i oprogramowania służącego do ich weryfikacji oraz opracowania statystycznego. Istnieje dość duży wybór programów kom-

puterowych (komercyjnych[2] i niekomercyjnych[3]) oferujących obliczenia statystyczne. Warto sprawdzić, czy uczelnia zakupiła taki program na potrzeby studentów w wersji na komputery osobiste lub do pracowni komputerowych. Sprzedawcy oferują również darmowe użytkowanie oprogramowania w okresie kilku miesięcy. Popularne arkusze kalkulacyjne również umożliwiają obliczenie niektórych statystyk. Najważniejsza rada dla korzystających z programów komputerowych dotyczy zalecenia, by sprawdzić, jak obliczano statystyki i jak powinny być one interpretowane. Informacje te znajdują się zawsze w narzędziu „Pomoc" (*Help*). Użyte oprogramowanie, opis obliczeń i wzór powinny znaleźć się w publikacji. Często jest tam również zamieszczona pomoc służąca do interpretacji uzyskanych wyników. Wybierając oprogramowanie, warto sprawdzić, czy importuje i eksportuje ono pliki pochodzące z innych źródeł, np. Excel, ESRI Shapefile, dBase. Nie zawsze jesteśmy w stanie przewidzieć, z jakiego oprogramowania i z jakich danych będziemy jeszcze korzystać.

Jeśli mamy już dane w elektronicznej wersji bazy danych, rozpoczynamy ich przetwarzanie. Pierwsze działania powinny służyć uporządkowaniu i poznaniu zebranego materiału (rozdział 2). Wówczas sprawdza się, czy wielkość próby jest zgodna z założeniami, czy nie wystąpiły błędy podczas zbierania danych lub ich wprowadzania. Wszelkie wartości wyraźnie różniące się od pozostałych („odstające") powinny być sprawdzone pod tym kątem. Za pomocą metod graficznych lub tabelarycznych można przeprowadzić wstępną analizę danych. W raporcie tabela statystyczna powinna zawierać dwa rodzaje danych: liczebności bezwzględne oraz wartości procentowe. Należy również pamiętać, aby miała numer porządkowy według kolejności jej występowania w tekście. Gdy mamy do dyspozycji oprogramowanie statystyczne, skraca się czas tworzenia wykresów i szeregów statystycznych.

Następnym etapem jest analiza jednej zmiennej dla wybranych cech (rozdział 4). Pozwala ona na określenie typu rozkładu danych, wartości średniej, odchylenia poszczególnych wartości od przeciętnej, a także asymetrii i koncentracji.

Jeśli chcemy poszukać zależności między cechami i prognozować nieznane wartości, przechodzimy do następnego, bardziej zaawansowanego

2 Statistica, IBM SPSS Statistics.

3 R, GNU, Rstudio.

etapu badań statystycznych, tj. analizy korelacji i regresji (rozdział 5). Wykresy rozrzutu powinny znaleźć się w tej części raportu badawczego. Korzystając z nich, należy zawsze mieć na uwadze, czy badamy całą zbiorowość czy próbę. Interpretując wyniki, pamiętajmy, że współczynniki korelacji pokazują współzależność dwóch cech, ale nie wskazują jej przyczyn i skutków. Ponadto każdy program statystyczny pozwala na ocenę istotności obliczonych współczynników. Jeśli nasze analizy obejmują dłuższy czas, to zalecane jest wybranie metody analizy dynamiki (rozdział 6).

Obliczenia powinny dać odpowiedź na postawione na wstępie pytania szczegółowe, niekiedy pozwalają jeszcze na dodatkowe wnioski, które warto umieścić w tekście.

W podsumowaniu (raporcie) badań należy zacząć od przypomnienia hipotezy i celów szczegółowych, a następnie opisać cechy zbiorowości statystycznej lub próby (jak była zbierana). Bardzo ważna jest struktura raportu. Najprościej tak ją uporządkować, aby nawiązywała do celów szczegółowych badania. Warto pamiętać, że dobry wykres pozwala lepiej zinterpretować zjawisko niż strona tekstu z jego opisem, np. piramida płci i wieku.

Na wstępie zazwyczaj przeprowadza się statystykę opisową, a później przechodzi do analizy dwóch zmiennych. W interpretacji wyników należy brać pod uwagę nie tylko wartości statystyk, lecz także wskazać ich siłę oraz istotność. Zaleca się ostrożne formułowanie wniosków. Jeśli mamy możliwość porównania wyników z innymi pracami naukowymi czy raportami na podobny temat, warto to zrobić i odnieść się do nich. Często na zakończenie badań odczuwamy pewien niedosyt. Nie wszystkie cele udało się osiągnąć, podczas badań pojawiły się dodatkowe interesujące pytania badawcze itd. A my musimy już skończyć pracę i oddać ją do recenzji i oceny. Nasze rozterki też są twórcze, mogą stać się przyczynkiem do przyszłych badań własnych albo innych badaczy. Dlatego można o nich wspomnieć w podsumowaniu.

Życzę wszystkim Studentkom i Studentom turystyki i rekreacji powodzenia!

LITERATURA

Babbie E., 2004, *Badania społeczne w praktyce*, Wydawnictwo Naukowe PWN, Warszawa.

Babbie E., 2013, *Podstawy badań społecznych*, Wydawnictwo Naukowe PWN, Warszawa.

Badanie rynku czeskiego. Raport z badania dla Polskiej Organizacji Turystycznej, 2014, https://www.pot.gov.pl (dostęp: 26.05.2015).

Bedyńska S., Brzezicka-Rotkiewicz A. (red.), 2007, *Statystyczny drogowskaz: praktyczny poradnik analizy danych w naukach społecznych na przykładach z psychologii*, Wydawnictwo Szkoły Wyższej Psychologii Społecznej „Academica", Warszawa.

Francuz P., Mackiewicz R., 2005, *Liczby nie wiedzą, skąd pochodzą. Przewodnik po metodologii i statystyce nie tylko dla psychologów*, Wydawnictwo Katolickiego Uniwersytetu Lubelskiego, Lublin.

Frankfort-Nachmias Ch., Nachmias D., 2001, *Metody badawcze w naukach społecznych*, Zysk i S-ka, Poznań.

Gardner R., 2001, *The Marshall Plan Fifty Years Later: Three What-Ifs and a When*, [w:] M. Schain (ed.), *The Marshall Plan: Fifty Years After*, Palgrave, Basingstoke, s. 119–129.

GUS, 1978, *Rocznik statystyczny 1977*, Główny Urząd Statystyczny, Warszawa.

GUS, 1993, *Rocznik statystyczny 1992*, Główny Urząd Statystyczny, Warszawa.

GUS, 1994, *Rocznik statystyczny 1993*, Główny Urząd Statystyczny, Warszawa.

GUS, 1995a, *Mały rocznik statystyczny 1994*, Główny Urząd Statystyczny, Warszawa.

GUS, 1995b, *Rocznik statystyczny 1994*, Główny Urząd Statystyczny, Warszawa.

GUS, 1995c, *Turystyka w 1994 r.*, Główny Urząd Statystyczny, Warszawa.

GUS, 1997, *Mały rocznik statystyczny 1996*, Główny Urząd Statystyczny, Warszawa.

GUS, 2000, *Rocznik statystyczny 1999*, Główny Urząd Statystyczny, Warszawa.

Hara T., 2008, *Quantitative tourism industry analysis: Introduction to input-output, social accounting matrix modeling and tourism satellite accounts*, Routledge, London.

Jażdżewska I., 2013, *Statystyka dla geografów*, Wydawnictwo Uniwersytetu Łódzkiego, Łódź.

Kendall M.G., Buckland W.R., 1986, *Słownik terminów statystycznych*, Państwowe Wydawnictwo Ekonomiczne, Warszawa.

Kraak M.-J., Ormeling F., 1998, *Wizualizacja danych przestrzennych*, Wydawnictwo Naukowe PWN, Warszawa.

Luszniewicz A., Słaby T., 1996, *Statystyka stosowana*, Polskie Wydawnictwo Ekonomiczne, Warszawa.

Mider D., Marcinkowska A., 2013, *Analiza danych ilościowych dla politologów. Praktyczne wprowadzenie z wykorzystaniem programu GNU PSPP*, ACAD, Warszawa.

Pasławski J., 2003, *Jak opracować kartogram*, wyd. 2, Uniwersytet Warszawski, Wydział Geografii i Studiów Regionalnych, Warszawa.

Pasławski J. (red.), 2006, *Wprowadzenie do kartografii i topografii*, Wydawnictwo Nowa Era, Wrocław.

Pociecha M., 2002, *Metody statystyczne w zarządzaniu turystyką*, Albis, Kraków.

Ratajski L., 1989, *Metodyka kartografii społeczno-ekonomicznej*, Państwowe Przedsiębiorstwo Wydawnictw Kartograficznych, Warszawa.

Ruch turystyczny w Łodzi i województwie łódzkim w 2011 roku, 2012, Regionalna Organizacja Turystyczna Województwa Łódzkiego, Łódź.

Runge J., 1992, *Wybrane zagadnienia analizy przestrzennej w badaniach geograficznych*, Wydawnictwo Uniwersytetu Śląskiego, Katowice.

Runge J., 2007, *Metody badań w geografii społeczno-ekonomicznej: elementy metodologii, wybrane narzędzia badawcze*, Wydawnictwo Uniwersytetu Śląskiego, Katowice.

Szkup R., 2003, *Kształtowanie podmiejskiej przestrzeni wypoczynkowej*, Wydawnictwo Uniwersytetu Łódzkiego, Łódź.

Uhorczak F., Ostrowski J., 1972, *Typogramy F. Uhorczaka jako środek graficznej prezentacji zjawisk wielocechowych*, „Polski Przegląd Kartograficzny", nr 4, s. 145–151.

Włodarczyk B., 2012, *Ruch turystyczny w województwie łódzkim w 2011 roku*, Regionalna Organizacja Turystyczna Województwa Łódzkiego, Łódź.

Zając K., 1988, *Zarys metod statystycznych*, Państwowe Wydawnictwo Ekonomiczne, Warszawa.

Założenia polityki ekologicznej miasta Łodzi, 1997, Wydział Ochrony Środowiska Urzędu Miasta Łodzi, Łódź.

Żyszkowska W., Spallek W., Borowicz D., 2012, *Kartografia tematyczna*, Wydawnictwo Naukowe PWN, Warszawa.

ZAŁĄCZNIKI

Załącznik 1. Tablica statystyczna. Rozkład *t*-Studenta

n	***α***				
	0,40	**0,10**	**0,05**	**0,01**	**0,001**
1	1,3764	6,3137	12,7062	63,6559	636,5776
2	1,0607	2,9200	4,3027	9,9250	31,5998
3	0,9785	2,3534	3,1824	5,8408	12,9244
4	0,9410	2,1318	2,7765	4,6041	8,6101
5	0,9195	2,0150	2,5706	4,0321	6,8685
6	0,9057	1,9432	2,4469	3,7074	5,9587
7	0,8960	1,8946	2,3646	3,4995	5,4081
8	0,8889	1,8595	2,3060	3,3554	5,0414
9	0,8834	1,8331	2,2622	3,2498	4,7809
10	0,8791	1,8125	2,2281	3,1693	4,5868
11	0,8755	1,7959	2,2010	3,1058	4,4369
12	0,8726	1,7823	2,1788	3,0545	4,3178
13	0,8702	1,7709	2,1604	3,0123	4,2209
14	0,8681	1,7613	2,1448	2,9768	4,1403

Załącznik 1 cd.

n	*α*				
	0,40	**0,10**	**0,05**	**0,01**	**0,001**
15	0,8662	1,7531	2,1315	2,9467	4,0728
16	0,8647	1,7459	2,1199	2,9208	4,0149
17	0,8633	1,7396	2,1098	2,8982	3,9651
18	0,8620	1,7341	2,1009	2,8784	3,9217
19	0,8610	1,7291	2,0930	2,8609	3,8833
20	0,8600	1,7247	2,0860	2,8453	3,8496
21	0,8591	1,7207	2,0796	2,8314	3,8193
22	0,8583	1,7171	2,0739	2,8188	3,7922
23	0,8575	1,7139	2,0687	2,8073	3,7676
24	0,8569	1,7109	2,0639	2,7970	3,7454
25	0,8562	1,7081	2,0595	2,7874	3,7251
26	0,8557	1,7056	2,0555	2,7787	3,7067
27	0,8551	1,7033	2,0518	2,7707	3,6895
28	0,8546	1,7011	2,0484	2,7633	3,6739
29	0,8542	1,6991	2,0452	2,7564	3,6595
30	0,8538	1,6973	2,0423	2,7500	3,6460
40	0,8507	1,6839	2,0211	2,7045	3,5510
50	0,8489	1,6759	2,0086	2,6778	3,4960
60	0,8477	1,6706	2,0003	2,6603	3,4602
70	0,8468	1,6669	1,9944	2,6479	3,4350
80	0,8461	1,6641	1,9901	2,6387	3,4164
90	0,8456	1,6620	1,9867	2,6316	3,4019
100	0,8452	1,6602	1,9840	2,6259	3,3905

Źródło: https://pl.wikisource.org (dostęp: 26.05.2015).

Załącznik 2. Tablica statystyczna. Rozkład *chi*-kwadrat

n	*α*				
	0,50	**0,10**	**0,05**	**0,01**	**0,001**
1	0,455	2,706	3,841	6,635	10,828
2	1,386	4,605	5,991	9,210	13,816
3	2,366	6,251	7,815	11,345	16,266
4	3,357	7,779	9,488	13,277	18,467
5	4,351	9,236	11,071	15,086	20,515
6	5,348	10,645	12,592	16,812	22,458
7	6,346	12,017	14,067	18,475	24,322
8	7,344	13,362	15,507	20,090	26,125
9	8,343	14,684	16,919	21,666	27,877
10	9,342	15,987	18,307	23,209	29,588
11	10,341	17,275	19,675	24,725	31,264
12	11,340	18,549	21,026	26,217	32,910
13	12,340	19,812	22,362	27,688	34,528
14	13,339	21,064	23,685	29,141	36,123
15	14,339	22,307	24,996	30,578	37,697
16	15,339	23,542	26,296	32,000	39,252
17	16,338	24,769	27,587	33,409	40,790
18	17,338	25,989	28,869	34,805	42,312
19	18,338	27,204	30,144	36,191	43,820
20	19,337	28,412	31,410	37,566	45,315
21	20,337	29,615	32,671	38,932	46,797
22	21,337	30,813	33,924	40,289	48,268
23	22,337	32,007	35,173	41,638	49,728
24	23,337	33,196	36,415	42,980	51,179
25	24,337	34,382	37,653	44,314	52,620
26	25,337	35,563	38,885	45,642	54,052
27	26,336	36,741	40,113	46,963	55,476
28	27,336	37,916	41,337	48,278	56,892
29	28,336	39,088	42,557	49,588	58,301

Załącznik 2 cd.

n	*α*				
	0,50	**0,10**	**0,05**	**0,01**	**0,001**
30	29,336	40,256	43,773	50,892	59,703
40	39,335	51,805	55,759	63,691	73,402
50	49,335	63,167	67,505	76,154	86,661
60	59,335	74,397	79,082	88,379	99,607
70	69,335	85,527	90,531	100,425	112,317
80	79,334	96,578	101,879	112,329	124,839
90	89,334	107,565	113,145	124,116	137,208
100	99,334	118,498	124,342	135,807	149,449

Źródło: https://pl.wikisource.org (dostęp: 26.05.2015).

Załącznik 3. Ankieta dla właścicieli działek letniskowych

Lokalizacja	**Skala**	**Fizjonomia zabudowy**	**Skala**
Wieś	N	Powierzchnia domu	I
Gmina	N	Liczba kondygnacji	I
Powiat	N	Liczba izb	I
Województwo	N	Materiał ścian: a	N
Współrzędne geograficzne	I	Pokrycie dachu: b	N
Powierzchnia w m^2	I	Dach: c	N
Wyposażenie w infrastrukturę: odpowiedź *tak/nie*	**Skala**	**Wyposażenie działek: odpowiedź *tak/nie***	**Skala**
Ogrzewanie	N	Ogrodzenie trwałe	N
Gaz	N	Dom letniskowy	N
Elektryczność	N	Ogród owocowo-warzywny	N
Wodociąg	N	Budynki gospodarcze	N
Kanalizacja	N	Garaż	N
Internet	N	Chodniki, podjazdy	N
Telewizja satelitarna	N	Urządzenia sportowe	N
Metryczka właścicieli	**Skala**	**Metryczka właścicieli**	**Skala**
Wiek	I	Data nabycia działki	I
Płeć	N	Pierwszy właściciel? / który?	I
Wykształcenie	P	Odległość z domu na działkę	I
Sytuacja majątkowa: d	P	Sytuacja zawodowa: e	N
Miasto	N		
Osiedle	N		

Objaśnienia: skale porządkowe: N – nominalna, P – porządkowa, I – ilorazowa; a – drewno, drewno + cegła, pustak, blacha, prefabrykaty, inne (jakie?) ...; b – blacha, dachówka bitumiczna, eternit, papa + smoła, gont, inne (jakie?) ...; c – płaski, jednospadowy, dwuspadowy, wielospadowy, inne (jakie?) ...; d – bardzo dobra, dobra, przeciętna, zła, bardzo zła; e – uczeń/student, pracujący..., emeryt/rencista.

Źródło: opracowano na podstawie Szkup (2003), zmodyfikowane.

Załącznik 4. Przykład wypełnionej karty inwentaryzacyjnej łódzkich murali

Numer inwentarzowy	005
Nazwa właściwa	Czytająca
Obraz/fotografia Nazwisko fotografa: Anna Napieralska Data wykonania: 2012 r.	
Współrzędne geograficzne	51,777952 N 19,469894 E
Adres: ulica, numer domu	Pomorska 67
Właściciel budynku/zarządca	miasto
Data realizacji	wrzesień 2011 r.
Liczba płaszczyzn	1
Orientacja ściany	zachodnia
Autor projektu/realizacja	Aryz/Fundacja Urban Forms
Stan techniczny	bardzo dobry
Opis	Ten mural, a dokładnie jego symetryczne odbicie posłużyło jako tło w jednej z plansz gry komputerowej Devil May Cry.

Źródło: opracowanie Anna Napieralska.

Załącznik 5. Liczby losowe

0 4 8 8 6 5 3 3 2 0 9 2 1 2 7 5 3 0 0 1 6 3 9 0 0 6 4 8 0
8 6 8 9 4 7 3 7 9 0 6 3 4 3 0 8 8 9 5 4 3 3 4 7 8 4 4 0 4
8 6 9 2 4 1 5 5 7 0 5 2 1 0 4 9 7 1 7 5 0 1 8 1 6 6 7 1 0
9 3 0 1 5 1 6 6 6 1 6 3 8 7 0 9 2 8 3 8 7 3 9 6 5 7 9 4 7
0 1 1 7 9 5 1 8 0 5 2 9 7 3 5 4 0 4 4 9 7 8 2 0 0 1 8 2 7
3 5 1 7 8 6 0 9 6 3 8 9 1 2 0 5 9 5 2 8 8 7 2 8 5 6 5 0 9
6 9 5 9 3 9 0 3 5 5 8 9 4 5 4 0 6 9 0 9 8 4 9 1 5 6 2 7 6
6 2 4 4 0 4 6 3 8 3 1 0 0 7 0 5 2 5 4 4 0 1 1 8 4 6 7 8 0
4 3 8 9 8 3 5 1 9 7 9 9 6 9 2 6 2 0 6 5 4 4 0 1 1 8 4 6 7
8 0 4 9 8 9 8 3 5 1 9 7 9 6 9 2 6 2 0 6 1 4 4 0 8 9 0 5 9
4 0 0 9 7 3 8 0 8 4 4 4 8 2 5 1 8 1 5 6 7 9 1 2 6 1 5 2 7
6 8 4 7 5 8 4 3 3 3 2 3 1 2 3 0 4 2 1 5 1 4 3 7 3 7 0 7 0
5 2 7 7 1 4 0 4 4 5 0 0 2 6 0 9 9 6 4 0 7 7 3 0 8 5 9 2 6
1 5 2 7 1 7 1 9 8 6 3 2 6 4 5 4 5 1 6 9 0 7 6 8 4 2 0 3 0
7 5 2 4 9 0 8 6 3 7 9 3 1 0 7 0 1 8 0 6 6 0 2 1 6 5 6 5 5
5 7 4 0 7 3 2 2 7 1 0 3 0 7 0 6 6 1 6 9 7 5 5 1 1 8 7 5 4
7 4 7 3 0 7 2 4 8 3 9 2 2 6 5 7 8 2 3 5 0 8 7 6 3 2 0 1 5
1 8 6 4 7 3 3 8 7 4 2 0 7 1 3 5 4 7 0 8 2 2 1 4 5 3 0 8 7
8 6 3 3 8 4 9 3 4 9 5 7 3 3 4 4 9 7 2 1 5 4 4 1 4 6 9 1 9
5 9 9 8 6 2 2 4 8 8 0 8 2 7 4 5 3 8 6 5 1 3 4 8 3 6 2 2 9
8 9 1 5 0 1 3 1 7 4 3 2 7 2 3 1 7 4 3 8 1 4 8 6 4 7 1 6 4
7 5 8 7 2 1 0 7 6 1 0 3 5 5 0 0 3 7 1 1 1 7 8 0 9 3 5 9 8
4 9 1 4 8 7 0 6 1 2 1 0 2 5 5 5 4 6 5 7 4 2 0 2 7 2 0 1 1
4 8 1 0 4 2 6 2 1 7 2 7 2 6 8 5 9 1 9 2 7 8 9 2 6 2 8 6 0
6 0 7 3 4 6 6 5 2 2 4 9 4 9 2 3 1 2 7 5 7 9 6 3 0 6 9 5 2
1 5 3 1 2 9 3 3 7 9 0 3 9 5 3 8 7 7 8 4 0 1 6 5 1 0 8 5 3
2 9 0 2 6 6 2 4 6 5 5 1 4 9 7 2 4 5 2 5 3 9 0 4 7 7 3 5 1
2 1 1 0 5 9 3 4 1 6 2 7 8 6 2 9 7 1 3 2 8 9 2

Źródło: Jażdżewska (2013), s. 205.

INDEKS TERMINÓW