ELEMENTY STATYSTYKI DLA STUDENTÓW UCZELNI MEDYCZNYCH

Anna Baranowska

ELEMENTY STATYSTYKI DLA STUDENTÓW UCZELNI MEDYCZNYCH

Nowoczesne ujęcie z opisem obliczeń w programach Excel, R i Statistica

Wydanie drugie poprawione

Oficyna Wydawnicza GiS
Wrocław 2022

Anna Baranowska
Wydział Farmaceutyczny
Gdański Uniwersytet Medyczny
anna.baranowska@gumed.edu.pl

Skład książki w systemie XƎLaTeX wykonała autorka.

ISBN 978–83–67234–02–3

Wydanie II poprawione, Wrocław 2022
Oficyna Wydawnicza GiS, s.c., www.gis.wroc.pl
Druk i oprawa: Drukarnia I-BIS Bierońscy Sp.k.

SPIS TREŚCI

Przedmowa

Proponowana Czytelnikowi książka pomyślana jest przede wszystkim jako podręcznik dla studentów (zwykle pierwszego roku) uczelni medycznych, którzy w planie swoich studiów mają przedmiot ze słowem "statystyka" lub "biostatystyka" w jego nazwie. Często dla takiego przedmiotu statystykę łączy się z innym przedmiotem, takim jak matematyka czy informatyka i na realizację jego programu przeznacza się od 30 do 45 godzin.

Celem nauczania na zajęciach z takiego przedmiotu jest przekazanie studentom podstawowych wiadomości ze statystyki i nauczenie ich korzystania z odpowiednich statystycznych programów komputerowych, aby mogli oni wykonywać za ich pomocą obliczenia statystyczne przydatne w ich przyszłej działalności zawodowej oraz umieć poprawnie interpretować wyniki badań statystycznych towarzyszących prezentowanym wynikom prac specjalistycznych publikowanych przez wielu autorów.

Mimo że podręcznik ten przeznaczony jest głównie dla studentów młodszych lat uczelni medycznych, to mam nadzieję, że może on być pomocny także dla osób prowadzących zajęcia ze statystyki z tymi studentami, studentów starszych lat oraz innych osób, które chcą zapoznać się z podstawami statystyki i przykładami jej zastosowań.

Autorzy niektórych podręczników elementarnej statystyki przeznaczonych dla wymienionych wyżej grup czytelników często ograniczają się do wypisania podstawowych wzorów, według których należy wykonać obliczenia, i opisu sposobu wykonania tych obliczeń, korzystając z jednego z komputerowych programów statystycznych. I nawet jeśli wprowadzają pewne pojęcia i metody będące treścią klasycznej teorii statystyki, to robią to w sposób zwięzły i niewystarczający do jej zrozumienia.

Mając na uwadze, że na wielu kierunkach w wielu uczelniach klasyczny sposób nauczania podstaw statystyki jest łączony z nowoczesnym, tj. z użyciem komputerowych programów statystycznych, wybrałam inne podejście. Ograniczając terminologię[1)] i liczbę wprowadzanych pojęć ze statystyki opisowej do tych niezbędnych, które pojawiają się w części zwanej wnioskowaniem statystycznym oraz minimalizując liczbę wzorów matematycznych, ale nie unikając ich, starałam się wystarczająco dokładnie omówić każde pojęcie i metodę oraz zilustrować je jednym lub więcej przykładami, a często też rysunkami.

Pisząc ten podręcznik, brałam pod uwagę również takie kwestie jak:

- Statystyka jako część matematyki, która w naukach medycznych jest tylko

[1)] Starałam się używać terminologii możliwie najbliższej (po tłumaczeniu) terminologii używanej w nowoczesnych angielskojęzycznych podręcznikach statystyki.

narzędziem (choć pożytecznym i ważnym), wymaga znajomości matematyki (algebry) przynajmniej z zakresu szkoły średniej (i nieco więcej), aby zrozumieć opis jej podstawowych pojęć i metod.

- Poziom przygotowania matematycznego studentów uczelni medycznych jest bardzo zróżnicowany, a biegłość studentów (nie tylko uczelni medycznych) w wykonywaniu nawet prostych obliczeń i przekształceń algebraicznych w obecnych czasach nie jest duża.
- Przedmiot, któremu poświęcona jest ta książka, jest tylko jednym z wielu przedmiotów (i to nie najważniejszym, patrząc na liczbę przypisanych jemu punktów ECTS) i student nie ma zbyt dużo wolnego czasu na naukę tego przedmiotu, ponieważ musi dzielić ten czas z innymi przedmiotami.
- Mała liczba godzin i często nabrzmiały sylabus zmusza prowadzących zajęcia do dużego tempa i z wyżej wymienionych powodów wielu studentów nie radzi sobie z tym tempem. Dlatego chcą już poza salą wykładową czy ćwiczeniową mieć źródło, które pozwoli im wrócić do tematu z zajęć, uzupełnić swoją wiedzę i powtórzyć obliczenia prowadzone na zajęciach, korzystając z innych przykładów.
- Żyjemy w czasach, kiedy komputer stał się powszechnym narzędziem, a wraz z nim mamy dostęp do różnych programów statystycznych zwanych też pakietami lub środowiskami, które wykonują bardzo szybko skomplikowane procedury statystyczne, nie mówiąc już o wykonywaniu żmudnych (i nudnych) prostych działań arytmetycznych, w które obfitują różne metody statystyczne. Nie korzystanie z ich możliwości dla większości nie tylko studentów jest niewyobrażalne. Ponadto, tak dobry program statystyczny jak R jest darmowy i można go zainstalować na każdym komputerze. Excel z dobrze rozbudowanymi funkcjami i procedurami statystycznymi jest częścią pakietu Office i z tego powodu jest dostępny dla wielu osób. W końcu, na wielu uczelniach dostępny jest zaawansowany, ale względnie łatwy w użytkowaniu ze względu na system okienkowy, pakiet Statistica.

Aby uwzględnić nie tylko wymienione wyżej kwestie postąpiłam następująco:

- Zakładając, że wielu studentów po raz pierwszy spotyka się z symbolem sumy $\sum$ i symbolem całki $\int$, oprócz podania pewnych elementarnych informacji w tekście podręcznika wprowadziłam Dodatek B, w którym w miarę "łagodny" sposób objaśnione jest ich użycie. Podałam też nieco informacji o pochodnej. W przypadku całki ograniczyłam się do bardzo ogólnych informacji i interpretacji geometrycznej całki oznaczonej.
- Ilustrując omówione wcześniej pojęcia i metody, przeprowadzam szczegółowo wszystkie kroki wykonywanych obliczeń, aby studenci byli w stanie

powtórzyć je dla innych przykładów. Łącznie, podręcznik zawiera ponad 150 przykładów, w tym 119 przykładów w formie zadań (dlatego po ich treści występuje słowo **Rozwiązanie**), oraz 59 rysunków[2)].

- Zakres specjalistycznej wiedzy medycznej w początkowym okresie studiów nie jest duży i aby nie utrudniać zrozumienia istoty ilustrowanego przykładem problemu, w większości przykładów używam języka niespecjalistycznego, mniej formalnego[3)], a jak najbardziej zbliżonego do języka potocznego.
- Ograniczając, jak już wcześniej powiedziałam, liczbę wprowadzonych pojęć[4)] zakładam, że po ich opanowaniu student nie powinien mieć trudności, aby sięgnąć do innego źródła i zapoznać się samodzielnie z potrzebnym mu pojęciem tutaj nie omówionym. Natomiast nie usiłowałam ograniczać liczby testów hipotez[5)], których według moich doświadczeń jest więcej niż można i należy omówić na zajęciach. Wybór odpowiedniej liczby testów do opanowania przez studentów powinien należeć do osoby prowadzącej zajęcia ze studentami.
- Wprowadziłam Dodatek A, w którym podaję szczegółowe instrukcje (lub krótkie skrypty w języku R) jak otrzymać rozwiązania w prawie wszystkich przykładach mających formę zadań, korzystając z programów Excel, R i pakietu Statistica. A gdy potrzebne są kwantyle lub wartości krytyczne odpowiednich rozkładów, to zachęcam studentów do korzystania ze statystycznych programów komputerowych zamiast tablic statystycznych, które przynajmniej w części dotyczącej podstawowych rozkładów mają dużą szansę podzielić los tablic logarytmicznych.

Ponadto chciałam, aby podręcznik był matematycznie ścisły a jednocześnie przystępny zarówno dla osób słabiej przygotowanych matematycznie, jak i tych z większymi umiejętnościami. Dlatego starałam się tę ścisłość zapewnić nie przez użycie zawiłych wzorów matematycznych, a przez krótkie komentarze mające postać uwag w tekście, a jeszcze częściej przez umieszczenie ich w przypisach. Poza tym, dla osób lepiej przygotowanych z matematyki i o szerszym zainteresowaniu statystyką, wydzieliłam różne części materiału poprzez użycie mniejszej czcionki i

[2)] Wszystkie rysunki zostały wykonane, korzystając z pakietu *ggplot2* z programu R. Należy zwrócić też uwagę na to, że z powodu dużej ilości rysunków, tabel i różnych wzorów komputerowy program dokonujący łamania i składu tekstu musiał przenieść niektóre rysunki w inne miejsce niż miejsce tekstu, w którym znajduje się odniesienie do tych rysunków.

[3)] Np. używam wymiennie zwrotu "waga ciała" i "masa ciała" uważając, że dopóki nie ważymy (się) na innych planetach nie prowadzi to do nieporozumień.

[4)] Mimo ograniczania liczby pojęć w podręczniku znajdziemy 76 numerowanych definicji oraz dużą liczbę zdefiniowanych pojęć bez jawnego użycia słowa definicja.

[5)] Z wyjątkiem pominięcia testu średniej z nierealistycznym założeniem, że znamy wariancję.

minimalne zmniejszenie szerokości tekstu. Osoby mniej zainteresowane przedmiotem mogą pominąć te części tekstu.

Pragnę zaznaczyć, że wszystkie dane z wyjątkiem tych, gdzie podane zostało źródło ich pochodzenia, są fikcyjne (zwykle wygenerowane generatorem liczb pseudolosowych). Starałam się tylko, aby ich wartości zbytnio nie odbiegały od wartości spotykanych w rzeczywistości. Dlatego, choć wnioski otrzymane ze statystycznej analizy tych danych są merytorycznie poprawne, nie należy traktować ich tak, jak traktuje się wnioski uzyskane z rzeczywistych danych.

Niebagatelną rolę w projektowaniu i wyborze sposobu prezentacji materiału w przedstawianej czytelnikowi książce odegrało moje kilkuletnie doświadczenie zdobyte podczas prowadzenia zajęć ze statystyki ze studentami I roku Wydziału Farmaceutycznego Gdańskiego Uniwersytetu Medycznego, którego jestem pracownikiem. Podczas prowadzenia tych zajęć dawało się mocno odczuć brak takiego podręcznika i potrzebę jego napisania.

Na koniec pragnę wyrazić swoją wdzięczność i podziękowanie dr. hab. Zbigniewowi Bartoszewskiemu za jego nieocenioną pomoc merytoryczną i techniczną podczas pisanie tego podręcznika. Podzielił się on ze mną swoim doświadczeniem uzyskanym na Politechnice Gdańskiej podczas prowadzenia zajęć ze statystyki na studiach I stopnia na Wydziale Fizyki Technicznej i Matematyki Stosowanej oraz Wydziale Zarządzania i Ekonomii, przejrzał dokładnie wszystkie rozdziały i udzielił wiele cennych wskazówek i sugestii.

W wydaniu drugim dokonałam drobnych zmian redakcyjnych oraz poprawiłam zauważone błędy i usterki.

Anna Baranowska

Gdańsk, marzec 2022

Wprowadzenie

Termin **statystyka**[6)] używany jest w wielu znaczeniach. My będziemy używać pojęcia statystyki tak, jak jest ono używane we współczesnych podręcznikach statystyki, tj. jako nauki o metodach gromadzenia danych, ich opracowania, prezentacji, analizy i na ich podstawie wyciągania wniosków.

Tradycyjnie, statystykę dzielimy na *statystykę opisową* i *wnioskowanie statystyczne* (statystykę matematyczną). Statystyka opisowa zajmuje się gromadzeniem, opracowaniem, prezentacją oraz analizą danych statystycznych. Posługuje się ona zarówno metodami graficznymi (histogramami i różnego rodzaju wykresami) oraz tabelarycznymi, jak i numerycznymi wielkościami (miarami takimi jak, np. średnia, mediana, moda, odchylenie standardowe czy kurtoza) opisującymi ważne cechy charakterystyczne[7)] badanego zbioru danych. Natomiast wnioskowanie statystyczne zajmuje się wnioskowaniem o własnościach lub cechach charakterystycznych zbioru zwanego *populacją* na podstawie własności lub cech charakterystycznych *próby* (najczęściej losowej) będącej podzbiorem populacji.

Definicja 1. ***Populacją*** *będziemy nazywać skończony lub nieskończony zbiór wszystkich członków lub elementów danej grupy posiadających jedną (lub więcej) wspólną cechę, której badaniem jesteśmy zainteresowani.*

Definicja 2. ***Próbą*** *(lub próbką) będziemy nazywać dowolny niepusty i skończony podzbiór populacji.*

Przykłady:

(a) ***populacją*** *może być zbiór wszystkich ampułek szczepionki wyprodukowanej przez dany zakład w ciągu określonego czasu;*

(b) ***populacją*** *może być zbiór wszystkich dzieci w kraju poddanych szczepieniu daną szczepionką w danym roku;*

(c) ***próbą*** *może być podzbiór 50 ampułek szczepionki wybranych z populacji wszystkich ampułek zdefiniowanej w punkcie* (a)*;*

(d) ***próbą*** *może być podzbiór 30 dziewczynek i 30 chłopców wybranych z populacji wszystkich dzieci poddanych szczepieniu daną szczepionką zdefiniowanej w punkcie* (b).

W statystyce operuje się pojęciem zmiennej.

[6)] Więcej informacji o terminie statystyka, włączając jego historię, można znaleźć w [15].

[7)] Zamiast terminu cecha charakterystyczna, używanego przez nas jako odpowiednika angielskiego terminu *characteristic*, używa się też terminu cecha statystyczna.

Definicja 3. ***Zmienną*** *będziemy nazywać cechę charakterystyczną (lub atrybut), którą można mierzyć lub poklasyfikować i która może przyjmować różne wartości (liczbowe lub inne) dla różnych członków lub elementów populacji.*

Przykłady:

(a) *jeśli interesuje nas populacja dzieci w wieku 7 lat, to taką zmienną może być ich wzrost;*

(b) *jeśli interesuje nas populacja wszystkich pacjentów (np. w całym kraju lub w danym regionie), którzy skorzystali co najmniej raz w danym roku z zabiegu chirurgicznego, to zmienną może być, np. płeć pacjenta, długość okresu po każdym zabiegu prowadzącym do uzyskania pełnej sprawności fizycznej, ogólne samopoczucie pacjenta po upływie określonego czasu po zabiegu określane, np. jako bez zmian w porównaniu z samopoczuciem przed zabiegiem, umiarkowana poprawa, duża poprawa, całkowite ustąpienie dolegliwości.*

Definicja 4. *Mówimy, że znamy* ***rozkład zmiennej****, jeśli wiemy jakie wartości ona przyjmuje i jak często je przyjmuje.*

Będziemy wyróżniać następujące typy zmiennych:

1. Zmienne typu *ilościowego* (mierzalne), tj. takie, które przyjmują wartości liczbowe pochodzące z pomiaru (w szerokim znaczeniu tego słowa) lub obserwacji, np. wysokość ciśnienia krwi, wielkość tętna, liczba krwinek czerwonych w 1 mm^3 krwi, waga ciała.

   Z kolei, spośród **zmiennych typu ilościowego wyróżniamy**

   (a) Zmienne *typu dyskretnego* (skokowego), tj. zmienne przyjmujące wartości liczbowe i zbiór tych wartości jest albo skończony albo nieskończony; jeśli jest nieskończony, to wartości te można ponumerować, używając do numeracji liczb naturalnych. Najczęściej w naszych rozważaniach będziemy mieli do czynienia ze zmiennymi typu dyskretnego przyjmującymi wartości ze skończonego zbioru liczb całkowitych, np. wspomniana wcześniej liczba krwinek czerwonych w 1 mm^3 krwi, liczba przyjętych pacjentów każdego dnia w wybranym okresie czasu w danym szpitalu.

   (b) Zmienne *typu ciągłego*, tj. zmienne przyjmujące wartości z całego przedziału $[a, b]$[8)] (skończonego lub nieskończonego), np. temperatura ciała pacjentów danego szpitala o określonej porze dnia (zwróćmy tutaj uwagę, że przyrządy do mierzenia temperatury, głównie cyfrowe, dają nam tylko przybliżoną wartość temperatury ciała, dlatego

[8)] Lub sumy takich przedziałów.

w praktyce otrzymujemy tylko skończony podzbiór temperatur pacjentów ze zbioru wszystkich możliwych temperatur pokrywających cały odcinek na prostej).

2. Zmienne typu *jakościowego* (niemierzalne), np. płeć pacjenta, kolor oczu, jednostka chorobowa, na którą cierpi pacjent, natężenie bólu odczuwanego przez pacjenta (słaby, umiarkowany, silny), stopień zaawansowania choroby (lekko chory, ciężko chory, bardzo ciężko chory), ocena stanu zdrowia noworodka wg skali Apgar. Jako synonimu dla tego typu zmiennych używa się terminu *typ kategorialny*. Jak widać z podanych przykładów, zmienne tego typu na ogół przyjmują wartości nienumeryczne a jeśli nawet zmiennym tym nadaje się wartości liczbowe, to pełnią one rolę kodów czy nazw kategorii (klas) i wykonywanie na nich działań arytmetycznych może nie mieć sensu.

Z kolei, spośród **zmiennych typu jakościowego (kategorialnego) wyróżniamy**

(a) Zmienne *typu nominalnego*, tj. zmienne, których wartości można zakwalifikować do poszczególnych (wzajemnie wykluczających się) kategorii, ale kategorii tych nie można w sposób naturalny uporządkować (np. poprzez wskazanie kierunku wzrostu natężenia danej cechy charakterystycznej). Przykłady zmiennych typu nominalnego to: płeć pacjenta, kolor oczu, jednostka chorobowa, na którą cierpi pacjent, grupy krwi.

(b) Zmienne *typu porządkowego*[9)], tj. zmienne, których wartości można zakwalifikować do poszczególnych (wzajemnie wykluczających się) kategorii i kategorie te mają naturalny porządek (np. można wskazać kierunek wzrostu natężenia danej cechy charakterystycznej lub można wstawić znak nierówności pomiędzy poszczególnymi kategoriami). Przykłady zmiennych typu porządkowego to: natężenie bólu odczuwanego przez pacjenta (słaby, umiarkowany, silny), stopień zaawansowania choroby (lekko chory, ciężko chory, bardzo ciężko chory), temperatura ciała (poniżej 37°C, pomiędzy 37°C i 39°C, powyżej 39°C).

[9)] Zauważmy, że nie tylko zmienne typu jakościowego mogą być zmiennymi typu porządkowego, ponieważ zmienne typu ilościowego są w naturalny sposób zawsze typu porządkowego z porządkiem wyznaczonym przez relację nierówności "$\leq$" w każdym zbiorze liczbowym.

Rozdział 1

STATYSTYKA OPISOWA

1.1. Rozkład częstości i graficzna prezentacja danych

1.1.1. Organizacja danych w postaci tabelarycznej

Zwykle, zebrane 'surowe' dane niewiele mówią o interesującej nas cesze (lub cechach), którą chcemy badać i dopiero po dokonaniu odpowiedniego przekształcenia tych danych możemy uzyskać z nich dużo więcej informacji.

Przykład 1.1.1. *W zakładzie pracy 60 pracownikom wybranej grupy zmierzono ciśnienie tętnicze krwi. Wyniki pomiaru ciśnienia tętniczego skurczowego (mm Hg) zawiera poniższa tabela.*

117	136	128	144	120	133	111	121	118	136	124	133
130	119	133	131	127	126	101	108	158	112	114	103
130	119	111	128	130	141	132	124	129	161	116	155
129	109	129	132	140	145	124	152	133	112	132	119
120	102	122	104	117	142	111	127	120	125	114	121

▲

Takie dane nazywane są też *szeregiem nieuporządkowanym*. Można je uporządkować, np. wybierając przedziały liczbowe tak, aby były one rozłączne, równej długości (ale czasami są wyjątki) i każdy element tego zbioru znalazł się dokładnie w jednym z tych przedziałów. Następnie liczymy ile elementów naszego zbioru zawiera każdy z tych przedziałów, otrzymując tzw. *rozkład częstości*, tj. **zbiór klas i odpowiadających im częstości**. W polskiej literaturze rozkład częstości nazywany jest *szeregiem rozdzielczym* (w tym przypadku *szeregiem rozdzielczym przedziałowym*), a częstości nazywane są też *liczebnościami* lub *licznościami*.

Dużo informacji na podstawie zebranych danych daje nam też *rozkład częstości względnych*. Otrzymamy go, jeśli częstości w poszczególnych klasach podzielimy przez liczbę elementów w zbiorze danych. Rozkład częstości względnych można też wyrazić w *procentach*, mnożąc częstości względne przez 100 (pamiętajmy, że symbol % to nic innego jak $1/100$). Rozkład częstości względnych w procentach nie jest standardowym elementem tabeli rozkładu częstości, ale czasami będziemy go umieszczać w tabeli, ponieważ wykorzystuje się go w wykresach kołowych.

Tabela 1.1 przedstawia rozkład częstości i rozkład częstości względnych ciśnienia tętniczego skurczowego poddanej badaniu grupy z Przykładu 1.1.1. Ponieważ prawy koniec przedziału klasowego (z wyjątkiem ostatniego) jest równy lewemu

końcowi przedziału następującego po nim, to dla zapewnienia rozłączności tych przedziałów będziemy zakładać, że jest to przedział lewostronnie domknięty[1)], tj. lewy koniec przedziału klasowego jest włączony do przedziału, a prawy nie.

Tabela 1.1

Rozkład częstości i częstości względnych ciśnienia tętniczego krwi w badanej grupie

Klasy	Częstość	Częstość względna	Częstość względna w (%)
$100-110$	6	0,1000	10,00
$110-120$	14	0,2333	23,33
$120-130$	18	0,3000	30,00
$130-140$	13	0,2167	21,67
$140-150$	5	0,0833	8,33
$150-160$	3	0,0500	5,00
$160-170$	1	0,0167	1,67
Suma	60	1,0000	100,00

Teraz łatwo jest, np. zauważyć, że zdecydowana większość pracowników ma ciśnienie w granicach od 110 do 140 mm Hg. Zauważmy, że w rozpatrywanym przykładzie zmienna, którą jest ciśnienie skurczowe krwi, jest typu ilościowego.

Przejdźmy teraz do przykładu rozkładu częstości zmiennej typu jakościowego (kategorialnego).

Przykład 1.1.2. *Wybrano dane 90 pacjentów i wypisano ich grupy krwi wraz z czynnikiem Rh (+ lub − przy grupie krwi). Dane te zawiera poniższa tabela.*

0+	0+	0−	0+	0+	0+	A+	A+	A+	0+	A−	0−
AB+	0+	B+	A+	A+	B−	B+	0+	B−	B+	A−	0+
0+	B+	0+	0+	0+	0+	B−	A+	A+	0+	0+	0+
0+	A+	B+	A+	0+	A+	0+	B+	A+	0−	0+	0+
B+	0+	A+	B+	AB−	0+	A+	0+	A+	A+	0+	0+
B+	AB−	0+	A+	A+	0+	A+	AB+	0−	0+	A+	0+
0+	B+	0−	A+	A+	0+	0+	B+	0+	A−	0+	B−
A+	0−	A+	0+	A+	B+						

Po policzeniu częstości otrzymujemy rozkład częstości (*szereg rozdzielczy punktowy*) i rozkład częstości względnych, które to zapisaliśmy w Tabeli 1.2.

Zauważmy, że w przeciwieństwie do zmiennej typu ilościowego, gdzie klasami były przedziały liczbowe zwane też *przedziałami klasowymi*, dla zmiennej typu jakościowego klasami są wszystkie wartości, które ta zmienna może przyjmować.

[1)] W wielu pakietach statystycznych możliwy jest wybór albo lewostronnie domkniętego przedziału klasowego albo prawostronnie domkniętego z wyborem tego drugiego jako opcją domyślną.

Tabela 1.2

Rozkład grup krwi w badanej grupie

Klasy	Częstość	Częstość względna	Częstość względna w (%)
0+	37	0,4111	41,11
0−	6	0,0667	6,67
A+	24	0,2667	26,67
A−	3	0,0333	3,33
B+	12	0,1333	13,33
B−	4	0,0444	4,44
AB+	2	0,0222	2,22
AB−	2	0,0222	2,22
Suma	90	0,9999[a)]	99,99[a)]

[a)] Ze względu na zaokrąglenia do czterech miejsc po przecinku suma względnych częstości wynosi 0,9999, a nie 1. Z powodu zaokrągleń suma częstości w procentach wyniosła 99,99%, a nie 100%. Jeśli jednak będziemy korzystać ze statystycznego programu komputerowego i jeśli nawet wyniki będą wyświetlane, np. z dwoma miejscami po przecinku, to obliczenia będą wykonywane jednak z dużą większą dokładnością i otrzymane wyniki będą na ogół poprawne, tj. tam gdzie mamy otrzymać 1 to otrzymamy 1, a gdzie 100% to otrzymamy 100%. ▲

Czyli dla takiej zmiennej mamy tyle klas ile różnych wartości ona przyjmuje. W naszym przykładzie klasami są wszystkie grupy krwi z dodanym znakiem '+' i '−' (czynnik Rh). Podkreślmy zatem, że przedziałów klasowych (liczbowych) używamy dla danych ilościowych *typu ciągłego* jak i *typu dyskretnego*. W tym ostatnim przypadku **tylko wtedy, gdy badana zmienna (cecha) przyjmuje wiele różnych wartości** (np. ponad 20 choć nie mamy tutaj jakiejś sztywnej reguły). Natomiast gdy dane reprezentują zmienną typu jakościowego, to klasami są wszystkie przyjmowane przez tę zmienną wartości.

W przypadku rozkładu częstości z przedziałami klasowymi lewy koniec przedziału nazywamy *dolną granicą przedziału klasowego*, a prawy koniec przedziału *górną granicą przedziału klasowego*. Różnicę między górną i dolną granicą przedziału klasowego nazywamy *szerokością* lub *rozpiętością* przedziału klasowego.

Podsumujmy w punktach proces tworzenia tabeli rozkładu częstości.

1. Wyznaczamy wartość największą i najmniejszą w zbiorze danych, jeśli mamy do czynienia z danymi typu ilościowego i obliczamy **rozstęp** w analizowanym zbiorze danych, tj. różnicę pomiędzy znalezioną wartością największą i najmniejszą, aby wyznaczyć przedział liczbowy, który należy podzielić na klasy. Często ten przedział liczbowy rozszerza się nieco, biorąc jako lewy koniec tego przedziału 'wygodną' dla nas liczbę na lewo od wartości najmniejszej w zbiorze danych, a jako prawy koniec tego przedziału 'wygodną' dla nas liczbę na prawo od wartości największej w zbiorze danych.

2. Wyznaczamy liczbę klas, jeśli mamy do czynienia z danymi typu ilościowego. Powszechnie uważa się, że przy konstrukcji rozkładu częstości liczba przedziałów klasowych powinna wynosić od 5 do 20. I choć nie ma tu sztywnej reguły, to bardzo ważnym jest utworzenie wystarczająco dużo klas, aby dobrze opisać dane. Jeśli zbiór danych zawiera n elementów, to rozsądnym jest wybieranie liczby przedziałów klasowych k, kierując się jedną z poniższych rekomendacji:

   (a) liczba klas k w przybliżeniu równa $\sqrt{n}$,
   (b) liczba klas k równa części całkowitej $5 \log_{10} n$,
   (c) liczba klas k równa części całkowitej $1 + 3{,}322 \log_{10} n$.

   Aby klasy były rozłączne należy zdecydować, czy przedziały klasowe mają być półdomknięte z lewej czy z prawej. Otrzymane przedziały klasowe wpisujemy w pierwszej kolumnie tabeli.

3. Następnie liczymy ile wartości (danych) znajduje się w poszczególnej klasie i uzyskane liczby nazywane częstościami (lub liczebnościami klas) wpisujemy w drugiej kolumnie tabeli na wysokości odpowiadających im klas.

4. Dzieląc otrzymane w poprzednim kroku częstości przez liczbę wszystkich danych, otrzymujemy częstości względne, które zapisujemy w trzeciej kolumnie.

5. Często przydatnym jest wyrażenie częstości względnej w procentach i zapisanie jej w czwartej kolumnie. Otrzymuje się ją z trzeciej kolumny, mnożąc występujące w niej liczby przez 100. Z częstości względnej w procentach korzystamy, tworząc na przykład wykres kołowy dla rozkładu częstości[2)].

Czasami przydatnym jest użycie tzw. *rozkładu skumulowanych częstości*, który pokazuje częstości nie w danej klasie, ale w przedziałach określonych nierównością "mniejszy lub równe" danej liczbie, zwykle równej górnej granicy poszczególnego przedziału klasowego. Uzyskuje się go, sumując kolejne częstości w klasach. Np. w Przykładzie 1.1.1 skumulowana częstość dla pierwszej klasy wynosi $0+6=6$, dla drugiej klasy jest równa $0+6+14=20$, dla trzeciej $0+6+14+18=38$ itd. Dla drugiej, trzeciej i dalszych klas skumulowane częstości można też uzyskać, dodając do poprzedniej skumulowanej częstości częstość dla danej klasy. W podanym przykładzie, skumulowane częstości 20, 38 można uzyskać, wykonując obliczenia $6+14=20$, $20+18=38$ itd. W podobny sposób tworzymy *rozkład skumulowanych częstości względnych* i *rozkład skumulowanych częstości w procentach*.

[2)] Tworzenie wykresu kołowego ma sens, jeśli liczba klas nie jest zbyt duża. W przeciwnym razie, szczególnie w czarno-białym druku, możemy otrzymać zbyt małe i mało odróżnialne od siebie wycinki koła, powodujące małą czytelność wykresu.

Wystarczy zamiast częstości użyć odpowiednio częstości względnych i częstości względnych w procentach.

Przykład 1.1.3. *Grupie* 60 *pacjentów z temperaturą ciała 39°C i wyższą podano lek, a następnie co godzinę mierzono temperaturę, aby sprawdzić po ilu godzinach temperatura ciała obniżyła się o co najmniej 1°C. Uzyskane dane zawiera poniższa tabela.*

3	3	4	4	3	3	4	2	3	2	2	3	3	3	1	4	3	1	3	1
3	2	2	4	3	5	3	3	2	3	2	4	1	4	5	5	3	3	6	2
3	3	1	2	2	2	1	4	5	5	1	2	4	2	3	5	4	4	3	4

Ponieważ rozstęp w tym zbiorze danych jest mały ($6 - 1 = 5$), to klasami będą tu pojedyncze wartości danych, tj. liczby 1, 2, 3, 4, 5, 6. Ze względu na to, że mierzony czas jest zmienną *typu ciągłego* zwykle tworzy się przedziały klasowe o długości 1, odejmując od liczb będących klasami 0,5 i dodając do nich 0,5 (w tym przypadku klasy i przedziały klasowe nie będą tożsame, jak było to wcześniej). Postępując tak, otrzymamy następującą tabelę rozkładu częstości

Tabela 1.3

Rozkłady częstości i skumulowane rozkłady częstości czasów reakcji organizmu na lek obniżający temperaturę ciała

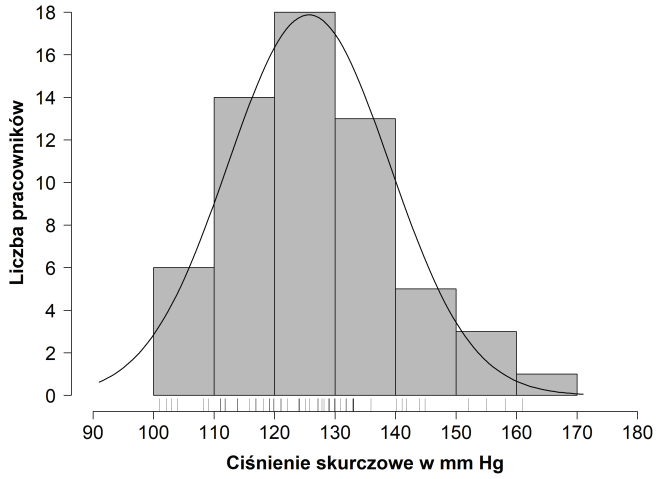
Klasy	Przedziały klasowe	Częstość	Skumulowana częstość	Częstość względna	Skumulowana częstość względna	Częstość wzgl.(%)
1	0,5 – 1,5	7	7	0,1167	0,1167	11,67
2	1,5 – 2,5	13	20	0,2167	0,3334	21, 67
3	2,5 – 3,5	21	41	0,3500	0,6834	35,00
4	3,5 – 4,5	12	53	0,2000	0,8834	20,00
5	4,5 – 5,5	6	59	0,1000	0,9834	10,00
6	5,5 – 6,5	1	60	0,0167	1,0001	1,67
Suma		60		1,0001[b)]		100,01[b)]

[b)] Zob. uwagi w przypisie *a*) pod Tabelą 1.2 na str. 18. ▲

1.1.2. Przykłady graficznej prezentacji danych

Jest bardzo wiele sposobów prezentacji danych za pomocą grafiki. W tej sekcji omówimy tylko niektóre z nich. Mając dane reprezentujące zmienną (cechę) **typu ilościowego** zorganizowane w tabelach z rozkładami częstości bezwzględnych, względnych czy też względnych w procentach (nieskumulowanych i skumulowanych), najprostszym i najczęstszym sposobem ich graficznej prezentacji jest

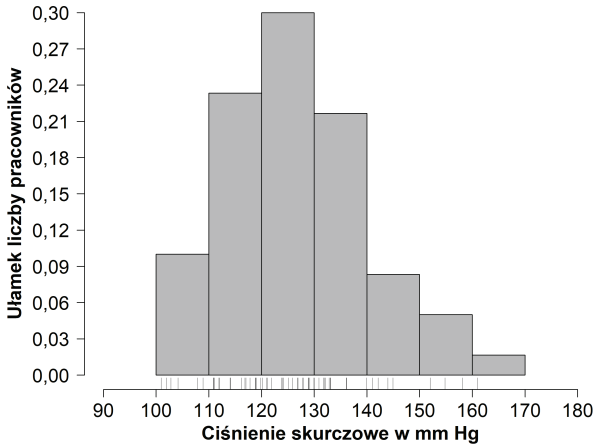
histogram. Składa się on z szeregu przylegających do siebie prostokątów umieszczonych na poziomej osi współrzędnych. Podstawy (szerokości) tych prostokątów wyznaczają szerokości przedziałów klasowych, a ich wysokość wyznaczają częstości (liczebności) elementów należących do danych przedziałów klasowych. Rys. 1.1 przedstawia histogram rozkładu częstości ciśnienia tętniczego skurczowego uzyskany na podstawie kolumny 1 i 2 Tabeli 1.1, a Rys. 1.2 przedstawia histogram rozkładu częstości względnych ciśnienia tętniczego skurczowego uzyskany na podstawie kolumny 1 i 3 tej samej tabeli. Wiele pakietów statystycznych daje możliwość wprowadzenia dodatkowych elementów do histogramu, np. do histogramu na Rys. 1.1 dodano przeskalowaną *krzywą rozkładu normalnego*[3)] (zwaną też *krzywą normalną*) najlepiej dopasowaną do analizowanego rozkładu. Zauważmy też, że w obu histogramach prostokąty zostały trochę odsunięte od osi poziomej i w uzyskanej w ten sposób przestrzeni znajdują się cienkie kreseczki wskazujące pozycje danych. Nie jest to jednak standardowy element histogramu.



Rys. 1.1. Histogram rozkładu częstości ciśnienia tętniczego skurczowego z nałożoną przeskalowaną krzywą rozkładu normalnego

Jeśli chodzi o dane typu jakościowego (kategorialnego), to do ich graficznej prezentacji zamiast histogramów używamy wykresów słupkowych. Są one podobne do histogramów. Dla tego typu danych szerokość prostokątów nie ma znaczenia i

[3)] O rozkładzie normalnym będzie mowa w dalszej części podręcznika. Gdy suma pól prostokątów histogramu, oznaczmy ją literą S, jest różna od 1, to krzywa rozkładu normalnego wymaga skalowania, tj. pomnożenia jej rzędnych (współrzędnych na osi pionowej) przez liczbę S.



Rys. 1.2. Histogram rozkładu względnej częstości ciśnienia tętniczego skurczowego

nie muszą przylegać jeden do drugiego. Ich szerokości powinny być równe, a ich wysokości reprezentują częstości w odpowiednich klasach. Oś pionowa[4)] powinna mieć podziałkę. Natomiast podziałka nie jest potrzebna na osi poziomej i zamiast niej pod prostokątami występują etykiety (nazwy klas). Rys. 1.3 przedstawia wykres słupkowy częstości grup krwi i czynnika Rh w badanej grupie osób, a Rys. 1.4 jest wykresem słupkowym względnej częstości grup krwi i czynnika Rh w badanej grupie osób.

Histogramów używamy również do graficznej prezentacji częstości skumulowanych. Przykład takiego histogramu dany jest na Rys. 1.5.

Do graficznej prezentacji danych używane są też wykresy kołowe[5)]. Przykład takiego wykresu dany jest na Rys. 1.6.

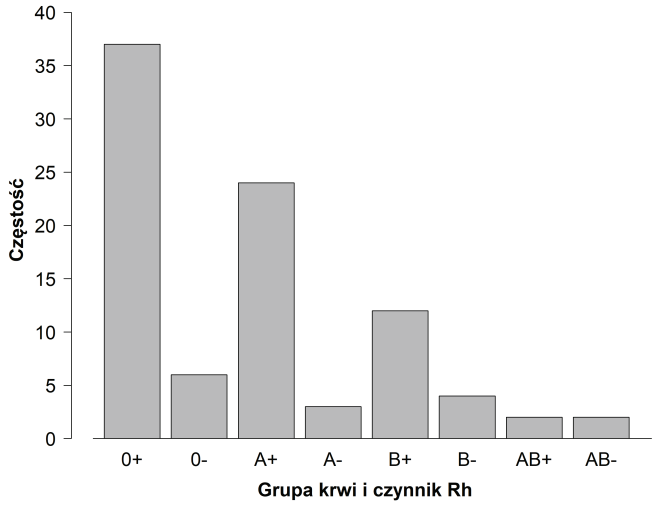
Dane lub bardziej precyzyjnie, informacje o interesującej nas zmiennej zawarte w danych, można też przedstawiać graficznie za pomocą wykresów liniowych. Jest wiele rodzajów wykresów liniowych. Omówimy tylko kilka ważniejszych z nich.

Wielobok[6)] częstości (włączając częstość względną jak i skumulowaną) to łamana łącząca punkty na płaszczyźnie, których pierwszą współrzędną są środki

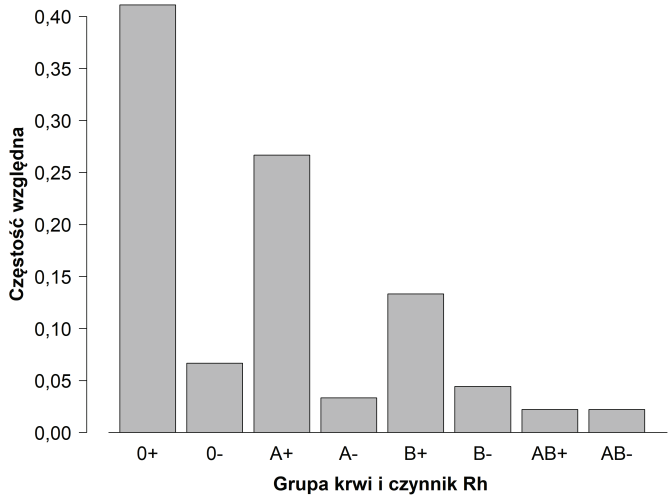
[4)] Czasami wygodniej jest odwrócić role osi, aby otrzymać poziomy wykres słupkowy i wtedy w tym zdaniu i następnym trzeba zamienić miejscami słowa "pionowa" i "pozioma".

[5)] Uważa się jednak, że wykresy kołowe nie są dobrą formą prezentacji informacji, ponieważ oko ludzkie dobrze sobie radzi z porównaniami, gdy użyta jest miara liniowa, a znacznie gorzej, gdy mamy porównywać względne miary pól.

[6)] Zamiast *wielobok* powinno się mówić *łamana*, chyba że zaznaczamy cały obszar ograniczony łamaną, osią poziomą i pionowymi odcinkami łączącymi końce łamanej z osią poziomą.



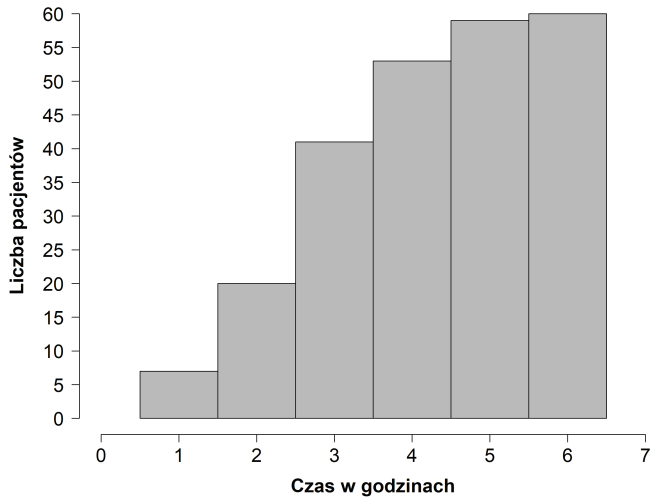
Rys. 1.3. Wykres słupkowy rozkładu częstości grup krwi i czynnika Rh



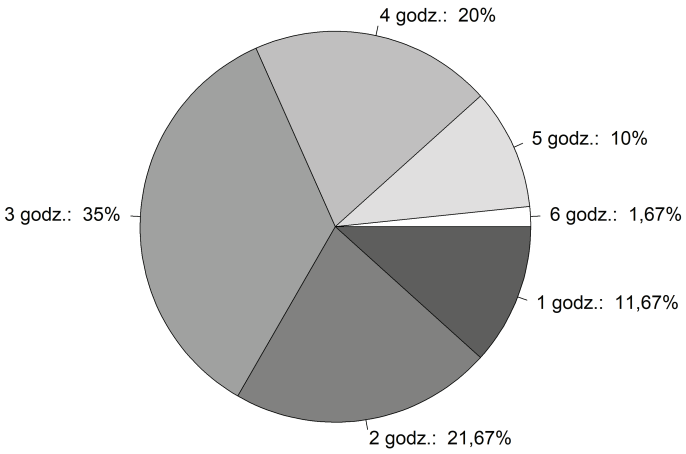
Rys. 1.4. Wykres słupkowy rozkładu częstości względnej grup krwi i czynnika Rh

przedziałów klasowych[7)], a drugą częstości. W zależności od tego czy mamy do czynienia z częstościami, częstościami względnymi czy też częstościami skumulowanymi będziemy mówić odpowiednio: o wieloboku częstości, wieloboku częstości

[7)] Jeśli klasy są jednoelementowe, to pierwszymi współrzędnymi są wartości danych, dla których podane są częstości.



Rys. 1.5. Histogram rozkładu częstości skumulowanych czasu potrzebnego po zażyciu leku do obniżenia temperatury ciała o co najmniej 1°C

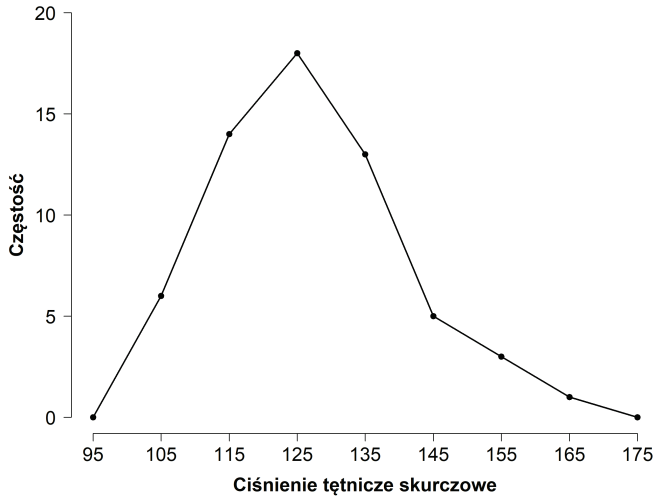


Rys. 1.6. Wykres kołowy rozkładu częstości czasu potrzebnego po zażyciu leku do obniżenia temperatury ciała o co najmniej 1°C

względnych i wieloboku częstości skumulowanych[8)]. Na osi poziomej z zasady zaznaczamy jeszcze jedną wartość na lewo od środka pierwszego przedziału częstości (na Rys. 1.7 jest nią wartość 95) w taki sposób, aby odległości wszystkich są-

[8)] Nazywanym też krzywą wzrostu częstości, a w języku angielskim *ogive*. *Uwaga:* Dla danych typu ciągłego zamiast środków przedziałów klasowych łączymy końce tych przedziałów.

siednich zaznaczonych współrzędnych były jednakowe, a w przypadku wieloboku częstości i częstości względnych zaznaczamy w podobny sposób jeszcze drugą wartość na prawo od środka ostatniego przedziału częstości (na Rys. 1.7 jest nią wartość 175). Tym dodanym współrzędnym przypisujemy zerowe częstości. **Wieloboki częstości mogą być użyte tylko do danych reprezentujących zmienne (cechy) typu ilościowego**.



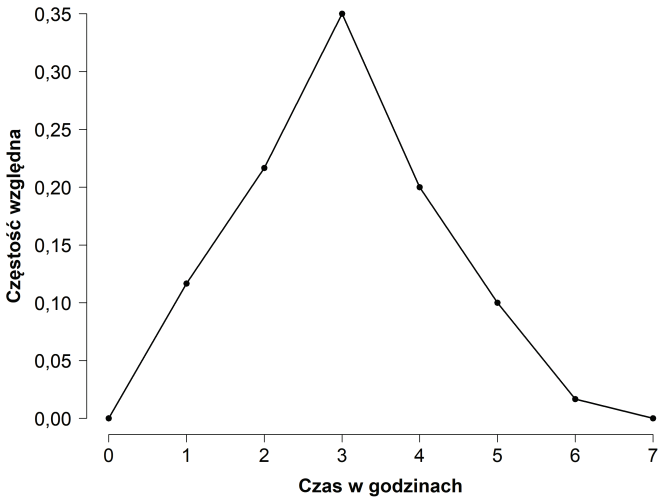
Rys. 1.7. Wielobok częstości dla ciśnienia tętniczego skurczowego

Rys. 1.7 przedstawia wykres wieloboku częstości dla ciśnienia tętniczego skurczowego z Przykładu 1.1.1 (dane w Tabeli 1.1), Rys. 1.8 pokazuje wielobok częstości względnych dla czasu potrzebnego po zażyciu leku do obniżenia temperatury ciała o co najmniej 1°C dla danych z Tabeli 1.3, a Rys. 1.9 prezentuje wielobok skumulowanych częstości dla czasu potrzebnego po zażyciu leku do obniżenia temperatury ciała o co najmniej 1°C dla danych z Tabeli 1.3.

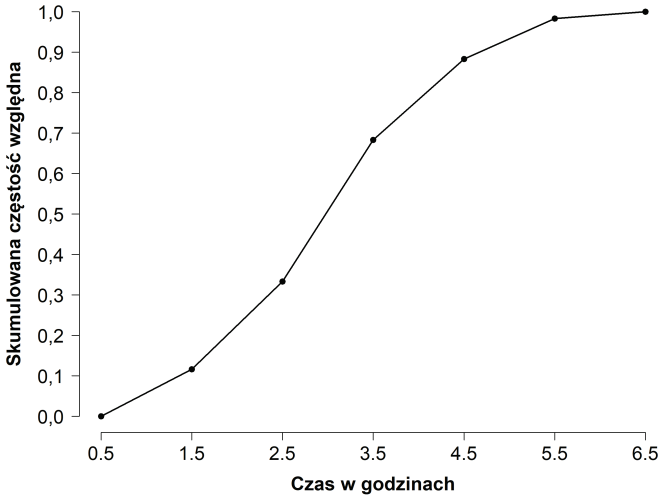
1.2. Opis statystyczny

1.2.1. Miary tendencji centralnej

W statystyce najczęściej mamy do czynienia z próbą wziętą z populacji w celu uzyskania z niej informacji o interesującej nas zmiennej (cesze) lub zmiennych. Następnie na podstawie tej informacji wnioskuje się o własnościach badanej cechy w całej populacji. Jeśli jednak populacja jest mała, to na ogół nie ma potrzeby wybierania próby z populacji, ponieważ interesującą nas informację możemy uzyskać,



Rys. 1.8. Wielobok częstości względnych dla czasu potrzebnego po zażyciu leku do obniżenia temperatury ciała o co najmniej 1°C



Rys. 1.9. Wielobok skumulowanych częstości względnych dla czasu potrzebnego po zażyciu leku do obniżenia temperatury ciała o co najmniej 1°C

wykorzystując całą populację. Definicje niektórych pojęć, które wprowadzimy w tej sekcji mogą się różnić w zależności od tego czy będą się one odnosić do próby czy do populacji. Liczbę elementów w próbie będziemy oznaczać literą n, a liczbę

elementów w populacji literą N.

Miary tendencji centralnej charakteryzują za pomocą liczb średni lub typowy poziom wartości badanej zmiennej (cechy). Trzy powszechnie używane miary tendencji centralnej to **średnia**[9]**, mediana i moda**. Zanim zdefiniujemy te trzy miary powiedzmy najpierw dla jakiego typu danych mają one sens:

- *średnią* definiuje się dla danych typu ilościowego (ciągłego i dyskretnego);
- *medianę* definiuje się dla danych typu ilościowego[10] (ciągłego i dyskretnego);
- *modę* definiuje się dla danych typu ilościowego i jakościowego; może ona jednak nie istnieć.

Definicja 1.2.1. *Dla zbioru danych reprezentujących wartości pewnej zmiennej (cechy)* ***średnia arytmetyczna*** *tej zmiennej lub krótko* ***średnia*** *to suma wszystkich wartości elementów tego zbioru podzielona przez liczbę jego elementów. Średnią dla próby $x_1, \ldots, x_n$ będziemy oznaczali symbolem $\bar{x}$, a średnią dla populacji $x_1, \ldots, x_N$ oznaczymy grecką literą μ. Mamy wobec tego następujące wzory:*

$$\text{Średnia dla próby:} \qquad \bar{x} = \frac{x_1 + x_2 + \cdots + x_n}{n}, \tag{1.1}$$

$$\text{Średnia dla populacji:} \qquad \mu = \frac{x_1 + x_2 + \cdots + x_N}{N}. \tag{1.2}$$

Zapis wzorów (1.1) i (1.2) można znacznie skrócić, jeśli użyjemy symbolu sumowania $\sum$ (grecka duża litera *sigma*)[11]. Wtedy oba wzory możemy zapisać

$$\bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i, \qquad \mu = \frac{1}{N}\sum_{i=1}^{N} x_i. \tag{1.3}$$

Wzór $\sum_{i=1}^{n} x_i$ z symbolem $\sum_{i=1}^{n}$, który zobaczymy, gdy wzór ten jest wyodrębniony z tekstu[12] (jak w (1.3) powyżej) i wzór $\sum_{i=1}^{n} x_i$ z symbolem $\sum_{i=1}^{n}$, który zobaczymy, gdy jest on w tej samej linijce wraz z innym tekstem oznacza, że należy zsumować wyrazy ciągu $x_1, x_2, \ldots, x_n$, zaczynając od wyrazu z indeksem (numerem) $i = 1$ i kończąc na wyrazie z indeksem n.

[9] W naszym tekście ograniczymy się tylko do średniej *arytmetycznej* choć w statystyce używa się też innych średnich, np. *geometrycznej*, *ważonej* itp.

[10] Czasami statystycy definiują medianę też dla danych typu jakościowego porządkowego.

[11] Więcej informacji o symbolu sumowania znajduje się w Dodatku B.

[12] Jak widać wyżej, taki wzór użyty w tekście powoduje, że odstęp linijki, w której się on znajduje, od dwóch sąsiednich linijek tekstu jest znacznie większy niż odstępy pomiędzy pozostałymi linijkami tekstu.

Przykład 1.2.1. *Poniższa tabela zawiera wagi noworodków (w kg) urodzonych w wybranym tygodniu w pewnym szpitalu. Obliczyć średnią wagę noworodka.*

4,09	3,79	3,68	3,58	4,05	3,69	2,85	3,18	3,41	3,71
3,76	3,11	3,91	3,80	2,97	2,96	2,70	4,01	3,72	2,97

Rozwiązanie:

$$\begin{aligned}\sum_{i=1}^{20} x_i = {} & 4{,}09 + 3{,}79 + 3{,}68 + 3{,}58 + 4{,}05 + 3{,}69 + 2{,}85 + 3{,}18 \\ & + 3{,}41 + 3{,}71 + 3{,}76 + 3{,}11 + 3{,}91 + 3{,}80 + 2{,}97 + 2{,}96 \\ & + 2{,}70 + 4{,}01 + 3{,}72 + 2{,}97 = 69{,}94.\end{aligned}$$

Zatem $$\bar{x} = \frac{\sum_{i=1}^{20} x_i}{20} = \frac{69{,}94}{20} = 3{,}497.$$ ▲

Często granice sumowania $i = 1$ i n we wzorze $\sum_{i=1}^{n} x_i$ opuszcza się, jeśli zmienna ma tylko jeden[13)] indeks, np. i, i pisze się wtedy $\sum x_i$, co oznacza, że należy zsumować wartości x_i dla wszystkich i.

Uwaga 1.2.1. *Jedną z głównych wad średniej jest jej wrażliwość na każdą wartość składnika sumy obliczanej podczas jej wyznaczania, tak że pojedyncza wartość danej mocno różniąca się od wartości pozostałych danych[14)] może znacznie zmienić wartość średniej.*

Uwaga 1.2.2. *Obliczanie średnich, choć łatwe do wykonania korzystając nawet z prostego kalkulatora, najczęściej będzie wykonywane przez programy statystyczne jako jeden z elementów wśród innych obliczeń statystycznych. Ale wiele programów statystycznych daje nam możliwość obliczyć ją oddzielnie.*

Wady średniej nie ma miara tendencji centralnej zwana **medianą**. Uporządkujmy niemalejąco, tj. od najmniejszej do największej, dane $x_1, \ldots, x_n$ i tak uporządkowany zbór oznaczymy jako zbiór $x_{(1)}, \ldots, x_{(n)}$. Na przykład dla zbioru $3, 7, 1$, gdzie $x_1 = 3$, $x_2 = 7$ i $x_3 = 1$, po uporządkowaniu otrzymamy zbiór $1, 3, 7$, gdzie $x_{(1)} = 1$, $x_{(2)} = 3$, $x_{(3)} = 7$.

Definicja 1.2.2. *W zbiorze danych reprezentujących wartości pewnej zmiennej i będących próbą* ***mediana*** *jest* ***wartością środkową*** *uporządkowanego niemalejąco lub nierosnąco wyjściowego zbioru danych. Zatem, gdy wartości $x_1, \ldots, x_n$*

[13)] Zmienne mogą mieć więcej indeksów, np. x_{ij} ma dwa indeksy i i j. Opuszczając je, nie wiedzielibyśmy czy mamy sumować po obu indeksach i i j czy tylko po jednym z nich.

[14)] Takie dane nazywamy danymi "odstającymi" (po angielsku *outliers*).

zmiennej w próbie tworzą po uporządkowaniu ciąg niemalejący $x_{(1)}, \ldots, x_{(n)}$*, to* ***medianę*** *z próby, oznaczaną symbolem* m_e*, definiuje wzór*

$$m_e = \begin{cases} x_{((n+1)/2)} & \text{gdy} \quad n \text{ jest nieparzyste,} \\ \frac{1}{2}\left(x_{(n/2)} + x_{((n/2)+1)}\right) & \text{gdy} \quad n \text{ jest parzyste.} \end{cases} \tag{1.4}$$

Podobnie definiujemy **medianę populacji** M_e. Zamieniając we wzorze (1.4) m_e na M_e i małe n na duże N, otrzymujemy

$$M_e = \begin{cases} x_{((N+1)/2)} & \text{gdy} \quad N \text{ jest nieparzyste,} \\ \frac{1}{2}\left(x_{(N/2)} + x_{((N/2)+1)}\right) & \text{gdy} \quad N \text{ jest parzyste.} \end{cases} \tag{1.5}$$

Przykład 1.2.2. *Wyznaczmy medianę w zbiorze danych z Przykładu 1.2.1.*

Rozwiązanie:

Po uporządkowaniu danych niemalejąco otrzymamy:

2,70	2,85	2,96	2,97	2,97	3,11	3,18	3,41	3,58	**3,68**
3,69	3,71	3,72	3,76	3,79	3,80	3,91	4,01	4,05	4,09

Ponieważ liczba elementów naszego zbioru jest parzysta ($n = 20$), to $x_{n/2} = x_{10} = 3{,}68$ a $x_{(n/2)+1} = x_{11} = 3{,}69$ (obie wartości w powyższej tabeli zaznaczyliśmy, zapisując je tłustym drukiem) i z drugiej linijki wzoru otrzymujemy

$$m_e = \frac{1}{2}(3{,}68 + 3{,}69) = 3{,}685.$$

▲

Przykład 1.2.3. *Niech następujące dane*

3, 5, 4, 6, 15, 4, 3

wyrażają liczbę poważnych zabiegów chirurgicznych po urazach spowodowanych wypadkami w kolejnych dniach jednego z tygodni lutego. Wyznaczyć średnią liczbę zabiegów chirurgicznych na jeden dzień oraz medianę dla tych danych.

Rozwiązanie:

Po uporządkowaniu[15)] (niemalejąco) naszego zbioru danych otrzymamy

$$3, 3, 4, \mathbf{4}, 5, 6, 15$$

Wykonując obliczenia, otrzymujemy

[15)] Zwróćmy uwagę na to, że średnia nie zależy od kolejności w jakiej występują dane.

$$\bar{x} = \frac{3+3+4+4+5+6+15}{7} = 5{,}71, \qquad m_e = x_{(4)} = 4.$$

Jak widzimy, obie te miary tendencji centralnej znacznie się różnią. Na średnią, w odróżnieniu od mediany, duży wpływ ma liczba 15, znacznie *odstająca* od pozostałych danych. Jeśli przyczyną tej dużej liczby interwencji chirurgicznych w porównaniu do reszty dni tygodnia było w tym dniu, np. nagłe oblodzenie jezdni i chodników, to mediana pewnie lepiej niż średnia przybliża średnią dzienną liczbę interwencji w innych tygodniach zimowych i nie zmieni się, jeśli w rozważanych danych zamiast 15 znajdzie się dowolna inna liczba większa od 3. Natomiast średnia, zależąc od wartości każdej obserwacji, pozwala nam zauważyć takie zmiany w danych i zwrócić na nie uwagę. ▲

Trzecia miara tendencji centralnej, tj. **moda** nazywana też **dominantą**, jest rzadziej używana niż dwie poprzednie.

Definicja 1.2.3. *Dla zbioru danych reprezentujących wartości pewnej zmiennej (cechy)* ***modą*** *nazywamy wartość, która występuje w tym zbiorze najczęściej.*

Niech zbiór danych (obserwacji) reprezentuje wartości pewnej zmiennej.

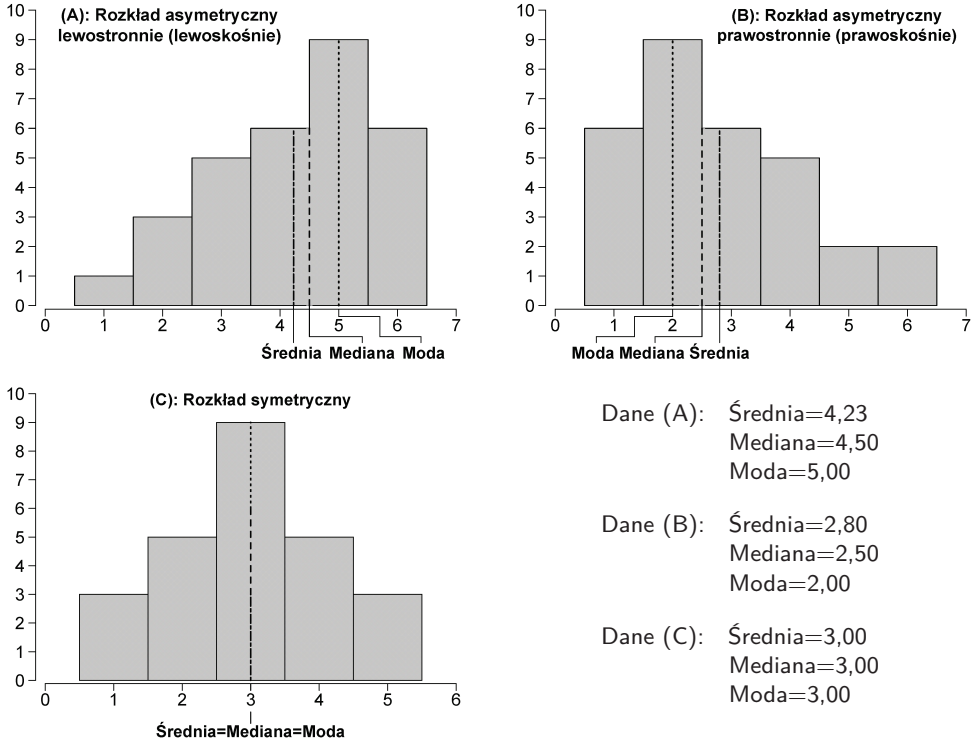
- Jeśli w zbiorze tym tylko jeden element występuje z największą częstością, to mówimy wtedy, że zmienna ma rozkład *jednomodalny*.
- Jeśli w zbiorze tym dwa elementy występują z tą samą największą częstością, to oba są modami i mówimy wtedy, że zmienna ma rozkład *dwumodalny* lub *bimodalny*.
- Jeśli w zbiorze tym więcej niż dwa elementy występują z tą samą największą częstością, to one wszystkie są modami i mówimy wtedy, że zmienna ma rozkład *wielomodalny*.
- Jeśli w zbiorze tym wszystkie elementy występują tylko raz, to dla tego zbioru moda nie istnieje.

Patrząc na Tabelę 1.3, widzimy, że w Przykładzie 1.1.3 modą jest liczba 3 godzin, ponieważ ta liczba godzin występuje najczęściej w badanym zbiorze i jej częstość występowania wynosi 21. Dla zbioru danych z Przykładu 1.2.1 modą jest ciężar noworodka równy 2,97 kg. Dla zbioru danych z Przykładu 1.2.3 modami są liczba 3 i liczba 4 interwencji chirurgicznych w ciągu dnia. Dla zbioru danych z Przykładu 1.1.2 (Tabela 1.2) modą jest grupa krwi 0+.

1.2.2. Kształt rozkładu częstości

Rozkład częstości może przyjmować różne kształty. Wyróżnia się trzy kształty rozkładu: **symetryczny**, **asymetryczny prawostronnie** zwany też *asymetrycznym*

prawoskośnie lub *asymetrycznym dodatnio* i **asymetryczny lewostronnie** zwany też *asymetrycznym lewoskośnie* lub *asymetrycznym ujemnie*. O dwóch ostatnich typach rozkładów mówimy, że są skośne. Przykładowe wykresy (histogramy) rozkładu symetrycznego, asymetrycznego prawostronnie i asymetrycznego lewostronnie pokazane są na Rys. 1.10. Jeśli zmienna ma rozkład mocno asymetryczny



Rys. 1.10. Typy rozkładów częstości (wyróżniliśmy położenie średniej, mediany i mody różnym typem linii, prowadząc je od górnej krawędzi odpowiedniego słupka (prostokąta) do punktu wskazującego dany parametr na osi poziomej)

prawostronnie (jak mierzyć skośność rozkładu powiemy później), to większość jej wartości wpada do przedziału na lewo od średniej i skupia się w lewym końcu rozkładu, a jego tzw. ogon znajduje się po prawej stronie wykresu. Ponadto, jej średnia i mediana są na prawo od mody oraz zwykle średnia jest też na prawo od mediany[16)].

Natomiast, jeśli zmienna ma rozkład mocno asymetryczny lewostronnie, to

[16)] Dla rozkładów asymetrycznych prawostronnie można jednak podać przykłady, dla których średnia jest na lewo od mediany, a dla asymetrycznych lewostronnie na prawo od mediany.

większość jej wartości wpada do przedziału na prawo od średniej i skupia się w prawym końcu rozkładu, a jego ogon znajduje się po lewej stronie wykresu. Ponadto, jej średnia i mediana są na lewo od mody oraz zwykle średnia jest też na lewo od mediany.

Dlatego w przypadkach mocnej asymetrii bardziej odpowiednią miarą tendencji centralnej będzie raczej mediana niż średnia.

1.2.3. Miary zmienności (rozproszenia, dyspersji)

Zacznijmy od następującego przykładu.

Przykład 1.2.4. *Wiersz A poniższej tabeli zawiera uporządkowane rosnąco czasy oczekiwania (w dniach) wybranych 15 pacjentów na zabieg w szpitalu w jednej miejscowości, a wiersz B tak samo uporządkowane czasy oczekiwania wybranych 15 pacjentów na zabieg w szpitalu w innej miejscowości. Wyznaczyć średnią i medianę dla czasów w wierszu A i czasów w wierszu B.*

A:	3	6	9	12	17	20	25	30	32	35	39	40	41	45	56
B:	13	13	14	15	20	21	26	30	31	34	36	37	38	40	42

Rozwiązanie:

Obliczenia są łatwe i po ich wykonaniu otrzymujemy

$$\bar{x}_A = \bar{x}_B = 27\tfrac{1}{3}, \qquad m_{e:A} = m_{e:B} = 30. \qquad \blacktriangle$$

Przykład ten pokazuje, że choć średnie i mediany dla obu zbiorów danych są takie same, to wartości danych w zbiorze A są bardziej rozproszone niż w zbiorze B. Dlatego najprostszą miarą zmienności (rozproszenia) jest wprowadzony wcześniej *rozstęp* zbioru danych. Rozstęp zbioru A wynosi 53 a rozstęp zbioru B równy jest 29. Wadą rozstępu jest jego zależność tylko od dwóch wartości w zbiorze danych, największej i najmniejszej, a nie bierze on pod uwagę wszystkich pozostałych wartości danych. Wydawać by się mogło, że lepszą miarą zmienności jest wzięcie średniej z odchyleń wszystkich danych od średniej, tj. $(\sum_{i=1}^{n}(x_i - \bar{x}))/n$. Jednak, nie jest trudno policzyć, że suma takich odchyleń zawsze jest równa 0. Dlatego, w podanym wzorze zamiast sumować $(x_i - \bar{x})$ sumuje się wartości bezwzględne odchyleń $|x_i - \bar{x}|$ i (dla próby) za miarę zmienności bierze się wielkość określoną wzorem

$$d = \frac{\sum_{i=1}^{n} |x - \bar{x}|}{n}, \tag{1.6}$$

którą nazywa się *odchyleniem przeciętnym*. My jednak będziemy zajmować się najczęściej używaną miarą zmienności, tj. *wariancją* oraz pierwiastkiem kwadratowym z wariancji zwanym *odchyleniem standardowym*.

Definicja 1.2.4. ***Wariancję*** *z próby $x_1, x_2, \ldots, x_n$ o średniej $\bar{x}$ oznaczamy symbolem*[17] *$\hat{s}^2$ i określamy wzorem*

$$\hat{s}^2 = \frac{(x_1 - \bar{x})^2 + (x_2 - \bar{x})^2 + \cdots + (x_n - \bar{x})^2}{n-1} = \frac{\sum_{i=1}^{n}(x_i - \bar{x})^2}{n-1}. \qquad (1.7)$$

Definicja 1.2.5. ***Odchylenie standardowe*** *z próby $x_1, x_2, \ldots, x_n$ o średniej $\bar{x}$ oznaczamy symbolem $\hat{s}$ i definiujemy jako pierwiastek kwadratowy z wariancji. Zatem $\hat{s} = \sqrt{\hat{s}^2}$ lub równoważnie*

$$\hat{s} = \sqrt{\frac{(x_1 - \bar{x})^2 + (x_2 - \bar{x})^2 + \cdots + (x_n - \bar{x})^2}{n-1}} = \sqrt{\frac{\sum_{i=1}^{n}(x_i - \bar{x})^2}{n-1}}. \qquad (1.8)$$

Dla większej ilości danych obliczanie wariancji i odchylenia standardowego jest uciążliwe, jeśli korzystamy z kalkulatora. Wtedy najlepiej skorzystać z któregoś z pakietów statystycznych. Jeśli liczba danych n nie jest zbyt duża i chcemy obliczyć te wielkości "ręcznie", to najlepiej skorzystać z następujących wzorów, które są równoważne odpowiednio wzorom (1.7) i (1.8), a wymagają mniej obliczeń.

$$\hat{s}^2 = \frac{n\left(\sum_{i=1}^{n} x_i^2\right) - \left(\sum_{i=1}^{n} x_i\right)^2}{n(n-1)}, \quad \hat{s} = \sqrt{\frac{n\left(\sum_{i=1}^{n} x_i^2\right) - \left(\sum_{i=1}^{n} x_i\right)^2}{n(n-1)}} \qquad (1.9)$$

Przykład 1.2.5. *Dla danych z Przykładu 1.2.4 obliczyć wariancję i odchylenie standardowe. Dla wygody przepiszmy te dane.*

A:	3	6	9	12	17	20	25	30	32	35	39	40	41	45	56
B:	13	13	14	15	20	21	26	30	31	34	36	37	38	40	42

Rozwiązanie:

Obliczenia wykonamy ze wzorów (1.7), (1.8) i (1.9) a wyniki umieścimy w tabelce.

[17] Dla zgodności z oznaczeniem odpowiedniego estymatora wariancji, który wprowadzimy później, stawiamy nad s daszek.

Zbiór A:					Zbiór B:			
Obs.	x_i	$x_i - \bar{x}$	$(x_i - \bar{x})^2$	x_i^2	y_i	$y_i - \bar{y}$	$(y_i - \bar{y})^2$	y_i^2
1	3	−24,3333	592,1111	9	13	−14,3333	205,4444	169
2	6	−21,3333	455,1111	36	13	−14,3333	205,4444	169
3	9	−18,3333	336,1111	81	14	−13,3333	177,7778	196
4	12	−15,3333	235,1111	144	15	−12,3333	152,1111	225
5	17	−10,3333	106,7778	289	20	−7,3333	53,7778	400
6	20	−7,3333	53,7778	400	21	−6,3333	40,1111	441
7	25	−2,3333	5,4444	625	26	−1,3333	1,7778	676
8	30	2,6667	7,1111	900	30	2,6667	7,1111	900
9	32	4,6667	21,7778	1024	31	3,6667	13,4444	961
10	35	7,6667	58,7778	1225	34	6,6667	44,4444	1156
11	39	11,6667	136,1111	1521	36	8,6667	75,1111	1296
12	40	12,6667	160,4444	1600	37	9,6667	93,4444	1369
13	41	13,6667	186,7778	1681	38	10,6667	113,7778	1444
14	45	17,6667	312,1111	2025	40	12,6667	160,4444	1600
15	56	28,6667	821,7778	3136	42	14,6667	215,1111	1764
Suma	410		3489,3333	14696	410		1559,3333	12766

Zatem zgodnie z wzorami, odpowiednio (1.7) i (1.8), mamy

$$\hat{s}_A^2 = \frac{3489{,}33333}{14} = 249{,}2381, \qquad \hat{s}_A = \sqrt{249{,}2381} = 15{,}7873,$$
$$\hat{s}_B^2 = \frac{1559{,}33333}{14} = 111{,}3810, \qquad \hat{s}_B = \sqrt{111{,}3810} = 10{,}5537.$$

Natomiast ze wzorów (1.9) otrzymujemy dokładnie te same wartości

$$\hat{s}_A^2 = \frac{15 \cdot 14696 - 410^2}{15 \cdot 14} = 249{,}2381, \qquad \hat{s}_A = \sqrt{249{,}2381} = 15{,}7873,$$
$$\hat{s}_B^2 = \frac{15 \cdot 12766 - 410^2}{15 \cdot 14} = 111{,}3810, \qquad \hat{s}_B = \sqrt{111{,}3810} = 10{,}5537.$$

Zwróćmy uwagę na fakt, że korzystając ze wzorów (1.9) zamiast (1.7) i (1.8) nie musimy wykonywać n odejmowań, a tylko jedno oraz dodatkowo wykonać dwa mnożenia i jedno potęgowanie. Ponadto, nie jest nam potrzebne obliczanie średniej i jeśli dane są liczbami całkowitymi, to wszystkie działania wykonujemy na liczbach całkowitych nawet, gdy średnia nie jest liczbą całkowitą. ▲

Uwaga 1.2.3. *We wzorach* (1.7) *i* (1.8) *mamy dzielenie przez* $n-1$*, a nie przez* n*. Wariancję z próby najczęściej obliczamy po to, aby użyć jej do oszacowania wariancji dla populacji, a nie dysponując wartością średniej* μ *dla populacji we wspomnianych wzorach, używamy średniej* $\bar{x}$ *dla próby. Próba zwykle też nie jest*

duża i w takich przypadkach wariancja w populacji zwykle będzie niedoszacowana, jeśli podzielimy sumę kwadratów odchyleń przez n. Dzieląc tę sumę przez $n-1$, otrzymamy trochę większe a przez to lepsze oszacowanie wariancji dla populacji[18)].

Zauważmy, że dokonując małych zmian we wzorach (1.9), będziemy mogli użyć ich do obliczania wariancji i odchylenia standardowego dla danych pogrupowanych. Dane pogrupowane, to dane, dla których mamy podane tylko przedziały, do których ich wartości należą oraz częstości, tj. liczby danych należących do poszczególnych przedziałów. W szczególności, danymi pogrupowanymi są dane, dla których podajemy tylko różne wartości zmiennej (cechy) x_i, $i = 1, \ldots, k$ oraz częstości n_i ich występowania w zbiorze danych. Wypiszmy też wzór na obliczenie średniej dla tego typu danych, choć nie jest on potrzebny do obliczenia wariancji i odchylenia standardowego (ale mając średnią, można ją w tym celu wykorzystać). Niech n oznacza sumę częstości, tj. $n = n_1 + \cdots + n_k$. Po dokonaniu odpowiednich zmian w (1.3) i (1.9) otrzymujemy

$$\overline{x} = \frac{\sum_{i=1}^{k} n_i x_i}{n}, \quad \hat{s}^2 = \frac{n\left(\sum_{i=1}^{k} n_i x_i^2\right) - \left(\sum_{i=1}^{k} n_i x_i\right)^2}{n(n-1)} = \frac{\left(\sum_{i=1}^{k} n_i x_i^2\right) - n\overline{x}^2}{n-1}, \quad \hat{s} = \sqrt{\hat{s}^2}. \tag{1.10}$$

Przykład 1.2.6. *Obliczyć średnią, wariancję i odchylenie standardowe dla pogrupowanych danych z Przykładu 1.1.3, biorąc wartości z pierwszej i trzeciej kolumny Tabeli 1.3 rozkładu częstości, tj. dla danych*

Klasa (x_i):	1	2	3	4	5	6
Częstość:	7	13	21	12	6	1

Rozwiązanie:

Korzystając z wyników w tabelce

Nr obserwacji	Klasy x_i	Częstości n_i	$n_i x_i$	$n_i x_i^2$
1	1	7	7	7
2	2	13	26	52
3	3	21	63	189
4	4	12	48	192
5	5	6	30	150
6	6	1	6	36
Suma		60	180	626

otrzymujemy

[18)] Precyzyjne uzasadnienie użycia $n-1$ zamiast n wymaga pojęcia *estymatora nieobciążonego*, które wprowadzimy później.

$$\bar{x} = \frac{180}{60} = 3, \qquad \hat{s}^2 = \frac{60 \cdot 626 - 180^2}{60 \cdot 59} = 1{,}4576, \qquad \hat{s} = \sqrt{1{,}4576} = 1{,}2073.$$

▲

Rozważmy teraz przykład dla danych pogrupowanych w przedziałach.

Przykład 1.2.7. *Obliczyć średnią, wariancję i odchylenie standardowe dla danych z Przykładu 1.1.1, korzystając z kolumny pierwszej i drugiej Tabeli 1.1 rozkładu częstości, tj. dla danych (P - przedziały, C - częstości)*

P:	$[100, 110)$	$[110, 120)$	$[120, 130)$	$[130, 140)$	$[140, 150)$	$[150, 160)$	$[160, 170)$
C:	6	14	18	13	5	3	1

Rozwiązanie:

Ponieważ teraz zamiast wartości zmiennej mamy tylko przedziały do których te wartości należą, to za wartości x_i badanej zmiennej (w tym przypadku wartości ciśnienia skurczowego) bierzemy środki podanych przedziałów. Dalsza część obliczeń jest dokładnie taka sama jak w poprzednim przykładzie.

Korzystając z wyników w tabelce poniżej, otrzymujemy: $\bar{x} = \frac{7600}{60} = 126{,}6667,$

Nr obserwacji	Wartości x_i	Częstości n_i	$n_i x_i$	$n_i x_i^2$
1	105	6	630	66150
2	115	14	1610	185150
3	125	18	2250	281250
4	135	13	1755	236925
5	145	5	725	105125
6	155	3	465	72075
7	165	1	165	27225
Suma		60	7600	973900

$$\hat{s}^2 = \frac{60 \cdot 973900 - 7600^2}{60 \cdot 59} = 190{,}3955, \qquad \hat{s} = \sqrt{190{,}3955} = 13{,}7984.$$

Pamiętajmy jednak, że mając dane z podanymi tylko przedziałami dla wartości zmiennej i częstościami, otrzymujemy tylko wartości **przybliżone** dla średniej, wariancji i odchylenia standardowego. Parametry te obliczamy w taki sposób, gdy nie mamy danych podanych tak jak w Przykładzie 1.1.1. Korzystając bowiem z wartości 60 pomiarów ciśnienia podanych w treści przykładu, otrzymamy prawdziwe wartości średniej, wariancji i odchylenia standardowego dla próby równe

$$\bar{x} = 125{,}6667, \qquad \hat{s}^2 = 179{,}2768, \qquad \hat{s} = \sqrt{179{,}2768} = 13{,}3894.$$ ▲

Podajmy teraz definicję **wariancji** i **odchylenia standardowego** dla populacji.

Definicja 1.2.6. ***Wariancję*** *populacji $x_1, x_2, \ldots, x_N$ o średniej μ oznaczamy symbolem σ^2 i określamy wzorem*

$$\sigma^2 = \frac{(x_1-\mu)^2 + (x_2-\mu)^2 + \cdots + (x_N-\mu)^2}{N} = \frac{\sum_{i=1}^{N}(x_i-\mu)^2}{N}. \qquad (1.11)$$

Definicja 1.2.7. ***Odchylenie standardowe*** *populacji $x_1, x_2, \ldots, x_N$ o średniej μ oznaczamy symbolem σ i definiujemy jako pierwiastek kwadratowy z wariancji, tj. $\sigma = \sqrt{\sigma^2}$ lub równoważnie wzorem*

$$\sigma = \sqrt{\frac{(x_1-\mu)^2 + (x_2-\mu)^2 + \cdots + (x_N-\mu)^2}{N}} = \sqrt{\frac{\sum_{i=1}^{N}(x_i-\mu)^2}{N}}. \qquad (1.12)$$

W przypadku danych pogrupowanych dla populacji możemy wyprowadzić odpowiedniki wzorów (1.10), które dla $N = n_1 + n_2 + \cdots + n_k$ będą miały postać

$$\mu = \frac{\sum_{i=1}^{k} n_i x_i}{N}, \quad \sigma^2 = \frac{N\left(\sum_{i=1}^{k} n_i x_i^2\right) - \left(\sum_{i=1}^{k} n_i x_i\right)^2}{N^2}, \quad \sigma = \sqrt{\sigma^2} \qquad (1.13)$$

Przykład 1.2.8. *W roku 2016 urodziło się w Polsce 382257 noworodków. Tabl. 78(108) na str. 294 Rocznika Demograficznego GUS 2017 zawiera następujące dane o wagach 382243*[19)] *noworodków, które cytujemy w poniższej tabelce.*

Przedział wagowy	Liczba urodzeń	Przedział wagowy	Liczba urodzeń
200 - 600	261	3000 - 3500	143999
600 - 1000	1012	3500 - 4000	119365
1000 - 1500	2124	4000 - 4500	35000
1500 - 2000	4649	4500 - 5000	4667
2000 - 2500	14268	5000 - 7000	427
2500 - 3000	56471		

Obliczyć średnią, wariancję i odchylenie standardowe.

Rozwiązanie:

Zbiorowość wszystkich noworodków urodzonych w 2016 roku możemy traktować jako populację. Aby można było zastosować wzory (1.13) trzeba wyznaczyć środki przedziałów wagowych podanych w tabelce i przyjąć je za wartości x_i. Ponieważ wagi noworodków powyżej 7000 g są rzadkością, a najniższa waga noworodka (dane z czasu pisania tego tekstu), który utrzymał się przy życiu wynosi 244 g, to wydaje się rozsądnym wziąć $[200, 600)$ jako pierwszy przedział i $[5000, 7000)$ jako ostatni przedział. Zatem korzystając ze wzorów (1.13), mamy

[19)] Dla 14 noworodków brak danych o ich wadze.

Nr obserwacji	Wartości x_i	Częstości n_i	$n_i x_i$	$n_i x_i^2$
1	400	261	104400	41760000
2	800	1012	809600	647680000
3	1250	2124	2655000	3318750000
4	1750	4649	8135750	14237562500
5	2250	14268	32103000	72231750000
6	2750	56471	155295250	427061937500
7	3250	143999	467996750	1520989437500
8	3750	119365	447618750	1678570312500
9	4250	35000	148750000	632187500000
10	4750	4667	22168250	105299187500
11	6000	427	2562000	15372000000
Suma		382243	1288198750	4469957877500

W rezultacie, ze wzorów (1.13) otrzymujemy

$$\mu = \frac{1288198750}{382243} \approx 3370{,}10,$$
$$\sigma^2 = \frac{382243 \cdot 4469957877500 - 1288198750^2}{382243^2} \approx 336419{,}05,$$
$$\sigma = \sqrt{336419{,}05} \approx 580{,}02.$$

▲

1.2.4. Współczynnik zmienności

Jeśli dwie próby reprezentują dwie zmienne typu ilościowego przyjmujące wartości liczbowe wyrażone w tych samych jednostkach, to można bezpośrednio porównywać obliczone dla nich wariancje czy też odchylenia standardowe i jeśli one są różne to mówić, że zmienność jednej zmiennej jest większa (lub mniejsza) od drugiej. Tak jest np., gdy jedną zmienną jest ciśnienie krwi skurczowe, a drugą rozkurczowe. Ale jeśli jedną zmienną jest, np. ciśnienie skurczowe, a drugą częstość tętna (puls), to bezpośrednie porównywanie ich wariancji nie ma sensu. Aby można było porównać zmienność w dwóch (lub więcej) zbiorach danych wprowadza się pojęcie **współczynnika zmienności**. Trzeba jednak dodać, że współczynnik ten jest wrażliwy na błędy wartości średniej.

Definicja 1.2.8. ***Współczynnik zmienności*** *dla próby $x_1, x_2, \ldots, x_n$ o średniej $\bar{x}$ oznaczamy symbolem V i określamy go wzorem*

$$V = \frac{\hat{s}}{\bar{x}} \qquad (\textit{lub w procentach:} \quad V = \frac{\hat{s}}{\bar{x}} \cdot 100\%). \tag{1.14}$$

Definicja 1.2.9. ***Współczynnik zmienności*** *dla populacji $x_1, x_2, \ldots, x_N$ o średniej μ oznaczamy symbolem υ i określamy go wzorem*

$$\upsilon = \frac{\sigma}{\mu} \qquad (\textit{lub w procentach:} \quad \upsilon = \frac{\sigma}{\mu} \cdot 100\%). \tag{1.15}$$

Przykład 1.2.9. *Pomiary pewnego parametru w grupie pacjentów obciążonych jedną z chorób dały średnią $\bar{x} = 53{,}85$ jednostek i odchylenie standardowe $\hat{s} = 12{,}37$ jednostek, a w grupie pacjentów nieobciążonych tą chorobą średnią $\bar{x} = 43{,}17$ jednostek i odchylenie standardowe $\hat{s} = 12{,}73$ jednostek. Obliczyć współczynniki zmienności dla obu grup i porównać je.*

Rozwiązanie:

Pierwsza grupa: $$V = \frac{12{,}37}{53{,}85} \cdot 100\% = 22{,}97\%.$$

Druga grupa: $$V = \frac{12{,}73}{43{,}17} \cdot 100\% = 29{,}49\%.$$

Różnica między współczynnikiem zmienności w grupie pacjentów nie obciążonych daną chorobą i w grupie pacjentów obciążonych tą chorobą wynosi 6,52%. ▲

Przykład 1.2.10. *W grupie pacjentów zmierzono poziom glukozy we krwi i na ich podstawie wyliczono średnią $\bar{x} = 104{,}59$ mg/dl i odchylenie standardowe $\hat{s} = 21{,}74$ mg/dl. Natomiast dla ich masy ciała średnia i odchylenie standardowe wyniosły odpowiednio $\bar{x} = 79{,}24$ kg i $\hat{s} = 7{,}22$ kg. Obliczyć współczynniki zmienności dla obu zmiennych i porównać je.*

Rozwiązanie:

Poziom glukozy: $$V = \frac{21{,}74}{104{,}59} \cdot 100\% = 20{,}79\%$$

Waga: $$V = \frac{7{,}22}{79{,}24} \cdot 100\% = 9{,}11\%$$

Różnica między współczynnikiem zmienności poziomu glukozy we krwi i współczynnikiem zmienności masy ciała w grupie pacjentów wynosi 11,68%. ▲

1.2.5. Miary pozycyjne

Jako miar pozycyjnych używamy kwantyli, percentyli[20)], decyli i kwartyli.

[20)] Percentyle w języku polskim nazywane są też centylami, szczególnie w kontekście siatek centylowych, np. siatka centylowa wysokości ciała chłopców w wieku 3–18 lat.

Definicja 1.2.10. *Dla zbioru danych reprezentujących wartości pewnej zmiennej (cechy) i liczby* $p \in (0,1)$[21] ***kwantylem rzędu*** $\mathbf{p}$ *nazywamy taką liczbę* $\mathbf{q_p}$*, że nie więcej niż* $100p\%$ *wartości danych jest mniejsza od* $\mathbf{q_p}$ *i nie więcej niż* $100(1-p)\%$ *danych jest większa od* $\mathbf{q_p}$.

Jest to jedna z definicji kwantyla jaką można spotkać. Zanim przejdziemy do przykładów, wyjaśnijmy dwie kwestie związane z powyższą definicją.

1. $100p\%$ danych, czyli liczba np elementów zbioru danych o liczebności n, gdzie $0 < p < 1$, na ogół nie jest liczbą całkowitą. Na przykład 70%, tj. 0,7 danych dla zbioru danych 3, 4, 5, 6, jest równe 2,8. Co miałoby oznaczać, że nie więcej niż 2,8 tych danych ma wartość mniejszą od danej liczby, np. 5? Wiemy co oznacza, że 2 wartości danych są mniejsze niż 5 - są to wartości 3 i 4. Dlatego, jeśli liczebność zbioru danych wynosi n i np nie jest liczbą całkowitą, to w powyższej definicji przez $100p\%$ liczby danych należy rozumieć liczbę danych równą zaokrągleniu w górę liczby np do najbliższej liczby całkowitej. Podobnie należy rozumieć $100(1-p)\%$ danych, gdy nie jest ona liczbą całkowitą.

2. Niestety, Definicja 1.2.10 nie określa jednoznacznie kwantyla danego rzędu. Biorąc przykład z czterema danymi z punktu 1 i $p = 0,7$, łatwo sprawdzić, że każda liczba z przedziału $[4,6)$ spełnia warunki bycia kwantylem $\mathbf{q_p}$ z Definicji 1.2.10. Dlatego można spotkać co najmniej 9 różnych definicji kwantyla wyznaczających jednoznacznie kwantyle q_p i wiele programów statystycznych pozwala wybierać za pomocą opcji odpowiednie definicje, a otrzymywane wartości domyślne kwantyli w różnych programach często są różne. Omówimy nieco dalej metodę wyznaczania kwantyli, które nie tylko spełniają warunki podanej definicji, ale są też wyznaczane jednoznacznie.

Właściwie, to percentyle, decyle i kwartyle możemy traktować jako szczególne przypadki kwantyli. Różnica pomiędzy kwantylami a percentylami polega na tym, że percentyle to są kwantyle, ale wyrażone w procentach zamiast w ułamkach z przedziału $[0,1]$. Tak więc kwantyl rzędu 0,37 to 37% percentyl. Decyle to 10%, 20%, ..., 90% percentyle. A kwartyle, to 25%, 50% i 75% percentyle. 25% percentyl nazywamy pierwszym kwartylem, 50% percentyl nazywamy drugim kwartylem, a 75% percentyl nazywamy trzecim kwartylem. Oznaczamy je odpowiednio symbolami Q_1, Q_2 i Q_3. Zauważmy też, że

$$\text{kwantyl rzędu } 0{,}5 = 50\% \text{ percentyl} = Q_2 = \textbf{mediana}.$$

[21] Nie wszystkie komputerowe programy statystyczne umożliwiają obliczanie kwantyli rzędu 0 i 1. Na przykład R i Excel umożliwiają, ale Statistica nie.

Jedna z metod wyznaczania kwantyla q_p dla zbioru danych $x_1, x_2, \ldots, x_n$ jest następująca (jest to metoda domyślna w programie Statistica):

Krok 1: Porządkujemy zbiór $x_1, x_2, \ldots, x_n$,
otrzymując ciąg niemalejący $x_{(1)}, x_{(2)}, \ldots, x_{(n)}$.

Krok 2: Obliczamy $c = n \cdot p$.

Krok 3: (a) Jeśli c nie jest liczbą całkowitą, to zaokrąglamy ją w górę do najbliższej liczby całkowitej. Oznaczmy otrzymaną w ten sposób liczbę symbolem $\lceil c \rceil$. Wtedy kwantyl q_p jest dany wzorem

$$q_p = x_{(\lceil c \rceil)}.$$

(b) Jeśli c jest liczbą całkowitą, to kwantyl q_p rzędu p jest dany wzorem

$$q_p = \frac{x_{(c)} + x_{(c+1)}}{2}.$$

Jeśli dla pewnego c okazałoby się, że $c = n$, to kładziemy w powyższym wzorze $x_{(c+1)} = x_{(c)}$.

Przykład 1.2.11. *Sumę nieobecności studentów na zajęciach ze statystyki w kolejnych miesiącach roku akademickiego daje poniższa tabelka. Wyznaczyć kwantyle rzędu* 0,05 *i* 0,95, 20%, 40% *i* 60% *percentyl oraz pierwszy, drugi i trzeci kwartyl dla tych danych.*

Miesiąc:	X	XI	XII	I	II	III	IV	V	VI
Nieobecności:	2	7	4	8	13	10	12	9	1

Rozwiązanie:

Krok 1: Porządkując dane od najmniejszej do największej, otrzymujemy

$$1, 2, 4, 7, 8, 9, 10, 12, 13.$$

Krok 2 i 3: Zamiast zapisywać c dla każdego p z osobna użyjmy wektora $p = (0{,}05, 0{,}20, \ldots, 0{,}95)$. Obliczamy $c = n \cdot p$ dla $n = 9$ i poszczególnych współrzędnych wektora p a wyniki umieszczamy w tabeli.

p:	0,05	0,20	0,25	0,40	0,50	0,60	0,75	0,95
c:	0,45	1,80	2,25	3,60	4,50	5,40	6,75	8,55
$\lceil c \rceil$:	1	2	3	4	5	6	7	9
$x_{(\lceil c \rceil)}$:	1	2	4	7	8	9	10	13

W trzecim wierszu tabelki wpisane są zaokrąglone w górę do liczb całkowitych wartości $c = n \cdot p$ z drugiego wiersza, a w czwartym wierszu mamy wartości elementów uporządkowanego zbioru z Kroku 1 o indeksach (numerach) wskazanych w trzecim wierszu. Zatem otrzymaliśmy

kwantyle: $q_{0,05} = 1, \quad q_{0,95} = 13,$

percentyle: $p_{20} = 2, \quad p_{40} = 7, \quad p_{60} = 9,$

kwartyle: $Q_1 = 4, \quad Q_2 = 8, \quad Q_3 = 10.$ ▲

Przykład 1.2.12. *Obliczyć* 50%, 60%, 70%, 80% *i* 90% *percentyle dla danych:*

Dane:	90	85	65	72	82	96	70	79	68	84
Dane uporządkowane $x_{(i)}$:	65	68	70	72	79	82	84	85	90	96

Rozwiązanie:

Wykonujemy te same kroki co w poprzednim przykładzie. Wyniki kroku 1 zostały umieszczone w tabelce z danymi. W kroku 2 zamieniamy procenty $100p\%$ występujące w $100p\%$ percentylach do obliczenia na setne p, tj. na prawdopodobieństwa, mnożąc je przez $1/100$. Obliczone p umieściliśmy w drugim wierszu tabeli. Następnie obliczamy $c = n \cdot p$ (dla $n = 10$ bo tyle mamy danych). Widzimy, że nie musimy zaokrąglać c, ponieważ są one całkowite. Wobec tego wykonujemy część (b) kroku 3 i wyniki wpisujemy w czwartym wierszu tabeli.

Percentyle do obliczenia:	50%	60%	70%	80%	90%
p:	0,50	0,60	0,70	0,80	0,90
c:	5	6	7	8	9
Obliczone percentyle $(x_{(c)} + x_{(c+1)})/2$:	80,5	83	84,5	87,5	93

▲

1.2.6. Uzupełnienia opisu statystycznego

Mając już zdefiniowane kwartyle, możemy teraz wrócić do miar rozproszenia i wprowadzić pojęcia rozstępu i odchylenia ćwiartkowego.

Definicja 1.2.11. ***Rozstępem ćwiartkowym*** *zwanym też* ***rozstępem międzykwartylowym*** *nazywamy różnicę pomiędzy trzecim a pierwszym kwartylem, tj. różnicę*

$$R_Q = Q_3 - Q_1. \tag{1.16}$$

Natomiast ***odchyleniem ćwiartkowym*** *zwanym też* ***odchyleniem kwartylowym*** *nazywamy połowę rozstępu ćwiartkowego, tj.*

$$Q = \frac{Q_3 - Q_1}{2} = \frac{R_Q}{2}. \tag{1.17}$$

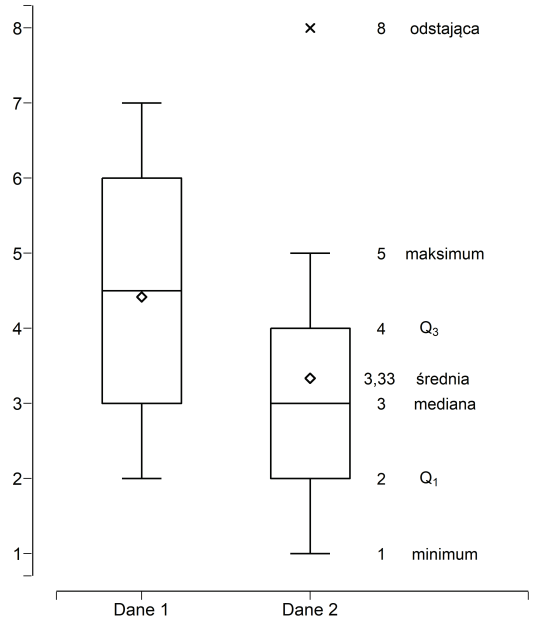
Przykład 1.2.13. *Dla danych z Przykładu 1.2.11 wyznaczyć rozstęp ćwiartkowy i odchylenie ćwiartkowe.*

Rozwiązanie:

$$R_Q = Q_3 - Q_1 = 10 - 4 = 6, \qquad Q = \frac{R_Q}{2} = \frac{6}{2} = 3. \qquad \blacktriangle$$

W statystyce, obserwacje znajdujące się w przedziale $[Q_1 - 3R_Q, Q_1 - 1{,}5R_Q)$ lub $(Q_3 + 1{,}5R_Q, Q_3 + 3R_Q]$ są uznawane jako elementy *umiarkowanie odstające*, a obserwacje znajdujące się na zewnątrz przedziału $[Q_1 - 3R_Q, Q_3 + 3R_Q]$ jako elementy *ekstremalnie odstające*.

Wykonując różne analizy danych z wykorzystaniem pakietów statystycznych, często otrzymujemy wykres **pudełkowy** nazywany też wykresem **pudełko–wąsy** lub wykresem **ramka–wąsy**, który w swojej standardowej wersji przedstawia tylko 5 informacji o zbiorze danych: wartość mediany, pierwszego i trzeciego kwartyla oraz wartość minimalną i maksymalną. Wartościami ponadstandardowymi mogą być średnia i obserwacje odstające. Mimo tak niewielu zawartych w takim wykresie informacji, potrafi on bardzo dużo powiedzieć o własnościach zbioru danych.



Rys. 1.11. Wykresy pudełko–wąsy. Na wykresie pudełkowym dla zbioru Dane 2 po jego prawej stronie wskazano pozycje wartości wszystkich statystyk[22)] oraz obserwacji odstającej wraz z ich wartościami

[22)] Średnia, mediana, kwartyle, wartość minimalna i maksymalna, jeśli pochodzą z próby losowej, nazywane są statystykami. Pojęcia próby losowej i statystyki wprowadzone będą w Rozdziale 3.

Na jego podstawie możemy zaobserwować poziom rozproszenia zbioru wartości obserwacji, sprawdzić czy zbiór ten jest symetryczny, zobaczyć jak bardzo oddalone od mediany są wartości minimalne i maksymalne. Na Rys. 1.11 mamy dwa wykresy pudełkowe otrzymane dla danych

Dane 1:	6	2	2	7	4	6	3	5	6	4	5	3
Dane 2:	4	3	1	2	8	2	1	4	5			

Na obu wykresach dodano symbol wskazujący położenie średniej, a do wykresu dla zbioru Dane 2 dodano symbol wskazujący położenie obserwacji odstającej, która nie była brana pod uwagę przy wyznaczaniu wartości maksymalnej dla tego zbioru danych.

Wróćmy teraz do wcześniej wspomnianych sposobów mierzenia skośności rozkładu zmiennej. Ponieważ istnieje więcej niż jedna miara skośności, to omówimy miarę, która jest najczęściej używana i którą wykorzystuje wiele pakietów statystycznych, (m.in. Excel, Statistica, SAS).

Definicja 1.2.12. *Dla zbioru danych $x_1, \ldots, x_n$* ***współczynnikiem skośności***[23)] *będziemy nazywać wielkość g określoną wzorem*

$$g = \frac{n\sum_{i=1}^{n}(x_i - \bar{x})^3}{(n-1)(n-2)\hat{s}^3}. \tag{1.18}$$

Uwaga 1.2.4. *Jeśli mamy do czynienia z danymi pogrupowanymi z liczbą klas k, częstościami $n_1, n_2, \ldots, n_k$ i $n = n_1 + \cdots + n_k$, to wzór* (1.18) *przyjmuje postać*

$$g = \frac{n\sum_{i=1}^{k} n_i(x_i - \bar{x})^3}{(n-1)(n-2)\hat{s}^3}. \tag{1.19}$$

Jeśli współczynnik skośności g jest równy 0, to zmienna, którą dane reprezentują, ma rozkład symetryczny (względem średniej), tj. dane te są rozłożone symetrycznie względem średniej. Jeśli $g < 0$, to rozkład zmiennej jest asymetryczny lewostronnie, a gdy $g > 0$, to rozkład zmiennej jest asymetryczny prawostronnie.

Przykład 1.2.14. *Rozpatrzmy dane dla konstrukcji wykresów na Rys. 1.10.*

	Wykres (A)						Wykres (B)						Wykres (C)				
Wartości x_i:	1	2	3	4	5	6	1	2	3	4	5	6	1	2	3	4	5
Częstości n_i:	1	3	5	6	9	6	6	9	6	5	2	2	3	5	9	5	3

Obliczyć dla nich współczynniki skośności.

[23)] Jest to tzw. *standaryzowany klasyczny współczynnik Fishera-Pearsona z poprawką na wielkość próby.*

Rozwiązanie:

Skorzystamy ze wzorów (1.10), aby obliczyć $\bar{x}$ i $\hat{s}$ oraz wzoru (1.19), aby obliczyć współczynnik skośności g. Wartości x_i i częstości n_i bierzemy z powyższej tabelki.

$$\bar{x}_A = 127/30 = 4{,}2333, \qquad \bar{x}_B = 84/30 = 2{,}8, \qquad \bar{x}_C = 75/25 = 3.$$

i	Wykres (A)			Wykres (B)			Wykres (C)		
	$n_i x_i$	$n_i x_i^2$	$n_i(x_i-\bar{x})^3$	$n_i x_i$	$n_i x_i^2$	$n_i(x_i-\bar{x})^3$	$n_i x_i$	$n_i x_i^2$	$n_i(x_i-\bar{x})^3$
1	1	1	−33,8027	6	6	−34,992	3	3	−24
2	6	12	−33,4181	18	36	−4,608	10	20	−5
3	15	45	−9,3802	18	54	0,048	27	81	0
4	24	96	−0,0762	20	80	8,640	20	80	5
5	45	225	4,0557	10	50	21,296	15	75	24
6	36	216	33,0838	12	72	65,536			
Suma	127	595	−39,5378	84	298	55,920	75	259	0

$$\hat{s}_A = \sqrt{\frac{30\cdot 595 - 127^2}{30\cdot 29}} = 1{,}4065, \qquad g_A = \frac{30\cdot(-39{,}5378)}{29\cdot 28\cdot 1{,}4065^3} = -0{,}5250,$$

$$\hat{s}_B = \sqrt{\frac{30\cdot 298 - 84^2}{30\cdot 29}} = 1{,}4716, \qquad g_B = \frac{30\cdot 55{,}920}{29\cdot 28\cdot 1{,}4716^3} = 0{,}6483,$$

$$\hat{s}_C = \sqrt{\frac{25\cdot 259 - 75^2}{25\cdot 24}} = 1{,}1902, \qquad g_C = \frac{25\cdot 0}{24\cdot 23\cdot 1{,}1902^3} = 0.$$

Zatem obliczone współczynniki skośności potwierdzają, że dane odpowiadające wykresowi (A) mają rozkład asymetryczny lewostronnie, a dane odpowiadające wykresowi (B) mają rozkład asymetryczny prawostronnie. Choć, biorąc pod uwagę rozmiary tych zbiorów danych, nie są to rozkłady mocno asymetryczne. Natomiast dane odpowiadające wykresowi (C) mają rozkład symetryczny. ▲

Inną miarą kształtu rozkładu, która mierzy spłaszczenie rozkładu, jest kurtoza.

Definicja 1.2.13. *Dla zbioru danych $x_1, \ldots, x_n$* ***kurtozę*** *k zwaną również* ***współczynnikiem ekscesu***[24)] *definiujemy wzorem*

$$k = \frac{n(n+1)\sum_{i=1}^{n}(x_i-\bar{x})^4 - 3(n-1)\left(\sum_{i=1}^{n}(x_i-\bar{x})^2\right)^2}{(n-1)(n-2)(n-3)\hat{s}^4}. \tag{1.20}$$

[24)] *Excess* w języku *ang.* to więcej, bardziej (niż ...), tutaj: bardziej spłaszczony w porównaniu z rozkładem normalnym, o którym będzie mowa w dalszej części tekstu. Część autorów rozróżnia pojęcie kurtozy i współczynnika ekscesu. Wtedy różnica pomiędzy kurtozą a współczynnikiem ekscesu wynosi 3, a kurtoza zmiennej losowej o rozkładzie normalnym jest równa 3. Przy naszej definicji zmienna losowa o rozkładzie normalnym ma kurtozę równą 0. Korzystając ze statystycznych programów komputerowych, trzeba zwracać uwagę na to, która definicja jest użyta.

Uwaga 1.2.5. *Jeśli mamy do czynienia z danymi pogrupowanymi z liczbą klas j, częstościami n_1, n_2, ..., n_j i $n = n_1 + \cdots + n_j$, to wzór* (1.20) *przyjmuje postać*

$$k = \frac{n(n+1)\sum_{i=1}^{j} n_i(x_i - \bar{x})^4 - 3(n-1)\left(\sum_{i=1}^{j} n_i(x_i - \bar{x})^2\right)^2}{(n-1)(n-2)(n-3)\hat{s}^4}. \qquad (1.21)$$

Rozkłady bardziej spłaszczone niż rozkład normalny mają kurtozę $k < 0$, a mniej spłaszczone niż rozkład normalny mają kurtozę $k > 0$.

Przykład 1.2.15. *Obliczyć kurtozę dla danych odpowiadających wykresowi (C) na Rysunku 1.10.*

Rozwiązanie:

Średnią obliczyliśmy w Przykładzie 1.2.14 i wyniosła ona $\bar{x} = 3$.

Nr obserwacji	Wartości x_i	Częstości n_i	$n_i(x_i - \bar{x})^2$	$n_i(x_i - \bar{x})^4$
1	1	3	12	48
2	2	5	5	5
3	3	9	0	0
4	4	5	5	5
5	5	3	12	48
Suma			34	106

Podstawiając do wzoru (1.21) otrzymane w tabelce sumy oraz obliczone ze wzoru (1.10) $\hat{s} = 1{,}1902$, otrzymamy

$$k = \frac{25 \cdot 26 \cdot 106 - 3 \cdot 24 \cdot 34^2}{24 \cdot 23 \cdot 22 \cdot 1{,}1902^4} = -0{,}5881. \qquad \blacktriangle$$

Przykład 1.2.16. *Obliczyć kurtozę dla danych z Przykładu 1.1.1.*

Rozwiązanie:

Ponieważ mamy 60 danych to do obliczenia kurtozy najlepiej użyć któregoś z pakietów statystycznych lub przynajmniej wykonać obliczenia pośrednie za pomocą innego programu komputerowego. W tym drugim przypadku otrzymamy:

$$\sum_{i=1}^{n}(x_i - \bar{x})^2 = 10577{,}3333, \quad \sum_{i=1}^{n}(x_i - \bar{x})^4 = 5957767{,}1111, \quad \hat{s} = 13{,}3894.$$

Podstawiając powyższe rezultaty do wzoru (1.21), otrzymujemy

$$k = \frac{60 \cdot 61 \cdot 5957767{,}1111 - 3 \cdot 59 \cdot 10577{,}3333^2}{59 \cdot 58 \cdot 57 \cdot 13{,}3894^4} = 0{,}3195. \qquad \blacktriangle$$

1.2.7. Współczynnik korelacji (Pearsona)

Mając dwa równoliczne zbiory obserwacji dwóch różnych cech, często interesuje nas czy istnieje jakaś zależność pomiędzy tymi dwiema cechami. Mając na przykład wagi ciał grupy ludzi i wysokość ich ciśnienia skurczowego, może nas interesować czy wraz ze wzrostem wagi wzrasta również ciśnienie krwi lub czy wyższym masom ciała odpowiada wyższy poziom cholesterolu.

Jeśli jedną interesującą nas zmienną (cechę) oznaczymy literą X, a drugą zmienną literą Y, to mając dane (realizacje zmiennej X) $x_1, \ldots, x_n$ i dane (realizacje zmiennej Y) $y_1, \ldots, y_n$, automatycznie mamy zbiór par $(x_1, y_1), \ldots, (x_n, y_n)$ i odpowiadające im punkty na płaszczyźnie $x0y$, które możemy narysować, otrzymując tzw. wykres punktowy. Taki wykres może nam wskazywać liniową zależność pomiędzy obiema zmiennymi, jeśli punkty te będą skupiać się wzdłuż pewnej prostej. Ta zależność nie musi być liniowa, a może być kwadratowa, jeśli punkty te będą skupiać się wzdłuż paraboli. Innego rodzaju zależności też mogą występować.

Liniową zależność można badać, używając **współczynnika korelacji**.

Definicja 1.2.14. *Oznaczmy symbolami $\hat{s}_x$ i $\hat{s}_y$ odpowiednio odchylenie standardowe dla danych $x_1, x_2, \ldots, x_n$ reprezentujących jedną cechę (zmienną) i odchylenie standardowe dla danych $y_1, y_2, \ldots, y_n$ reprezentujących drugą cechę (zmienną). Wtedy dla danych $x_1, x_2, \ldots, x_n$ i $y_1, y_2, \ldots, y_n$* ***współczynnik korelacji (Pearsona)*** *definiujemy wzorem*

$$r = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{(n-1)\hat{s}_x\hat{s}_y} = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_{i=1}^{n}(x_i - \bar{x})^2 \sum_{i=1}^{n}(y_i - \bar{y})^2}}. \tag{1.22}$$

Gdy $r > 0$, to mówimy, że zmienne są **dodatnio skorelowane**, a gdy $r < 0$, to mówimy, że zmienne są **ujemnie skorelowane**.

Występującą w definicji współczynnika korelacji r wielkość

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

którą oznaczyliśmy symbolem $\operatorname{cov}(x, y)$, nazywamy **kowariancją** zmiennych x i y. Tak więc współczynnik korelacji liniowej r możemy zapisać też w postaci

$$r = \frac{\operatorname{cov}(x, y)}{\hat{s}_x\hat{s}_y}.$$

Współczynnik korelacji ma następujące własności:

1. $$r \in [-1, 1].$$
2. Jeśli $r > 0$, to większym wartościom jednej cechy odpowiadają (średnio) większe wartości drugiej cechy.

3. Jeśli $r < 0$, to większym wartościom jednej cechy odpowiadają (średnio) mniejsze wartości drugiej cechy.

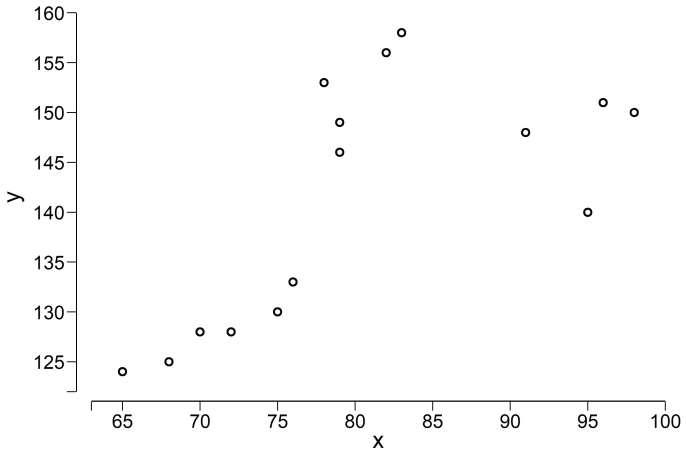
Uwaga 1.2.6. *Współczynnik korelacji Pearsona można obliczać tylko dla danych typu ilościowego. Dlatego istnieją jeszcze inne współczynniki korelacji, które można obliczać dla danych typu porządkowego, np. współczynnik korelacji Spearmana czy Kendalla. Aby policzyć współczynnik korelacji Spearmana, wystarczy policzyć rangi*[25] *danych* $x_1, x_2, \ldots, x_n$ *i* $y_1, y_2, \ldots, y_n$ *i we wzorze* (1.22) *dane te zastąpić ich rangami*[26]*.*

Przykład 1.2.17. *15 pacjentów zważono oraz zmierzono im skurczowe ciśnienie tętnicze. Ich wagi ciała w kg i ciśnienie skurczowe w mm Hg przedstawia tabelka.*

Waga:	75	76	70	96	79	95	91	68	78	72	79	83	98	82	65
Ciśnienie:	130	133	128	151	146	140	148	125	153	128	149	158	150	156	124

Narysować wykres punktowy zależności między wagą ciała a ciśnieniem skurczowym i obliczyć współczynnik korelacji pomiędzy tymi zmiennymi.

Rozwiązanie:



Rys. 1.12. Wykres punktowy zależności pomiędzy zmienną x reprezentującą wagę ciała a zmienną y reprezentującą ciśnienie skurczowe

Patrząc na wykres na Rys. 1.12, nie jest łatwo zauważyć liniową zależność pomiędzy wagą ciała a ciśnieniem skurczowym. Nie możemy jednak z góry odrzucić

[25] O rangach będziemy mówić później, omawiając testy hipotez.

[26] Gdy w żadnym z dwóch zbiorów danych nie ma jednakowych rang, stosowany do nich wzór (1.22) przekształca się do prostszego, często podawanego w podręcznikach, wzoru na współczynnik korelacji Spearmana postaci $r_S = 1 - (6 \sum_{i=1}^{n} d_i^2)/(n(n^2-1))$, gdzie $d_i = \text{ranga}(x_i) - \text{ranga}(y_i)$.

liniowej zależności pomiędzy tymi dwiema zmiennymi.
Policzmy więc współczynnik korelacji Pearsona. Aby go wyznaczyć potrzebne są nam średnie $\bar{x}$ i $\bar{y}$. Odpowiednie sumy do ich wyznaczenia znajdują się w poniższej tabeli.

Obs.	x_i	$x_i - \bar{x}$	$(x_i - \bar{x})^2$	y_i	$y_i - \bar{y}$	$(y_i - \bar{y})^2$	$(x_i - \bar{x})(y_i - \bar{y})$
1	75	−5,4667	29,8844	130	−11,2667	126,9378	61,5911
2	76	−4,4667	19,9511	133	−8,2667	68,3378	36,9244
3	70	−10,4667	109,5511	128	−13,2667	176,0044	138,8578
4	96	15,5333	241,2844	151	9,7333	94,7378	151,1911
5	79	−1,4667	2,1511	146	4,7333	22,4044	−6,9422
6	95	14,5333	211,2178	140	−1,2667	1,6044	−18,4089
7	91	10,5333	110,9511	148	6,7333	45,3378	70,9244
8	68	−12,4667	155,4178	125	−16,2667	264,6044	202,7911
9	78	−2,4667	6,0844	153	11,7333	137,6711	−28,9422
10	72	−8,4667	71,6844	128	−13,2667	176,0044	112,3244
11	79	−1,4667	2,1511	149	7,7333	59,8044	−11,3422
12	83	2,5333	6,4178	158	16,7333	280,0044	42,3911
13	98	17,5333	307,4178	150	8,7333	76,2711	153,1244
14	82	1,5333	2,3511	156	14,7333	217,0711	22,5911
15	65	−15,4667	239,2178	124	−17,2667	298,1378	267,0578
Suma	1207		1515,7333	2119		2044,9333	1194,1333

$$\bar{x} = \frac{1207}{15} = 80{,}4667, \qquad \bar{y} = \frac{2119}{15} = 141{,}2667.$$

Zatem
$$r = \frac{1194{,}1333}{\sqrt{1515{,}7333 \cdot 2044{,}9333}} = 0{,}6783.$$

Obliczony współczynnik korelacji wskazuje na większą niż umiarkowaną liniową zależność (dodatnią korelację) pomiędzy masą ciała a ciśnieniem skurczowym w badanej grupie pacjentów. ▲

Indeks